Causal Inference for Social and Engineering Systems

by

Anish Agarwal

B.S., M.S. in Computer Science from California Institute of Technology, June 2015

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

May 2022

Signature of Author: ______________________________

Anish Agarwal
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by: ______________________________

Devavrat Shah
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: ______________________________

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Causal Inference for Social and Engineering Systems

by Anish Agarwal

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

What will happen to Y if we do A?

A variety of meaningful social and engineering questions can be formulated this way: What will happen to a patient's health if they are given a new therapy? What will happen to a country's economy if policy-makers legislate a new tax? What will happen to a data center's latency if a new congestion control protocol is used? We explore how to answer such counterfactual questions using observational data—which is increasingly available due to digitization and pervasive sensors—and/or very limited experimental data. The two key challenges are: (i) counterfactual prediction in the presence of latent confounders; (ii) estimation with modern datasets which are high-dimensional, noisy, and sparse.

The key framework we introduce is connecting causal inference with tensor completion. In particular, we represent the various potential outcomes (i.e., counterfactuals) of interest through an order-3 tensor. The key theoretical results presented are: (i) Formal identification results establishing under what missingness patterns, latent confounding, and structure on the tensor is recovery of unobserved potential outcomes possible. (ii) Introducing novel estimators to recover these unobserved potential outcomes and proving they are finite-sample consistent and asymptotically normal.

Finally, we discuss connections between matrix/tensor completion and time series analysis; we believe this could serve as a basis to do counterfactual forecasting.

Thesis Supervisor: Devavrat Shah
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

"Work for someone you want to emulate 20 years from now".

As I struggled immensely with what to do with myself in the years during and right after undergrad, someone wisely gave me the advice above. And I must say each of the mentors I have had during my PhD has more than fulfilled this role. First and foremost, I have to thank my PhD advisors, Alberto Abadie, Munther Dahleh, and Devavrat Shah, who have all made this PhD such an immensely pleasurable experience *and* have still managed to push me to grow intellectually and emotionally. I've come to appreciate doing both at the same time is a rare trait in a mentor—one I very much hope to emulate in my career.

I would have to thank Alberto even if I never had gotten the opportunity to meet him, as his idea of synthetic controls has been foundational to so much of the work I have done in my thesis. It would not be a stretch to describe synthetic controls as simple, elegant, and enormously influential—having gotten to know Alberto over the past few years, it would be even less of a stretch to describe him with the same words. Further, the unconditional support and encouragement Alberto always gives me, has instilled in me the confidence to take on the challenge of doing research at the intersection of EECS and Economics, and bridging these two communities. I'm greatly looking forward to working together with him on many research problems in the coming years, and having him as a lifelong mentor. Most importantly, I hope Alberto (and his family) and I get to ski together soon!

Munzer has always been an unwavering source of calm and good humor right from the start of my thesis. No matter how unnecessarily stressed I got due to the ebbs and flows of the PhD journey, my weekly jaunts to Munzer's office invariably centered me. We discussed everything ranging from the nuances of a paper, to where machine learning was going to be in the next decade, and even gossip about MIT and world politics (which many a times felt one and the same). I especially have to thank Munzer as my first breakthrough with research came from a prescient idea that he had—while sitting at at airport terminal—about how does one design marketplaces for data. He gave me just enough rope to be creative and run with it without losing my way; that experience instilled in me the confidence to do independent research. A major part of Munzer's life at MIT the past 5 years has been the creation and running of IDSS, an incredibly forward-thinking and ambitious initiative that I've had the privilege to be a part of. I can safely say I would not have been able to write a thesis titled "causal inference for social and engineering systems" if not for the intellectual vision that Munzer had for IDSS, and the amazing set of faculty and students he brought together as part of it.

It is very difficult for me to convey in words just how grateful I am to Devavrat. I greatly admire him both as a researcher and as a human being. Intellectually, I've never met anyone who so broadly, yet intimately, sees the entire workflow from the germination of a theoretical idea all the way to its manifestation as a practical product used at a company. It simply leaves me in awe and has fundamentally defined my taste in research questions. As a human being, his energy and optimism are both boundless and infectious, as is the thought he puts into the well-being of his PhD advisees. I came to MIT 100% certain that I did not intend to pursue an academic career. However, the relationship I have had with him—and I'm sure many of his other advisees have had too—left me with a strong desire to have the same experience with students of my own. I can safely say this is the biggest reason why I've had a change of heart and now plan to pursue academia. The fondest memories of my PhD have been my squash games with him, and the non-fat, medium-sized, lattes from Vester he used to treat me to after (a large-sized latte if we then had to work on a proof together). The seed of almost any good idea I had during graduate school was planted during these weekly outings. At this point, I consider Devavrat not just my PhD advisor, but a close friend and confidant. I'm certain I will be bugging him for advice, squash games, and lattes for many decades to come.

Two other mentors I have to thank are Vishal Misra and Vasilis Syrgkanis. Vishal, like Devavrat, has a fantastic knack for identifying good research problems and how it translates to practical impact. Also during my job market search, he has been a constant source of support and the next step in my academic career is in no small part due to him. For this I will always be grateful. I spent a wonderful summer at Microsoft Research with Vasilis—I don't think I have met anyone who thinks quite as quickly as him nor anyone who has an encyclopedic knowledge of so many different fields. The project I worked on with him during the summer remains one of my favorite results from my PhD. I also have to thank him for agreeing to be in my thesis committee, despite being in Greece. I hope I get a chance to collaborate and learn from him over the coming years.

"You're the average of the five people closest to you".

This is another saying that resonates greatly with me and there are indeed five people who have been an intimate part of my PhD journey. On the research side, this thesis could not have been completed without Abdullah Alomar, Dennis Shen, and Rahul Singh. I met Abdullah when he was a master's student at MIT and we've worked on numerous projects together. The final chapter of this thesis was written in close collaboration with him. Seeing him grow both intellectually and personally over the past few years has been a privilege to be part of, and the day he got into the PhD program at MIT was a very proud day for me too. Dennis is my academic brother and I feel very fortunate to have gone through all the highs and lows of the PhD with him. Indeed, three of the four chapters in this thesis are as much his work as they are mine. I am certain we will be lifelong collaborators and friends. Rahul is not only one of my closest academic collaborators but more importantly one of my closest friends and confidants. Our trips to Oaxaca, Vancouver, and Singapore are some of my favorite memories from the PhD. He was also patient enough to teach me and get me excited about causal inference despite my best efforts to the contrary. For this I am incredibly grateful and certain he will have a similar impact on many students of his own in the coming years—I do sincerely hope one day he stops teasing me for once calling causal inference "philosophical hogwash".

On the personal side, I have to thank my two flatmates Lorenzo Masoero and Fraser Macdonald. First of all, I could not have asked for two better people to have gotten through the worst of Covid with together. Our evening soirees consisting of making dinner and watching the Wes Anderson canon made the pandemic as pleasant as it could have been. I started the PhD journey with Lorenzo, whom I met during our visit days. I would not have gotten through the interminable first year of the PhD without him, which was rife with the never-ending problem sets and existential angst of whether we made a mistake by going to graduate school. Lorenzo's single-minded pursuit of excellence and thoughtfulness in whatever he does, be it skiing, sailing, music, food, building community around him (and Bayesian nonparametrics) is unparalleled, and something I will always look up to. My friendship with Fraser was an unlikely one. We came from different worlds and us initially living together was more convenience than intention. However, fast-forward to five years later, I cannot imagine not having him as a friend. I've not met anyone who so genuinely shares in the highs and lows of life with the people closest to him. Every night I looked forward to reveling together in our living room, celebrating the small wins of the day over a beer, amongst other things. Over the years, I've learned that Fraser's capacity to find incredible wit and humor in any subject is only matched by his ability to inhale a bag of potato chips at 2am in the morning. I must also thank Fraser for introducing me to Erika, his now fiancé, who herself has become a dear friend. Getting to see their relationship beautifully develop over the course of my PhD has been a joy.

"It takes a village".

This thesis was inspired in so many different ways by the larger community at MIT and more broadly speaking. Devavrat and Munzer's research group were great fun to be a part of. The numerous dinners, weekly research talks, and the casual chats in 32D-666 will remain very fond memories. I also have to the thank the MIT econometrics group. In particular, Victor Chernozhukov, Anna Mikusheva, and Whitney Newey for welcoming me as a member despite my coming from a different department. The weekly metrics student lunches is a wonderful tradition that I hope continues for a long time to come. I also have to thank my many, many collaborators: Mohammad Alizadeh, Arwa Al-Anwary, Varkey Alumootil, Jehangir Amjad, Romain Cosson, Jessy Han, Pouya Hamadian, Thibaut Horel, Arash, Nasr-Esfahany, Maryann Rui, Tuhin Sarkar, Dogyoon Song, Dylan Sleeper, Chandler Squires, Andy Tsai, Caroline Uhler, Madeline Wong, Zhi Xu, and Cindy Yang. I learned something from each and everyone one of you. I would be remissed if I did not thank the LIDS/IDSS staff who made the place run remarkably efficiently; a special thank you to Lynne Dell, Alina Man, and Brian Jones.

Lastly, I have to thank those dearest to me. Chloe, you've been through this entire PhD journey with me putting up with all of my idiosyncrasies. You did so despite me telling you I'm moving from California to Boston 3 weeks before I left, even though we had only dated for a few months at that point. I'm very proud of how we've stuck together and built a community around us despite all the uncertainty that is part and parcel of one's late twenties—not to mention the unique and considerable challenge we faced of being long-distance through all of it. Due to you, I ran a marathon, learned how to ski and sail, and had countless travels adventures from Lake Tahoe to Sri Lanka during my PhD. You're the most loving, caring, hard-working person I know. I love you. I have to also thank Tauji, Amma, Bua, and Fufaji. You've been my family in the U.S. from the moment I stepped foot in this country 12 years ago. Tauji, the amount you sacrifice and care for your family is truly special. It was wonderful getting to live with you in Newton to both begin and end my PhD.

To end, I have to thank my brother Abhi and my parents. Abhi, it has been such a pleasure seeing you grow in your time in the U.S., and somehow also end up doing a PhD in statistics and machine learning. I've never met someone who works as hard or has such a single-minded devotion to the problem at hand, and I cannot wait to see what you do next. I'm incredibly proud of you. To my parents, I am immeasurably indebted to you. Despite how headstrong you know I can be, this entire journey could not have happened without your constant guidance and support. To my dad, thank you for being such a solid rock in my life. There is no other person I trust as much to help me get through a problem I am having, be it ensuring I have the right visas to not be in the U.S. illegally, to long-term advice on career and family. To my mom, you are my closest confidant and my best friend. Whenever anything of note happens, be it is good or bad, you are the first person I want to tell. During my PhD, the gush of love and affection I got from you both during my yearly pilgrimages back home to Singapore every December always replenished me, and was absolutely vital for me to take on the year ahead—I also have to thank Mary for being a member of our Singapore family all these years. Mom, dad, you both went through innumerable challenges immigrating to Singapore to give both Abhi and me the best opportunities to succeed. This PhD in some sense is an important stop along this multi-generational journey we are on as a family.

Contents

List of Figures

Chapter 1

Introduction

The aim of this thesis is to lay the theoretical foundations for the design of systems that make automated data-driven decisions via counterfactual reasoning, i.e., *what will happen to Y if we do A?* A variety of meaningful socio-economic and engineering questions can be formulated this way. To name a few: What will happen to a country's economy if policy-makers legislate a new tax? What will happen to a patient's health if they are given a new therapy? What will happen to a company's revenue if a new discount is introduced? What will happen to a data center's latency if a new congestion control protocol is used?

Challenge: counterfactual reasoning without building simulators or experimentation. Traditional approaches to do counterfactual reasoning require either building complex models/simulators of the system of interest (e.g., "Gazebo" in robotics, "NS-3" in networking, "Spice" in circuits) or doing extensive experimentation (e.g., clinical trials in medicine, A/B testing in e-commerce). However, as the complexity of such systems and the level of personalization now required grows, building precise simulators and doing experimentation are both becoming increasingly infeasible. For example, running a randomized control trial (RCT) to identify a personalized therapy for a patient sub-population is prohibitively expensive and time consuming. As a further example, to evaluate the efficacy of new congestion control policies for modern data centers, both building reliable simulators and running experiments is very challenging. In other cases, such as evaluating the effect of a social policy on a geographic region, collecting experimental data might be entirely infeasible. Thus, our goal is to build counterfactual decision-making systems using *observational data*, which is far more readily available (and/or extremely limited experimental data). The challenge is that building counterfactual models from observational data requires carefully thinking about *confounding*, i.e., hidden correlation between which decisions are chosen and the outcome of interest. See Figure 1.1 below for a classic example of how unobserved confounding can derail learning causal relationships.

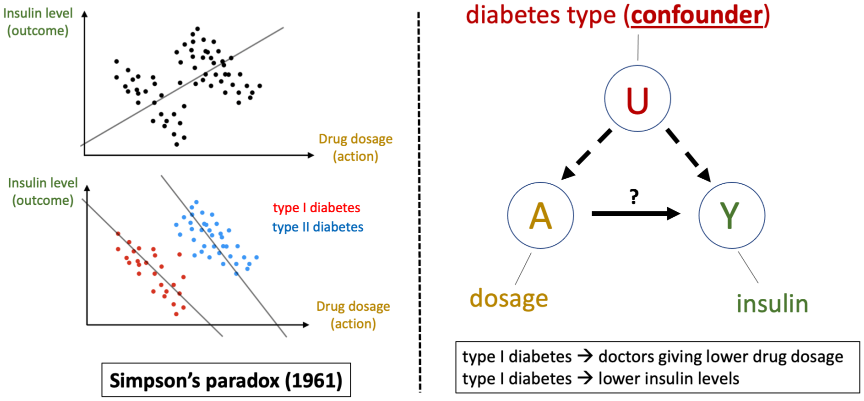


Figure 1.1: Imagine a dataset of patients who have diabetes and their insulin level after different drug dosages. A naive regression would lead to the conclusion that the higher the drug dosage, the higher the insulin level after, which would be absurd. However, after clustering data by patients who have Type I and II diabetes, we reach the opposite conclusion for *both* clusters. This phenomena is classically known as Simpson's paradox. Here diabetes type is a hidden confounder.

Indeed, addressing confounding is at the heart of learning a causal relationship between two entities. In addition to confounding, modern datasets are inherently *high-dimensional, noisy, and sparse*, which adds to the statistical and engineering challenge of building reliable models.

Opportunity: we are collecting observational data at an unprecedented scale. Due to digitization and pervasive sensors, we are collecting observational data at an exponentially increasing pace. Just through a smartphone, companies have detailed data of the purchasing and engagement behavior of their customers, under a variety of contexts. As a further example, companies now collect detailed network trace log data across geographically distributed data centers applying different congestion control protocols. Such data can be viewed as a large collection of *natural experiments* concurrently occurring. Thus, a major opportunity towards counterfactual decision-making is to design methods that *share information* across these natural experiments, and carefully deal with the challenges of confounding, high-dimensionality, noise, and sparsity in the associated datasets. Proposing novel frameworks and methods to do so is the focus of this thesis.

■ 1.1 How to Share Information Across (Natural) Experiments?

Potential outcomes viewed through a tensor. Causal inference is the study of recovering unobserved "potential outcomes"—the counterfactual of what would potentially have happened if an intervention had occurred—using observational and experimental data. One way to represent potential outcomes is through a matrix, where rows index different units (e.g., individuals, sub-populations, geographic regions) and columns index different interventions (e.g., discounts, health therapies, socio-economic policies). If we collect multiple measurements of units, say over time, then the various potential outcomes can be represented through an order-3 tensor across units, interventions and measurements. See Figure 1.2 for a visualization of this potential outcomes tensor. If we could fully observe this potential outcomes matrix/tensor, we would of course be able to effectively do counterfactual decision-making. However, without the ability to do extensive experimentation and build a reliable model of the system of interest, we only observe a sparse, noisy, subset of the entries of this tensor using observational data. See Figure 1.3 of typical sparsity patterns that show up in practice.

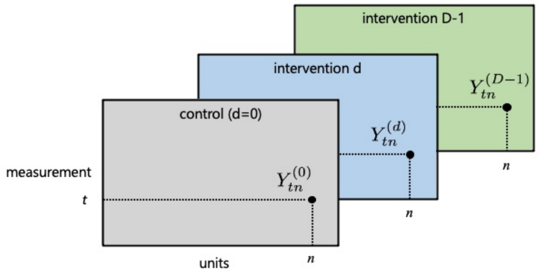
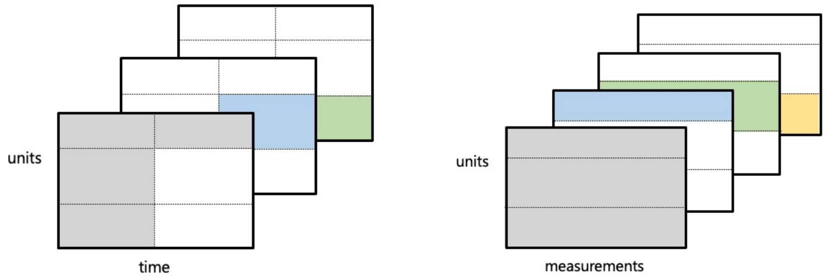


Figure 1.2: Potential outcomes tensor.



(a) Panel data setting.

(b) Data-efficient RCT.

Figure 1.3: Observations are represented by colored blocks while unobservable counterfactuals are represented by white blocks.

Matrix/tensor completion. Matrix/tensor completion is exactly the study of recovering an underlying matrix/tensor from a sparse subset of noisy observations. Traditionally, estimators for such problems assume entries of the matrix/tensor are "missing completely at random" (MCAR), i.e., each entry is revealed at random, independent of everything else, with uniform probability. This is likely unrealistic due to the presence of latent confounders, i.e., unobserved factors that determine both the entries of the underlying matrix and the missingness pattern in the observed matrix. In general, these confounders yield "missing not at random" (MNAR) data. Thus, in this thesis we focus on designing methods that do matrix/tensor completion with MNAR data, and study under what sparsity patterns, latent confounding, and structure on the matrix/tensor (e.g. low-rank) is recovery of unobserved potential outcomes possible. Indeed, it is both commonly said that

"Causal inference is a missing data problem."
&
"Matrix/tensor completion is a missing data problem."

Latent factor model for matrix/tensor completion. Let $[Y_{tn}^{(d)}] \in \mathbb{R}^{T \times N \times D}$ denote the collection of potential outcomes for N units, T measurements, and D interventions. The key assumption made in this thesis is that this collection of potential outcomes admits the following decomposition:

$$Y_{tn}^{(d)} = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell} \lambda_{d\ell} + \varepsilon_{tn}^{(d)}, \tag{1.1}$$

where r is the canonical polyadic (CP) rank, $u_t, v_n, \lambda_d \in \mathbb{R}^r$ are latent factors associated with the t-th measurement, n-th unit, and d-th intervention, respectively, and $\varepsilon_{tn}^{(d)}$ is a mean zero residual term. These latent factors are essentially the hidden structure across units, measurements, and interventions; the CP rank r implicitly captures the level of diversity across these dimensions. We note that such a factorization always exists, but the key assumption we make is that the CP rank r is much smaller than N, T, D. This latent structure in the data is exactly what will allow us to pool information across different natural experiments to answer counterfactuals for a given unit under an unseen intervention. In Chapter 5, when we lay the foundations for how to do counterfactual forecasting, we make an additional assumption that the time factor $u_{t\ell}$ for $\ell in[r]$ is explicitly modeled as a time series. That is, for $\ell \in [r]$,

$$u_{t,\ell} = \sum_{j=1}^{g} \alpha_j u_{t-j,\ell}. \tag{1.2}$$

This autoregressive time series model for $u_{\cdot,\ell}$ is known as a linear recurrent formula, and (approximately) admits a wide variety of time series dynamics include any finite sum and product of harmonics, polynomials, exponentials, and sufficiently smooth periodic function. The spatio-temporal model captured by (1.1) and (1.2) allows us to (1) effectively share information across units, time, and interventions, (2) forecast *future* counterfactual potential outcomes.

Goal: estimating unobserved potential outcomes. In the various works that comprise this thesis, the common goal is to recover $\mathbb{E}[Y_{tn}^{(d)}]$ for various subsets of $[N], [T], [D]$. In Chapter 3, we focus on estimating and doing inference for

$$\theta_n^{(d)} = \frac{1}{|\mathcal{T}_{\text{post}}|} \sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}[Y_{tn}^{(d)}], \text{ for } n \in [N], d \in [D],$$

where $\mathcal{T}_{\text{post}} \subset [T]$ is some subset of measurements of interest. That is, the expected potential outcome of unit n under intervention d, averaged over the measurements $t \in \mathcal{T}_{\text{post}}$. In Chapter 4, we generalize the results in Chapter 3 and estimate

$$\mathbb{E}[Y_{tn}^{(d)}], \text{ for } n \in [N], d \in [D], t \in [T],$$

i.e., for *each* t, n, d, *entry-wise*. In Chapter 5, we focus on estimating

$$\mathbb{E}[Y_{tn}^{(0)}], \text{ for } t > T,$$

i.e., forecasting potential outcomes under a given intervention for time steps in the *future* (here we take T to be "present day").

■ 1.2 Thesis Overview

■ 1.2.1 Principal Component Regression (Chapter 2)

Motivation. In Chapter 2, we show how estimating $\mathbb{E}[Y_{tn}^{(d)}]$ can effectively be reduced to the problem of high-dimensional error-in-variables (EIV) regression, a challenging setting where the number of covariates can be much larger than the number of measurements, and covariates are corrupted by noise. Intuitively, $Y_{tn}^{(d)}$ can be thought of as a noisy observation of $\mathbb{E}[Y_{tn}^{(d)}]$ (due to the presence of $\varepsilon_{tn}^{(d)}$). Thus, using the observed $Y_{tn}^{(d)}$ as covariates to learn a regression model boils down to EIV regression. In short, in Agarwal et al. (2019b, 2021e,d), which forms the basis for Chapter 2, we show principal component regression (PCR) is surprisingly effective for EIV regression. In doing so, we provide a theoretical justification for PCR, a very popular regression method at least since the 1980's Jolliffe (1986), but with minimal formal analysis. As we will see, the theoretical analysis we do for PCR underpins a lot of the algorithmic development and analysis we do in the subsequent chapters.

Setup. We consider the setup of EIV regression in a high-dimensional fixed design setting. Formally, we observe a labeled dataset of size n, denoted as $\{(y_i, z_i) : i \leq n\}$. Here, $y_i \in \mathbb{R}$ represents the response variable, also known as the label or target. For any $i \geq 1$, we posit that

$$y_i = \langle x_i, \beta^* \rangle + \varepsilon_i, \tag{1.3}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown model parameter, $x_i \in \mathbb{R}^p$ is the associated covariate, and $\varepsilon_i \in \mathbb{R}$ is the response noise. Unlike traditional regression settings where $z_i = x_i$, the error-in-variables regression setting reveals a corrupted version of the covariate x_i. Precisely, for any $i \geq 1$, let $z_i \in \mathbb{R}^p$ be given by

$$z_i = (x_i + w_i) \circ \pi_i, \tag{1.4}$$

where $w_i \in \mathbb{R}^p$ is the covariate measurement noise and $\pi_i \in \{0, 1\}^p$ is a binary observation mask with $\circ$ denoting component-wise multiplication, i.e., we observe the k-th component of z_i if $\pi_{ik} = 1$ and 0 otherwise. Further, we consider a high-dimensional setting where both n and p are growing with n possibly much smaller than p.

PCR Algorithm. In a nutshell, PCR is a two-stage process: first, PCR "de-noises" the observed in-sample (training) covariate matrix $\boldsymbol{Z} = [z_i^T] \in \mathbb{R}^{n \times p}$ via principal component analysis (PCA), i.e., PCR replaces $\boldsymbol{Z}$ by its low-rank approximation. This PCA steps aim to retain only the part of the spectra that corresponds to signal. Then, PCR regresses $y = [y_i] \in \mathbb{R}^n$ with respect to this low-rank approximation to produce the model estimate $\widehat{\beta}$. We are interested in the following natural questions about the estimation quality of PCR in a high-dimensional error-in-variables setting: (1) amongst the many observationally equivalent models in the high-dimensional setting, is there a model that PCR identifies consistently? (2) given noisy and partially observed out-of-sample (test) covariates, how can PCR be methodologically extended to accurately predict the expected test response variables, i.e., under what conditions does PCR generalize?

Geometric intuition. The intuition behind using PCR is that if the expected potential outcomes have a low-rank structure (i.e., $\mathbb{E}[Y_{tn}^{(d)}] = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell} \lambda_{d\ell}$) then $\boldsymbol{X} = [x_i]_{i \in [n]}$ will have low-rank structure too—all of the signal is captured in the first few singular values in the spectral space. The noise matrix $\boldsymbol{W} = [w_i]_{i \in [n]}$ on the hand will be spectrally diffuse due to the independence of w_i across measurements. That is,

"signal is spectrally concentrated while noise is spectrally diffuse".

See Figure 1.4 for a visual depiction of the typical spectrum of signal vs. noise for low-rank $\boldsymbol{X}$.

Key Results. We establish three main results for PCR.

(1) Parameter estimation. We establish that PCR consistently estimates the unique minimum ℓ_2-norm model parameter amongst all feasible models as per (1.3), i.e., PCR implicitly regularizes. Notably, the minimum ℓ_2-norm β^* is of primary interest from the perspective of prediction. In line with the geometric intuition from Figure 1.4, we define a "signal-to-noise" ratio of the true covariates $\boldsymbol{X}$ as $\text{snr} := \rho s_r / (\sqrt{n} + \sqrt{p})$, where s_r is the smallest non-zero singular value of $\boldsymbol{X}$ and ρ is the probability of observing each entry of $\boldsymbol{Z}$. We establish that PCR consistently estimates the minimum ℓ_2-norm β^*, i.e.,

$$\left\| \widehat{\beta} - \beta^* \right\|_2 \xrightarrow{p} \tilde{o}(1),$$

if snr is growing sufficiently quickly, i.e., $\text{snr} = \omega(\sqrt{\log(np)})$, Further, we establish minimax optimality of the PCR estimator for the EIV setting—if $\text{snr} = O(\sqrt{\log(np)})$, then

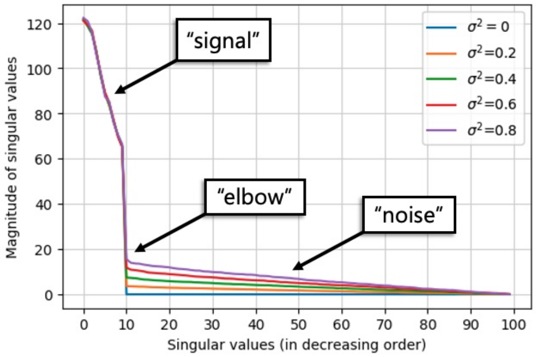


Figure 1.4: Simulation displays the spectrum of $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W} \in \mathbb{R}^{100\times 100}$. Here, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$, where the entries of $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{100\times 10}$ are sampled independently from $\mathcal{N}(0,1)$; further, the entries of $\boldsymbol{W}$ are sampled independently from $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 \in \{0, 0.2, \dots, 0.8\}$. Across varying levels of noise σ^2, there is a steep drop-off in magnitude from the top to remaining singular values—this marks the "elbow" point. As seen from the figure, the top singular values of $\boldsymbol{Z}$ correspond closely with that of $\boldsymbol{X}$ ($\sigma^2 = 0$), and the remaining singular values are induced by $\boldsymbol{W}$. Thus, $\text{rank}(\boldsymbol{Z}) \approx \text{rank}(\boldsymbol{X}) = 10$.

no estimator can consistently estimate β^*. The key theoretical results in Chapter 3 and 4 crucially build upon this parameter estimation result.

(2) Out-of-sample prediction. We establish that PCR achieves vanishing out-of-sample prediction error, even in the presence of corrupted out-of-sample covariates. To the best of our knowledge, the standard works in the error-in-variables literature do not provide prediction error guarantees in the presence of corrupted test covariates. Additionally, since we consider a fixed design setting, we do *not* make any distributional assumptions on the data generating process of the true train and test covariates to arrive at our result. We hope this provides a novel perspective on learning with covariate shifts, an important topic in the statistics and econometrics literatures. As with our parameter estimation result, we establish that our test prediction error is controlled by the signal-to-noise ratio corresponding to the test covariates, defined as $\text{snr}_{\text{test}} := \rho s'_{r'}/(\sqrt{m} + \sqrt{p})$, where $s'_{r'}$ is the smallest non-zero singular value of the test covariates $\boldsymbol{X}'$ and m is the size of the test set. We establish that PCR's test prediction error vanishes, i.e.,

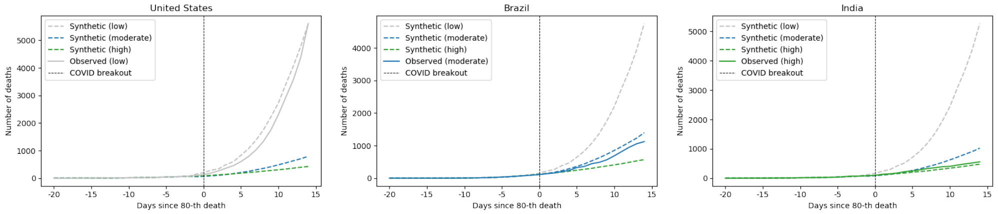
$$\left\| \boldsymbol{X}'\beta^* - \widehat{\boldsymbol{X}'}\widehat{\beta} \right\|_2 \xrightarrow{p} \tilde{o}(1),$$

as long as $\text{snr}, \text{snr}_{\text{test}}$ are growing sufficiently quickly. Note that $\widehat{\boldsymbol{X}'}$ is produced by doing PCA on the noisy test covariates $\boldsymbol{Z}' = \boldsymbol{X}' + \boldsymbol{W}'$. In the special case when the underlying train and test covariates have a well-balanced spectra and $m, p = \Theta(n)$, we show that the squared ℓ_2-norm test prediction error rate is $\tilde{O}(1/n)$.

(3) "Online" variant of PCR. In Chapter 5, we show that a simple variant of the PCR algorithm provided in Chapter 2 is able to handle "online" data. This is where the noisy test covariates $z'_i = x'_i + w'_i$ for $i \in [m]$ arrive one at a time in an online fashion and we have to make an immediate prediction for $\langle x'_i, \beta^* \rangle$. The "online" setting is harder as there is no way to explicitly de-noise z'_i for $i \in [m]$; in contrast, in the "offline" setting considered above, all of the test covariates $\boldsymbol{Z}' = [z'_i]_{i \in [m]}$ are available in batch and so we can collectively de-noise them via PCA. Surprisingly we show that taking the inner product of z'_i with $\widehat{\beta}$ (learned using the training covariates $\boldsymbol{Z}$) *implicitly* de-noises z'_i. Hence, our online variant of PCR is simply to multiply each z'_i that arrives with $\widehat{\beta}$. We establish this algorithm has regret (i.e., ℓ_2-norm test prediction error) scaling as $\tilde{O}(1/n)$ when $m, p = \Theta(n)$. This online variant of PCR is a crucial step in our time series forecasting results in Chapter 5.

■ 1.2.2 Synthetic Interventions (Chapter 3)

Motivation. We briefly motivate and describe the synthetic interventions (SI) framework through a policy application we applied it to: the goal was to estimate the counterfactual COVID-19 morbidity rate across countries if they had enacted different mobility restricting interventions (measured using Google's mobility reports Google (2020)). The challenge in doing so was that heterogeneous characteristics of a country (e.g., the government structure, demographics, cultural leanings) are "confounders", i.e., they impact *both* the social distancing policies that were put into place and the observed health outcomes. The SI estimator implicitly corrects for this confounding by leveraging observational data across nations. For example, to estimate the U.S.'s health outcomes under a policy it did not implement, the estimator builds a "synthetic U.S." by leveraging data of other countries that did go through that policy. See Figure 1.5 for a depiction of the synthetic US, Brazil and India produced by the SI estimator under different levels of mobility restriction (details in Agarwal et al. (2021c)). In doing so, we also extend the synthetic controls method, a widely used frameworks in econometrics, to the multiple intervention setting, an important open problem in the literature Abadie (2020). The simplicity and robustness of the SI estimator is what allows us to use it in a range of applications. These include: clinical trial design, policy-evaluation, development economics, "synthetic" A/B testing in e-commerce, and synthetic biology Agarwal et al. (2021c); Squires et al. (2021); Agarwal et al..



(a) United States **(b)** Brazil **(c)** India

Figure 1.5: Counterfactual predictions of COVID-19 related morbidity counts under different mobility restriction levels.

Setup. We consider a setting with $N \geq 1$ units and $D \geq 1$ interventions. For each unit and intervention pair, there are $T \geq 1$ outcomes/measurements of interest. Unless stated otherwise, we index units with $n \in [N] := \{1, \dots, N\}$, outcomes with $t \in [T]$, and interventions with $d \in [D]_0$.[1] Recall in Figure 1.2, we encode the potential outcomes into a order-3 tensor whose dimensions correspond to units, measurements, and interventions. We describe our observations through $\boldsymbol{Y} = [Y_{tnd}] \in \{\mathbb{R} \cup \star\}^{T \times N \times D}$, where $\star$ indicates a missing entry. We assume $\boldsymbol{Y}$ obeys the following sparsity pattern.

Required observation pattern in potential outcomes tensor in SI. We assume we observe the same $T_0 \leq T$ outcomes for all units under the same intervention. Without loss of generality, let this intervention be $d = 0$, and let the indices corresponding to these T_0 measurements be $\mathcal{T}_{pre} := \{1, \dots, T_0\}$. That is, we observe $Y_{tn0} = Y_{tn}^{(0)}$ for all $n \in [N]$ and $t \in [T_0]$. Further, for every intervention d, there is a non-empty subset of units, $\mathcal{I}^{(d)} \subset [N]$, for which we observe $T_1 \leq T$ measurements. Let $N_d = |\mathcal{I}^{(d)}|$. Without loss of generality, we assume the indices corresponding to these T_1 measurements are $\mathcal{T}_{post} := \{T - T_1 + 1, \dots, T\}$. That is, we observe $Y_{tnd} = Y_{tn}^{(d)}$ for $d \in [D]_0$, $n \in \mathcal{I}^{(d)}$, and $t \in \mathcal{T}_{post}$. For all other entries of $\boldsymbol{Y}$, we assume $Y_{tnd} = \star$.

Recall from Figure 1.3 two typical sparsity patterns seen in causal inference settings that meet the required observation pattern above.

Goal. As stated earlier, the goal in this chapter is to estimate $Y_{tn}^{(d)}$ for all $n \in [N]$ and $d \in [D]$,

$$\theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}[Y_{tn}^{(d)}].$$

[1] Let $[X]_0 = \{0, 1, \dots, X - 1\}$ and $[X] = \{1, \dots, X\}$ for any positive integer X.

SI Algorithm. The SI estimator recovers a given $\theta_n^{(d)}$ with valid confidence intervals via a three-step procedure, where each step has a simple closed form expression. Below, we give an overview of the method. For a given (n, d) pair, the first step is to estimate a linear model, $w^{(n,d)} \in \mathbb{R}^{N_d}$ such that for all $t \in \mathcal{T}_{\text{pre}}$,

$$Y_{tn0} \approx \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} Y_{tj0}.$$

Specifically, we use principal component regression (PCR) (analyzed in Chapter 2) to learn $w^{(n,d)}$ by linearly regressing $\{Y_{tn0} : t \in \mathcal{T}_{\text{pre}}\}$ on $\{Y_{tj0} : t \in \mathcal{T}_{\text{pre}}, j \in \mathcal{I}^{(d)}\}$. Subsequently $\theta_n^{(d)}$ is estimated as

$$\begin{aligned}\widehat{\mathbb{E}}[Y_{tn}^{(d)}] &= \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} Y_{tjd}, \quad \text{for } t \in \mathcal{T}_{\text{post}}, \\ \widehat{\theta}_n^{(d)} &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \widehat{\mathbb{E}}[Y_{tn}^{(d)}].\end{aligned}$$

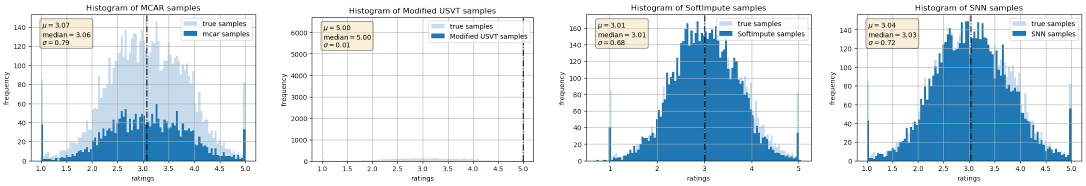
Key results. We establish an identification result for $\widehat{\theta}_n^{(d)}$, i.e., the counterfactual outcome $\widehat{\theta}_n^{(d)}$ can be expressed as a function of *observed outcomes*. Importantly, the SI estimator allows for latent confounders that determine how interventions are assigned. Further, we prove finite-sample consistency and asymptotic normality of the estimator. That is,

$$\begin{aligned}|\widehat{\theta}_n^{(d)} - \theta_n^{(d)}| &\xrightarrow{p} \tilde{O}\left(\frac{1}{\min(T_0^{1/4}, \sqrt{T_1}, \sqrt{N_d})}\right), \\ \frac{\sqrt{T_1}(\widehat{\theta}_n^{(d)} - \theta_n^{(d)})}{\tilde{\sigma}} &\xrightarrow{d} \mathcal{N}(0, 1),\end{aligned}$$

where $\tilde{\sigma}$ can be consistently estimated. In doing so, we establish novel identification, estimation and inference results for the widely used synthetic controls framework Abadie (2020) as well.

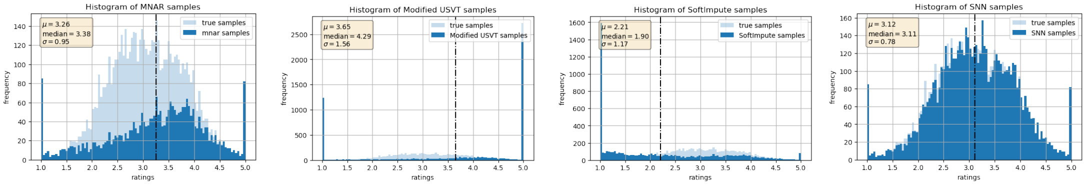
■ 1.2.3 Causal Matrix Completion (Chapter 4)

Motivation. Recommendations have become an integral part of modern social and engineering systems, and can strongly alter the purchasing and engagement behavior of individuals. The goal of recommender systems is to optimally recommend previously unrated items that an individual is likely to prefer. This problem can be reduced to matrix completion, where rows index users and columns index items; each missing user-item entry corresponds to the potential rating a user would give to that item had they rated it. To motivate the importance of studying the missingness mechanism, we showcase two experiments (details in Agarwal et al. (2021b)), one with MCAR and the other with MNAR data in Figures 1.6a and 1.7a. We use three matrix completion algorithms to recover the distribution of true ratings given a subset of revealed ratings: (i) Universal singular value thresholding (`USVT`) a popular spectral based method; (ii) Softimpute (`softImpute`), a popular optimization based method; (iii) "synthetic nearest neighbors" (`SNN`), our proposed method. In Figure 1.6, under MCAR, `softImpute` and `SNN` both accurately recover the true distribution, while `USVT` cannot. In Figure 1.7, under MNAR, `SNN` continues faithful recovery, but `softImpute` is significantly biased. This underscores how MNAR data can bias recommendations and the need for a rigorous framework to tackle it. That is exactly what this chapter aims to do.



(a) True, revealed. **(b)** `USVT`. **(c)** `softImpute`. **(d)** `SNN`.

Figure 1.6: MCAR: `softImpute` and `SNN` recover true distribution faithfully; note different scale for `USVT`.



(a) True, revealed . **(b)** `USVT`. **(c)** `softImpute`. **(d)** `SNN`.

Figure 1.7: MNAR: `SNN` faithfully recovers true distribution with MNAR data.; note different scales for `USVT` & `softImpute`.

Setup. We consider a signal matrix $\boldsymbol{A} = [A_{ij}] \in \mathbb{R}^{m \times n}$, a noise matrix $\boldsymbol{E} = [\varepsilon_{ij}] \in \mathbb{R}^{m \times n}$, and a propensity score matrix $\boldsymbol{P} = [p_{ij}] \in [0,1]^{m \times n}$. All three matrices are entirely latent, i.e., unobserved. Let $\boldsymbol{Y} = [Y_{ij}] \in \mathbb{R}^{m \times n}$ denote the "noisy" version of $\boldsymbol{A}$, with $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{A}$; we denote $\varepsilon_{ij} = Y_{ij} - A_{ij}$. We assume $\boldsymbol{Y}$ itself is partially observed. In particular, we denote $\boldsymbol{D} = [D_{ij}] \in \{0,1\}^{m \times n}$ with $\mathbb{E}[\boldsymbol{D}] = \boldsymbol{P}$ as the missingness mask matrix that indicates which entries of $\boldsymbol{Y}$ are observed. For convenience, we encode our observations into

$\widetilde{\boldsymbol{Y}} = [\widetilde{Y}_{ij}] \in \{\mathbb{R} \cup \{\star\}\}^{m \times n}$ such that for $(i, j) \in [m] \times [n]$,

$$\widetilde{Y}_{ij} = \begin{cases} Y_{ij}, & \text{if } D_{ij} = 1 \\ \star, & \text{otherwise.} \end{cases}$$

In words, if $D_{ij} = 1$ then A_{ij} is noisily observed, and if $D_{ij} = 0$ then A_{ij} remains unknown.

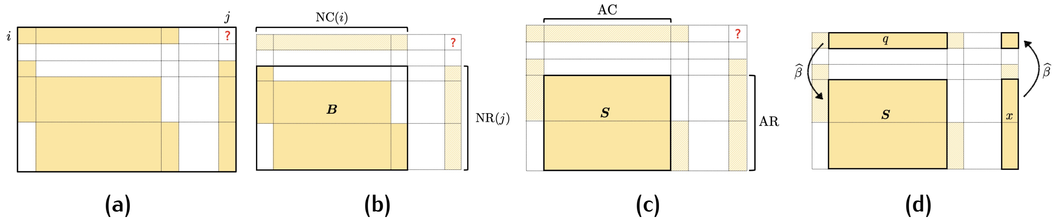
In terms of the type of MNAR data this work considers, we allow for $\boldsymbol{D}$ and $\boldsymbol{Y}$ to be dependent, provided $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{A}$, where $\boldsymbol{A}$ is latent. In fact, we allow $\boldsymbol{D}$ to be any arbitrary function of $\boldsymbol{A}$, random or deterministic, subject to suitable observation patterns. Notably, our framework also allows the entries in $\boldsymbol{D}$ to be dependent with each other across both rows and columns, and the minimum value of $\boldsymbol{P}$ to be 0, which are important departures from the current matrix completion literature. Under these conditions, we propose an algorithm that provably recovers $\boldsymbol{A}$ from $\widetilde{\boldsymbol{Y}}$ with entry-wise (i.e., max-norm) guarantees.

For concreteness, let us return to the recommender system example. In line with the potential outcomes framework, A_{ij} can be interpreted as $\mathbb{E}[Y^{(ij)}]$, i.e., the expected potential rating a user would give to an item if they had rated it. $\boldsymbol{P}$ dictates the probability these expected ratings are revealed; $\boldsymbol{Y}$ in relation to $\boldsymbol{A}$ then models the inherent randomness in how users rate items; that is, $\boldsymbol{Y}$ can be interpreted as a "noisy" instance of $\boldsymbol{A}$.

SNN algorithm. `SNN` is a simple two-step algorithm which combines the nearest neighbors (i.e., collaborative filtering) approach for matrix completion with the SI approach in Chapter 3. See Figure 1.8 for a visual depiction of the algorithm. `SNN` draws inspiration from the popular K-Nearest Neighbour (`KNN`) algorithm. However, the key assumption underlying `KNN` is that there *do exist* K rows that are close to identical to the i-th row, with respect to some pre-defined metric. However, it is not necessary that these K rows exist *even* for a rank 1 matrix. As a simple example, consider a matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ where $M_{i\cdot} = [i, 2i, \dots, ni]$. By construction $\boldsymbol{M}$ is rank 1, but for any row, there does not exist any other row that is close to it in a mean squared sense; hence, it has no nearest neighbours.

`SNN` overcomes this hurdle A_{ij} by taking an average of the observed outcomes for column j that are associated with the K "'synthetic" neighbors of row i. Each of the K "synthetic" neighbors of row i is constructed from NR(j), where the k-th synthetic neighboring row is formed by a linear combination, defined by $\widehat{\beta}^{(k)}$, of the rows in AR$^{(k)}$ (which is the k-th) partition of AR. As in SI, $\widehat{\beta}^{(k)}$ is learned via PCR.

Key results. We prove entry-wise finite-sample consistency and asymptotic normality of



(a) (b) (c) (d)

Figure 1.8: We visually depict the various quantities needed to define the SNN algorithm. Figure 4.5a depicts a particular sparsity pattern in our matrix $\widetilde{Y}$ with entry (i, j) missing. Figure 4.5b depicts NR(j) and NC(i). Figure 4.5c depicts AR, AC, and S. Figure 4.5d depicts the SNN algorithm with $K = 1$; for $K > 1$, we partition the rows in S into K mutually disjoint sets.

the SNN estimator for matrix completion with MNAR data (and MCAR data as a special case), an unresolved problem in the literature. That is,

$$|\hat{A}_{ij} - A_{ij}| \xrightarrow{p} \tilde{O}\Big(\frac{1}{\sqrt{K}}\Big),$$

$$\frac{\sqrt{K}(\hat{A}_{ij} - A_{ij})}{\tilde{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

How does this generalize SI? The setup in SI (Chapter 3) can be made a special case by effectively *flattening* the tensor into a matrix; rows of the induced matrix still correspond to units, but a column is a double index for a measurement and an intervention, i.e., the (i, j, d)-th entry of the tensor corresponds to the $(i, (j, d))$-th entry of the induced matrix. Given this reduction, we generalize the framework, algorithm, and theoretical results in SIin the following ways. First, we formally extend the SI framework, to recover matrices under more general missingness patterns.. Doing so allows us to apply our framework to a wider variety of applications such as recommender systems, while the SI framework was introduced in the context of personalized policy evaluation and synthetic A/B testing. Third, this work establishes point-wise finite-sample consistency and asymptotic normality of our proposed SNN algorithm, which was absent in Agarwal et al. (2021c) with respect to the SI algorithm. Indeed, in the context of the panel data literature, establishing point-wise asymptotic normality for each unit, (intervention, time)-tuple is of independent interest.

■ 1.2.4 Multivariate Singular Spectrum Analysis (Chapter 5)

Motivation. A large focus of the causal inference literature is estimating the counterfactual of *what would have happened* under an unseen intervention, i.e., causal imputation. Another meaningful question is estimating *what will happen* in the future under a collection of possible different interventions, i.e., causal forecasting. Doing this effectively has a variety of applications, e.g., forecasting customer demand under different discounting policies to optimize supply-chain management, forecasting network latency under different congestion control policies to optimize infrastructure resource planning. To do this, we have to combine models and algorithms for causal inference with those for time series forecasting. For example, say we collect data of many units over time under different interventions. The estimators SI Agarwal et al. (2021c) and SNN Agarwal et al. (2021b), discussed earlier, are shown to perform accurate causal imputation under a latent low-rank factor model, i.e., there is a low-dimensional factor for each unit, time period, and intervention—recall the model in (1.1). To do causal forecasting, a natural extension of such a latent factor model is to explicitly model the latent time factors as a time series (e.g., an autoregressive process)—recall the mode in (1.2). That is, a *spatio-temporal factor model*. Towards this, in Agarwal et al. (2022, 2019a, 2021a), under a novel spatio-temporal factor model across units and time, we propose and analyze a variant of Multivariate singular spectrum analysis (mSSA), which is a very popular method to impute and forecast a multivariate time series. However, despite its heavy use in practice, the theoretical properties of mSSA are not well understood. For the variant of mSSA we introduce, we establish a rigorous finite-sample analysis of its imputation and out-of-sample forecasting properties; such a finite-sample analysis of mSSA has been missing from the literature.

The hope is to eventually extend this spatio-temporal model and the variant of mSSA we propose to do counterfactual forecasting by incorporating data collected across different interventions as well.

Setup. We consider a discrete time setting with time indexed as $t \in \mathbb{Z}$. For $N \in \mathbb{N}$, let the collection $f_n : \mathbb{Z} \to \mathbb{R}, \;\; n \in [N] := \{1, \dots, N\}$ be the latent time series of interest. For $t \in [T]$ and $n \in [N]$, we observe $X_n(t)$ where for $\rho \in (0, 1]$,

$$X_n(t) = \begin{cases} f_n(t) + \eta_n(t) & \text{with probability } \rho \\ \star & \text{with probability } 1 - \rho. \end{cases} \tag{1.5}$$

Here $\star$ represents a missing observation and $\eta_n(t)$ represents the per-step noise, which we assume to be an independent (across t, n) mean-zero random variable. Though $\eta_n(t)$ is independent, we note that the underlying time series, $f_n(\cdot)$, is of course strongly dependent across t, n. Indeed the presence of per-step noise $\eta_n(t)$ and missing values (denoted by

⋆) represent an additional challenge of measurement error in our setup. As discussed earlier, the generic spatio-temporal factor model for $f_n(\cdot), n \in [N]$ described in (1.2) *without* additional noise $\eta_n(\cdot)$ or missingness already provides an expressive model for a time series including any finite sum of products of harmonics and polynomials, any differentiable periodic function, and any Hölder continuous function.

Goal. Our objective is two-folds, for $n \in [N]$: (i) imputation – estimating $f_n(t)$ for all $t \in [T]$; (ii) out-of-sample forecasting – predicting $f_n(t)$ for $t > T$.

mSSA algorithm. See Figure 1.9 for a visual depiction of the key steps of the variant of mSSA we propose. They key steps are as follows: (1) transform time series $X_n(t),\ t \in [T]$ into an $L \times T/L$ matrix where the entry of the matrix in row $i \in [L]$ and column $j \in [T/L]$ is $X_n(i + (j - 1) \text{ x } L)$. This matrix induced by the time series is called the Page matrix, and we denote it as $\text{P}(X_n, T, L)$. (2) Take $\text{P}(X_n, T, L)$ for $n \in [N]$ and concatenate them column-wise—this induced stacked page matrix is denoted as $\text{SP}((X_1, \ldots, X_N), T, L)$. (3) Simply do PCR on $\text{SP}((X_1, \ldots, X_N), T, L)$. The first step of PCA (also called hard singular value thresholding (HSVT)) *implicitly* de-noises $\text{SP}((X_1, \ldots, X_N), T, L)$ and that is the only step associated need for imputation. (4) The linear model learned via PCR can then be used to do out-of-sample forecasting.

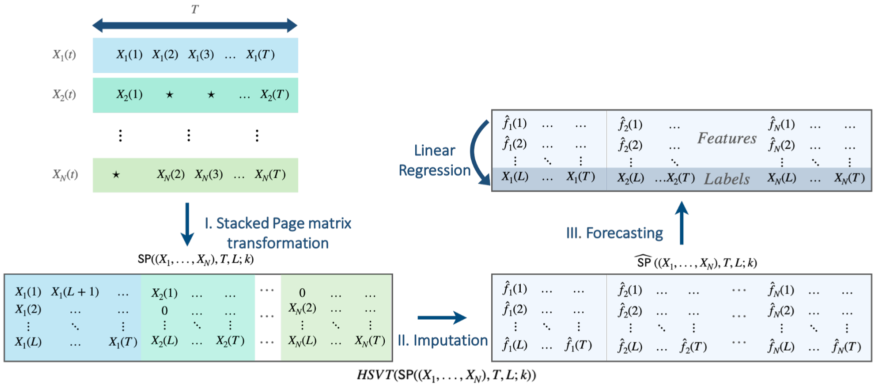


Figure 1.9: Key steps of our proposed variant of the mSSA algorithm.

Key results. Under the spatio-temporal factor model in (1.2), we show the mean-squared imputation error ($\text{ImpErr}(N, T)$) and out-of-sample forecasting error $\text{TestForErr}(N, T)$ scale as follows

$$\text{ImpErr}(N, T), \text{TestForErr}(N, T) = \tilde{O}\Big(\frac{1}{\sqrt{\min(N, T)T}}\Big)$$

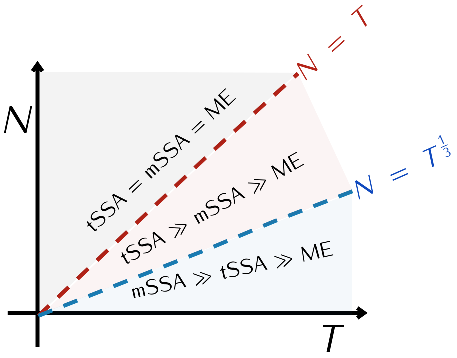


Figure 1.10: Relative effectiveness of tSSA, mSSA, ME for varying N, T.

In comparison, the univariate version of mSSA, called SSA, has error scaling as $\tilde{O}\Big(\frac{1}{\sqrt{T}}\Big)$. And at least for imputation, if one does matrix completion directly on the time series data without creating the Page matrix first, the error scales as $\tilde{O}\Big(\min(N, T)\Big)$. Hence our analysis suggests mSSA exploits both the temporal and spatial structure in the data.

Lastly, in Chapter 5, we we propose a novel tensor variant of SSA, termed tSSA, which exploits recent developments in the tensor estimation literature. In tSSA, rather than doing a column-wise stacking of the Page Matrices induced by each of the N time series to form a larger matrix, we instead view each Page matrix as a slice of a $L \times T/L \times N$ order-three tensor. In other words, the entry of the tensor with indices $i \in [L], j \in [T/L]$ and $n \in [N]$ equals the entry of $\mathrm{P}(X_n, L, T)$ with indices i, j. With respect to imputation error, we characterize the relative performance of tSSA, mSSA, and "vanilla" matrix estimation (ME). We find that when $N = o(T^{1/3})$, mSSA outperforms tSSA; when $T^{1/3} = o(N)$, $N = o(T)$ tSSA outperforms mSSA; when $T = o(N)$, standard matrix estimation methods are equally as effective as mSSA and tSSA. See Figure 1.10 for a graphical depiction of the various regimes. In addition to being a basis for counterfactual forecasting, we hope this work motivates future inquiry into the connections between the classical field of time series analysis and the modern, growing field of matrix/tensor estimation.

Chapter 2

On Principal Component Regression

■ 2.1 Introduction

We consider the setup of error-in-variables regression in a high-dimensional fixed design setting. Formally, we observe a labeled dataset of size n, denoted as $\{(y_i, z_i) : i \leq n\}$. Here, $y_i \in \mathbb{R}$ represents the response variable, also known as the label or target. For any $i \geq 1$, we posit that

$$y_i = \langle x_i, \beta^* \rangle + \varepsilon_i, \tag{2.1}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown model parameter, $x_i \in \mathbb{R}^p$ is the associated covariate, and $\varepsilon_i \in \mathbb{R}$ is the response noise. Unlike traditional regression settings where $z_i = x_i$, the error-in-variables regression setting reveals a corrupted version of the covariate x_i. Precisely, for any $i \geq 1$, let $z_i \in \mathbb{R}^p$ be given by

$$z_i = (x_i + w_i) \circ \pi_i, \tag{2.2}$$

where $w_i \in \mathbb{R}^p$ is the covariate measurement noise and $\pi_i \in \{0, 1\}^p$ is a binary observation mask with $\circ$ denoting component-wise multiplication, i.e., we observe the k-th component of z_i if $\pi_{ik} = 1$ and 0 otherwise. Further, we consider a high-dimensional setting where both n and p are growing with n possibly much smaller than p.

Our interest is in analyzing the performance of the classical method of principal component regression (PCR) for this scenario. In a nutshell, PCR is a two-stage process: first, PCR "de-noises" the observed in-sample (train) covariate matrix $\boldsymbol{Z} = [z_i^T] \in \mathbb{R}^{n \times p}$ via principal component analysis (PCA), i.e., PCR replaces $\boldsymbol{Z}$ by its low-rank approximation. Then, PCR regresses $y = [y_i] \in \mathbb{R}^n$ with respect to this low-rank approximation to produce the model estimate $\widehat{\beta}$. We are interested in the following natural questions about the estimation quality of PCR in a high-dimensional error-in-variables setting: (1) amongst the many feasible models in the high-dimensional setting, is there a model that PCR identifies consistently? (2) given noisy and partially observed out-of-sample (test) covariates, how can PCR be methodologically extended to accurately predict the expected test response variables, i.e., under what conditions does PCR generalize?

■ 2.1.1 Contributions

Model identification. As the main contribution of this work, we establish that PCR consistently estimates the unique minimum ℓ_2-norm model parameter amongst all feasible models as per (2.1), i.e., PCR implicitly regularizes. Notably, the minimum ℓ_2-norm β^* is of primary interest from the perspective of prediction (see Section 2.5.1 for a discussion on this). We define a "signal-to-noise" ratio of the true covariates $\boldsymbol{X} = [x_i^T] \in \mathbb{R}^{n \times p}$ as snr $:= \rho s_r / (\sqrt{n} + \sqrt{p})$, where s_r is the smallest non-zero singular value of $\boldsymbol{X}$ and ρ is the probability of observing each entry of $\boldsymbol{Z}$. Theorem 2.5.1 establishes that PCR consistently estimates the minimum ℓ_2-norm β^* if snr is growing sufficiently quickly, i.e., snr $= \omega(\sqrt{\log(np)})$, ignoring dependencies on β^* and the rank of $\boldsymbol{X}$.

Near optimality. We establish a minimax lower bound for parameter estimation in our setting of interest in Theorem 2.5.2. This result suggests that if snr $= O(1)$, then the parameter estimation error is lower bounded by a positive, absolute constant. That is, PCR is near minimax optimal.

Out-of-sample prediction. We establish that PCR achieves vanishing out-of-sample prediction error, even in the presence of corrupted out-of-sample covariates (Theorem 2.5.3 and Corollary 2.5.2). To the best of our knowledge, the standard works in the error-in-variables literature do not provide prediction error guarantees in the presence of corrupted test covariates. Additionally, since we consider a fixed design setting, we do *not* make any distributional assumptions on the data generating process of the true train and test covariates to arrive at our result. Rather, we introduce a natural linear algebraic condition (Assumption 5) relating the train and test covariates. We hope this provides a novel perspective on learning with covariate shifts, an important topic in the statistics and econometrics literatures. In contrast, popular tools to understand the generalization properties of estimators, such as Rademacher complexity analyses, commonly operate under distributional assumptions. As with our parameter estimation result, we establish that our test prediction error is controlled by the signal-to-noise ratio corresponding to the test covariates, defined as $\text{snr}_{\text{test}} := \rho s'_{r'}/(\sqrt{m} + \sqrt{p})$, where $s'_{r'}$ is the smallest non-zero singular values of test covariates $\boldsymbol{X}'$ and m is the size of the test set. Theorem 2.5.3 then states that PCR's test prediction error vanishes as long as $\text{snr}, \text{snr}_{\text{test}}$ are growing sufficiently quickly. In the special case when the underlying train and test covariates have a well-balanced spectra and $m, p = \Theta(n)$, we show that the squared ℓ_2-norm test prediction error rate is $\tilde{O}(1/n)$ (Corollary 2.5.2). This improves upon the best known test prediction error rate of $\tilde{O}(1/\sqrt{n})$ for PCR, as established in Agarwal et al. (2019b, 2021e). Alas, the results in these prior works consider a transductive learning setting with i.i.d. covariates. In contrast, our work goes beyond this restrictive setup by allowing for a fixed design setting. For a detailed comparison, please refer to Section 2.2.

Counterfactual predictions for synthetic controls. An important motivation for this work is that of synthetic controls, a popular method for counterfactual predictions with observational data Abadie et al. (2010); Abadie and Gardeazabal (2003); Abadie (2020). Specifically, counterfactual prediction in synthetic controls corresponds to out-of-sample prediction in the setting of this work: (a) the factor model structure utilized in the synthetic controls literature implies the low-rank structure of the underlying covariate matrix; (b) the per-step idiosyncratic shocks correspond to the error-in-variables; and (c) out-of-sample prediction corresponds to counterfactual prediction of the outcomes under control. In light of this, our results add to the synthetic controls literature in the following ways. We show consistency in recovering the vector of counterfactual outcomes in terms of the mean-squared error (MSE). In particular, we provide a “fast-rate" analysis, which improves upon the “slow-rate" analysis in prior works Agarwal et al. (2019b, 2021e). Further, our analysis considers a fixed design setting, i.e., the latent factors can be deterministic, rather than being sampled independently as considered in prior works Agarwal et al. (2019b, 2021e). Finally, the near optimality of PCR suggested by our minimax result implies the near optimality of the robust synthetic controls method of Amjad et al. (2018) and suggests its attractivness amongst the many variants in the synthetic controls literature (e.g., Abadie et al. (2010); Abadie and Gardeazabal (2003); Doudchenko and Imbens (2016a); Athey et al. (2021)).

■ 2.1.2 Organization

The remainder of this paper is organized as follows. In Section 2.2, we discuss related works. In Section 2.3, we formally describe our setup and assumptions. We then describe the PCR algorithm in Section 2.4 followed by its parameter estimation and out-of-sample prediction error bounds in Section 2.5. In Section 2.6, we discuss applications of this work to the synthetic controls literature. To reinforce our theoretical findings, we provide illustrative simulations in Section 2.7. We conclude and discuss important future directions of research in Section 5.8.

■ 2.1.3 Notation

For any matrix $\boldsymbol{A} \in \mathbb{R}^{a \times b}$, we denote its operator (spectral), Frobenius, and max element-wise norms as $\boldsymbol{A}_2$, $\boldsymbol{A}_F$, and $\boldsymbol{A}_{\max}$, respectively. By rowspan($\boldsymbol{A}$), we denote the subspace of $\mathbb{R}^b$ spanned by the rows of $\boldsymbol{A}$. For any vector $v \in \mathbb{R}^a$, let v_p denote its ℓ_p-norm. If v is a random variable, we define its sub-gaussian (Orlicz) norm as v_{ψ_2}. Let $\circ$ denote component-wise multiplication and let $\otimes$ denote the outer product. For any two numbers $a, b \in \mathbb{R}$, we use $a \wedge b$ to denote $\min(a, b)$ and $a \vee b$ to denote $\max(a, b)$. Further, let $[a] = \{1, \dots, a\}$ for any integer a. Let f and g be two functions defined on the same space. We say that $f(n) = O(g(n))$ if and only if there exists a positive real number M and a real number n_0 such that for all $n \geq n_0, |f(n)| \leq M|g(n)|$. Analogously we say $f(n) = \Theta(g(n))$ if and only if there exists positive real numbers m, M such that for all $n \geq n_0,\ m|g(n)| \leq |f(n)| \leq M|g(n)|$; $f(n) = o(g(n))$ if for any $m > 0$, there exists n_0 such that for all $n \geq n_0, |f(n)| \leq m|g(n)|$; $f(n) = \omega(g(n))$ if for any $m > 0$, there exists n_0 such that for all $n \geq n_0, |f(n)| \geq m|g(n)|$. $\tilde{O}(\cdot)$ is defined analogously to $O(\cdot)$, but ignores log dependencies.

■ 2.2 Related works

In this section, we discuss related prior works. We also provide a detailed discussion of our key assumptions and connect them to the assumptions made in these prior works. In Section 2.2.1, we discuss our work in the context of the PCR literature. In Section 2.2.3, we present a detailed comparison with the high-dimensional error-in-variables literature. In Section 2.2.4, we briefly discuss linear regression with hidden confounders, an important topic in econometrics.

■ 2.2.1 Principal Component Regression (PCR)

PCR, as a method, was introduced in Jolliffe (1982). Despite the ubiquity of PCR in practice, the formal literature on PCR is surprisingly sparse. Notable works include Bair et al. (2006); Chao et al. (2019); Agarwal et al. (2019b, 2021e).

Model Identification.

In Agarwal et al. (2019b, 2021e), the authors present finite-sample analyses for the prediction error of PCR in a high-dimensional error-in-variables setting, but do not provide any analysis for parameter estimation. In contrast, this work establishes that PCR identifies the unique model with minimum ℓ_2-norm, i.e., PCR implicitly regularizes, and provide non-asymptotic rates of convergence.

Fixed Design.

In Agarwal et al. (2019b, 2021e), the authors consider a transductive learning setting, where *both* in-sample and out-of-sample covariates are accessible upfront. Importantly, Agarwal et al. (2019b, 2021e) assumes i.i.d. covariates, which allows them to leverage the techniques of Rademacher complexity analyses to establish their prediction error bounds. This work, on the other hand, considers the classical supervised learning setup, where test covariates are not revealed during training. Further, we consider a fixed design setting where the in-sample and out-of-sample covariates do *not* need to obey the same distribution. Rather, we establish that PCR achieves consistent test prediction in a distribution-free setting as long as a natural linear algebraic constraint is satisfied between the train and test covariates (Assumption 5).

Rates of Convergence.

In Agarwal et al. (2019b, 2021e), the authors show that when $m, p = \Theta(n)$, PCR's out-of-sample prediction error decays as $\tilde{O}(1/\sqrt{n})$. As noted in Agarwal et al. (2019b, 2021e), this "slow" rate likely arises from their analysis via Rademacher complexity arguments.

In contrast, we leverage our model identification result to prove that PCR's out-of-sample prediction error decays at the "fast" rate of $\tilde{O}(1/n)$.

Minimax Lower Bound.

Unlike prior works, we introduce a minimax lower bound for parameter estimation. This allows us to establish the near minimax optimality of PCR with respect to parameter estimation. In the process, we introduce a natural notion of signal-to-noise ratio for this setting.

■ 2.2.2 Functional Principal Component Analysis (fPCA)

We also take note of a related literature on functional principal component analysis (fPCA), which is a natural generalization of PCA to infinite-dimensional operators (see Yao et al. (2005); Hall et al. (2006); Li and Hsing (2010); Descary et al. (2019)). Typically in this literature, it is assumed that we observe n randomly sampled trajectories at p locations (carefully chosen from a grid with minor perturbations) forming an $n \times p$ data matrix, say D. The $p \times p$ matrix, $D^T D$, is the empirical proxy of the underlying covariance kernel that corresponds to these random trajectories. Under appropriate structural assumptions on these trajectories, despite the high-dimensionality, the $D^T D$ matrix can be represented as the additive sum of a low-rank matrix and a noise matrix. This resembles the setting of low-rank matrix estimation with a key difference being that *all* entries are observed. In the work of Descary et al. (2019), the low-rank component is estimated by performing an explicit rank minimization, which is known to be computationally hard. The functional (or trajectory) approximation from this low-rank estimation is obtained by smoothing (or interpolation)—this is where the careful choice of locations in a grid plays an important role. The estimation error is provided with respect to the normalized Frobenius norm (i.e., Hilbert-Schmidt norm when discretized). Finally, the fPCA literature considers the setting where $n \to \infty$ for a given p or at best $n \gg p$.

In comparison, PCR, as we argue, utilizes hard singular value thresholding (HSVT), a popular method in the matrix estimation toolkit, to recover the low-rank matrix; such an approach is computationally efficient and even yields a closed form solution. Indeed, PCR, as introduced in the original work Jolliffe (1982), is precisely HSVT followed by ordinary least squares (OLS). As a result, unlike the standard fPCA setup, PCR allows for sparsity in the observed covariate matrix since HSVT is precisely designed to recover the underlying low-rank matrix in the presence of both noisy and missing entries. As shown in this work, for OLS to not only recover the minimum ℓ_2-norm model parameter faithfully but also to obtain meaningful prediction error bounds, the matrix estimation error needs to be bounded with respect to a stronger norm than the normalized Frobenius norm, concretely $\cdot_{2,\infty}$ (recall that $(1/\sqrt{np})\cdot_F \leq (1/\sqrt{n})\cdot_{2,\infty}$). That is, the typical error bound for $(1/\sqrt{np})\cdot_F$ is not sufficient to provide guarantees for PCR with error-in-variables. Finally, the setup of this work, i.e., error-in-variables in high dimension, allows for both $n \ll p$ and $n \gg p$; the current fPCA literature only allows for $n \gg p$. Indeed, for this closely related line of fPCA literature, our work suggests a few interesting directions for future research: (1) allow the sampling locations to be different across the n measurements, provided there is sufficient overlap rather than requiring them to be the same; (2) allow for n trajectories and p observed locations per trajectory to scale simultaneously rather than a requirement of $n \gg p$; (3) extend fPCA guarantees for computationally efficient methods like HSVT.

There has also been work on functional principal component regression (fPCR), which extends PCR to allow for β^* to be an infinite-dimensional parameter as opposed to a high-dimensional parameter that is considered in this work. In particular, Hall and Horowitz (2007) and Cai and Hall (2006) consider the problem of parameter estimation and prediction error for fPCR, respectively. However, they focus on the setting *without* error-in-variables. As noted above, parameter estimation and test prediction error at the fast rate of $\tilde{O}(1/n)$ for PCR with error-in-variables, even in the finite-dimensional case, has remained elusive. Extending our results on parameter estimation and prediction error for fPCR with error-in-variables remains interesting future work.

■ 2.2.3 Error-in-variables

In what follows, we highlight a few key comparisons between this work and prominent works in the high-dimensional error-in-variables literature, cf. Loh and Wainwright (2012), Datta and Zou (2017), Rosenbaum and Tsybakov (2010), Rosenbaum and Tsybakov (2013), Belloni et al. (2017a), Belloni et al. (2017b), Chen and Caramanis (2012), Chen and Caramanis (2013), Kaul and Koul (2015).

Out-of-Sample Predictions.

The focus of the literature has been on parameter estimation. As such, these works do not provide extensions of their algorithms to produce predictions in the presence of noisy and partially observed test covariates; hence, even with knowledge of the exact β^*, it is unclear how these previous results can be extended to establish generalization error bounds. In contrast, PCR naturally handles this setting, as shown in Section 2.4.

Knowledge of Noise Distribution.

The algorithms furnished in prior works explicitly utilize knowledge of the *unknown* covariance of $\boldsymbol{W}$ to recover β^*. In particular, these algorithms typically "correct" the covariance of $\boldsymbol{Z}$ by subtracting this noise covariance, i.e., $\boldsymbol{Z}^T\boldsymbol{Z} - \mathbb{E}[\boldsymbol{W}^T\boldsymbol{W}]$. To carry out such a correction, one must assume access to either oracle knowledge of $\mathbb{E}[\boldsymbol{W}^T\boldsymbol{W}]$ or a good data-driven estimator for it. As noted by Chen and Caramanis (2013), such an estimator can be costly or simply infeasible in many practical settings. PCR, on the other hand, does not require any such knowledge. Formally, the first step in PCR, which finds a low-rank approximation of $\boldsymbol{Z}$, *implicitly* de-noises the covariates without utilizing knowledge of the noise distribution. The trade-off in PCR is that our results only hold if the number of retained singular components is chosen to be the rank of $\boldsymbol{X}$. However, as we will see in Section 2.4.4, there exists numerous data-driven methods to choose this hyper-parameter. Indeed, an interesting future direction is to analyze PCR when this hyper-parameter is misspecified.

Operating Assumptions.

Below, we compare and relate our primary structural assumptions (Assumptions 2 and 6, 7) with those typically made in the literature.

Low-rank vis-á-vis sparsity. Arguably, one of the most popularly endowed structures in high-dimensional regression is sparsity in the model parameter, β^*. Specifically, it is commonly posited that β^* is r-sparse, i.e., β^* has at most r nonzero entries (Loh and Wainwright (2012); Datta and Zou (2017); Rosenbaum and Tsybakov (2010) to name a few). In contrast, this work assumes that the underlying training covariate matrix $\boldsymbol{X}$ is low-rank (Assumption 2), i.e., the spectral profile of $\boldsymbol{X}$ is described by r nonzero singular values. These notions of sparsity are related. In particular, if rank(X) $= r$, then there exists an r-sparse $\widetilde{\beta}$ such that $X\beta^* = X\widetilde{\beta}$ (see Proposition 3.4 of Agarwal et al. (2021e)); meanwhile, if β^* is r-sparse, then it is not hard to verify that there exists a $\tilde{X}$ of rank r that also provides equivalent responses. In other words, one needs sparsity of some form for consistent estimation in a high-dimensional setting, and the two perspectives described above can be viewed as complementary to each other. We note that the low-rank assumption on X can be tested in a data-driven way by simply inspecting the singular values of Z, as described in Section 2.4.4. Further, it is well-established that (approximately) low-rank matrices are abundant in real-world data science applications (see Udell and Townsend (2019); Xu (2017a) and references therein).

Well-balanced spectra vis-á-vis restricted eigenvalue condition. A secondary condition that is often assumed in the literature captures the amount of "information spread" across the rows and columns of the covariates X, which leads to a bound on the smallest singular value of X. Specifically, prior works often assume that a type of restricted eigenvalue condition (see Definitions 1 and 2 in Loh and Wainwright (2012)) is satisfied for the empirical estimate of the covariance of X. In comparison, to obtain "fast" rate results, this work assumes the spectra of X is well-balanced (Assumptions 6 and 7). We emphasize that the assumption of a well-balanced spectra is *not* necessary for consistent estimation, but rather is one example that yields a reasonable signal-to-noise ratio, which guarantees both vanishing parameter estimation *and* out-of-sample prediction errors at a "fast" rate of $\tilde{O}(1/n)$. We note that in many previous works in the high-dimensional regression literature, the restricted eigenvalue condition, or variants of it, are shown to hold with high probability (w.h.p.) if each row of X is sampled i.i.d. (or at least, independently) from a mean zero sub-gaussian distribution. Such a data generating process also implies that the largest and smallest singular values of X are $\tilde{O}(\sqrt{n} + \sqrt{p})$. However, if we assume that X has rank r and each entry of $X = \Theta(1)$, then one can easily verify that the largest singular value of X is $\Omega(\sqrt{np/r})$. This difference in the typical magnitude of the largest singular value reflects the difference in applications in which a restricted eigenvalue assumption versus a low-rank assumption is likely to hold. The restricted eigenvalue assumption is particularly suited in engineering applications such as compressed sensing where one gets to *design* X. The applications arising in the social or life sciences primarily involve performing inference with observational data. In such settings, a low-rank assumption on X is likely more suitable to capture the latent structure amongst the observed covariates. However, the well-balanced spectra condition is similar to the restricted eigenvalue condition in that it requires the smallest non-zero singular value of X to be of the same order as that of the largest singular value (i.e., $s_1, s_r = \Theta(\sqrt{np/r})$, where s_1, s_r denotes the largest and smallest singular values of $\boldsymbol{X}$, respectively).

Indeed, such assumptions or analogous versions to it are pervasive across many fields. For instance, within the econometrics factor model literature, it is standard to assume that the factor structure is separated from the idiosyncratic errors (e.g., Assumption A of Bai and Ng (2020)); within the robust covariance estimation literature, this assumption is closely related to the notion of pervasiveness (see Fan et al. (2018)); within the matrix/tensor estimation literature, it is assumed that the non-zero singular values are of the same order to achieve minimax optimal rates (e.g., Cai et al. (2019)). The well-balanced spectra has also been shown to hold w.h.p. for the embedded Gaussians model, which is a canonical probabilistic generating process used to analyze probabilistic PCA (see Tipping and Bishop (1999); Bishop (1999) and Proposition 4.2 of Agarwal et al. (2021e)). Ultimately, the well-balanced spectra and restricted eigenvalue conditions require the signal is well-spread across the covariates; for a detailed comparison between the two, please see Section 3.5 in Agarwal et al. (2021e). Finally, a practical benefit of Assumptions 6 and 7 is that, like Assumption 2, they can be empirically verified following the same procedure described in Section 2.4.4.

■ 2.2.4 Linear Regression with Hidden Confounding

The high-dimensional error-in-variables regression setup is related to linear regression with hidden confounding, a common model within the causal inference and econometrics literatures (see Guo et al. (2020); Ćevid et al. (2020) and references therein). As noted by Guo et al. (2020), a particular class of error-in-variables models can be reformulated as linear regression with hidden confounding. Using our notation, they consider a high-dimensional model where the rows of $\boldsymbol{X}$ are sampled i.i.d. As such, $\boldsymbol{X}$ can be full-rank, but $\boldsymbol{W}$ is assumed to have low-rank structure. Here, the aim is to estimate β^*, which is assumed to be sparse. In comparison, we place the low-rank assumption on $\boldsymbol{X}$ and assume the rows of $\boldsymbol{W}$ are sampled independently and can be of full-rank. Notably, for this setup, Ćevid et al. (2020) "deconfounds" the observed covariates $\boldsymbol{Z}$ by a spectral transformation of its singular values. Indeed, it is interesting future work to analyze PCR for this important and closely related setting.

■ 2.3 Setup

In this section, we provide a precise description of our problem, including our observations and assumptions.

■ 2.3.1 Observation Model

As described in Section 4.1, we have access to n labeled observations $\{(y_i, z_i) : i \leq n\}$, which we will refer to as our in-sample (train) data; recall that x_i corresponds to the *latent* covariate with respect to z_i. Collectively, we assume (2.1) and (2.2) are satisfied. In addition, we observe $m \geq 1$ unlabeled out-of-sample (test) covariates; for $i \in \{n+1, \dots, n+m\}$, we only observe the noisy covariates z_i, which again correspond to the latent covariates x_i, but we do not have access to the associated response variables y_i.

Throughout, let $\boldsymbol{X} = [x_i^T : i \leq n] \in \mathbb{R}^{n \times p}$ and $\boldsymbol{X}' = [x_i^T : i > n] \in \mathbb{R}^{m \times p}$ represent the underlying train and test covariates, respectively. Similarly, let $\boldsymbol{Z} = [z_i^T : i \leq n] \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Z}' = [z_i^T : i > n] \in \mathbb{R}^{m \times p}$ represent their noisy and partially observed counterparts.

■ 2.3.2 Modeling Assumptions

We make the following assumptions.

Assumption 1 (Bounded covariates). $\|\boldsymbol{X}\|_{\max} \leq 1$, $\|\boldsymbol{X}'\|_{\max} \leq 1$.

Assumption 2 (Covariate rank). $\text{rank}(\boldsymbol{X}) = r$, $\text{rank}(\boldsymbol{X}') = r'$.

Assumption 3 (Response noise). *$\{\varepsilon_i : i \leq n\}$ are a sequence of independent mean zero subgaussian random variables with $\|\varepsilon_i\|_{\psi_2} \leq \sigma$.*

Assumption 4 (Covariate noise). *$\{w_i : i \leq n+m\}$ are a sequence of independent mean zero subgaussian random vectors with $\|w_i\|_{\psi_2} \leq K$ and $\|\mathbb{E}[w_i \otimes w_i]\|_2 \leq \gamma^2$. Further, π_i is a vector of independent Bernoulli variables with parameter $\rho \in (0, 1]$.*

Assumption 5 (Subspace inclusion)**.** *The rowspace of $\boldsymbol{X}'$ is contained within that of $\boldsymbol{X}$, i.e.,* $\text{rowspan}(\boldsymbol{X}') \subseteq \text{rowspan}(\boldsymbol{X})$.

We note that ε_i, w_i, π_i across $i \geq 1$ are not only mutually independent, but also independent of $\boldsymbol{X}$.

■ 2.4 Principal Component Regression

We describe PCR, as introduced in Jolliffe (1982), with a variation to handle missing data. To that end, let $\hat{\rho}$ denote the fraction of observed entries in $\boldsymbol{Z}$. We define $\widetilde{\boldsymbol{Z}} = (1/\hat{\rho})\boldsymbol{Z} = \sum_{i=1}^{n \wedge p} \widehat{s}_i \widehat{u}_i \otimes \widehat{v}_i$, where $\widehat{s}_i \in \mathbb{R}$ are the singular values (arranged in decreasing order) and $\widehat{u}_i \in \mathbb{R}^n, \widehat{v}_i \in \mathbb{R}^p$ are the left and right singular vectors, respectively.

■ 2.4.1 Parameter Estimation

For a given algorithmic parameter $k \in [n \wedge p]$, PCR estimates the model parameter as

$$\widehat{\beta} = \Big(\sum_{i=1}^{k} (1/\widehat{s}_i) \widehat{v}_i \otimes \widehat{u}_i \Big) y. \tag{2.3}$$

■ 2.4.2 Out-of-sample Prediction

Let $\hat{\rho}'$ denote the proportion of observed entries in $\boldsymbol{Z}'$. As before, let $\widetilde{\boldsymbol{Z}}' = (1/\hat{\rho}')\boldsymbol{Z}' = \sum_{i=1}^{m \wedge p} \widehat{s}'_i \widehat{u}'_i \otimes \widehat{v}'_i$, where $\widehat{s}'_i \in \mathbb{R}$ are the singular values and $\widehat{u}'_i \in \mathbb{R}^m, \widehat{v}'_i \in \mathbb{R}^p$ are the left and right singular vectors, respectively. Given algorithmic parameter $\ell \in [m \wedge p]$, let $\widetilde{\boldsymbol{Z}}'^{\ell} = \sum_{i=1}^{\ell} \widehat{s}'_i \widehat{u}'_i \otimes \widehat{v}'_i$, and define the test response estimates as $\widehat{y}' = \widetilde{\boldsymbol{Z}}'^{\ell} \widehat{\beta}$.

If the responses are known to belong to a bounded interval, say $[-b, b]$ for some $b > 0$, then the entries of $\widehat{y}'$ are truncated as follows: for every $i > n$,

$$\widehat{y}_i = \begin{cases} -b & \text{if } \widehat{y}_i \leq -b, \\ \widehat{y}_i & \text{if } -b < \widehat{y}_i < b, \\ b & \text{if } b \leq \widehat{y}_i. \end{cases}$$

2.4.3 Properties of PCR

We state some useful properties of PCR, which we will use extensively throughout this work. These are well-known results, discussed at length in Chapter 17 of Roman (2008) and Chapter 6.3 of Strang (2006).

Property 2.4.1. *Let* $\widetilde{Z}^k = \sum_{i=1}^{k} \widehat{s}_i \widehat{u}_i \otimes \widehat{v}_i$*. Then* $\widehat{\beta}$*, as given in* (2.3)*, also satisfies*

1. $\widehat{\beta}$ *is the unique solution of the following program:*

$$\begin{aligned} &\textit{minimize} \quad \left\|\beta\right\|_2 \quad \textit{over} \quad \beta \in \mathbb{R}^p \\ &\textit{such that} \ \ \beta \in \operatorname*{arg\,min}_{\beta' \in \mathbb{R}^p} y - \widetilde{Z}^k \beta'^2_2. \end{aligned}$$

2. $\widehat{\beta} \in \mathrm{rowspan}(\widetilde{Z}^k)$.

2.4.4 Choosing k & When to use PCR

In general, the correct number of principal components k to use is not known a priori. This is a well-studied problem in the low-rank matrix estimation literature and there exists a suite of principled methods to choose k. These include visual inspections of the plotted singular values (Cattell (1966)), cross-validation (Wold (1978); Owen and Perry (2009)), Bayesian methods (Hoff (2007)), and "universal" thresholding schemes that preserve singular values above a precomputed threshold (Chatterjee (2015); Donoho and Gavish (2013)). A common argument for these approaches is rooted in the underlying assumption that the smallest non-zero singular value of $\boldsymbol{X}$ (i.e., signal) is well-separated from the largest singular value of $\boldsymbol{W}$ (i.e., noise). Specifically, under reasonable signal-to-noise scenarios, Weyl's inequality implies that a "sharp" threshold or gap should exist between the top r singular values and remaining singular values of the observed data $\widetilde{Z}$. This gives rise to a natural "elbow" point and suggests choosing a threshold within this gap, which the methods described above are designed to accomplish. For a graphical depiction of the elbow, please see Figure 3.3.

As such, for a practitioner, a natural data-driven diagnostic of when to use PCR is to simply plot the singular values of $\widetilde{Z}$. If the spectrum does not exhibit this elbow structure (i.e., low-rankness), then PCR (as is) may not be the best suited estimation procedure.

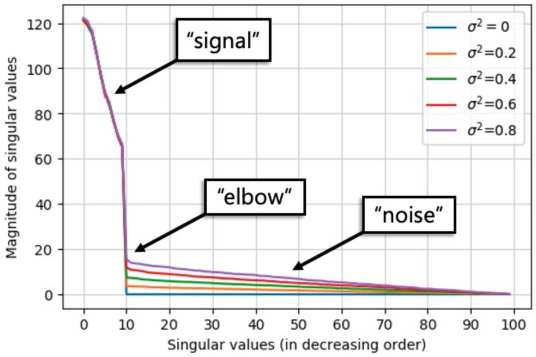


Figure 2.1: Simulation displays the spectrum of $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W} \in \mathbb{R}^{100\times100}$. Here, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$, where the entries of $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{100\times10}$ are sampled independently from $\mathcal{N}(0,1)$; further, the entries of $\boldsymbol{W}$ are sampled independently from $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 \in \{0, 0.2, \dots, 0.8\}$. Across varying levels of noise σ^2, there is a steep drop-off in magnitude from the top to remaining singular values—this marks the "elbow" point. As seen from the figure, the top singular values of $\boldsymbol{Z}$ correspond closely with that of $\boldsymbol{X}$ ($\sigma^2 = 0$), and the remaining singular values are induced by $\boldsymbol{W}$. Thus, $\text{rank}(\boldsymbol{Z}) \approx \text{rank}(\boldsymbol{X}) = 10$.

■ 2.5 Main Results

We state PCR's parameter estimation and generalization properties in this section. For the remainder of the paper, $C(K,\gamma,\sigma) > 0$ will denote any constant that depends only on K, γ, σ, and $C, c > 0$ will denote absolute constants. The values of $C(K,\gamma,\sigma)$, C, and c may change from line to line or even within a line.

■ 2.5.1 Parameter Estimation

In the high-dimensional framework, recall that there are infinitely many feasible models that can satisfy (2.1). Thus, the question remains whether there is a β^* that is recovered by PCR? To the best of our knowledge, despite the popularity of PCR, such a question has yet to be answered in an error-in-variables setting.

We find that PCR recovers the unique model with minimum ℓ_2-norm, i.e., $\beta^* \in \text{rowspan}(\boldsymbol{X})$. We note that this uniqueness follows since every element in the column space of a matrix is associated with a unique element in its row space coupled with any element in its null space. Thus, for the purposes of prediction, it suffices to consider this particular β^* (see Roman (2008), Strang (2006) for details). Further, recall from Property 2.4.1 that PCR enforces $\widehat{\beta} \in \text{rowspan}(\widetilde{\boldsymbol{Z}}^k)$. Hence, if $k = r$ and the rowspace of $\widetilde{\boldsymbol{Z}}^r$ is "close" to the rowspace of $\boldsymbol{X}$, then this suggests $\widehat{\beta} \approx \beta^*$. We highlight that the "noise" in $\boldsymbol{Z}$ arises from the missingness pattern induced by π_i and the measurement error $\boldsymbol{W}$; meanwhile, the "signal" in $\boldsymbol{Z}$ comes from $\boldsymbol{X}$, where its strength is captured by the magnitude of its singular values. This naturally leads to the following "signal-to-noise" ratio definition:

$$\mathrm{snr} := \frac{\rho s_r}{\sqrt{n} + \sqrt{p}}. \tag{2.4}$$

Here, s_r is the smallest non-zero singular value of the signal matrix $\boldsymbol{X}$, ρ determines the fraction of observed entries, and $\sqrt{n} + \sqrt{p}$ is induced by the perturbation in the singular values due to the noise matrix $\boldsymbol{W}$. Indeed, as one would expect, the signal strength s_r scales linearly with ρ. Moreover, from standard concentration results for random sub-gaussian matrices, it follows that $\boldsymbol{W}_2 = \tilde{O}(\sqrt{n} + \sqrt{p})$ (see Lemma 2.12.8). With this notation, we state the main result.

Theorem 2.5.1. *Suppose Assumptions 1, 2, 3, 4 hold. Consider* $\beta^* \in \mathrm{rowspan}(\boldsymbol{X})$ *and PCR with* $k = r$. *Let* $\rho \geq c\frac{\log^2(np)}{np}$, $\mathrm{snr} \geq C(K, \gamma, \sigma)$, $\beta^*{}_2 = \Omega(1), \beta^*{}_1 = O(\sqrt{p})$. *Then with probability at least* $1 - O(1/(np)^{10})$,

$$\widehat{\beta} - \beta^{*2}_2 \leq C(K, \gamma, \sigma) \log(np) \cdot \left(\frac{r\beta^{*2}_2}{\mathrm{snr}^2} + \frac{\beta^{*2}_1}{\mathrm{snr}^4} \right). \tag{2.5}$$

Interpretation. For added interpretability, we suppress dependencies on K, γ, σ for the following discussion. One can then verify that Theorem 2.5.1 implies that a sufficient condition for PCR's parameter estimation error to vanish w.h.p. is

$$\frac{\mathrm{snr}}{\sqrt{r\log(np)}\beta^*{}_2 \ \vee \ \log^{1/4}(np)\beta^{*1/2}_1} \to \infty.$$

Conversely, in Theorem 2.5.2 below, we establish that if $\mathrm{snr} = O(1)$, then the parameter estimation error is lower bounded by a positive, absolute constant; in this sense, Theorem 2.5.1 is nearly tight. The key to establishing this result is to show that the Gaussian location model problem (see Wu (2020)) can be written as an instance of error-in-variables regression.

Theorem 2.5.2. *Suppose* $n = O(p)$ *and* $\mathrm{snr} = O(1)$. *Then,*

$$\inf_{\widehat{\beta}} \sup_{\beta^* \in \mathbb{B}_2} \mathbb{E}\widehat{\beta} - \beta^{*2}_2 = \Omega(1),$$

where $\mathbb{B}_2 = \{v \in \mathbb{R}^p : v_2 \leq 1\}$.

The minimax bound in Theorem 2.5.2 utilizes $\rho = 1$ and does not necessarily capture the refined dependence on ρ. Observe that (2.5) suggests that the error decays as ρ^{-4}. While this dependency on ρ may not be optimal, similar dependencies have appeared in the error bound within the error-in-variables literature, e.g., see Loh and Wainwright (2012) and references therein. Returning to Theorem 2.5.1, if $\text{snr} = \Omega(\rho\sqrt{(n \wedge p)/r})$, then (2.5) simplifies as

$$\widehat{\beta} - \beta^{*2}_{2} \leq \frac{C(K, \gamma, \sigma) r^2 \log np}{\rho^4 (n \wedge p)} \beta^{*2}_{2}. \tag{2.6}$$

Next, we describe a natural setup under which the conditions leading to (2.6) hold. To that end, we introduce Assumption 6 (recall that its interpretation is discussed in detail in Section 2.2.3).

Assumption 6 (Balanced spectra: training covariates)**.** *The r non-zero singular values s_i of $\boldsymbol{X}$ satisfy $s_i = \Theta(\sqrt{np/r})$.*

Corollary 2.5.1. *Let the setup of Theorem 2.5.1 and Assumption 6 hold. Then with probability at least $1 - O(1/(np)^{10})$,*

$$\widehat{\beta} - \beta^{*2}_{2} \leq \frac{C(K, \gamma, \sigma) r^2 \log(np)}{\rho^4 (n \wedge p)} \beta^{*2}_{2},$$

where $C(K, \gamma, \sigma)$ is a large enough constant dependent on K, γ, σ.

Proof. By Assumption 6, we have $s_r = \Theta(\sqrt{\frac{np}{r}})$. This yields

$$\text{snr} = \frac{\rho s_r}{\sqrt{n} + \sqrt{p}} \geq \frac{c\rho\sqrt{np}}{\sqrt{r(n+p)}} \geq c\rho\sqrt{(n \wedge p)/r},$$

i.e., $\text{snr} = \Omega(\rho\sqrt{(n \wedge p)/r})$. Using this lower bound on snr, the assumptions $\beta^*{}_2 = \Omega(1)$ and $\beta^*{}_1 = O(\sqrt{p})$, and simplifying (2.5), we complete the proof of Corollary 2.5.1. ■

■ 2.5.2 Out-of-sample Prediction Error

Next, we bound PCR's out-of-sample prediction error in the presence of corrupted unseen data, defined as

$$\mathrm{MSE}_{\text{test}} := \frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_{n+i} - \langle x_{n+i}, \beta^* \rangle)^2. \tag{2.7}$$

We define some more useful quantities. Let $s_\ell, s'_\ell \in \mathbb{R}$ be the ℓ-th singular values of $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. Recall from Section 2.4 that $\widehat{s}_\ell, \widehat{s}'_\ell$ are the ℓ-th singular values of $\widetilde{\boldsymbol{Z}}$ and $\widetilde{\boldsymbol{Z}}'$, respectively. Analogous to (2.4), we define a signal-to-noise ratio for the test covariates:

$$\mathrm{snr}_{\text{test}} := \frac{\rho s'_{r'}}{\sqrt{m} + \sqrt{p}}. \tag{2.8}$$

In Theorem 2.5.3, we bound $\mathrm{MSE}_{\text{test}}$ both in probability and in expectation with respect to these quantities.

Theorem 2.5.3. *Let the setup of Theorem 2.5.1 and Assumption 5 hold. Consider PCR with $\ell = r'$. Let $\rho \geq c\frac{\log^2(mp)}{mp}$. Then, with probability at least $1 - O(1/((n \wedge m)p)^{10})$,*

$$\mathrm{MSE}_{\text{test}} \leq C(K, \gamma, \sigma)\log((n \vee m)p)\left(\frac{r\left(1 \vee \frac{p}{m}\right)\beta^{*2}_1}{\rho^2 \mathrm{snr}^2} + \frac{r(n \vee p)\beta^{*2}_1}{\mathrm{snr}^4} + \frac{r\beta^{*2}_1}{\mathrm{snr}^2_{\text{test}} \wedge m} + \frac{\sqrt{n}\beta^*{}_1}{\mathrm{snr}^2}\right).$$

Further, for all $i > n$, if $\langle x_i, \beta^* \rangle \in [-b, b]$ and $\widehat{y}'$ is appropriately truncated, then

$$\mathbb{E}[\mathrm{MSE}_{\text{test}}] \leq C(K, \gamma, \sigma) r \log((n \vee m)p)\left(\frac{\left(1 \vee \frac{p}{m}\right)}{\rho^2 \mathrm{snr}^2} + \frac{(n \vee p)}{\mathrm{snr}^4} + \frac{1}{\mathrm{snr}^2_{\text{test}} \wedge m}\right)\beta^{*2}_1 + \frac{Cb^2}{((n \wedge m)p)^{10}}.$$

Interpretation. For interpretability, we suppress dependencies on σ, K, γ, and assume $p = \Theta(m)$ and $m \to \infty$. One can then verify that Theorem 2.5.3 implies that a sufficient condition for PCR's expected test prediction error to vanish is

$$\frac{\rho\mathrm{snr} \wedge \mathrm{snr}_{\text{test}}}{\sqrt{r\log((n \vee m)p)}\beta^*{}_1} \to \infty, \qquad \frac{\mathrm{snr}}{(r\log(np))^{1/4}(n \vee p)^{1/4}\beta^{*1/2}_1} \to \infty.$$

As with Theorem 2.5.1, we specialize Theorem 2.5.3 in Corollary 2.5.2 to the setting where $\mathrm{snr} = \Omega(\rho\sqrt{(n \wedge p)/r})$ and $\mathrm{snr}_{\text{test}} = \Omega(\rho\sqrt{(m \wedge p)/r})$. A sufficient condition for these lower bounds on snr and $\mathrm{snr}_{\text{test}}$ to hold is if the non-zero singular values of $\boldsymbol{X}$ and $\boldsymbol{X}'$ are well-balanced, i.e., Assumptions 6 and 7 hold.

Assumption 7 (Balanced spectra: testing covariates)**.** *The r' non-zero singular values s'_i of $\boldsymbol{X}'$ satisfy $s'_i = \Theta(\sqrt{mp/r'})$.*

Corollary 2.5.2. *Let the setup of Corollary 2.5.1 and Theorem 2.5.3 hold. Further, let Assumption 7 hold. Then with probability at least $1 - O(1/((n \wedge m)p)^{10})$,*

$$\mathrm{MSE}_{\text{test}} \leq \frac{C(K,\gamma,\sigma)r^3 \log((n \vee m)p)}{\rho^4} \left(\left(\frac{1 \vee \frac{p}{m}}{n \wedge p} + \frac{n \vee p}{(n \wedge p)^2} + \frac{1}{m} \right) \beta^{*2}_{1} + \left(\frac{\sqrt{n}}{n \wedge p} \right) \beta^{*}{}_{1} \right).$$

Further,

$$\mathbb{E}[\mathrm{MSE}_{\text{test}}] \leq \frac{C(K,\gamma,\sigma)r^3 \log((n \vee m)p)}{\rho^4} \left(\frac{1 \vee \frac{p}{m}}{n \wedge p} + \frac{n \vee p}{(n \wedge p)^2} + \frac{1}{m} \right) \beta^{*2}_{1} + \frac{Cb^2}{((n \wedge m)p)^{10}}.$$

Proof. Using identical arguments to those used in the proof of Corollary 2.5.1, we have that Assumption 6 implies $\text{snr} \geq c\rho\sqrt{(n \wedge p)/r}$, and Assumptions 5 and 7 imply $\text{snr}_{\text{test}} \geq c\rho\sqrt{(m \wedge p)/r}$ Plugging these lower bounds on snr and snr_{test} into the bounds in Theorem 2.5.3 and simplifying completes the proof. ■

Interpretation. For the following discussion, we suppress dependencies on K, γ, σ, r and log factors, assume $\rho = \Theta(1)$ and only consider the scaling with respect to n, m, p. Hence, Corollary 2.5.2 implies that if $p = o(n(n \wedge m)), n = o(p^2)$,[1] then the out-of-sample prediction error vanishes to zero both in expectation and w.h.p., as $n, m, p \to \infty$. If we make the additional assumption that $n = \Theta(p)$ and $p = \Theta(m)$, then Corollary 2.5.2 implies the error scales as $\tilde{O}(1/n)$ in expectation. This improves upon the best known rate of $\tilde{O}(1/\sqrt{n})$, established in Agarwal et al. (2019b, 2021e) (notably, these works do not provide a high probability bound). We re-emphasize that we consider a fixed design setting; as such, our generalization error bounds do not make any distributional assumptions on $\boldsymbol{X}$ and $\boldsymbol{X}'$, which Agarwal et al. (2019b, 2021e) require in order to leverage standard Rademacher tools for their analysis. Finding the optimal relative scalings of n, m, p to achieve vanishing prediction error is left for future work.

A Complementary Perspective on Generalization

As discussed above, Theorem 2.5.3 and Corollary 2.5.2 do *not* require any distributional assumptions on the in- and out-of-sample covariates, but rather rely on a *purely linear*

[1] Practically speaking, this condition is not binding, i.e., if $n = \Omega(p^2)$, then we can sample a subset of the training data to satisfy it. Therefore, this condition is likely an artefact of our analysis.

algebraic condition given by Assumption 5. Intuitively, because we consider $\beta^* \in \text{rowspan}(\boldsymbol{X})$, it follows that generalization is achievable if the rowspace of the out-of-sample covariates lies wthin that of the in-sample covariates, i.e., the out-of-sample covariates are at most as "rich" or "complex" as the in-sample covariates used for learning in a linear algebraic sense. As we have seen in our test error results and will shortly see in our simulations in Section 2.7.3, Assumption 5 is the key condition that allows PCR to generalize, even if the in-sample and out-of-sample data obey different distributions. Thus, Assumption 5 offers a complementary, distribution-free perspective to generalization, and has possible implications to transfer learning and learning with covariate shifts.

■ 2.6 Synthetic Controls

In this section, we overview the synthetic controls framework, connect it to the (high-dimensional) error-in-variables with fixed design, and state our results from Section 2.5 in this context.

■ 2.6.1 Setup

The synthetic controls framework considers the panel data setting, where observations of units are collected over time. More formally, let there be $p+1$ units indexed as $\{0, \dots, p\}$. Without loss of generality, let unit 0 be our *target* unit of interest; we refer to the remaining p units as *donors*. Consider two interventions: *control* and *treatment*. For all $p+1$ units, we observe their outcomes under control for the first n time steps. For target unit 0, we observe its outcomes under treatment for time steps $\{n+1, \dots, n+m\}$, but continue to observe outcomes under control for the donor units. As such, we refer to the first n time steps as the *pre-intervention* period and the remaining m time steps as the *post-intervention* period. Our interest is to estimate the target unit's counterfactual (unobserved) outcomes under control during the post-intervention period.

Donor observations under control. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{X}' \in \mathbb{R}^{m \times p}$ represent the ground truth outcomes under control associated with the p donor units during the pre- and post-intervention periods, respectively, i.e., the jth column of $\boldsymbol{X}$ and $\boldsymbol{X}'$ represents the underlying outcomes under control for the jth unit for the first n and last m time steps, respectively. Rather than observing $\boldsymbol{X}$ and $\boldsymbol{X}'$, however, we only have access to $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$

and $\boldsymbol{Z}' \in \mathbb{R}^{m \times p}$, which are noisy (and potentially sparse) realizations of $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. The distributional characteristics of $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ obey the conditions described in Section 2.3. Thus, using the potential outcomes language of Neyman (1923) and Rubin (1974), we refer to $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ as the matrices of potential outcomes under control, and $\boldsymbol{X}$ and $\boldsymbol{X}'$ as the matrices of expected potential outcomes under control for the donor units.

Target unit observations under control. For target unit 0, let $y_i \in \mathbb{R}$ for $i \in [n+m]$ denote the potential outcome under control at the ith time step. As stated above, since the target unit is exposed to treatment during time steps $n+1, \dots, n+m$, we only observe its outcomes under control during the pre-intervention period, i.e., we observe $y_{\text{pre}} = [y_i : i \leq n] \in \mathbb{R}^n$. We summarize the synthetic controls objective as follows: given $\{y_{\text{pre}}, \boldsymbol{Z}, \boldsymbol{Z}'\}$, the interest is to recover the target unit's expected counterfactual outcomes under control during the post-intervention period, $\mathbb{E}[y_{\text{post}}] \in \mathbb{R}^m$, where $y_{\text{post}} = [y_i : n+1 \leq i \leq n+m]$.

Existence of synthetic controls. The key modeling premise within the synthetic controls framework is that the target unit's outcomes under control can be expressed as some linear combination of outcomes under control of the donor units. In our setup, this translates to the existence of a linear model $\beta^* \in \mathbb{R}^p$ satisfying

$$y_i = \langle x_i, \beta^* \rangle + \varepsilon_i,$$

for all $i \in [n+m]$; here, $\varepsilon_i \in \mathbb{R}$ models the idiosyncratic randomness in the potential outcomes. Notably, we remark that the typical factor model structure utilized in the synthetic controls literature (e.g., Abadie et al. (2010); Abadie and Gardeazabal (2003); Abadie (2020) and references therein), implies that the concatenated matrix of dimensions $(n+m) \times (p+1)$, whose first row is given by $(\mathbb{E}[y_{\text{pre}}], \mathbb{E}[y_{\text{post}}])$ and remaining rows are given by $(\boldsymbol{X}, \boldsymbol{X}')$, is low-rank. As such, it follows that $\boldsymbol{X}$ and $\boldsymbol{X}'$ are necessarily low-rank, and β^* exists with high probability (see Agarwal et al. (2021e) for details on the latter note). Moreover, because we consider a fixed design setting, we do not enforce the latent time factors associated with the pre- and post-intervention periods to be sampled i.i.d., as is standard in the literature, cf. Agarwal et al. (2021e). This allows us to model settings with underlying time trends or shifting ideologies. In summary, the objective in synthetic controls is identical to that of out-of-sample prediction in (high-dimensional) error-in-variables with fixed design.

■ 2.6.2 Robust Synthetic Controls (RSC)

The robust synthetic controls (RSC) method of Amjad et al. (2018) uses PCR to estimate β^*. Specifically, it produces $\widehat{\beta}$ as per the method described in Section 2.4.1 using $\{y_{\text{pre}}, \boldsymbol{Z}\}$. Subsequently, RSC produces $[\widehat{y}_i : n+1 \leq i \leq n+m]$ using $\boldsymbol{Z}'$ with $\widehat{\beta}$, as per the method described in Section 2.4.2. Below, we use our formal results on PCR, as stated in Section 2.5, to establish statistical guarantees of the RSC method in the fixed design setting, which has been absent.

Theoretical results. Recall that the aim of synthetic controls is to estimate the counterfactual outcomes under control for the target unit during the post-intervention period. As such, the primary performance goal is to minimize the post-intervention (out-of-sample) prediction error, MSE_{test}, as defined in (2.7). Corollary 2.5.2 implies the following result.

Theorem 2.6.1. *Suppose Assumptions 1 to 7 hold. Consider* $\beta^* \in \text{rowspan}(\boldsymbol{X})$ *and let* $k = r$ *for PCR within RSC. Let* $\rho \geq c\frac{\log^2(np)}{np}$, $\text{snr}, \text{snr}_{test} \geq C(K, \gamma, \sigma)$, $\beta^*{}_2 = \Omega(1), \beta^*{}_1 = O(\sqrt{p})$. *Then with probability at least* $1 - O(1/((n \wedge m)p)^{10})$,

$$\text{MSE}_{\text{test}} \leq \frac{C(K,\gamma,\sigma) r^3 \log((n \vee m)p)}{\rho^4} \left(\left(\frac{1 \vee \frac{p}{m}}{n \wedge p} + \frac{n \vee p}{(n \wedge p)^2} + \frac{1}{m} \right) \beta^{*2}_{1} + \left(\frac{\sqrt{n}}{n \wedge p} \right) \beta^*{}_1 \right).$$

Further,

$$\mathbb{E}[\text{MSE}_{\text{test}}] \leq \frac{C(K,\gamma,\sigma) r^3 \log((n \vee m)p)}{\rho^4} \left(\frac{1 \vee \frac{p}{m}}{n \wedge p} + \frac{n \vee p}{(n \wedge p)^2} + \frac{1}{m} \right) \beta^{*2}_{1} + \frac{Cb^2}{((n \wedge m)p)^{10}}.$$

Implications and comparison with literature. For the setting of synthetic controls, it is reasonable to consider $\beta^*{}_1 = O(1)$. As explained in Section 2.5.2, under the interpretation of Corollary 2.5.2, if $\rho = \Theta(1)$, $p = o(n(n \wedge m)), n = o(p^2)$, then the counterfactual prediction error vanishes to zero both in expectation and w.h.p., as $n, m, p \to \infty$. If we make the additional assumption that $n = \Theta(p), \rho = \Theta(1)$ and $p = \Theta(m)$, then $\mathbb{E}[\text{MSE}_{\text{test}}] = \tilde{O}(1/n)$. This improves upon the best known rate of $\tilde{O}(1/\sqrt{n})$, established in Agarwal et al. (2019b, 2021e), which considers a random design setting where the latent time factors are sampled i.i.d. (notably, they also do not provide a high probability bound).

In the synthetic controls literature, numerous variants of the original approach Abadie et al. (2010); Abadie and Gardeazabal (2003) have been proposed, see Abadie (2020). While it is not obvious which of these methods provides the best performance a priori, our result suggests that RSC is a natural candidate, especially in the presence of noisy observations and low-rank structure (which can be empirically tested).

■ 2.7 Simulations

In this section, we present illustrative simulations to support our theoretical results.

■ 2.7.1 Parameter Estimation

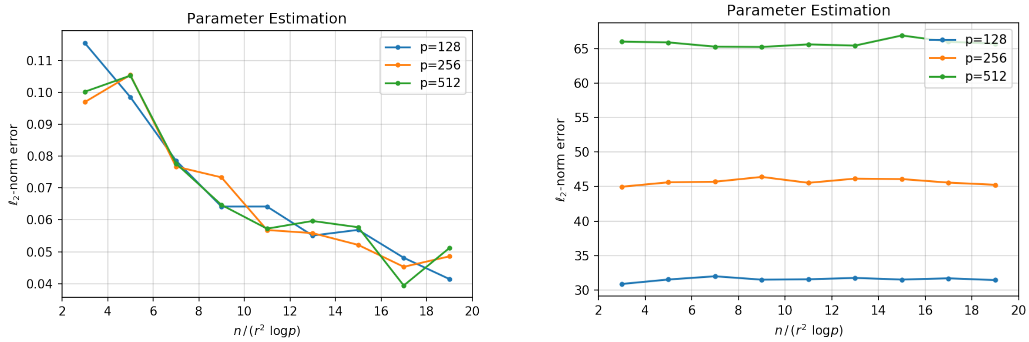
The purpose of this simulation is to demonstrate that PCR does indeed identify the unique minimum ℓ_2-norm β^*.

Generative Model

We construct covariates $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ via the classical probabilistic PCA model, cf. Tipping and Bishop (1999). That is, we first generate $\boldsymbol{X}_r \in \mathbb{R}^{n \times r}$ by independently sampling each entry from a standard normal distribution. Then, we sample a transformation matrix $\boldsymbol{Q} \in \mathbb{R}^{r \times p}$, where each entry is uniformly and independently sampled from $\{-1/\sqrt{r}, 1/\sqrt{r}\}$. The final matrix then takes the form $\boldsymbol{X} = \boldsymbol{X}_r \boldsymbol{Q}$. We choose $\text{rank}(\boldsymbol{X}) = r = p^{\frac{1}{3}}$, where $p \in \{128, 256, 512\}$.

Next, we generate $\beta \in \mathbb{R}^p$ by sampling from a multivariate standard normal vector with independent entries. The noiseless response vector $a \in \mathbb{R}^n$ is defined to be $a = \boldsymbol{X}\beta$. Finally, as motivated by Property 2.4.1, the minimum ℓ_2-norm model of interest, β^*, is computed as $\beta^* = \boldsymbol{X}^\dagger a$, where $\boldsymbol{X}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{X}$.

We consider an additive noise model. Specifically, the entries of $\varepsilon \in \mathbb{R}^n$ are sampled i.i.d. from a normal distribution with mean 0 and variance $\sigma^2 = 0.2$. The entries of $\boldsymbol{W} = [w_i^T] \in \mathbb{R}^{n \times p}$ are sampled in an identical fashion. We then define our observed response vector as $y = a + \varepsilon$ and observed covariate matrix as $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$. For simplicity, we do not mask any of the entries.



(a) ℓ_2-norm error of $\widehat{\beta}$ with respect to the min. ℓ_2-norm solution of (2.1), i.e., β^*.

(b) ℓ_2-norm error of $\widehat{\beta}$ with respect to a random solution of (2.1).

Figure 2.2: Plots of ℓ_2-norm errors, i.e., $\widehat{\beta} - \beta^*{}_2$ in (2.2a) and $\widehat{\beta} - \beta_2$ in (2.2b), versus the rescaled sample size $n/(r^2 \log p)$ after running PCR with rank $r = p^{\frac{1}{3}}$. As predicted by Theorem 2.5.1, the curves for different values of p under (2.2a) roughly align and decay to zero as n increases.

Results

Using the observations $(y, \boldsymbol{Z})$, we perform PCR to yield $\widehat{\beta}$. To show that PCR can accurately recover β^*, we compute the ℓ_2-norm parameter estimation error, or root-mean-squared-error (RMSE), with respect to β^* and β in Figures 2.2a and 2.2b, respectively. As suggested by Figure 2.2a, the RMSE with respect to β^* roughly aligns for different values of p, after rescaling the sample size as $n/(r^2 \log p)$, and decays to zero as the sample size increases; this is predicted by Theorem 2.5.1. On the other hand, Figure 2.2b shows that the RMSE with respect to β stays roughly constant across different values of p. Therefore, as established in Agarwal et al. (2019b), PCR performs implicit regularization by not only de-noising the observed covariates, but also finding the minimum-norm solution.

■ 2.7.2 Out-of-sample Prediction: PCR vs. Ordinary Least Squares

The purpose of this simulation is to demonstrate the benefit of the implicit de-noising effect of PCR vs. OLS.

Generative Model

For each experiment, we let $n = m = p = 1000$. We generate training and testing covariates $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{R}^{1000 \times 1000}$, respectively, with $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}') = 10$ and $\text{rowspan}(\boldsymbol{X}') \subseteq$

rowspan($\boldsymbol{X}$), i.e., Assumption 5 holds. To do so, we sample $\boldsymbol{U}, \boldsymbol{U}', \boldsymbol{V} \in \mathbb{R}^{1000 \times 10}$ by independently sampling each entry from a standard normal distribution. Then, we define $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ and $\boldsymbol{X}' = \boldsymbol{U}'\boldsymbol{V}^T$.

We then generate $\beta \in \mathbb{R}^{1000}$ as in Section 2.7.1, and use it to produce $a = \boldsymbol{X}\beta$ and $a' = \boldsymbol{X}'\beta$. Similarly, we generate the response noise $\varepsilon \in \mathbb{R}^{1000}$ and covariate noises $\boldsymbol{W}, \boldsymbol{W}' \in \mathbb{R}^{1000 \times 1000}$ by independently sampling each entry from a normal distribution with mean zero and variance σ^2, where $\sigma^2 \in \{0.1, 0.2, \dots 1.0\}$. Again, for simplicity, we do not mask any of the entries. We then define our observed response as $y = a + \varepsilon$, and the observed training and testing covariates as $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$ and $\boldsymbol{Z}' = \boldsymbol{X}' + \boldsymbol{W}'$, respectively.

Results

Using the observations $(y, \boldsymbol{Z}, \boldsymbol{Z}')$, we perform PCR as in Section 2.4.2 to produce $\widehat{y}'_{\text{pcr}} \in \mathbb{R}^{1000}$. The OLS out-of-sample estimates are produced using the same algorithm as in Section 2.4.2 *without* the singular value thresholding step on either $\boldsymbol{Z}$ or $\boldsymbol{Z}'$, i.e., we do not de-noise the training nor testing covariates. The estimates produced from OLS are defined as $\widehat{y}'_{\text{ols}} \in \mathbb{R}^{1000}$. In both PCR and OLS, we do not truncate the estimated entries. For any estimate $\widehat{y}' \in \mathbb{R}^{1000}$, we define the out-of-sample mean squared error (MSE) as $(1/1000)\widehat{y}' - a'^2_2$. In Figure 2.3, as we vary the level of response and covariate noise σ^2, we plot the MSE of $\widehat{y}'_{\text{pcr}}$ versus that of $\widehat{y}'_{\text{ols}}$. The MSE of OLS is between three to four orders of magnitude larger than that of PCR across all noise levels. We remark that even when $\sigma^2 = 0.1$, the error of OLS is almost three orders of magnitude larger than PCR – this indicates the significant level of bias that is introduced even with minimal measurement error. In essence, this stresses the importance of de-noising the training and testing covariates via singular value thresholding.

■ 2.7.3 Out-of-sample Prediction: Robustness of PCR to Distribution Shifts

The purpose of this simulation is to demonstrate that PCR can generalize even when the testing covariates are not only corrupted, but also sampled from a different distribution than the training covariates.

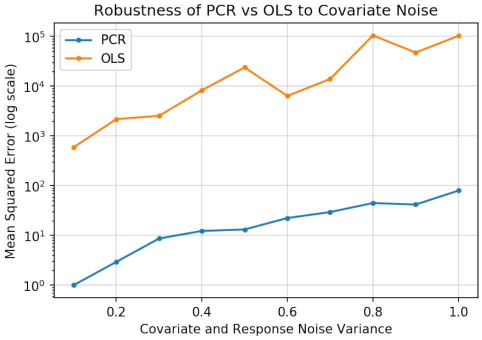


Figure 2.3: MSE plot of $\widehat{y}'_{\text{pcr}}$ (blue) versus $\widehat{y}'_{\text{ols}}$ (orange) as we increase the level of covariate and response noises. While PCR's error scales gracefully with the level of noise, OLS suffers large amounts of bias, even in the presence of small amounts of measurement error.

Generative Model

Throughout, we let $n = m = p = 1000$. We generate the training covariates as in Section 2.7.2, i.e., $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$, where the entries of $\boldsymbol{U}, \boldsymbol{V}$ are sampled independently from a standard normal distribution. Next, we generate four different out-of-sample covariates, defined as $\boldsymbol{X}'_{N_1}, \boldsymbol{X}'_{N_2}, \boldsymbol{X}'_{U_1}, \boldsymbol{X}'_{U_2}$ via the following procedure: We independently sample the entries of $\boldsymbol{U}'_{N_1}$ from a standard normal distribution, and define $\boldsymbol{X}'_{N_1} = \boldsymbol{U}'_{N_1}\boldsymbol{V}^T$. We define $\boldsymbol{X}'_{N_2} = \boldsymbol{U}'_{N_2}\boldsymbol{V}^T$ similarly with the entries of $\boldsymbol{U}'_{N_2}$ sampled from a normal distribution with mean zero and variance 5. Next, we independently sample the entries of $\boldsymbol{U}'_{U_1}$ from a uniform distribution with support $[-\sqrt{3}, \sqrt{3}]$, and define $\boldsymbol{X}'_{U_1} = \boldsymbol{U}'_{U_1}\boldsymbol{V}^T$. We define $\boldsymbol{X}'_{U_2} = \boldsymbol{U}'_{U_2}\boldsymbol{V}^T$ similarly with the entries of $\boldsymbol{U}'_{U_2}$ sampled from a uniform distribution with support $[-\sqrt{15}, \sqrt{15}]$.

By construction, we note that the mean and variance of the entries in $\boldsymbol{X}'_{U_1}$ match that of $\boldsymbol{X}'_{N_1}$; an analogous relationship holds between $\boldsymbol{X}'_{U_2}$ and $\boldsymbol{X}'_{N_2}$. Further, while $\boldsymbol{X}'_{N_1}$ follows the same distribution as that of $\boldsymbol{X}$, we note that there is a clear distribution shift from $\boldsymbol{X}$ to $\boldsymbol{X}'_{U_1}, \boldsymbol{X}'_{N_2}, \boldsymbol{X}'_{U_2}$.

We proceed to generate β as in Section 2.7.2. We then define $a'_{N_1} = \boldsymbol{X}'_{N_1}\beta$, and define $a'_{N_2}, a'_{U_1}, a'_{U_2}$ analogously. Further, the response noise ε and covariate noises $\boldsymbol{W}, \boldsymbol{W}'$ are constructed in the same fashion as described in Section 2.7.2, where the variance again follows $\sigma^2 \in \{0.1, 0.2, \dots 1.0\}$. We define the training responses as $y = \boldsymbol{X}\beta + \varepsilon$ and observed training covariates as $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$. The first set of observed testing covariates is defined as $\boldsymbol{Z}'_{N_1} = \boldsymbol{X}'_{N_1} + \boldsymbol{W}'$, with analogous definitions for $\boldsymbol{Z}'_{N_2}, \boldsymbol{Z}'_{U_1}, \boldsymbol{Z}'_{U_2}$.

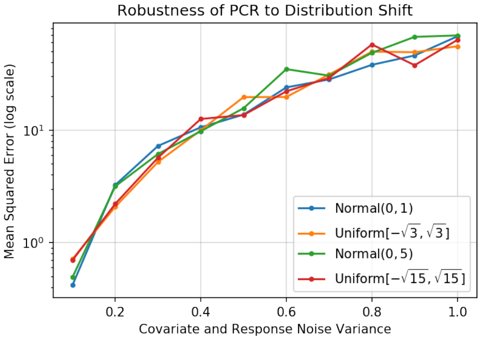


Figure 2.4: MSE plot of multiple PCR estimates – $\widehat{y}'_{N_1}$, $\widehat{y}'_{U_1}$, $\widehat{y}'_{N_2}$, and $\widehat{y}'_{U_2}$ – as we shift the distribution of the out-of-sample covariates, while ensuring Assumption 5 holds. Pleasingly, the MSE remains closely matched across all noise levels and distribution shifts.

Results

Using the observations $(y, \boldsymbol{Z}, \boldsymbol{Z}'_{N_1})$, we perform PCR to produce $\widehat{y}'_{N_1}$. We produce $\widehat{y}'_{N_2}, \widehat{y}'_{U_1}, \widehat{y}'_{U_2}$ analogously. We define MSE as in Section 2.7.2 with each estimate compared against its corresponding latent response, e.g., $\widehat{y}'_{N_1}$ against a'_{N_1}. Figure 2.4 shows the MSE of $\widehat{y}'_{N_1}$, $\widehat{y}'_{U_1}$, $\widehat{y}'_{N_2}$, and $\widehat{y}'_{U_2}$ as we vary σ^2. Pleasingly, despite the changes in the data generating process of the out-of-sample responses we evaluate on, the MSE for all four experiments closely matches across all noise levels. This motivates Assumption 5 as the key requirement for generalization, at least for PCR, rather than distributional invariance between the training and testing covariates.

■ 2.7.4 Out-of-sample Prediction: Subspace Inclusion vs. Distributional Invariance

The purpose of this simulation is to further illustrate that subspace inclusion (Assumption 5) is the key structure that enables PCR to successfully generalize, and not necessarily distributional invariance between the training and testing covariates.

Generative Model

As before, we let $n = m = p = 1000$. We continue to generate the training covariates as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ following the procedure in Section 2.7.2. We now generate two different testing

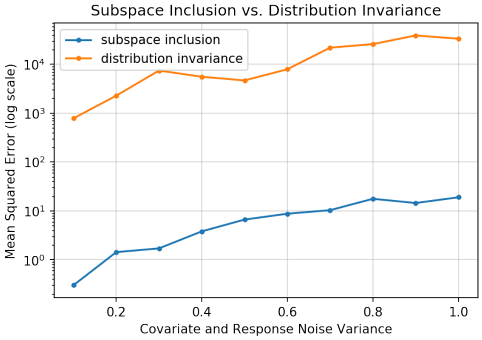


Figure 2.5: Plots of PCR's MSE under two situations: when Assumption 5 holds but distributional invariance is violated (blue), and when Assumption 5 is violated but distributional invariance holds (orange). Across varying levels of noise, the former condition achieves a much lower MSE.

covariates. First, we generate $\boldsymbol{X}'_1 = \boldsymbol{U}'\boldsymbol{V}^T$, where the entries of $\boldsymbol{U}'$ are independently sampled from a normal distribution with mean zero and variance 5. As such, it follows that Assumption 5 immediately holds between $\boldsymbol{X}'_1$ and $\boldsymbol{X}$, though they do not obey the same distribution. Next, we generate $\boldsymbol{X}'_2 = \boldsymbol{U}\boldsymbol{V}'^T$, where the entries of $\boldsymbol{V}'$ are independently sampled from a standard normal (just as in $\boldsymbol{V}$). In doing so, we ensure that $\boldsymbol{X}'_2$ and $\boldsymbol{X}$ follow the same distribution, though Assumption 5 no longer holds.

We generate β as in Section 2.7.2, and define $a'_1 = \boldsymbol{X}'_1\beta$ and $a'_2 = \boldsymbol{X}'_2\beta$. We also generate $\varepsilon, \boldsymbol{W}, \boldsymbol{W}'$ as in Section 2.7.2. In turn, we define the training data as $y = \boldsymbol{X}\beta + \varepsilon$ and $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, and testing data as $\boldsymbol{Z}'_1 = \boldsymbol{X}'_1 + \boldsymbol{W}'$ and $\boldsymbol{Z}'_2 = \boldsymbol{X}'_2 + \boldsymbol{W}'$.

Results

We apply PCR under two scenarios. First, we apply PCR using $(y, \boldsymbol{Z}, \boldsymbol{Z}'_1)$ to yield $\widehat{y}'_1$, and once again using $(y, \boldsymbol{Z}, \boldsymbol{Z}'_2)$ to yield $\widehat{y}'_2$. We define MSE as in Section 2.7.2 with each estimate compared against its corresponding latent response, e.g., $\widehat{y}'_1$ against a'_1. Figure 2.5 shows the MSE of $\widehat{y}'_1$ and $\widehat{y}'_2$ across varying levels of noise. As we can see, when Assumption 5 holds yet distributional invariance is violated, the corresponding MSE of $\widehat{y}'_1$ is almost three orders of magnitude smaller than that of $\widehat{y}'_2$, where Assumption 5 is violated but distributional invariance holds. This reinforces that the key structure required for PCR (and possibly other linear estimators) to generalize is Assumption 5, and not necessarily distributional invariance, as is typically assumed in the statistical learning literature.

■ 2.8 Conclusion

In this work, we analyze the standard method of PCR in a high-dimensional error-in-variables fixed design setting. As our main contributions, we establish that PCR identifies the unique model parameter with minimum ℓ_2-norm, is near minimax optimal, and achieves vanishing out-of-sample prediction under a natural linear algebraic relationship between the train and test covariates. Notably, both our parameter estimation and generalization error results are distribution-free. As an important consequence, our results also provide guarantees for counterfactual prediction for synthetic controls under fixed design settings. To the best of our knowledge, our out-of-sample (counterfactual) prediction guarantees in fixed design settings have been elusive in both the high-dimensional error-in-variables and synthetic controls literatures.

As an important future direction of research, it remains to establish bounds when the covariates are only approximately low-rank, i.e., there exists a matrix $\boldsymbol{A}$ such that $\text{rank}(\boldsymbol{A}) = r$ and $\boldsymbol{X} \approx \boldsymbol{A}$ in some norm. Our analysis of PCR in such a setting suggests an additional error term of the form $\boldsymbol{V}_{r,\perp} \boldsymbol{V}_{r,\perp}^T \beta^*{}_2$; here, the columns of $\boldsymbol{V}_{r,\perp} \in \mathbb{R}^{p \times (p-r)}$ form an orthonormal basis that is orthogonal to the top r right singular vectors of $\boldsymbol{X}$, and β^* is again the minimum norm model. To justify our postulation, recall that $\beta^* \in \text{rowspan}(\boldsymbol{X})$. Thus, it follows that $\boldsymbol{V}_{r,\perp} \boldsymbol{V}_{r,\perp}^T \beta^*{}_2$ is precisely the unavoidable parameter estimator error by taking a rank r approximation of $\boldsymbol{X}$. Hence, it stands to reason that *soft* singular value thresholding (SVT), which appropriately down-weights the singular values of $\widetilde{\boldsymbol{Z}}$, may be a more appropriate algorithmic approach as opposed to the *hard* SVT approach in PCR. Further, as stated earlier, extending our analysis to study the performance of PCR when the number of singular values retained is misspecified remains interesting future work.

Lastly, we believe another important future line of research is to bridge our out-of-sample prediction error analysis with recent exciting work on analyzing the generalization of over-parameterized estimators. Our key enabling assumption is that the rowspace of the test covariates lies within that of the training covariates, i.e., the test covariates are no more "complex" than the training covariates in a linear algebraic sense. In comparison, recent techniques to bound the generalization error of modern statistical estimators focus on the complexity of the learning algorithm itself, and assume the data generating process produces i.i.d. samples. Hence, a likely fruitful approach to produce tighter generalization error bounds for more complex non-linear settings would be to exploit both the complexity of the learning algorithm *and* the relative complexity of the test covariates compared to the training covariates—possibly by adapting our subspace inclusion condition with an appropriate non-linear notion.

■ 2.9 Proof of Theorem 2.5.1

We start with some useful notations. Let $y = X\beta^* + \varepsilon$ be the vector notation of (2.1) with $y = [y_i : i \le n] \in \mathbb{R}^n$, $\varepsilon = [\varepsilon_i : i \le n] \in \mathbb{R}^n$. Throughout, let $\boldsymbol{X} = \boldsymbol{USV}^T$. Recall that the SVD of $\widetilde{\boldsymbol{Z}} = 1/(\widehat{\rho})\boldsymbol{Z} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$. Its truncation using the top k singular components is denoted as $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{V}}_k^T$.

Further, we will often use the following bound: for any $\boldsymbol{A} \in \mathbb{R}^{a\times b}$, $v \in \mathbb{R}^b$,

$$\boldsymbol{A}v_2 = \sum_{j=1}^{b} \boldsymbol{A}_{\cdot j}v_j{}_2 \le (\max_{j\le b} \boldsymbol{A}_{\cdot j}{}_2)(\sum_{j=1}^{b}|v_j|) = \boldsymbol{A}_{2,\infty}v_1, \tag{2.9}$$

where $\boldsymbol{A}_{2,\infty} = \max_j \boldsymbol{A}_{\cdot j}{}_2$ with $\boldsymbol{A}_{\cdot j}$ representing the j-th column of $\boldsymbol{A}$.

As discussed in Section 2.5.1, we shall denote β^* as the unique minimum ℓ_2-norm model parameter satisfying (2.1); equivalently, this can be formulated as $\beta^* \in \text{rowspan}(\boldsymbol{X})$. As a result, it follows that

$$\boldsymbol{V}_\perp^T\beta^* = \boldsymbol{0}, \tag{2.10}$$

where $\boldsymbol{V}_\perp$ represents a matrix of orthornormal basis vectors that span the nullspace of $\boldsymbol{X}$.

Similarly, let $\widehat{\boldsymbol{V}}_{k,\perp} \in \mathbb{R}^{p\times(p-k)}$ be a matrix of orthonormal basis vectors that span the nullspace of $\widetilde{\boldsymbol{Z}}^k$; thus, $\widehat{\boldsymbol{V}}_{k,\perp}$ is orthogonal to $\widehat{\boldsymbol{V}}_k$. Then,

$$\begin{aligned}\widehat{\beta} - \beta^*{}_2^2 &= \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T(\widehat{\beta} - \beta^*) + \widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^T(\widehat{\beta} - \beta^*)_2^2 \\ &= \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T(\widehat{\beta} - \beta^*)_2^2 + \widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^T(\widehat{\beta} - \beta^*)_2^2 \\ &= \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T(\widehat{\beta} - \beta^*)_2^2 + \widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^T\beta^*{}_2^2.\end{aligned} \tag{2.11}$$

Note that in the last equality we have used Property 2.4.1, which states that $\widehat{\boldsymbol{V}}_{k,\perp}^T\widehat{\beta} = \boldsymbol{0}$. Next, we bound the two terms in (2.11).

Bounding $\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*)_2^2$. To begin, note that

$$\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*)_2^2 = \widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*)_2^2, \tag{2.12}$$

since $\widehat{\boldsymbol{V}}_k$ is an isometry. Next, consider

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*)_2^2 &\le 2\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*{}_2^2+2\boldsymbol{X}\beta^*-\widetilde{\boldsymbol{Z}}^k\beta^*{}_2^2\\ &\le 2\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*{}_2^2+2\boldsymbol{X}-\widetilde{\boldsymbol{Z}}^k{}_{2,\infty}^2\beta^*{}_1^2,\end{aligned}$$

where we used (3.47). Recall that $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{V}}_k^T$. Therefore,

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*)_2^2 &= (\widehat{\beta}-\beta^*)^T\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{\Sigma}}_k^2\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*)\\ &= (\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*))^T\widehat{\boldsymbol{\Sigma}}_k^2(\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*))\\ &\ge \widehat{s}_k^2\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*)_2^2.\end{aligned}$$

Therefore using (3.43), we conclude that

$$\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T(\widehat{\beta}-\beta^*)_2^2 \le \frac{2}{\widehat{s}_k^2}\left(\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*{}_2^2+\boldsymbol{X}-\widetilde{\boldsymbol{Z}}^k{}_{2,\infty}^2\beta^*{}_1^2\right). \tag{2.13}$$

Next, we bound $\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*{}_2$.

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-y_2^2 &= \widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*-\varepsilon_2^2\\ &= \widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*{}_2^2+\varepsilon_2^2-2\langle\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*,\varepsilon\rangle.\end{aligned} \tag{2.14}$$

By Property 2.4.1 we have,

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-y_2^2 \le \widetilde{\boldsymbol{Z}}^k\beta^*-y_2^2 &= (\widetilde{\boldsymbol{Z}}^k-\boldsymbol{X})\beta^*-\varepsilon_2^2\\ &= (\widetilde{\boldsymbol{Z}}^k-\boldsymbol{X})\beta^*{}_2^2+\varepsilon_2^2-2\langle(\widetilde{\boldsymbol{Z}}^k-\boldsymbol{X})\beta^*,\varepsilon\rangle.\end{aligned} \tag{2.15}$$

From (3.50) and (3.51), we have

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k\widehat{\beta}-\boldsymbol{X}\beta^*{}_2^2 &\le (\widetilde{\boldsymbol{Z}}^k-\boldsymbol{X})\beta^*{}_2^2+2\langle\widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*),\varepsilon\rangle\\ &\le \boldsymbol{X}-\widetilde{\boldsymbol{Z}}^k{}_{2,\infty}^2\beta^*{}_1^2+2\langle\widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*),\varepsilon\rangle,\end{aligned} \tag{2.16}$$

where we used (3.47). From (3.49) and (3.52), we conclude that

$$\widehat{V}_k \widehat{V}_k^T (\widehat{\beta} - \beta^*)_2^2 \le \frac{4}{\widehat{s}_k^2} \Big(X - \widetilde{Z}^k{}_{2,\infty}^2 \beta^*{}_1^2 + \langle \widetilde{Z}^k (\widehat{\beta} - \beta^*), \varepsilon \rangle \Big). \tag{2.17}$$

Bounding $\widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T \beta^*{}_2^2$*.* Consider

$$\begin{aligned} \widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T \beta^*{}_2 &= (\widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T - V_\perp V_\perp^T)\beta^* + V_\perp V_\perp^T \beta^*{}_2 \\ &\overset{(a)}{=} (\widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T - V_\perp V_\perp^T)\beta^*{}_2 \\ &\le \widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T - V_\perp V_\perp^T{}_2 \ \beta^*{}_2, \end{aligned} \tag{2.18}$$

where (a) follows from $V_\perp^T \beta^* = \mathbf{0}$ due to (2.10). Then,

$$\begin{aligned} \widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T - V_\perp V_\perp^T &= (\mathcal{D} - V_\perp V_\perp^T) - (\mathcal{D} - \widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T) \\ &= V V^T - \widehat{V}_k \widehat{V}_k^T. \end{aligned} \tag{2.19}$$

From (2.18) and (2.19), it follows that

$$\widehat{V}_{k,\perp} \widehat{V}_{k,\perp}^T \beta^*{}_2 \le V V^T - \widehat{V}_k \widehat{V}_k^T{}_2 \ \beta^*{}_2. \tag{2.20}$$

Bringing together (2.11), (3.53), and (2.20). Collectively, we obtain

$$\begin{aligned} \widehat{\beta} - \beta^*{}_2^2 &\le V V^T - \widehat{V}_k \widehat{V}_k^T{}_2^2 \ \beta^*{}_2^2 \\ &\quad + \frac{4}{\widehat{s}_k^2} \Big(X - \widetilde{Z}^k{}_{2,\infty}^2 \ \beta^*{}_1^2 + \langle \widetilde{Z}^k (\widehat{\beta} - \beta^*), \varepsilon \rangle \Big). \end{aligned} \tag{2.21}$$

Key lemmas. We state the key lemmas bounding each of the terms on the right hand side of (2.21). This will help us conclude the proof of Theorem 2.5.1. The proofs of these lemmas are presented in Sections 2.12.1, 2.12.2, 2.12.3, 2.12.4.

Lemma 2.9.1. *Consider the setup of Theorem 2.5.1, and PCR with parameter $k = r =$ rank(X). Then, for any $t > 0$, the following holds with probability at least $1 - \exp\big(-t^2\big)$:*

$$U U^T - \widehat{U}_r \widehat{U}_r^T{}_2 \le C(K, \gamma) \frac{\sqrt{n} + \sqrt{p} + t}{\rho s_r},$$

$$VV^T - \widehat{V}_r\widehat{V}_r^T{}_2 \le C(K,\gamma)\frac{\sqrt{n}+\sqrt{p}+t}{\rho s_r}.$$

Here, $s_r > 0$ represents the r-th singular value of $\boldsymbol{X}$.

Lemma 2.9.2. *Consider PCR with parameter $k = r$ and $\rho \ge c\frac{\log^2 np}{np}$. Then with probability at least $1 - O(1/(np)^{10})$,*

$$\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r{}_{2,\infty}^2 \le C(K,\gamma)\Big(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r+\sqrt{r}\,\log(np)}{\rho^2}\Big) + C\frac{\log(np)}{\rho\, p}.$$

Lemma 2.9.3. *If $\rho \ge c\frac{\log^2 np}{np}$, then for any $k \in [n \wedge p]$, we have with probability at least $1 - O(1/(np)^{10})$,*

$$|\widehat{s}_k - s_k| \le \frac{C(K,\gamma)(\sqrt{n}+\sqrt{p})}{\rho} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}}s_k.$$

Lemma 2.9.4. *Given $\widetilde{\boldsymbol{Z}}^r$, the following holds with probability at least $1 - O(1/(np)^{10})$ with respect to randomness in ε:*

$$\langle\widetilde{\boldsymbol{Z}}^r(\widehat{\beta}-\beta^*),\varepsilon\rangle \le \sigma^2 r + C\sigma\sqrt{\log(np)}\,(\sigma\sqrt{r} + \sigma\sqrt{\log(np)} + \beta^*{}_1(\sqrt{n} + \widetilde{\boldsymbol{Z}}^r - \boldsymbol{X}_{2,\infty})).$$

Completing the proof of Theorem 2.5.1. Using Lemma 2.9.4, the following holds with probability at least $1 - O(1/(np)^{10})$:

$$\begin{aligned}
&\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r{}_{2,\infty}^2\beta^*{}_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\beta}-\beta^*),\varepsilon\rangle \\
&\quad\le \boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r{}_{2,\infty}^2\beta^*{}_1^2 + C\sigma\sqrt{\log(np)}\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r{}_{2,\infty}\beta^*{}_1 + C\sigma^2\log(np) \\
&\qquad\quad + C\sigma\sqrt{\log(np)}(\sqrt{n}\beta^*{}_1 + s\sigma\sqrt{r}) + \sigma^2 r \\
&\quad\le C(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k{}_{2,\infty}\beta^*{}_1 + \sigma\sqrt{\log(np)})^2 + C\sigma\sqrt{\log(np)}(\sqrt{n}\beta^*{}_1 + \sigma\sqrt{r}) + \sigma^2 r \\
&\quad\le C\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k{}_{2,\infty}^2\beta^*{}_1^2 + C\sigma^2(\log(np)+r) + C\sigma\sqrt{n\log(np)}\beta^*{}_1. \qquad (2.22)
\end{aligned}$$

Using (2.21) and (2.22), we have with probability at least $1 - O(1/(np)^{10})$,

$$\begin{aligned}
\widehat{\beta} - \beta^*{}_2^2 &\le VV^T - \widehat{V}_k\widehat{V}_k^T{}_2^2\ \beta^*{}_2^2 \\
&\qquad + C\frac{\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k{}_{2,\infty}^2\beta^*{}_1^2 + \sigma^2(\log(np)+r) + \sigma\sqrt{n\log(np)}\beta^*{}_1}{\widehat{s}_r^2}. \qquad (2.23)
\end{aligned}$$

Using Lemma 2.9.1 in (2.23), we have with probability at least $1 - O(1/(np)^{10})$,

$$\widehat{\beta} - \beta^{*2}_{2} \leq C(K, \gamma)\frac{n+p}{\rho^2 s_r^2}\beta^{*2}_{2} + C\frac{\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^{k2}_{2,\infty}\beta^{*2}_{1}}{\widehat{s}_r^2} + C\frac{\sigma^2(\log(np) + r) + \sigma\sqrt{n\log(np)}\beta^{*}{}_{1}}{\widehat{s}_r^2}. \tag{2.24}$$

By Lemma 2.9.3 with $k = r$, and since $\rho \geq c\frac{\log^2 np}{np}$ and $\text{snr} \geq C(K, \gamma, \sigma)$, we have that

$$\begin{aligned}\frac{|\widehat{s}_r - s_r|}{s_r} &\leq \frac{C(K,\gamma)(\sqrt{n} + \sqrt{p})}{\rho s_r} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} \\ &= \frac{C(K,\gamma)}{\text{snr}} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} \leq \frac{1}{2}.\end{aligned}$$

As a result,

$$s_r/2 \leq \widehat{s}_r \leq 3s_r/2. \tag{2.25}$$

Now, using the definition of snr as per (2.4) and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we have

$$\frac{n+p}{\rho^2 s_r^2} \leq \frac{1}{\text{snr}^2}. \tag{2.26}$$

Using (2.25), (2.26), the assumption that $\beta^{*}{}_{1} = O(\sqrt{p})$, and observing that $\beta^{*2}_{1} \leq p\beta^{*2}_{2}$, we obtain

$$\frac{\sigma^2(\log(np) + r)}{\widehat{s}_r^2} \leq C\frac{\sigma^2\rho^2 r\log(np)}{\text{snr}^2(n \vee p)} \tag{2.27}$$

$$\frac{\sigma\sqrt{n\log(np)}}{\widehat{s}_r^2}\beta^{*}{}_{1} \leq C\frac{\sigma\rho^2\sqrt{np\log(np)}}{\text{snr}^2(n \vee p)} \leq C\frac{\sigma\sqrt{\log(np)}}{\text{snr}^2} \tag{2.28}$$

$$\frac{(n+p)(n + \sqrt{n}\,\log(np))}{\rho^4 s_r^2\widehat{s}_r^2}\beta^{*2}_{1} \leq C\frac{n\log(np)}{\text{snr}^4(n \vee p)}\beta^{*2}_{1} \leq C\frac{\log(np)}{\text{snr}^4}\beta^{*2}_{1} \tag{2.29}$$

$$\frac{r + \sqrt{r}\,\log(np)}{\rho^2\widehat{s}_r^2}\beta^{*2}_{1} \leq C\frac{rp\log(np)}{\text{snr}^2(n \vee p)}\beta^{*2}_{2} \leq C\frac{r\log(np)}{\text{snr}^2}\beta^{*2}_{2} \tag{2.30}$$

$$\frac{\log(np)}{\rho\widehat{s}_r^2 p}\beta^{*2}_{1} \leq C\frac{p\rho\log(np)}{\text{snr}^2 p(n \vee p)}\beta^{*2}_{2} \leq C\frac{\log(np)}{\text{snr}^2(n \vee p)}\beta^{*2}_{2}. \tag{2.31}$$

Plugging Lemma 2.9.3, (2.26), (2.27), (2.28), (2.29), (2.30), (2.31) into (2.24), using the assumption $\beta^*{}_2 = \Omega(1)$, and simplifying, we obtain

$$\widehat{\beta} - \beta^{*2}_{2} \leq C(K, \gamma, \sigma)\log(np) \cdot \left(\frac{r\beta^{*2}_{2}}{\text{snr}^2} + \frac{\beta^{*2}_{1}}{\text{snr}^4} \right).$$

This completes the proof of Theorem 2.5.1.

■ 2.9.1 Proof of Lemma 2.9.1

Recall that $\boldsymbol{U}, \boldsymbol{V}$ denote the left and right singular vectors of $\boldsymbol{X}$ (equivalently, $\rho\boldsymbol{X}$), respectively; meanwhile, $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k$ denote the top k left and right singular vectors of $\widetilde{\boldsymbol{Z}}$ (equivalently, $\boldsymbol{Z}$), respectively. Further, observe that $\mathbb{E}[\boldsymbol{Z}] = \rho\boldsymbol{X}$ and let $\tilde{\boldsymbol{W}} = \boldsymbol{Z} - \rho\boldsymbol{X}$. To arrive at our result, we recall Wedin's Theorem Wedin (1972).

Theorem 2.9.1 (Wedin's Theorem). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times p}$, let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ and $\boldsymbol{B} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$ be their respective SVDs. Let $\boldsymbol{U}_k, \boldsymbol{V}_k$ (respectively, $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k$) correspond to the truncation of $\boldsymbol{U}, \boldsymbol{V}$ (respectively, $\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}$) that retains the columns corresponding to the top k singular values of $\boldsymbol{A}$ (respectively, $\boldsymbol{B}$). Let s_k denote the k-th singular value of $\boldsymbol{A}$. Then,*

$$\max\left(\boldsymbol{U}_k\boldsymbol{U}_k^T - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T{}_2, \boldsymbol{V}_k\boldsymbol{V}_k^T - \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^T{}_2 \right) \leq \frac{2\left\|\boldsymbol{A} - \boldsymbol{B}\right\|_2}{s_k - s_{k+1}}.$$

Using Theorem 3.15.1 for $k = r$, it follows that

$$\max\left(\boldsymbol{U}\boldsymbol{U}^T - \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{U}}_r^T{}_2, \boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r^T{}_2 \right) \leq \frac{2\tilde{\boldsymbol{W}}_2}{\rho s_r}, \tag{2.32}$$

where s_r is the smallest nonzero singular value of $\boldsymbol{X}$. Next, we obtain a high probability bound on $\tilde{\boldsymbol{W}}_2$. To that end,

$$\frac{1}{n}\tilde{\boldsymbol{W}}_2^2 = \frac{1}{n}\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}_2 \leq \frac{1}{n}\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}} - \mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 + \frac{1}{n}\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2. \tag{2.33}$$

We bound the two terms in (2.71) separately. We recall the following lemma, which is a direct extension of Theorem 4.6.1 of Vershynin (2018) for the non-isotropic setting, and we present its proof for completeness in Section 2.12.5.

Lemma 2.9.5 (Independent sub-gaussian rows). *Let $\boldsymbol{A}$ be an $n \times p$ matrix whose rows A_i are independent, mean zero, sub-gaussian random vectors in $\mathbb{R}^p$ with second moment matrix*

$\mathbf{\Sigma} = (1/n)\mathbb{E}[\boldsymbol{A}^T\boldsymbol{A}]$. *Then for any* $t \geq 0$, *the following inequality holds with probability at least* $1 - \exp\left(-t^2\right)$*:*

$$\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A} - \mathbf{\Sigma}_2 \leq K^2 \max(\delta, \delta^2), \quad \textit{where } \delta = C\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}; \tag{2.34}$$

here, $K = \max_i \left\|A_i\right\|_{\psi_2}$.

The matrix $\tilde{\boldsymbol{W}} = \boldsymbol{Z} - \rho\boldsymbol{X}$ has independent rows by Assumption 4. We state the following Lemma about the distribution property of the rows of $\tilde{\boldsymbol{W}}$, the proof of which can be found in Section 2.12.6.

Lemma 2.9.6. *Let Assumption 4 hold. Then,* $z_i - \rho x_i$ *is a sequence of independent, mean zero, sub-gaussian random vectors satisfying* $z_i - \rho x_{i\psi_2} \leq C(K+1)$.

From Lemmas 2.12.5 and 2.12.6, with probability at least $1 - \exp\left(-t^2\right)$,

$$\frac{1}{n}\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}} - \mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 \leq C(K+1)^2(1 + \frac{p}{n} + \frac{t^2}{n}). \tag{2.35}$$

Finally, we claim the following bound on $\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2$, the proof of which is in Section 2.12.7.

Lemma 2.9.7. *Let Assumption 4 hold. Then, we have*

$$\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 \leq C(K+1)^2 n(\rho - \rho^2) + n\rho^2\gamma^2.$$

From (2.71), (2.73) and Lemma 2.12.7, it follows that with probability at least $1 - \exp\left(-t^2\right)$ for any $t > 0$, we have

$$\tilde{\boldsymbol{W}}_2^2 \leq C(K+1)^2(n + p + t^2) + n(\rho(1-\rho)(K+1)^2 + \rho^2\gamma^2).$$

For this, we conclude the following lemma.

Lemma 2.9.8. *For any* $t > 0$, *the following holds with probability at least* $1 - \exp\left(-t^2\right)$*:*

$$\boldsymbol{Z} - \rho\boldsymbol{X}_2 \leq C(K, \gamma)(\sqrt{n} + \sqrt{p} + t).$$

Using the above and (2.70), we conclude the proof of Lemma 2.9.1.

■ 2.9.2 Proof of Lemma 2.9.2

We want to bound $\mathbf{X} - \widetilde{\mathbf{Z}}^{k}{}^2_{2,\infty}$. To that end, let $\Delta_j = \mathbf{X}_{\cdot j} - \widetilde{\mathbf{Z}}^k_{\cdot j}$ for any $j \in [p]$. Our interest is in bounding $\Delta_{j}{}^2_2$ for all $j \in [p]$. Consider,

$$\widetilde{\mathbf{Z}}^k_{\cdot j} - \mathbf{X}_{\cdot j} = (\widetilde{\mathbf{Z}}^k_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}) + (\widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j} - \mathbf{X}_{\cdot j}).$$

Now, note that $\widetilde{\mathbf{Z}}^k_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}$ belongs to the subspace spanned by column vectors of $\widehat{\mathbf{U}}_k$, while $\widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j} - \mathbf{X}_{\cdot j}$ belongs to its orthogonal complement with respect to $\mathbb{R}^n$. As a result,

$$\widetilde{\mathbf{Z}}^k_{\cdot j} - \mathbf{X}_{\cdot j}{}^2_2 = \widetilde{\mathbf{Z}}^k_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}{}^2_2 + \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j} - \mathbf{X}_{\cdot j}{}^2_2. \tag{2.36}$$

Bounding $\widetilde{\mathbf{Z}}^k_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}{}^2_2$. Recall that $\widetilde{\mathbf{Z}} = (1/\hat{\rho})\mathbf{Z} = \widehat{\mathbf{U}} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{V}}^T$, and hence $\mathbf{Z} = \hat{\rho} \widehat{\mathbf{U}} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{V}}^T$. Consequently,

$$\begin{aligned}\frac{1}{\hat{\rho}} \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{Z}_{\cdot j} &= \frac{1}{\hat{\rho}} \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{Z} \mathbf{e}_j = \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \widehat{\mathbf{U}} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{V}}^T \mathbf{e}_j \\ &= \widehat{\mathbf{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{V}}_k^T \mathbf{e}_j = \widetilde{\mathbf{Z}}^k_{\cdot j}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\widetilde{\mathbf{Z}}^k_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j} &= \frac{1}{\hat{\rho}} \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{Z}_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j} \\ &= \frac{1}{\hat{\rho}} \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T (\mathbf{Z}_{\cdot j} - \rho \mathbf{X}_{\cdot j}) + \Big(\frac{\rho - \hat{\rho}}{\hat{\rho}}\Big) \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}.\end{aligned}$$

Therefore,

$$\begin{aligned}\widetilde{\mathbf{Z}}^k_{\cdot j} - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}{}^2_2 &\le \frac{2}{\hat{\rho}^2} \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T (\mathbf{Z}_{\cdot j} - \rho \mathbf{X}_{\cdot j})^2_2 + 2\Big(\frac{\rho - \hat{\rho}}{\hat{\rho}}\Big)^2 \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T \mathbf{X}_{\cdot j}{}^2_2 \\ &\le \frac{2}{\hat{\rho}^2} \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T (\mathbf{Z}_{\cdot j} - \rho \mathbf{X}_{\cdot j})^2_2 + 2\Big(\frac{\rho - \hat{\rho}}{\hat{\rho}}\Big)^2 \mathbf{X}_{\cdot j}{}^2_2,\end{aligned}$$

where we have used the fact that $\widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T{}_2 = 1$. Recall that $\mathbf{U} \in \mathbb{R}^{n \times r}$ represents the left singular vectors of $\mathbf{X}$. Thus,

$$\widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T (\mathbf{Z}_{\cdot j} - \rho \mathbf{X}_{\cdot j})^2_2 \le 2(\widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T - \mathbf{U}\mathbf{U}^T)(\mathbf{Z}_{\cdot j} - \rho \mathbf{X}_{\cdot j})^2_2 + 2\mathbf{U}\mathbf{U}^T (\mathbf{Z}_{\cdot j} - \rho \mathbf{X}_{\cdot j})^2_2$$

$$\leq 2\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T - \boldsymbol{U}\boldsymbol{U}^T{}_2^2\; \boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}{}_2^2 + 2\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}){}_2^2.$$

By Assumption 1, we have that $\boldsymbol{X}_{\cdot j}{}_2^2 \leq n$. This yields

$$\begin{aligned}\widetilde{\boldsymbol{Z}}_{\cdot j}^k - \widehat{U}_k\widehat{\boldsymbol{U}}_k^T \boldsymbol{X}_{\cdot j}{}_2^2 &\leq \frac{4}{\hat{\rho}^2}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T - \boldsymbol{U}\boldsymbol{U}^T{}_2^2\; \boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}{}_2^2 \\ &\quad + \frac{4}{\hat{\rho}^2}\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}){}_2^2 + 2n\left(\frac{\rho - \hat{\rho}}{\hat{\rho}}\right)^2. \qquad (2.37)\end{aligned}$$

We now state Lemmas 2.12.9 and 2.12.10. Their proofs are in Sections 2.12.8 and 2.12.9, respectively.

Lemma 2.9.9. *For any $\alpha > 1$,*

$$\mathbb{P}\left(\rho/\alpha \leq \widehat{\rho} \leq \alpha\rho\right) \geq 1 - 2\exp\left(-\frac{(\alpha-1)^2 np\rho}{2\alpha^2}\right).$$

Therefore, for $\rho \geq c\frac{\log^2 np}{np}$, we have with probability $1 - O(1/(np)^{10})$

$$\frac{\rho}{2} \leq \hat{\rho} \leq 2\rho \quad \textit{and} \quad \left(\frac{\rho - \hat{\rho}}{\hat{\rho}}\right)^2 \leq C\frac{\log(np)}{\rho np}.$$

Lemma 2.9.10. *Consider any matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times\ell}$ with $1 \leq \ell \leq n$ such that its columns $\boldsymbol{Q}_{\cdot j}$ for $j \in [\ell]$ are orthonormal vectors. Then for any $t > 0$,*

$$\begin{aligned}&\mathbb{P}\left(\max_{j\in[p]} \left\|\boldsymbol{Q}\boldsymbol{Q}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})\right\|_2^2 \geq \ell C(K+1)^2 + t\right) \\ &\qquad \leq p \cdot \exp\left(-c\min\left(\frac{t^2}{C(K+1)^4\ell}, \frac{t}{C(K+1)^2}\right)\right).\end{aligned}$$

Subsequently, with probability $1 - O(1/(np)^{10})$,

$$\max_{j\in[p]} \left\|\boldsymbol{Q}\boldsymbol{Q}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})\right\|_2^2 \leq C(K+1)^2(\ell + \sqrt{\ell}\log(np)).$$

Both terms $\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}{}_2^2$ and $\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}){}_2^2$ can be bounded by Lemma 2.12.10: for the first term $\boldsymbol{Q} = \mathcal{D}$, and for the second term $\boldsymbol{Q} = \boldsymbol{U}$. In summary, with probability $1 - O(1/(np)^{10})$, we have

$$\max_{j\in[p]} \left\|\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}\right\|_2^2 \leq C(K+1)^2(n + \sqrt{n}\log(np)), \qquad (2.38)$$

and

$$\max_{j\in[p]} \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})_2^2 \leq C(K+1)^2(r + \sqrt{r}\log(np)). \tag{2.39}$$

Using (2.75), (2.76), (2.77), and Lemmas 2.9.1 and 2.12.9 with $k = r$, we conclude that with probability $1 - O(1/(np)^{10})$,

$$\max_{j\in[p]} \widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{U}_k \widehat{U}_k^T \boldsymbol{X}_{\cdot j2}^{\;2} \leq C(K,\gamma)\Big(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r+\sqrt{r}\,\log(np)}{\rho^2}\Big) + C\frac{\log(np)}{\rho\, p}. \tag{2.40}$$

Bounding $\widehat{U}_k \widehat{U}_k^T \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j2}^{\;2}$. Recalling $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, we obtain $\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{X}_{\cdot j} = \boldsymbol{X}_{\cdot j}$ since $\boldsymbol{U}\boldsymbol{U}^T$ is the projection onto the column space of $\boldsymbol{X}$. Therefore,

$$\begin{aligned}\widehat{U}_k \widehat{U}_k^T \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j2}^{\;2} &= \widehat{U}_k \widehat{U}_k^T \boldsymbol{X}_{\cdot j} - \boldsymbol{U}\boldsymbol{U}^T \boldsymbol{X}_{\cdot j2}^{\;2} \\ &\leq \widehat{U}_k \widehat{U}_k^T - \boldsymbol{U}\boldsymbol{U}^{T\,2}_{\;2}\ \boldsymbol{X}_{\cdot j2}^{\;2}.\end{aligned}$$

Using Property 1, note that $\boldsymbol{X}_{\cdot j2}^{\;2} \leq n$. Thus using Lemma 2.9.1 with $k = r$, we have that with probability at least $1 - O(1/(np)^{10})$, we have

$$\widehat{U}_k \widehat{U}_k^T \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j2}^{\;2} \leq C\frac{n(n+p)}{\rho^2 s_r^2}. \tag{2.41}$$

Concluding. From (2.74), (2.78), and (2.79), we claim with probability at least $1 - O(1/(np)^{10})$

$$\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^{k\,2}_{\;2,\infty} \leq C(K,\gamma)\Big(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r+\sqrt{r}\,\log(np)}{\rho^2}\Big) + C\frac{\log(np)}{\rho\, p}.$$

This completes the proof of Lemma 2.9.2.

■ 2.9.3 Proof of Lemma 2.9.3

To bound $\widehat{s}_k$, we recall Weyl's inequality.

Lemma 2.9.11 (Weyl's inequality). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, let σ_i and $\widehat{\sigma}_i$ be the i-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then* *for all $i \in [m \wedge n]$,*

$$|\sigma_i - \widehat{\sigma}_i| \le \left\| \boldsymbol{A} - \boldsymbol{B} \right\|_2 .$$

Let $\tilde{s}_k$ be the k-th singular value of $\boldsymbol{Z}$. Then, $\widehat{s}_k = (1/\hat{\rho})\tilde{s}_k$ since it is the k-th singular value of $\widetilde{\boldsymbol{Z}} = (1/\hat{\rho})\boldsymbol{Z}$. By Lemma 5.14.4, we have

$$|\tilde{s}_k - \rho s_k| \le \boldsymbol{Z} - \rho \boldsymbol{X}_2;$$

recall that s_k is the k-th singular value of $\boldsymbol{X}$. As a result,

$$\begin{aligned} |\widehat{s}_k - s_k| &= \frac{1}{\hat{\rho}} |\tilde{s}_k - \hat{\rho} s_k| \\ &\le \frac{1}{\hat{\rho}} |\tilde{s}_k - \rho s_k| + \frac{|\rho - \hat{\rho}|}{\hat{\rho}} s_k \\ &\le \frac{\boldsymbol{Z} - \rho \boldsymbol{X}_2}{\hat{\rho}} + \frac{|\rho - \hat{\rho}|}{\hat{\rho}} s_k . \end{aligned}$$

From Lemma 2.12.8 and Lemma 2.12.9, it follows that with probability at least $1 - O(1/(np)^{10})$,

$$|\widehat{s}_k - s_k| \le \frac{C(K, \gamma)(\sqrt{n} + \sqrt{p})}{\rho} + C \frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} s_k .$$

This completes the proof of Lemma 2.9.3.

■ 2.9.4 Proof of Lemma 2.9.4

We need to bound $\langle \widetilde{\boldsymbol{Z}}^k (\widehat{\beta} - \beta^*), \varepsilon \rangle$. To that end, we recall that $\widehat{\beta} = \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y$, $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T$, and $y = \boldsymbol{X}\beta^* + \varepsilon$. Thus,

$$\widetilde{\boldsymbol{Z}}^k \widehat{\beta} = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X} \beta^* + \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon .$$

Therefore,

$$\langle \widetilde{\boldsymbol{Z}}^k (\widehat{\beta} - \beta^*), \varepsilon \rangle = \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X} \beta^*, \varepsilon \rangle + \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon, \varepsilon \rangle - \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T \beta^*, \varepsilon \rangle . \qquad (2.42)$$

Now, ε is independent of $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{\Sigma}}_k, \widehat{\boldsymbol{V}}_k$ since $\widetilde{\boldsymbol{Z}}^k$ is determined by $\boldsymbol{Z}$, which is independent of ε. As a result,

$$\begin{aligned}
\mathbb{E}[\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon, \varepsilon\rangle] &= \mathbb{E}[\varepsilon^T \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon] \\
&= \mathbb{E}[\operatorname{tr}\Big(\varepsilon^T \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon\Big)] = \mathbb{E}[\operatorname{tr}\Big(\varepsilon\varepsilon^T \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T\Big)] \\
&= \operatorname{tr}\Big(\mathbb{E}[\varepsilon\varepsilon^T]\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T\Big) \le C\operatorname{tr}\Big(\sigma^2 \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T\Big) \\
&= C\sigma^2 \widehat{\boldsymbol{U}}_k{}_F^2 = C\sigma^2 k. \qquad (2.43)
\end{aligned}$$

Therefore, it follows that

$$\mathbb{E}[\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\beta} - \beta^*), \varepsilon\rangle] \le C\sigma^2 k, \qquad (2.44)$$

where we used the fact $\mathbb{E}[\varepsilon] = \boldsymbol{0}$. To obtain a high probability bound, using Lemma 5.14.2 it follows that for any $t > 0$

$$\mathbb{P}\Big(\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X}\beta^*, \varepsilon\rangle \ge t\Big) \le \exp\Big(-\frac{ct^2}{n\beta^*{}_1^2\sigma^2}\Big) \qquad (2.45)$$

due to Assumption 3, and

$$\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X}\beta^*{}_2 \le \boldsymbol{X}\beta^*{}_2 \le \boldsymbol{X}_{2,\infty}\beta^*{}_1 \le \sqrt{n}\beta^*{}_1;$$

note that we have used the fact that $\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T$ is a projection matrix and $\boldsymbol{X}_{2,\infty} \le \sqrt{n}$ due to Assumption 1. Similarly, for any $t > 0$

$$\mathbb{P}\Big(\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T \beta^*, \varepsilon\rangle \ge t\Big) \le \exp\Big(-\frac{ct^2}{\sigma^2(n + \widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}_{2,\infty}^2)\beta^*{}_1^2}\Big), \qquad (2.46)$$

due to Assumption 3, and

$$\begin{aligned}
\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T \beta^*{}_2 = (\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\beta^* + \boldsymbol{X}\beta^*{}_2 \;&\le (\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\beta^*{}_2 + \boldsymbol{X}\beta^*{}_2 \\
&\le (\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}_{2,\infty} + \boldsymbol{X}_{2,\infty})\beta^*{}_1.
\end{aligned}$$

Finally, using Lemma 5.14.3 and (3.61), it follows that for any $t > 0$

$$\mathbb{P}\Big(\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon, \varepsilon\rangle \ge \sigma^2 k + t\Big) \le \exp\Big(-c\min\Big(\frac{t^2}{k\sigma^4}, \frac{t}{\sigma^2}\Big)\Big), \qquad (2.47)$$

since $\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T$ is a projection matrix and by Assumption 3.

From (3.57), (3.62), (3.64), and (3.66), we conclude that with probability at least $1 - O(1/(np)^{10})$,

$$\langle \widetilde{\boldsymbol{Z}}^k (\widehat{\beta} - \beta^*), \varepsilon \rangle \leq \sigma^2 k + C\sigma\sqrt{\log(np)}(\sigma\sqrt{k} + \sigma\sqrt{\log(np)} + \beta^*{}_1(\sqrt{n} + \widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}_{2,\infty})).$$

This completes the proof of Lemma 2.9.4.

■ 2.9.5 Proof of Lemma 2.12.5

As mentioned earlier, the proof presented here is a natural extension of that for Theorem 4.6.1 in Vershynin (2018) for the non-isotropic setting. Recall that

$$\|\boldsymbol{A}\| = \max_{x \in S^{p-1}, y \in S^{n-1}} \langle \boldsymbol{A}x, y \rangle,$$

where S^{p-1}, S^{n-1} denote the unit spheres in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively. We start by bounding the quadratic term $\langle \boldsymbol{A}x, y \rangle$ for a finite set x, y obtained by placing 1/4-net on the unit spheres, and then use the bound on them to bound $\langle \boldsymbol{A}x, y \rangle$ for all x, y over the spheres.

Step 1: Approximation. We will use Corollary 4.2.13 of Vershynin (2018) to establish a 1/4-net of $\mathcal{N}$ of the unit sphere S^{p-1} with cardinality $|\mathcal{N}| \leq 9^p$. Applying Lemma 4.4.1 of Vershynin (2018), we obtain

$$\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{\Sigma}_2 \leq 2 \max_{x \in \mathcal{N}} \left| \langle (\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{\Sigma})x, x \rangle \right| = 2 \max_{x \in \mathcal{N}} \left| \frac{1}{n}\boldsymbol{A}x_2^2 - x^T\boldsymbol{\Sigma}x \right|.$$

To achieve our desired result, it remains to show that

$$\max_{x \in \mathcal{N}} \left| \frac{1}{n}\boldsymbol{A}x_2^2 - x^T\boldsymbol{\Sigma}x \right| \leq \frac{\epsilon}{2},$$

where $\epsilon = K^2 \max(\delta, \delta^2)$.

Step 2: Concentration. Let us fix a unit vector $x \in S^{p-1}$ and write

$$\|\boldsymbol{A}x\|_2^2 - x^T\boldsymbol{\Sigma}x = \sum_{i=1}^n \left(\langle A_i, x \rangle^2 - \mathbb{E}[\langle A_i, x \rangle^2] \right) =: \sum_{i=1}^n \left(Y_i^2 - \mathbb{E}[Y_i^2] \right).$$

Since the rows of A are assumed to be independent sub-gaussian random vectors with $A_{i\psi_2} \leq K$, it follows that $Y_i = \langle A_i, x \rangle$ are independent sub-gaussian random variables with $Y_{i\psi_2} \leq K$. Therefore, $Y_i^2 - \mathbb{E}[Y_i^2]$ are independent, mean zero, sub-exponential random variables with

$$Y_i^2 - \mathbb{E}[Y_i^2]_{\psi_1} \leq C Y_i^2{}_{\psi_1} \leq C Y_{i\psi_2}^2 \leq CK^2.$$

As a result, we can apply Bernstein's inequality (see Theorem 5.14.1) to obtain

$$\begin{aligned}
\mathbb{P}\left(\left| \frac{1}{n} \|Ax\|_2^2 - x^T \mathbf{\Sigma} x \right| \geq \frac{\epsilon}{2} \right) &= \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (Y_i^2 - \mathbb{E}[Y_i^2]) \right| \geq \frac{\epsilon}{2} \right) \\
&\leq 2 \exp\left(-c \min\left(\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2} \right) n \right) \\
&= 2 \exp\left(-c \delta^2 n \right) \\
&\leq 2 \exp\left(-c C^2 (p + t^2) \right),
\end{aligned}$$

where the last inequality follows from the definition of δ in (2.72) and because $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$.

Step 3: Union bound. We now apply a union bound over all elements in the net. Specifically,

$$\mathbb{P}\left(\max_{x \in \mathcal{N}} \left| \frac{1}{n} \|Ax\|_2^2 - x^T \mathbf{\Sigma} x \right| \geq \frac{\epsilon}{2} \right) \leq 9^p \cdot 2 \exp\left(-c C^2 (p + t^2) \right) \leq 2 \exp\left(-t^2 \right),$$

for large enough C. This concludes the proof.

■ 2.9.6 Proof of Lemma 2.12.6

Recall that $z_i = (x_i + w_i) \circ \pi_i$, where w_i is an independent mean zero subgaussian vector with $w_{i\psi_2} \leq K$ and π_i is a vector of independent Bernoulli variables with parameter ρ. Hence, $\mathbb{E}[z_i - \rho x_i] = \mathbf{0}$ and is independent across $i \in [n]$. The only remaining item is a bound on $z_i - \rho x_{i\psi_2}$. To that end, note that

$$\begin{aligned}
z_i - \rho x_{i\psi_2} &= x_i \circ \pi_i + w_i \circ \pi_i - \rho x_{i\psi_2} \\
&\leq x_i \circ (\rho \mathbf{1} - \pi_i)_{\psi_2} + w_i \circ \pi_{i\psi_2}.
\end{aligned}$$

Now, $(\rho\mathbf{1} - \pi_i)$ is independent, zero mean random vector whose absolute value is bounded by 1, and is component-wise multiplied by x_i which are bounded in absolute value by 1 as per Assumption 1. That is, $x_i \circ (\rho\mathbf{1} - \pi_i)$ is a zero mean random vector where each component is independent and bounded in absolute value by 1. That is, $\cdot_{\psi_2} \leq C$.

For $w_i \circ \pi_i$, note that w_i and π_i are independent vectors and the coordinates of π_i have support $\{0, 1\}$. Therefore, from Lemma 2.12.12, it follows that $w_i \circ \pi_{i\psi_2} \leq w_{i\psi_2} \leq K$ by Assumption 4. The proof of Lemma 2.12.6 is complete by choosing a large enough C.

Lemma 2.9.12. *Suppose that $Y \in \mathbb{R}^n$ and $P \in \{0,1\}^n$ are independent random vectors. Then,*

$$Y \circ P_{\psi_2} \leq Y_{\psi_2}.$$

Proof. Given a binary vector $P \in \{0,1\}^n$, let $I_P = \{i \in [n] : P_i = 1\}$. Observe that

$$Y \circ P = \sum_{i\in I_P} e_i \otimes e_i Y.$$

Here, $\circ$ denotes the Hadamard product (entry-wise product) of two matrices. By definition of the ψ_2-norm,

$$Y_{\psi_2} = \sup_{u\in\mathbb{S}^{n-1}} u^T Y_{\psi_2} = \sup_{u\in\mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_Y[\exp\left(|u^T Y|^2/t^2\right)] \leq 2\}.$$

Let $u_0 \in \mathbb{S}^{n-1}$ denote the maximum-achieving unit vector (such a u_0 exists because $\inf\{\cdots\}$ is continuous with respect to u and $\mathbb{S}^{n-1}$ is compact). Now,

$$\begin{aligned}
Y \circ P_{\psi_2} &= \sup_{u\in\mathbb{S}^{n-1}} u^T Y \circ P_{\psi_2} \\
&= \sup_{u\in\mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{Y,P}[\exp\left(|u^T Y \circ P|^2/t^2\right)] \leq 2\} \\
&= \sup_{u\in\mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_P[\mathbb{E}_Y[\exp\left(|u^T Y \circ P|^2/t^2\right) \mid P]] \leq 2\} \\
&= \sup_{u\in\mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_P[\mathbb{E}_Y[\exp(|u^T \sum_{i\in I_P} e_i \otimes e_i Y|^2/t^2) \mid P]] \leq 2\} \\
&= \sup_{u\in\mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_P[\mathbb{E}_Y[\exp(|(\sum_{i\in I_P} e_i \otimes e_i u)^T Y|^2/t^2) \mid P]] \leq 2\}.
\end{aligned}$$

For any $u \in \mathbb{S}^{n-1}$, observe that

$$\mathbb{E}_Y[\exp(|(\sum_{i\in I_P} e_i \otimes e_i u)^T Y|^2/t^2) \mid P] \leq \mathbb{E}_Y[\exp\left(|u_0^T Y|^2/t^2\right)].$$

Therefore, taking supremum over $u \in \mathbb{S}^{n-1}$, we obtain

$$Y \circ P_{\psi_2} \leq Y_{\psi_2}.$$

■

■ 2.9.7 Proof of Lemma 2.12.7

Consider

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}] &= \sum_{i=1}^n \mathbb{E}[(z_i - \rho x_i) \otimes (z_i - \rho x_i)] \\ &= \sum_{i=1}^n \mathbb{E}[z_i \otimes z_i] - \rho^2(x_i \otimes x_i) \\ &= \sum_{i=1}^n (\rho - \rho^2)\mathrm{diag}(x_i \otimes x_i) + (\rho - \rho^2)\mathrm{diag}(\mathbb{E}[w_i \otimes w_i]) + \rho^2\mathbb{E}[w_i \otimes w_i].\end{aligned}$$

Note that $\mathrm{diag}(\boldsymbol{X}^T\boldsymbol{X})_2 \leq n$ due to Assumption 1. Using Assumption 4, it follows that $\mathrm{diag}(\mathbb{E}[w_i \otimes w_i])_2 \leq CK^2$. By Assumption 4, we have $\mathbb{E}[w_i \otimes w_i]_2 \leq \gamma^2$. Therefore,

$$\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 \leq Cn(\rho - \rho^2)(K+1)^2 + n\rho^2\gamma^2.$$

This completes the proof of Lemma 2.12.7.

■ 2.9.8 Proof of Lemma 2.12.9

By the Binomial Chernoff bound, for $\alpha > 1$,

$$\mathbb{P}\left(\widehat{\rho} > \alpha\rho\right) \leq \exp\left(-\frac{(\alpha-1)^2}{\alpha+1} np\rho\right) \quad \text{and} \quad \mathbb{P}\left(\widehat{\rho} < \rho/\alpha\right) \leq \exp\left(-\frac{(\alpha-1)^2}{2\alpha^2} np\rho\right).$$

By the union bound,

$$\mathbb{P}\left(\rho/\alpha \leq \widehat{\rho} \leq \alpha\rho\right) \geq 1 - \mathbb{P}\left(\widehat{\rho} > \alpha\rho\right) - \mathbb{P}\left(\widehat{\rho} < \rho/\alpha\right).$$

Noticing $\alpha + 1 < 2\alpha < 2\alpha^2$ for all $\alpha > 1$, we obtain the desired bound claimed in Lemma 2.12.9. To complete the remaining claim of Lemma 2.12.9, we consider an α that satisfies

$$(\alpha - 1)^2 \leq C\frac{\log(np)}{\rho np},$$

for a constant $C > 0$. Thus,

$$1 - C\frac{\sqrt{\log(np)}}{\sqrt{\rho np}} \leq \alpha \leq 1 + C\frac{\sqrt{\log(np)}}{\sqrt{\rho np}}.$$

Then, with $\rho \geq c\frac{\log^2 np}{np}$, we have that $\alpha \leq 2$. Further by choosing $C > 0$ large enough, we have

$$\frac{(\rho - \hat{\rho})^2}{\hat{\rho}^2} \leq C\frac{\log(np)}{\rho np}.$$

holds with probability at least $1 - O(1/(np)^{10})$. This completes the proof of Lemma 2.12.9.

2.9.9 Proof of Lemma 2.12.10

By definition $\boldsymbol{QQ}^T \in \mathbb{R}^{n \times n}$ is a rank ℓ matrix. Since $\boldsymbol{Q}$ has orthonormal column vectors, the projection operator has $\boldsymbol{QQ}^T{}_2 = 1$ and $\boldsymbol{QQ}^T{}_F^2 = \ell$. For a given $j \in [p]$, the random vector $\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}$ is such that it has zero mean, independent components that are sub-gaussian by Assumption 4. For any $i \in [n], j \in [p]$, we have by property of ψ_2 norm, $z_{ij} - \rho x_{ij}{}_{\psi_2} \leq z_i - \rho x_i{}_{\psi_2}$ which is bounded by $C(K+1)$ using Lemma 2.12.6. Recall the Hanson-Wright inequality (Vershynin (2018)):

Theorem 2.9.2 (Hanson-Wright inequality). *Let $\zeta \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then for any $t > 0$,*

$$\mathbb{P}\left(\left|\zeta^T \boldsymbol{A}\zeta - \mathbb{E}[\zeta^T \boldsymbol{A}\zeta]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{L^4\left\|\boldsymbol{A}\right\|_F^2}, \frac{t}{L^2\left\|\boldsymbol{A}\right\|_2}\right)\right),$$

where $L = \max_{i \in [n]} \left\|\zeta_i\right\|_{\psi_2}$.

Now with $\zeta = \boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}$ and the fact that $\boldsymbol{Q}^T \boldsymbol{Q} = \mathcal{D} \in \mathbb{R}^{\ell \times \ell}$, $\boldsymbol{Q}\boldsymbol{Q}^T \zeta_2^2 = \zeta^T \boldsymbol{Q}\boldsymbol{Q}^T \zeta$. Therefore, by Theorem 2.12.3, for any $t > 0$,

$$\boldsymbol{Q}\boldsymbol{Q}^T \zeta_2^2 \leq \mathbb{E}[\zeta^T \boldsymbol{Q}\boldsymbol{Q}^T \zeta] + t,$$

with probability at least $1 - \exp\Big(- c \min\big(\frac{t}{C(K+1)^2}, \frac{t^2}{C(K+1)^4 \ell}\big)\Big)$. Now,

$$\begin{aligned}
\mathbb{E}[\zeta^T \boldsymbol{Q}\boldsymbol{Q}^T \zeta] &= \sum_{m=1}^{\ell} \mathbb{E}[(\boldsymbol{Q}_{\cdot m}^T \zeta)^2] \\
&\overset{(a)}{=} \sum_{m=1}^{\ell} \mathrm{Var}(\boldsymbol{Q}_{\cdot m}^T \zeta) \\
&\overset{(b)}{=} \sum_{m=1}^{\ell} \sum_{i=1}^{n} \boldsymbol{Q}_{im}^2 \mathrm{Var}(\zeta_i) \\
&\overset{(c)}{\leq} C(K+1)^2 \ell,
\end{aligned}$$

where $\zeta = \boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}$, and hence (a) follows from $\mathbb{E}[\zeta] = \mathbb{E}[\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}] = \boldsymbol{0}$, (b) follows from ζ having independent components and (c) follows from each component of ζ having ψ_2-norm bounded by $C(K+1)$. Therefore, it follows by union bound that for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\max_{j \in [p]} \left\|\boldsymbol{Q}\boldsymbol{Q}^T (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\right\|_2^2 \geq \ell C(K+1)^2 + t\Big) \\
&\quad \leq p \cdot \exp\Big(- c \min\Big(\frac{t^2}{C(K+1)^4 \ell}, \frac{t}{C(K+1)^2}\Big)\Big).
\end{aligned}$$

This completes the proof of Lemma 2.12.10.

■ 2.10 Proof of Theorem 2.5.2

Broadly, we proceed in three steps: (i) stating the Gaussian location model (GLM) and an associated minimax result; (ii) reducing GLM to an instance of error-in-variables regression; (iii) establishing a minimax result on the parameter estimation error of error-in-variables using the GLM minimax result.

Gaussian location model. Below, we introduce the GLM setting through a well-known minimax result.

Lemma 2.10.1 (Theorem 12.4 of Wu (2020)). *Let $\theta \sim \mathcal{N}(\theta^*, \sigma^2 \boldsymbol{I}_p)$, where $\boldsymbol{I}_p \in \mathbb{R}^{p \times p}$ is the identity matrix and $\theta, \theta^* \in \mathbb{R}^p$. Given θ, let $\widehat{\theta}$ be any estimator of θ. Then,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{B}_2} \mathbb{E}\widehat{\theta} - \theta^{*}{}^{2}_{2} = \Theta(\sigma^2 p \wedge 1).$$

Reducing GLM to error-in-variables. We will now show how an instance of GLM can be reduced to an instance of error-in-variables. Towards this, we follow the setup of Lemma 2.10.1 and define $\beta^* = \theta^*$, $\beta = \theta$, and $s = 1/\sigma$. For convenience, we write $\beta = \beta^* + \eta$, where the entries of η are independent Gaussian r.v.s with mean zero and variance $1/s^2$; hence $\beta \sim \mathcal{N}(\beta^*, (1/s^2)\boldsymbol{I}_p)$. Now, recall that the error-in-variables setting reveals a response vector $y = \boldsymbol{X}\beta^* + \varepsilon$ and covariate $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, where the parameter estimation objective is to recover β^* from $(y, \boldsymbol{Z})$. Below, we construct instances of these quantities using β, β^* as follows:

(i) Let the SVD of $\boldsymbol{X}$ be defined as $\boldsymbol{X} = su \otimes v$, where $u = (1, 0, \dots, 0)^T \in \mathbb{R}^n$ and $v = \beta^*$. Note by construction, $\text{rank}(\boldsymbol{X}) = 1$ and $\beta^* \in \text{rowspan}(\boldsymbol{X})$.

(ii) To construct y, we first sample $\varepsilon \in \mathbb{R}^n$ whose entries are independent standard normal r.v.s. Next, we define $y = su + \varepsilon$. From (i), we note that $\boldsymbol{X}\beta^* = su$ such that y can be equivalently expressed as $y = \boldsymbol{X}\beta^* + \varepsilon$.

(iii) Let $\boldsymbol{Z} = su \otimes \beta$. By construction, it follows that $\boldsymbol{Z} = \boldsymbol{X} + su \otimes \eta$. Note that $\boldsymbol{W} = su \otimes \eta$ is an $n \times p$ matrix whose entries in the first row are independent standard normal r.v.s and the remaining entries are zero.

Establishing minimax parameter estimation result. As stated above, the error-in-variables parameter estimation task is to construct $\widehat{\beta}$ from $(y, \boldsymbol{Z})$ such that $\widehat{\beta} - \beta^{*}{}_{2}$ vanishes as n, p grow. Using the above reduction combined with Lemma 2.10.1, it follows that

$$\inf_{\widehat{\beta}} \sup_{\beta^* \in \mathbb{B}_2} \mathbb{E}\widehat{\beta} - \beta^{*}{}^{2}_{2} = \Theta(p/s^2 \wedge 1).$$

To attain our desired result, it suffices to establish that $p/s^2 = \Omega(1)$. By (2.4) and under the assumption $n = O(p)$, we have that $s^2 \leq 2\text{snr}^2(n + p) \leq c\text{snr}^2 p$ for some $c > 0$.

As such, if snr $= O(1)$, then the minimax error is bounded below by a constant. This completes the proof.

■ 2.11 Proof of Theorem 2.5.3

Recall that $\boldsymbol{X}'$ and $\boldsymbol{Z}'$ denote the latent and observed testing covariates, respectively. We denote the SVD of the former as $\boldsymbol{X}' = \boldsymbol{U}'\boldsymbol{S}'(\boldsymbol{V}')^T$. Let s'_ℓ be the ℓ-th singular value of $\boldsymbol{X}'$. Further, recall that $\widetilde{\boldsymbol{Z}}' = (1/\widehat{\rho}')\boldsymbol{Z}'$, and its rank ℓ truncation is denoted as $\widetilde{\boldsymbol{Z}}'^{\ell}$. Our interest is in bounding $\widetilde{\boldsymbol{Z}}'^{\ell}\widehat{\beta} - \boldsymbol{X}'\beta^*{}_2$. Towards this, consider

$$\begin{aligned}\widetilde{\boldsymbol{Z}}'^{\ell}\widehat{\beta} - \boldsymbol{X}'\beta^{*2}_{2} &= \widetilde{\boldsymbol{Z}}'^{\ell}\widehat{\beta} - \widetilde{\boldsymbol{Z}}'^{\ell}\beta^* + \widetilde{\boldsymbol{Z}}'^{\ell}\beta^* - \boldsymbol{X}'\beta^{*2}_{2} \\ &\leq 2\widetilde{\boldsymbol{Z}}'^{\ell}(\widehat{\beta} - \beta^*)^2_2 + 2(\widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}')\beta^{*2}_{2}. \end{aligned} \tag{2.48}$$

We shall bound the two terms on the right hand side of (2.48) next.

Bounding $\widetilde{\boldsymbol{Z}}'^{\ell}(\widehat{\beta} - \beta^*)^2_2$. Since $\widetilde{\boldsymbol{Z}}'^{\ell} = (1/\widehat{\rho}')\boldsymbol{Z}'^{\ell}$, we have

$$\begin{aligned}\widetilde{\boldsymbol{Z}}'^{\ell}(\widehat{\beta} - \beta^*)^2_2 &= \frac{1}{(\widehat{\rho}')^2}\boldsymbol{Z}'^{\ell}(\widehat{\beta} - \beta^*)^2_2 \\ &= \frac{1}{(\widehat{\rho}')^2}(\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}' + \rho\boldsymbol{X}')(\widehat{\beta} - \beta^*)^2_2 \\ &\leq \frac{2}{(\widehat{\rho}')^2}(\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}')(\widehat{\beta} - \beta^*)^2_2 + 2\left(\frac{\rho}{\widehat{\rho}'}\right)^2\boldsymbol{X}'(\widehat{\beta} - \beta^*)^2_2. \end{aligned} \tag{2.49}$$

Now, note that $\boldsymbol{Z}' - \boldsymbol{Z}'^{\ell}{}_2$ is the $(\ell + 1)$-st largest singular value of $\boldsymbol{Z}'$. Therefore, by Weyl's inequality (Lemma 5.14.4), we have for any $\ell \geq r'$,

$$\boldsymbol{Z}' - \boldsymbol{Z}'^{\ell}{}_2 \leq \boldsymbol{Z}' - \rho\boldsymbol{X}'{}_2.$$

In turn, this gives

$$\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}'{}_2 \leq \boldsymbol{Z}'^{\ell} - \boldsymbol{Z}'{}_2 + \boldsymbol{Z}' - \rho\boldsymbol{X}'{}_2 \leq 2\boldsymbol{Z}' - \rho\boldsymbol{X}'{}_2.$$

Thus, we have

$$(\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}')(\widehat{\beta} - \beta^*)^2_2 \leq 4\boldsymbol{Z}' - \rho\boldsymbol{X}'^2_2\ \widehat{\beta} - \beta^{*2}_{2}. \tag{2.50}$$

Recall that $\boldsymbol{V}$ and $\boldsymbol{V}_\perp$ span the rowspace and nullspace of $\boldsymbol{X}$, respectively. By Assumption 5, it follows that $\boldsymbol{V}'^T\boldsymbol{V}_\perp = \mathbf{0}$ and hence $\boldsymbol{X}'\boldsymbol{V}_\perp\boldsymbol{V}_\perp^T = \mathbf{0}$. As a result,

$$\begin{aligned}\boldsymbol{X}'(\widehat{\beta} - \beta^*)_2^2 &= \boldsymbol{X}'(\boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{V}_\perp\boldsymbol{V}_\perp^T)(\widehat{\beta} - \beta^*)_2^2 \\ &= \boldsymbol{X}'\boldsymbol{V}\boldsymbol{V}^T(\widehat{\beta} - \beta^*)_2^2 \\ &\leq \boldsymbol{X}'^2_2\ \boldsymbol{V}\boldsymbol{V}^T(\widehat{\beta} - \beta^*)_2^2.\end{aligned}$$

Recalling that $\widehat{\boldsymbol{V}}_r$ denotes the top r right singular vectors of $\widetilde{\boldsymbol{Z}}^r$, consider

$$\begin{aligned}\boldsymbol{V}\boldsymbol{V}^T(\widehat{\beta} - \beta^*)_2^2 &= (\boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^T + \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^T)(\widehat{\beta} - \beta^*)_2^2 \\ &\leq 2\boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^{T2}_2\ \widehat{\beta} - \beta^{*2}_2 + 2\widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^T(\widehat{\beta} - \beta^*)_2^2.\end{aligned}$$

From (3.53) and above, we obtain

$$\begin{aligned}\boldsymbol{V}\boldsymbol{V}^T(\widehat{\beta} - \beta^*)_2^2 \leq C\boldsymbol{V}\boldsymbol{V}^T &- \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^{T2}_2\ \widehat{\beta} - \beta^{*2}_2 \\ &+ \frac{C}{\widehat{s}_r^2}\Big(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^{r2}_{2,\infty}\beta^{*2}_1 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\beta} - \beta^*), \varepsilon\rangle\Big).\end{aligned}$$

Thus,

$$\begin{aligned}\boldsymbol{X}'(\widehat{\beta} - \beta^*)_2^2 \leq C\boldsymbol{X}'^2_2\ \boldsymbol{V}\boldsymbol{V}^T &- \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^{T2}_2\ \widehat{\beta} - \beta^{*2}_2 \\ &+ \frac{C\boldsymbol{X}'^2_2}{\widehat{s}_r^2}\Big(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^{r2}_{2,\infty}\beta^{*2}_1 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\beta} - \beta^*), \varepsilon\rangle\Big). \qquad (2.51)\end{aligned}$$

In summary, plugging (2.50) and (2.51) into (2.49), we have

$$\begin{aligned}\widetilde{\boldsymbol{Z}}'^{,\ell}(\widehat{\beta} - \beta^*)_2^2 \leq\ & \frac{C}{(\widehat{\rho}')^2}\boldsymbol{Z}' - \rho\boldsymbol{X}'^2_2\ \widehat{\beta} - \beta^{*2}_2 \\ &+ C\left(\frac{\rho}{\widehat{\rho}'}\right)^2\boldsymbol{X}'^2_2\ \boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^{T2}_2\ \widehat{\beta} - \beta^{*2}_2 \\ &+ \frac{C\rho^2\boldsymbol{X}'^2_2}{(\widehat{\rho}')^2\widehat{s}_r^2}\Big(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^{r2}_{2,\infty}\beta^{*2}_1 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\beta} - \beta^*), \varepsilon\rangle\Big). \qquad (2.52)\end{aligned}$$

Bounding $(\widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}')\beta^{*2}_2$. Using inequality (3.47),

$$(\widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}')\beta^{*2}_2 \leq \widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}'^2_{2,\infty}\beta^{*2}_1. \qquad (2.53)$$

Combining. Incorporating (2.52) and (2.53) into (2.48) with $\ell = r'$ yields

$$\widetilde{\boldsymbol{Z}}'^{r'}\widehat{\beta} - \boldsymbol{X}'\beta^{*2}_{2} \leq \Delta_1 + \Delta_2, \tag{2.54}$$

where

$$\begin{aligned}\Delta_1 &= \frac{C}{(\hat{\rho}')^2}\boldsymbol{Z}' - \rho\boldsymbol{X}'^{2}_{2}\ \widehat{\beta} - \beta^{*2}_{2} + C\Big(\frac{\rho s'_1}{\hat{\rho}'}\Big)^2 \boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r{}^{T2}_{2}\ \widehat{\beta} - \beta^{*2}_{2} \\ &\quad + 2\boldsymbol{X}' - \widetilde{\boldsymbol{Z}}'^{r'2}_{2,\infty}\beta^{*2}_{1}, \\ \Delta_2 &= C\Big(\frac{\rho s'_1}{\hat{\rho}'\widehat{s}_r}\Big)^2\Big(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^{r2}_{2,\infty}\beta^{*2}_{1} + \langle\widetilde{\boldsymbol{Z}}^{r}(\widehat{\beta} - \beta^*), \varepsilon\rangle\Big).\end{aligned}$$

Note that (2.54) is a deterministic bound. We will now proceed to bound Δ_1 and Δ_2, first in high probability then in expectation.

Bound in high-probability. We first bound Δ_1. First we note that by adapting Lemma 2.12.9 with $\hat{\rho}'$ in place of $\hat{\rho}$, we obtain with probability at least $1 - O(1/(mp)^{10})$,

$$\rho/2 \leq \hat{\rho}' \leq \rho. \tag{2.55}$$

By adapting Lemma 2.12.8 for $\boldsymbol{Z}', \boldsymbol{X}'$ in place of $\boldsymbol{Z}, \boldsymbol{X}$, we have with probability at least $1 - O(1/(mp)^{10})$,

$$\boldsymbol{Z}' - \rho\boldsymbol{X}'_2 \leq C(K, \gamma)(\sqrt{m} + \sqrt{p}).$$

Hence, using Theorem 2.5.1 and (2.55), we have that with probability at least $1 - O(1/((n \wedge m)p)^{10})$

$$\frac{1}{(\hat{\rho}')^2 m}\boldsymbol{Z}' - \rho\boldsymbol{X}'^{2}_{2}\widehat{\beta} - \beta^{*2}_{2} \leq \frac{C(K, \gamma, \sigma)\log np}{\rho^2} \cdot (1 + \frac{p}{m})\left(\frac{r\beta^{*2}_{2}}{\mathrm{snr}^2} + \frac{\beta^{*2}_{1}}{\mathrm{snr}^4}\right) \tag{2.56}$$

Note, $s'_1 = O(\sqrt{mp})$ which follows by Assumption 1. Using this bound on s'_1 and recalling Lemma 2.9.1, (2.26), and Theorem 2.5.1, it follows that with probability at least $1 - O(1/(np)^{10})$

$$\begin{aligned}&\Big(\frac{\rho s'_1}{\hat{\rho}'}\Big)^2\frac{1}{m}\boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r^{T2}{}_{2}\widehat{\beta} - \beta^{*2}_{2} \\ &\quad\leq C(K, \gamma)\frac{mp}{m}\frac{n + p}{\rho^2 s_r^2}\widehat{\beta} - \beta^{*2}_{2} \leq C(K, \gamma, \sigma)\log np \cdot \left(\frac{rp\beta^{*2}_{2}}{\mathrm{snr}^4} + \frac{p\beta^{*2}_{1}}{\mathrm{snr}^6}\right).\end{aligned} \tag{2.57}$$

Next, we adapt Lemma 2.9.2 for $\widetilde{Z}', X'$ in place of $\widetilde{Z}, X$ with $\ell = r'$. If $\rho \geq c\frac{\log^2 mp}{mp}$, then with probability at least $1 - O(1/(mp)^{10})$

$$\begin{aligned}
&\frac{1}{m}X' - \widetilde{Z}'^{r'}{}^2_{2,\infty}\beta^*{}^2_1 \\
&\leq \frac{C(K,\gamma)}{m}\Big(\frac{(m+p)(m+\sqrt{m}\log(mp))}{\rho^4(s'_r)^2} + \frac{r'+\sqrt{r'}\log(mp)}{\rho^2}\Big) + C\frac{\log(mp)}{\rho p}\beta^*{}^2_1 \\
&\leq \frac{C(K,\gamma)\log(mp)}{\rho^2}\Big(\frac{1}{\text{snr}^2_{\text{test}}} + \frac{r'}{m}\Big)\beta^*{}^2_1. \qquad (2.58)
\end{aligned}$$

Note that the above uses the inequality $\frac{m+p}{\rho^2(s'_{r'})^2} \leq \frac{1}{\text{snr}^2_{\text{test}}}$, which follows from the definition of $\text{snr}^2_{\text{test}}$ in (2.8). Hence, by using (2.56), (2.57), (2.58), we conclude that with probability at least $1 - O(1/((n \wedge m)p)^{10})$,

$$\begin{aligned}
\frac{\Delta_1}{m} \leq C(K,\gamma,\sigma)\log((n\vee m)p)\Bigg[&\frac{r(1+\frac{p}{m})\beta^*{}^2_2}{(\rho^2\text{snr}^2)} + \frac{1}{\text{snr}^4}\left((1+\frac{p}{m})\frac{\beta^*{}^2_1}{\rho^2} + rp\beta^*{}^2_2\right) \\
&+ \frac{p\beta^*{}^2_1}{\text{snr}^6} + \left(\frac{1}{\text{snr}^2_{\text{test}}} + \frac{r'}{m}\right)\frac{\beta^*{}^2_1}{\rho^2}\Bigg]. \qquad (2.59)
\end{aligned}$$

We will now bound Δ_2. As per (2.22), with probability at least $1 - O(1/(np)^{10})$,

$$\begin{aligned}
&X - \widetilde{Z}^r{}^2_{2,\infty}\beta^*{}^2_1 + \langle\widetilde{Z}^r(\widehat{\beta} - \beta^*), \varepsilon\rangle \\
&\leq CX - \widetilde{Z}^r{}^2_{2,\infty}\beta^*{}^2_1 + C\sigma^2(r + \log(np)) + C\sigma\sqrt{n\log(np)}\beta^*{}_1. \qquad (2.60)
\end{aligned}$$

Recalling Lemma 2.9.2 and the definition of snr, we have that with probability at least $1 - O(1/(np)^{10})$,

$$X - \widetilde{Z}^r{}^2_{2,\infty} \leq \frac{C(K,\gamma)\log(np)}{\rho^2}\Big(\frac{n}{\text{snr}^2} + r\Big). \qquad (2.61)$$

Using (2.25), (2.26), (2.55), we have

$$\Big(\frac{\rho s'_1}{\widehat{\rho}'\widehat{s}_r}\Big)^2 \leq \frac{C(s'_1)^2\rho^2}{\text{snr}^2(n+p)}. \qquad (2.62)$$

Therefore, (2.55), (2.60), (2.61), and (2.62), the bound $s'_1 = O(\sqrt{mp})$, and the assumption

${\beta^*}_2 = \Omega(1)$ imply that with probability at least $1 - O(1/((n \wedge m)p)^{10})$,

$$\begin{aligned} \frac{\Delta_2}{m} &\leq \frac{C(K,\gamma)(s_1')^2 \log(np)\rho^2}{m(n+p)\mathrm{snr}^2} \left(\frac{{\beta^*}_1^2}{\rho^2}(\frac{n}{\mathrm{snr}^2} + r) + \sigma^2 r + \sigma\sqrt{n}{\beta^*}_1 \right) \\ &\leq \frac{C(K,\gamma,\sigma)\log(np)}{\mathrm{snr}^2} \left(\frac{n{\beta^*}_1^2}{\mathrm{snr}^2} + r{\beta^*}_1^2 + \sqrt{n}{\beta^*}_1 \right). \end{aligned} \tag{2.63}$$

Incorporating (2.59) and (2.63) into (2.54),

$$\begin{aligned} \frac{\Delta_1 + \Delta_2}{m} \leq C(K,\gamma,\sigma)\log((n \vee m)p) \Bigg[&\frac{1}{\mathrm{snr}^2} \left(\frac{r(1+\frac{p}{m}){\beta^*}_2^2}{\rho^2} + r{\beta^*}_1^2 + \sqrt{n}{\beta^*}_1 \right) \\ &+ \frac{1}{\mathrm{snr}^4} \left(\left(\frac{1}{\rho^2}(1+\frac{p}{m}) + n \right) {\beta^*}_1^2 + rp{\beta^*}_2^2 \right) \\ &+ \frac{p{\beta^*}_1^2}{\mathrm{snr}^6} + \left(\frac{1}{\mathrm{snr}_{\text{test}}^2} + \frac{r'}{m} \right) {\beta^*}_1^2 \Bigg]. \end{aligned} \tag{2.64}$$

Observing that ${\beta^*}_2 \leq {\beta^*}_1$ and using the assumption that $\mathrm{snr} \geq C(K,\gamma,\sigma)$, we have

$$\frac{1}{\mathrm{snr}^2} \left(\frac{r(1+\frac{p}{m}){\beta^*}_2^2}{\rho^2} + r{\beta^*}_1^2 \right), \ \frac{1}{\mathrm{snr}^4} \left(\left(\frac{1}{\rho^2}(1+\frac{p}{m}) \right) {\beta^*}_1^2 \right) \leq \frac{1}{\mathrm{snr}^2} \left(\frac{r(1+\frac{p}{m}){\beta^*}_1^2}{\rho^2} \right) \tag{2.65}$$

$$\frac{1}{\mathrm{snr}^4} \left(n{\beta^*}_1^2 + rp{\beta^*}_2^2 \right), \ \frac{p{\beta^*}_1^2}{\mathrm{snr}^6} \leq \frac{1}{\mathrm{snr}^4} \left(r(n \vee p){\beta^*}_1^2 \right) \tag{2.66}$$

$$\left(\frac{1}{\mathrm{snr}_{\text{test}}^2} + \frac{r'}{m} \right) {\beta^*}_1^2 \leq \frac{r{\beta^*}_1^2}{\mathrm{snr}_{\text{test}}^2 \wedge m}, \tag{2.67}$$

where in the last equality, we used Assumption 5. Using (2.65), (2.66), (2.67) in (2.64) and simplifying concludes the high-probability bound.

Bound in expectation. Here, we assume that $\{\langle x_i, \beta^* \rangle \in [-b, b] : i > n\}$. As such, we enforce $\{\widehat{y}_i \in [-b, b] : i > n\}$. With (2.54), this yields

$$\mathrm{MSE}_{\text{test}} \leq \frac{1}{m} \widetilde{\boldsymbol{Z}}'^{r'} \widehat{\beta} - \boldsymbol{X}' {\beta^*}_2^2 \leq \frac{1}{m}(\Delta_1 + \Delta_2).$$

We define $\mathcal{E}$ as the event such that the bounds in (2.56), (2.57), (2.58), (2.55), (2.61), and Lemma 2.9.3 hold. Thus, if $\mathcal{E}$ occurs, then (2.59) implies that

$$\frac{1}{m}\mathbb{E}[\Delta_1 | \mathcal{E}] \leq C(K,\gamma,\sigma)\log((n \vee m)p) \Bigg[\frac{r(1+\frac{p}{m}){\beta^*}_2^2}{(\rho^2 \mathrm{snr}^2)} + \frac{1}{\mathrm{snr}^4} \left((1+\frac{p}{m})\frac{{\beta^*}_1^2}{\rho^2} + rp{\beta^*}_2^2 \right) \tag{2.68}$$

$$+ \frac{p\beta^{*2}_{1}}{\mathrm{snr}^6} + \left(\frac{1}{\mathrm{snr}^2_{\mathrm{test}}} + \frac{r'}{m}\right)\frac{\beta^{*2}_{1}}{\rho^2}\Bigg].$$

Next, we bound $\mathbb{E}[\Delta_2 | \mathcal{E}]$. To do so, observe that ε is independent of the event $\mathcal{E}$. Thus, by (3.60), we have

$$\begin{aligned}\mathbb{E}[\langle \widetilde{\boldsymbol{Z}}^r(\widehat{\beta} - \beta^*), \varepsilon\rangle | \mathcal{E}] &= \mathbb{E}[\langle \widehat{\boldsymbol{U}}_r \widehat{\boldsymbol{U}}_r^T \boldsymbol{X}\beta^*, \varepsilon\rangle + \langle \widehat{\boldsymbol{U}}_r \widehat{\boldsymbol{U}}_r^T \varepsilon, \varepsilon\rangle - \langle \widehat{\boldsymbol{U}}_r \widehat{\boldsymbol{\Sigma}}_r \widehat{\boldsymbol{V}}_r^T \beta^*, \varepsilon\rangle | \mathcal{E}] \\ &= \mathbb{E}[\langle \widehat{\boldsymbol{U}}_r \widehat{\boldsymbol{U}}_r^T \varepsilon, \varepsilon\rangle | \mathcal{E}] \le C\sigma^2 r.\end{aligned}$$

Combining the above inequality with (2.61),

$$\frac{1}{m}\mathbb{E}[\Delta_2 | \mathcal{E}] \le C(K, \gamma, \sigma)\log(np)\left(\frac{r\beta^{*2}_{1}}{\mathrm{snr}^2} + \frac{n\beta^{*2}_{1}}{\mathrm{snr}^4}\right). \tag{2.69}$$

Due to truncation, observe that $\mathrm{MSE}_{\mathrm{test}}$ is always bounded above by $4b^2$. Thus,

$$\begin{aligned}\mathbb{E}[\mathrm{MSE}_{\mathrm{test}}] &\le \mathbb{E}[\mathrm{MSE}_{\mathrm{test}} | \mathcal{E}] + \mathbb{E}[\mathrm{MSE}_{\mathrm{test}} | \mathcal{E}^c]\ \mathbb{P}(\mathcal{E}^c) \\ &\le \frac{1}{m}\mathbb{E}[\Delta_1 + \Delta_2 | \mathcal{E}] + Cb^2\left(1/(np)^{10} + 1/(mp)^{10}\right).\end{aligned}$$

Plugging (2.65), (2.66), (2.67) into (2.68); then using that bound along with (2.69) in the inequality above completes the proof.

■ 2.12 Helpful Concentration Inequalities

In this section, we state and prove a number of helpful concentration inequalities used to establish our primary results.

Lemma 2.12.1. *Let X be a mean zero, sub-gaussian random variable. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\exp(\lambda X) \le \exp\left(C\lambda^2 \|X\|^2_{\psi_2}\right).$$

Lemma 2.12.2. *Let $X_1, \dots, X_n$ be independent, mean zero, sub-gaussian random variables. Then,*

$$\sum_{i=1}^n X^2_{i\,\psi_2} \le C\sum_{i=1}^n \|X_i\|^2_{\psi_2}.$$

Theorem 2.12.1 (Bernstein's inequality). *Let $X_1, \dots, X_n$ be independent, mean zero,* *sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} X_i\Big| \geq t\Big) \leq 2\exp\Big(-c\min\Big(\frac{t^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\Big)\Big),$$

where $c > 0$ is an absolute constant.

Lemma 2.12.3 (Modified Hoeffding Inequality). *Let $X \in \mathbb{R}^n$ be random vector with independent mean-zero sub-Gaussian random coordinates with $X_{i\psi_2} \leq K$. Let $a \in \mathbb{R}^n$ be another random vector that satisfies $a_2 \leq b$ almost surely for some constant $b \geq 0$. Then for all $t \geq 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} a_i X_i\Big| \geq t\Big) \leq 2\exp\Big(-\frac{ct^2}{K^2 b^2}\Big),$$

where $c > 0$ is a universal constant.

Proof. Let $S_n = \sum_{i=1}^{n} a_i X_i$. Then applying Markov's inequality for any $\lambda > 0$, we obtain

$$\begin{aligned}\mathbb{P}\left(S_n \geq t\right) &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda t)\right)\\ &\leq \mathbb{E}\left[\exp(\lambda S_n)\right] \cdot \exp(-\lambda t)\\ &= \mathbb{E}_a\left[\mathbb{E}\left[\exp(\lambda S_n) \mid a\right]\right] \cdot \exp(-\lambda t).\end{aligned}$$

Now, conditioned on the random vector a, observe that

$$\mathbb{E}\left[\exp(\lambda S_n)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\Big(CK^2\lambda^2 a_2^2\Big) \leq \exp\Big(CK^2\lambda^2 b^2\Big),$$

where the equality follows from conditional independence, the first inequality by Lemma 2.13.1, and the final inequality by assumption. Therefore,

$$\mathbb{P}\left(S_n \geq t\right) \leq \exp\Big(CK^2\lambda^2 b^2 - \lambda t\Big).$$

Optimizing over λ yields the desired result:

$$\mathbb{P}\left(S_n \geq t\right) \leq \exp\Big(-\frac{ct^2}{K^2 b^2}\Big).$$

Applying the same arguments for $-\langle X, a\rangle$ gives a tail bound in the other direction. ■

Lemma 2.12.4 (Modified Hanson-Wright Inequality)**.** *Let $X \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with $X_{i\psi_2} \leq K$. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a random matrix satisfying $\boldsymbol{A}_2 \leq a$ and $\boldsymbol{A}_F^2 \leq b$ almost surely for some $a, b \geq 0$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\left|X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X]\right| \geq t\right) \leq 2 \cdot \exp\left(-c \min\left(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\right)\right).$$

Proof. The proof follows similarly to that of Theorem 6.2.1 of Vershynin (2018). Using the independence of the coordinates of X, we have the following useful diagonal and off-diagonal decomposition:

$$X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X] = \sum_{i=1}^n \left(A_{ii} X_i^2 - \mathbb{E}[A_{ii} X_i^2]\right) + \sum_{i \neq j} A_{ij} X_i X_j.$$

Therefore, letting

$$p = \mathbb{P}\left(X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X] \geq t\right),$$

we can express

$$p \leq \mathbb{P}\Big(\sum_{i=1}^n \left(A_{ii} X_i^2 - \mathbb{E}[A_{ii} X_i^2]\right) \geq t/2\Big) + \mathbb{P}\Big(\sum_{i \neq j} A_{ij} X_i X_j \geq t/2\Big) =: p_1 + p_2.$$

We will now proceed to bound each term independently.

Step 1: diagonal sum. Let $S_n = \sum_{i=1}^n (A_{ii} X_i^2 - \mathbb{E}[A_{ii} X_i^2])$. Applying Markov's inequality for any $\lambda > 0$, we have

$$\begin{aligned} p_1 &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda t/2)\right) \\ &\leq \mathbb{E}_{\boldsymbol{A}} \mathbb{E}\left[\left[\exp(\lambda S_n) \mid \boldsymbol{A}\right]\right] \cdot \exp(-\lambda t/2). \end{aligned}$$

Since the X_i are independent, sub-Gaussian random variables, $X_i^2 - \mathbb{E}[X_i^2]$ are independent mean-zero sub-exponential random variables, satisfying

$$\left\|X_i^2 - \mathbb{E}[X_i^2]\right\|_{\psi_1} \leq C \left\|X_i^2\right\|_{\psi_1} \leq C \left\|X_i\right\|_{\psi_2}^2 \leq C K^2.$$

Conditioned on $\boldsymbol{A}$ and optimizing over λ using standard arguments, yields

$$p_1 \leq \exp\Big(- c \min\Big(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\Big)\Big).$$

Step 2: off-diagonals. Let $S = \sum_{i \neq j} A_{ij} X_i X_j$. Again, applying Markov's inequality for any $\lambda > 0$, we have

$$p_2 = \mathbb{P}\left(\exp(\lambda S) \geq \exp(\lambda t/2)\right) \leq \mathbb{E}_{\boldsymbol{A}}\left[\mathbb{E}\left[\exp(\lambda S) \mid \boldsymbol{A}\right]\right] \cdot \exp(-\lambda t/2).$$

Let g be a standard multivariate gaussian random vector. Further, let X' and g' be independent copies of X and g, respectively. Conditioning on $\boldsymbol{A}$ yields

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda S)\right] &\leq \mathbb{E}\Big[\exp\Big(4\lambda X^T \boldsymbol{A} X'\Big)\Big] && \text{(by Decoupling Remark 6.1.3 of Vershynin (2018))} \\
&\leq \mathbb{E}\Big[\exp\Big(C_1 \lambda g^T A g'\Big)\Big] && \text{(by Lemma 6.2.3 of Vershynin (2018))} \\
&\leq \exp\Big(C_2 \lambda^2 \left\|A\right\|_F^2\Big) && \text{(by Lemma 6.2.2 of Vershynin (2018))} \\
&\leq \exp\Big(C_2 \lambda^2 b\Big),
\end{aligned}$$

where $|\lambda| \leq c/a$. Optimizing over λ then gives

$$p_2 \leq \exp\Big(- c \min\Big(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\Big)\Big).$$

Step 3: combining. Putting everything together completes the proof. ■

■ 2.12.1 Proof of Lemma 2.9.1

Recall that $\boldsymbol{U}, \boldsymbol{V}$ denote the left and right singular vectors of $\boldsymbol{X}$ (equivalently, $\rho \boldsymbol{X}$), respectively; meanwhile, $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k$ denote the top k left and right singular vectors of $\widetilde{\boldsymbol{Z}}$ (equivalently, $\boldsymbol{Z}$), respectively. Further, observe that $\mathbb{E}[\boldsymbol{Z}] = \rho \boldsymbol{X}$ and let $\tilde{\boldsymbol{W}} = \boldsymbol{Z} - \rho \boldsymbol{X}$. To arrive at our result, we recall Wedin's Theorem Wedin (1972).

Theorem 2.12.2 (Wedin's Theorem). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times p}$, let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ and $\boldsymbol{B} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$ be their respective SVDs. Let $\boldsymbol{U}_k, \boldsymbol{V}_k$ (respectively, $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k$) correspond to the truncation of $\boldsymbol{U}, \boldsymbol{V}$ (respectively, $\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}$) that retains the columns corresponding to the top k singular* *values of* $\boldsymbol{A}$ *(respectively,* $\boldsymbol{B}$*). Let* s_k *denote the* k*-th singular value of* $\boldsymbol{A}$*. Then,*

$$\max\left(U_k U_k^T - \widehat{U}_k \widehat{U}_k^T{}_2, V_k V_k^T - \widehat{V}_k \widehat{V}_k^T{}_2\right) \leq \frac{2\left\|\boldsymbol{A}-\boldsymbol{B}\right\|_2}{s_k - s_{k+1}}.$$

Using Theorem 3.15.1 for $k = r$, it follows that

$$\max\left(\boldsymbol{U}\boldsymbol{U}^T - \widehat{\boldsymbol{U}}_r \widehat{\boldsymbol{U}}_r^T{}_2, \boldsymbol{V}\boldsymbol{V}^T - \widehat{\boldsymbol{V}}_r \widehat{\boldsymbol{V}}_r^T{}_2\right) \leq \frac{2\tilde{\boldsymbol{W}}_2}{\rho s_r}, \tag{2.70}$$

where s_r is the smallest nonzero singular value of $\boldsymbol{X}$. Next, we obtain a high probability bound on $\tilde{\boldsymbol{W}}_2$. To that end,

$$\frac{1}{n}\tilde{\boldsymbol{W}}_2^2 = \frac{1}{n}\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}_2 \leq \frac{1}{n}\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}} - \mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 + \frac{1}{n}\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2. \tag{2.71}$$

We bound the two terms in (2.71) separately. We recall the following lemma, which is a direct extension of Theorem 4.6.1 of Vershynin (2018) for the non-isotropic setting, and we present its proof for completeness in Section 2.12.5.

Lemma 2.12.5 (Independent sub-gaussian rows). *Let* $\boldsymbol{A}$ *be an* $n \times p$ *matrix whose rows* A_i *are independent, mean zero, sub-gaussian random vectors in* $\mathbb{R}^p$ *with second moment matrix* $\boldsymbol{\Sigma} = (1/n)\mathbb{E}[\boldsymbol{A}^T\boldsymbol{A}]$*. Then for any* $t \geq 0$*, the following inequality holds with probability at least* $1 - \exp\left(-t^2\right)$*:*

$$\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{\Sigma}_2 \leq K^2 \max(\delta, \delta^2), \quad \textit{where } \delta = C\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}; \tag{2.72}$$

here, $K = \max_i \left\|A_i\right\|_{\psi_2}$*.*

The matrix $\tilde{\boldsymbol{W}} = \boldsymbol{Z} - \rho\boldsymbol{X}$ has independent rows by Assumption 4. We state the following Lemma about the distribution property of the rows of $\tilde{\boldsymbol{W}}$, the proof of which can be found in Section 2.12.6.

Lemma 2.12.6. *Let Assumption 4 hold. Then,* $z_i - \rho x_i$ *is a sequence of independent, mean zero, sub-gaussian random vectors satisfying* $z_i - \rho x_{i\psi_2} \leq C(K+1)$*.*

From Lemmas 2.12.5 and 2.12.6, with probability at least $1 - \exp\left(-t^2\right)$,

$$\frac{1}{n}\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}} - \mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 \leq C(K+1)^2(1 + \frac{p}{n} + \frac{t^2}{n}). \tag{2.73}$$

Finally, we claim the following bound on $\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2$, the proof of which is in Section 2.12.7.

Lemma 2.12.7. *Let Assumption 4 hold. Then, we have*

$$\mathbb{E}[\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}}]_2 \leq C(K+1)^2 n(\rho-\rho^2) + n\rho^2\gamma^2.$$

From (2.71), (2.73) and Lemma 2.12.7, it follows that with probability at least $1-\exp\big(-t^2\big)$ for any $t>0$, we have

$$\tilde{\boldsymbol{W}}_2^2 \leq C(K+1)^2(n+p+t^2) + n(\rho(1-\rho)(K+1)^2+\rho^2\gamma^2).$$

For this, we conclude the following lemma.

Lemma 2.12.8. *For any $t>0$, the following holds with probability at least $1-\exp\big(-t^2\big)$:*

$$\boldsymbol{Z}-\rho\boldsymbol{X}_2 \leq C(K,\gamma)(\sqrt{n}+\sqrt{p}+t).$$

Using the above and (2.70), we conclude the proof of Lemma 2.9.1.

■ 2.12.2 Proof of Lemma 2.9.2

We want to bound $\boldsymbol{X}-\widetilde{\boldsymbol{Z}}^{k}{}^2_{2,\infty}$. To that end, let $\Delta_j = \boldsymbol{X}_{\cdot j} - \widetilde{\boldsymbol{Z}}^k_{\cdot j}$ for any $j\in[p]$. Our interest is in bounding $\Delta_j{}^2_2$ for all $j\in[p]$. Consider,

$$\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \boldsymbol{X}_{\cdot j} = (\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}) + (\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}).$$

Now, note that $\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}$ belongs to the subspace spanned by column vectors of $\widehat{\boldsymbol{U}}_k$, while $\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}$ belongs to its orthogonal complement with respect to $\mathbb{R}^n$. As a result,

$$\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \boldsymbol{X}_{\cdot j}{}^2_2 = \widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}{}^2_2 + \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}{}^2_2. \tag{2.74}$$

Bounding $\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}{}^2_2$. Recall that $\widetilde{\boldsymbol{Z}} = (1/\hat{\rho})\boldsymbol{Z} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$, and hence $\boldsymbol{Z} = \hat{\rho}\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T$.

Consequently,

$$\begin{aligned}\frac{1}{\hat{\rho}}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{Z}_{\cdot j} &= \frac{1}{\hat{\rho}}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{Z}\boldsymbol{e}_j = \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^T\boldsymbol{e}_j \\ &= \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{V}}_k^T\boldsymbol{e}_j = \widetilde{\boldsymbol{Z}}^k_{\cdot j}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j} &= \frac{1}{\hat{\rho}}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{Z}_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j} \\ &= \frac{1}{\hat{\rho}}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}) + \Big(\frac{\rho-\hat{\rho}}{\hat{\rho}}\Big)\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}.\end{aligned}$$

Therefore,

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}{}_2^2 &\le \frac{2}{\hat{\rho}^2}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 + 2\Big(\frac{\rho-\hat{\rho}}{\hat{\rho}}\Big)^2\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}{}_2^2 \\ &\le \frac{2}{\hat{\rho}^2}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 + 2\Big(\frac{\rho-\hat{\rho}}{\hat{\rho}}\Big)^2\boldsymbol{X}_{\cdot j}{}_2^2,\end{aligned}$$

where we have used the fact that $\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T{}_2 = 1$. Recall that $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ represents the left singular vectors of $\boldsymbol{X}$. Thus,

$$\begin{aligned}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 &\le 2(\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T - \boldsymbol{U}\boldsymbol{U}^T)(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 + 2\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 \\ &\le 2\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T - \boldsymbol{U}\boldsymbol{U}^T{}_2^2\ \boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}{}_2^2 + 2\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2.\end{aligned}$$

By Assumption 1, we have that $\boldsymbol{X}_{\cdot j}{}_2^2 \le n$. This yields

$$\begin{aligned}\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\boldsymbol{X}_{\cdot j}{}_2^2 &\le \frac{4}{\hat{\rho}^2}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T - \boldsymbol{U}\boldsymbol{U}^T{}_2^2\ \boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}{}_2^2 \\ &\quad + \frac{4}{\hat{\rho}^2}\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 + 2n\Big(\frac{\rho-\hat{\rho}}{\hat{\rho}}\Big)^2. \qquad (2.75)\end{aligned}$$

We now state Lemmas 2.12.9 and 2.12.10. Their proofs are in Sections 2.12.8 and 2.12.9, respectively.

Lemma 2.12.9. *For any $\alpha > 1$,*

$$\mathbb{P}\left(\rho/\alpha \le \widehat{\rho} \le \alpha\rho\right) \ge 1 - 2\exp\Big(-\frac{(\alpha-1)^2 np\rho}{2\alpha^2}\Big).$$

Therefore, for $\rho \geq c\frac{\log^2 np}{np}$, we have with probability $1 - O(1/(np)^{10})$

$$\frac{\rho}{2} \leq \hat{\rho} \leq 2\rho \quad \textit{and} \quad \left(\frac{\rho - \hat{\rho}}{\hat{\rho}}\right)^2 \leq C\frac{\log(np)}{\rho np}.$$

Lemma 2.12.10. *Consider any matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times \ell}$ with $1 \leq \ell \leq n$ such that its columns $\boldsymbol{Q}_{\cdot j}$ for $j \in [\ell]$ are orthonormal vectors. Then for any $t > 0$,*

$$\mathbb{P}\left(\max_{j \in [p]} \left\|\boldsymbol{Q}\boldsymbol{Q}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})\right\|_2^2 \geq \ell C(K+1)^2 + t\right)$$
$$\leq p \cdot \exp\left(-c\min\left(\frac{t^2}{C(K+1)^4\ell}, \frac{t}{C(K+1)^2}\right)\right).$$

Subsequently, with probability $1 - O(1/(np)^{10})$,

$$\max_{j \in [p]} \left\|\boldsymbol{Q}\boldsymbol{Q}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})\right\|_2^2 \leq C(K+1)^2(\ell + \sqrt{\ell}\log(np)).$$

Both terms $\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}{}_2^2$ and $\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2$ can be bounded by Lemma 2.12.10: for the first term $\boldsymbol{Q} = \mathcal{D}$, and for the second term $\boldsymbol{Q} = \boldsymbol{U}$. In summary, with probability $1 - O(1/(np)^{10})$, we have

$$\max_{j \in [p]} \left\|\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}\right\|_2^2 \leq C(K+1)^2(n + \sqrt{n}\log(np)), \tag{2.76}$$

and

$$\max_{j \in [p]} \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})_2^2 \leq C(K+1)^2(r + \sqrt{r}\log(np)). \tag{2.77}$$

Using (2.75), (2.76), (2.77), and Lemmas 2.9.1 and 2.12.9 with $k = r$, we conclude that with probability $1 - O(1/(np)^{10})$,

$$\max_{j \in [p]} \widetilde{\boldsymbol{Z}}_{\cdot j}^k - \widehat{U}_k\widehat{U}_k^T\boldsymbol{X}_{\cdot j}{}_2^2 \leq C(K,\gamma)\left(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r + \sqrt{r}\,\log(np)}{\rho^2}\right) + C\frac{\log(np)}{\rho\, p}. \tag{2.78}$$

Bounding $\widehat{U}_k\widehat{U}_k^T\boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}{}_2^2$. Recalling $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, we obtain $\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{X}_{\cdot j} = \boldsymbol{X}_{\cdot j}$ since $\boldsymbol{U}\boldsymbol{U}^T$ is the projection onto the column space of $\boldsymbol{X}$. Therefore,

$$\begin{aligned}\widehat{U}_k\widehat{U}_k^T\boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}{}_2^2 &= \widehat{U}_k\widehat{U}_k^T\boldsymbol{X}_{\cdot j} - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{X}_{\cdot j}{}_2^2 \\ &\leq \widehat{U}_k\widehat{U}_k^T - \boldsymbol{U}\boldsymbol{U}^T{}_2^2\ \boldsymbol{X}_{\cdot j}{}_2^2.\end{aligned}$$

Using Property 1, note that $\boldsymbol{X}_{\cdot j}{}_2^2 \leq n$. Thus using Lemma 2.9.1 with $k = r$, we have that with probability at least $1 - O(1/(np)^{10})$, we have

$$\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}{}_2^2 \leq C \frac{n(n+p)}{\rho^2 s_r^2}. \tag{2.79}$$

Concluding. From (2.74), (2.78), and (2.79), we claim with probability at least $1 - O(1/(np)^{10})$

$$\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k{}_{2,\infty}^2 \leq C(K,\gamma)\Big(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r + \sqrt{r}\,\log(np)}{\rho^2}\Big) + C\frac{\log(np)}{\rho p}.$$

This completes the proof of Lemma 2.9.2.

■ 2.12.3 Proof of Lemma 2.9.3

To bound $\widehat{s}_k$, we recall Weyl's inequality.

Lemma 2.12.11 (Weyl's inequality). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, let σ_i and $\widehat{\sigma}_i$ be the i-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then for all $i \in [m \wedge n]$,*

$$|\sigma_i - \widehat{\sigma}_i| \leq \left\|\boldsymbol{A} - \boldsymbol{B}\right\|_2.$$

Let $\tilde{s}_k$ be the k-th singular value of $\boldsymbol{Z}$. Then, $\widehat{s}_k = (1/\hat{\rho})\tilde{s}_k$ since it is the k-th singular value of $\widetilde{\boldsymbol{Z}} = (1/\hat{\rho})\boldsymbol{Z}$. By Lemma 5.14.4, we have

$$|\tilde{s}_k - \rho s_k| \leq \boldsymbol{Z} - \rho \boldsymbol{X}_2;$$

recall that s_k is the k-th singular value of $\boldsymbol{X}$. As a result,

$$\begin{aligned}
|\widehat{s}_k - s_k| &= \frac{1}{\hat{\rho}}|\tilde{s}_k - \hat{\rho} s_k| \\
&\leq \frac{1}{\hat{\rho}}|\tilde{s}_k - \rho s_k| + \frac{|\rho - \hat{\rho}|}{\hat{\rho}} s_k \\
&\leq \frac{\boldsymbol{Z} - \rho \boldsymbol{X}_2}{\hat{\rho}} + \frac{|\rho - \hat{\rho}|}{\hat{\rho}} s_k.
\end{aligned}$$

From Lemma 2.12.8 and Lemma 2.12.9, it follows that with probability at least $1 - O(1/(np)^{10})$,

$$|\widehat{s}_k - s_k| \leq \frac{C(K,\gamma)(\sqrt{n}+\sqrt{p})}{\rho} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} s_k.$$

This completes the proof of Lemma 2.9.3.

■ 2.12.4 Proof of Lemma 2.9.4

We need to bound $\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*), \varepsilon\rangle$. To that end, we recall that $\widehat{\beta} = \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y$, $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T$, and $y = \boldsymbol{X}\beta^* + \varepsilon$. Thus,

$$\widetilde{\boldsymbol{Z}}^k \widehat{\beta} = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X}\beta^* + \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon.$$

Therefore,

$$\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*), \varepsilon\rangle = \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X}\beta^*, \varepsilon\rangle + \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon, \varepsilon\rangle - \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{V}}_k^T \beta^*, \varepsilon\rangle. \tag{2.80}$$

Now, ε is independent of $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{\Sigma}}_k, \widehat{\boldsymbol{V}}_k$ since $\widetilde{\boldsymbol{Z}}^k$ is determined by $\boldsymbol{Z}$, which is independent of ε. As a result,

$$\begin{aligned}
\mathbb{E}[\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon, \varepsilon\rangle] &= \mathbb{E}[\varepsilon^T \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon] \\
&= \mathbb{E}[\operatorname{tr}\Big(\varepsilon^T \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \varepsilon\Big)] = \mathbb{E}[\operatorname{tr}\Big(\varepsilon\varepsilon^T \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T\Big)] \\
&= \operatorname{tr}\Big(\mathbb{E}[\varepsilon\varepsilon^T]\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T\Big) \leq C\operatorname{tr}\Big(\sigma^2 \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T\Big) \\
&= C\sigma^2 \widehat{\boldsymbol{U}}_{k}{}_F^2 = C\sigma^2 k.
\end{aligned} \tag{2.81}$$

Therefore, it follows that

$$\mathbb{E}[\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\beta}-\beta^*), \varepsilon\rangle] \leq C\sigma^2 k, \tag{2.82}$$

where we used the fact $\mathbb{E}[\varepsilon] = \boldsymbol{0}$. To obtain a high probability bound, using Lemma 5.14.2 it follows that for any $t > 0$

$$\mathbb{P}\left(\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T \boldsymbol{X}\beta^*, \varepsilon\rangle \geq t\right) \leq \exp\left(-\frac{ct^2}{n\beta^{*2}_{1}\sigma^2}\right) \tag{2.83}$$

due to Assumption 3, and

$$\widehat{U}_k\widehat{U}_k^T\boldsymbol{X}\beta^*{}_2 \leq \boldsymbol{X}\beta^*{}_2 \leq \boldsymbol{X}_{2,\infty}\beta^*{}_1 \leq \sqrt{n}\beta^*{}_1;$$

note that we have used the fact that $\widehat{U}_k\widehat{U}_k^T$ is a projection matrix and $\boldsymbol{X}_{2,\infty} \leq \sqrt{n}$ due to Assumption 1. Similarly, for any $t > 0$

$$\mathbb{P}\left(\langle \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{V}}_k^T\beta^*, \varepsilon\rangle \geq t\right) \leq \exp\left(-\frac{ct^2}{\sigma^2(n+\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}_{2,\infty}^2)\beta^{*2}_1}\right), \tag{2.84}$$

due to Assumption 3, and

$$\begin{aligned}\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{V}}_k^T\beta^*{}_2 = (\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\beta^* + \boldsymbol{X}\beta^*{}_2 \ &\leq (\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\beta^*{}_2 + \boldsymbol{X}\beta^*{}_2 \\ &\leq (\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}_{2,\infty} + \boldsymbol{X}_{2,\infty})\beta^*{}_1.\end{aligned}$$

Finally, using Lemma 5.14.3 and (3.61), it follows that for any $t > 0$

$$\mathbb{P}\left(\langle \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T\varepsilon, \varepsilon\rangle \geq \sigma^2 k + t\right) \leq \exp\left(-c\min\left(\frac{t^2}{k\sigma^4}, \frac{t}{\sigma^2}\right)\right), \tag{2.85}$$

since $\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^T$ is a projection matrix and by Assumption 3.

From (3.57), (3.62), (3.64), and (3.66), we conclude that with probability at least $1 - O(1/(np)^{10})$,

$$\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\beta} - \beta^*), \varepsilon\rangle \leq \sigma^2 k + C\sigma\sqrt{\log(np)}(\sigma\sqrt{k} + \sigma\sqrt{\log(np)} + \beta^*{}_1(\sqrt{n} + \widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}_{2,\infty})).$$

This completes the proof of Lemma 2.9.4.

■ 2.12.5 Proof of Lemma 2.12.5

As mentioned earlier, the proof presented here is a natural extension of that for Theorem 4.6.1 in Vershynin (2018) for the non-isotropic setting. Recall that

$$\|\boldsymbol{A}\| = \max_{x \in S^{p-1}, y \in S^{n-1}} \langle \boldsymbol{A}x, y\rangle,$$

where S^{p-1}, S^{n-1} denote the unit spheres in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively. We start by bounding the quadratic term $\langle \boldsymbol{A}x, y\rangle$ for a finite set x, y obtained by placing 1/4-net on the unit spheres, and then use the bound on them to bound $\langle \boldsymbol{A}x, y\rangle$ for all x, y over the spheres.

Step 1: Approximation. We will use Corollary 4.2.13 of Vershynin (2018) to establish a 1/4-net of $\mathcal{N}$ of the unit sphere S^{p-1} with cardinality $|\mathcal{N}| \leq 9^p$. Applying Lemma 4.4.1 of Vershynin (2018), we obtain

$$\frac{1}{n}A^TA - \mathbf{\Sigma}_2 \leq 2\max_{x\in\mathcal{N}}\left|\langle(\frac{1}{n}A^TA - \mathbf{\Sigma})x, x\rangle\right| = 2\max_{x\in\mathcal{N}}\left|\frac{1}{n}Ax_2^2 - x^T\mathbf{\Sigma}x\right|.$$

To achieve our desired result, it remains to show that

$$\max_{x\in\mathcal{N}}\left|\frac{1}{n}Ax_2^2 - x^T\mathbf{\Sigma}x\right| \leq \frac{\epsilon}{2},$$

where $\epsilon = K^2\max(\delta, \delta^2)$.

Step 2: Concentration. Let us fix a unit vector $x \in S^{p-1}$ and write

$$\|Ax\|_2^2 - x^T\mathbf{\Sigma}x = \sum_{i=1}^n\left(\langle A_i, x\rangle^2 - \mathbb{E}[\langle A_i, x\rangle^2]\right) =: \sum_{i=1}^n\left(Y_i^2 - \mathbb{E}[Y_i^2]\right).$$

Since the rows of A are assumed to be independent sub-gaussian random vectors with $A_{i\psi_2} \leq K$, it follows that $Y_i = \langle A_i, x\rangle$ are independent sub-gaussian random variables with $Y_{i\psi_2} \leq K$. Therefore, $Y_i^2 - \mathbb{E}[Y_i^2]$ are independent, mean zero, sub-exponential random variables with

$$Y_i^2 - \mathbb{E}[Y_i^2]_{\psi_1} \leq C Y_i^2{}_{\psi_1} \leq C Y_{i\psi_2}^2 \leq CK^2.$$

As a result, we can apply Bernstein's inequality (see Theorem 5.14.1) to obtain

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\|Ax\|_2^2 - x^T\mathbf{\Sigma}x\right| \geq \frac{\epsilon}{2}\right) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n(Y_i^2 - \mathbb{E}[Y_i^2])\right| \geq \frac{\epsilon}{2}\right)\\ &\leq 2\exp\left(-c\min\left(\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2}\right)n\right)\\ &= 2\exp\left(-c\delta^2 n\right)\\ &\leq 2\exp\left(-cC^2(p+t^2)\right),\end{aligned}$$

where the last inequality follows from the definition of δ in (2.72) and because $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$.

Step 3: Union bound. We now apply a union bound over all elements in the net. Specifically,

$$\mathbb{P}\left(\max_{x\in\mathcal{N}}\left|\frac{1}{n}\left\|Ax\right\|_2^2 - x^T\mathbf{\Sigma}x\right| \geq \frac{\epsilon}{2}\right) \leq 9^p \cdot 2\exp\left(-cC^2(p+t^2)\right) \leq 2\exp\left(-t^2\right),$$

for large enough C. This concludes the proof.

■ 2.12.6 Proof of Lemma 2.12.6

Recall that $z_i = (x_i + w_i) \circ \pi_i$, where w_i is an independent mean zero subgaussian vector with $w_{i\psi_2} \leq K$ and π_i is a vector of independent Bernoulli variables with parameter ρ. Hence, $\mathbb{E}[z_i - \rho x_i] = \mathbf{0}$ and is independent across $i \in [n]$. The only remaining item is a bound on $z_i - \rho x_{i\psi_2}$. To that end, note that

$$\begin{aligned} z_i - \rho x_{i\psi_2} &= x_i \circ \pi_i + w_i \circ \pi_i - \rho x_{i\psi_2} \\ &\leq x_i \circ (\rho\mathbf{1} - \pi_i)_{\psi_2} + w_i \circ \pi_{i\psi_2}. \end{aligned}$$

Now, $(\rho\mathbf{1} - \pi_i)$ is independent, zero mean random vector whose absolute value is bounded by 1, and is component-wise multiplied by x_i which are bounded in absolute value by 1 as per Assumption 1. That is, $x_i \circ (\rho\mathbf{1} - \pi_i)$ is a zero mean random vector where each component is independent and bounded in absolute value by 1. That is, $\cdot_{\psi_2} \leq C$.

For $w_i \circ \pi_i$, note that w_i and π_i are independent vectors and the coordinates of π_i have support $\{0, 1\}$. Therefore, from Lemma 2.12.12, it follows that $w_i \circ \pi_{i\psi_2} \leq w_{i\psi_2} \leq K$ by Assumption 4. The proof of Lemma 2.12.6 is complete by choosing a large enough C.

Lemma 2.12.12. *Suppose that $Y \in \mathbb{R}^n$ and $P \in \{0, 1\}^n$ are independent random vectors. Then,*

$$Y \circ P_{\psi_2} \leq Y_{\psi_2}.$$

Proof. Given a binary vector $P \in \{0, 1\}^n$, let $I_P = \{i \in [n] : P_i = 1\}$. Observe that

$$Y \circ P = \sum_{i \in I_P} e_i \otimes e_i Y.$$

Here, $\circ$ denotes the Hadamard product (entry-wise product) of two matrices. By definition of the ψ_2-norm,

$$Y_{\psi_2} = \sup_{u \in \mathbb{S}^{n-1}} u^T Y_{\psi_2} = \sup_{u \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_Y[\exp\Big(|u^T Y|^2/t^2\Big)] \le 2\}.$$

Let $u_0 \in \mathbb{S}^{n-1}$ denote the maximum-achieving unit vector (such a u_0 exists because $\inf\{\cdots\}$ is continuous with respect to u and $\mathbb{S}^{n-1}$ is compact). Now,

$$\begin{aligned}
Y \circ P_{\psi_2} &= \sup_{u \in \mathbb{S}^{n-1}} u^T Y \circ P_{\psi_2} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{Y,P}[\exp\Big(|u^T Y \circ P|^2/t^2\Big)] \le 2\} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_P[\mathbb{E}_Y[\exp\Big(|u^T Y \circ P|^2/t^2\Big) \mid P]] \le 2\} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_P[\mathbb{E}_Y[\exp(|u^T \sum_{i \in I_P} e_i \otimes e_i Y|^2/t^2) \mid P]] \le 2\} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_P[\mathbb{E}_Y[\exp(|(\sum_{i \in I_P} e_i \otimes e_i u)^T Y|^2/t^2) \mid P]] \le 2\}.
\end{aligned}$$

For any $u \in \mathbb{S}^{n-1}$, observe that

$$\mathbb{E}_Y[\exp(|(\sum_{i \in I_P} e_i \otimes e_i u)^T Y|^2/t^2) \mid P] \le \mathbb{E}_Y[\exp\Big(|u_0^T Y|^2/t^2\Big)].$$

Therefore, taking supremum over $u \in \mathbb{S}^{n-1}$, we obtain

$$Y \circ P_{\psi_2} \le Y_{\psi_2}.$$

■

■ 2.12.7 Proof of Lemma 2.12.7

Consider

$$\begin{aligned}
\mathbb{E}[\tilde{W}^T \tilde{W}] &= \sum_{i=1}^{n} \mathbb{E}[(z_i - \rho x_i) \otimes (z_i - \rho x_i)] \\
&= \sum_{i=1}^{n} \mathbb{E}[z_i \otimes z_i] - \rho^2 (x_i \otimes x_i)
\end{aligned}$$

$$= \sum_{i=1}^{n} (\rho - \rho^2)\mathrm{diag}(x_i \otimes x_i) + (\rho - \rho^2)\mathrm{diag}(\mathbb{E}[w_i \otimes w_i]) + \rho^2 \mathbb{E}[w_i \otimes w_i].$$

Note that $\mathrm{diag}(\boldsymbol{X}^T \boldsymbol{X})_2 \leq n$ due to Assumption 1. Using Assumption 4, it follows that $\mathrm{diag}(\mathbb{E}[w_i \otimes w_i])_2 \leq CK^2$. By Assumption 4, we have $\mathbb{E}[w_i \otimes w_i]_2 \leq \gamma^2$. Therefore,

$$\mathbb{E}[\tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{W}}]_2 \leq Cn(\rho - \rho^2)(K+1)^2 + n\rho^2\gamma^2.$$

This completes the proof of Lemma 2.12.7.

2.12.8 Proof of Lemma 2.12.9

By the Binomial Chernoff bound, for $\alpha > 1$,

$$\mathbb{P}\left(\widehat{\rho} > \alpha\rho\right) \leq \exp\left(-\frac{(\alpha-1)^2}{\alpha+1} np\rho\right) \quad \text{and} \quad \mathbb{P}\left(\widehat{\rho} < \rho/\alpha\right) \leq \exp\left(-\frac{(\alpha-1)^2}{2\alpha^2} np\rho\right).$$

By the union bound,

$$\mathbb{P}\left(\rho/\alpha \leq \widehat{\rho} \leq \alpha\rho\right) \geq 1 - \mathbb{P}\left(\widehat{\rho} > \alpha\rho\right) - \mathbb{P}\left(\widehat{\rho} < \rho/\alpha\right).$$

Noticing $\alpha + 1 < 2\alpha < 2\alpha^2$ for all $\alpha > 1$, we obtain the desired bound claimed in Lemma 2.12.9. To complete the remaining claim of Lemma 2.12.9, we consider an α that satisfies

$$(\alpha - 1)^2 \leq C\frac{\log(np)}{\rho np},$$

for a constant $C > 0$. Thus,

$$1 - C\frac{\sqrt{\log(np)}}{\sqrt{\rho np}} \leq \alpha \leq 1 + C\frac{\sqrt{\log(np)}}{\sqrt{\rho np}}.$$

Then, with $\rho \geq c\frac{\log^2 np}{np}$, we have that $\alpha \leq 2$. Further by choosing $C > 0$ large enough, we have

$$\frac{(\rho - \hat{\rho})^2}{\hat{\rho}^2} \leq C\frac{\log(np)}{\rho np}.$$

holds with probability at least $1 - O(1/(np)^{10})$. This completes the proof of Lemma 2.12.9.

■ 2.12.9 Proof of Lemma 2.12.10

By definition $\boldsymbol{Q}\boldsymbol{Q}^T \in \mathbb{R}^{n\times n}$ is a rank ℓ matrix. Since $\boldsymbol{Q}$ has orthonormal column vectors, the projection operator has $\boldsymbol{Q}\boldsymbol{Q}^T{}_2 = 1$ and $\boldsymbol{Q}\boldsymbol{Q}^T{}^2_F = \ell$. For a given $j \in [p]$, the random vector $\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}$ is such that it has zero mean, independent components that are sub-gaussian by Assumption 4. For any $i \in [n], j \in [p]$, we have by property of ψ_2 norm, $z_{ij} - \rho x_{ij}{}_{\psi_2} \leq z_i - \rho x_i{}_{\psi_2}$ which is bounded by $C(K+1)$ using Lemma 2.12.6. Recall the Hanson-Wright inequality (Vershynin (2018)):

Theorem 2.12.3 (Hanson-Wright inequality). *Let $\zeta \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then for any $t > 0$,*

$$\mathbb{P}\left(\left|\zeta^T \boldsymbol{A}\zeta - \mathbb{E}[\zeta^T \boldsymbol{A}\zeta]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{L^4\left\|\boldsymbol{A}\right\|_F^2}, \frac{t}{L^2\left\|\boldsymbol{A}\right\|_2}\right)\right),$$

where $L = \max_{i\in[n]}\left\|\zeta_i\right\|_{\psi_2}$.

Now with $\zeta = \boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}$ and the fact that $\boldsymbol{Q}^T\boldsymbol{Q} = \mathcal{D} \in \mathbb{R}^{\ell\times\ell}$, $\boldsymbol{Q}\boldsymbol{Q}^T\zeta^2_2 = \zeta^T\boldsymbol{Q}\boldsymbol{Q}^T\zeta$. Therefore, by Theorem 2.12.3, for any $t > 0$,

$$\boldsymbol{Q}\boldsymbol{Q}^T\zeta_2^2 \leq \mathbb{E}[\zeta^T\boldsymbol{Q}\boldsymbol{Q}^T\zeta] + t,$$

with probability at least $1 - \exp\left(-c\min\left(\frac{t}{C(K+1)^2}, \frac{t^2}{C(K+1)^4\ell}\right)\right)$. Now,

$$\begin{aligned}\mathbb{E}[\zeta^T\boldsymbol{Q}\boldsymbol{Q}^T\zeta] &= \sum_{m=1}^{\ell}\mathbb{E}[(\boldsymbol{Q}_{\cdot m}^T\zeta)^2]\\ &\overset{(a)}{=} \sum_{m=1}^{\ell}\mathrm{Var}(\boldsymbol{Q}_{\cdot m}^T\zeta)\\ &\overset{(b)}{=} \sum_{m=1}^{\ell}\sum_{i=1}^{n}\boldsymbol{Q}_{im}^2\mathrm{Var}(\zeta_i)\\ &\overset{(c)}{\leq} C(K+1)^2\ell,\end{aligned}$$

where $\zeta = \boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}$, and hence (a) follows from $\mathbb{E}[\zeta] = \mathbb{E}[\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}] = \boldsymbol{0}$, (b) follows from ζ having independent components and (c) follows from each component of ζ having

ψ_2-norm bounded by $C(K+1)$. Therefore, it follows by union bound that for any $t > 0$,

$$\mathbb{P}\Big(\max_{j \in [p]} \left\| \boldsymbol{Q}\boldsymbol{Q}^T (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}) \right\|_2^2 \geq \ell C (K+1)^2 + t \Big)$$
$$\leq p \cdot \exp\left(-c \min\left(\frac{t^2}{C(K+1)^4 \ell}, \frac{t}{C(K+1)^2} \right) \right).$$

This completes the proof of Lemma 2.12.10.

■ 2.13 Helpful Concentration Inequalities

In this section, we state and prove a number of helpful concentration inequalities used to establish our primary results.

Lemma 2.13.1. *Let X be a mean zero, sub-gaussian random variable. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E} \exp(\lambda X) \leq \exp\left(C\lambda^2 \|X\|_{\psi_2}^2 \right).$$

Lemma 2.13.2. *Let $X_1, \ldots, X_n$ be independent, mean zero, sub-gaussian random variables. Then,*

$$\sum_{i=1}^n X_{i\psi_2}^2 \leq C \sum_{i=1}^n \|X_i\|_{\psi_2}^2 .$$

Theorem 2.13.1 (Bernstein's inequality). *Let $X_1, \ldots, X_n$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right| \geq t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}} \right) \right),$$

where $c > 0$ is an absolute constant.

Lemma 2.13.3 (Modified Hoeffding Inequality). *Let $X \in \mathbb{R}^n$ be random vector with independent mean-zero sub-Gaussian random coordinates with $X_{i\psi_2} \leq K$. Let $a \in \mathbb{R}^n$ be another random vector that satisfies $a_2 \leq b$ almost surely for some constant $b \geq 0$. Then for all $t \geq 0$,*

$$\mathbb{P}\left(\left| \sum_{i=1}^n a_i X_i \right| \geq t \right) \leq 2 \exp\left(-\frac{ct^2}{K^2 b^2} \right),$$

where $c > 0$ is a universal constant.

Proof. Let $S_n = \sum_{i=1}^n a_i X_i$. Then applying Markov's inequality for any $\lambda > 0$, we obtain

$$\begin{aligned} \mathbb{P}\left(S_n \geq t\right) &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda t)\right) \\ &\leq \mathbb{E}\left[\exp(\lambda S_n)\right] \cdot \exp(-\lambda t) \\ &= \mathbb{E}_a\left[\mathbb{E}\left[\exp(\lambda S_n) \mid a\right]\right] \cdot \exp(-\lambda t). \end{aligned}$$

Now, conditioned on the random vector a, observe that

$$\mathbb{E}\left[\exp(\lambda S_n)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\Big(CK^2\lambda^2 a_2^2\Big) \leq \exp\Big(CK^2\lambda^2 b^2\Big),$$

where the equality follows from conditional independence, the first inequality by Lemma 2.13.1, and the final inequality by assumption. Therefore,

$$\mathbb{P}\left(S_n \geq t\right) \leq \exp\Big(CK^2\lambda^2 b^2 - \lambda t\Big).$$

Optimizing over λ yields the desired result:

$$\mathbb{P}\left(S_n \geq t\right) \leq \exp\left(-\frac{ct^2}{K^2 b^2}\right).$$

Applying the same arguments for $-\langle X, a\rangle$ gives a tail bound in the other direction. ■

Lemma 2.13.4 (Modified Hanson-Wright Inequality). *Let $X \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with $X_{i\psi_2} \leq K$. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a random matrix satisfying $\boldsymbol{A}_2 \leq a$ and $\boldsymbol{A}_F^2 \leq b$ almost surely for some $a, b \geq 0$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\left|X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X]\right| \geq t\right) \leq 2 \cdot \exp\left(-c \min\left(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\right)\right).$$

Proof. The proof follows similarly to that of Theorem 6.2.1 of Vershynin (2018). Using the independence of the coordinates of X, we have the following useful diagonal and off-diagonal decomposition:

$$X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X] = \sum_{i=1}^n \Big(A_{ii} X_i^2 - \mathbb{E}[A_{ii} X_i^2]\Big) + \sum_{i \neq j} A_{ij} X_i X_j.$$

Therefore, letting

$$p = \mathbb{P}\left(X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X] \geq t\right),$$

we can express

$$p \leq \mathbb{P}\Big(\sum_{i=1}^{n}\left(A_{ii}X_i^2 - \mathbb{E}[A_{ii}X_i^2]\right) \geq t/2\Big) + \mathbb{P}\Big(\sum_{i \neq j} A_{ij}X_iX_j \geq t/2\Big) =: p_1 + p_2.$$

We will now proceed to bound each term independently.

Step 1: diagonal sum. Let $S_n = \sum_{i=1}^{n}(A_{ii}X_i^2 - \mathbb{E}[A_{ii}X_i^2])$. Applying Markov's inequality for any $\lambda > 0$, we have

$$\begin{aligned} p_1 &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda t/2)\right) \\ &\leq \mathbb{E}_{\boldsymbol{A}}\mathbb{E}\left[[\exp(\lambda S_n) \mid \boldsymbol{A}]\right] \cdot \exp(-\lambda t/2). \end{aligned}$$

Since the X_i are independent, sub-Gaussian random variables, $X_i^2 - \mathbb{E}[X_i^2]$ are independent mean-zero sub-exponential random variables, satisfying

$$\left\|X_i^2 - \mathbb{E}[X_i^2]\right\|_{\psi_1} \leq C\left\|X_i^2\right\|_{\psi_1} \leq C\left\|X_i\right\|_{\psi_2}^2 \leq CK^2.$$

Conditioned on $\boldsymbol{A}$ and optimizing over λ using standard arguments, yields

$$p_1 \leq \exp\left(-c\min\left(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\right)\right).$$

Step 2: off-diagonals. Let $S = \sum_{i \neq j} A_{ij}X_iX_j$. Again, applying Markov's inequality for any $\lambda > 0$, we have

$$p_2 = \mathbb{P}\left(\exp(\lambda S) \geq \exp(\lambda t/2)\right) \leq \mathbb{E}_{\boldsymbol{A}}\left[\mathbb{E}\left[\exp(\lambda S) \mid \boldsymbol{A}\right]\right] \cdot \exp(-\lambda t/2).$$

Let g be a standard multivariate gaussian random vector. Further, let X' and g' be independent copies of X and g, respectively. Conditioning on $\boldsymbol{A}$ yields

$$\begin{aligned} \mathbb{E}\left[\exp(\lambda S)\right] &\leq \mathbb{E}\left[\exp\left(4\lambda X^T \boldsymbol{A} X'\right)\right] && \text{(by Decoupling Remark 6.1.3 of Vershynin (2018))} \\ &\leq \mathbb{E}\left[\exp\left(C_1\lambda g^T \boldsymbol{A} g'\right)\right] && \text{(by Lemma 6.2.3 of Vershynin (2018))} \\ &\leq \exp\left(C_2\lambda^2\left\|\boldsymbol{A}\right\|_F^2\right) && \text{(by Lemma 6.2.2 of Vershynin (2018))} \end{aligned}$$

$$\leq \exp\Big(C_2\lambda^2 b\Big),$$

where $|\lambda| \leq c/a$. Optimizing over λ then gives

$$p_2 \leq \exp\left(-c\min\left(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\right)\right).$$

Step 3: combining. Putting everything together completes the proof. ■

Chapter 3

Synthetic Interventions

■ 3.1 Introduction

There is a growing interest in personalized decision-making, where the goal is to select an optimal intervention for each unit from a collection of interventions. For example in e-commerce, a common goal of an online platform is to determine which discount level is best suited (e.g., to increase engagement levels) for each customer sub-population. Such customer sub-populations are usually created based on customer demographics and their prior interactions with the platform, and the clustering is done to ensure that customers within a sub-population behave similarly, while allowing for significant heterogeneity across sub-populations. At its core, estimating the best "personalized" intervention requires evaluating the impact of each of the $D \geq 1$ discount levels on each of the $N \geq 1$ customer sub-populations. One way to estimate the best personalized intervention is to conduct $N \times D$ randomized experiments corresponding to each combination. In particular, for each sub-population and discount level, some number of randomly chosen customers from that sub-population are assigned that discount level and their corresponding outcomes are measured. Conducting these $N \times D$ personalized experiments would allow one to estimate all $N \times D$ causal parameters. However, this can become very expensive, if not infeasible, as D and N grow. This motivates the need for "data-efficient" experiment design where $N \times D$ causal parameters can be inferred from far fewer experiments.

Similar questions arise in program evaluation, where one may want to design a governmental policy that is particularly suited to the socio-economic realities of a geographic location. For example, in the context of COVID-19, policy-makers from N countries might be interested in evaluating which of D mobility restricting policies is best suited for their country. There is an additional challenge in program evaluation compared to the experimental setting as only observational data is readily available. In particular, each country can only undergo at most one of the D policies. Further, the policy implemented in any given country will likely be confounded with the characteristics of that country (e.g., the government structure, population density and demographics, cultural leanings), which itself may impact the outcomes observed under the policy. Further, these confounding variables might not be fully known or observed.

In summary, our interest is in evaluating the $N \times D$ causal parameters associated with N units and D interventions based on potentially confounded, limited observations that do not scale with $N \times D$.

Special instances of this question have a rich literature within econometrics and beyond. In particular, the sub-problem of estimating what would have happened to a "treated" unit (i.e., undergoes an intervention) under "control" (i.e., absence on an intervention) has been studied through the prominent frameworks of differences-in-differences (DID) Ashenfelter and Card (1984); Bertrand et al. (2004); Angrist and Pischke (2009) and synthetic controls (SC) Abadie et al. (2010); Abadie and Gardeazabal (2003). These frameworks live within the framework of panel data settings, where one gets repeated measurements of units. By considering "control" as one of the D interventions, we see that DID and SC both allow recovery of at most N causal parameters, namely the counterfactual of what would have been to a treated unit had it remained under control. However, towards the broader goal of personalized decision-making, one needs to answer counterfactual questions beyond just what would have happened under control, as discussed in the e-commerce and COVID-19 examples above. Indeed, extending the SC framework to estimate all $N \times D$ causal parameters has been posed as an important open question in Abadie (2020). A goal of this work is to address this open question.

■ 3.1.1 Overview of Synthetic Interventions Framework

Consider a setting with $N \geq 1$ units and $D \geq 1$ interventions. For each unit and intervention pair, there are $T \geq 1$ outcomes/measurements of interest. Unless stated otherwise, we index units with $n \in [N] := \{1, \ldots, N\}$, outcomes with $t \in [T]$, and interventions with $d \in [D]_0$.[1] Following the causal framework of Neyman (1923) and Rubin (1974), we denote $Y_{tn}^{(d)} \in \mathbb{R}$ as the potential outcome for unit n and measurement t under intervention d. We encode the set of all potential outcomes $\{Y_{tn}^{(d)}\}$ into an order-3

[1] Let $[X]_0 = \{0, 1, \ldots, X-1\}$ and $[X] = \{1, \ldots, X\}$ for any positive integer X.

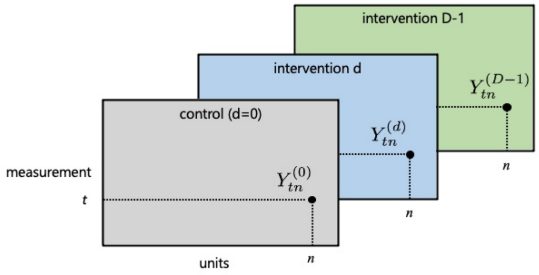


Figure 3.1: Potential outcomes tensor.

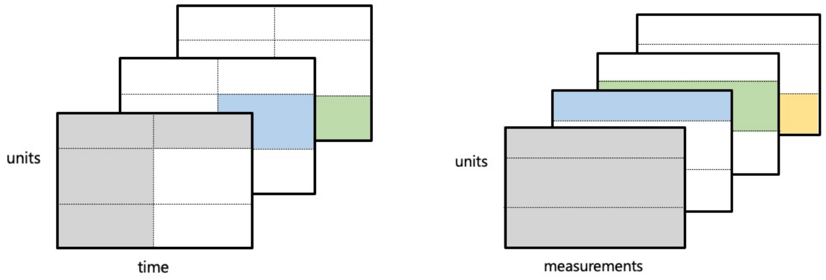
tensor whose dimensions correspond to units, measurements, and interventions. See Figure 3.1 for a visualization of this potential outcomes tensor.

We describe our observations through $\boldsymbol{Y} = [Y_{tnd}] \in \{\mathbb{R} \cup \star\}^{T \times N \times D}$, where $\star$ indicates a missing entry. We assume $\boldsymbol{Y}$ obeys the following sparsity pattern.

Assumption 8 (Observation pattern, SUTVA). *We observe the same $T_0 \leq T$ outcomes for all units under the same intervention. Without loss of generality, let this intervention be $d = 0$, and let the indices corresponding to these T_0 measurements be $\mathcal{T}_{\text{pre}} := \{1, \dots, T_0\}$. That is, we observe $Y_{tn0} = Y_{tn}^{(0)}$ for all $n \in [N]$ and $t \in [T_0]$. Further, for every intervention d, there is a non-empty subset of units, $\mathcal{I}^{(d)} \subset [N]$, for which we observe $T_1 \leq T$ measurements. Without loss of generality, we assume the indices corresponding to these T_1 measurements are $\mathcal{T}_{\text{post}} := \{T - T_1 + 1, \dots, T\}$. That is, we observe $Y_{tnd} = Y_{tn}^{(d)}$ for $d \in [D]_0$, $n \in \mathcal{I}^{(d)}$, and $t \in \mathcal{T}_{\text{post}}$. For all other entries of $\boldsymbol{Y}$, we assume $Y_{tnd} = \star$.*

Observation pattern. As per Assumption 8, we observe all N units under $d = 0$ for $t \in \mathcal{T}_{\text{pre}}$. For $d > 0$ and $t \in \mathcal{T}_{\text{post}}$, however, it is sufficient to only observe outcomes for $N_d = |\mathcal{I}^{(d)}| \geq 1$ units, provided N_d is sufficiently large. Below, we connect our observation pattern with what is typically assumed in the SC literature. In particular, as with our setting, SC assumes that all N units are observed under $d = 0$ for $t \in \mathcal{T}_{\text{pre}}$. Since $d = 0$ often represents control and t typically indexes the t^{th} time step, this time frame is often referred to as the "pre-intervention" period. During the "post-intervention" period, i.e., $t \in \{T_0 + 1, \dots, T\}$ with $T_1 = T - T_0$, each unit is then observed under one of the D interventions. See Figure 3.2a for a visual depiction of the typical observation pattern in the SC literature. Indeed, we use $\mathcal{T}_{\text{pre}}$ and $\mathcal{T}_{\text{post}}$ to be in line with the SC literature.

However, our setup allows for more general observation patterns compared to SC as we do not strictly need a separate pre- and post-intervention period. This is particularly



(a) Panel data setting.

(b) Proposed data-efficient RCT.

Figure 3.2: Observations are represented by colored blocks while unobservable counterfactuals are represented by white blocks. We remark that the tensors in 3.2a and 3.2b have been transposed to better align with the standard panel data setups.

relevant in multi-arm randomized control trials (RCTs), like in the e-commerce setting described earlier, where there may not be a pre-intervention period. Further, in many such RCTs setting, units can be simultaneously observed under multiple interventions; for example, e-commerce platforms can randomly choose different individuals from the same customer sub-populations to undergo different discount levels (here, a unit is defined as a customer sub-population). Hence, the sets $\mathcal{I}^{(d_1)}$ and $\mathcal{I}^{(d_2)}$ do not have to be mutually exclusive. However, despite being able to simultaneously observe units under multiple interventions, recall our goal is to learn all $N \times D$ causal parameters using far fewer number of experiments. We show this is possible if we observe all units under the same intervention for some number of measurements, as formalized in Assumption 8. See Figure 3.2b for a visual depiction of the observation pattern in our proposed data-efficient RCT.

Target causal parameter. The goal is to estimate $Y_{tn}^{(d)}$ for all $n \in [N]$ and $t \in \mathcal{T}_{\text{post}}$ for all $d \in [D]$. Specifically, we are interested in the estimation of the causal parameter

$$\theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}[Y_{tn}^{(d)}].$$

That is, the expected potential outcome of unit n under intervention d, averaged over the measurements $t \in \mathcal{T}_{\text{post}}$. In the setting of SC, this would correspond to estimating the average potential outcome of each unit under each intervention during the post-intervention period. We emphasize that in observational settings, the potential outcomes $Y_{tn}^{(d)}$ might be correlated with which entries of $\boldsymbol{Y}$ are observed, i.e., there is confounding. This serves as an additional challenge in defining the appropriate causal framework to estimate $\theta_n^{(d)}$.

Synthetic interventions (SI) estimator: an overview. The SI estimator recovers a given $\theta_n^{(d)}$ with valid confidence intervals via a three-step procedure, where each step has a simple closed form expression. Below, we give an overview of the method. For a given (n, d) pair, the first step is to estimate a linear model, $w^{(n,d)} \in \mathbb{R}^{N_d}$ such that for all $t \in \mathcal{T}_{\text{pre}}$,

$$Y_{tn0} \approx \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} Y_{tj0}.$$

Specifically, we use principal component regression (PCR) to learn $w^{(n,d)}$ by linearly regressing $\{Y_{tn0} : t \in \mathcal{T}_{\text{pre}}\}$ on $\{Y_{tj0} : t \in \mathcal{T}_{\text{pre}}, j \in \mathcal{I}^{(d)}\}$. Subsequently $\theta_n^{(d)}$ is estimated as

$$\widehat{\mathbb{E}}[Y_{tn}^{(d)}] = \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} Y_{tjd}, \quad \text{for } t \in \mathcal{T}_{\text{post}}, \tag{3.1}$$

$$\widehat{\theta}_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \widehat{\mathbb{E}}[Y_{tn}^{(d)}]. \tag{3.2}$$

The precise details of the SI estimator are in Section 3.3. Though the SI method produces an estimate of the entire trajectory $\widehat{\mathbb{E}}[Y_{tn}^{(d)}]$ for $t \in \mathcal{T}_{\text{post}}$ as defined in (3.1), our theoretical results are focused on proving consistency and asymptotic normality for $\widehat{\theta}_n^{(d)}$. Further, the SI estimator comes with a data-driven hypothesis test to verify when one can accurately learn a model $w^{(n,d)}$ using outcomes under $d = 0$ and $t \in \mathcal{T}_{\text{pre}}$, and then transfer it to estimate counterfactual outcomes for measurements $\mathcal{T}_{\text{post}}$ and $d \in [D]_0$.

Challenge in estimating beyond control: causal transportability with panel data. For ease of discussion, we restrict our attention here to the setting of SC where we have a pre- and post-intervention period. Identification and estimation of $\theta_n^{(d)}$ across all (n, d) complements the existing SC literature, which focuses on estimating $\{\theta_n^{(0)} : n \notin \mathcal{I}^{(0)}\}$, i.e., the counterfactual average potential outcome under control for a treated unit, otherwise referred to as the “treatment effect on a treated unit”. However, the SI framework allows one to estimate causal parameters of the form $\theta_n^{(d)}$ for $d \neq 0$ and $n \in \mathcal{I}^{(0)}$, this corresponds to the “treatment effect for an untreated unit”; estimating the best personalized treatment for a unit that is thus far untreated is clearly of interest, but is not identifiable nor estimable in the SC framework. We discuss the fundamental challenge in estimating $\theta_n^{(d)}$ for $d \neq 0$ next.

Indeed, if we restrict ourselves to the task of estimating $\{\theta_n^{(0)} : n \notin \mathcal{I}^{(0)}\}$, we see that the

SI estimator in essence reduces to the standard SC estimator.[2] A reasonable question that may arise is whether $\theta_n^{(d)}$ across (n, d) can be estimated by simply fitting $N \times D$ separate SC estimators, one for each pair. *This is not possible*. In SC, $w^{(n,0)}$ is learned using outcomes data under control and then is only applied on outcomes under control. However, to estimate $\theta_n^{(d)}$ for $d \neq 0$, this requires learning a model using outcomes under control, but then applying it to outcomes under intervention $d \neq 0$—recall (3.2). This begs the pivotal question: "when does the structure between units under control continue to hold in other intervention regimes"? The SC framework does not have an answer for this question thus far and it has indeed been posed as an important open question in Abadie (2020). As we will see, a tensor factor model across time, units, *and interventions* provides one natural answer to this question.

We highlight that the problem of when one can learn a model under one interventional regime and transfer it to another has gained interest across many fields and has been known by terms including: "causal transportability", "transfer learning", "learning with distribution shift". We believe the SI framework provides one answer to this problem for the panel data setting, and more broadly as well.

■ 3.1.2 Related Works

There are two key conceptual frameworks that are foundational to this work: SC and latent factor models. Given the sizeable literature, we provide a brief overview of closely related works from each topic.

As noted earlier, when restricted to $d = 0$, i.e., estimating $\theta_n^{(0)}$ for a subset of units $n \notin \mathcal{I}^{(0)}$, the SI method is effectively SC as introduced in the seminal works of Abadie et al. (2010); Abadie and Gardeazabal (2003), but with a different form of regularization. Precisely, for this restricted setting, SI is identical to the "robust synthetic controls" (RSC) method of Amjad et al. (2018) and further analyzed in Agarwal et al. (2021e). RSC is a variant of SC, which de-noises observations and imputes missing values via matrix completion, e.g., by using singular value thresholding. Other variants of SC have been considered, including in Hsiao et al. (2012); Doudchenko and Imbens (2016b); Athey et al. (2021); Li and Bell (2017); Xu (2017b); Amjad et al. (2018, 2019); Li (2018); Arkhangelsky et al. (2020); Bai and Ng (2020); Ben-Michael et al. (2020); Chan and Kwok (2020);

[2] The key difference being that rather than restricting $w^{(n,0)}$ to be convex as is classically done, we learn this linear model using PCR. We motivate the suitability of PCR in Section 3.3.

Chernozhukov et al. (2020b); Fernández-Val et al. (2020). In Section 3.8, we provide a detailed comparison with closely related works; in particular, we focus on the recent works of Arkhangelsky et al. (2020); Bai and Ng (2020); Chernozhukov et al. (2020b); Agarwal et al. (2021e) as they highlight some of the primary points of comparison of our work with the SC literature.

A critical aspect that enables SC is the structure between units and time under control ($d = 0$). One elegant encoding of this structure is through a (matrix) factor model (also known as an interactive fixed effect model), Chamberlain (1984); Liang and Zeger (1986); Arellano and Honore (2000); Bai (2003, 2009); Pesaran (2006); Moon and Weidner (2015, 2017). Recent works have connected factor models with low-rank matrices, which are prevalent within the matrix completion (MC) literature, cf. Candès and Tao (2010); Recht (2011); Chatterjee (2015). As such, several works of late, including Athey et al. (2021); Amjad et al. (2018, 2019); Agarwal et al. (2021e); Fernández-Val et al. (2020); Arkhangelsky et al. (2020), have developed estimators that are guided by MC principles to estimate and analyze causal parameters related to $\theta_n^{(0)}$, by directly learning from the observed outcomes rather than relying on additional covariates.

■ 3.1.3 Contributions & Organization of Paper

Section 3.3: Methodological. We propose the SI estimator, which recovers all $N \times D$ causal parameters $\theta_n^{(d)}$, with valid confidence intervals, under the observation pattern described in Assumption 8. For the special case where $N_d = N/D$, SI estimates the $N \times D$ parameters using $T_0 \times N + T_1 \times N$ observations, which is at most $2 \times T \times N$, independent of D.

Section 4.6.3: Empirical. We empirically assess the efficacy of the SI framework through two real-world case studies: (i) Experimental—evaluating the feasibility of running data-efficient A/B tests using data from a large e-commerce platform. We call our proposed experimental design "synthetic A/B testing". (ii) Observational—evaluating the impact of mobility restrictions on COVID-19 related morbidity outcomes.

Section 3.5: Theoretical. We introduce a novel tensor factor model over potential outcomes, which is a natural generalization of the matrix factor model. Under this setting, we prove the identification of all $N \times D$ causal parameters, and establish finite-sample consistency and asymptotic normality of the SI estimator. Collectively, our results provide an answer to the open question of extending SC to multiple interventions, Abadie (2020).

Section 3.6: Robustness check for SI (and SC). We propose a data-driven hypothesis test to check for the feasibility of when a model learned can be transferred across interventional regimes and measurements. We provide guarantees for both its Type I and Type II errors. We use the test in the canonical SC case studies of (i) terrorism in Basque Country and (ii) California's Proposition 99 Abadie et al. (2010); Abadie and Gardeazabal (2003).

Section 3.7: Simulations. We run simulations which support our theoretical consistency and asymptotic normality results.

Section 3.8: Comparison with SC literature. By restricting our attention to estimating $\theta_n^{(0)}$ for $n \notin \mathcal{I}^{(0)}$, our identification and inference results immediately give new results for the SC literature. To better contextualize our work, we present a detailed comparison with some representative works in the SC literature.

Section 3.9: Causal inference & tensor completion. Our proposed tensor factor model over potential outcomes serves as a connection between causal inference and the growing field of tensor completion. Traditionally, the goal in tensor completion is to recover a tensor from its partial, noisy observations, where its entries are missing completely at random. In contrast, our work suggests that the SI method can be viewed as solving a "causal" variant of the tensor completion problem where there is confounding, i.e., the missingness pattern is correlated with the entries of the potential outcomes tensor. Indeed, we hope this provides a framework for future inquiry towards obtaining a clearer understanding of the trade-offs between sample complexity, statistical accuracy, experiment design, and the role of computation in estimating various causal parameters.

Notations. See Section 3.2 for formal definitions of the standard notations we use.

■ 3.2 Standard Notation

For a matrix $\boldsymbol{A} \in \mathbb{R}^{a \times b}$, we denote its transpose as $\boldsymbol{A}' \in \mathbb{R}^{b \times a}$. We denote the operator (spectral) and Frobenius norms of $\boldsymbol{A}$ as $\boldsymbol{A}_{\text{op}}$ and $\boldsymbol{A}_F$, respectively. The columnspace (or range) of $\boldsymbol{A}$ is the span of its columns, which we denote as $\mathcal{R}(\boldsymbol{A}) = \{v \in \mathbb{R}^a : v = \boldsymbol{A}x, x \in \mathbb{R}^b\}$. The rowspace of $\boldsymbol{A}$, given by $\mathcal{R}(\boldsymbol{A}')$, is the span of its rows. Recall that the nullspace of $\boldsymbol{A}$ is the set of vectors that are mapped to zero under $\boldsymbol{A}$. For any vector $v \in \mathbb{R}^a$, let v_p denote its ℓ_p-norm. Further, we define the inner product between vectors $v, x \in \mathbb{R}^a$ as $\langle v, x \rangle = \sum_{\ell=1}^{a} v_\ell x_\ell$. If v is a random variable, we define its sub-Gaussian (Orlicz) norm as v_{ψ_2}. Let $[a] = \{1, \dots, a\}$ for any integer a.

Let f and g be two functions defined on the same space. We say that $f(n) = O(g(n))$ if and only if there exists a positive real number M and a real number n_0 such that for all $n \geq n_0, |f(n)| \leq M|g(n)|$. Analogously we say: $f(n) = \Theta(g(n))$ if and only if there exists positive real numbers m, M such that for all $n \geq n_0,\ m|g(n)| \leq |f(n)| \leq M|g(n)|$; $f(n) = o(g(n))$ if for any $m > 0$, there exists n_0 such that for all $n \geq n_0, |f(n)| \leq m|g(n)|$.

We adopt the standard notations and definitions for stochastic convergences. As such, we denote $\xrightarrow{d}$ and $\xrightarrow{p}$ as convergences in distribution and probability, respectively. We will also make use of O_p and o_p, which are probabilistic versions of the commonly used deterministic O and o notations. More formally, for any sequence of random vectors X_n, we say $X_n = O_p(a_n)$ if for every $\varepsilon > 0$, there exists constants C_ε and n_ε such that $\mathbb{P}(X_{n2} > C_\varepsilon a_n) < \varepsilon$ for every $n \geq n_\varepsilon$; equivalently, we say $(1/a_n)X_n$ is "uniformly tight" or "bounded in probability". Similarly, $X_n = o_p(a_n)$ if for all $\varepsilon, \varepsilon' > 0$, there exists n_ε such that $\mathbb{P}(X_{n2} > \varepsilon' a_n) < \varepsilon$ for every $n \geq n_\varepsilon$. Therefore, $X_n = o_p(1) \iff X_n \xrightarrow{p} 0$. Additionally, we use the "plim" probability limit operator: $\text{plim } X_n = a \iff X_n \xrightarrow{p} a$. We say a sequence of events $\mathcal{E}_n$, indexed by n, holds "with high probability" (w.h.p.) if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$, i.e., for any $\varepsilon > 0$, there exists a n_ε such that for all $n > n_\varepsilon$, $\mathbb{P}(\mathcal{E}_n) > 1 - \varepsilon$. More generally, a multi-indexed sequence of events $\mathcal{E}_{n_1, \dots, n_d}$, with indices $n_1, \dots, n_d$ with $d \geq 1$, is said to hold w.h.p. if $\mathbb{P}(\mathcal{E}_{n_1, \dots, n_d}) \to 1$ as $\min\{n_1, \dots, n_d\} \to \infty$. We also use $\mathcal{N}(\mu, \sigma^2)$ to denote a normal or Gaussian distribution with mean μ and variance σ^2—we call it *standard* normal or Gaussian if $\mu = 0$ and $\sigma^2 = 1$.

■ 3.3 Synthetic Interventions: Estimator

In Section 3.3.1, we formally introduce the SI estimator. In Section 3.3.2, we discuss some practical heuristics for its implementation and supplement it with a more technical discussion to motivate and justify its various steps. Without loss of generality, we will focus on estimating $\theta_n^{(d)}$ for a given (n, d) pair.

Additional notation. Here, we introduce necessary notation required to formally state the estimator. Formally, let

$$Y_{\text{pre},n} = [Y_{tn0} : t \in \mathcal{T}_{\text{pre}}] \in \mathbb{R}^{T_0}$$

represent the vector of observed outcomes for unit n under $d = 0$ for $t \in \mathcal{T}_{\text{pre}}$. Let

$$\begin{aligned} \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} &= [Y_{tj0} : t \in \mathcal{T}_{\text{pre}},\ j \in \mathcal{I}^{(d)}] \in \mathbb{R}^{T_0 \times N_d}, \\ \boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}} &= [Y_{tjd} : t \in \mathcal{T}_{\text{post}},\ j \in \mathcal{I}^{(d)}] \in \mathbb{R}^{T_1 \times N_d}, \end{aligned}$$

represent the observed outcomes for units within $\mathcal{I}^{(d)}$ for the *pre* and *post* measurements. Note $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ is constructed using observed outcomes under intervention 0, while $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$ is constructed using outcomes under intervention d. We define the singular value decomposition (SVD) of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ as

$$\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} = \sum_{\ell=1}^{M} \widehat{s}_\ell \widehat{u}_\ell \widehat{v}_\ell',$$

where $M = \min\{T_0, N_d\}$, $\widehat{s}_\ell \in \mathbb{R}$ are the singular values (arranged in decreasing order), and $\widehat{u}_\ell \in \mathbb{R}^{T_0}, \widehat{v}_\ell \in \mathbb{R}^{N_d}$ are the left and right singular vectors, respectively. Note $\widehat{s}_\ell, \widehat{u}_\ell, \widehat{v}_\ell'$ are actually indexed by the measurements in $\mathcal{T}_{\text{pre}}$ and units in $\mathcal{I}^{(d)}$, but we suppress this dependence to increase readability.

■ 3.3.1 Estimator

The SI estimator is a simple three-step procedure, each with a closed-form expression. It has one hyper-parameter $k \in [M]$ that quantifies the number of singular components of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ to retain. The third step shows how to estimate the confidence interval for $\widehat{\theta}_n^{(d)}$; for simplicity, we pick the 95% confidence interval.

1. Learn a linear model $w^{(n,d)}$ between unit n and $\mathcal{I}^{(d)}$.

$$\widehat{w}^{(n,d)} = \Big(\sum_{\ell=1}^{k} (1/\widehat{s}_\ell) \widehat{v}_\ell \widehat{u}_\ell' \Big) Y_{\text{pre},n}. \tag{3.3}$$

2. Estimate $\theta_n^{(d)}$:

$$\widehat{\mathbb{E}}[Y_{tn}^{(d)}] = \sum_{j \in \mathcal{I}^{(d)}} \widehat{w}_j^{(n,d)} Y_{tjd}, \quad \text{for } t \in \mathcal{T}_{\text{post}} \tag{3.4}$$

$$\widehat{\theta}_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \widehat{\mathbb{E}}[Y_{tn}^{(d)}]. \tag{3.5}$$

3. Produce the 95% confidence interval:

$$\theta_n^{(d)} \in \left[\widehat{\theta}_n^{(d)} \pm \frac{1.96 \cdot \widehat{\sigma} \widehat{w}^{(n,d)}{}_2}{\sqrt{T_1}} \right], \tag{3.6}$$

where

$$\widehat{\sigma}^2 = \frac{1}{T_0} \left\| Y_{\text{pre},n} - \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} \right\|_2^2. \tag{3.7}$$

■ 3.3.2 Discussion of the SI Estimator

Below, we discuss the three steps of the SI estimator. Our goal here is to provide (i) practical heuristics of when and how to implement the SI estimator in practice and (ii) a more technical justification for each step.

Step 1: Estimating $\widehat{w}^{(n,d)}$

The first step of the SI estimator, given in (3.3), estimates a linear relationship $\widehat{w}^{(n,d)}$ between the observed outcomes $Y_{\text{pre},n}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ by doing singular value thresholding (SVT) on the matrix $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ and subsequently running ordinary least squares (OLS) on the resulting matrix and $Y_{\text{pre},n}$. This is also known in the literature as principal component regression (PCR) Agarwal et al. (2021e,d). The number of singular values retained is a hyper-parameter $k \in \min\{T_0, N_d\}$. Below we justify when and why PCR is appropriate in the setting of latent factor models and how to choose k.

Why PCR and how to choose k? The SI estimator, like various other methods such as SC, DID and its variants, learn a linear model, i.e., $\widehat{w}^{(n,d)}$ between the target unit n and the “donor” units in $\mathcal{I}^{(d)}$. However to ensure the linear model is not “overfit”, the different variants of these methods suggest different ways to regularize the linear fit. In particular,

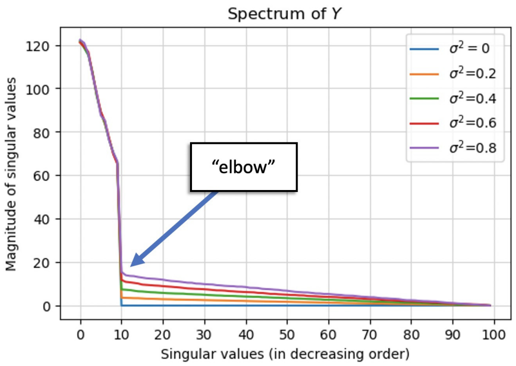


Figure 3.3: Simulation displays the spectrum of $\boldsymbol{Y} = \mathbb{E}[\boldsymbol{Y}] + \boldsymbol{E} \in \mathbb{R}^{100\times100}$. Here, $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{U}\boldsymbol{V}'$, where the entries of $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{100\times10}$ are sampled independently from $\mathcal{N}(0,1)$; further, the entries of $\boldsymbol{E}$ are sampled independently from $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 \in \{0, 0.2, \dots, 0.8\}$. Across varying levels of noise σ^2, there is a steep drop-off in magnitude from the top to remaining singular values—this marks the "elbow" point. As seen from the figure, the top singular values of $\boldsymbol{Y}$ correspond closely with that of $\mathbb{E}[\boldsymbol{Y}]$ ($\sigma^2 = 0$), and the remaining singular values are induced by $\boldsymbol{E}$. Thus, $\text{rank}(\boldsymbol{Y}) \approx \text{rank}(\mathbb{E}[\boldsymbol{Y}]) = 10$.

in SC, it is standard to impose that the linear fit is convex, i.e., the coefficients of $\widehat{w}^{(n,d)}$ are non-negative and sum to 1. More recent works impose an ℓ_1-penalty term to the linear fit, e.g., Chernozhukov et al. (2020b), which is known as LASSO in the literature, and it promotes sparsity in the learned $\widehat{w}^{(n,d)}$, i.e., forces most of the coefficients of $\widehat{w}^{(n,d)}$ to be 0. PCR can be seen as another form of regularization, where rather than regularizing $\widehat{w}^{(n,d)}$ directly, it imposes "spectral sparsity", i.e., sets most of the singular values of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ to be 0, before learning a linear model. Indeed, PCR is particularly effective as a regularization technique if $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ is (approximately) low-rank—for further discussion on this point, please refer to the technical discussion below and Agarwal et al. (2021e,d).

With regards to selecting the hyper-parameter k, there exist a number of principled heuristics to choose it and we name a few here. Perhaps the most popular data-driven approach is simply to use cross-validation, where the pre-intervention data is our training set and the post-intervention data is our validation set. Another standard approach is to use a "universal" thresholding scheme that preserves the singular values above a precomputed threshold (see Chatterjee (2015) and Donoho and Gavish (2013)). Finally, a "human-in-the-loop" approach is to inspect the spectral characteristics of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$, and choose k to be the natural "elbow" point that partitions the singular values into those of large and small magnitudes. For a graphical depiction of such an elbow point, see Figure 3.3. We observe this clear elbow point in both case studies in Section 4.6.3. As such, if this approximate low-rank spectral profile is not in the data, then using PCR is likely not appropriate.

Technical discussion. Here, we give a more technical justification of when we expect an elbow point in the spectrum of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$. The key assumption we make in Section 3.5.1 is that $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ is a low-rank matrix. If $\boldsymbol{E}_{\text{pre},\mathcal{I}^{(d)}} = \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ has independent sub-Gaussian rows, then random matrix theory informs us that the singular values of $\boldsymbol{E}_{\text{pre},\mathcal{I}^{(d)}}$ are much smaller in magnitude compared to those of the signal matrix $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$. Hence, it is likely that a "sharp" threshold or "large" gap exists between the top singular values associated with $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$. Specifically, the largest singular value of $\boldsymbol{E}_{\text{pre},\mathcal{I}^{(d)}}$ scales as $O_p(\sqrt{T_0} + \sqrt{N_d})$, cf. Vershynin (2018). In comparison, if the entries of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ are $\Theta(1)$ and its nonzero singular values are of the same magnitude, then they will scale as $\Theta(\sqrt{T_0 N_d})/r_{\text{pre}})$, where r_{pre} is the rank of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, which is $\gg \sqrt{T_0} + \sqrt{N_d}$ when both T_0 and N_d are growing.

This low-rank structure is also what motivates the existence of a a linear model $w^{(n,d)}$. Consider $\boldsymbol{Y}_{\text{pre},[N]} = [Y_{tj0} : t \in \mathcal{T}_{\text{pre}}, j \in [N]] \in \mathbb{R}^{T_0 \times N}$. If $\mathbb{E}[\boldsymbol{Y}_{\text{pre},[N]}]$ is low rank, it suggests that $\mathbb{E}[Y_{\text{pre},n}]$, the column in $\mathbb{E}[\boldsymbol{Y}_{\text{pre},[N]}]$ corresponding to unit n, can be well-approximated as a linear combination of a few other columns of the matrix $\mathbb{E}[\boldsymbol{Y}_{\text{pre},[N]}]$; in particular, it can be well-approximated by the columns in the sub-matrix $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$. This linear combination of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ that represents $\mathbb{E}[Y_{\text{pre},n}]$ is precisely $w^{(n,d)}$. Given the model, we do not get to observe $\mathbb{E}[Y_{\text{pre},n}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, rather only their "noisy" versions given by $Y_{\text{pre},n}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$. This setting is also known in the literature as "error-in-variables" regression. As discussed above, by running PCR on $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$, we effectively "de-noise" it to approximately recover $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, and subsequently we show $\widehat{w}^{(n,d)}$ is close to $w^{(n,d)}$—see Lemma 4.9.1.

Step 2: Estimating $\widehat{\theta}_n^{(d)}$

When can we "transfer" $w^{(n,d)}$ across interventional regimes and measurements? Step 2 of the SI estimator, defined in (3.4) and (3.5), applies the learned model $\widehat{w}^{(n,d)}$ on $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$ to produce the counterfactual estimate $\widehat{\theta}_n^{(d)}$. This should lead to some pause as we are learning $\widehat{w}^{(n,d)}$ using $Y_{\text{pre},n}, \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$, which correspond to outcomes for $d = 0$ and for $t \in \mathcal{T}_{\text{pre}}$. However, we are applying $\widehat{w}^{(n,d)}$ on $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$ which correspond to outcomes for any $d \in [D]_0$ and for $t \in \mathcal{T}_{\text{post}}$. Thus, Step 2 of the estimation procedure implicitly assumes $\mathbb{E}[Y_{\text{post},n}^{(d)}] = \mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]w^{(n,d)}$!

In effect, Step 2 of SI estimator is a form of *causal transportation*. We find a necessary condition for such causal transportation to be feasible is that the "complexity" of $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$ is less than that of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$. To build intuition, consider the setting where $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ is a matrix of 0's, i.e., it has a degenerate rowspace. Then, it is easy to see that we cannot effectively learn $w^{(n,d)}$ using *any* regression procedure. Hence, we require a natural condition that the span of the right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$ (i.e., its rowspace) lies within that of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$—see Assumption 15. Of course, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ are not observed, but we can estimate their respective rowspaces via $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$. The data-driven hypothesis test we propose in Section 3.6.1 uses $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ to check whether this subspace inclusion property holds. The discussion above suggests that before running Step 2 of the SI estimator, one should first run this hypothesis test as a robustness check to verify if it feasible to transfer the learned model $\widehat{w}^{(n,d)}$ between $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ and $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$. If it fails, then causal transportation across is likely infeasible.

Technical discussion. We now motivate and formally introduce the test statistic that underpins our proposed hypothesis test. Let $r_{\text{pre}} = \text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}])$, and let $r_{\text{post}} = \text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}])$. Let $\boldsymbol{V}_{\text{pre}} \in \mathbb{R}^{N_d \times r_{\text{pre}}}$ denote the right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$; analogously, define $\boldsymbol{V}_{\text{post}} \in \mathbb{R}^{N_d \times r_{\text{post}}}$ with respect to $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$. Let $\widehat{\boldsymbol{V}}_{\text{pre}} \in \mathbb{R}^{N_d \times r_{\text{pre}}}$ and $\widehat{\boldsymbol{V}}_{\text{post}} \in \mathbb{R}^{N_d \times r_{\text{post}}}$ denote their respective estimates, which are constructed from the top r_{pre} and r_{post} right singular vectors of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ and $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$, respectively. As discussed above, note if causal transportation as required by Step 2 of the SI estimator is to hold, then we require that the span of $\boldsymbol{V}_{\text{post}}$ is contained within that of $\boldsymbol{V}_{\text{pre}}$. Since do not observe $\boldsymbol{V}_{\text{pre}}$ and $\boldsymbol{V}_{\text{post}}$, a a natural test statistic, $\widehat{\tau}$, is the distance between the $\widehat{\boldsymbol{V}}_{\text{pre}}$ and $\widehat{\boldsymbol{V}}_{\text{post}}$:

$$\widehat{\tau} = (\boldsymbol{I} - \widehat{\boldsymbol{V}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}}) \widehat{\boldsymbol{V}}_{\text{post}}{}_F^2. \tag{3.8}$$

Indeed, if $\widehat{\boldsymbol{V}}_{\text{pre}} = \boldsymbol{V}_{\text{pre}}$, $\widehat{\boldsymbol{V}}_{\text{post}} = \boldsymbol{V}_{\text{post}}$, and the span of $\widehat{\boldsymbol{V}}_{\text{post}}$ is within that of $\widehat{\boldsymbol{V}}_{\text{pre}}$, then $\widehat{\tau} = 0$. To account for noise, for a given significance level $\alpha \in (0, 1)$, we reject the validity of Step 2 of the SI method to estimate $\widehat{\theta}_n^{(d)}$, if $\widehat{\tau} > \tau(\alpha) \geq 0$ for $\tau(\alpha)$ as defined in (3.20). See Section 3.6 for details and formal results.

Step 3: Uncertainty Quantification of $\widehat{\theta}_n^{(d)}$

The confidence interval we produce in (3.6) is a direct implication of our asymptotic normality result given in Theorem 4.5.2. The variance scales with two quantities $\widehat{w}^{(n,d)}{}_2$

and $\widehat{\sigma}$. Both quantities are standard. In particular, $\widehat{w}^{(n,d)}{}_2$ is the ℓ_2-norm our estimate of the linear model and $\widehat{\sigma}$ as defined in (3.7) is essentially our in-sample "training" error. This suggest that if our in-sample training error as produced by PCR is large, then naturally our uncertainty of the estimated $\widehat{\theta}_n^{(d)}$ grows accordingly.

■ 3.4 Empirics

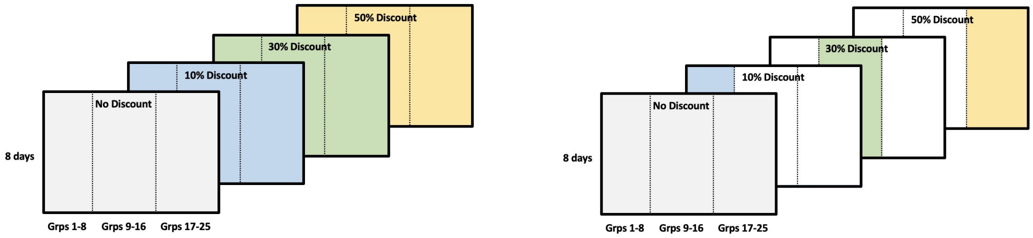
We present two real-world case-studies to assess the efficacy of the SI framework and explore possible applications both in experimental and observational settings.

■ 3.4.1 Synthetic A/B Testing via SI

Background

We use data collected from a large e-commerce company that segmented its users into $N = 25$ customer sub-populations/groups (i.e., units), with approximately $10,000$ individual users per group, based on the historical engagement of a user, as measured by time and money spent on the platform. The aim of the company was to learn how different discount levels (i.e., interventions) affected the engagement levels of each of the 25 customer groups. The levels were 10%, 30%, and 50% discounts over the regular subscription cost (control or 0% discount.). Thus, there are total of $D = 4$ interventions. The A/B test was performed by *randomly* partitioning users in each of the $N = 25$ customer groups into 4 subgroups; these subgroups corresponded to either one of the 3 discounts strategies or control. User engagement in each of these $N \times D = 25 \times 4 = 100$ experiments was measured daily over $T = 8$ days. We only had access to the average engagement level of all the customers in each of the 100 experiments. That is, we had access to 100 trajectories each of length 8.

This e-commerce A/B test is particularly suited to validate the SI framework as we observe the engagement levels of each customer group under each of the three discounts and control, i.e., for each customer group, we observe all possible "counterfactual" outcomes. Further, it is an experimental setting and thus there is no latent confounding in how interventions were assigned. Given that we have access to all possible potential outcomes of interest, we can test the efficacy of the SI framework by exposing data from a limited number of experiments to the SI estimator, and see how well it can recreate the customer



(a) Actual experimental setup.

(b) SI experimental setup.

Figure 3.4: Experimental setups for the e-commerce A/B testing case study. Observations are represented by colored blocks while unobservable counterfactuals (held-out test sets) are represented by white blocks.

engagement outcomes in the hidden experiments.

Experimental Setup

To simulate how a *data-efficient* synthetic A/B test (i.e., RCT) can be run, we randomly partition the 25 customer groups into three clusters, denoted as customer groups 1-8, 9-16, and 17-25. For all 25 customer groups, we give the SI estimator access to the outcomes under 0% discount (i.e., control). For the 10% discount level, we give the SI estimator access to data from customer groups 1-8, but hold out the corresponding data for groups 9-25. In other words, the SI estimator does not get to observe the trajectories of groups 9-25 under a 10% discount. Using the observed trajectories of customer groups 1-8, we separately apply the SI estimator to create synthetic user engagement trajectories for each of the 9-25 customer groups under a 10% discount. Analogously, we only give the SI estimator data for the 30% and 50% discount, for customer groups 9-16 and 17-25, respectively. Note here $T_0 = T_1 = T$.

See Figure 3.4a and 3.4b for a visual depiction of the full A/B test carried out, and our proposed synthetic A/B test, respectively. Given this setup, out interest is in recovering the following causal parameters: $\theta_n^{(30\%)}$, $\theta_n^{(50\%)}$ for groups 1-8; $\theta_n^{(10\%)}$, $\theta_n^{(50\%)}$ for groups 9-16; $\theta_n^{(10\%)}$, $\theta_n^{(30\%)}$ for groups 17-25.

Applying the SI Estimator

Feasibility checks. As discussed in Section 3.3.2, we run two feasibility checks: (i) inspect that the appropriately defined $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ is approximately low-rank; (ii) run the hypothesis

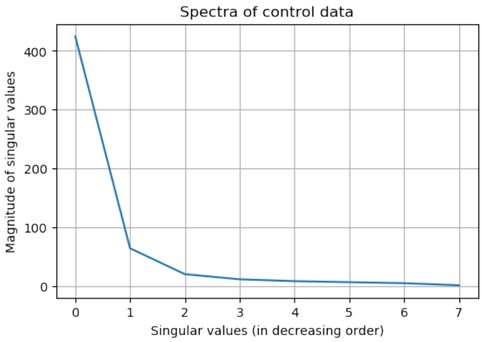


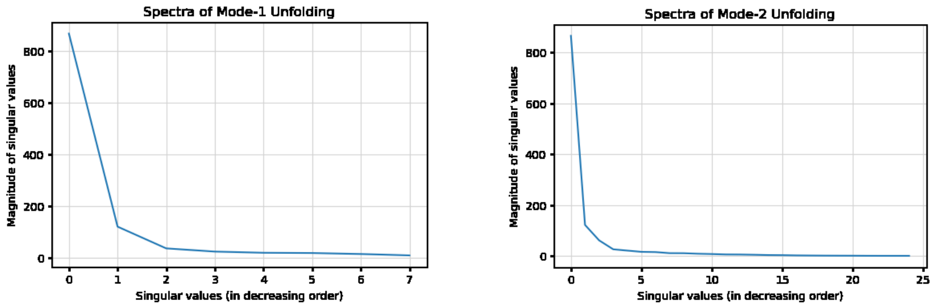
Figure 3.5: Spectral profile of control data for groups 1-25.

test to check for subspace inclusion, a necessary condition for causal transportablity for the SI estimator. Let $d = \{0, \dots, 3\}$ correspond to discount levels $\{0\%, 10\%, 30\%, 50\%\}$, respectively. We plot the spectrum of $\boldsymbol{Y}_{\text{pre},[N]} = [\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} : d \in \{0, \dots, 3\}]$ in Figure 3.5 and observe that it is approximately low-rank. In Table 3.1, we show the hypothesis test results for the three discounts. The hypothesis test passes for each discount at a significance level of $\alpha = 0.05$. Thus, this dataset passes both feasibility checks.

Further, since we have access to the customer engagement outcomes from all 100 experiments, we can inspect the spectral profile of the induced tensor of potential outcomes. This allows to check whether the latent factor model across units, measurements, *and interventions holds*—see Assumption 9 for a formal description of the tensor factor model. Specifically, let $\boldsymbol{Y} = [Y_{tnd} : t \in [8], n \in [25], d \in [4]_0]$ be a order-3 tensor, where Y_{tnd} represents the observed engagement level for customer group n, on day t under discount policy d. Consider the mode-1 and mode-2 unfolding of $\boldsymbol{Y}$, which result into $8 \times (25 \times 4)$ and $25 \times (8 \times 4)$ matrices, respectively. We plot the spectra of the mode-1 and mode-2 unfoldings of $\boldsymbol{Y}$ as shown in Figure 3.6. For both mode-1 and mode-2 unfoldings of the tensor $\boldsymbol{Y}$, over 99% of the spectral energy is captured by the top two singular values. This suggests that $\boldsymbol{Y}$ has a low canonical polyadic tensor rank, (Farias and Li, 2019, Proposition 1).

Empirical results. To quantify the accuracy of counterfactual predictions, we need a meaningful baseline to compare against. To that end, we use the following squared error metric for any (n, d) pair:

$$\text{SE}_n^{(d)} = 1 - \frac{(\theta_n^{(d)} - \widehat{\theta}_n^{(d)})^2}{(\theta_n^{(d)} - (1/T_1 N_d) \sum_{t \in \mathcal{T}_{\text{post}}} \sum_{j \in \mathcal{I}^{(d)}} Y_{tjd})^2}. \tag{3.9}$$



(a) Spectra of mode-1 unfolding.

(b) Spectra of mode-2 unfolding.

Figure 3.6: In both 4.5b and 4.5c, the top two singular values of the mode-1 and mode-2 unfoldings of the tensor capture more than 99% of the spectral energy.

Intervention	Hypo. Test ($\alpha = 0.05$)	$\text{SE}^{(d)}$
10% discount	Pass	0.98
30% discount	Pass	0.99
50% discount	Pass	0.99

Table 3.1: A/B test case study: hypothesis test and prediction accuracy results.

In words, the numerator in the right most term on the right-hand side of (3.9) represents the squared error associated with the SI estimate $\widehat{\theta}_n^{(d)}$. We coin the denominator, $(1/N_d) \sum_{j \in \mathcal{I}^{(d)}} Y_{tjd}$, as the "RCT estimator"; this is the average outcome across all units within subgroup $\mathcal{I}^{(d)}$. If the units are indeed homogeneous (i.e., they react similarly to intervention d), then the RCT estimator will be a good predictor of $\theta_n^{(d)}$. Therefore, $\text{SE}_n^{(d)} > 0$ indicates the success of the SI estimator over the RCT baseline and (3.9) can be interpreted as a modified R^2 statistic with respect to this baseline. In effect, the $\text{SE}_n^{(d)}$ captures the gain by "personalizing" the prediction to the target unit using SI over the natural baseline of taking the average outcome over $\mathcal{I}^{(d)}$.

Across the three discounts, SI achieves a median $\text{SE}^{(d)}$ of at least 0.98 as denoted in Table 3.1. $\text{SE}^{(d)}$ is calculated as the median of $\text{SE}_n^{(d)}$ over all $n \notin \mathcal{I}^{(d)}$. We see that the SI estimator far outperforms the RCT estimator. This indicates significant heterogeneity amongst the customer groups in how they respond to discounts, and thus reinforces the need for personalization.

Takeaways: Towards Data-efficient Synthetic A/B Testing

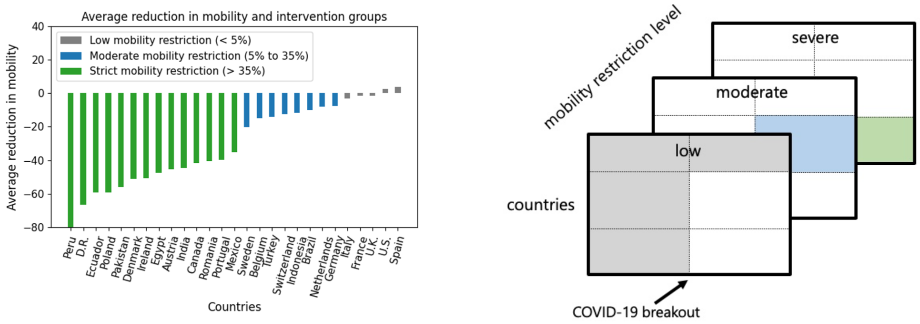
In this A/B testing framework, the e-commerce company implemented a total of 100 distinct experiments—one experiment for each of the 25 customer groups under each of the 4 interventions. In contrast, SI only required observations from 50 experiments to produce counterfactual estimates for the remaining 50 experiments accurately. More generally, as discussed in Section 3.1, an RCT requires $N \times D$ experiments to estimate the optimal "personalized" intervention for every unit. Meanwhile, the synthetic A/B testing setup as described here, requires access to data for only $2N$ experiments: in the first N experiments, all units are under the same interventional regime, say control ($d = 0$); next, divide all N units into D partitions each of size N/D, and assign intervention d to units in the d-th partition, which leads to another N experiments. Using these $2N$ experiments, we can recover the causal parameters associated with each of the N units for the remaining $D - 2$ interventions. Thus, the number of required experiments does not scale with D, which can become significant as the number of interventions, i.e, level of personalization, grows. This efficiency can be significant especially when experimentation is costly and/or unethical (e.g., clinical trials for personalized medicine). Lastly, we emphasize the SI estimator did not use additional covariates about the customer groups or discounts. Rather, it only required a unique identifier (UID) for each customer group and discount.

■ 3.4.2 Mobility Restriction Policies and COVID-19 Morbidity Outcomes

Background

At the onset of the COVID-19 pandemic across the globe, the policies that different nations enacted were primarily targeted at restricting mobility to curb the spread of the virus. A question of interest to policy-makers is how effective was mobility restriction, as it can help understand the trade-offs between health outcomes and economic impact of these policies. Although it is infeasible to run experiments in this setting, we explore how the SI framework can leverage readily available observational data from across the globe to help answer these questions. Below, we list and justify our key modeling decisions.

Outcome metric: daily death counts. Due to its relative reliability and availability, we use



(a) (b)

Figure 3.7: Figure 3.7a displays the average reduction in mobility and assigned intervention groups. Figure 3.7b shows the observation pattern according to these assignments.

daily COVID-19 related morbidity outcomes as our outcome variable of interest, taken from Dong et al. (2020). The other standard metric, number of daily infections, is much less reliable due to the inconsistencies in testing and reporting across regions.

Intervention: change in mobility rate. At the start of the pandemic, each country implemented numerous policies to combat the spread of COVID-19. This makes it difficult to analyze any particular policy (e.g., stay-at-home orders vs. schools shutting down) in isolation. Thus a key assumption we make is that at the start of the pandemic, almost all such policies had been directed towards restricting how individuals move and interact. That is, we assume the effect of these various policies on COVID-19 related morbidity outcomes is solely mediated via the level of mobility restriction. To that end, we use Google's mobility reports Google (2020) to study the change in a country's mobility compared to their respective national baseline from January 2020. Thus, we adopt mobility as our notion of intervention, and investigate how a country's change in mobility level potentially affects COVID-19 related morbidity outcomes. The additional challenge here compared to the A/B testing case study is that we only have access to observational data. That is, there may be observed and/or latent characteristics associated with the country that might both influence the mobility restriction policy enacted (e.g., population demographics, cultural trends, governmental structure) as well as the health outcomes observed.

Experimental Setup

Pre- and post-intervention periods. Recall from Assumption 8 that to apply SI, we require a set of outcome measurements for all units that are under a common intervention. Towards that, using Google's mobility reports, we verify that 20 days prior to cumulative 80 COVID-19 related deaths in a country (and any time before), none of the $N = 27$ countries we select in this case study restricted mobility. Thus, rather than using a particular calendar date, we choose the day a country has cumulative 80 COVID-19 deaths as "Day 0". Henceforth, we refer to the pre- and post-intervention periods as the days before and after Day 0, respectively. In particular, $T_0 = 20$, and we measure post-intervention outcomes for the first 15 days from the onset of the pandemic in a country, i.e., $T_1 = 15$.

Categorizing countries by intervention received: average (lagged) mobility score. Studies have shown that there is a median lag of 20 days from the onset of infection to the day of death (e.g., see Wilson et al. (2020)). Thus, a country's death count on a particular day is a result of the infection levels from approximately 20 days prior. In order to analyze the effect of a mobility restricting intervention from "Day 0" onwards, we consider a country's mobility score from Day -20 to Day -1. Given that Google's mobility score changes every day, we take its average in a given country from Day -20 to Day -1, and then bucket it into the $D = 3$ intervention groups defined as follows; see Figure 3.7a for a visual depiction of this clustering. For a given country, we define: (a) "*low mobility restriction*" as a reduction in mobility that is less than 5% compared to the national baseline from January 2020; (b) "*moderate mobility restriction*" as a reduction in mobility between 5% to 35% compared to the national baseline from January 2020; (c) "*severe mobility restriction*" as a reduction in mobility greater than 35% compared to the national baseline from January 2020. We remark that by discretizing mobility from a numerical trajectory over 20 days into these three categorical buckets is a significant modeling choice and it coarsens the possible causal parameters of interest. See Figure 3.7b for a visual depiction of the observation pattern.

Applying the SI Estimator

Feasibility checks. We run the same two feasibility checks as in the A/B testing case study. In particular, we (i) inspect the spectrum of $\boldsymbol{Y}_{\text{pre},[N]} = [\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} : d \in$

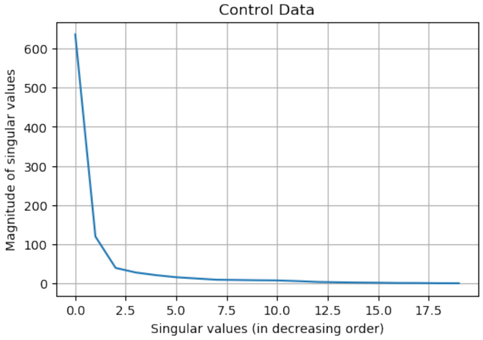


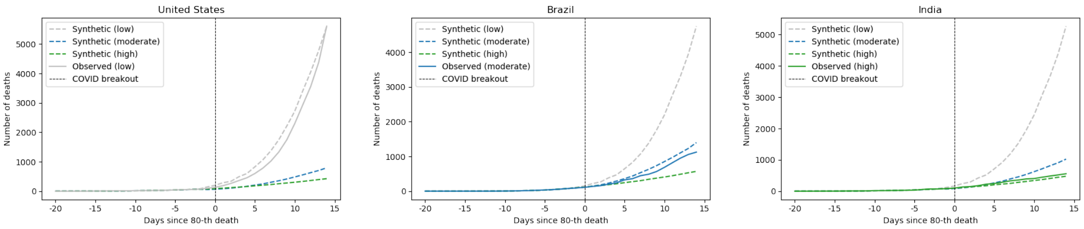
Figure 3.8: Spectral profile of control data across all countries.

{low, moderate, severe}]; (ii) run the hypothesis test for subspace inclusion. Let $d = \{0, 1, 2\}$ correspond to the mobility restrictions {"low", "moderate", "severe"} as described above, respectively. We plot the spectrum of $\boldsymbol{Y}_{\text{pre},[N]}$ in Figure 3.8, which clearly exhibits low-rank structure. In Table 3.2, we show that our dataset passes the hypothesis test for all mobility restriction levels at a significance level of $\alpha = 0.05$. Thus, both feasibility checks pass.

Empirical results. We apply SI using the setup above to produce counterfactual predictions of the daily COVID-19 morbidity outcomes for the 15 days following Day 0 under the three mobility restriction levels for each country. In this case study, since we do not have access to the counterfactual trajectory of a country for a mobility restricting intervention it did not actually go through, the best we can do is leave-one-out cross validation. For a country n that undergoes mobility restricting intervention d in the post-intervention period (i.e., $n \in \mathcal{I}^{(d)}$), we hide the post-intervention data for that country, and we see how well SI is able to recreate that trajectory using only that country's pre-intervention data. We then use the estimated trajectory from SI to produce $\text{SE}_n^{(d)}$, as defined in (3.9). Subsequently, we define $\text{SE}^{(d)}$ as the median $\text{SE}_n^{(d)}$ over $n \in \mathcal{I}^{(d)}$. In Table 3.2, we see that the median $\text{SE}^{(d)}$ is 0.46, 0.80, 0.08 for the low, moderate, and severe mobility restriction, respectively. This indicates that there is indeed significant heterogeneity amongst the countries in how mobility affects the COVID-19 related morbidity outcomes at least with respect to the low and moderate restriction. For severe mobility restriction, the SI estimator only slightly outperforms the RCT estimate; this is likely due to the fact that countries that underwent a severe mobility restriction all had very few COVID-19 related deaths in the post-intervention period, and thus there was not much heterogeneity in their trajectories.

Intervention	Hypo. Test ($\alpha = 0.05$)	SE$^{(d)}$
low mobility restriction	Pass	0.46
moderate mobility restriction	Pass	0.80
severe mobility restriction	Pass	0.08

Table 3.2: COVID-19 case study: hypothesis test and prediction accuracy results.



(a) United States **(b)** Brazil **(c)** India

Figure 3.9: Counterfactual predictions of COVID-19 related morbidity counts under different mobility restriction levels.

For every mobility restriction level, we display the three synthetic trajectories produced by SI for one representative country, using only the pre-intervention outcomes for each given country. This corresponds to producing two counterfactual trajectories per country, for the two mobility interventions it did not actually go through, and one trajectory for the intervention it did actually go through, which serves as cross-validation. Hence, in this case study, we produce $2 \times 27 = 54$ counterfactual trajectories. For the low mobility restricting regime, we show results for USA in Figure 3.9a. The dashed lines on Days 0-15 are the predicted values under all possible mobility restriction levels and the solid line represents the true national death trajectory. Pleasingly, the SI predictions closely match the observed morbidity outcomes in the post-intervention period. Similarly, for the moderate and severe mobility restricting interventions, we display results for Brazil and India in Figures 3.9b and 3.9c, respectively. Again, the cross-validation trajectory produced from SI closely matches the observed morbidity outcomes under all different interventions. We note similar results hold generally across all countries we use in this study. We display the results for USA, Brazil, and India as they are the largest countries within each intervention group.

Further, we emphasize that we produce these trajectories using *only* their COVID-19 related death outcomes and the Google mobility report to categorize which intervention bucket they belong to in the post-intervention period. That is, we do not use any additional covariates about the countries or the various interventions.

Takeaways: Towards Personalized Program Evaluation

Importantly, the SI model of each country is fit in the pre-intervention period, when no intervention has yet occurred. Still, the learned model transfers to an interventional setting, i.e., when the interventions take effect within the donor countries. This helps validate the SI framework. An "optimistic" conclusion one can draw from the figures above is that, uniformly across all countries, there is a significant drop in the number of deaths with even a "moderate" drop in mobility (i.e, a 5-35% drop in mobility compared to the national baseline). After this point, gains by further restricting mobility seem to be diminishing. Of course, with any such conclusions, it needs to be rigorously cross-validated with other studies of a similar nature; further these estimated counterfactual trajectories are only valid up to the modeling decisions made and under appropriate causal assumptions. In Section 3.5, we provide a formal causal framework for SI and our associated consistency and normality results for the SI estimator.

■ 3.5 Formal Results

In this section, we present formal results. In Section 3.5.1, we introduce our causal framework. In Section 3.5.2, we establish identification of the causal parameter of interest, $\theta_n^{(d)}$ under this framework. In Sections 3.5.3 and 3.5.4, we establish finite-sample consistency and asymptotic normality of the SI estimator, respectively. Finally, in Section 3.5.5, we interpret and discuss our key assumptions and formal results.

■ 3.5.1 Causal Framework

Tensor Factor Model

Below we introduce a novel tensor factor model for potential outcomes across units, measurements, and interventions.

Definition 3.5.1 (Tensor factor model). *For any unit $n \in [N]$, intervention $d \in [D]_0$ and outcome $t \in [T]$, the expectation of potential outcome $Y_{tn}^{(d)}$ satisfies the factor structure*

$$\mathbb{E}[Y_{tn}^{(d)}] = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell} w_{d\ell},$$

where $r \geq 1$ represents the 'rank' or the dimension of the latent factors, and $u_t, v_n, w_d \in \mathbb{R}^r$ represent the latent factors associated with the t-th measurement, n-th unit, and d-th intervention, respectively.

The tensor factor model is a natural generalization of the matrix factor model traditionally considered in the literature. Specifically, when restricted to any specific intervention d, we can re-write

$$\mathbb{E}[Y_{tn}^{(d)}] = \langle u_t^{(d)}, v_n \rangle \tag{3.10}$$

where $u_t^{(d)} = [u_{t\ell} w_{d\ell} : \ell \in [r]] \in \mathbb{R}^r$ and $v_n = [v_{n\ell} : \ell \in [r]] \in \mathbb{R}^r$. Indeed, (3.10) states that the expected potential outcomes for each intervention should satisfy the traditional matrix factor model. However, the critical additional assumption is that the unit n specific factor v_n remains invariant across interventions. In this work, we shall require this weaker condition, which is implied by the tensor factor model.

Assumption 9 (Invariant unit factors). *For any given (t, n, d),*

$$Y_{tn}^{(d)} = \langle u_t^{(d)}, v_n \rangle + \varepsilon_{tn}^{(d)}, \tag{3.11}$$

where $u_t^{(d)} \in \mathbb{R}^r$ is the latent factor specific to (t, d); $v_n \in \mathbb{R}^r$ is the latent factor specific to n; and $\varepsilon_{tn}^{(d)} \in \mathbb{R}$ is a mean zero residual term specific to (t, n, d).

Assumption 9 essentially states that the collection of latent factors

$$\mathcal{LF} := \left\{ u_t^{(d)}, v_n : (t, n, d) \right\}$$

determine the expected potential outcomes $\{\mathbb{E}[Y_{tn}^{(d)}] : (t, n, d)\}$. Thus, the distribution of the potential outcomes is captured through the residual term $\{\varepsilon_{tn}^{(d)} : (t, n, d)\}$.

Latent Factors can be Latent Confounders

Collectively, let the intervention assignments be denoted as

$$\mathcal{D} = \{(t, n, d) : \; Y_{tnd} \neq \star, \text{ i.e., } Y_{tn}^{(d)} \text{ is observed}\}.$$

In an ideal setting, we wish to have $\mathcal{D}$ and $Y_{tn}^{(d)}$ be independent, as is the case in a RCT. However, in observational studies, the interventions and potential outcomes can be dependent, which is known as confounding. We consider a setting where the confounders can be latent, but their impact on interventions and potential outcomes is mediated through the latent factors. That is, we assume conditioned on the latent factors, $\mathcal{LF}$, the potential outcomes and interventions are conditionally independent.

Assumption 10 (Selection on latent factors)**.** *Conditioned on the unobserved latent factors $\mathcal{LF}$, the interventions $\mathcal{D}$ and $\{\varepsilon_{tn}^{(d)} : t \in [T], n \in [N], d \in [D]_0\}$ are independent. That is, for all (t, n, d),*

$$\mathbb{E}[\varepsilon_{tn}^{(d)} | \mathcal{LF}] = 0, \qquad \varepsilon_{tn}^{(d)} \perp\!\!\!\perp \mathcal{D} \mid \mathcal{LF}.$$

Strictly speaking, our identification results only require $\mathbb{E}[\varepsilon_{tn}^{(d)} | \mathcal{LF}, \mathcal{D}] = 0$, which is known as conditional mean independence. However, we state it as $\varepsilon_{tn}^{(d)} \perp\!\!\!\perp \mathcal{D} \mid \mathcal{LF}$ to increase the interpretability of the conditional exogeneity assumption we make. Assumptions 9 and 10 collectively imply that $Y_{tn}^{(d)} \perp\!\!\!\perp \mathcal{D} \mid \mathcal{LF}$. Hence, this conditional independence condition can be thought of as "selection on latent factors", which is analogous to "selection on observables". Similar conditional independence assumptions have been considered in Athey et al. (2021); Kallus et al. (2018).

In the context of the COVID-19 case study described in Section 3.4.2, it suggests that the mobility restriction each country went through can be dependent on latent factors (e.g. national politics, cultural trends, population demographics), which might also influence the COVID-19 morbidity outcomes under different interventions. However, Assumption 10 requires that the collection of latent factors is rich enough that conditional on it, the potential health outcomes of a country and the mobility restriction interventions are independent.

■ 3.5.2 Causal Parameter & Identification

Target causal parameter. We can now formally define our target causal parameter. For any given (n, d), we aim to estimate

$$\theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}\left[Y_{tn}^{(d)} \;\middle|\; \{u_t^{(d)}, v_n : t \in \mathcal{T}_{\text{post}}\} \right]. \tag{3.12}$$

That is, for each unit n and intervention d, our interest is in the average expected potential outcomes over $\mathcal{T}_{\text{post}}$.

Identification. Next, we argue that our target causal parameter can be written as a function of observed outcomes, i.e., we establish identification of our causal parameter. Let $\mathcal{E} = \{\mathcal{LF}, \mathcal{D}\}$ refer to the collection of latent factors and intervention assignments. To that end, we make an additional mild assumption.

Assumption 11 (Linear span inclusion). *Given (n, d) and conditioned on $\mathcal{E}$, $v_n \in \text{span}(\{v_j : j \in \mathcal{I}^{(d)}\})$, i.e., there exists $w^{(n,d)} \in \mathbb{R}^{N_d}$ such that $v_n = \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} v_j$.*

Given Assumption 9, the linear span inclusion requirement as stated in Assumption 11 is rather weak. For example, consider the case that $span(\{v_j : j \in \mathcal{I}^{(d)}\}) = \mathbb{R}^r$. Since $v_n \in \mathbb{R}^r$, Assumption 11 would immediately hold. Note Theorem 4.6.1 of Vershynin (2018) implies that if $\{v_n\}_{n \in [N]}$ are sampled as independent, mean zero, sub-Gaussian vectors, then $\text{span}(\{v_j : j \in \mathcal{I}^{(d)}\}) = \mathbb{R}^r$ holds with high probability as $N^{(d)}$ grows. More intuitively, Assumption 11 requires that the intervention assignment $\mathcal{D}$ is such that sufficiently many units undergo intervention d, and their unit factors are collectively “diverse” enough so that their span includes the latent factor associated with any other unit. Now we state the formal identification result.

Theorem 3.5.1. *Given (n, d), let Assumptions 8 to 11 hold. Then, given $w^{(n,d)}$,*

$$\mathbb{E}[Y_{tn}^{(d)} | u_t^{(d)}, v_n] = \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}\left[Y_{tjd} | \mathcal{E}\right] \quad \text{for } t \in [T], \tag{3.13}$$

$$\theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{post}} \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}\left[Y_{tjd} | \mathcal{E}\right]. \tag{3.14}$$

■ 3.5.3 Finite-sample Consistency

Additional Assumptions. We state additional assumptions required to establish statistical guarantees for our estimation procedure.

Assumption 12 (Sub-Gaussian noise). *Conditioned on $\mathcal{E}$, for any (t, n, d), $\varepsilon_{tn}^{(d)}$ are zero-mean independent sub-Gaussian random variables with $\text{Var}[\varepsilon_{tn}^{(d)} | \mathcal{E}] = \sigma^2$ and $\varepsilon_{tn}^{(d)} | \mathcal{E}_{\psi_2} \leq C\sigma$ for some constant $C > 0$.*

Assumption 13 (Boundededness). *For any (t, n, d), $\mathbb{E}[Y_{tn}^{(d)} | \mathcal{E}] \in [-1, 1]$.*[3]

[3] The precise bound $[-1, 1]$ is without loss of generality, i.e., it can be extended to $[a, b]$ for $a, b \in \mathbb{R}$ with $a \leq b$.

Assumption 14 (Well-balanced spectra). *For any d, let $1 \leq r_{\text{pre}} \leq r$ be the rank of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$ and $s_1 \geq \dots s_{r_{\text{pre}}} > 0$ be its nonzero singular values. We assume the singular values are well-balanced, i.e., for universal constants $c, c' > 0$, $s_{r_{\text{pre}}}/s_1 \geq c$ with $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}]_F^2 \geq c' N_d T_0$.*

Assumption 15 (Subspace inclusion). *For any d, the rowspace of $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}|\mathcal{E}]$ lies within that of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$.*

Finite-sample consistency. Now we state the finite-sample guarantee which establishes that the estimator described in Section 3.3.1 yields a consistent estimate of the causal parameter for any given unit-intervention pair. To simplify notation, we will henceforth absorb dependencies on σ into the constant within $O_p(\cdot)$, defined in Section 3.2.

Theorem 3.5.2. *Given (n, d), let Assumptions 8 to 15 hold. Further, suppose $k = r_{\text{pre}} =$ $\text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}])$, where k is defined as in (3.3). Then conditioned on $\mathcal{E}$,*

$$\widehat{\theta}_n^{(d)} - \theta_n^{(d)} = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{T_0^{1/4}} + \frac{\widetilde{w}^{(n,d)}{}_2}{\sqrt{T_1}} + \frac{\widetilde{w}^{(n,d)}{}_1 r_{\text{pre}}^{3/2}\sqrt{\log(T_0 N_d)}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}}\right).$$

Here,

$$\widetilde{w}^{(n,d)} = \boldsymbol{V}_{\text{pre}}\boldsymbol{V}_{\text{pre}}' w^{(n,d)}, \tag{3.15}$$

where $\boldsymbol{V}_{\text{pre}} \in \mathbb{R}^{N_d \times r_{\text{pre}}}$ represents right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$ and $w^{(n,d)}$ is defined as in Theorem 4.3.1. We assume $\widetilde{w}^{(n,d)}{}_2 \geq c$, where $c > 0$ is a universal constant.[4]

■ 3.5.4 Asymptotic Normality

We establish that the estimate is asymptotically normal around the true causal parameter. This will justify the confidence interval defined in (3.6) of Section 3.3.

Theorem 3.5.3. *Given (n, d), let the setup of Theorem 4.5.1 hold. Let $\widetilde{w}^{(n,d)}$ be defined as in (3.15). Suppose (i) $T_0, T_1, N_d \to \infty$; (ii) $r_{\text{pre}}^2 \log(T_0 N_d) = o(\min\{T_0, N_d\})$; and (iii)*

$$T_1 = o\left(\widetilde{w}^{(n,d)2}_2 \cdot \min\left\{\frac{\sqrt{T_0}}{r_{\text{pre}}}, \frac{\min\{T_0, N_d\}}{\widetilde{w}^{(n,d)2}_1 r_{\text{pre}}^3 \log(T_0 N_d)}\right\}\right). \tag{3.16}$$

[4] We believe that the log factors within our results could be removed with careful analysis.

Then conditioned on $\mathcal{E}$,

$$\frac{\sqrt{T_1}}{\sigma \widetilde{w}^{(n,d)}{}_2}\left(\widehat{\theta}_n^{(d)} - \theta_n^{(d)}\right) \xrightarrow{d} \mathcal{N}(0,1)\,. \tag{3.17}$$

Further, for $\widehat{w}^{(n,d)}$ and $\widehat{\sigma}^2$ defined in (3.3) and (3.7), respectively, we have

$$\widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)} = O_p\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}}\sqrt{\log(T_0 N_d)}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}}\right), \tag{3.18}$$

$$\widehat{\sigma}^2 - \sigma^2 = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{\sqrt{T_0}} + \frac{r_{\text{pre}}\sqrt{\log(T_0 N_d)}\widetilde{w}^{(n,d)}{}_1}{\min\{\sqrt{T_0}, \sqrt{N_d}\}}\right). \tag{3.19}$$

■ 3.5.5 Discussion

On Assumptions 14 and 15

Assumption 14 requires that the nonzero singular values of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$ are well-balanced. Within the econometrics factor model analyses and matrix completion literature, Assumption 14 is analogous to incoherence-style conditions, e.g., Assumption A of Bai and Ng (2020). It is also closely related to the notion of pervasiveness (see Fan et al. (2018)). Additionally, Assumption 14 has been shown to hold with high probability for the canonical probabilistic generating process used to analyze probabilistic principal component analysis in Bishop (1999) and Tipping and Bishop (1999); here, the observations are assumed to be a high-dimensional embedding of a low-rank matrix with independent sub-Gaussian entries (see Proposition 4.2 of Agarwal et al. (2021e)). Lastly, we highlight that the assumption of a gap between the top few singular values of observed matrix of interest and the remaining singular values has been widely adopted in the econometrics literature of large dimensional factor analysis dating back to Chamberlain and Rothschild (1983).

Next, we discuss Assumption 15. The goal of the SI framework is to "causally transport" a model learned under one intervention and a set of outcome measurements to other interventions and measurements. This requires analyzing the "generalization" properties of the SI estimator, a term commonly used in the statistical learning literature. However, the potential outcomes under different interventions are likely to arise from different distributions, which makes analyzing when it generalizes much more challenging; traditional generalization analyses require making stringent distributional assumptions, e.g., all measurements are sampled i.i.d. Thus, to understand when such causal transportation is feasible, we need a different framework of generalization. Assumption 15 precisely provides such a framework; it is a purely linear algebraic requirement, and makes no distributional assumptions about the latent factors. Indeed, learning with distribution shifts is an active area of research in causal inference, machine learning, and statistics (e.g., learning with covariate shift, transfer learning). Our hope is that the SI framework, and Assumption 15 more specifically, might provide a meaningful way to think about generalization in such settings.

Learning $N \times D$ causal parameters with $O(N)$ observations

In this work, our interest is in estimating $N \times D$ causal parameters: $\theta_n^{(d)}$ for all (n, d). If all $Y_{tn}^{(d)}$ for $t \in \mathcal{T}_{\text{post}}$ were observed, their empirical mean would concentrate around $\theta_n^{(d)}$ with error scaling as $O(1/\sqrt{T_1})$. That is, to obtain an additive error of $O(\delta)$ for any $\delta > 0$, such an estimation procedure would require $T_1 = \Omega(\delta^{-2})$. Therefore, for all $N \times D$ pairs, it would require $N \times D \times O(\delta^{-2})$ observations.

Theorem 4.5.1 establishes that by observing $N \times T_0 + (\sum_{d=1}^{D} N_d) \times T_1$ observations in total, we can estimate $\theta_n^{(d)}$ for any n within error $O(\max(T_0^{-1/4}, T_1^{-1/2}, N_d^{-1/2}))$. As a special case, consider $N_d = N/D$. Then, for any $\delta > 0$, with $T_0 = \Omega(\delta^{-4})$, $T_1 = \Omega(\delta^{-2})$ and $N_d = \Omega(\delta^{-2})$, $\theta_n^{(d)}$ can be estimated within additive error $O(\delta)$ with observations $N \times O(\delta^{-4})$, independent of D. Thus, SI enables learning all $N \times D$ causal parameters using $O(N)$ observations, independent of D.

Relative scaling of N, T and D. Theorems 4.5.1 and 4.5.2 suggests that for SI to do meaningful estimation only two of the three dimensions, N, T need to be scaling relatively quickly compared to D; e.g. in the case where $N_d \to \infty$. However, depending on the relative scalings of N, T and D, variants of the SI algorithm might be more suitable. For example, it might be more apt to regress on interventions rather than units if $N = o(D)$. We explore this further in Section 3.9.

■ 3.6 A Hypothesis Test for Subspace Inclusion

In Section 3.6.1, we formally define the subspace inclusion hypothesis test; this serves as a robustness check of Assumption 15, which enables our learned model to be transferred across interventions and measurements. In Section 3.6.2, we apply this test to two seminal case studies from the SC literature and discuss our findings.

3.6.1 Hypothesis Test & Formal Results

Notation. Recall $r_{\text{pre}} = \text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}])$ and $r_{\text{post}} = \text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}])$. Further, recall $\boldsymbol{V}_{\text{pre}} \in \mathbb{R}^{N_d \times r_{\text{pre}}}$ denotes the right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$. Define $\boldsymbol{V}_{\text{post}} \in \mathbb{R}^{N_d \times r_{\text{post}}}$ with respect to $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$ analogously. Finally, let $\widehat{\boldsymbol{V}}_{\text{pre}} \in \mathbb{R}^{N_d \times r_{\text{pre}}}$ and $\widehat{\boldsymbol{V}}_{\text{post}} \in \mathbb{R}^{N_d \times r_{\text{post}}}$ denote their respective estimates, which are constructed from the top r_{pre} and r_{post} right singular vectors of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ and $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$, respectively.

Hypothesis Test

Consider hypotheses

$$H_0: \ \text{span}(\boldsymbol{V}_{\text{post}}) \subseteq \text{span}(\boldsymbol{V}_{\text{pre}}) \quad \text{and} \quad H_1: \ \text{span}(\boldsymbol{V}_{\text{post}}) \nsubseteq \text{span}(\boldsymbol{V}_{\text{pre}}).$$

Recall the test statistic $\widehat{\tau}$ as defined in (3.8) in Section 3.3.2, and the motivation for its usage. More formally, we define the test as follows: for any significance level $\alpha \in (0,1)$,

$$\text{Retain } H_0 \text{ if } \ \widehat{\tau} \leq \tau(\alpha) \quad \text{and} \quad \text{Reject } H_0 \text{ if } \ \widehat{\tau} > \tau(\alpha).$$

Here, $\tau(\alpha)$ is the critical value, which we define for some absolute constant $C \geq 0$:

$$\tau(\alpha) = \frac{C\sigma^2 r_{\text{post}} \phi^2_{\text{pre}}(\alpha/2)}{s^2_{r_{\text{pre}}}} + \frac{C\sigma^2 r_{\text{post}} \phi^2_{\text{post}}(\alpha/2)}{\varsigma^2_{r_{\text{post}}}} + \frac{C\sigma r_{\text{post}} \phi_{\text{pre}}(\alpha/2)}{s_{r_{\text{pre}}}}, \tag{3.20}$$

where $\phi_{\text{pre}}(a) = \sqrt{T_0} + \sqrt{N_d} + \sqrt{\log(1/a)}$; $\phi_{\text{post}}(a) = \sqrt{T_1} + \sqrt{N_d} + \sqrt{\log(1/a)}$; and s_ℓ, ς_ℓ are the ℓ-th singular values of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}|\mathcal{E}]$, respectively.

Type I and Type II Error Guarantees

Given our choice of $\widehat{\tau}$ and $\tau(\alpha)$, we control both Type I and Type II errors of our test.

Theorem 3.6.1. *Let Assumptions 8, 9, 10, 12 hold. Fix any $\alpha \in (0,1)$. Then conditioned on $\mathcal{E}$, there exists an absolute constant $C \geq 0$, defined in (3.20), such that the Type I error* *is bounded as $\mathbb{P}(\widehat{\tau} > \tau(\alpha) | H_0) \leq \alpha$. To bound the Type II error, suppose the following additional condition holds:*

$$r_{\text{post}} > \boldsymbol{V}_{\text{pre}} \boldsymbol{V}'_{\text{pre}} \boldsymbol{V}_{\text{post}}{}_F^2 + 2\tau(\alpha) + \frac{C \sigma r_{\text{post}} \phi_{\text{post}}(\alpha/2)}{\varsigma_{r_{\text{post}}}}. \tag{3.21}$$

Then, the Type II error is bounded as $\mathbb{P}(\widehat{\tau} \leq \tau(\alpha) | H_1) \leq \alpha$.

The particular C for which Theorem 3.6.1 holds depends on the underlying distribution of $\varepsilon_{tn}^{(d)}$, which determines the distribution of the potential outcomes $Y_{tn}^{(d)}$. C can be made explicit for certain classes of distributions; as an example, Corollary 3.6.1 specializes Theorem 3.6.1 to when $\varepsilon_{tn}^{(d)}$ are normally distributed.

Corollary 3.6.1. *Consider the setup of Theorem 3.6.1 with $C = 4$. Let $\varepsilon_{tn}^{(d)}$ be normally distributed for all (t, n, d). Then, $\mathbb{P}(\widehat{\tau} > \tau(\alpha) | H_0) \leq \alpha$ and $\mathbb{P}(\widehat{\tau} \leq \tau(\alpha) | H_1) \leq \alpha$.*

We now argue (3.21) is not a restrictive condition. Conditioned on H_1, observe that $r_{\text{post}} > \boldsymbol{V}_{\text{pre}} \boldsymbol{V}'_{\text{pre}} \boldsymbol{V}_{\text{post}}{}_F^2$ always holds. If Assumption 14 holds and the nonzero singular values of $\mathbb{E}[\boldsymbol{Y}_{\text{post}, \mathcal{I}^{(d)}}]$ are well-balanced, then one can easily verify that the latter two terms on the right-hand side of (3.21) decay to zero as T_0, T_1, N_d grow.

Computing $\tau(\alpha)$

Computing $\tau(\alpha)$ requires estimating (i) σ^2; (ii) $r_{\text{pre}}, r_{\text{post}}$; (iii) $s_{r_{\text{pre}}}, \varsigma_{r_{\text{post}}}$. We provide an estimator for σ in (3.7) and establish its consistency in Theorem 4.5.2. Further, recall that Lemma 3.13.1 establishes that the singular values of $\boldsymbol{Y}_{\text{pre}, \mathcal{I}^{(d)}}$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre}, \mathcal{I}^{(d)}}]$ must be close. Thus, we can use the spectra of $\boldsymbol{Y}_{\text{pre}, \mathcal{I}^{(d)}}$ as a good proxy to estimate r_{pre} and $s_{r_{\text{pre}}}$. Analogous arguments hold for $\boldsymbol{Y}_{\text{post}, \mathcal{I}^{(d)}}$ with respect to $\mathbb{E}[\boldsymbol{Y}_{\text{post}, \mathcal{I}^{(d)}}]$. Further, note that Corollary 3.6.2 specializes $\tau(\alpha)$ under Assumption 14.

Corollary 3.6.2. *Let the setup of Theorem 3.6.1 hold. Suppose Assumption 14 holds. Further, suppose that conditioned on $\mathcal{E}$, the r_{post} nonzero singular values ς_i of $\mathbb{E}[\boldsymbol{Y}_{\text{post}, \mathcal{I}^{(d)}}]$ are well-balanced, i.e., $\varsigma_i^2 = \Theta(T_1 N_d / r_{\text{post}})$. Then, $\tau(\alpha) = O\left(\frac{\sqrt{\log(1/\alpha)}}{\min\{\sqrt{T_0}, \sqrt{T_1}, \sqrt{N_d}\}}\right)$.*

If we consider the noiseless case, $\varepsilon_{tn}^{(d)} = 0$, we note that $\tau(\alpha) = 0$. More generally, if the spectra of $\mathbb{E}[\boldsymbol{Y}_{\text{pre}, \mathcal{I}^{(d)}}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{post}, \mathcal{I}^{(d)}}]$ are well-balanced, then Corollary 3.6.2 establishes that $\tau(\alpha) = o(1)$, even in the presence of noise. We remark that Corollary 3.6.1 allows for exact constants in the definition of $\tau(\alpha)$ under the Gaussian noise model.

A Practical Heuristic

Here, we provide a complementary approach to computing $\tau(\alpha)$, used in Squires et al. (2021). To build intuition, observe that $\widehat{\tau}$ represents the remaining spectral energy of $\widehat{V}_{\text{post}}$ not contained within span$(\widehat{V}_{\text{pre}})$. Further, we note $\widehat{\tau}$ is trivially bounded by r_{post} since the columns of $\widehat{V}_{\text{post}}$ are orthonormal. Thus, one can fix some fraction $\alpha \in (0, 1)$ and reject H_0 if $\widehat{\tau} > \tau(\alpha)$, where $\tau(\alpha) = \alpha \cdot r_{\text{post}}$. In words, if more than α fraction of the spectral energy of $\widehat{V}_{\text{post}}$ lies outside the span of $\widehat{V}_{\text{pre}}$, then the alternative test rejects H_0. We remark that this variant is likely more robust compared to its exact computation counterpart, which requires estimating several "nuisance" quantities described above in order to estimate (3.20) and varies with the underlying modeling assumptions we make about $\varepsilon_{tn}^{(d)}$ and the singular values, $s_{r_{\text{pre}}}, \varsigma_{r_{\text{post}}}$. As such, we employ the heuristic variant in our case studies presented in Section 4.6.3.

3.6.2 Subspace Inclusion Test Applied to Synthetic Controls

We revisit two seminal case studies within the SC literature: (i) an evaluation on the impact of terrorism in Basque Country Abadie and Gardeazabal (2003); (ii) an evaluation on the impact of California's Proposition 99 on tobacco consumption Abadie et al. (2010). In particular, we apply our subspace inclusion hypothesis test to study the potential feasibility of counterfactual inference in both studies. Indeed, these studies have been used extensively to explain the utility of the SC method and have subsequently served as benchmarks in following works, cf. Amjad et al. (2018); Athey et al. (2021); Arkhangelsky et al. (2020); Agarwal et al. (2021e), and more broadly as well. As such, we hope our findings not only motivate the usage of this test, but also spark the development of additional robustness tests to stress test the causal conclusions drawn from these studies and beyond.

Terrorism in Basque Country

Background & setup. In 1968, the first Basque Country victim of terrorism was claimed; however, it was not until the mid-1970s did the terrorist activity become more rampant Abadie and Gardeazabal (2003). To study the economic ramifications of terrorism on Basque Country, we use the per-capita GDP of $N = 18$ Spanish regions from 1955-1997, i.e., $T = 43$. Among these regions, there is one treated region, Basque Country, which experienced terrorism; the other 17 regions are considered control regions as they were relatively unaffected by terrorism. Translating to our notation, we have $D = 2$ with $d = 0$ representing control and $d = 1$ representing treatment (i.e., terrorism); in turn, we have $T_0 = 14$ and $T_1 = 29$. Without loss of generality, we index Basque Country as unit $n = 1$. Our interest is to estimate $\theta_1^{(0)}$.

We note that the original work of Abadie and Gardeazabal (2003) uses 13 additional predictor variables for each region, including demographic information pertaining to one's educational status, and average shares for six industrial sectors. We do not utilize this additional information here, i.e., we only use information related the outcome of interest, i.e., per-capita GDP.

Hypothesis test results. We apply our heuristic variant of the hypothesis test, introduced in Section 3.6.1, with $\alpha = 0.05$. We find that our test statistic $\widehat{\tau} = 0.01$ and $\tau(\alpha) = 0.05$. Since $\widehat{\tau} < \tau(\alpha)$, our tests suggest that the SI method is suitable for this study, and also supports the suitability of and conclusions drawn from previous works utilizing SC-like methods for this case study.

California Proposition 99

Background & setup. In 1988, California introduced the first modern-time large-scale anti-tobacco legislation in the United States Abadie et al. (2010). To analyze the effect of California's anti-tobacco legislation, we use the annual per-capita cigarette consumption at the state level for $N = 39$ states from 1970-2000, i.e., $T = 31$. With the exception of California, the other 38 states included in this study neither adopted an anti-tobacco program or raised cigarette sales taxes by 50 cents or more. As such, these states are considered the control states and California is considered the treated state. Translating to our notations, we have $D = 2$ with $d = 0$ representing control and $d = 1$ representing treatment (i.e., Proposition 99); in turn, we have $T_0 = 29$ and $T_1 = 12$. Without loss of generality, we index California as unit $n = 1$. Our interest is to estimate $\theta_1^{(0)}$.

It is worth noting that the original work in Abadie et al. (2010) uses six additional covariates per state, e.g., retail price, beer consumption per capita, and percentage of individuals between ages of 15-24. We do not include these variables in our study.

Hypothesis test results. We apply our heuristic variant of the hypothesis test with

$\alpha = 0.05$. We find that our test statistic $\widehat{\tau} = 1.64$ and $\tau(\alpha) = 0.15$. Since $\widehat{\tau} > \tau(\alpha)$, our test suggests that the SI estimator is likely ill-suited to produce a reliable estimate of $\theta_1^{(0)}$; in fact, our test only passes for $\alpha \geq 0.48$. Since the robust synthetic controls estimator of Amjad et al. (2018); Agarwal et al. (2021e) is a special case of the SI estimator, this puts into question the conclusions drawn in Amjad et al. (2018); Agarwal et al. (2021e). More generally, this may shed some doubt on the suitability of any linear predictor, including the original SC estimator and many of its variants (e.g. Athey et al. (2021); Arkhangelsky et al. (2020)), used to investigate the impact of Proposition 99. Thus, we believe this result may be of independent interest to the SC literature to revisit the conclusions drawn from these prior works and to further extend the toolkit of robustness/sensitivity analysis tests of when one can transfer a model across time periods (and interventional regimes).

■ 3.7 Simulations

In this section, we present illustrative simulations to reinforce our theoretical results. In particular, these simulations suggest that the finite-sample estimation error bounds are tight, asymptotic Gaussian approximation is accurate, and the assumptions behind our theoretical results are necessary. Below, we present a brief overview of the setup for each simulation as well as the primary takeaway, and relegate the details to Section 3.11.

■ 3.7.1 Consistency

The purpose of this section is to study the finite-sample properties of the SI estimator.

Setup. We generate $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$, $w^{(n,d)}$, $\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]w^{(n,d)}$, and $\theta_n^{(d)} = (1/T_1)\sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}[Y_{tj}^{(d)}]w_j^{(n,d)}$ in such a way that our operating assumptions hold; namely, Assumptions 9, 11, 13, 14, 15. Further, we sample $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$, $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$, and $Y_{\text{pre},n}$ while respecting Assumptions 8, 10, 12. Our objective is to estimate $\theta_n^{(d)}$ from $(\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}, \boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}, Y_{\text{pre},n})$. In order to showcase the error rate of the SI estimator, we vary the length of $\mathcal{T}_{\text{post}}$ while keeping $T_0 = N_d$ fixed. That is, we choose $T_1 = 200$ and vary the number of post-intervention samples as ρT_1, where $\rho \in \{0.1, 0.2, \dots, 1.0\}$. As such, we index our causal parameter as $\theta_n^{(d)}(\rho)$ and estimates as $\widehat{\theta}_n^{(d)}(\rho)$.

Results. We perform 100 iterations for each ρ and plot the mean absolute errors (MAEs),

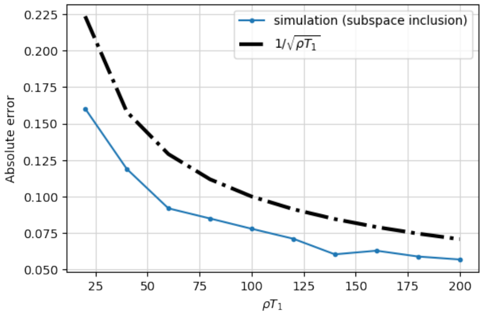


Figure 3.10: Plot of mean absolute errors across 100 iterations for every ρ, which corresponds to different post-intervention lengths. As implied by Theorem 4.5.1, the error decays as $O_p(1/\sqrt{\rho T_1})$.

$|\widehat{\theta}_n^{(d)}(\rho) - \theta_n^{(d)}(\rho)|$, in Figure 3.10. As the figure shows, the MAE of $\widehat{\theta}_n^{(d)}(\rho)$ decays as the post-intervention period ρT_1 increases. Moreover, given the choice of $T_0 = N_d$, the error effectively scales as $O_p(1/\sqrt{\rho T_1})$, which matches the implication of Theorem 4.5.1. Therefore, this simulation demonstrates that the estimator described in Section 3.3 produces a consistent estimate of the underlying causal parameter if Assumptions 8 to 15 hold.

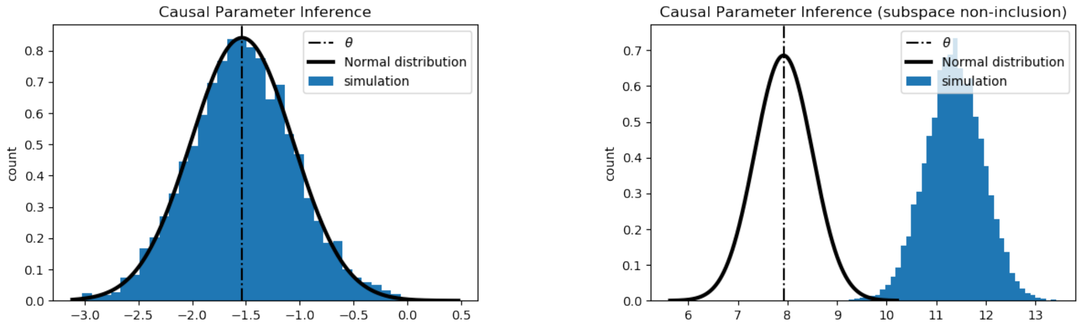
■ 3.7.2 Asymptotic Normality

In this section, we study the asymptotic normality properties of the SI estimator, as well as the importance of subspace inclusion (Assumption 15).

Subspace Inclusion Holds

In the following simulation, we will enforce Assumption 15 to hold between the pre- and post-intervention data. However, we will allow the pre- and post-intervention data to be sampled from different distributions.

Setup. We consider a binary intervention model $D = 2$, where the pre-intervention data will be observed under control $d = 0$, while the post-intervention data will be observed under treatment $d = 1$. In order to separate treatment from control, we will sample $\{Y_{tj}^{(0)}\}$ and $\{Y_{tj}^{(1)}\}$ from different distributions. Next, we generate $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}] = [Y_{tj}^{(0)} : t \in \mathcal{T}_{\text{pre}}, j \in \mathcal{I}^{(1)}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}] = [Y_{tj}^{(1)} : t \in \mathcal{T}_{\text{post}}, j \in \mathcal{I}^{(1)}]$, $w^{(n,1)}$, $\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]w^{(n,1)}$, and $\theta_n^{(1)} = (1/T_1)\sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}[Y_{tj}^{(1)}]w_j^{(n,1)}$ in such a way that our operating assumptions



(a) Assumption 15 holds.

(b) Assumption 15 fails.

Figure 3.11: In 3.11a, the histogram of estimates follows the theoretical asymptotic normal distribution. In 3.11b, the histogram of estimates does not follow the theoretical asymptotic normal distribution.

hold; namely, Assumptions 9, 11, 13, 14, 15. Further, we sample $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}$, $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$, and $Y_{\text{pre},n}$ while respecting Assumptions 8, 10, 12. Our objective is to estimate $\theta_n^{(1)}$ from $(\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}, \boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}, Y_{\text{pre},n})$. We re-emphasize that $(\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}, Y_{\text{pre},n})$ used to learn the model follow a different distribution from $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$ used for predictions; however, subspace inclusion between the pre- and post-intervention is upheld.

Results. We run 5000 iterations and display the histogram of estimates $\widehat{\theta}_n^{(1)}$ in Figure 3.11a. As implied by Theorem 4.5.2, the histogram is very well-approximated by a normal distribution with mean $\theta_n^{(1)}$ and variance $(1/\sqrt{T_1})\sigma^2\|\widetilde{w}^{(n,1)}\|_2^2$. This figure demonstrates that even if the pre- and post-intervention potential outcomes follow different distributions, our estimates remain normally distributed around the true causal parameter. That is, it is feasible to learn $\widehat{w}^{(n,d)}$ under one intervention setting (e.g., control), and then transfer the learned model to a different intervention regime, which may obey a different distribution, provided subspace inclusion holds.

Subspace Inclusion Fails

Next, we show $\widehat{\theta}_n^{(d)}$ is non-trivially biased when Assumption 15 fails.

Setup. We again consider a binary intervention model $D = 2$, where the pre-intervention data will be observed under control $d = 0$, while the post-intervention data will be observed under treatment $d = 1$. In order to separate treatment from control, we will sample $\{Y_{tj}^{(0)}\}$

and $\{Y_{tj}^{(1)}\}$ from different distributions. Next, we generate $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}] = [Y_{tj}^{(0)} : t \in \mathcal{T}_{\text{pre}}, j \in \mathcal{I}^{(1)}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}] = [Y_{tj}^{(1)} : t \in \mathcal{T}_{\text{post}}, j \in \mathcal{I}^{(1)}]$, $w^{(n,1)}$, $\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]w^{(n,1)}$, and $\theta_n^{(1)} = (1/T_1)\sum_{t\in\mathcal{T}_{\text{post}}} \mathbb{E}[Y_{tj}^{(1)}]w_j^{(n,1)}$ in such a way that Assumptions 9, 11, 13, 14 hold. Crucially, however, we will violate Assumption 15. Further, we sample $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}$, $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$, and $Y_{\text{pre},n}$ while respecting Assumptions 8, 10, 12. Our objective is to estimate $\theta_n^{(1)}$ from $(\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}, \boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}, Y_{\text{pre},n})$.

Results. We perform 5000 iterations and plot the histogram of estimates in Figure 3.11b. Unlike Figure 3.11a, the histogram is not well-approximated by the normal distribution with mean $\theta_n^{(1)}$ but instead has non-trivial bias. The juxtaposition of these two figures reinforces the importance of Assumption 15 for successful counterfactual inference (i.e., generalization).

■ 3.8 Comparison with Synthetic Controls Literature

The goal of this section is to better contextualize our assumptions and results within the SC literature, by doing a detailed comparison with some representative works.

Causal parameter in SC: treatment effect on the treated. For any (n, d) pair, let $\bar{Y}_n^{(d)} = \frac{1}{T_1}\sum_{t\in\mathcal{T}_{\text{post}}} Y_{tn}^{(d)}$. In addition, let $\tau_n^{(d_1,d_2)} = \bar{Y}_n^{(d_1)} - \bar{Y}_n^{(d_2)}$ denote the (relative) treatment effect between interventions d_1 and d_2 for unit n, averaged over the post-intervention period. The most closely related causal parameter in the SC literature to our work is $\tau_n^{(d,0)}$ for $n \in \mathcal{I}^{(d)}$ and $d \neq 0$. This is referred to as the unit specific treatment effect on the treated, averaged over the post-intervention period. Recall $\bar{Y}_n^{(d)}$ is observed for $n \in \mathcal{I}^{(d)}$. As such, the goal in these works is to estimate the counterfactual potential outcomes $\bar{Y}_n^{(0)}$ for $n \notin \mathcal{I}^{(0)}$, i.e., what would have happened to a "treated" unit n had it remained under control. The quantity $\bar{Y}_n^{(0)}$ is closely related to $\theta_n^{(0)}$; the slight difference is that $\theta_n^{(0)} = \mathbb{E}[\bar{Y}_n^{(0)} | \{u_t^{(0)}, v_n : t \in \mathcal{T}_{\text{post}}\}]$, where the expectation is taken with respect to the mean zero residual term and conditioned on the latent factors. Many of the works listed below *implicitly* condition on either the latent factors or directly on the observations Y_{tn}, if a factor model is not assumed.

We re-emphasize the SI framework allows for identification and inference of $\theta_n^{(d)}$ for $d \neq 0$ and all $n \in [N]$, which these previous works do not consider. This enables some of the key applications we study, such as synthetic A/B testing described in Section 4.6.3.

Representative works. Arguably, $\tau_n^{(d,0)}$ is the most common casual parameter considered in the SC literature, e.g., see Hsiao et al. (2012); Doudchenko and Imbens (2016b); Athey et al. (2021); Li and Bell (2017); Xu (2017b); Li (2018); Bai and Ng (2020); Chan and Kwok (2020); Chernozhukov et al. (2020b); Fernández-Val et al. (2020). We restrict our attention to four recent works, Bai and Ng (2020), Chernozhukov et al. (2020b), Arkhangelsky et al. (2020), and Agarwal et al. (2021e) as they highlight some of the primary points of comparison of SI with the SC literature. Again, given the vastness of the literature, we underscore that these comparisons are by no means exhaustive.[5]

3.8.1 Comparison with Chernozhukov et al.

Overview of assumptions. There are two key assumptions made in Chernozhukov et al. (2020b). First, the authors assume the existence of $w^{(n,0)} \in \mathbb{R}^{N_0}$ such that for all time t and conditioned[6] on any sampling of $\{Y_{tj}^{(0)} : j \in \mathcal{I}^{(0)}\}$,

$$Y_{tn}^{(0)} = \sum_{j \in \mathcal{I}^{(0)}} w_j^{(n,0)} Y_{tj}^{(0)} + \varepsilon_{tn}, \tag{3.22}$$

where $\mathbb{E}[\varepsilon_{tn}] = 0$ and $\mathbb{E}[\varepsilon_{tn} Y_{tj}^{(0)}] = 0$. Second, they assume the existence of an oracle estimator for $w^{(n,0)}$, denoted as $\widehat{w}^{(n,0)}$, such that $w^{(n,0)} - \widehat{w}^{(n,0)}{}_2 = o(1)$.

The two most common restrictions placed on $w^{(n,0)}$, for which there exist estimators with formal performance guarantees, include (i) $w^{(n,0)}$ is convex, i.e., $\sum_{j \in \mathcal{I}^{(0)}} w_j^{(n,0)} = 1$ and $w_j^{(n,0)} \geq 0$, cf. Abadie et al. (2010); Abadie and Gardeazabal (2003); Doudchenko and Imbens (2016b); (ii) $w^{(n,0)}$ is approximately sparse, i.e., $w^{(n,0)}{}_1 = O(1)$, which is a relatively weaker assumption, cf. Raskutti et al. (2011); Chernozhukov et al. (2020a). Under these assumptions, the authors provide a flexible and practical t-test based inference procedure to estimate $\bar{Y}_n^{(0)}$, which utilizes the oracle estimator $\widehat{w}^{(n,0)}$. We compare these assumptions with the SI framework.

Functional form in (3.22). Note that under Assumption 9, we can equivalently write

[5] A related causal parameter considered in many of these works is the time specific treatment effect on the treated, averaged over all treated units. This reduces to estimating $\bar{Y}_t^{(0)} = (N - N_0)^{-1} \sum_{n \notin \mathcal{I}^{(0)}} Y_{tn}^{(0)}$ for a particular $t > T_0$. Estimating this quantity is quite similar to estimating $\bar{Y}_n^{(0)}$. In particular, we can estimate $\bar{Y}_t^{(0)}$ by simply transposing the observations and then running the same estimator used to estimate $\bar{Y}_n^{(0)}$.

[6] The conditioning on $\{Y_{tj}^{(0)} : j \in \mathcal{I}^{(0)}\}$ is implicit in Chernozhukov et al. (2020b).

Assumption 11 as follows:

$$\mathbb{E}[Y_{tn}^{(d)}] = \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}[Y_{tj}^{(d)}] \quad \text{for all } d. \tag{3.23}$$

Compared to (3.22) and for $d = 0$, we operate under a weaker functional form assumption as given in (3.23). To see this, simply take expectations on both sides of (3.22) with respect to ε_{tn}, which implies (3.23). That is, we do not require the linear relationship $w_j^{(n,0)}$ to hold between the *noisy* potential outcomes $Y_{tn}^{(0)}$ and $\{Y_{tj}^{(0)} : j \in \mathcal{I}^{(0)}\}$. Rather, we make the weaker assumption that the relationship only holds between the *expected* potential outcomes.

Restrictions on $w^{(n,0)}$. As discussed, theoretical guarantees in previous works require $w^{(n,0)}$ to be approximately sparse, or even more stringently, convex. Further, the estimators used to learn $w^{(n,0)}$ require *explicit knowledge* of the restriction placed on $w^{(n,0)}$. For example, if $w^{(n,0)}$ is assumed to be convex, then convex regression is used; if $w^{(n,0)}$ is assumed to be approximately sparse, then a ℓ_1-norm regularizer is used. In comparison, the estimator we propose to learn $w^{(n,0)}$ in (3.3) does not impose any such restriction on $w^{(n,0)}$, even in the high-dimensional setting where $N_0 > \max\{T_0, T_1\}$. Indeed, Lemma 4.9.1 establishes that our estimator consistently learns the *unique* minimum ℓ_2-norm $w^{(n,d)}$, denoted as $\widetilde{w}^{(n,d)}$. Further, our consistency and asymptotic normality results *implicitly* scale with the ℓ_1- and ℓ_2-norm of $\widetilde{w}^{(n,d)}$. In particular, the error in our consistency result of Theorem 4.5.1 implicitly scales with $\widetilde{w}^{(n,d)}{}_1$ and the variance in our asymptotic normality result of Theorem 4.5.2 implicitly scales with $\widetilde{w}^{(n,d)}{}_2$.

We do, however, assume that there exists a latent factor model between the potential outcomes (Assumption 9); such an assumption is not made in Chernozhukov et al. (2020b). To overcome high-dimensionality, where $w^{(n,d)}$ is not uniquely specified, the estimator in (3.3) directly exploits the low-rank structure induced by this factor model in Assumption 9.

■ 3.8.2 Comparison with Bai et al.

Overview of assumptions. In Bai and Ng (2020), they consider the following factor model:

$$\begin{aligned} Y_{tn}^{(0)} &= \langle x_{tn}, \beta \rangle + \langle F_t, \Lambda_n \rangle + \varepsilon_{tn} \qquad (3.24)\\ Y_{tn}^{(d)} &= \alpha_{tn}^{(d)} + Y_{tn}^{(0)}, \ \text{ for } d \neq 0. \end{aligned}$$

Here, $\Lambda_n \in \mathbb{R}^{r_1}$ is a unit n specific latent factor and $F_t \in \mathbb{R}^{r_1}$ is a time t specific latent factor associated with intervention $d = 0$ (control); $x_{tn} \in \mathbb{R}^k$ is a vector of *observed covariates*; and $\beta \in \mathbb{R}^k$ is a latent factor acting on x_{tn}, which crucially is invariant across (t, n). The authors make appropriate assumptions on the factor loadings $\boldsymbol{\Lambda} = [\Lambda_n] \in \mathbb{R}^{r_1 \times N}$ and $\boldsymbol{F} = [F_t] \in \mathbb{R}^{r_1 \times T}$, and on the residual term ε_{tn}. See Assumptions A to D in Bai and Ng (2020) for details. In essence, these assumptions require (i) $\boldsymbol{\Lambda}, \boldsymbol{F}$ to satisfy an incoherence-like condition, which implies a bounded operator norm (e.g., $\mathbb{E}[\boldsymbol{F}^4_{\text{op}}], \boldsymbol{\Lambda}_{\text{op}} \leq M$, where M does not scale with N_0, T); (ii) each of the r singular vectors of $\boldsymbol{\Lambda}, \boldsymbol{F}$ are identifiable (e.g., the non-zero singular values of $\boldsymbol{\Lambda}, \boldsymbol{F}$ are distinct); (iii) ε_{tn} across (t, n) have sufficiently light tails (e.g., $\mathbb{E}[|\varepsilon_{tn}|^8] < M$) and are weakly correlated.

Further, the authors propose an estimator composed of the following main steps: (i) estimate $\widehat{\beta}$ using the observations $\{Y_{tn}, x_{tn}\}$ across (t, n); (ii) estimate $\widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{F}}$ using the residuals of $\{Y_{tn} - x_{tn}\widehat{\beta}\}$; (iii) estimate $\bar{Y}^{(0)}_n$ using $\{x_{tn}, \widehat{\beta}, \widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{F}}\}$. Below, we compare their primary assumptions with that of the SI framework.

Factor model in (3.24). First, we compare the factor model assumed in (3.24) with that in (3.11) of Assumption 9. Under (3.24), if we further assume that $\langle x_{tn}, \beta \rangle$ admits a latent factorization given by $\langle x_{tn}, \beta \rangle = \langle \tilde{x}_t, \tilde{x}_n \rangle$ with $\tilde{x}_t, \tilde{x}_n \in \mathbb{R}^{k_1}$, then we can write (3.24) as a special case of (3.11). To see this, let $d = 0$ and define $u^{(0)}_t, v_n \in \mathbb{R}^r$ in (3.11) as

$$u^{(0)}_t = (F_t, \tilde{x}_t), \;\; v_n = (\Lambda_n, \tilde{x}_n), \tag{3.25}$$

where $r = r_1 + k_1$. We stress that we do not require access to $\tilde{x}_t$ or $\tilde{x}_n$ to express the model in Bai and Ng (2020) as an instance of (3.11); instead, we require that such a factorization of $\langle x_{tn}, \beta \rangle$ *implicitly* exists. However, if one does not assume such a latent factorization, then first learning β using observed covariates and then estimating $\boldsymbol{\Lambda}, \boldsymbol{F}$ via the residuals, is likely necessary. In Section 3.10, we show how to incorporate covariates in SI.

In addition, if one wants to estimate $\{\theta^{(d)}_n : d \neq 0\}$ in the SI framework (which we underscore is not considered in the SC literature), we require the added assumption that $\alpha^{(d)}_n$ factorizes into a (time, intervention) specific and unit specific latent factor, i.e.,

$$\alpha^{(d)}_{tn} = \langle \tilde{\alpha}^{(d)}_t, \tilde{\alpha}_n \rangle, \tag{3.26}$$

where $\tilde{\alpha}_t^{(d)}, \tilde{\alpha}_n \in \mathbb{R}^{r_2}$. Then, for $d \neq 0$, we can define $u_t^{(d)}, v_n \in \mathbb{R}^r$ in (3.11) as

$$u_t^{(d)} = (F_t, \tilde{x}_t, \tilde{\alpha}_t^{(d)}), \;\; v_n = (\Lambda_n, \tilde{x}_n, \tilde{\alpha}_n), \tag{3.27}$$

where $r = r_1 + r_2 + k_1$.[7]

Assumption on Λ, F. Since our ultimate goal is to estimate $\theta_n^{(d)}$, it is not necessary to accurately estimate either of the latent factor loadings $(\boldsymbol{\Lambda}, \boldsymbol{F})$. In fact, our proposed estimator in Section 3.3 explicitly circumvents estimating these two quantities. In comparison, Bai and Ng (2020) establish their theoretical guarantees by requiring explicit bounds on the estimation qualities of $(\boldsymbol{\Lambda}, \boldsymbol{F})$, which naturally require making more stringent assumptions on the latent factor loadings; for example, they require each singular vector of $(\boldsymbol{\Lambda}, \boldsymbol{F})$ to be identifiable, which itself requires all of the non-zero singular values to be distinct.

■ 3.8.3 Comparison with Arkhangelsky et al. and Agarwal et al.

We compare with these recent works together as they both consider approximately low-rank factor models for the expected potential outcomes, a generalization of what we do.

Comparison with Arkhangelsky et al. (2020). They consider a binary intervention model:

$$\begin{aligned} \boldsymbol{Y}^{(0)} &= \boldsymbol{L} + \boldsymbol{E}, \\ \boldsymbol{Y}^{(1)} &= \boldsymbol{Y}^{(0)} + \boldsymbol{A}, \end{aligned}$$

where $\boldsymbol{Y}^{(0)} = [Y_{tn}^{(0)}] \in \mathbb{R}^{T \times N}$ is the matrix of potential outcomes under control; $\boldsymbol{L} = [L_{tn}] \in \mathbb{R}^{T \times N}$ encodes the latent factor model under control (in our notation, $L_{tn} = \langle u_t^{(0)}, v_n \rangle$); $\boldsymbol{E} = [\varepsilon_{tn}] \in \mathbb{R}^{T \times N}$ encodes the mean zero residuals; $\boldsymbol{Y}^{(1)} = [Y_{tn}^{(1)}] \in \mathbb{R}^{T \times N}$ is the matrix of potential outcomes under $d = 1$; and $\boldsymbol{A} = [\alpha_{tn}] \in \mathbb{R}^{T \times N}$ encodes the treatment effect. The authors propose the "synthetic difference-in-differences" estimator for $\frac{1}{T_1(N-N_0)} \sum_{t \in \mathcal{T}_{\text{post}}} \sum_{n \notin \mathcal{I}^{(0)}} \alpha_{tn} | \boldsymbol{L}$, and establish rigorous asymptotic normality results.

[7] In the SC literature, a closely related factor model is $Y_{tn}^{(0)} = \langle \beta_t, x_n \rangle + \langle F_t, \Lambda_n \rangle + \varepsilon_{tn}$ and $Y_{tn}^{(d)} = \alpha_{tn}^{(d)} + Y_{tn}^{(0)}$ for $d \neq 0$, cf. Abadie et al. (2010); Abadie and Gardeazabal (2003). Here, $\beta_t \in \mathbb{R}^k$ is a latent factor, $x_n \in \mathbb{R}^k$ is an observed unit specific covariate, and $F_t, \Lambda_n \in \mathbb{R}^{r_1}$ and $\alpha_{tn}^{(d)} \in \mathbb{R}$ are defined as in (3.25). As such, similar to (3.25), we can write this factor model as a special case of (3.11), where $u_t^{(0)} = (\beta_t, F_t)$ and $v_n = (x_n, \Lambda_n)$. Again, if one wants to estimate $\theta_{tn}^{(d)}$, then we need to assume an additional factorization of $\alpha_{tn}^{(d)}$, just as in (3.26) and (3.27).

If we continue to restrict our attention to the estimation of counterfactuals under control, then the two primary differences between this work and Arkhangelsky et al. (2020) are the assumptions made on the (i) spectral profile of the factor models and (ii) relationship between the target and donors. In particular, we consider a low-rank $\boldsymbol{L}$ (Assumption 9) with a well-balanced spectra (Assumption 14); the latter effectively assumes a lower bound on the smallest non-zero singular value of $\boldsymbol{L}$. In comparison, Arkhangelsky et al. (2020) makes a weaker assumption that $\boldsymbol{L}$ is *approximately* low-rank, i.e., its $\sqrt{\min\{T_0, N_0\}}$-th singular value is sufficiently small (see Assumption 3 of their work). However, as in Abadie et al. (2010), they require a convex relationship to hold between the target and donors, while we make the weaker assumption that a linear relationship holds (see the discussion under Assumption 11 for why a linear relationship is directly implied by a low-rank factor model).

Comparison with Agarwal et al. (2021e). They analyze the recently proposed "robust SC" (RSC) estimator of Amjad et al. (2018), which is closely related to our proposed estimator in Section 3.3 if restricted to estimating $\theta_n^{(0)}$ for $n \notin \mathcal{I}^{(0)}$. The also consider an approximately low-rank $\boldsymbol{L}$; specifically, they consider two natural generating processes that induce an approximately low-rank $\boldsymbol{L} = [L_{tn}]$: (i) the spectra of $\boldsymbol{L}$ is geometrically decaying (i.e., the k-th singular value of $\boldsymbol{L}$, denoted as s_k, obeys $s_k = s_1 c^{k-1}$, where $c \in (0,1)$; (ii) $\boldsymbol{L}$ follows a generalized factor model, i.e., $L_{tn} = g(\rho_t, \omega_n)$, where $g(\cdot,\cdot)$ is a latent, Hölder continuous function, and $\rho_t, \omega_n \in [0,1]^\ell$ are latent variables associated with time t and unit n (an exact low-rank model is a special case of this where g is bi-linear). Agarwal et al. (2021e) establish finite-sample consistency for the RSC estimator with respect to a similar causal parameter as that of $\theta_n^{(0)}$ (see Section 4 of Agarwal et al. (2021e) for details). However, they do not provide asymptotic normality results for their target causal parameter.

Future directions. An interesting future research direction is to build upon these works to study the trade-offs between the assumptions placed on the latent factor model (e.g., spectral profile), relationship between the target and donors, and subsequent inference guarantees one can establish for various target causal parameters (e.g., extending the model and estimator of Arkhangelsky et al. (2020) to estimate $\theta_n^{(d)}$ for any $n \notin \mathcal{I}^{(d)}$ and $d \neq 0$). Further, another direction that would be of interest is staggered adoption, where the duration of the pre-intervention period can vary amongst the units, i.e., T_0 is specific to each unit. This setup for example was considered in Athey et al. (2021).

3.8.4 Statistical Guarantees

As stated earlier, the works listed at the beginning of Section 3.8 use some combination (and/or variation) of the assumptions discussed in Sections 3.8.1 to 3.8.3 to prove formal statistical guarantees. To prove asymptotic normality, the works of Chernozhukov et al. (2020b); Bai and Ng (2020); Arkhangelsky et al. (2020), make additional assumptions on the relative scalings between T_0, T_1, N_0. We also make similar assumptions but with respect to N_d, rather than N_0, to estimate $\widehat{\theta}_n^{(d)}$ for $d \neq 0$. In particular, Chernozhukov et al. (2020b) requires (i) $T_0, T_1 \to \infty$; (ii) $T_0/T_1 \to c_0$, where $c_o \in [0, \infty]$. Bai and Ng (2020) requires (i) $T_0, T_1, N \to \infty$; (ii) $\sqrt{N}/\min\{N_0, T_0\} \to 0$; and (iii) $\sqrt{T_0 + T_1}/\min\{N_0, T_0\} \to 0$. To estimate $\theta_n^{(d)}$, we require (i) $T_0, T_1, N_d \to \infty$; (ii) $T_1/\min\{N_d, \sqrt{T_0}\} \to 0$.

Arkhangelsky et al. (2020) allows for the additional flexibility that they only require the product $(N - N_0)T_1 \to \infty$. As an implication, $N - N_0$ can be a constant as long as T_1 grows; hence, if $N - N_0 = 1$, then this is equivalent to $\theta_n^{(0)}$. On the other extreme, they allow for T_1 to be fixed as long as $N - N_0$ grows; one can verify that our results hold under this setting by using the estimate $\frac{1}{N-N_0} \sum_{n \notin \mathcal{I}^{(0)}} \sum_{j \in \mathcal{I}^{(0)}} \widehat{w}_j^{(n,0)} Y_{tj}$. To establish their results, Arkhangelsky et al. (2020) assume certain relative scalings between $T_0, T_1, N_0, (N - N_0)$ (see Assumption 2 of Arkhangelsky et al. (2020)), which require that $T_1(N - N_0)$ do not grow "too quickly" compared to T_0 and N_0. We note that finding the optimal relative scalings between T_0, T_1, N_0 (or N_d) remains interesting future work.

3.9 Causal Inference & Tensor Completion

In this section, we re-interpret the classical potential outcomes framework through the lens of tensors. Specifically, consider an order-3 tensor with axes that correspond to measurements, units, and interventions. Each entry of this tensor is associated with the potential outcome for a specific measurements, unit, and intervention. Recall Figure 3.1 for a graphical depiction of this tensor. Therefore, estimating unobserved potential outcomes, the fundamental task in causal inference, is equivalent to estimating various missing entries of this order-3 potential outcomes tensor. Recall from Figure 3.2 how different observational and experimental studies that are prevalent in causal inference can be equivalently posed as tensor completion with different sparsity patterns. Indeed, imputing entries of a tensor that are noisy and/or missing is the goal of tensor completion (TC).

In Section 3.9.1, we discuss how important concepts in causal inference have a related notion in tensor completion. In Section 3.9.2, we show how low-rank tensor factor models, prevalent in the TC literature can guide future algorithmic design for causal inference. In Section 3.9.3, we pose what algorithmic advances are required in the TC literature to allow it to more directly connect with causal inference.

■ 3.9.1 Encoding Causal Inference as Tensor Completion

Causal parameters and TC error metrics. Here, we discuss the relationship between causal inference with different target causal parameters and TC under different error metrics. The first step in causal inference is to define a target causal parameter, while the first step in tensor completion is to define an error metric between the underlying and estimated tensors. Below, we discuss a few important connections between these two concepts.

To begin, consider as the causal parameter the average potential outcome under intervention d across all T measurements and N units.[8] Then, estimating this parameter can equivalently be posed as requiring a tensor completion method with a Frobenius-norm error guarantee for the d-th slice of the potential outcomes tensor with dimension $T \times N$, normalized by $1/\sqrt{TN}$. As such, a uniform bound for this causal parameter over all D interventions would in turn require a guarantee over the max (normalized) Frobenius-norm error for each of the D slices. Another causal parameter is unit n's potential outcome under intervention d averaged over all T measurements (recall, this is $\theta_n^{(d)}$). Analogously, this translates to the ℓ_2-norm error guarantee of the n-th column of the d-th tensor slice, normalized by $1/\sqrt{T}$. A uniform bound over all N units for the d-th intervention would then correspond to a $\ell_{2,\infty}$-norm error (defined in (3.47)) for the d-th tensor slice. As a final example, let the target causal parameter be unit n's potential outcome under intervention d and measurement t. This would require a TC method with a max-norm (entry-wise) error of the d-th matrix slice. Similar as above, a uniform bound over all measurements, units, and interventions corresponds to a max-norm error over the entire tensor.

[8] If there is a pre- and post-intervention period, then the target causal parameter is typically restricted to the T_1 post-intervention measurements.

■ 3.9.2 Learning Across Interventions via Tensor Factor Models

Tensor factor model. We start by recalling Definition 3.5.1 for a low-rank tensor factor model. Let $[Y_{tn}^{(d)}] \in \mathbb{R}^{T \times N \times D}$ denote an order-3 tensor of potential outcomes, which we assume admits the following decomposition:

$$Y_{tn}^{(d)} = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell} \lambda_{d\ell} + \varepsilon_{tn}^{(d)}, \tag{3.28}$$

where r is the canonical polyadic (CP) rank, $u_t, v_n, \lambda_d \in \mathbb{R}^r$ are latent factors associated with the t-th measurement, n-th unit, and d-th intervention, respectively, and $\varepsilon_{tn}^{(d)}$ is a mean zero residual term. We note that such a factorization always exists, but the key assumption is that the CP rank r is much smaller than N, T, D.

Algorith design guided by tensor factor models. The factorization in Assumption 9 is *implied* by the factorization assumed by a low-rank tensor as given in (3.28). In particular, Assumption 9 does not require the additional factorization of the (time, intervention) factor $u_t^{(d)}$ as $\langle u_t, \lambda_d \rangle$, where u_t is a time specific factor and λ_d is an intervention specific factor. An important question we pose is whether it is feasible to design estimators that exploit an implicit factorization of $u_t^{(d)} = \langle u_t, \lambda_d \rangle$. Indeed, a recent follow-up work, Squires et al. (2021) finds that rather than regressing on units, using a variant of SI which regresses across interventions leads to better empirical results in the setting of single cell therapies. Further, for example Shah and Yu (2019) directly exploit the tensor factor structure in (3.28) to provide max-norm error bounds for TC under a uniformly missing at random sparsity pattern. We leave as an open question whether one can exploit the further latent factor structure in (3.28) over that in Assumption 9 to operate under less stringent causal assumptions or data requirements, and/or identify and estimate more involved causal parameters.

■ 3.9.3 Need for a New Tensor Completion Toolkit for Causal Inference

The TC literature has grown tremendously because it provides an expressive formal framework for a large number of emerging applications. In particular, this literature quantifies the number of samples required and the computational complexity of different estimators to achieve theoretical guarantees for a given error metric (e.g., Frobenius-norm error over the entire tensor)—indeed, this trade-off is of central importance in the emerging sub-discipline at the interface of computer science, statistics, and machine learning. Given our preceding discussions connecting causal inference and TC, a natural question is whether we can apply the current toolkit of tensor completion to understand statistical and computational trade-offs in causal inference.

We believe a direct transfer of the techniques and analyses used in TC is not immediately possible for the following reasons. First, most results in the TC literature assume uniformly randomly missing entries over all $T \times N \times D$ elements of the tensor. In comparison, as seen in Figure 3.2, causal inference settings frequently induce a "missing not at random" sparsity pattern. Further, this literature typically studies the Frobenius-norm error across the entire tensor. However, as discussed in Section 3.9.1, most meaningful causal parameters require more refined error metrics over the tensor.

Hence, we pose two important and related open questions: (i) what block sparsity patterns and structural assumptions on the potential outcomes tensor allow for faithful recovery with respect to a meaningful error metric for causal inference, and (ii) if recovery is possible, what are the fundamental statistical and computational trade-offs that are achievable? An answer to these questions will formally bridge causal inference with TC, as well as computer science and machine learning more broadly.

■ 3.10 Covariates

In this section, we discuss how access to meaningful covariate information about each unit can help improve learning of the model. To this end, let $\boldsymbol{X} = [X_{kn}] \in \mathbb{R}^{K \times N}$ denote the matrix of covariates across units, i.e., X_{kn} denotes the k-th descriptor (or feature) of unit n. One approach towards incorporating covariates into the Synthetic Interventions estimation procedure described in Section 3.3.1, is to impose the following structure on $\boldsymbol{X}$.

Assumption 16 (Covariate structure). *For any $k \in [K]$ and $n \in [N]$, let $X_{kn} = \langle \varphi_k, v_n \rangle + \zeta_{kn}$. Here, $\varphi_k \in \mathbb{R}^r$ represents a latent factor specific to descriptor k, $v_n \in \mathbb{R}^r$ is the unit latent factor as defined in* (3.11)*, and $\zeta_{kn} \in \mathbb{R}$ is a mean zero measurement noise specific to descriptor k and unit n.*

Interpretation. The key structure we impose in Assumption 16 is that the covariates

X_{kn} have the same latent unit factors v_n as the potential outcomes $Y_{tn}^{(d)}$. Thus, given a target unit n and subgroup $\mathcal{I}^{(d)}$, this allows us to express unit n's covariates as a linear combination of the covariates associated with units within $\mathcal{I}^{(d)}$ via the *same* linear model that describes the relationship between their respective potential outcomes (formalized in Proposition 3.10.1 below). One notable flexibility of our covariate model is that the observations of covariates can be *noisy* due to the presence of measurement noise ζ_{kn}.

Proposition 3.10.1. *Given $n \in [N]$, $d \in [D]_0$, let Assumptions 11 and 16 hold. Then, conditioned on $\mathcal{E}_\varphi = \mathcal{E} \cup \{\varphi_k : k \in [K]\}$, we have for all k,*

$$\mathbb{E}[X_{kn} \,|\, \mathcal{E}_\varphi] = \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}[X_{kj} \,|\, \mathcal{E}_\varphi],$$

where recall $w^{(n,d)}$ as defined in Assumptions 11.

Proof. Proof is immediate by plugging Assumption 11 into Assumption 16. ■

Adapting the Synthetic Interventions estimator. Proposition 3.10.1 suggests a natural modification to the model learning stage of the Synthetic Interventions estimator presented in Section 3.3.1. In particular, similar to the estimation procedure of Abadie et al. (2010), we propose appending the covariates to the pre-intervention outcomes. Formally, we define $X_n = [X_{kn}] \in \mathbb{R}^K$ as the vector of covariates associated with the unit $n \in [N]$; analogously, we define $\boldsymbol{X}_{\mathcal{I}^{(d)}} = [X_{kj} : j \in \mathcal{I}^{(d)}] \in \mathbb{R}^{K \times N_d}$ as the matrix of covariates associated with units within $\mathcal{I}^{(d)}$. We further define $Z_n = [Y_{\text{pre},n}, X_n] \in \mathbb{R}^{T_0+K}$ and $\boldsymbol{Z}_{\mathcal{I}^{(d)}} = [\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}, \boldsymbol{X}_{\mathcal{I}^{(d)}}] \in \mathbb{R}^{(T_0+K) \times N_d}$ as the concatenation of pre-intervention outcomes and features associated with the target unit n and subgroup $\mathcal{I}^{(d)}$, respectively. We denote the SVD of $\boldsymbol{Z}_{\mathcal{I}^{(d)}}$ as

$$\boldsymbol{Z}_{\mathcal{I}^{(d)}} = \sum_{\ell=1}^{M} \widehat{\lambda}_\ell \widehat{\mu}_\ell \widehat{\nu}_\ell',$$

where $M = \min\{(T_0 + K), N_d\}$, $\widehat{\lambda}_\ell \in \mathbb{R}$ are the singular values (arranged in decreasing order), and $\widehat{\mu}_\ell \in \mathbb{R}^{T_0+K}$, $\widehat{\nu}_\ell \in \mathbb{R}^{N_d}$ are the left and right singular vectors, respectively. Then, we define the modified model parameter estimator as

$$\widehat{w}^{(n,d)} = \Big(\sum_{\ell=1}^{k} (1/\widehat{\lambda}_\ell) \widehat{\nu}_\ell \widehat{\mu}_\ell' \Big) Z_n.$$

The remainder of the algorithm, as described in Section 3.3.1, proceeds as is with the new estimate $\widehat{w}^{(n,d)}$.

Theoretical implications. Let in addition to setup of Theorem 4.5.1 and Assumption 16 hold. Then, it can be verified that the statistical guarantees in Sections 3.5.3, 3.5.4 and 3.6.1 continue to hold with T_0 being replaced by $T_0 + K$. In essence, adding covariates into the model-learning stage augments the data size, which can improve the estimation rates. For example, in Theorem 4.5.1, the term $T_0^{-\frac{1}{4}}$ in the bound on error is replaced by $(T_0 + K)^{-\frac{1}{4}}$.

■ 3.11 Simulation Details

In this section, we present the details of our simulations in Section 3.7.

■ 3.11.1 Consistency

Below, we provide details for our consistency simulation in Section 3.7.1.

Generative Model for Synthetic Data

We let $N_d = |\mathcal{I}^{(d)}| = 200$ and $r = 15$, where r is defined in Assumption 9. We define the latent unit factors associated with $\mathcal{I}^{(d)}$ as $\boldsymbol{V}_{\mathcal{I}^{(d)}} \in \mathbb{R}^{N_d \times r}$ (refer to (3.11)), where its entries are independently sampled from a standard normal distribution.

Pre-intervention data. We choose $T_0 = 200$ and $r_{\text{pre}} = 10$. We define the latent pre-intervention time factors under control ($d = 0$) as $\boldsymbol{U}_{\text{pre}} \in \mathbb{R}^{T_0 \times r}$, which is sampled as follows: (i) let $\boldsymbol{A} \in \mathbb{R}^{T_0 \times r_{\text{pre}}}$, where its entries are independently sampled from a standard normal; (ii) let $\boldsymbol{Q} \in \mathbb{R}^{r_{\text{pre}} \times (r - r_{\text{pre}})}$, where its entries are first independently sampled from a uniform distribution over $[0, 1]$, and then its columns are normalized to sum to one; (iii) define $\boldsymbol{U}_{\text{pre}} = [\boldsymbol{A}, \boldsymbol{AQ}]$ as the concatenation of $\boldsymbol{A}$ and $\boldsymbol{AQ}$. By construction, $\boldsymbol{U}_{\text{pre}}$ has rank r_{pre} w.h.p., which we empirically verify. Next, we define $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] = \boldsymbol{U}_{\text{pre}} \boldsymbol{V}'_{\mathcal{I}^{(d)}} \in \mathbb{R}^{T_0 \times N_d}$. Again by construction, we note that $\text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]) = r_{\text{pre}}$ w.h.p., which we empirically verify. We then generate the model $w^{(n,d)} \in \mathbb{R}^{N_d}$ from a multivariate standard normal distribution, and define $\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]w^{(n,d)} \in \mathbb{R}^{T_0}$.

Post-intervention data. We choose $T_1 = 200$. We sample the post-intervention time factors as follows: (i) let $\boldsymbol{P} \in \mathbb{R}^{T_1 \times T_0}$, where its entries are first independently sampled from a uniform distribution over $[0, 1]$, and then its rows are normalized to sum to one; (ii) define $\boldsymbol{U}_{\text{post}} = \boldsymbol{P}\boldsymbol{U}_{\text{pre}} \in \mathbb{R}^{T_1 \times r}$. We then define $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}] = \boldsymbol{U}_{\text{post}}\boldsymbol{V}'_{\mathcal{I}^{(d)}} \in \mathbb{R}^{T_1 \times N_d}$. To study the effect of the post-intervention period length, we will treat it as a variable. As such, we define $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)] \in \mathbb{R}^{\rho T_1 \times N_d}$ as the first ρT_1 rows of $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$, where $\rho \in \{0.1, 0.2, \ldots, 1.0\}$. For every ρ, we define $\theta_n^{(d)}(\rho)$ using $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)]$ and $w^{(n,d)}$.

Interpretation of data generating process. We now motivate the construction of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$. Recall that the SI framework allows potential outcomes from different interventions to be sampled from different distributions. As such, we construct $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$ such that they follow a different distribution to that of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$. This allows us to study when the model learnt using pre-intervention data "generalizes" to a post-intervention regime generated from a different distribution. However, we note that by construction, Assumption 15 holds w.h.p. between $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ for every ρ. We empirically verify all three conditions.

Observations. We generate $Y_{\text{pre},n}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ by adding independent noise entries from a normal distribution with mean zero and variance $\sigma^2 = 0.3$ to $\mathbb{E}[Y_{\text{pre},n}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, respectively. For every ρ, we generate $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)$ by applying the same additive noise model to $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)]$.

Verifying assumptions. We note that our data generating process ensures that Assumptions 8, 10, 12 hold. In addition, we empirically verify Assumptions 13 and 14. Further, for Assumption 9, we note that our pre- and post-intervention (expected) potential outcomes associated with $\mathcal{I}^{(d)}$ were all generated using $\boldsymbol{V}_{\mathcal{I}^{(d)}}$; thus, their variations only arise due to the sampling procedure for their respective latent time-intervention factors. Given that $\mathbb{E}[Y_{\text{pre},n}]$ and $\theta_n^{(d)}(\rho)$ were both defined using $w^{(n,d)}$, Assumption 11 holds.

Simulation Setup

We perform 100 iterations for each ρ. The potential outcomes, $\mathbb{E}[Y_{\text{pre},n}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)]$ are fixed, but the idiosyncratic shocks are re-sampled every iteration to yield new (random) outcomes. For each iteration, we use $(Y_{\text{pre},n}, \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}})$ to learn $\widehat{w}^{(n,d)}$, as given by (3.3). Then, we use $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}(\rho)$ and $\widehat{w}^{(n,d)}$ to yield $\widehat{\theta}_n^{(d)}(\rho)$, as given by (3.5). The mean absolute errors (MAEs), $|\widehat{\theta}_n^{(d)}(\rho) - \theta_n^{(d)}(\rho)|$ are plotted in Figure 3.10.

■ 3.11.2 Asymptotic Normality: Subspace Inclusion Holds

In this section, we describe the setup for Section 3.7.2.

Generative Model for Synthetic Data

We consider the binary $D = 2$ intervention model. Our interest is in estimating $\theta_n^{(1)}$. Our data generating process will be such that the pre- and post-intervention data will obey different distributions. Towards this, we choose $N_1 = |\mathcal{I}^{(1)}| = 400$ and $r = 15$. We define $\boldsymbol{V} \in \mathbb{R}^{N_1 \times r}$, where its entries are independently sampled from a standard normal.

Pre-intervention data. We choose $T_0 = 400$, and define the latent pre-intervention time factors under control as $\boldsymbol{U}_{\text{pre}} \in \mathbb{R}^{T_0 \times r}$, where its entries are sampled independently from a standard normal. Next, we define $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}] = \boldsymbol{U}_{\text{pre}} \boldsymbol{V}' \in \mathbb{R}^{T_0 \times N_1}$. By construction, $\text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]) = r$ w.h.p., which we empirically verify. We then generate the model $w^{(n,1)} \in \mathbb{R}^{N_1}$ from a standard normal, and define $\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}] w^{(n,1)} \in \mathbb{R}^{T_0}$. We define $\widetilde{w}^{(n,1)} = \mathbb{E}[Y_{\text{pre},\mathcal{I}^{(1)}}]^{\dagger} \mathbb{E}[Y_{\text{pre},n}]$, where $\dagger$ is the pseudo-inverse.

Post-intervention data. We choose $T_1 = 20$, and generate post-intervention time factors under $d = 1$ as follows: We define $\boldsymbol{U}_{\text{post}} \in \mathbb{R}^{T_1 \times r}$, where its entries are independently sampled uniformly over $[-\sqrt{3}, \sqrt{3}]$. We then define $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}] = \boldsymbol{U}_{\text{post}} \boldsymbol{V}' \in \mathbb{R}^{T_1 \times N_1}$. Finally, we define $\theta_n^{(1)}$ using $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$ and $w^{(n,1)}$.

Interpretation of data generating process. We note $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$ follow different distributions to reflect that the pre- and post-intervention potential outcomes are associated with different interventions; the former with $d = 0$ and the latter with $d = 1$. However, by construction, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$ satisfies Assumption 15 with respect to $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$ w.h.p., which we empirically verify.

Observations. We generate $Y_{\text{pre},n}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}$ by adding independent noise from a normal distribution with mean zero and variance $\sigma^2 = 0.5$ to $\mathbb{E}[Y_{\text{pre},n}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$, respectively. We generate $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$ by applying the same additive noise model to $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$.

Verifying assumptions. As before, Assumptions 8 to 14 hold by construction.

Simulation Setup

We perform 5000 iterations, where $\mathbb{E}[Y_{\text{pre},n}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(0)}}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$ are fixed throughout, but the idiosyncratic shocks are re-sampled to generate new (random) outcomes. Within each iteration, we first use $(Y_{\text{pre},n}, \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}})$ to fit $\widehat{w}^{(n,1)}$, as in (3.3). Next, we use $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$ and $\widehat{w}^{(n,1)}$ to yield $\widehat{\theta}_n^{(1)}$, as in (3.5). The resulting histogram is displayed in Figure 3.11a.

■ 3.11.3 Asymptotic Normality: Subspace Inclusion Fails

Next, we describe the setup for Section 3.7.2.

Generative Model for Synthetic Data

We continue analyzing the binary $D = 2$ intervention model. We let $N_1 = 400$, $r = 15$, and generate $\boldsymbol{V}_{\mathcal{I}^{(1)}} \in \mathbb{R}^{N_1 \times r}$ by independently sampling its entries from a standard normal.

Pre-intervention data. We choose $T_0 = 400$ and $r_{\text{pre}} = 12$. We construct $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$ using $\boldsymbol{V}_{\mathcal{I}^{(1)}}$ identically to that in Section 3.11.1, such that $\text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]) = r_{\text{pre}}$ w.h.p., which we empirically verify. As before, we generate $w^{(n,1)} \in \mathbb{R}^{N_1}$ from a standard normal and define $\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]w^{(n,1)} \in \mathbb{R}^{T_0}$, as well as $\widetilde{w}^{(n,1)} = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]^{\dagger}\mathbb{E}[Y_{\text{pre},n}]$.

Post-intervention data. We choose $T_1 = 20$, and define the post-intervention time factors under $d = 1$ as $\boldsymbol{U}_{\text{post}} \in \mathbb{R}^{T_1 \times r}$, where its entries are sampled independently from a standard normal. Next, we define $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}] = \boldsymbol{U}_{\text{post}}\boldsymbol{V}'_{\mathcal{I}^{(1)}} \in \mathbb{R}^{T_1 \times N_1}$. By construction, $\text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]) = r$ w.h.p., which we empirically verify. Since $r_{\text{pre}} < r$, Assumption 15 fails between $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$. We define $\theta_n^{(1)}$ using $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$ and $w^{(n,1)}$.

Observations. We generate $Y_{\text{pre},n}$ and $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}$ by adding independent noise from a normal distribution with mean zero and variance $\sigma^2 = 0.5$ to $\mathbb{E}[Y_{\text{pre},n}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$, respectively.

We generate $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$ by applying the same noise model to $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$.

Verifying assumptions. As before, Assumptions 8 to 14 hold by construction.

3.11.4 Simulation Setup

We perform 5000 iterations, where $\mathbb{E}[Y_{\text{pre},n}], \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}]$ are fixed, but the idiosyncratic shocks are re-sampled. In each iteration, we use $(Y_{\text{pre},n}, \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(1)}})$ to fit $\widehat{w}^{(n,1)}$, and then use $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(1)}}$ and $\widehat{w}^{(n,1)}$ to yield $\widehat{\theta}_n^{(1)}$. The resulting histogram is displayed in Figure 3.11b.

3.12 Proof of Theorem 4.3.1

In what follows, the descriptors above the equalities will denote the assumption used, e.g., $A1$ represents Assumption 1:

$$\begin{aligned}
&\mathbb{E}[Y_{tn}^{(d)} | u_t^{(d)}, v_n] \\
&\overset{A2}{=} \mathbb{E}[\langle u_t^{(d)}, v_n\rangle + \varepsilon_{tn}^{(d)} \mid u_t^{(d)}, v_n] \\
&\overset{A3}{=} \langle u_t^{(d)}, v_n\rangle \mid \{u_t^{(d)}, v_n\} \\
&= \langle u_t^{(d)}, v_n\rangle \mid \mathcal{E} \\
&\overset{A4}{=} \langle u_t^{(d)}, \sum_{j\in\mathcal{I}^{(d)}} w_j^{(n,d)} v_j\rangle \mid \mathcal{E} \\
&\overset{A3}{=} \sum_{j\in\mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}\left[(\langle u_t^{(d)}, v_j\rangle + \varepsilon_{tj}^{(d)}) \mid \mathcal{E}\right] \\
&\overset{A2}{=} \sum_{j\in\mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}[Y_{tj}^{(d)} | \mathcal{E}] \\
&\overset{A1}{=} \sum_{j\in\mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}\left[Y_{tjd} | \mathcal{E}\right].
\end{aligned}$$

The third equality follows since $\langle u_t^{(d)}, v_n\rangle$ is deterministic conditioned on $\{u_t^{(d)}, v_n\}$. Recalling (3.12), one can verify (3.14) using the same the argument used to prove (3.13), but where we begin by conditioning on the set of latent factors $\{u_t^{(d)}, v_n : t \in \mathcal{T}_{\text{post}}\}$ rather than just $\{u_t^{(d)}, v_n\}$.

3.13 Perturbation of Singular Values

In Lemma 3.13.1, we argue that the singular values of $\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}$ and $\mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$ are close.

Lemma 3.13.1. *Let Assumptions 8, 9, 10, 12, 13 and 14 hold. Then for any given $d \in [D]_0$, conditioned on $\mathcal{E}$, for any $\zeta > 0$ and $i \le \min\{T_0, N_d\}$, $|s_i - \widehat{s}_i| \le C\sigma(\sqrt{T_0} + \sqrt{N_d} + \zeta)$ with probability at least $1 - 2\exp\big(-\zeta^2\big)$, $C > 0$ is an absolute constant. Here, for $i \le \min\{T_0, N_d\}$, $s_i, \widehat{s}_i$ are singular values in non-increasing order of $\mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}|\mathcal{E}]$ and $\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}$ respectively.*

For ease of notation, we suppress the conditioning on $\mathcal{E}$ for the remainder of the proof. To bound the gap between s_i and $\widehat{s}_i$, we recall the following well-known results.

Lemma 3.13.2 (Weyl's inequality). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, let σ_i and $\widehat{\sigma}_i$ be the i-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then for all $i \le \min\{m, n\}$, $|\sigma_i - \widehat{\sigma}_i| \le \boldsymbol{A} - \boldsymbol{B}_{\mathrm{op}}$.*

Lemma 3.13.3 (Sub-Gaussian Matrices: Theorem 4.4.5 of Vershynin (2018)). *Let $\boldsymbol{A} = [A_{ij}]$ be an $m \times n$ random matrix where the entries A_{ij} are independent, mean zero, sub-Gaussian random variables. Then for any $t > 0$, we have $\boldsymbol{A}_{\mathrm{op}} \le CK(\sqrt{m} + \sqrt{n} + t)$ w.p. at least $1 - 2\exp\big(-t^2\big)$. Here, $K = \max_{i,j} A_{ij\psi_2}$, and $C > 0$ is an absolute constant.*

By Lemma 5.14.4, we have for any $i \le \min\{T_0, N_d\}$, $|s_i - \widehat{s}_i| \le \boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}]_{\mathrm{op}}$. Recalling Assumption 12 and applying Lemma 3.13.3, we conclude for any $t > 0$ and some absolute constant $C > 0$, $|s_i - \widehat{s}_i| \le C\sigma(\sqrt{T_0} + \sqrt{N_d} + t)$ w.p. at least $1 - 2\exp\big(-t^2\big)$. This completes the proof.

3.14 Helper Lemmas

We state two helper Lemmas needed for establishing Theorem 4.5.1.

Lemma 3.14.1. *Given $n \in [N]$ and $d \in [D]_0$, let the setup of Theorem 4.3.1 and Assumption 15 hold. Then,*

$$\theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \sum_{j \in \mathcal{I}^{(d)}} \widetilde{w}_j^{(n,d)} \mathbb{E}[Y_{tjd} | \mathcal{E}],$$

where $\mathcal{T}_{\text{post}} = \{T - T_1 + 1, \ldots, T\}$ and $\widetilde{w}^{(n,d)}$ is defined as in (3.15).

Proof. Recall that $\boldsymbol{V}_{\text{pre}} \in \mathbb{R}^{N_d \times r_{\text{pre}}}$ represents right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$. We also recall $\widetilde{w}^{(n,d)} = \boldsymbol{V}_{\text{pre}} \boldsymbol{V}_{\text{pre}}' w^{(n,d)}$, where $w^{(n,d)}$ is defined as in Theorem 4.3.1. Let $\boldsymbol{V}_{\text{post}} \in \mathbb{R}^{N_d \times r_{\text{post}}}$ denote the right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$. Assumption 15 implies

$$\boldsymbol{V}_{\text{post}} = \boldsymbol{V}_{\text{pre}} \boldsymbol{V}_{\text{pre}}' \boldsymbol{V}_{\text{post}}. \tag{3.29}$$

Therefore, we have

$$\begin{aligned} \mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]\widetilde{w}^{(n,d)} &= \mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]\boldsymbol{V}_{\text{pre}} \boldsymbol{V}_{\text{pre}}' w^{(n,d)} \\ &= \mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}] w^{(n,d)}, \end{aligned} \tag{3.30}$$

where we use (3.29) in the last equality. Hence, we conclude

$$\theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \sum_{j \in \mathcal{I}^{(d)}} w_j^{(n,d)} \mathbb{E}[Y_{tjd}] = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \sum_{j \in \mathcal{I}^{(d)}} \widetilde{w}_j^{(n,d)} \mathbb{E}[Y_{tjd}].$$

The first equality follows from (3.14) in Theorem 4.3.1. The second equality follows from (3.30). This completes the proof. ∎

Lemma 3.14.2 (Corollary 5.1 of Agarwal et al. (2021d)). *Given unit $n \in [N]$ and intervention $d \in [D]_0$, let Assumptions 8 to 15 hold. Further, suppose $k = r_{\text{pre}} =$ $\text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} | \mathcal{E}])$, where k is defined as in (3.3). Then, conditioned on $\mathcal{E}$,*

$$\widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)} = O_p\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}} \sqrt{\log(T_0 N_d)}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}} \right).$$

Proof. We first re-state Corollary 5.1 of Agarwal et al. (2021d).

Lemma (Corollary 5.1 of Agarwal et al. (2021d)). *Let the setup of Lemma 4.9.1 hold. Then* *with probability at least* $1 - O(1/(T_0 N_d)^{10})$

$$\widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)}{}_2^2 \le C(\sigma)\frac{r_{\text{pre}}^2 \log(T_0 N_d)}{\min\{T_0, N_d\}}\widetilde{w}^{(n,d)}{}_2^2, \tag{3.31}$$

where $C(\sigma)$ *is a constant that only depends on* σ*.*

This is seen by adapting the notation in Agarwal et al. (2021d) to that used in this paper. In particular, $y = Y_{\text{pre},n}, \boldsymbol{X} = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}], \widetilde{\boldsymbol{Z}} = \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}, \widehat{\beta} = \widehat{w}^{(n,d)}, \beta^* = \widetilde{w}^{(n,d)}$, where $y, \boldsymbol{X}, \widetilde{\boldsymbol{Z}}, \widehat{\beta}, \beta^*$ are the notations used in Agarwal et al. (2021d).[9] Further, we also use the fact that $\boldsymbol{X}\beta^*$ in the notation of Agarwal et al. (2021d) (i.e., $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]\widetilde{w}^{(n,d)}$) in our notation) equals $\mathbb{E}[Y_{\text{pre},n}]$. This follows from

$$\mathbb{E}[Y_{\text{pre},n}] = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]\widetilde{w}^{(n,d)}, \tag{3.32}$$

which follows from (3.13) in Theorem 4.3.1 and (3.30). We conclude by noting (3.31) implies

$$\widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)} = O_p\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}}\sqrt{\log(T_0 N_d)}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}}\right).$$

$\blacksquare$

3.15 Proof of Theorem 4.5.1

For ease of notation, we suppress the conditioning on $\mathcal{E}$ for the remainder of the proof. Let $C(v_p) = \max\{1, v_p\}$ for any $v \in \mathbb{R}^a$, and let $\mathcal{T}_{\text{post}} = \{T - T_1 + 1, \dots, T\}$. For any $t \in \mathcal{T}_{\text{post}}$, let $Y_{t,\mathcal{I}^{(d)}} = [Y_{tjd} : j \in \mathcal{I}^{(d)}] \in \mathbb{R}^{N_d}$ and $\varepsilon_{t,\mathcal{I}^{(d)}} = [\varepsilon_{tj}^{(d)} : j \in \mathcal{I}^{(d)}] \in \mathbb{R}^{N_d}$. Note that the rows of $\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}$ are formed by $\{Y_{t,\mathcal{I}^{(d)}} : t \in \mathcal{T}_{\text{post}}\}$. Additionally, let $\Delta^{(n,d)} = \widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)}$. Finally, for any matrix $\boldsymbol{A}$ with orthonormal columns, let $\mathcal{P}_A = \boldsymbol{A}\boldsymbol{A}'$ denote the projection matrix onto the subspace spanned by the columns of $\boldsymbol{A}$.

[9] Since we do not consider missing values in this work, $\widetilde{\boldsymbol{Z}} = \boldsymbol{Z}$, where $\boldsymbol{Z}$ is the notation used in Agarwal et al. (2021d).

By (3.5) and Lemma 3.14.1 (implied by Theorem 4.3.1), we have

$$\begin{aligned}\widehat{\theta}_n^{(d)} - \theta_n^{(d)} &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \left(\langle Y_{t,\mathcal{I}^{(d)}}, \widehat{w}^{(n,d)} \rangle - \langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \widetilde{w}^{(n,d)} \rangle \right) \\ &= \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \left(\langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \Delta^{(n,d)} \rangle + \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \widetilde{w}^{(n,d)} \rangle + \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \Delta^{(n,d)} \rangle \right), \quad (3.33)\end{aligned}$$

where we have used $Y_{t,\mathcal{I}^{(d)}} = \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}] + \varepsilon_{t,\mathcal{I}^{(d)}}$. From Assumption 15, it follows that $\boldsymbol{V}_{\text{post}} = \mathcal{P}_{V_{\text{pre}}} \boldsymbol{V}_{\text{post}}$, where $\boldsymbol{V}_{\text{pre}}, \boldsymbol{V}_{\text{post}}$ are the right singular vectors of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$. Hence, $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}] = \mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}] \mathcal{P}_{V_{\text{pre}}}$. As such, for any $t \in \mathcal{T}_{\text{post}}$,

$$\langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \Delta^{(n,d)} \rangle = \langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} \rangle. \quad (3.34)$$

Plugging (3.34) into (3.33) yields

$$\begin{aligned}\widehat{\theta}_n^{(d)} - \theta_n^{(d)} = \frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} &\left(\langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} \rangle \right. \\ &\left. + \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \widetilde{w}^{(n,d)} \rangle + \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \Delta^{(n,d)} \rangle \right). \quad (3.35)\end{aligned}$$

Below, we bound the three terms on the right-hand side (RHS) of (3.35) separately.

Bounding term 1. By Cauchy-Schwartz inequality, observe that

$$\langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} \rangle \leq \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}]_2 \; \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)}{}_2.$$

Under Assumption 13, we have $\mathbb{E}[Y_{t,\mathcal{I}^{(d)}}]_2 \leq \sqrt{N_d}$. As such,

$$\frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} \rangle \leq \sqrt{N_d} \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)}{}_2.$$

Hence, it remains to bound $\mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)}{}_2$. Towards this, we state the following lemma. Its proof can be found in Section 3.15.1.

Lemma 3.15.1. *Consider the setup of Theorem 4.5.1. Then,*

$$\mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} = O_p \left(\frac{r_{pre}^{1/2}}{\sqrt{N_d} T_0^{\frac{1}{4}}} + \frac{\widetilde{w}^{(n,d)}{}_1 r_{\text{pre}}^{3/2} \sqrt{\log(T_0 N_d)}}{\sqrt{N_d} \cdot \min\{\sqrt{T_0}, \sqrt{N_d}\}} \right).$$

Using Lemma 3.15.1, we obtain

$$\frac{1}{T_1}\sum_{t\in\mathcal{T}_{\text{post}}}\langle\mathbb{E}[Y_{t,\mathcal{I}^{(d)}}],\mathcal{P}_{V_{\text{pre}}}\Delta^{(n,d)}\rangle = O_p\left(\frac{r_{\text{pre}}^{1/2}}{T_0^{\frac{1}{4}}} + \frac{\widetilde{w}^{(n,d)}{}_1 r_{\text{pre}}^{3/2}\sqrt{\log(T_0N_d)}}{\min\{\sqrt{T_0},\sqrt{N_d}\}}\right). \tag{3.36}$$

This concludes the analysis for the first term.

Bounding term 2. We begin with a simple lemma.

Lemma 3.15.2. *Let γ_t be a sequence of independent mean zero sub-Gaussian random variables with variance σ^2. Then, $\frac{1}{T}\sum_{t=1}^{T}\gamma_t = O_p\left(\frac{\sigma^2}{\sqrt{T}}\right)$.*

Proof. Immediately holds by Hoeffding's lemma (Lemma 5.14.2). ■

By Assumptions 9 and 12, we have for any $t\in\mathcal{T}_{\text{post}}$,

$$\mathbb{E}[\langle\varepsilon_{t,\mathcal{I}^{(d)}},\widetilde{w}^{(n,d)}\rangle] = 0, \quad \text{Var}(\langle\varepsilon_{t,\mathcal{I}^{(d)}},\widetilde{w}^{(n,d)}\rangle) = \sigma^2\widetilde{w}^{(n,d)}{}_2^2.$$

Since $\langle\varepsilon_{t,\mathcal{I}^{(d)}},\widetilde{w}^{(n,d)}\rangle$ are independent across t, Lemma 3.15.2 yields

$$\frac{1}{T_1}\sum_{t\in\mathcal{T}_{\text{post}}}\langle\varepsilon_{t,\mathcal{I}^{(d)}},\widetilde{w}^{(n,d)}\rangle = O_p\left(\frac{\widetilde{w}^{(n,d)}{}_2}{\sqrt{T_1}}\right). \tag{3.37}$$

Bounding term 3. First, we define the event $\mathcal{E}_1$ as

$$\mathcal{E}_1 = \left\{\Delta^{(n,d)}{}_2 = O\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}}\sqrt{\log(T_0N_d)}}{\min\{\sqrt{T_0},\sqrt{N_d}\}}\right)\right\}.$$

By Lemma 4.9.1, $\mathcal{E}_1$ occurs w.h.p. (defined in Section 3.2). Next, we define $\mathcal{E}_2$ as

$$\mathcal{E}_2 = \left\{\frac{1}{T_1}\sum_{t\in\mathcal{T}_{\text{post}}}\langle\varepsilon_{t,\mathcal{I}^{(d)}},\Delta^{(n,d)}\rangle = O\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}}\sqrt{\log(T_0N_d)}}{\sqrt{T_1}\min\{\sqrt{T_0},\sqrt{N_d}\}}\right)\right\}.$$

Now, condition on $\mathcal{E}_1$. By Assumptions 9 and 12, we have for any $t\in\mathcal{T}_{\text{post}}$,

$$\mathbb{E}[\langle\varepsilon_{t,\mathcal{I}^{(d)}},\Delta^{(n,d)}\rangle] = 0$$
$$\text{Var}(\langle\varepsilon_{t,\mathcal{I}^{(d)}},\Delta^{(n,d)}\rangle) = \sigma^2\Delta^{(n,d)}{}_2^2.$$

The above uses the fact that $\widehat{w}^{(n,d)}$ depends on $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$ and $Y_{\text{pre},n}$ and hence are independent of $\varepsilon_{t,\mathcal{I}^{(d)}}$ for all $t \in \mathcal{T}_{\text{post}}$. Given that $\langle \varepsilon_{t,\mathcal{I}^{(d)}}, \Delta^{(n,d)} \rangle$ are independent across t, Lemmas 4.9.1 and 3.15.2 imply $\mathcal{E}_2 | \mathcal{E}_1$ occurs w.h.p. Further, we note

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1)\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1^c)\mathbb{P}(\mathcal{E}_1^c) \geq \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1)\mathbb{P}(\mathcal{E}_1). \tag{3.38}$$

Since $\mathcal{E}_1$ and $\mathcal{E}_2|\mathcal{E}_1$ occur w.h.p, it follows from (3.38) that $\mathcal{E}_2$ occurs w.h.p. As a result,

$$\frac{1}{T_1} \sum_{t \in \mathcal{T}_{\text{post}}} \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \Delta^{(n,d)} \rangle = O_p\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}} \sqrt{\log(T_0 N_d)}}{\sqrt{T_1} \min\{\sqrt{T_0}, \sqrt{N_d}\}} \right). \tag{3.39}$$

Collecting terms. Incorporating (3.36), (3.37), (3.39) into (3.35), and simplifying yields

$$\widehat{\theta}_n^{(d)} - \theta_n^{(d)} = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{T_0^{\frac{1}{4}}} + \frac{\widetilde{w}^{(n,d)}{}_2}{\sqrt{T_1}} + \frac{\widetilde{w}^{(n,d)}{}_1 r_{\text{pre}}^{3/2} \sqrt{\log(T_0 N_d)}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}} \right).$$

This concludes the proof of Theorem 4.5.1.

■ 3.15.1 Proof of Lemma 3.15.1

We begin by introducing some helpful notations: let $\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} = \sum_{\ell=1}^{r_{\text{pre}}} \widehat{s}_\ell \widehat{u}_\ell \widehat{v}_\ell'$ be the rank r_{pre}-approximation of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$. More compactly, $\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} = \widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}}$. To establish Lemma 3.15.1, consider the following decomposition:

$$\mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} = (\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}}) \Delta^{(n,d)} + \mathcal{P}_{\widehat{V}_{\text{pre}}} \Delta^{(n,d)}.$$

We proceed to bound each term separately.

Bounding term 1. Recall $\boldsymbol{A}v_2 \leq \boldsymbol{A}_{\text{op}} v_2$ for any $\boldsymbol{A} \in \mathbb{R}^{a \times b}$ and $v \in \mathbb{R}^b$. Thus,

$$(\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}}) \Delta^{(n,d)}{}_2 \leq \mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}}{}_{\text{op}} \Delta^{(n,d)}{}_2. \tag{3.40}$$

To control the above term, we state a helper lemma that bounds the distance between the subspaces spanned by the columns of $\boldsymbol{V}_{\text{pre}}$ and $\widehat{\boldsymbol{V}}_{\text{pre}}$. Its proof is given in Section 3.15.2.

Lemma 3.15.3. *Consider the setup of Theorem 4.5.1. Then for any $\zeta > 0$, the following* *holds w.p. at least* $1 - 2\exp\big(-\zeta^2\big)$*:* $\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}\,\text{op}} \leq \frac{C\sigma(\sqrt{T_0}+\sqrt{N_d}+\zeta)}{s_{r_{\text{pre}}}}$*, where* $s_{r_{\text{pre}}}$ *is the* r_{pre}*-th singular value of* $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ *with* $r_{\text{pre}} = \text{rank}(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}])$*, and* $C > 0$ *is an absolute constant.*

Applying Lemma 3.15.3 with Assumption 14, we have

$$\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}\,\text{op}} = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}}\right). \tag{3.41}$$

Substituting (3.41) and the bound in Lemma 4.9.1 into (3.40), we obtain

$$(\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\Delta^{(n,d)} = O_p\left(\frac{\widetilde{w}^{(n,d)}{}_2 r_{\text{pre}}^{3/2}\sqrt{\log(T_0 N_d)}}{\min\{T_0, N_d\}}\right). \tag{3.42}$$

Bounding term 2. To begin, since $\widehat{\boldsymbol{V}}_{\text{pre}}$ is an isometry, it follows that

$$\mathcal{P}_{\widehat{V}_{\text{pre}}}\Delta^{(n,d)}{}^2_2 = \widehat{\boldsymbol{V}}'_{\text{pre}}\Delta^{(n,d)}{}^2_2. \tag{3.43}$$

We upper bound $\widehat{\boldsymbol{V}}'_{\text{pre}}\Delta^{(n,d)}{}^2_2$ as follows: consider

$$\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\Delta^{(n,d)}{}^2_2 = (\widehat{\boldsymbol{V}}'_{\text{pre}}\Delta^{(n,d)})'\widehat{\boldsymbol{\Sigma}}^2_{r_{\text{pre}}}(\widehat{\boldsymbol{V}}'_{\text{pre}}\Delta^{(n,d)}) \geq \widehat{s}^2_{r_{\text{pre}}}\,\widehat{\boldsymbol{V}}'_{\text{pre}}\Delta^{(n,d)}{}^2_2. \tag{3.44}$$

Using (3.44) and (3.43) together implies

$$\mathcal{P}_{\widehat{V}_{\text{pre}}}\Delta^{(n,d)}{}^2_2 \leq \frac{\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\Delta^{(n,d)}{}^2_2}{\widehat{s}^2_{r_{\text{pre}}}}. \tag{3.45}$$

To bound the numerator in (3.45), note

$$\begin{aligned}
&\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\Delta^{(n,d)}{}^2_2 \\
&\leq 2\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}]^2_2 + 2\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\widetilde{w}^{(n,d)}{}^2_2 \\
&= 2\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}]^2_2 + 2(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] - \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}})\widetilde{w}^{(n,d)}{}^2_2,
\end{aligned} \tag{3.46}$$

where we have used (3.32). To further upper bound the the second term on the RHS above, we use the following inequality: for any $\boldsymbol{A} \in \mathbb{R}^{a\times b}$, $v \in \mathbb{R}^b$,

$$\boldsymbol{A} v_2 = \sum_{j=1}^{b} \boldsymbol{A}_{\cdot j} v_j 2 \le (\max_{j\le b} \boldsymbol{A}_{\cdot j} 2)(\sum_{j=1}^{b} |v_j|) = \boldsymbol{A}_{2,\infty} v_1, \tag{3.47}$$

where $\boldsymbol{A}_{2,\infty} = \max_j \boldsymbol{A}_{\cdot j 2}$ and $\boldsymbol{A}_{\cdot j}$ represents the j-th column of $\boldsymbol{A}$. Thus,

$$(\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] - \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}})\widetilde{w}^{(n,d)2}_{\ 2} \le \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] - \boldsymbol{Y}^{r_{\text{pre}}\ 2}_{\text{pre},\mathcal{I}^{(d)}\,2,\infty} \widetilde{w}^{(n,d)2}_{\ 1}. \tag{3.48}$$

Substituting (3.46) into (3.45) and subsequently using (3.48) implies

$$\begin{aligned} &\mathcal{P}_{\widehat{V}_{\text{pre}}} \Delta^{(n,d)2}_{\ 2} \\ &\le \frac{2}{\widehat{s}^2_{r_{\text{pre}}}} \Big(\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}]^2_2 + \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] - \boldsymbol{Y}^{r_{\text{pre}}\ 2}_{\text{pre},\mathcal{I}^{(d)}\,2,\infty} \widetilde{w}^{(n,d)2}_{\ 1} \Big). \end{aligned} \tag{3.49}$$

Next, we bound $\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}]^2_2$. To this end, observe that

$$\begin{aligned} &\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - Y_{\text{pre},n}{}^2_2 = \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}] - \varepsilon_{\text{pre},n}{}^2_2 \\ &= \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}]^2_2 + \varepsilon_{\text{pre},n}{}^2_2 - 2\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}], \varepsilon_{\text{pre},n}\rangle. \end{aligned} \tag{3.50}$$

We call upon Property 4.1 of Agarwal et al. (2021d), which states that $\widehat{w}^{(n,d)}$, as given by (3.3), is the unique solution to the following program:

$$\begin{aligned} &\text{minimize} \quad \|w\|_2 \quad \text{over} \quad w \in \mathbb{R}^{N_d} \\ &\text{such that} \quad w \in \arg\min_{\omega\in\mathbb{R}^{N_d}} Y_{\text{pre},n} - \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \omega^2_2. \end{aligned}$$

This, along with (3.32), implies that

$$\begin{aligned} &\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - Y_{\text{pre},n}{}^2_2 \le \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widetilde{w}^{(n,d)} - Y_{\text{pre},n}{}^2_2 \\ &= \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widetilde{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}] - \varepsilon_{\text{pre},n}{}^2_2 \\ &= \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widetilde{w}^{(n,d)} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]\widetilde{w}^{(n,d)} - \varepsilon_{\text{pre},n}{}^2_2 \\ &= (\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}])\widetilde{w}^{(n,d)2}_{\ 2} + \varepsilon_{\text{pre},n}{}^2_2 - 2\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widetilde{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}], \varepsilon_{\text{pre},n}\rangle. \end{aligned} \tag{3.51}$$

From (3.50) and (3.51), we have

$$\begin{aligned} &\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} - \mathbb{E}[Y_{\text{pre},n}]^2_2 \\ &\le (\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}])\widetilde{w}^{(n,d)2}_{\ 2} + 2\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \Delta^{(n,d)}, \varepsilon_{\text{pre},n}\rangle \end{aligned}$$

$$\leq \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]^2_{2,\infty} \widetilde{w}^{(n,d)2}_{\;1} + 2\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \Delta^{(n,d)}, \varepsilon_{\text{pre},n} \rangle, \tag{3.52}$$

where we used (3.47) in the second inequality above. Using (3.49) and (3.52), we obtain

$$\begin{aligned} &\mathcal{P}_{\widehat{V}_{\text{pre}}} \Delta^{(n,d)2}_{\;2} \\ &\leq \frac{4}{\widehat{s}^2_{r_{\text{pre}}}} \Big(\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]^2_{2,\infty} \widetilde{w}^{(n,d)2}_{\;1} + \langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \Delta^{(n,d)}, \varepsilon_{\text{pre},n} \rangle \Big). \end{aligned} \tag{3.53}$$

We now state two helper lemmas that will help us conclude the proof. The proof of Lemmas 3.15.4 and 3.15.5 are given in Appendices 3.15.3 and 3.15.4, respectively.

Lemma 3.15.4 (Lemma 7.2 of Agarwal et al. (2021d)). *Let Assumptions 8, 9, 10, 12, 13, 14 hold. Suppose $k = r_{\text{pre}}$, where k is defined as in (3.3). Then,* $\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]_{2,\infty} = O_p\left(\frac{\sqrt{r_{\text{pre}} T_0 \log(T_0 N_d)}}{\min\{\sqrt{T_0}, \sqrt{N_d}\}} \right)$.

Lemma 3.15.5. *Let Assumptions 8 to 14 hold. Then, given $\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}$, the following holds with respect to the randomness in $\varepsilon_{\text{pre},n}$:*

$$\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \Delta^{(n,d)}, \varepsilon_{\text{pre},n} \rangle = O_p\left(r_{\text{pre}} + \sqrt{T_0} + \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]_{2,\infty} \widetilde{w}^{(n,d)}{}_1 \right).$$

Incorporating Lemmas 3.13.1, 3.15.4, 3.15.5, and Assumption 14 into (3.53), we conclude

$$\mathcal{P}_{\widehat{V}_{\text{pre}}} \Delta^{(n,d)} = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{\sqrt{N_d} T_0^{\frac{1}{4}}} + \frac{r_{\text{pre}} \widetilde{w}^{(n,d)}{}_1 \sqrt{\log(T_0 N_d)}}{\sqrt{N_d} \cdot \min\{\sqrt{T_0}, \sqrt{N_d}\}} \right). \tag{3.54}$$

Collecting terms. Combining (3.42) and (3.54), and noting $v_2 \leq v_1$ for any v, we conclude

$$\mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{\sqrt{N_d} T_0^{\frac{1}{4}}} + \frac{\widetilde{w}^{(n,d)}{}_1 r^{3/2}_{\text{pre}} \sqrt{\log(T_0 N_d)}}{\sqrt{N_d} \cdot \min\{\sqrt{T_0}, \sqrt{N_d}\}} \right).$$

■ 3.15.2 Proof of Lemma 3.15.3

We recall the well-known singular subspace perturbation result by Wedin.

Theorem 3.15.1 (Wedin's Theorem Wedin (1972)). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, let $\boldsymbol{V}, \widehat{\boldsymbol{V}} \in \mathbb{R}^{n \times n}$ denote their respective right singular vectors. Further, let $\boldsymbol{V}_k \in \mathbb{R}^{n \times k}$ (respectively,*

$\widehat{\boldsymbol{V}}_k \in \mathbb{R}^{n\times k}$) correspond to the truncation of $\boldsymbol{V}$ (respectively, $\widehat{\boldsymbol{V}}$), respectively, that retains the columns corresponding to the top k singular values of $\boldsymbol{A}$ (respectively, $\boldsymbol{B}$). Let s_i represent the i-th singular value of $\boldsymbol{A}$. Then, $\mathcal{P}_{V_k} - \mathcal{P}_{\widehat{V}_k}{}_{\mathrm{op}} \le \frac{2\boldsymbol{A}-\boldsymbol{B}_{\mathrm{op}}}{s_k - s_{k+1}}$.

Recall that $\widehat{\boldsymbol{V}}_{\mathrm{pre}}$ is formed by the top r_{pre} right singular vectors of $\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}$ and $\boldsymbol{V}_{\mathrm{pre}}$ is formed by the top r_{pre} signular vectors of $\mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}]$. Therefore, Theorem 3.15.1 gives $\mathcal{P}_{\widehat{V}_{\mathrm{pre}}} - \mathcal{P}_{V_{\mathrm{pre}}}{}_{\mathrm{op}} \le \frac{2\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}]_{\mathrm{op}}}{s_{r_{\mathrm{pre}}}}$, where we used $\mathrm{rank}(\mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}]) = r_{\mathrm{pre}}$, and hence $s_{r_{\mathrm{pre}}+1} = 0$. By Assumption 12 and Lemma 3.13.3, we can further bound the inequality above. In particular, for any $\zeta > 0$, we have w.p. at least $1 - 2\exp\left(-\zeta^2\right)$, $\mathcal{P}_{\widehat{V}_{\mathrm{pre}}} - \mathcal{P}_{V_{\mathrm{pre}}}{}_{\mathrm{op}} \le \frac{C\sigma(\sqrt{T_0}+\sqrt{N_d}+\zeta)}{s_{r_{\mathrm{pre}}}}$, where $C > 0$ is an absolute constant.

■ 3.15.3 Proof of Lemma 3.15.4

We first re-state Lemma 7.2 of Agarwal et al. (2021d).

Lemma. *Let the setup of Lemma 3.15.4 hold. Then w.p. at least $1 - O(1/(T_0 N_d)^{10})$,*

$$\begin{aligned}\mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}] - \boldsymbol{Y}^{r_{\mathrm{pre}}}_{\mathrm{pre},\mathcal{I}^{(d)}}{}^{2}_{2,\infty}&\\ \le C(\sigma)\left(\frac{(T_0+N_d)(T_0+\sqrt{T_0}\log(T_0N_d))}{s^2_{r_{\mathrm{pre}}}} + r_{\mathrm{pre}} + \sqrt{r_{\mathrm{pre}}}\log(T_0N_d)\right) + \frac{\log(T_0N_d)}{N_d},&\end{aligned} \tag{3.55}$$

where $C(\sigma)$ is a constant that only depends on σ.

This is seen by adapting the notation in Agarwal et al. (2021d) to that used in this paper. In particular, $\boldsymbol{X} = \mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}]$, $\widetilde{\boldsymbol{Z}}^r = \boldsymbol{Y}^{r_{\mathrm{pre}}}_{\mathrm{pre},\mathcal{I}^{(d)}}$, where $\boldsymbol{X}, \widetilde{\boldsymbol{Z}}^r$ are the notations used in Agarwal et al. (2021d) with $r = r_{\mathrm{pre}}$.

Next, we simplify (3.55) using Assumption 14. As such, w.p. at least $1 - O(1/(T_0N_d)^{10})$,

$$\mathbb{E}[\boldsymbol{Y}_{\mathrm{pre},\mathcal{I}^{(d)}}] - \boldsymbol{Y}^{r_{\mathrm{pre}}}_{\mathrm{pre},\mathcal{I}^{(d)}}{}^{2}_{2,\infty} \le C(\sigma)\left(\frac{r_{\mathrm{pre}}T_0\log(T_0N_d)}{\min\{T_0, N_d\}}\right).$$

This concludes the proof of Lemma 3.15.4.

■ 3.15.4 Proof of Lemma 3.15.5

Throughout this proof, $C, c > 0$ will denote absolute constants, which can change from line to line or even within a line. Recall $\widehat{w}^{(n,d)} = \widehat{\boldsymbol{V}}_{\mathrm{pre}}\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{pre}}\widehat{\boldsymbol{U}}'_{\mathrm{pre}}Y_{\mathrm{pre},n}$ and $Y_{\mathrm{pre},n} = \mathbb{E}[Y_{\mathrm{pre},n}] + \varepsilon_{\mathrm{pre},n}$.

Thus,

$$\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} = \widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}} \widehat{\boldsymbol{V}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}^{-1}_{\text{pre}} \widehat{\boldsymbol{U}}'_{\text{pre}} Y_{\text{pre},n} = \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}] + \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}. \tag{3.56}$$

Therefore,

$$\begin{aligned}
&\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} (\widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)}), \varepsilon_{\text{pre},n} \rangle && (3.57)\\
&= \langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}], \varepsilon_{\text{pre},n} \rangle + \langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}, \varepsilon_{\text{pre},n} \rangle - \langle \widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}} \widetilde{w}^{(n,d)}, \varepsilon_{\text{pre},n} \rangle.
\end{aligned}$$

Note that $\varepsilon_{\text{pre},n}$ is independent of $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}$, and thus also independent of $\widehat{\boldsymbol{U}}_{\text{pre}}$, $\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}$. Therefore,

$$\mathbb{E}\Big[\langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}], \varepsilon_{\text{pre},n} \Big] = 0, \tag{3.58}$$

$$\mathbb{E}\Big[\langle \widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}} \widetilde{w}^{(n,d)}, \varepsilon_{\text{pre},n} \rangle\Big] = 0 \tag{3.59}$$

Moreover, using the cyclic property of the trace operator, we obtain

$$\begin{aligned}
&\mathbb{E}[\langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}, \varepsilon_{\text{pre},n} \rangle] = \mathbb{E}[\varepsilon'_{\text{pre},n} \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}] = \mathbb{E}[\operatorname{tr}\Big(\varepsilon'_{\text{pre},n} \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}\Big)]\\
&= \mathbb{E}[\operatorname{tr}\Big(\varepsilon_{\text{pre},n} \varepsilon'_{\text{pre},n} \mathcal{P}_{\widehat{U}_{\text{pre}}}\Big)] = \operatorname{tr}\Big(\mathbb{E}[\varepsilon_{\text{pre},n} \varepsilon'_{\text{pre},n}] \mathcal{P}_{\widehat{U}_{\text{pre}}}\Big) = \operatorname{tr}\Big(\sigma^2 \mathcal{P}_{\widehat{U}_{\text{pre}}}\Big) = \sigma^2 \widehat{\boldsymbol{U}}_{\text{pre}}{}^2_F = \sigma^2 r_{\text{pre}}. \quad (3.60)
\end{aligned}$$

Note that the above also uses (i) the mean zero and coordinate-wise independence of $\varepsilon_{\text{pre},n}$; (ii) orthonormality of $\widehat{\boldsymbol{U}}_{\text{pre}} \in \mathbb{R}^{n \times r_{\text{pre}}}$. Therefore, it follows that

$$\mathbb{E}[\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \Delta^{(n,d)}, \varepsilon_{\text{pre},n} \rangle] = \sigma^2 r_{\text{pre}}. \tag{3.61}$$

Next, to obtain high probability bounds for the inner product term, we use the following lemmas, the proofs of which can be found in Appendix A of Agarwal et al. (2021d).

Lemma 3.15.6 (Modified Hoeffding's Lemma). *Let $X \in \mathbb{R}^n$ be r.v. with independent mean-zero sub-Gaussian random coordinates with $X_{i\psi_2} \leq K$. Let $a \in \mathbb{R}^n$ be another random vector that satisfies $a_2 \leq b$ for some constant $b \geq 0$. Then for any $\zeta \geq 0$, $\mathbb{P}\Big(\Big|\sum_{i=1}^n a_i X_i\Big| \geq \zeta\Big) \leq 2\exp\Big(-\frac{c\zeta^2}{K^2 b^2}\Big)$, where $c > 0$ is a universal constant.*

Lemma 3.15.7 (Modified Hanson-Wright Inequality). *Let $X \in \mathbb{R}^n$ be a r.v. with independent mean-zero sub-Gaussian coordinates with $X_{i\psi_2} \leq K$. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a random matrix satisfying $\boldsymbol{A}_{\text{op}} \leq a$ and $\boldsymbol{A}^2_F \leq b$ for some $a, b \geq 0$. Then for any $\zeta \geq 0$,*

$$\mathbb{P}\Big(\Big|X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X]\Big| \geq \zeta\Big) \leq 2\exp\Big(-c\min\Big(\frac{\zeta^2}{K^4 b}, \frac{\zeta}{K^2 a}\Big)\Big).$$

Using Lemma 5.14.2 and (3.58), and Assumptions 9 and 12, it follows that for any $\zeta > 0$

$$\mathbb{P}\left(\langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}], \varepsilon_{\text{pre},n} \rangle \geq \zeta\right) \leq \exp\left(- \frac{c\zeta^2}{T_0 \sigma^2} \right). \tag{3.62}$$

Note that the above also uses $\mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}]_2 \leq \mathbb{E}[Y_{\text{pre},n}]_2 \leq \sqrt{T_0}$, which follows from the fact that $\mathcal{P}_{\widehat{U}_{\text{pre}}}{}_{\text{op}} \leq 1$ and Assumption 13. Further, (3.62) implies

$$\langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}], \varepsilon_{\text{pre},n} \rangle = O_p(\sqrt{T_0}). \tag{3.63}$$

Similarly, using (3.59), we have for any $\zeta > 0$

$$\mathbb{P}\left(\langle \widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}} \widetilde{w}^{(n,d)}, \varepsilon_{\text{pre},n} \rangle \geq \zeta\right) \leq \exp\left(- \frac{c\zeta^2}{\sigma^2 (T_0 + \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]^2_{2,\infty} \widetilde{w}^{(n,d)}{}^2_1)} \right) \tag{3.64}$$

where we use the fact that

$$\begin{aligned}
\widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}} \widetilde{w}^{(n,d)}{}_2 &= \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widetilde{w}^{(n,d)} \pm \mathbb{E}[Y_{\text{pre},n}]_2 \\
&= (\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]) \widetilde{w}^{(n,d)} + \mathbb{E}[Y_{\text{pre},n}]_2 \\
&\leq (\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]) \widetilde{w}^{(n,d)}{}_2 + \mathbb{E}[Y_{\text{pre},n}]_2 \\
&\leq \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]_{2,\infty} \widetilde{w}^{(n,d)}{}_1 + \sqrt{T_0}.
\end{aligned}$$

In the inequalities above, we use $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] \widetilde{w}^{(n,d)} = \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] w^{(n,d)} = \mathbb{E}[Y_{\text{pre},n}]$, which follows from (3.30) and (3.47). Then, (3.64) implies

$$\langle \widehat{\boldsymbol{U}}_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\text{pre}} \widehat{\boldsymbol{V}}'_{\text{pre}} \widetilde{w}^{(n,d)}, \varepsilon_{\text{pre},n} \rangle = O_p\left(\boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]_{2,\infty} \widetilde{w}^{(n,d)}{}_1 + \sqrt{T_0} \right) \tag{3.65}$$

Finally, using Lemma 5.14.3, (3.61), Assumptions 9 and 12, we have for any $\zeta > 0$

$$\mathbb{P}\left(\langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}, \varepsilon_{\text{pre},n} \rangle \geq \sigma^2 r_{\text{pre}} + \zeta\right) \leq \exp\left(- c \min\left(\frac{\zeta^2}{\sigma^4 r_{\text{pre}}}, \frac{\zeta}{\sigma^2} \right) \right), \tag{3.66}$$

where we have used $\mathcal{P}_{\widehat{U}_{\text{pre}}}{}_{\text{op}} \leq 1$ and $\mathcal{P}_{\widehat{U}_{\text{pre}}}{}^2_F = r_{\text{pre}}$. Then, (3.66) implies

$$\langle \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n}, \varepsilon_{\text{pre},n} \rangle = O_p\left(r_{\text{pre}} \right). \tag{3.67}$$

From (3.57), (3.63), (3.65), and (3.67), we conclude that

$$\langle \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \Delta^{(n,d)}, \varepsilon_{\text{pre},n} \rangle = O_p\left(r_{\text{pre}} + \sqrt{T_0} + \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]_{2,\infty} \widetilde{w}^{(n,d)}{}_1 \right).$$

■ 3.16 Proof of Theorem 4.5.2

For ease of notation, we suppress the conditioning on $\mathcal{E}$ for the remainder of the proof. Additionally, let $\mathcal{T}_{\text{post}} = \{T - T_1 + 1, \dots, T\}$ and $\Delta^{(n,d)} = \widehat{w}^{(n,d)} - \widetilde{w}^{(n,d)}$. We start by establishing (3.17). To begin, we scale the left-hand side (LHS) of (3.35) by $\sqrt{T_1}$ and analyze each of the three terms on the right-hand side (RHS) of (3.35) separately. To address the first term on the RHS of (3.35), we scale (3.36) by $\sqrt{T_1}/(\sigma\|\widetilde{w}^{(n,d)}\|_2)$ and recall our assumption on T_1 given by (3.16). We then obtain

$$\frac{1}{\sqrt{T_1}\sigma\|\widetilde{w}^{(n,d)}\|_2} \sum_{t \in \mathcal{T}_{\text{post}}} \langle \mathbb{E}[Y_{t,\mathcal{I}^{(d)}}], \mathcal{P}_{V_{\text{pre}}} \Delta^{(n,d)} \rangle = o_p(1). \tag{3.68}$$

To address the second term on the RHS of (3.35), we scale (3.37) by $\sqrt{T_1}\sigma\|\widetilde{w}^{(n,d)}\|_2$. Since $\langle \varepsilon_{t,\mathcal{I}^{(d)}}, \widetilde{w}^{(n,d)} \rangle$ are independent across t, the Lindeberg–Lévy Central Limit Theorem (see Billingsley (1986)) yields

$$\frac{1}{\sqrt{T_1}\sigma\|\widetilde{w}^{(n,d)}\|_2} \sum_{t \in \mathcal{T}_{\text{post}}} \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \widetilde{w}^{(n,d)} \rangle \xrightarrow{d} \mathcal{N}(0,1). \tag{3.69}$$

To address the third term on the RHS of (3.35), we scale (3.39) by $\sqrt{T_1}\sigma\|\widetilde{w}^{(n,d)}\|_2$ and recall the assumption $\log(T_0 N_d) = o\left(\min\{T_0, N_d\}/(C^2(\|\widetilde{w}^{(n,d)}\|_2) r_{\text{pre}}^2)\right)$. This yields

$$\frac{1}{\sqrt{T_1}\sigma\|\widetilde{w}^{(n,d)}\|_2} \sum_{t \in \mathcal{T}_{\text{post}}} \langle \varepsilon_{t,\mathcal{I}^{(d)}}, \Delta^{(n,d)} \rangle = o_p(1). \tag{3.70}$$

Finally, scaling (3.35) by $\sqrt{T_1}$ and collecting (4.19), (4.20), (4.21), we conclude

$$\frac{\sqrt{T_1}}{\sigma\|\widetilde{w}^{(n,d)}\|_2}(\widehat{\theta}_n^{(d)} - \theta_n^{(d)}) \xrightarrow{d} \mathcal{N}(0,1).$$

This establishes (3.17).

(3.18) is immediate from Lemma 4.9.1.

Next, we establish (3.19). Using the definition of $\widehat{w}^{(n,d)}$ in (3.3), $Y_{\text{pre},\mathcal{I}^{(d)}}\widehat{w}^{(n,d)} = Y^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}\widehat{w}^{(n,d)}$, where recall that $Y^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}}$ is obtained by truncating SVD of $Y_{\text{pre},\mathcal{I}^{(d)}}$ by retaining top r_{pre} components. Substituting this equality into our definition of the variance estimator in (3.7), we get $\widehat{\sigma}^2 = \frac{1}{T_0} Y_{\text{pre},n} - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)2}_{2}$. Next, consider

$$\begin{aligned} & Y_{\text{pre},n} - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)2}_{2} \\ & = \mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)2}_{2} + \varepsilon_{\text{pre},n}{}^{2}_{2} + 2\langle \varepsilon_{\text{pre},n}, (\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)}) \rangle. \end{aligned} \tag{3.71}$$

Below, we analyze each term on the RHS of (3.71) separately.

Bounding $\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)}{}_2$. In order to control this term, we incorporate Lemmas 3.15.4 and 3.15.5 into the bound in (3.52) to obtain

$$\frac{1}{\sqrt{T_0}} (\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)}) = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{\sqrt{T_0}} + \frac{r_{\text{pre}} \sqrt{\log(T_0 N_d)} \widetilde{w}^{(n,d)}{}_1}{\min\{\sqrt{T_0}, \sqrt{N_d}\}} \right). \tag{3.72}$$

Bounding $\varepsilon_{\text{pre},n2}$. Since the entries of $\varepsilon_{\text{pre},n}$ are independent mean zero sub-Gaussian random variables, it follows from Lemma 3.15.2 that

$$\frac{1}{T_0} \varepsilon_{\text{pre},n}{}^{2}_{2} - \sigma^2 = O_p\left(\frac{1}{\sqrt{T_0}} \right). \tag{3.73}$$

Bounding $\langle \varepsilon_{\text{pre},n}, (\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)}) \rangle$. From (3.56), we have $\boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)} = \mathcal{P}_{\widehat{U}_{\text{pre}}} (\mathbb{E}[Y_{\text{pre},n}] + \varepsilon_{\text{pre},n})$. Hence,

$$\begin{aligned} & \langle \varepsilon_{\text{pre},n}, (\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r\text{pre}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)}) \rangle \\ & \quad = \langle \varepsilon_{\text{pre},n}, \mathbb{E}[Y_{\text{pre},n}] \rangle - \langle \varepsilon_{\text{pre},n}, \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}] \rangle - \langle \varepsilon_{\text{pre},n}, \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n} \rangle. \end{aligned} \tag{3.74}$$

By Lemma 5.14.2 and Assumption 12, it follows that for any $\zeta > 0$

$$\mathbb{P}\left(\langle \varepsilon_{\text{pre},n}, \mathbb{E}[Y_{\text{pre},n}] \rangle \geq \zeta \right) \leq \exp\left(-\frac{c\zeta^2}{T_0 \sigma^2} \right), \tag{3.75}$$

$$\mathbb{P}\left(\langle \varepsilon_{\text{pre},n}, \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}] \rangle \geq \zeta \right) \leq \exp\left(-\frac{c\zeta^2}{T_0 \sigma^2} \right). \tag{3.76}$$

Note that we have used $\mathcal{P}_{\widehat{U}_{\text{pre}}}{}_{\text{op}} \leq 1$ and Assumption 13 to obtain $\mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}]_2 \leq \mathbb{E}[Y_{\text{pre},n}]_2 \leq \sqrt{T_0}$. Together, (3.75) and (3.76) then imply that

$$\langle \varepsilon_{\text{pre},n}, \mathbb{E}[Y_{\text{pre},n}] \rangle = O_p(\sqrt{T_0}), \quad \langle \varepsilon_{\text{pre},n}, \mathcal{P}_{\widehat{U}_{\text{pre}}} \mathbb{E}[Y_{\text{pre},n}] \rangle = O_p(\sqrt{T_0}). \tag{3.77}$$

From (3.66), we have for any $\zeta > 0$,

$$\mathbb{P}\left(\langle \varepsilon_{\text{pre},n}, \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n} \rangle \geq \sigma^2 r_{\text{pre}} + \zeta\right) \leq \exp\left(- c \min\left(\frac{\zeta^2}{\sigma^4 r_{\text{pre}}}, \frac{\zeta}{\sigma^2} \right)\right),$$

which implies that

$$\langle \varepsilon_{\text{pre},n}, \mathcal{P}_{\widehat{U}_{\text{pre}}} \varepsilon_{\text{pre},n} \rangle = O_p(r_{\text{pre}}). \tag{3.78}$$

Plugging (3.77) and (3.78) into (3.74), we obtain

$$\frac{1}{T_0} \langle \varepsilon_{\text{pre},n}, (\mathbb{E}[Y_{\text{pre},n}] - \boldsymbol{Y}^{r_{\text{pre}}}_{\text{pre},\mathcal{I}^{(d)}} \widehat{w}^{(n,d)}) \rangle = O_p\left(\frac{r_{\text{pre}}}{T_0} + \frac{1}{\sqrt{T_0}} \right). \tag{3.79}$$

Collecting terms. Normalizing (3.71) by T_0 and subsequently incorporating the bounds in (3.72), (3.73), and (3.79), we conclude

$$\widehat{\sigma}^2 - \sigma^2 = O_p\left(\frac{\sqrt{r_{\text{pre}}}}{\sqrt{T_0}} + \frac{r_{\text{pre}}\sqrt{\log(T_0 N_d)} \widetilde{w}^{(n,d)}{}_1}{\min\{\sqrt{T_0}, \sqrt{N_d}\}} \right).$$

■ 3.17 Proof of Theorem 3.6.1

For ease of notation, we suppress the conditioning on $\mathcal{E}$ for the remainder of the proof. We make use of the following notation: for any matrix $\boldsymbol{A}$ with orthonormal columns, let $\mathcal{P}_A = \boldsymbol{A}\boldsymbol{A}'$ denote the projection matrix onto the subspace spanned by the columns of $\boldsymbol{A}$. Additionally, we follow the notation established in Section 3.6.1. In particular, we recall $\phi_{\text{pre}}(a) = \sqrt{T_0} + \sqrt{N_d} + \sqrt{\log(1/a)}$; $\phi_{\text{post}}(a) = \sqrt{T_1} + \sqrt{N_d} + \sqrt{\log(1/a)}$; and s_ℓ, ς_ℓ are the ℓ-th singular values of $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]$ and $\mathbb{E}[\boldsymbol{Y}_{\text{post},\mathcal{I}^{(d)}}]$, respectively. Finally, we let $C \geq 0$ denote an absolute constant, whose value can change from line to line or even within a line.

Type I error. We first bound the Type I error, which anchors on Lemma 3.17.1, stated below. The proof of Lemma 3.17.1 can be found in Section 3.17.2.

Lemma 3.17.1. *Suppose H_0 is true. Then,*

$$\begin{aligned}\widehat{\tau} = (\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}_F^2 + &(\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})_F^2 \\ &+ 2\langle(\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}, (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F.\end{aligned} \tag{3.80}$$

We proceed to bound each term on the right-hand side of (3.80) independently.

Bounding $(\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}_F^2$. By Lemma 3.15.3, we have w.p. at least $1 - \alpha_1$,

$$(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}_F^2 \leq \mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}{}_{\text{op}}^2 \;\; \widehat{\boldsymbol{V}}_{\text{post}}{}_F^2 \leq \frac{C\sigma^2 r_{\text{post}}\phi_{\text{pre}}^2(\alpha_1)}{s_{r_{\text{pre}}}^2}. \tag{3.81}$$

Note that we have used the fact that $\widehat{\boldsymbol{V}}_{\text{post}}{}_F^2 = r_{\text{post}}$.

Bounding $(\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})_F^2$. Observe that $(\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})$ is a projection matrix, and hence $\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}}{}_{\text{op}} \leq 1$. By adapting Lemma 3.15.3 for $\widehat{\boldsymbol{V}}_{\text{post}}, \boldsymbol{V}_{\text{post}}$ in place of $\widehat{\boldsymbol{V}}_{\text{pre}}, \boldsymbol{V}_{\text{pre}}$, we have w.p. at least $1 - \alpha_2$

$$\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}{}_F^2 \leq r_{\text{post}}\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}{}_{\text{op}}^2 \leq \frac{C\sigma^2 r_{\text{post}}\phi_{\text{post}}^2(\alpha_2)}{\varsigma_{r_{\text{post}}}^2}. \tag{3.82}$$

Note that we have used the following: (i) $\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}{}_F = \sin\Theta_F$, where $\sin\Theta \in \mathbb{R}^{r_{\text{post}} \times r_{\text{post}}}$ is a matrix of principal angles between the two projectors (see Absil et al. (2006)), which implies $\text{rank}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}) \leq r_{\text{post}}$; (ii) the standard norm inequality $\boldsymbol{A}_F \leq \sqrt{\text{rank}(\boldsymbol{A})}\boldsymbol{A}_{\text{op}}$ for any matrix $\boldsymbol{A}$. Using the result above, we have

$$(\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})_F^2 \leq \boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}}{}_{\text{op}}^2 \;\; \mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}{}_F^2 \leq \frac{C\sigma^2 r_{\text{post}}\phi_{\text{post}}^2(\alpha_2)}{\varsigma_{r_{\text{post}}}^2}. \tag{3.83}$$

Bounding $\langle(\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}},\ (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F$. Using the cyclic property of the trace operator, we have that

$$\begin{aligned}\langle(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}},\ (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F &= \text{tr}\,(\widehat{\boldsymbol{V}}'_{\text{post}}(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})(\boldsymbol{I} - \mathcal{P}_{V_{pre}})\widehat{\boldsymbol{V}}_{\text{post}}) \\ &= \text{tr}\,((\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})(\boldsymbol{I} - \mathcal{P}_{V_{pre}})\mathcal{P}_{\widehat{V}_{\text{post}}}).\end{aligned} \tag{3.84}$$

Note that $\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}$ is symmetric, and $\boldsymbol{I} - \mathcal{P}_{V_{pre}}$ and $\mathcal{P}_{\widehat{V}_{\text{post}}}$ are both symmetric positive semidefinite (PSD). As a result, Lemmas 3.15.3 and 3.17.6 yield w.p. at least $1 - \alpha_1$

$$\begin{aligned} &\operatorname{tr}\left((\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})(\boldsymbol{I} - \mathcal{P}_{V_{pre}})\mathcal{P}_{\widehat{V}_{\text{post}}}\right) \leq \mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}{}_{\text{op}} \operatorname{tr}\left((\boldsymbol{I} - \mathcal{P}_{V_{pre}})\mathcal{P}_{\widehat{V}_{\text{post}}}\right) \\ &\leq \mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}{}_{\text{op}} \boldsymbol{I} - \mathcal{P}_{V_{pre}}{}_{\text{op}} \operatorname{tr}\left(\mathcal{P}_{\widehat{V}_{\text{post}}}\right) \leq \frac{C\sigma r_{\text{post}} \phi_{\text{pre}}(\alpha_1)}{s_{r_{\text{pre}}}}. \end{aligned} \tag{3.85}$$

Again, to arrive at the above inequality, we use $\boldsymbol{I} - \mathcal{P}_{V_{pre}}{}_{\text{op}} \leq 1$ and $\operatorname{tr}(\mathcal{P}_{\widehat{V}_{\text{post}}}) = r_{\text{post}}$.

Collecting terms. Collecting (3.81), (3.83), and (3.85) with $\alpha_1 = \alpha_2 = \alpha/2$, w.p. at least $1 - \alpha$,

$$\widehat{\tau} \leq \frac{C\sigma^2 r_{\text{post}} \phi^2_{\text{pre}}(\alpha/2)}{s^2_{r_{\text{pre}}}} + \frac{C\sigma^2 r_{\text{post}} \phi^2_{\text{post}}(\alpha/2)}{\varsigma^2_{r_{\text{post}}}} + \frac{C\sigma r_{\text{post}} \phi_{\text{pre}}(\alpha/2)}{s_{r_{\text{pre}}}}.$$

Defining the upper bound as $\tau(\alpha)$ completes the bound on the Type I error.

Type II error. Next, we bound the Type II error. We will leverage Lemma 3.17.2, the proof of which can be found in Section 3.17.3.

Lemma 3.17.2. *The following equality holds:* $\widehat{\tau} = r_{\text{post}} - c_1 - c_2$, *where*

$$\begin{aligned} c_1 &= \boldsymbol{V}_{pre} \boldsymbol{V}'_{pre} \boldsymbol{V}_{post}{}^2_F \\ c_2 &= (\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}) \widehat{\boldsymbol{V}}_{\text{post}}{}^2_F + \mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})^2_F \\ &\quad + 2\langle (\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}) \widehat{\boldsymbol{V}}_{\text{post}}, \mathcal{P}_{V_{\text{pre}}} \widehat{\boldsymbol{V}}_{\text{post}} \rangle_F + 2\langle \mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}), \mathcal{P}_{V_{\text{pre}}} \mathcal{P}_{V_{\text{post}}} \rangle_F. \end{aligned} \tag{3.86}$$

We proceed to bound each term on the right hand side of (3.86) separately.

Bounding $(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}) \widehat{\boldsymbol{V}}_{\text{post}}{}^2_F$. From (3.81), we have that w.p. at least $1 - \alpha_1$,

$$(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}}) \widehat{\boldsymbol{V}}_{\text{post}}{}^2_F \leq \frac{C\sigma^2 r_{\text{post}} \phi^2_{\text{pre}}(\alpha_1)}{s^2_{r_{\text{pre}}}}. \tag{3.87}$$

Bounding $\mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})^2_F$. Using the inequality $\boldsymbol{A}\boldsymbol{B}_F \leq \boldsymbol{A}_{\text{op}} \boldsymbol{B}_F$ for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, as well as the bound in (3.82), we have w.p. at least $1 - \alpha_2$,

$$\mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})^2_F \leq \frac{C\sigma^2 r_{\text{post}} \phi^2_{\text{post}}(\alpha_2)}{\varsigma^2_{r_{\text{post}}}}. \tag{3.88}$$

Bounding $\langle(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}, \mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F$. Using an identical argument used to create the bounds in (3.84) and (3.85), but replacing $\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}}$ with $\mathcal{P}_{V_{\text{pre}}}$, we obtain w.p. at least $1 - \alpha_1$

$$\langle(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}, \mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F \le \frac{C\sigma r_{\text{post}}\phi_{\text{pre}}(\alpha_1)}{s_{r_{\text{pre}}}}. \tag{3.89}$$

Bounding $\langle\mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}), \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}\rangle_F$. Like in the argument to produce the bound in (3.85), we use Lemmas 3.15.3 and 3.17.6 to get that w.p. at least $1 - \alpha_2$,

$$\begin{aligned}&\langle\mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}), \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}\rangle_F = \operatorname{tr}((\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})\mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}) \\ &= \operatorname{tr}((\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})\mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}) \le \mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}\text{op}}\ \mathcal{P}_{V_{\text{pre}}\text{op}}\operatorname{tr}(\mathcal{P}_{V_{\text{post}}}) \le \frac{C\sigma r_{\text{post}}\phi_{\text{post}}(\alpha_2)}{\varsigma_{r_{\text{post}}}}.\end{aligned} \tag{3.90}$$

Collecting terms. Combining (3.87), (3.88), (3.89), (3.90) with $\alpha_1 = \alpha_2 = \alpha/2$, and using the definition of $\tau(\alpha)$, we have that w.p. at least $1 - \alpha$,

$$c_2 \le \tau(\alpha) + \frac{C\sigma r_{\text{post}}\phi_{\text{post}}(\alpha/2)}{\varsigma_{r_{\text{post}}}}.$$

Hence, along with using Lemma 3.17.2, it follows that w.p. at least $1 - \alpha$,

$$\widehat{\tau} \ge r_{\text{post}} - c_1 - \tau(\alpha) - \frac{C\sigma r_{\text{post}}\phi_{\text{post}}(\alpha/2)}{\varsigma_{r_{\text{post}}}}. \tag{3.91}$$

Now, suppose r_{post} satisfies (3.21), which implies that H_1 must hold. Then, (3.91) and (3.21) together imply $\mathbb{P}(\widehat{\tau} > \tau(\alpha)|H_1) \ge 1 - \alpha$. This completes the proof.

■ 3.17.1 Proof of Corollary 3.6.1

We utilize Lemmas 3.17.3 and 3.17.4, which are sharp versions of Lemmas 3.13.3 and 3.15.3.

Lemma 3.17.3 (Gaussian Matrices: Theorem 7.3.1 of Vershynin (2018))**.** *Let the setup of Lemma 3.13.3 hold. Assume A_{ij} are Gaussian r.v.s with variance σ^2. Then for any $t > 0$, $A_{\text{op}} \le \sigma(\sqrt{m} + \sqrt{n} + t)$ w.p. at least $1 - 2\exp(-t^2)$.*

Lemma 3.17.4. *Let the setup of Lemma 3.15.3 hold. Further, assume ε_{tn} are Gaussian* *r.v.s with variance σ^2. Then for any $\alpha \in (0,1)$, we have w.p. at least $1-\alpha$,*

$$\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}\,\text{op}} \le \frac{2\sigma\phi_{\text{pre}}(\alpha)}{s_{r_{\text{pre}}}}, \quad \mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}\,\text{op}} \le \frac{2\sigma\phi_{\text{post}}(\alpha)}{\varsigma_{r_{\text{post}}}}.$$

Proof. The proof is identical to that of Lemma 3.15.3 except $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} - \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}]_{\text{op}}$ is now bounded above using Lemma 3.17.3. ■

The remainder of the proof of Corollary 3.6.1 is identical to that of Theorem 3.6.1.

■ 3.17.2 Proof of Lemma 3.17.1

$\widehat{\tau} = (\boldsymbol{I} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F = (\boldsymbol{I} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}} - (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}} + (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F = (\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}} + (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F = (\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F + (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F + 2\langle(\mathcal{P}_{V_{\text{pre}}} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}, (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F.$

Under H_0, it follows that $(\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\boldsymbol{V}_{\text{post}} = 0$. As a result, $(\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F = (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\mathcal{P}_{\widehat{V}_{\text{post}}}{}^2_F = (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\mathcal{P}_{\widehat{V}_{\text{post}}} - (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})\mathcal{P}_{V_{\text{post}}}{}^2_F = (\boldsymbol{I} - \mathcal{P}_{V_{\text{pre}}})(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})^2_F$.

Applying these two sets of equalities above together completes the proof.

■ 3.17.3 Proof of Lemma 3.17.2

Because the columns of $\widehat{\boldsymbol{V}}_{\text{post}}$ are orthonormal, $r_{\text{post}} = \widehat{\boldsymbol{V}}_{\text{post}}{}^2_F = \mathcal{P}_{\widehat{V}_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F + (\boldsymbol{I} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F$. Therefore, it follows that

$$\widehat{\tau} = (\boldsymbol{I} - \mathcal{P}_{\widehat{V}_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F = r_{\text{post}} - \mathcal{P}_{\widehat{V}_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F. \tag{3.92}$$

Now, consider the second term of the equality above.

$$\begin{aligned}\mathcal{P}_{\widehat{V}_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F &= \mathcal{P}_{\widehat{V}_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}} - \mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}} + \mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F \\ &= (\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F + \mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F + 2\langle(\mathcal{P}_{\widehat{V}_{\text{pre}}} - \mathcal{P}_{V_{\text{pre}}})\widehat{\boldsymbol{V}}_{\text{post}}, \mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}\rangle_F. \end{aligned} \tag{3.93}$$

Further, analyzing the second term of (3.93), we note that

$$\begin{aligned}\mathcal{P}_{V_{\text{pre}}}\widehat{\boldsymbol{V}}_{\text{post}}{}^2_F &= \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{\widehat{V}_{\text{post}}}{}^2_F = \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}} + \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}{}^2_F \\ &= \mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}})^2_F + \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}{}^2_F + 2\langle\mathcal{P}_{V_{\text{pre}}}(\mathcal{P}_{\widehat{V}_{\text{post}}} - \mathcal{P}_{V_{\text{post}}}), \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}\rangle_F. \end{aligned} \tag{3.94}$$

Incorporating (3.93) and (3.94) into (3.92), and recalling $c_1 = \mathcal{P}_{V_{\text{pre}}}\mathcal{P}_{V_{\text{post}}}{}_F^2 = \boldsymbol{V}_{\text{pre}}\boldsymbol{V}'_{\text{pre}}\boldsymbol{V}_{\text{post}}{}_F^2$ completes the proof.

■ 3.17.4 Helper Lemmas

Lemma 3.17.5. *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric PSD matrices. Then,* $\text{tr}(\boldsymbol{A}\boldsymbol{B}) \geq 0$.

Proof. Let $\boldsymbol{B}^{1/2}$ denote the square root of $\boldsymbol{B}$. Since $\boldsymbol{A} \succeq 0$, we have $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}\big(\boldsymbol{A}\boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2}\big) = \text{tr}\big(\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{B}^{1/2}\big) = \sum_{i=1}^{n}(\boldsymbol{B}^{1/2}e_i)'\boldsymbol{A}(\boldsymbol{B}^{1/2}e_i) \geq 0$, ■

Lemma 3.17.6. *If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is a symmetric PSD matrix, then* $\text{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{A}) \cdot \text{tr}(\boldsymbol{B})$, *where $\lambda_{\max}(\boldsymbol{A})$ is the top eigenvalue of $\boldsymbol{A}$.*

Proof. Since $\boldsymbol{A}$ is symmetric, it follows that $\lambda_{\max}(\boldsymbol{A})\boldsymbol{I} - \boldsymbol{A} \succeq 0$. As a result, applying Lemma 3.17.5 yields $\text{tr}((\lambda_{\max}(\boldsymbol{A})\boldsymbol{I} - \boldsymbol{A})\boldsymbol{B}) = \lambda_{\max}(\boldsymbol{A}) \cdot \text{tr}(\boldsymbol{B}) - \text{tr}(\boldsymbol{A}\boldsymbol{B}) \geq 0$. ■

Chapter 4

Causal Matrix Completion

■ 4.1 Introduction

Matrix completion is the study of recovering an underlying matrix from its noisy and partial observations. Given its widespread applicability, the field of matrix completion has grown tremendously in recent years. To establish statistical guarantees for the various algorithms that exist for matrix completion, it is typically assumed that: (i) the underlying noiseless matrix has latent structure, e.g., it is low-rank, and (ii) the entries of this matrix are missing completely at random (MCAR), i.e., an entry is missing independent of everything else and with uniform probability. However, numerous modern applications of interest violate the latter assumption. Below, we consider two motivating examples.

First, arguably the most well-known application of matrix completion is recommender systems, which are ubiquitous in modern online platforms. Typically, data is collected in the form of a matrix, where the rows index users and columns index items; the (i, j)-th entry, therefore, corresponds to the rating supplied by user i for item j. In such scenarios, observations are often subject to *selection-biases*. For instance, in movie recommendations, a fan of fantasy fiction will almost certainly watch and highly rate the *Harry Potter* series. Similarly, in restaurant recommendations, a vegetarian is unlikely to enjoy nor rate a steakhouse restaurant. While these examples demonstrate self-selection biases from the end of the users, systems also exhibit targeted suggestions. For example, when a user searches for trails at the Grand Canyon, an ad placement system is more likely to display an ad for hiking boots than wedding shoes; in turn, this can increase the user's likelihood to purchase and rate hiking boots. In all of these cases, the user's preferences and/or the system's beliefs in its users' preferences, influence the sparsity pattern of the observation matrix.

A second example is panel data settings in econometrics. Here, observations of units (e.g., individuals, geographic locations) are collected over time as they undergo different interventions (e.g., promotions, socio-economic policies). The induced matrix has rows index units and columns index time-intervention pairs; the $(i, (a, t))$-th entry then corresponds to the potential outcome of unit i under the a-th intervention at time step t; here (a, t) represents the j-th column, i.e., columns are double indexed by both intervention and time. As with recommender systems, observations in panel data settings are unlikely to occur completely at random. For instance, policy-makers strategically recommend programs that are designed to achieve certain desirable outcomes based on numerous socio-economic factors surrounding the geographic region under their purview. Further, competing programs with disagreeing agendas cannot be simultaneously adopted for a specific region during the same time period, i.e., if the $(i, (a, t))$-th entry is observed, then the $(i, (a', t))$-th entry must be missing. Notably, similar matrices and observation patterns can arise in sequential decision-making paradigms within machine learning such as online learning, contextual bandits, and reinforcement learning with time-intervention pairs being replaced by state-action pairs.

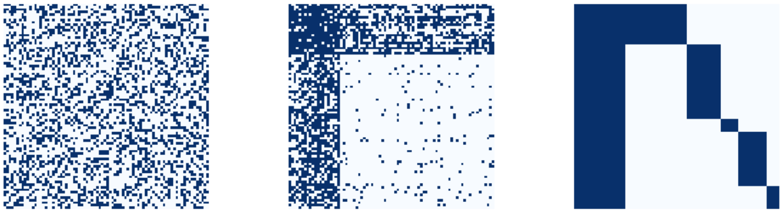
In both examples, the missingness pattern of the matrix is dependent on the underlying values in that matrix, and observing the outcome of one entry can alter the probability of observing another. That is, the entries are missing *not* at random (MNAR). To address the above challenges, there has been exciting recent progress on matrix completion with MNAR data, including Schnabel et al. (2016); Ma and Chen (2019); Zhu et al. (2019); Sportisse et al. (2020a,b); Wang et al. (2020); Yang et al. (2021); Bhattacharya and Chatterjee (2021). Through numerous empirical studies, these works have shown that algorithms that account for MNAR data outperform conventional algorithms that are designed for MCAR data. With respect to theoretical analysis, however, critical aspects of matrix completion with MNAR data remain to be explored. In particular, as highlighted in Ma and Chen (2019), there are two common limiting assumptions in the literature: (i) the revelation of each entry in the matrix is independent of all other entries, and (ii) each entry has a nonzero probability of being observed.

Another recent exciting line of work that we build upon is that of panel data and matrix completion, see Amjad et al. (2018, 2019); Arkhangelsky et al. (2019); Bai and Ng (2020); Fernández-Val et al. (2020); Athey et al. (2021); Agarwal et al. (2019b, 2021e,d,c); Agarwal and Singh (2021). Some of these works allow for MNAR data and entries of a matrix to be deterministically missing. However, they consider very restricted sparsity patterns that are not particularly suitable for important applications of matrix completion. For example, the most common sparsity pattern considered in the panel data literature is where for a given row i, if a column j is missing, then entries for all columns $j' > j$ in row i are also missing; such a pattern is unlikely to arise in recommendation systems or sequential decision-making. Further, to the best of our knowledge, none of these works within the panel data literature provide meaningful results for matrix completion with MCAR data. The statistical parameters these works aim to estimate are also less meaningful for these other applications of matrix completion. The most common statistical parameter these works consider is the average outcome for all missing entries in a given row i; in say recommendation systems, this would correspond to the average rating a user i would have given for all movies they did not rate. This is not particularly meaningful for an online platform—ideally, a platform would like to do accurate inference for each (i, j) pair.

The focus of this work is to propose a formal causal framework and an algorithm with provable guarantees to analyze matrix completion with MNAR data where the probability that an entry of the matrix is missing can: (i) depend on the underlying values in the matrix itself; (ii) depend on which other entries are missing; (iii) potentially be deterministically zero. Further, we want to allow for more general missingness patterns and estimate more refined statistical parameters than considered in the panel data literature thus far.

■ 4.1.1 How the Missingness Mechanism can Bias Inference: A Teaser

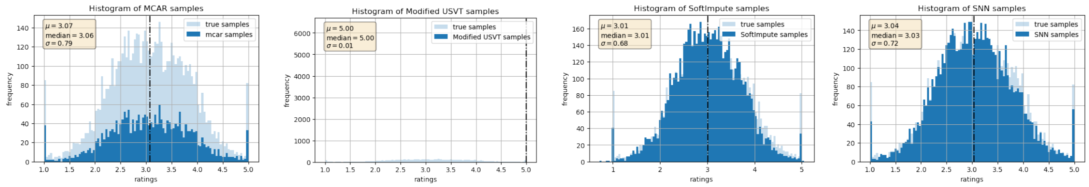
As further motivation for why it is important to carefully think about the underlying mechanism for why data is missing, we now provide illustrative empirical simulations. In particular we run three experiments, each with a different mechanism for how data is missing. In Experiment 1, data is missing via a MCAR mechanism i.e., each entry is missing independently at random with probability 0.35; the induced sparsity pattern is depicted in Figure 4.1a. In Experiment 2, data is missing in a MNAR fashion, i.e., each entry has a different probability of being missing; the induced sparsity pattern is depicted in Figure 4.1b. However, we ensure key assumptions made thus far in the matrix completion literature with MNAR data are maintained; in particular, (i) the revelation of entries are entry-wise independent and (ii) each entry has a nonzero probability of being observed. In Experiment 3, data is missing in a MNAR fashion, but we violate conditions (i) and (ii) above; the induced sparsity pattern is depicted in Figure 4.1c. For exact details on the missingness mechanism in Experiment 2 and 3, refer to Section 4.6.2 and 4.6.2, respectively.



(a) MCAR. **(b)** Limited MNAR. **(c)** General MNAR.

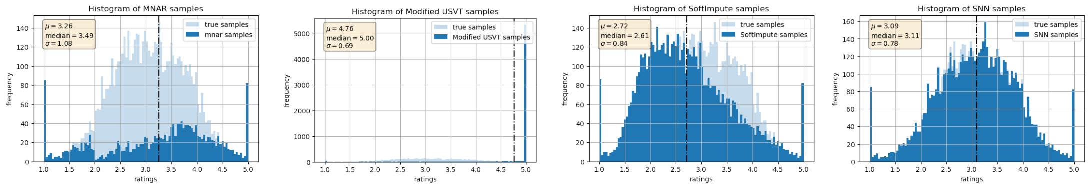
Figure 4.1: Empirical sparsity pattern under different missingess mechanisms.

In all experiments, we first create a sample of true "ratings", which are invariant across all three experiments. We enforce these ratings to go from 1 to 5, as is standard in many online platforms. The distribution of true/revealed ratings are plotted in light/dark blue in Figures 4.2a, 4.3a, and 4.4a, respectively. As expected, the distribution of the revealed ratings in the MCAR setup matches that of the true ratings. However, the set of ratings that are revealed in both MNAR settings are severely biased, i.e., their distribution does not match that of the true underlying ratings.



(a) True, revealed. **(b)** (Modified) `USVT`. **(c)** (Modified) `softImpute`. **(d)** `SNN`.

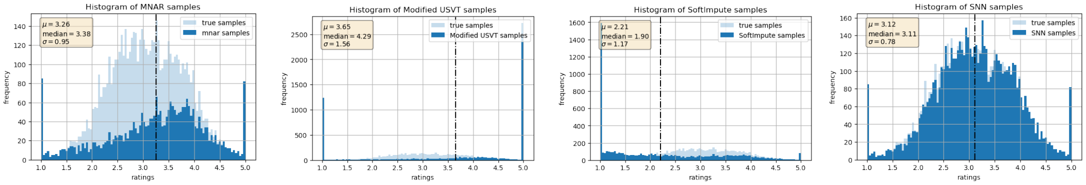
Figure 4.2: MCAR: recovered ratings distributions under (modified) `USVT`, (modified) `softImpute`, and `SNN`.



(a) True, revealed. **(b)** (Modified) `USVT`. **(c)** (Modified) `softImpute`. **(d)** `SNN`.

Figure 4.3: Limited MNAR: recovered ratings distributions under (modified) `USVT`, (modified) `softImpute`, and `SNN`.

We use three matrix completion algorithms and see whether they can recover the distribution of true ratings given the revelead entries in all three experiments. The algorithms are: Universal singular value thresholding (`USVT`) Chatterjee (2015), which is a popular



(a) True, revealed. **(b)** (Modified) `USVT`. **(c)** (Modified) `softImpute`. **(d)** `SNN`.

Figure 4.4: More general MNAR: recovered ratings distributions under (modified) `USVT`, (modified) `softImpute`, and `SNN`.

spectral based method;[1] Softimpute (`softImpute`) Hastie et al. (2015), which is a popular optimization based method; Synthetic nearest neighbours (`SNN`), which is our proposed method for matrix completion with MNAR data, and is a combination of the approach taken in nearest neighbour style and panel data methods in econometrics. `USVT` and `softImpute` are not designed for MNAR data, as is, but we de-bias them for MNAR data as is done in Bhattacharya and Chatterjee (2021) and Ma and Chen (2019), respectively. See details in Section 4.6.3.

We see that in Figure 4.2, under the MCAR setting, `softImpute` and `SNN` both recover the distribution of true ratings very well, while `USVT` cannot. Once we go to the limited MNAR setting, depicted in Figure 4.3, where conditions (i) and (ii) are upheld, `SNN` is still able to recover the underlying distribution of true ratings, but now both `softImpute` and `USVT` have non-negligible bias. In the general MNAR setting, depicted in Figure 4.4, `SNN` continues to accurately recover the distribution, but the bias of `softImpute` is significantly worsened.

This empirical illustration highlights the sensitivity of these traditional matrix completion methods to the missingness mechanism and strongly motivates the need for a rigorous framework for tackling the general MNAR setting where conditions (i) and (ii) above are violated. Providing such a framework is what we set out to do in this work.

■ 4.1.2 Problem Statement

We now formally introduce our setup. Consider a signal matrix $\boldsymbol{A} = [A_{ij}] \in \mathbb{R}^{m \times n}$, a noise matrix $\boldsymbol{E} = [\varepsilon_{ij}] \in \mathbb{R}^{m \times n}$, and a propensity score matrix $\boldsymbol{P} = [p_{ij}] \in [0, 1]^{m \times n}$. All three matrices are entirely latent, i.e., unobserved. Let $\boldsymbol{Y} = [Y_{ij}] \in \mathbb{R}^{m \times n}$ denote the

[1] Surprisingly, we find that the original `USVT` algorithm performs better in all three experiments. See Section 4.7.

"noisy" version of $\boldsymbol{A}$, with $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{A}$; we denote $\varepsilon_{ij} = Y_{ij} - A_{ij}$. We assume $\boldsymbol{Y}$ itself is partially observed. In particular, we denote $\boldsymbol{D} = [D_{ij}] \in \{0,1\}^{m \times n}$ with $\mathbb{E}[\boldsymbol{D}] = \boldsymbol{P}$ as the missingness mask matrix that indicates which entries of $\boldsymbol{Y}$ are observed. For convenience, we encode our observations into $\widetilde{\boldsymbol{Y}} = [\widetilde{Y}_{ij}] \in \{\mathbb{R} \cup \{\star\}\}^{m \times n}$ such that for $(i,j) \in [m] \times [n]$,

$$\widetilde{Y}_{ij} = \begin{cases} Y_{ij}, \text{ if } D_{ij} = 1 \\ \star, \text{ otherwise.} \end{cases} \tag{4.1}$$

In words, if $D_{ij} = 1$ then A_{ij} is noisily observed, and if $D_{ij} = 0$ then A_{ij} remains unknown. For concreteness, let us return to the recommender system example. Here, $\boldsymbol{A}$ represents the expected rating for every user-item pair and $\boldsymbol{P}$ dictates the probability these expected ratings are revealed, both of which are unknown. $\boldsymbol{Y}$ in relation to $\boldsymbol{A}$ then models the inherent randomness in how users rate items; that is, $\boldsymbol{Y}$ can be interpreted as a "noisy" instance of $\boldsymbol{A}$. Another interpretation of what ε_{ij} represents is that many online platform only allow users to input integer valued ratings (e.g. integer between 1 to 5 or a binary 0/1). Hence, Y_{ij} can be interpreted as a "noisy" *discretized* observation of A_{ij}, which may actually be *continuous* (i.e., lie within the continuous interval $[1,5]$ or $[0,1]$). Observationally, we have access to $\boldsymbol{D}$ and $\widetilde{\boldsymbol{Y}}$; the former refers to the collection of ratings users have supplied to the system while the latter refers to the corresponding realized "noisy" ratings. Finally, we remark that (4.1) also agrees with standard panel data setups in econometrics, where each observation is assumed to be corrupted by an idiosyncratic shock, which is represented by ε_{ij}.

In terms of the type of MNAR data this work considers, we allow for $\boldsymbol{D}$ and $\boldsymbol{Y}$ to be dependent, provided $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{A}$, where $\boldsymbol{A}$ is latent. In fact, we allow $\boldsymbol{D}$ to be any arbitrary function of $\boldsymbol{A}$, random or deterministic, subject to suitable observation patterns which we discuss in the forthcoming sections. Notably, our framework also allows the entries in $\boldsymbol{D}$ to be dependent with each other across both rows and columns, and the minimum value of $\boldsymbol{P}$ to be 0, which are important departures from the current matrix completion literature. Under these conditions, we propose an algorithm that provably recovers $\boldsymbol{A}$ from $\widetilde{\boldsymbol{Y}}$ with entry-wise (i.e., max-norm) guarantees.

■ 4.1.3 Contributions & Paper Organization

Section 4.2: Related works. We provide an overview of the current literature on matrix completion under the different models of missingness proposed by Rubin (1976); Little and

Rubin (2019): (i) missing completely at random (MCAR); (ii) missing at random (MAR); (iii) missing not at random (MNAR). We note the usage of the terms MCAR, MAR, and MNAR is inconsistent across the previous works on matrix completion, and so we hope that our literature survey helps give a more comprehensive and unified overview of the different regimes of missingness considered in these works.

Section 4.3: Causal framework for matrix completion. We propose a formal causal framework for matrix completion using the language of potential outcomes, see Neyman (1923); Rubin (1974). We interpret $\boldsymbol{Y}$ as the matrix of potential outcomes and $\boldsymbol{P}$ as the matrix of intervention assignments. Building upon the recent work of Agarwal et al. (2021c), we propose a framework that allows (i) correlation between $\boldsymbol{D}$ and $\boldsymbol{Y}$, i.e., hidden confounding; (ii) correlation between the entries of $\boldsymbol{D}$; (iii) the minimum value of $\boldsymbol{P}$ to be 0, i.e., entries of $\widetilde{\boldsymbol{Y}}$ can be deterministically missing; (iv) $\boldsymbol{P}$ to not exhibit low-dimensional structure as is required in the panel data literature, i.e., we consider significantly more general missingness patterns. To the best of our knowledge, our framework, and associated algorithm, is the first within the MNAR matrix completion literature that allows for conditions (i)-(iv) to simultaneously hold. Additionally, we do *not* make any parametric or distributional assumptions on $\boldsymbol{P}$, as is common in previous works on matrix completion. Nevertheless, we establish an identification result in Theorem 4.3.1, which effectively states that $\boldsymbol{A}$ can be learned from $\widetilde{\boldsymbol{Y}}$ in an entry-wise sense. We believe our proposed framework provides a unified causal view for a variety of applications that can be posed as matrix completion problems with MNAR data.

Section 4.4: An algorithmic solution. We combine the nearest neighbours approach for matrix completion —popularly known as collaborative filtering—with the synthetic controls approach for panel data, to design a novel two-step algorithm, which we call "synthetic nearest neighbors" (SNN), to estimate $\boldsymbol{A}$ from $\widetilde{\boldsymbol{Y}}$. Pleasingly, each step of SNN enjoys a simple closed-form solution. In order to efficiently execute SNN in practice, we provide an algorithm to automatically find the "neighbors" for any (i, j) pair in a data-driven manner. To do so, we relate this task to the well-known problem of finding the "maximum" biclique in a bipartite graph. Since SNN is a generalization of the recently proposed synthetic interventions (SI) estimator of Agarwal et al. (2021c), which itself is a generalization of the popular synthetic controls algorithm of Abadie et al. (2010); Abadie and Gardeazabal (2003), this subroutine may be of independent interest to the synthetic controls and panel data literatures.

Section 4.5: Theoretical results. We establish entry-wise finite-sample consistency and asymptotic normality of SNN, i.e., we provide theoretical guarantees for A_{ij} for each (i, j) pair. Hence, our analysis implies new theoretical results, in a max-norm sense, for the literature on matrix completion with MNAR data. As a special case, this also provides novel entry-wise finite-sample consistency and asymptotic normality results for the traditional matrix completion with MCAR data literature. Collectively, our identification, consistency, and asymptotic normality results, coupled with SNN, can be seen as a generalization of the SI framework proposed in Agarwal et al. (2021c).

Section 5.6: Experimental validation. We run comprehensive experiments, both with simulated and real-world data, to test the empirical efficacy of SNN against a collection of state-of-the-art matrix completion algorithms for MNAR data. Some key takeaways are as follows: (i) SNN is robust to the various forms of missingness across all experiments, while the previous methods are relatively sensitive to it. (ii) we find the approaches to de-bias estimators for MNAR data are not particularly effective, i.e., their performance is similar to their MCAR analogues; this is in line with the empirical findings of Ma and Chen (2019).

■ 4.1.4 Notations

For a matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, we denote its operator (spectral), nuclear, Frobenius, and max element-wise norms as $\boldsymbol{X}_2$, $\boldsymbol{X}_*$, $\boldsymbol{X}_F$, and $\boldsymbol{X}_{\max}$, respectively. For a matrix $\boldsymbol{X}$ with orthonormal columns, let $\mathcal{P}_X = \boldsymbol{X}\boldsymbol{X}^T$ denote the projection matrix onto the subspace spanned by the columns of $\boldsymbol{X}$. For a vector $v \in \mathbb{R}^m$, let v_p denote its ℓ_p-norm. For a random variable v, we define its sub-gaussian (Orlicz) norm as v_{ψ_2}. Let $\circ$ denote component-wise multiplication and let $\otimes$ denote the outer product. For a positive integer a, let $[a] = \{1, \ldots, a\}$. For index sets $\mathcal{I}_1 \subseteq [m]$ and $\mathcal{I}_2 \subseteq [n]$, let $\boldsymbol{X}_{\mathcal{I}_1, \mathcal{I}_2}$ denote the $|\mathcal{I}_1| \times |\mathcal{I}_2|$ sub-matrix of $\boldsymbol{X}$ whose rows and columns are indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. As a shorthand, let $\boldsymbol{X}_{\mathcal{I}_1, \cdot}$ denote the $|\mathcal{I}_1| \times n$ sub-matrix of $\boldsymbol{X}$ that retains the columns of $\boldsymbol{X}$ but only considers those rows indexed by $\mathcal{I}_1$; we define $\boldsymbol{X}_{\cdot, \mathcal{I}_2}$ analogously. Unless stated otherwise, we index rows with $i \in [m]$ and columns with $j \in [n]$.

Let f and g be two functions defined on the same space. We say $f(n) = O(g(n))$ if and only if there exists a positive real number M and a real number n_0 such that for all $n \geq n_0, |f(n)| \leq M|g(n)|$. Analogously we say: $f(n) = \Theta(g(n))$ if and only if there exists positive real numbers m, M such that for all $n \geq n_0,\ m|g(n)| \leq |f(n)| \leq M|g(n)|$; $f(n) = o(g(n))$ if for any $m > 0$, there exists n_0 such that for all $n \geq n_0, |f(n)| \leq m|g(n)|$.

We adopt the standard notations and definitions for stochastic convergences. As such, we denote $\xrightarrow{d}$ and $\xrightarrow{p}$ as convergences in distribution and probability, respectively. We will also make use of O_p and o_p, which are probabilistic versions of the commonly used deterministic O and o notations. More formally, for any sequence of random vectors X_n, we say $X_n = O_p(a_n)$ if for every $\varepsilon > 0$, there exists constants C_ε and n_ε such that $\mathbb{P}(X_{n2}> C_\varepsilon a_n) < \varepsilon$ for every $n \geq n_\varepsilon$; equivalently, we say $(1/a_n)X_n$ is "uniformly tight" or "bounded in probability". Similarly, $X_n = o_p(a_n)$ if for all $\varepsilon, \varepsilon' > 0$, there exists n_ε such that $\mathbb{P}(X_{n2}> \varepsilon' a_n) < \varepsilon$ for every $n \geq n_\varepsilon$. Therefore, $X_n = o_p(1) \iff X_n \xrightarrow{p} 0$. Additionally, we denote: $\text{plim } X_n = a \iff X_n \xrightarrow{p} a$. We say a sequence of events $\mathcal{E}_n$, indexed by n, holds "with high probability" (w.h.p.) if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$, i.e., for any $\varepsilon > 0$, there exists a n_ε such that for all $n > n_\varepsilon$, $\mathbb{P}(\mathcal{E}_n) > 1 - \varepsilon$. More generally, a multi-indexed sequence of events $\mathcal{E}_{n_1,\dots,n_d}$, with indices $n_1, \dots, n_d$ with $d \geq 1$, is said to hold w.h.p. if $\mathbb{P}(\mathcal{E}_{n_1,\dots,n_d}) \to 1$ as $\min\{n_1, \dots, n_d\} \to \infty$. We also use $\mathcal{N}(\mu, \sigma^2)$ to denote a normal or Gaussian distribution with mean μ and variance σ^2—we call it *standard* normal if $\mu = 0$ and $\sigma^2 = 1$.

4.2 Related Works

Given the vastness of the matrix completion literature, we do not strive to do an exhaustive review of it. Instead, we focus on a few representative works that propose and analyze algorithms designed for the three different models of missingness: MCAR, MAR, and MNAR. In Section 4.2.1, we give an overview of the type of algorithms for matrix completion studied thus far in existing works. In Section 4.2.2, we discuss the different models of missingness considered in the matrix completion literature, and representative algorithms for these various models. Finally, in Section 4.2.3, we discuss the growing literature exploring the intersection of matrix completion and causal inference; in particular, the panel data literature in econometrics.

4.2.1 Overview of Matrix Completion Algorithms

Algorithms for matrix completion broadly fall into two classes: empirical risk minimization (ERM) methods and matching (i.e., collaborative filtering) methods, with ERM methods being relatively more popular. We give an overview of both class of methods below.

Empirical Risk Minimization (ERM) Methods. Empirical risk minimization (ERM) is arguably the de facto approach to recover the underlying signal matrix $\boldsymbol{A}$ given $\widetilde{\boldsymbol{Y}}$. Specifically, ERM approaches aim to solve the following program:

$$\text{minimize} \quad \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} d(T_{ij}, Q_{ij}) + \lambda \text{ regularize}(\boldsymbol{Q}). \tag{4.2}$$

Here, $\Omega \subseteq [m] \times [n]$, $d(\cdot,\cdot)$ is an appropriate distance measure (e.g., squared loss), T_{ij} is a "simple" transformation of $\widetilde{Y}_{ij}$ (e.g. $\mathbb{1}(D_{ij} = 1) \cdot \widetilde{Y}_{ij}$), regularize$(\cdot)$ is a regularization term and $\lambda > 0$ is the regularization hyper-parameter. For certain algorithms, they replace the regularizer (i.e., set $\lambda = 0$) with a constraint, constraint$(\cdot)$.

In order to prove statistical guarantees about these various estimators, structure is placed on $\boldsymbol{A}$. The assumptions made guide the specific choices of the above parameters, which then define the algorithm. For instance, if the singular values of $\boldsymbol{A}$ are assumed to be moderately sparse (i.e., only few are non-zero), then a natural convex regularizer would penalize solutions with large nuclear norm, i.e., regularize$(\boldsymbol{Q}) = \boldsymbol{Q}_*$ Candès and Tao (2010); Recht (2011). Indeed, choosing $\Omega = \{(i,j) : D_{ij} = 1\}$ as the collection of observed entries, $T_{ij} = \widetilde{Y}_{ij}$, and $d(\cdot,\cdot)$ as the squared loss yields the popular `softImpute` algorithm of Mazumder et al. (2010); Hastie et al. (2015). As another example, if $\boldsymbol{A}$ is assumed to be exactly low-rank, then a natural constraint would be the rank of the output matrix. More specifically, contraint$(\boldsymbol{Q})$ can be defined as rank$(\boldsymbol{Q}) \leq \mu$ for some pre-specified integer $\mu > 0$. Then, choosing $\Omega = [m] \times [n]$, $T_{ij} = \mathbb{1}(D_{ij} = 1) \cdot \widetilde{Y}_{ij}$, and $d(\cdot,\cdot)$ as the squared loss yields a suite of spectral based methods Keshavan et al. (2010a,b); Gavish and Donoho (2014); Chatterjee (2015). Other notable algorithms within the broader ERM class include maximum-margin matrix factorization (`MMMF`) Srebro et al. (2004), probabilistic matrix factorization (`PMF`) Mnih and Salakhutdinov (2008), and `SVD++` Koren (2008) to name a few.

Broadly speaking, it is commonly assumed that $\boldsymbol{A}$ follows some form of a latent variable model; in particular, $A_{ij} = f(u_i, v_j)$, where f is a sufficiently "smooth" latent function (e.g., Hölder continuous), and u_i, v_j are low-dimensional latent variables associated with row i and column j, respectively. Such latent variable models imply that $\boldsymbol{A}$ is (approximately) low-rank, i.e., $A_{ij} \approx \langle u_i, v_j \rangle$, where $u_i, v_j \in \mathbb{R}^r$ and $r \ll \min\{m, n\}$, e.g., Xu (2017a); Udell and Townsend (2019); Agarwal et al. (2021e). For an excellent overview on standard assumptions made on $\boldsymbol{A}$ and the subsequent guarantees proven for the estimation error, please refer to Davenport and Romberg (2016).

When every entry is revealed with uniform probability (i.e., $p_{ij} = p$), (4.2) is an unbiased estimate of the full loss function with all entries revealed (i.e., $\boldsymbol{D}$ is an all ones matrix). When p_{ij} are nonuniform, however, recent works have provably and empirically shown that (4.2) is biased Schnabel et al. (2016); Ma and Chen (2019). As such, these works advocate to de-bias the standard ERM objective by re-weighting each observation inversely by its propensity score p_{ij}. This technique is often known in the causal inference literature as inverse propensity scoring (IPS) or weighting (IPW), see Imbens and Rubin (2015); Little and Rubin (2019). This yields the following adapted program:

$$\text{minimize} \sum_{(i,j)\in\Omega} (1/\widehat{p}_{ij})\ d(T_{ij}, Q_{ij}) + \lambda\ \text{regularize}(\boldsymbol{Q}), \tag{4.3}$$

where $\widehat{p}_{ij}$ is an estimate of p_{ij}. In words, (4.3) requires learning $\boldsymbol{P}$ prior to carrying out the standard ERM of (4.2). Faithful matrix recovery under more general missingness patterns thus requires structure on not only $\boldsymbol{A}$, but also $\boldsymbol{P}$ and $\boldsymbol{D}$. We overview standard assumptions on these quantities in Section 4.2.2.

Matching methods. For traditional applications of matrix completion, such as recommendation systems, K nearest neighbour (KNN) methods have been popular (e.g., Goldberg et al. (1992); Linden et al. (2003); Kleinberg and Sandler (2008); Koren and Bell (2015); Lee et al. (2016); Chen et al. (2018)). In KNN, to impute a missing entry (i,j), the first step is to select K rows for which the entry in the j-th column is not missing. Of all the rows for which the j-th column is not missing, the K rows are selected such that they are the “closest” to row i. In particular, a hyper-parameter of KNN is the metric that is chosen to define “closeness” between any two given rows; the most commonly used metric is the mean squared distance between the commonly revealed entries for a given two rows. Once these K “neighbour rows” are chosen, the estimate for the missing entry (i,j) is the average $\frac{1}{K}\sum_{k\in\text{neighbour rows}} \widetilde{Y}_{kj}$. An attractive quality of these KNN methods is that they do not require imputing missing values by 0. A related literature that shares similarities with KNN is that of synthetic controls Abadie et al. (2010); Abadie and Gardeazabal (2003). A key difference is that to impute (i,j), uniform weights (i.e., $1/K$) are not used for the neighbouring rows; classically in synthetic controls, these weights are constrained to lie within the simplex, i.e., the weights are non-negative and sum to 1 (if the weights are restricted to be $1/K$, this is known in the panel data literature as “difference-in-differences”). However, as discussed earlier, synthetic controls methods have been designed to handle restricted sparsity patterns naturally arising in the panel data setting. Given the growing literature on synthetic controls, we do a detailed literature review of it in Section 4.2.3.

■ 4.2.2 Three Models of Missingness

Below, we utilize the useful taxonomy set in Rubin (1976); Little and Rubin (2019) to discuss the three primary mechanisms that lead to missing data and how previous works fit within these regimes.

Missing completely at random (MCAR). MCAR is the most standard model of missingness assumed in the matrix completion literature and is characterized by the following properties: (i) $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{Y}$; (ii) $D_{ij} \perp\!\!\!\perp D_{ab}$ for all $(i,j) \neq (a,b)$; (iii) $p_{ij} = p > 0$ for all (i,j). In words, MCAR assumes each element of $\boldsymbol{D}$ is an independent and identically distributed (i.i.d.) Bernoulli random variable (r.v.) with parameter $p \in (0,1]$. This implies that the missingness pattern is independent of the values in $\boldsymbol{Y}$. We note that this condition $p_{ij} > 0$ is known in the causal inference literature as "positivity", see Imbens and Rubin (2015). It follows that the maximum likelihood estimator $\widehat{p}_{ij} = \widehat{p}$ for all (i,j), where $\widehat{p}$ is the fraction of observed entries in $\widetilde{\boldsymbol{Y}}$. As previously mentioned, given MCAR data, (4.2) is an unbiased estimator of the ideal loss function where all entries observed. Though MCAR is likely unrealistic outside experimental settings, the MCAR regime remains a popular abstraction in machine learning and statistics to study the inherent trade-offs between the observation probability p, properties of the noise $\boldsymbol{E}$, and the structure imposed on the signal $\boldsymbol{A}$, in terms of the estimation error between $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}$. Methods such as singular value thresholding explicitly impute missing values in $\boldsymbol{Y}$ (denoted as $\star$) by 0 and re-weight all non-missing values in $\boldsymbol{Y}$ by $1/\widehat{p}$, where $\widehat{p}$ is the fraction of observed entries. This can be interpreted as a form of uniform IPW. Other methods such as nuclear norm minimization, alternating least squares, and nearest neighbour methods do not require imputing missing values by 0. However, existing theoretical analysis of these algorithms do still require that $\mathbb{E}[D_{ij}] = p$, and that D_{ij} is independent of the all other randomness in the model.

Missing at random (MAR). [2] MAR is a more challenging setting than MCAR. The three key assumptions of MAR are as follows. (i) $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{Y} \mid \mathcal{O}$, where $\mathcal{O}$ represents observed covariates about the rows and columns of the matrix (e.g., covariates about users and movies in the context of recommender systems)—concretely, these observed variables, $\mathcal{O}$,

[2] Many works in the matrix completion literature do not differentiate between MAR and MNAR, and call both regimes MNAR. We differentiate between them to be more in line with models of missingness proposed by Rubin (1976); Little and Rubin (2019).

often include features or covariates $(X_i, \tilde{X}_j)$, which are associated with row i and column j, respectively, and observed outcomes $\widetilde{Y}_{ij}$. (ii) $D_{ij} \perp\!\!\!\perp D_{ab}$ for all $(i, j) \neq (a, b)$. (iii) $p_{ij} > 0$ for all (i, j). Here, the entries of $\boldsymbol{D}$ continue to obey positivity and remain independent Bernoulli r.v.'s.

Below, we overview two popular propensity estimation techniques of Schnabel et al. (2016). To aid the following discussion, let $\boldsymbol{X} = \{(X_i, \tilde{X}_j) : (i, j) \in [m] \times [n]\}$ denote the set of observed features, and $\boldsymbol{H}$ denote the set of hidden features. The first approach is via Naive Bayes, which assumes that $p_{ij} = \mathbb{E}[D_{ij} | \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{Y}] = \mathbb{E}[D_{ij} | \widetilde{Y}_{ij}]$. Under this assumption, the maximum likelihood estimator $\widehat{p}_{ij}$ can be solved using Bayes formula; however, such an approach requires a small sample of MCAR data, see Schnabel et al. (2016). The second estimation strategy is based on logistic regression. Here, it is assumed that there exists model parameters ϕ such that $p_{ij} = \mathbb{E}[D_{ij} | \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{Y}] = \mathbb{E}[D_{ij} | X_i, \tilde{X}_j, \phi]$; within the causal inference literature, this is often known as "selection on observables", see Imbens and Rubin (2015). Typically, it is posited that $\phi = (\omega_1, \omega_2, \alpha, \gamma)$ and $\mathbb{E}[D_{ij} | X_i, \tilde{X}_j, \phi] = \sigma(\langle \omega_1, X_i \rangle + \langle \omega_2, \tilde{X}_j \rangle + \alpha_i + \gamma_j)$, where $\sigma(\cdot)$ takes a simple parametric form such as the sigmoid function. Some notable works in the MAR literature include Liang et al. (2016); Wang et al. (2018a,b, 2019).

Missing not at random (MNAR). MNAR is the most challenging missingness model in matrix completion with a comparatively sparser literature. In its fullest generality, in MNAR the following conditions are allowed: (i) $\boldsymbol{D}$ can depend on $\boldsymbol{Y}$ and other unobserved variables; (ii) D_{ij} can be correlated with D_{ab} for all $(i, j) \neq (a, b)$; (iii) $\min p_{ij} = 0$. The first condition implies that $\boldsymbol{D}$ and $\boldsymbol{Y}$ remain dependent even conditional on observed covariates. The second condition allows the revelation of one outcome to alter the probability of another outcome being revealed. Finally, the third condition can restrict certain outcomes from ever being revealed. Hence, the literature has thus far only considered a *limited* version of MNAR with conditions cf. Ma and Chen (2019); Bhattacharya and Chatterjee (2021); Yang et al. (2021). In particular, they continue to make the following assumptions: p_{ij} is a (nice) function solely of latent factors associated with entry (i, j); each entry of $\boldsymbol{D}$ is an independent (not necessarily identically distributed) Bernoulli r.v. with a strictly positive probability of being revealed, which are the assumptions as in MAR. These assumption are what allow the weighted ERM framework of (4.3) to continue being valid. The methods proposed in Ma and Chen (2019); Bhattacharya and Chatterjee (2021); Yang et al. (2021) work for this *limited* MNAR setting by positing that $\boldsymbol{P}$ is (approximately) low-rank, and recovers $\boldsymbol{P}$ from $\boldsymbol{D}$ via matrix completion algorithms. This is a generalization of the MAR setting as such an approach circumvents the requirement of meaningful auxiliary features $\boldsymbol{X}$ to conduct propensity score estimation. Additional works within the MNAR literature include Zhu et al. (2019); Sportisse et al. (2020a,b); Wang et al. (2020).

As previously mentioned, our work operates under greater generality than the *limited* MNAR regime thus far considered in the literature. More specifically, our framework allows $\boldsymbol{D}$ and $\boldsymbol{Y}$ to be dependent, provided $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{A}$, and for $\boldsymbol{D}$ to be any arbitrary function of $\boldsymbol{A}$, subject to suitable observation patterns. We also allow for conditions (ii) and (iii) described above to hold, i.e., the entries in $\boldsymbol{D}$ can be highly correlated and the minimum probability of observation can be deterministically set to 0. In Section 4.3, we will formally introduce our causal framework to rigorously discuss these properties.

Summary of matrix completion results. Across the various models of missingness, the key theoretical results for low-rank matrix completion typically have error bounds that scale in the following form (see Davenport and Romberg (2016)):

$$\frac{1}{mn}\widehat{\boldsymbol{A}} - \boldsymbol{A}_F^2 = O\left(\frac{1}{\text{poly}(p_{\min})} \cdot \frac{\text{poly}(r)}{\min(m,n)^{1-\delta}}\right)$$

for $\delta \geq 0$ and where poly$(\cdot)$ denotes polynomial dependence. Here, $p_{\min} = \min p_{ij}$ and r refers to the (approximate) rank of $\boldsymbol{A}$. The most studied metric in the literature is the average error across all entries, $(1/mn)\widehat{\boldsymbol{A}} - \boldsymbol{A}_F^2$, though recent works have begun to analyze stronger metrics such as the maximum average error across all columns, $(1/m)\widehat{\boldsymbol{A}} - \boldsymbol{A}_{2,\infty}^2$ (e.g., Agarwal et al. (2019b, 2021e,d); Agarwal and Singh (2021)), and the maximum entry-wise error, $\widehat{\boldsymbol{A}} - \boldsymbol{A}_{\max}$ (e.g., Lee et al. (2016)). Crucially, all of these error bounds scale with the inverse of poly$(p_{\min})$. As discussed above, this immediately rules out settings where $p_{\min} = 0$, i.e., condition (iii) of MNAR above. Finally, we remark that the literature studying the asymptotic properties of $\widehat{\boldsymbol{A}} - \boldsymbol{A}$ (e.g., proving asymptotic normality) is relatively small. Some notable works on the asymptotic analyses of matrix completion estimators under MCAR include Chen et al. (2019); Cai et al. (2020); Bhattacharya and Chatterjee (2021).

■ 4.2.3 Panel Data and Matrix Completion

In Section 4.3, we propose a causal framework for matrix completion that draws inspiration from the rich and growing literature in econometrics on panel data and matrix completion;

some relevant works include Amjad et al. (2018, 2019); Arkhangelsky et al. (2019); Bai and Ng (2020); Fernández-Val et al. (2020); Athey et al. (2021); Agarwal et al. (2019b, 2021e,d,c); Agarwal and Singh (2021). As is common in matrix completion, these works impose a (approximate) low-rank factor model on the signal matrix (i.e., $\boldsymbol{A}$), also known as an interactive fixed effects model, to capture structure across units and time (i.e., the rows and columns of the matrix, respectively).

Panel data & matrix completion: an overview. As described in Section 4.1, the sparsity structure considered in these works is one where for each row i, there is a column $j_i \in [n]$ such that $D_{ij} = 1$ for all $j < j_i$ and $D_{ij} = 0$ for $j \geq j_i$. That is, all entries for a given row i are observed till some column j_i, after which they are all missing. The motivation for such a sparsity pattern comes from socio-economic policy making where Y_{ij} represents unit i's potential outcome at time step j under "control", i.e., if no socio-economic intervention has yet been applied on unit i. The time steps $[1, j_i - 1]$ represent the period when unit i is under control, and time steps $[j_i, n]$ represent the period when unit i has undergone an intervention. Hence, Y_{ij} for $j > j_i$ is missing and the goal is to estimate the counterfactual of what would have happened to unit i had it remained under control during $[j_i, n]$. This particular setting is also known in the econometrics literature as "synthetic contorls" Abadie et al. (2010); Abadie and Gardeazabal (2003). The statistical/causal parameter that is most commonly studied is for a "treated" unit i, to estimate $\frac{1}{n-j_i} \sum_{j=j_i}^{n} Y_{ij}$. That is, the average potential outcome of unit i under control during the "post-intervention" period. Most of these works make the additional assumption that each unit either remains under control for the entire time period under consideration, or undergoes an intervention at a time step that is common across all units. Athey et al. (2021) is one notable work that allows for different post-intervention periods for each unit.

Connections to matrix completion with MNAR data. An attractive quality of this literature is that in some ways it allows for more relaxed conditions on $\boldsymbol{D}$ and $\boldsymbol{P}$ than those considered in the matrix completion with MNAR data literature discussed earlier, see Ma and Chen (2019); Yang et al. (2021); Sportisse et al. (2020b,a); Wang et al. (2019). In particular, the panel data literature allows the entries of $\boldsymbol{D}$ to be correlated, e.g., if $D_{ij} = 0$, then $D_{ij'} = 0$ for $j' > j$. Further, $\min p_{ij}$ is allowed to be 0 and j_i is allowed to depend on $\boldsymbol{A}$. On the other hand, the sparsity pattern considered in the panel data literature is far more restrictive compared to the works on matrix completion with MNAR data—as discussed above, in panel data settings, all columns for a given row are observed till a specific point, after which they are all missing (i.e., $D_{ij} = 1$ for all $j < j_i$ and $D_{ij} = 0$ for $j \geq j_i$).

Note that this also implies that $\boldsymbol{P}$ is low-rank. Such a sparsity pattern is unrealistic for many important applications for matrix completion, including recommendation systems and sequential decision-making. Further, it is not straightforward to see how the target statistical/causal parameter $\frac{1}{n-j_i}\sum_{j=j_i}^{n} Y_{ij}$ is particularly meaningful outside the synthetic controls literature. Hence, our aim with this work is to combine the best of both worlds, where we: (i) allow entries of $\boldsymbol{D}$ to be correlated; (ii) allow $\min p_{ij} = 0$; (iii) make no parametric assumptions about $\boldsymbol{P}$; (iv) allow $\boldsymbol{P}$ to not be low-rank; (v) allow for general missingness patterns in the matrix that includes MCAR data as a special case. Further the target parameter we aim to estimate (in expectation) is *each entry* Y_{ij} for every (i, j) pair. Also, by formally bridging the panel data literature to more classical applications of matrix completion such as recommendation systems, we hope this spurs further investigation into the unexplored connections between these two fields.

Comparison with synthetic interventions. Our proposed framework framework builds upon the recent work of Agarwal et al. (2021c), called synthetic interventions (SI). SI is a causal inference method to do tensor completion with MNAR data, where the dimensions of the order-3 tensor of interest are units, measurements, and interventions. That is, an entry Y_{ijd} of the tensor considered in SI refers to the potential outcome of the i-th unit, its j-th measurement, under the d-th intervention. Their setup can be made a special case of ours by effectively *flattening* the tensor into a matrix, where the rows of the induced matrix still correspond to units, but a column is a double index for a measurement and an intervention, i.e., the (i, j, d)-th entry of the tensor corresponds to the $(i, (j, d))$-th entry of the induced matrix. Given this simple reduction, we generalize the framework, algorithm, and theoretical results in Agarwal et al. (2021c) in the following ways. First, we formally extend the SI framework, to recover matrices under more general missingness patterns than that considered in Agarwal et al. (2021c). Doing so allows us to apply our framework to a wider variety of applications such as recommender systems, while the SI framework was introduced in the context of personalized policy evaluation and synthetic A/B testing. Third, this work establishes point-wise finite-sample consistency and asymptotic normality of our proposed SNN algorithm, which was absent in Agarwal et al. (2021c) with respect to the SI algorithm. Indeed, in the context of the panel data literature, establishing point-wise asymptotic normality for each unit, (intervention, time)-tuple is of independent interest.

■ 4.3 A Causal Framework for Matrix Completion

In this section, we develop a formal causal framework for matrix completion with MNAR data. In Section 4.3.1, we show how to causally interpret matrix completion with MNAR data using the language of potential outcomes in Section 4.3.1. We then state and justify our assumptions in Section 4.3.2, define our causal estimand in Section 4.3.3, and present our identification result in Section 4.3.4.

■ 4.3.1 Potential Outcomes

We follow the potential outcomes framework of Neyman (1923); Rubin (1974). In particular, we let the r.v. $Y_{ij} \in \mathbb{R}$, as defined in Section 4.1.2, denote the potential outcome associated with each pair (i, j) if it is revealed. For instance, in the case of recommender systems, Y_{ij} can be interpreted as the rating user i *would have* given to item j had they rated it. In the context of healthcare for example, Y_{ij} could represent patient i's health metric of interest (e.g. heart rate) *had they* been given treatment j. Finally, in the case of panel data setting, as discussed in Section 4.2.3, $Y_{i,(a,t)}$ can denote the metric of interest for unit i (e.g. revenue generated, socio-economic indicator), if they *would have* received the a-th socio-economic policy at time step t; here, (a, t) represents the j-th column.

If $D_{ij} = 1$, then by (4.1) we see that we actually do observe the (i, j)-th potential outcome, i.e. $\widetilde{Y}_{ij} = Y_{ij}$. That is, in the language of potential outcomes, we can interpret $\boldsymbol{D}$ as the matrix of intervention assignments. Through this perspective, we remark that (4.1) is an implicit assumption that is known in the causal inference literature as "consistency" or "stable-unit-treatment-value assumption" (SUTVA). As discussed earlier, the fact that $\boldsymbol{Y} \not\perp\!\!\!\perp \boldsymbol{D}$ (e.g. a user's preference for a movie can determine whether they rate it) means that the potential outcomes are not independent of the intervention assignments. This dependence is known in the causal inference literature as "confounding". Lastly, as alluded to earlier, we generalize the standard potential outcomes framework in that a given unit can receive multiple interventions. Traditionally, it is assumed that a unit receives exactly one intervention. However, in applications like movie recommendation systems, a user can "intervene" and rate multiple movies. Lastly, this framework also generalizes panel data settings, as we allow each unit to receive different interventions at different time steps; as discussed earlier, it is typically assumed that units are in control for a period of time, and then some subset of units receive one intervention for the remaining time steps.

■ 4.3.2 Assumptions

Below, we state our causal assumptions and then provide their corresponding interpretations.

Assumption 17 (Low-rank factor model)**.** *For every pair* (i, j)*, let*

$$Y_{ij} = \langle u_i, v_j \rangle + \varepsilon_{ij},$$

where $u_i, v_j \in \mathbb{R}^r$ *are latent vectors. Equivalently, we say* $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{V}^T + \boldsymbol{E}$*, where* u_i *refers to the* i*-th row of* $\boldsymbol{U} \in \mathbb{R}^{m \times r}$*, and* v_j *refers to the* j*-th row of* $\boldsymbol{V} \in \mathbb{R}^{n \times r}$*.*

Assumption 18 (Selection on latent factors)**.** *We have that for any intervention assignment* $\boldsymbol{D}$*,*

$$\mathbb{E}[\boldsymbol{E}|\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}] = 0$$

Neighbourhood rows and columns. For the remainder of this work, for a given column j, we refer to $\text{NR}(j) = \{a \in [m] : D_{aj} = 1\}$ as “neighborhood rows”, i.e., rows where entries in column j are not missing. Similarly, for a given row i, we refer to $\text{NC}(i) = \{b \in [n] : D_{ib} = 1\}$ as “neighborhood columns”, i.e., columns where entries in row i are not missing. See Figure 4.5b for a visual depiction of $\text{NR}(j)$ and $\text{NC}(i)$.

Assumption 19 (Linear span inclusion)**.** *Conditioned on* $\boldsymbol{D}$*, for a given pair* (i, j) *and any* $\mathcal{I} \subseteq NR(j)$*, if* $|\mathcal{I}| \geq \mu$*, then* u_i *lies in the linear row span of* $\boldsymbol{U}_{\mathcal{I}}$*, i.e., there exists a* $\beta \in \mathbb{R}^{|\mathcal{I}|}$ *such that*

$$u_i = \sum_{\ell \in \mathcal{I}} \beta_\ell u_\ell$$

Interpretation of Assumptions 17 to 19 By the tower law, Assumption 18 implies that $\mathbb{E}[\boldsymbol{E}|\boldsymbol{U}, \boldsymbol{V}] = 0$. This together with Assumption 17 posits that $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{V}]$ is a low-rank matrix with rank r. As discussed in Section 4.2, this is a standard assumption within the matrix completion literature. Next, we remark that Assumption 18, coupled with Assumption 17, implies that

$$\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{U}, \boldsymbol{V}] = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}].$$

That is, the potential outcomes are mean independent of the intervention assignments, conditioned on the latent row and column factors. This has been termed as "selection on latent factors", see Agarwal et al. (2021c). Similar conditional independence conditions have been explored in Athey et al. (2021); Kallus et al. (2018). Lastly, given Assumption 17, it follows that Assumption 19 is rather mild. To see this, suppose $\text{span}(\{u_\ell : \ell \in \mathcal{I}\}) = \mathbb{R}^r$, i.e., $\text{rank}(\boldsymbol{U}_{\mathcal{I},\cdot}) = r$. Then, Assumption 19 immediately holds as $u_i \in \mathbb{R}^r$. More generally, if the rows of $\boldsymbol{U}$ are randomly sampled sub-gaussian vectors, then $\text{span}(\{u_\ell : \ell \in \mathcal{I}\}) = \mathbb{R}^r$ for any set $\mathcal{I}$ holds w.h.p., provided $\mu \geq r$ is chosen to be sufficiently large; see Vershynin (2018) for details.

4.3.3 Target Causal Estimand

Define

$$\boldsymbol{A} := \mathbb{E}[\boldsymbol{Y} | \boldsymbol{U}, \boldsymbol{V}]. \tag{4.4}$$

Note that given Assumptions 17 and 18, the definition of $\boldsymbol{A}$ in (4.4) is consistent with the definition of $\boldsymbol{A}$ used in Section 4.1.2. We are now equipped to define our target causal estimand, which is A_{ij}; for the remainder, of the paper we focus on a particular pair (i, j), without loss of generality. Note given Assumptions 17 and 18, we can write

$$A_{ij} := \mathbb{E}[Y_{ij} | u_i, v_j].$$

In words, A_{ij} translates as the expected potential outcome for the (i, j)-th pair, conditioned on its row and column latent vectors (u_i, v_j). For instance, returning to recommender systems, A_{ij} represents the expected rating user i would supply for item j, conditioned on the latent features that characterize user i and item j. In panel data settings, letting $j = (a, t)$, A_{ij} represents the potential outcome of unit i had it received the a-th intervention at time step t.

4.3.4 Identification

The following identification results establishes that each entry of $\boldsymbol{A}$ can be learned from observable quantities, i.e., from $\widetilde{\boldsymbol{Y}}$. Practically speaking, this means that matrix completion with MNAR data for any pair (i, j) is possible.

Theorem 4.3.1. *Let Assumptions 17 to 19 hold. For a given pair (i, j) and $\mathcal{I} \subseteq NR(j)$ with*

$|\mathcal{I}| \geq \mu$, suppose β defined with respect to $\mathcal{I}$ as in Assumption 19, is known. Then,

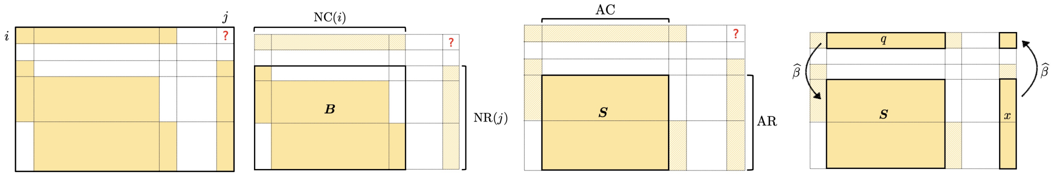
$$A_{ij} = \sum_{\ell \in \mathcal{I}} \beta_\ell \mathbb{E}[\widetilde{Y}_{\ell j} \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}].$$

Interpretation. Theorem 4.3.1 states that despite the missingness pattern being MNAR, if Assumptions 17 to 19 hold, and given knowledge of the linear model parameter β, the causal estimand A_{ij} can be expressed in terms of quantities that can be estimated from observed data, namely $\mathbb{E}[\widetilde{\boldsymbol{Y}}_{\mathcal{I},j}]$; this is known in the causal inference literature as "identification". Note, $\mathcal{I}$ is deterministic given $\boldsymbol{D}$. The key requirement of the missingness pattern $\boldsymbol{D}$ is that $\mathcal{I} \subseteq \text{NR}(j)$ is sufficiently large, which is parameterized by μ, i.e., we require $\mu \gg r$ where r is the rank of $\boldsymbol{A}$. That is, the number of rows for which column j is observed is sufficiently large. Thus, Theorem 4.3.1 suggests that the key quantity that enables the recovery of A_{ij} is β. In Section 4.4, we provide an algorithm to estimate β, which in turn, allows us to estimate A_{ij}.

■ 4.4 SNN: Matrix Completion with MNAR Data

In this section, we introduce an algorithm, synthetic nearest neighbors (SNN), for matrix completion with MNAR data. Towards this, we introduce helpful notation that will be used for the remainder of this work. Again, without loss of generality, we consider imputing the (i, j)-th entry of the matrix .

Notation. Let AR $\subseteq$ NR(j) and $\sim \subseteq$ NC(i) denote a subset of rows and columns, respectively, of $\widetilde{\boldsymbol{Y}}$ that satisfy $D_{ab} = 1$ for all $(a, b) \in \text{AR} \times \sim$. We refer to AR and $\sim$ as the "anchor rows" and "anchor columns" of pair (i, j), respectively. Collectively, AR and $\sim$ form a fully observed sub-matrix of $\widetilde{\boldsymbol{Y}}$; for ease of notation. We refer to this $|\text{AR}| \times |\sim|$ sub-matrix as $\boldsymbol{S} := [\widetilde{Y}_{ab} : (a, b) \in \text{AR} \times \sim]$. See Figure 4.5c for a visual depiction of AR, $\sim$, and $\boldsymbol{S}$. Note, by construction $\boldsymbol{S}$ is such that if entries from row a are present in $\boldsymbol{S}$, then $D_{aj} = 1$; similarly, if entries from column b are present in $\boldsymbol{S}$, then $D_{ib} = 1$ Additionally, let $q := [\widetilde{Y}_{ib} : b \in \sim]$ and $x := [\widetilde{Y}_{aj} : a \in \text{AR}]$. $q \in \mathbb{R}^{|\sim|}$ refers to the columns in row i which correspond to $\sim$; similarly, $x \in \mathbb{R}^{|\text{AR}|}$ refers to the rows in column j which correspond to AR. By construction, all the elements in q and x are not missing. See Figure 4.5d for a visual depiction of q and x.



(a) (b) (c) (d)

Figure 4.5: We visually depict the various quantities needed to define the SNN algorithm. Figure 4.5a depicts a particular sparsity pattern in our matrix $\widetilde{Y}$ with entry (i, j) missing. Figure 4.5b depicts $\mathrm{NR}(j)$ and $\mathrm{NC}(i)$. Figure 4.5c depicts AR, $\sim$, and $\boldsymbol{S}$. Figure 4.5d depicts the SNN algorithm with $K = 1$; for $K > 1$, we partition the rows in S into K mutually disjoint sets.

■ 4.4.1 Algorithm

We now present SNN in Algorithm 1 to impute the (i, j)-th entry. It has $K \in \mathbb{N}$ and $\lambda^{(k)} \in \mathbb{R}$ for $k \in [K]$ as hyper-parameters.

Algorithm 1 SNN(i, j)

```
  Input: $\{\lambda^{(k)} : k \in [K]\}$, $\{(\sim^{(k)}, \mathrm{AR}^{(k)}) : k \in [K]\}$ with mutually disjoint sets $\{\mathrm{AR}^{(k)} : k \in [K]\}$.
  for $k \in [K]$ do
      1. Define $\boldsymbol{S}^{(k)} = [\widetilde{Y}_{ab} : (a, b) \in \mathrm{AR}^{(k)} \times \sim^{(k)}]$
      2. Compute $\boldsymbol{S}^{(k)} \leftarrow \sum_{\ell \geq 1} \widehat{\tau}_\ell^{(k)} \widehat{u}_\ell^{(k)} \otimes \widehat{v}_\ell^{(k)}$
      3. Compute $\widehat{\beta}^{(k)} \leftarrow \left( \sum_{\ell \leq \lambda^{(k)}} (1/\widehat{\tau}_\ell^{(k)}) \widehat{u}_\ell^{(k)} \otimes \widehat{v}_\ell^{(k)} \right) q^{(k)}$
      4. Compute $\widehat{A}_{ij}^{(k)} \leftarrow \langle x^{(k)}, \widehat{\beta}^{(k)} \rangle$
  end for
  4. Output $\widehat{A}_{ij} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \widehat{A}_{ij}^{(k)}$
```

Note, for ease of notation, in Algorithm 1 we suppress the dependence on i and j in the definitions of $\{(\sim^{(k)}, \mathrm{AR}^{(k)}) : k \in [K]\}$, $\boldsymbol{S}^{(k)}$, $\widehat{\beta}^{(k)}$, $q^{(k)}$, and $x^{(k)}$. That is, these quantities will change depending on which (i, j)-th entry of the matrix we aim to impute. We continue to suppress this dependence for the remainder of the paper. For a visual depiction of the SNN algorithm for $K = 1$, refer to Figure 4.5d. For $K > 1$, we simply re-run the SNN algorithm seperately for the K disjoint subsets $\{\mathrm{AR}^{(k)} : k \in [K]\}$, and take the average of the estimates $\widehat{A}_{ij}^{(k)}$ for $k \in [K]$, produced by each iteration.

Interpretation. SNN draws inspiration from the popular K Nearest Neighbour (KNN) algorithm, described in Section 4.2. However, the key assumption underlying KNN is that there *do exist* K rows that are close to identical to the i-th row, with respect to some pre-defined metric. However, it is not necessary that these K rows exist *even* for a rank 1 matrix. As a simple example, consider a matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ where $M_{i\cdot} = [i, 2i, \dots, ni]$. By construction $\boldsymbol{M}$ is rank 1, but for any row, there does not exist any other row that is close to it in a mean squared sense; hence, it has no nearest neighbours.

The SNN algorithm overcomes this hurdle by first constructing K "synthetic" neighbors of row i from $\text{NR}(j)$, where the k-th synthetic neighboring row is formed by a linear combination, defined by $\widehat{\beta}^{(k)}$, of the rows in $\text{AR}^{(k)}$. Then, similar to KNN, SNN estimates A_{ij} by taking an average of the observed outcomes for column j that are associated with the K synthetic neighbors of row i. In words, $\widehat{\beta}^{(k)}$ is precisely the set of estimated linear weights that best recreates the observed outcomes of row i from the rows in $\text{AR}^{(k)}$, using observations from the columns in $\sim^{(k)}$. This idea of matching rows via a linear re-weighting takes inspiration from the synthetic controls literature—see Section 4.2.3 for details. To ensure the linear fit is appropriately regularized, a spectral sparsity constraint is imposed on $\boldsymbol{S}^{(k)}$, which is parameterized by $\lambda^{(k)}$. This constrained regression is known in the literature as principal component regression (PCR) (see Agarwal et al. (2019b, 2021e,d,c); Agarwal and Singh (2021)). We note that in lieu of requiring that there exist K close neighbouring rows as in KNN, SNN requires that the i-th row lies in the linear span of the rows in $\text{AR}^{(k)}$; that is, given Assumption 19 holds, we require $|\text{AR}^{(k)}| \geq \mu$. Note that for the matrix $\boldsymbol{M}$ described above, for any particular row, all other rows satisfy this linear span inclusion condition, i.e., $\mu = 1$ since $\boldsymbol{M}$ is rank 1.

Choosing $\lambda^{(k)}$ There exist a number of principled heuristics to select the hyper-parameter $\lambda^{(k)}$, and we name a few here. As is standard within the statistics and ML literatures, the most popular data-driven approach is to use cross-validation. Another common approach is to use a universal thresholding scheme that preserves the singular values above a precomputed threshold (see Gavish and Donoho (2014); Chatterjee (2015)). Finally, a human-in-the-loop approach is to inspect the spectral characteristics of $\boldsymbol{S}^{(k)}$ and choose $\lambda^{(k)}$ to be the natural "elbow" point that partitions the singular values into those of large and small magnitudes; in such a setting, the large magnitude singular values, which typically correspond to signal, are retained while the small magnitude singular values, which are often induced by noise, are filtered out. See the exposition on choosing the hyper-parameter for PCR in Agarwal et al. (2019b, 2021e,d,c).

Another Perspective on SNN. SNN imputes A_{ij} by building synthetic neighbors of row i from $\text{NR}(j)$. In Proposition 1, we demonstrate that A_{ij} can be equivalently estimated by building synthetic neighbors of column j from $\text{NC}(i)$ through a simple "transposition" of

Algorithm 1.

Proposition 1. *Consider any $k \in [K]$ and let $\widehat{\beta}^{(k)}$ be defined as in Algorithm 1. Further, let*

$$\widehat{\alpha}^{(k)} = \Big(\sum_{\ell \leq \lambda^{(k)}} (1/\widehat{\tau}_{\ell}^{(k)}) \widehat{v}_{\ell}^{(k)} \otimes \widehat{u}_{\ell}^{(k)}\Big) x^{(k)}.$$

Then,

$$\langle x^{(k)}, \widehat{\beta}^{(k)} \rangle = \langle q^{(k)}, \widehat{\alpha}^{(k)} \rangle.$$

■ 4.4.2 Finding Anchor Rows and Columns

Note the `SNN` algorithm takes as input $\{(\sim^{(k)}, \text{AR}^{(k)}) : k \in [K]\}$. However, the question remains that given the matrix $\boldsymbol{D}$, how to find these anchor rows and columns, with the additional constraint that the K set of anchor rows $\{\text{AR}^{(k)} : k \in [K]\}$ are mutually joint. In Section 4.4.2, we provide a practical algorithm `AnchorSubMatrix` in Algorithm 2 to find $\{(\sim^{(k)}, \text{AR}^{(k)}) : k \in [K]\}$ for a given pair (i, j). In Section 4.4.2, we discuss some motivating applications where anchor rows and columns are naturally induced.

Algorithmically Finding Anchor Rows and Columns via Maximum Biclique Search

In particular, we reduce our task of finding anchor rows and columns to a well-known problem in the graph theory literature known as finding "maximum bicliques". We briefly explain how to do this simple reduction. We first introduce some standard notation from graph theory. Let $\mathcal{G} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{E})$ denote a bipartite graph, where $(\mathcal{V}_1, \mathcal{V}_2)$ are the disjoint vertex sets and $\mathcal{E} \in \mathcal{V}_1 \times \mathcal{V}_2$ is the edge set, i.e., $(v_1, v_2) \in \mathcal{E}$ if there an edge between v_1 and v_2. Another way of representing $\mathcal{G}$ is via a bipartite incidence matrix $\boldsymbol{B} \in \{0, 1\}^{|\mathcal{V}_1| \times |\mathcal{V}_2|}$ (or adjacency matrix). In particular, $B_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$. If a sub-graph of $\mathcal{G}$ is complete, also called a biclique, then we denote it as $\mathcal{BC} \subset \mathcal{G}$, i.e., there is an edge between any pair of nodes $(v_1, v_2) \in \mathcal{BC}$. Now to see how to do the reduction between finding anchor rows and columns to the maximum biclique problem, recall $\boldsymbol{D} \in \{0, 1\}^{m \times n}$ is our matrix of intervention assignments. Note, $\boldsymbol{D}$ immediately induces a bipartite graph with $|\mathcal{V}_1| = m$ and $|\mathcal{V}_2| = n$, i.e., the vertex sets $\mathcal{V}_1$ and $\mathcal{V}_2$ correspond to the rows and columns of $\boldsymbol{D}$, respectively. We define $\mathcal{E}$ as follows, $(v_i, v_j) \in \mathcal{E}$ if $D_{ij} = 1$. In other words, the incidence matrix $\boldsymbol{B} \in \{0,1\}^{m \times n}$ induced by this graph is exactly equal to $\boldsymbol{D}$, i.e., $B_{ij} = 1$ if and only if $D_{ij} = 1$.

Given this reduction, we now describe how to practically implement the `AnchorSubMatrix` algorithm. We assume access to two algorithms: `createGraph` and `maxBiclique`. The former, `createGraph` : $\boldsymbol{B} \to \mathcal{G}$, takes as input a bipartite incidence matrix $\boldsymbol{B}$ (or adjacency matrix) and returns a bipartite graph $\mathcal{G}$; we note that the Python package `NetworkX` is an excellent resource to generate such graphs. The latter, `maxBiclique` : $\mathcal{G} \to \{\mathcal{BC}^{(\ell)}\}_{\ell \in [L]}$, takes as input a bipartite graph $\mathcal{G}$ and returns a set of L maximal bicliques $\{\mathcal{BC}^{(\ell)}\}_{\ell \in [L]}$; we refer the interested reader to Alexe et al. (2003); Zhang et al. (2014); Lyu et al. (2020); Lu et al. (2020) and references therein for example algorithms.

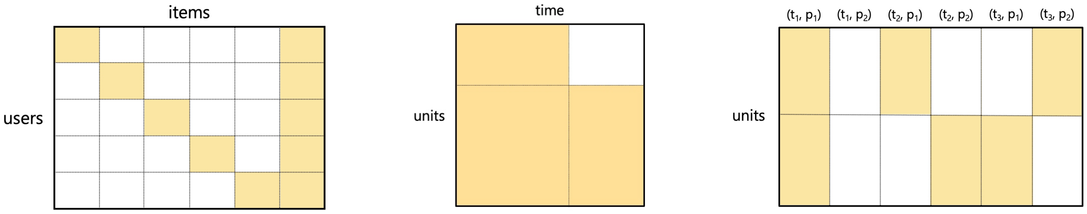
Algorithm 2 `AnchorSubMatrix`(i, j)

```
Input: createGraph, maxBiclique
1. Find $\text{NR}(j)$ and $\text{NR}(i)$
2. Assign $\boldsymbol{B} \leftarrow [D_{ab} : (a,b) \in \text{NR}(j) \times \text{NR}(i)]$
3. Generate $\mathcal{G} \leftarrow$ createGraph$(\boldsymbol{B})$
4. Compute $\{\mathcal{BC}^{(\ell)} = (\mathcal{V}_1^{(\ell)}, \mathcal{V}_2^{(\ell)}, \mathcal{E}^{(\ell)})\}_{\ell \in [L]} \leftarrow$ maxBiclique$(\mathcal{G})$
5. Assign $\mathcal{BC}^* = (\mathcal{V}_1^*, \mathcal{V}_2^*, \mathcal{E}^*) \leftarrow \arg\max \min\{|\mathcal{V}_1^{(\ell)}|, |\mathcal{V}_2^{(\ell)}|\}$ over $\ell \in [L]$
6. Output $\text{AR} \leftarrow \mathcal{V}_1^*$ and $\sim \leftarrow \mathcal{V}_2^*$
```

Given $(\sim, \text{AR})$ from Algorithm 2, we can construct $\{(\sim^{(k)}, \text{AR}^{(k)}) : k \in [K]\}$ as follows: First, we assign $\sim^{(k)} \leftarrow \sim$ for every k, i.e., the anchor columns for each subgroup k are all identically equal to $\sim$. Second, we (randomly) partition AR into K subgroups of equal size and then assign $\text{AR}^{(k)}$ as the k-th subgroup of AR such that $|\text{AR}^{(k)}| \sim |\text{AR}|/K$; in doing so, we ensure that $\{\text{AR}^{(k)} : k \in [K]\}$ are mutually disjoint sets. Note, for the purposes of theoretical analysis, we do not necessarily need to have $\sim^{(k)}$ be identical across all K. In Section 4.5, we show how the estimation error of `SNN` scales with $|\text{AR}^{(k)}|$ and $|\sim^{(k)}|$. In short, our theoretical results suggest that we want $\{(\sim^{(k)}, \text{AR}^{(k)}) : k \in [K]\}$ to be large on average; It is sufficient that we choose $|\sim^{(k)}|, |\text{AR}^{(k)}|$ such that $\min_{k \in [K]}\{|\sim^{(k)}|, |\text{AR}^{(k)}|\}$ is as large as possible; this is essence what Step 5 of Algorithm 2 is doing.

Applications where Anchor Rows and Columns are Naturally Induced

In this section, we discuss the typical sparsity pattern in recommender systems and sequential decision-making paradigms, which include panel data settings, reinforcement learning, and sequential A/B testing. We argue why these applications have a sparsity



(a) Recommender systems.

(b) Panel data.

(c) Sequential decision-making.

Figure 4.6: In both 4.6a, 4.6b, and 4.6c, observed entries are shown in yellow while unobserved entries are shown in white. Further, in 4.6c, the columns are indexed by (time, policy) tuples; here, t_ℓ and p_ℓ denote the ℓ-th time period and policy, respectively.

pattern where anchor rows and columns are naturally induced.

Recommender systems. As stated earlier, one of the key motivating applications for matrix completion is recommender systems. It has been noted in Ma and Chen (2019) that real-world recommender systems exhibit block-sparse structure; further the sparsity pattern is such that there is dependent missingness (i.e., $D_{ij} \not\perp\!\!\!\perp D_{ab}$ and zero probability of observing certain entries (i.e., $p_{\min} = 0$). An extreme version of this selection-bias would induce a sparsity structure as shown in Figure 4.6a. Within the context of movie recommender systems, a narrative for this missingness pattern is one where users only watch films that belong to genre(s) that they like and nothing else. However, in many recommender system applications, there exists a dense sub-matrix which corresponds to items that all users commonly rate—this corresponds to the rightmost columns of Figure 4.6a. This could occur if say a platform ask new users to indicate a subset of films that they enjoy. Indeed, this is a common practice for online platforms such as *Hulu, Netflix, StitchFix* to quickly learn a new user's preferences in order to provide a "warm-start" to their recommendation engine. Alternatively, many a time there are a small subset of iconic films (e.g., *Titanic* or *Star Wars*) that a large majority of users have watched. In this example in Figure 4.6a, all users can be used as anchor rows, and the set of items that are commonly rated across all users can be used as anchor columns. Further, we remark that in this example $\boldsymbol{P}$ is not low-rank, thus violating the key assumption required to learn p_{ij} in Ma and Chen (2019); Cai et al. (2020); Bhattacharya and Chatterjee (2021).

Sequential decision-making. As described earlier, in sequential decision-making, data is collected across units (e.g., individuals, customer types, geographic locations) over time in a sequential manner, where each unit is likely to be observed under a single or small set of interventions out of many at any time period. Many sequential decision-making problems can be phrased this way, including (i) panel data settings in econometrics; (ii) reinforcement learning and its variants (e.g., online learning, contextual bandits); (iii) sequential A/B testing. In (ii), an intervention denotes both the action picked and the observed state for that given time period; meanwhile in (iii), platforms run experiments on different customer types in a sequential and/or adaptive manner over time. The induced matrix in these settings has rows index units and columns index time-intervention pairs. It is common in many of these sequential decision-making settings that there is a time period when all units are under the same intervention. This is usually done to collect "control" data about each unit to establish its baseline. For example, in an e-commerce setting, companies commonly estimate the baseline engagement level of a customer to understand the treatment effect of a discount policy; similarly, in clinical trials, pharmaceutical companies collect health metrics of patients to establish the treatment effect a particular therapy has. Further, the assumption that such a control period exists is standard in the synthetic controls literature. For an illustration of the sparsity pattern in the induced matrix with a control period, see Figure 4.6b. Hence, this "control" period in sequential decision-making can serve as our anchor columns, and all units can serve as anchor rows.

4.5 Theoretical Results

Below, we establish the statistical properties of the SNN algorithm. Without loss of generality, we consider a specific pair (i, j). Recall from our discussion earlier, we suppress dependencies on (i, j), e.g., all anchor rows and columns $\text{AR}^{(k)}, \sim^{(k)}$ are defined with respect to (i, j). In Section 4.5.1, we state additional assumptions required to establish the theoretical results. In Sections 4.5.2 and 4.5.3, we establish finite-sample consistency and asymptotic normality of the SNN algorithm for a given entry (i, j). In Sections 4.5.4 and 4.5.5, we discuss our assumptions and theoretical results, respectively.

Notation. For every vector $v \in \mathbb{R}^a$, let v_p denotes its ℓ_p-norm. For the remainder of this work, let $\mathcal{E} = \{\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}\}$, i.e., the collection of latent factors and the observed missingness pattern. Recall the definition of $\boldsymbol{S}^{(k)}$, $\text{AR}^{(k)}$ and $\sim^{(k)}$ from Section 4.4.1. Moreover, for every $k \in [K]$, we denote the SVD of $\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}]$ as

$$\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}] = \sum_{\ell=1}^{r^{(k)}} \tau_\ell^{(k)} u_\ell^{(k)} \otimes v_\ell^{(k)};$$

here, $r^{(k)} = \text{rank}(\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}])$. We denote $\boldsymbol{U}^{(k)} \in \mathbb{R}^{|\text{AR}^{(k)}| \times r^{(k)}}$ and $\boldsymbol{V}^{(k)} \in \mathbb{R}^{|\sim^{(k)}| \times r^{(k)}}$ as the matrices of left and right singular vectors, respectively, i.e., $u_\ell^{(k)} \in \mathbb{R}^{|\text{AR}^{(k)}|}$ and $v_\ell^{(k)} \in \mathbb{R}^{|\sim^{(k)}|}$ form the ℓ-th columns of $\boldsymbol{U}^{(k)}$ and $\boldsymbol{V}^{(k)}$, respectively.

■ 4.5.1 Additional Assumptions

We state additional assumptions required to establish guarantees for the `SNN` algorithm. In Section 4.5.4 we provide interpretations for Assumptions 22 and 23; Assumptions 20 and 21 are relatively standard and self-explanatory. Below, k is indexed over $[K]$, where recall K is a hyper-parameter of the `SNN` algorithm.

Assumption 20 (Sub-gaussian noise). *Conditioned on $\mathcal{E}$, ε_{ij} are independent sub-gaussian mean-zero r.v.s with $\mathbb{E}[\varepsilon_{ij}^2] = \sigma_{ij}^2 \leq \sigma^2$ and $\varepsilon_{ij\psi_2} \leq C\sigma_{ij}$ for some constants $C > 0$ and $\sigma > 0$.*

Assumption 21 (Bounded expected potential outcomes). *Conditioned on $\mathcal{E}$, $A_{ij} \in [-1, 1]$.*[3]

Assumption 22 (Well-balanced spectra). *Conditioned on $\mathcal{E}$ and given a pair (i, j) as well as subgroup k, the $r^{(k)}$ nonzero singular values $\tau_\ell^{(k)}$ of $\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}]$ are well-balanced, i.e., there exist universal constants $c, c' > 0$ that satisfy*

$$\tau_{r^{(k)}}^{(k)} / \tau_1^{(k)} \geq c, \quad \mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}]_F^2 \geq c' |\text{AC}^{(k)}| \cdot |\text{AR}^{(k)}|.$$

Assumption 23 (Subspace inclusion). *Conditioned on $\mathcal{E}$ and given a pair (i, j) as well as subgroup k,*

$$\mathbb{E}[x^{(k)} \mid \mathcal{E}] \in \text{colspan}(\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}]),$$

where we recall $x^{(k)}$ is defined in Section 4.4.1.

■ 4.5.2 Finite-sample Consistency

The following result establishes that the `SNN` algorithm outputs entry-wise consistent estimates of $\boldsymbol{A}$, i.e., we establish consistency in $\cdot_{\max}$-norm. To simplify notation, we will henceforth absorb dependencies on σ into the constant within $O_p(\cdot)$. That is, we assume there exists an absolute constant $C \geq 0$ such that $\sigma \leq C$.

[3] The precise bound $[-1, 1]$ is without loss of generality, i.e., it can be extended to $[a, b]$ for any $a, b \in \mathbb{R}$ with $a \leq b$.

Theorem 4.5.1. *Conditioned on $\mathcal{E}$, for a given pair (i, j) and subgroup $k \in [K]$, suppose $|\mathrm{AR}^{(k)}|\geq \mu$ and let Assumptions 17 to 23 hold. Further, let $K = o(\min_k|\mathrm{AC}^{(k)}|^{10}|\mathrm{AR}^{(k)}|^{10})$. Finally, for each k, let $\lambda^{(k)} = rank(\mathbb{E}[\boldsymbol{S}^{(k)}])$, where $\lambda^{(k)}$ is defined as in Algorithm 1. Then,*

$$\widehat{A}_{ij} - A_{ij} = O_p\left(\frac{1}{K}\left\{\sum_{k=1}^{K}\frac{(r^{(k)})^{1/2}}{|\mathrm{AC}^{(k)}|^{1/4}} + \sum_{k=1}^{K}\frac{(r^{(k)})^{3/2}\widetilde{\beta}^{(k)}{}_1\log^{1/2}(|\mathrm{AC}^{(k)}||\mathrm{AR}^{(k)}|)}{\min\{|\mathrm{AC}^{(k)}|^{1/2}, |\mathrm{AR}^{(k)}|^{1/2}\}} + \left[\sum_{k=1}^{K}\widetilde{\beta}^{(k)}{}_2^2\right]^{1/2}\right\}\right).$$

where $\widetilde{\beta}^{(k)} = \mathcal{P}_{U^{(k)}}\beta^{(k)}$ is the projection of $\beta^{(k)}$ onto the subspace spanned by the columns of $\boldsymbol{U}^{(k)}$. We assume $\widetilde{\beta}^{(k)}{}_2\geq c$, for some absolute constant $c \geq 0$.

Corollary 4.5.1. *Suppose $|\sim^{(k)}|, |AR^{(k)}|= N$ for all $k \in [K]$. Let $\beta_{\max,2} = \max_k\widetilde{\beta}^{(k)}{}_2$, $\beta_{\max,1} = \max_k\widetilde{\beta}^{(k)}{}_1$, and $r_{\max} = \max_k r^{(k)}$. Let the setup of Theorem 4.5.1 hold. Then,*

$$\widehat{A}_{ij} - A_{ij} = O_p\left(\frac{r_{\max}^{1/2}}{N^{1/4}} + \frac{r_{\max}^{3/2}\cdot\beta_{\max,1}\cdot\log^{1/2}(N)}{N^{1/2}} + \frac{\beta_{\max,2}}{\sqrt{K}}\right)$$

Note, Theorem 4.5.1 does not require $N \to \infty$ to establish consistency of the SNN estimator. Rather, that $|\mathrm{AR}^{(k)}|, |\sim^{(k)}|$ is growing *on average* (ignoring logarithmic factors and dependence on $\beta^{(k)}, r^{(k)}, \sigma$). However, we state Corollary 4.5.1 to help further interpret our results in Section 4.5.5.

Implication for matrix completion with MCAR data. Proposition 2 below shows that SNN provides uniform entry-wise consistency for matrix completion with MCAR data as a special case if p, the probability of observing an entry, is sufficiently large.

Proposition 2 (SNN for matrix completion with MCAR data)**.** *Let the setup of Theorem 4.5.1 hold. Further, let $m = n = L$. Assume each entry (i, j) is revealed with uniform probability $p \in (0, 1]$, independent of everything else. Fix any $\delta > 0$. Let*

$$p \geq \left(\frac{Q}{L}\right)^{\frac{1}{Q^2}}$$

with $Q = C^*\delta^{-6}$, where C^* is a function only of $\beta^{(k)}, r^{(k)}$ for $k \in [K]$, σ, and $\log(L)$.

Then with probability at least $1 - \frac{C}{L^8}$, where $C > 0$ is an absolute constant, there exists sufficient anchor rows and columns, $AR^{(k)}, \sim^{(k)}$, such that uniformly for all $(i, j) \in [m] \times [n]$,

$$\widehat{A}_{ij} - A_{ij} = O_p(\delta).$$

Hence, for any fixed $p > 0$, we have that $\widehat{A}_{ij} - A_{ij} = o(1)$ uniformly for all $(i, j) \in [m] \times [n]$ as $L \to \infty$.

■ 4.5.3 Asymptotic Normality

The following establishes that the entry-wise estimate $\widehat{A}_{ij}$ of the SNN algorithm is asymptotically normal around the target causal parameter A_{ij}.

Theorem 4.5.2. *For a given pair (i, j) and subgroup k, let the setup of Theorem 4.5.1 hold. Define*

$$(\tilde{\sigma}^{(k)})^2 := \sum_{\ell \in \mathrm{AR}^{(k)}} (\widetilde{\beta}_\ell^{(k)} \sigma_{\ell j})^2$$

Further, let the following conditions holds

- *(i)* $K \to \infty$;
- *(ii)* $|\mathrm{AC}^{(k)}|, |\mathrm{AR}^{(k)}| \to \infty$ *for each k;*
- *(iii)* $r^{(k)} \widetilde{\beta}^{(k)2}_1 \log\left(|\mathrm{AC}^{(k)}||\mathrm{AR}^{(k)}|\right) = o(\min\{|\mathrm{AC}^{(k)}|, |\mathrm{AR}^{(k)}|\})$ *for each k;*
- *(iv)*

$$\sum_{k=1}^{K} \left(\frac{(r^{(k)})^{1/2}}{|\mathrm{AC}^{(k)}|^{1/4}} + \frac{(r^{(k)})^{3/2} \widetilde{\beta}^{(k)}{}_1 \log^{1/2}(|\mathrm{AC}^{(k)}||\mathrm{AR}^{(k)}|)}{\min\{|\mathrm{AC}^{(k)}|^{1/2}, |\mathrm{AR}^{(k)}|^{1/2}\}} \right) = o\left(\left[\sum_{k=1}^{K} (\tilde{\sigma}^{(k)})^2 \right]^{1/2} \right) \tag{4.5}$$

Then conditioned on $\mathcal{E}$,

$$\frac{K(\widehat{A}_{ij} - A_{ij})}{\left[\sum_{k=1}^{K} (\tilde{\sigma}^{(k)})^2\right]^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1) .$$

Remark 4.5.1. *Recall the notation in Corollary 4.5.1. Then one can easily verify a sufficient property for condition (iii) in Theorem 4.5.2 is*

$$r_{\max} \cdot \beta^2_{\max,1} \cdot \log(N) = o(N)$$

Further, let $\tilde{\sigma}_{\min} = \min_k \tilde{\sigma}^{(k)}$. Then one can easily verify a sufficient property for condition (iv) in Theorem 4.5.2 is

$$K = o\left(\tilde{\sigma}_{\min} \cdot \min\left\{\frac{N^{1/2}}{r_{\max}}, \frac{N}{r_{\max}^3 \cdot \beta_{\max,1}^2 \cdot \log(N)}\right\}\right) \tag{4.6}$$

If we ignore dependence on logarithmic factors and on $\beta_{\max,1}, r_{\max}, \tilde{\sigma}_{\min}$, (4.6) essentially requires that

$$K = o(N^{1/2}).$$

Practically, this can be interpreted as saying that to ensure valid confidence intervals, the number of synthetic nearest neighbours, i.e., K, we construct in `SNN` *cannot scale too quickly relative to the number of anchor rows and columns, i.e., $|\mathrm{AR}^{(k)}|, |\mathrm{AC}^{(k)}|$.*

■ 4.5.4 Discussion of Assumptions

Interpretation of Assumption 22. Assumption 22 requires that the nonzero singular values of $\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}]$ are well-balanced. Such an assumption is quite standard with the econometrics factor model and matrix completion literature. For example, it is analogous to incoherence-style conditions; see Assumption A of Bai and Ng (2020) and the discussion of theoretical results in Agarwal et al. (2021e). It is also closely related to the notion of pervasiveness, see Proposition 3.2 of Fan et al. (2018). Indeed, the assumption that there is a gap between the top few singular values of a matrix of interest, and the remaining singular values has been widely adopted in the econometrics literature of large dimensional factor analysis dating back to Chamberlain and Rothschild (1983). Crucially though, these works within econometrics (e.g. Bai and Ng (2020), Fan et al. (2018), Chamberlain and Rothschild (1983)) aim to accurately estimate the factors themselves, which require making additional assumptions about the spectra of the matrix of interest to ensure these factors are uniquely identifiable. Instead we simply require that these low-rank factors exist, but do not explicitly require accurately estimating them. Assumption 22 has also been shown to hold with high-probability for the canonical probabilistic generating process used to analyze probabilistic principal component analysis in Bishop (1999) and Tipping and Bishop (1999); here, the observations are assumed to be a high-dimensional embedding of a low-rank matrix with independent sub-Gaussian entries (see Proposition 4.2 of Agarwal et al. (2021e)). Within the matrix/tensor completion literature, for an overview of where the well-balanced spectra assumption is utilized, see Cai et al. (2021) and references therein. Practically speaking, Assumption 22 can be empirically validated by plotting the spectrum of $\boldsymbol{S}^{(k)}$, defined in Algorithm 1; if there is a natural "elbow" point in the singular spectrum of $\boldsymbol{S}^{(k)}$, i.e., there are a relatively small number of singular values that have a large and approximately equal magnitude, and the remaining singular values are significantly smaller, then Assumption 22 is likely to hold. For further discussion of this empirical robustness check, please refer to the related discussion in Agarwal et al. (2021c).

Interpretation of Assumption 23. Recall from Algorithm 1 that we learn the model $\widehat{\beta}^{(k)}$ by regressing $q^{(k)}$ on $\boldsymbol{S}^{(k)}$. $\widehat{A}_{ij}^{(k)}$ is then estimated by applying the model $\widehat{\beta}^{(k)}$ on the outcomes in $x^{(k)}$ (i.e., the entries in the j-th column of the rows $\mathrm{AR}^{(k)}$). The key question that remains is why would a model learned between $q^{(k)}$ and $\boldsymbol{S}^{(k)}$, *generalize* well to accurately estimate $A_{ij}^{(k)}$ using $\langle x^{(k)}, \widehat{\beta}^{(k)} \rangle$. Normally, in statistical learning, such generalization requires making distributional assumptions about the training data (i.e., $\boldsymbol{S}^{(k)}$) and the testing data (i.e., $x^{(k)}$). For example, each column of $\boldsymbol{S}^{(k)}$ and $x^{(k)}$ are sampled i.i.d. However, we do not want to make such an assumption as it is unrealistic in setting such as recommendation systems, e.g., the ratings users give different movies is likely to be neither identically nor independently distributed. Indeed, by conditioning on $\mathcal{E}$, we are implicitly conditioning on $\boldsymbol{U}$ and $\boldsymbol{V}$, which requires our analysis to be *instance dependent*, i.e., has to hold for the specific sampling of the latent factors $\boldsymbol{U}$ and $\boldsymbol{V}$. To circumvent making any distribution assumptions, we make the natural assumption that in expectation, $x^{(k)}$ lies within the linear span of $\boldsymbol{S}^{(k)}$. Such a condition is necessary as well for generalization, e.g., if every entry of $\mathbb{E}[\boldsymbol{S}^{(k)} \mid \mathcal{E}]$ is equal to 0, then no meaningful model $\widehat{\beta}^{(k)}$ can learned. Such an assumption has also been explored in Agarwal et al. (2021d,c); Agarwal and Singh (2021). In particular, in Agarwal et al. (2021c) the authors provide a data-driven hypothesis test to verify when such a condition holds.

■ 4.5.5 Discussion of Results

To ease the discussion of the interpretation of the results, we will ignore dependence on logarithmic factors, and $\beta^{(k)}, r^{(k)}, \sigma$.

Sample complexity. Note that even if $D_{ij} = 1$, estimating A_{ij} is not straightforward; we never get to observe A_{ij}, rather we only observe Y_{ij}, where $Y_{ij} = A_{ij} + \varepsilon_{ij}$. That is, even if $D_{ij} = 1$, our observation of A_{ij} is corrupted by noise and we only get a single sample of it.

Remarkably, despite having access to (at most) a single noisy sample of A_{ij}, the estimate $\widehat{A}_{ij}$ produced by the `SNN` algorithm is consistent and asymptotically normal around A_{ij}. Of course, this is assuming a low-rank factor model and a suitable observation pattern. Hypothetically, if we get K independent noisy samples of A_{ij}, denoted by $Y_{ij}^{(1)}, \dots, Y_{ij}^{(K)}$, the maximum likelihood estimator would be the empirical mean, $\frac{1}{K}\sum_{k=1}^{K} Y_{ij}^{(k)}$. In this hypothetical scenario, this empirical mean would concentrate around A_{ij} with error scaling as $O(K^{-1/2})$, i.e., with this estimation procedure to obtain an additive error of $O(\delta)$, we would need $K = \Omega(\delta^{-2})$ independent copies.

Now in comparison to the hypothetical scenario above where we have access to K independent samples, Corollary 4.5.1 effectively establishes that with access to at most $N^2 \times K$ observations, the error of the `SNN` estimator scales as $O(\max(N^{-1/4}, K^{-1/2}))$; this is assuming $|\sim^{(k)}|, |\mathrm{AR}^{(k)}|= N$ for all $k \in [K]$, as in Corollary 4.5.1. This implies for any $\delta > 0$, A_{ij} can be estimated to within an additive error of $O(\delta)$, if $N = \Omega(\delta^{-1/4})$ and $K = \Omega(\delta^{-2})$. Hence, compared to if we had K independent noisy samples of each A_{ij}, we pay an additional cost of N^2 in terms of the number of samples needed, and $N^{-1/4}$ in terms of the estimation error rate even though we either do not observe a sample of A_{ij} (i.e., it is missing), or only observe a single, noisy instantiation of it in $\widetilde{Y}_{ij}$. That is, to obtain estimation error of $O(\delta)$, it requires $O(\delta^{-2} \times \delta^{-4})$ observations (across different entries).

Further, in the hypothetical scenario where we get K independent noisy copies for each (i, j), if we wanted to estimate A_{ij} to within error $O(\delta)$ for all (i, j), this would require $m \times n \times K$ observations, with $K = \Omega(\delta^{-2})$. In contrast, for `SNN`, if we assume that for all (i, j), we can use the same set of anchor rows and columns, i.e., $\{|\sim^{(k)}|, |\mathrm{AR}^{(k)}|\}_{k\in[K]}$ can be chosen to be the same for all (i, j), then one can easily verify that the number of observations we need to recover each A_{ij} to within error $O(\delta)$ is at most $N^2 \times K + m \times N + n \times (N \times K)$,[4] with $N = \Omega(\delta^{-1/4})$ and $K = \Omega(\delta^{-2})$. See Figure 4.7 for a visual depiction of the observation pattern for which this holds. Thus, for any fixed $\delta > 0$, we can recover every entry A_{ij} to within additive error $O(\delta)$, with access to only $O(m + n)$ observations, rather than $O(m \times n)$ observations as would be naively required.

Connections to causal transportability, transfer learning, learning with distribution shift. We note that this problem of generalizing well without making an i.i.d assumption is known by a variety of terms across many fields of study; these include "causal transportability", "transfer learning", "learning with distribution shift". Given that subspace

[4] Technically, we only need $N^2 \times K + (m - N - K) \times N + (n - N) \times (N \times K)$.

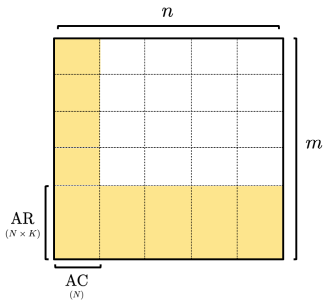


Figure 4.7: Sparsity pattern for which minimum number of observations required for entry-wise recovery.

inclusion, i.e., Assumption 23 holds, we show that generalization is possible without making any distributional assumptions about the underlying signal matrix $\boldsymbol{A}$. Indeed, our theoretical results in Theorem 4.5.1 and 4.5.2 can be interpreted as point-wise out-of-sample generalization error bounds, which are distribution free (i.e., instance dependent). This might be of independent interest.

■ 4.6 Experiments

The objective of this section is to compare the imputation accuracy of `SNN` against the state-of-the-art matrix completion algorithms for MNAR data. We describe these algorithms in Section 4.6.1. We do two case studies. In Section 4.6.2, we apply these various algorithms in the setting of recommender systems with different missingness patterns. In Section 4.6.3, we do the same but using data from a classic panel data case study in the econometrics literature called “California Prop 99” Abadie et al. (2010).

■ 4.6.1 Benchmark Matrix Completion Algorithms

In particular, we compare two types of algorithms for matrix completion against `SNN`; we choose these benchmarks to be in line with those considered in Ma and Chen (2019). The first group of algorithms does not account for entries being MNAR; these include `PMF` (Mnih and Salakhutdinov (2008)), `SVD` (Funk (2006)), `SVD++` (Koren (2008)), `softImpute` (Hastie et al. (2015)), and `KNN` (Lee et al. (2016)); we remark that the algorithm proposed in Athey et al. (2021) is similar to `softImpute` with the addition of separate fixed effects terms. In particular, both the algorithm design and associated analysis of these algorithms is for

MCAR data. In contrast, the second group does account for the limited MNAR setting as described in Section 4.2.2; these include `MaxNorm` (Cai and Zhou (2016)), `ExpoMF` (Liang et al. (2016)), and `WTN` (Srebro and Salakhutdinov (2010)).

With the exception of `ExpoMF`, we further consider IPW-variants of the other benchmark algorithms, i.e., for each algorithm, we first de-bias the loss function given in (4.3) via propensity scores. We do not do so with `ExpoMF` as their algorithm does not lend itself to be de-based via propensity scores in a straightforward manner (also see Ma and Chen (2019)).The propensity scores are estimated in two ways, which are in line with the MAR and limited MNAR setting described in Section 4.2.2. (i) MAR setting: We provide meaningful additional covariates $(X_i, \tilde{X}_j)$ for row i and column j and use logistic regression to learn $\widehat{p}_{ij}$; if a matrix completion algorithm is de-biased in this way, we add `LR` in front of it, e.g., `LR-PMF` means the `PMF` algorithm is used to estimate $\widehat{\boldsymbol{A}}$ and the loss function is de-biased using logistic regression. (ii) Limited MNAR setting: We do not provide additional covariates and directly estimate $\widehat{p}_{ij}$ using the observed mask matrix $\boldsymbol{D}$; this is done using the `1bitMC` algorithm in Davenport et al. (2014) and algorithms de-biased in this manner has a prefix of `1bitMC` added to them.; this approach to de-bias MNAR data is in line with what is proposed in Ma and Chen (2019); Yang et al. (2021); Bhattacharya and Chatterjee (2021).

We consider two error metrics, root mean-squared-error (RMSE) and mean-absolute-error (MAE). For all benchmark algorithms, we use 5-fold cross validation to tune their hyper-parameters through grid search for every error metric, i.e., for each benchmark algorithm, we find its best performing hyper-parameters with respect to RMSE and MAE on the validation set and report the error metric-specific hyper-parameters on the test set for each error metric. For `SNN`, we choose $K = 1$ and $\lambda^{(1)}$ as per Gavish and Donoho (2014), i.e., we do not tune the hyper-parameters of `SNN` nor do we optimize it for each error metric. We emphasize that only the algorithms with the `LR` prefix use the additional row and column covariates $X_i, \tilde{X}_j$.

■ 4.6.2 Recommendation Systems

We begin with recommendation systems, which is arguably the canonical matrix completion application. Through the recommendation systems setting, we present two MNAR missingness patterns—one obeys the standard assumptions on MNAR in the literature (which we often refer to as limited MNAR) while the other considers a more general

MNAR setting. To better understand the effect of the underlying mechanism which leads to missingness on each algorithm's ability to perform imputation, we consider the "noiseless" case, i.e., $\widetilde{Y}_{ij} = A_{ij}$ if $D_{ij} = 1$ and $\widetilde{Y}_{ij} = \star$ otherwise. We study the effect of additional noise ε_{ij} in the panel data setting in Section 4.6.3.

Limited MNAR Setting: Positivity & Independent Missingness

In our first illustration, our observation pattern reflects the self-selection bias phenomena where most users tend to provide ratings if they particularly liked or disliked an item. However, they are much less inclined to provide a rating for an item that they are lukewarm about. Our simulated setup also consists of "core users" and "core movies". We use core users to represent movie fanatics or critics, for instance, who provide explicit feedback for a significant number of films. In the setting of movie recommendations, we use core items to represent iconic movies such as *Star Wars* or *Titanic* that have influenced future films and popular culture, and are largely viewed by the general audience. These can also represent the subset of items that online platforms such as *Hulu, Netflix, Stichfix* display to new users when prompting for their preferences.

Experimental setup. We consider $m = 80$ users and $n = 80$ movies. We choose the dimension of the latent space as $r = 5$. We generate the latent user matrix $\boldsymbol{U} \in \mathbb{R}^{m \times r}$ as follows: (i) we first choose $m_{\text{core}} = 20$ core users and construct $\boldsymbol{U}_0 \in \mathbb{R}^{m_{\text{core}} \times r}$ by sampling entries i.i.d. from a standard normal distribution; (ii) next, we construct $\boldsymbol{U}_1 = \boldsymbol{B}\boldsymbol{U}_0 \in \mathbb{R}^{(m-m_{\text{core}}) \times r}$, where the entries in $\boldsymbol{B} \in \mathbb{R}^{(m-m_{\text{core}}) \times m_{\text{core}}}$ are sampled i.i.d. from a Dirichlet distribution, which ensures that the new factors lie in the same intervals as the factors in $\boldsymbol{U}_0$. In doing so, every row of $\boldsymbol{U}_1$, representing the latent factors corresponding to the "standard" users, is a linear combination of that of core users $\boldsymbol{U}_0$, i.e., every standard user can be expressed as a weighted combination of core users. We then define $\boldsymbol{U} = [\boldsymbol{U}_0, \boldsymbol{U}_1]$ such that the first m_{core} rows of $\boldsymbol{U}$ correspond to the core users. We construct $\boldsymbol{V} = [\boldsymbol{V}_0, \boldsymbol{V}_1] \in \mathbb{R}^{n \times r}$ similarly, where $\boldsymbol{V}_0 \in \mathbb{R}^{n_{\text{core}} \times r}$ and $\boldsymbol{V}_1 \in \mathbb{R}^{(n-n_{\text{core}}) \times r}$ represent the matrix of latent factors associated with core movies and standard movies, respectively; here, we choose $n_{\text{core}} = 20$. We form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^T \in \mathbb{R}^{m \times n}$ and scale the values to lie within the interval $[1, 5]$; by construction, $\boldsymbol{A}$ is a low-rank matrix.

Finally, we generate user and movie covariates matrices $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{Q}_1 \in \mathbb{R}^{m \times 3}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{Q}_2 \in \mathbb{R}^{n \times 3}$, where the entries in $\boldsymbol{Q}_1 \in \mathbb{R}^{r \times 3}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{r \times 3}$ sampled i.i.d. from a standard normal $\mathcal{N}(0, 1)$; additionally, we normalize the columns in $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ to have unit ℓ_2-norm.

Next, we describe our generative model for the propensity matrix $\boldsymbol{P} \in \mathbb{R}^{m \times n}$. Without loss of generality, we denote $\mathcal{C}_{\text{core}} := \{(i, j) : i \leq m_{\text{core}}, j \leq n_{\text{core}}\}$ as the subset of core users and core movies, $\mathcal{C}_{\text{user}} := \{(i, j) : i \leq m_{\text{core}}, j > n_{\text{core}}\}$ as the subset of core users and standard movies, $\mathcal{C}_{\text{item}} := \{(i, j) : i > m_{\text{core}}, j \leq n_{\text{core}}\}$ as the subset of standard users and core movies, and $\mathcal{C}_{\text{standard}} := \{(i, j) : i > m_{\text{core}}, j > n_{\text{core}}\}$ as the subset of standard users and standard movies. These will represent our four cohorts of interest. Next, for some $t \in (1, 5)$, $\kappa_{ij} > 0$, and $\alpha_{ij} \in (0, 1]$,

$$p_{ij} = \begin{cases} \kappa_{ij} \cdot \alpha_{ij}^{A_{ij}-1}, & \text{if } A_{ij} \in [1, t] \\ \kappa_{ij} \cdot \alpha_{ij}^{5-A_{ij}}, & \text{if } A_{ij} \in (t, 5]. \end{cases}$$

In our setting, we choose our threshold $t = 2.3$. Here, α_{ij} is a parameter that controls the MNAR effect: $\alpha_{ij} = 1$ is MCAR while $\alpha_{ij} \to 0$ only reveals 1 and 5 rated movies. We choose $\alpha_{ij} = 0.7$ for $(i, j) \in \mathcal{C}_{\text{core}}$, $\alpha_{ij} = 0.35$ for $(i, j) \in \mathcal{C}_{\text{user}}, \mathcal{C}_{\text{item}}$, and $\alpha_{ij} = 0.1$ for $(i, j) \in \mathcal{C}_{\text{standard}}$. For every (i, j) pair, κ_{ij} is set so that the expected number of revealed ratings within the cohort is equal to some value. We choose the expected number of observations within $\mathcal{C}_{\text{core}}$ as 90%, within $\mathcal{C}_{\text{user}}$ as 70%, within $\mathcal{C}_{\text{item}}$ as 70%, and within $\mathcal{C}_{\text{standard}}$ as 5%. This sampling process ensures the two key assumptions in the limited MNAR setting of the entries of $\boldsymbol{D}$ being independent and $p_{\min} > 0$ are satisfied. See Figure 4.1b for a visual depiction of empirical sparsity pattern under this missingness mechanism.

Results. In the following simulations, we obey the generative process above. In particular, we sample $\boldsymbol{A}$ and $\boldsymbol{P}$ once, as well as $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, and perform 10 experimental repeats where the only randomization lies in the sparsity pattern, i.e., we observe 10 independent realizations of $\boldsymbol{D}$. We report the average RMSEs and MAEs, as well as their respective standard deviations, over the 10 experimental runs in Table 4.1. We find that with respect to MAE, `SNN` achieves the best result along with `MaxNorm` (and its variants); with respect to RMSE, `SNN` is a close second with `SVD++` (and its variants), after `MaxNorm` (and its variants). Although positivity and independence between entries in $\boldsymbol{D}$ are upheld, we remark that debiasing via `1bitMC` and `LR-` do not always yield stronger results, e.g., see `PMF` and `softImpute`.

A More General MNAR Setting: Violating Positivity & Independence Assumptions

In this simulation, we violate two key assumptions in the current literature on MNAR data: (i) positivity and (ii) independence between the entries in $\boldsymbol{D}$. Towards this, we continue the notion of core movies, for which all users provide ratings. For the remaining movies, users only provide ratings if a movie belong to their favorite genre. This deterministically sets every entry in $\boldsymbol{P}$ (and thus $\boldsymbol{D}$) to either 0 or 1, and correlates the entries in $\boldsymbol{D}$, which yields a sparsity pattern similar to that shown in Figure 4.6a. See Figure 4.1c for a visual depiction of empirical sparsity pattern under this missingness mechanism.

Experimental setup. In particular, we consider $m = 80$ users and $n = 80$ items. We choose the dimension of the latent space as $r = 5$. We generate the latent user matrix $\boldsymbol{U} \in \mathbb{R}^{m \times r}$ by sampling entries i.i.d. from a standard normal distribution. To generate the latent item matrix $\boldsymbol{V} \in \mathbb{R}^{n \times r}$, we first choose $n_{\text{core}} = 30$ core items (to be defined in greater detail below) and construct $\boldsymbol{V}_0 \in \mathbb{R}^{n_{\text{core}} \times r}$ by sampling entries i.i.d. from a standard normal. Next, we construct $\boldsymbol{V}_1 = \boldsymbol{B}\boldsymbol{V}_0 \in \mathbb{R}^{(n-n_{\text{core}}) \times r}$, where the entries in $\boldsymbol{B} \in \mathbb{R}^{(n-n_{\text{core}}) \times n_{\text{core}}}$ are sampled i.i.d. from a Dirichlet distribution. In doing so, every row of $\boldsymbol{V}_1$ is a linear combination of rows in $\boldsymbol{V}_0$, i.e., every item can be expressed as a weighted combination of core items. We then define $\boldsymbol{V} = [\boldsymbol{V}_0, \boldsymbol{V}_1]$ such that the first n_{core} rows of $\boldsymbol{V}$ correspond to the core items. We form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^T \in \mathbb{R}^{m \times n}$ and scale the values to lie within the interval $[1, 5]$. Finally, we generate user and item feature matrices $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{Q}_1 \in \mathbb{R}^{m \times 10}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{Q}_2 \in \mathbb{R}^{n \times 10}$, where the entries in $\boldsymbol{Q}_1 \in \mathbb{R}^{r \times 10}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{r \times 10}$ sampled i.i.d. from a standard normal $\mathcal{N}(0, 1)$; additionally, we normalize the columns in $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ to have unit ℓ_2-norm. We generate higher dimensional covariates $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ to see if improves the relative performance of the MAR algorithms, denoted by the prefix LR, which use this additional information to estimate the propensities.

To describe the generating process for the observation pattern $\boldsymbol{D}$, we begin by providing an interpretation of the above quantities. First, we interpret r as the number of latent genres. In turn, the (i, k)-th entry in $\boldsymbol{U}$ can be interpreted as user i's preference for genre k; similarly, the (j, k)-th entry in $\boldsymbol{V}$ can be interpreted as the level to which item j is composed of genre k. We consider the setting where all users provide ratings for all core items, i.e., $D_{ij} = 1$ for every user $i \in [m]$ and core item $j \in [n_{\text{core}}]$. For the remaining entries in $\boldsymbol{D}$, we posit that every user will only rate items from their favorite genre. More specifically, given the above interpretation, we define user i's favorite genre $k^*(i)$ as $k^*(i) = \arg\max_{k \in [r]} U_{ik}$; similarly, we classify an item j as belonging to genre $k^\sharp(j)$ if

$k^\sharp(j) = \arg\max_{k\in[r]} V_{jk}$. Hence, for every user $i \in [m]$ and non-core item $j > n_{\text{core}}$, we have $D_{ij} = 1$ if $k^*(i) = k^\sharp(j)$ and 0 otherwise. We underscore that this model violates the standard operating assumptions within the current MNAR literature as entries in $\boldsymbol{D}$ are deterministically set to 0 (i.e., the minimum element in $\boldsymbol{P}$ is 0), and are dependent on one another.

Results. In the following simulations, we obey the generative process above. In particular, we sample $\boldsymbol{V}$ and $\boldsymbol{X}$ once, and perform 10 experimental repeats where the only randomization lies in the re-sampling of $\boldsymbol{U}$; this is done to model new users coming into the system with the movies fixed. We report the average RMSEs and MAEs, as well as their respective standard deviations, over the 10 experimental runs in Table 4.1. We find that SNN achieves the best RMSE and MAE, with 1bitMC-MaxNorm as a close second with respect to RMSE and MAE. As with the limited MNAR setting experiment, we find that de-biasing does not always improve results. This is reasonable given that our generative process violates the typical assumptions underlying propensity estimation methods. The relative improvement of SNN shows its robustness to the general MNAR setting, where entry-wise positivity and independence of $\boldsymbol{D}$ are violated. The fact that KNN performs relatively poorly indicates that matching via linear weights is indeed more expressive than matching with uniforms weights as done in KNN. WE also note that the various state-of-the-art algorithms are still relatively robust to the general MNAR setting. This may warrant further investigation into the potential gap between theory and practice on the robustness of these methods to different missingness patterns.

■ 4.6.3 Panel Data

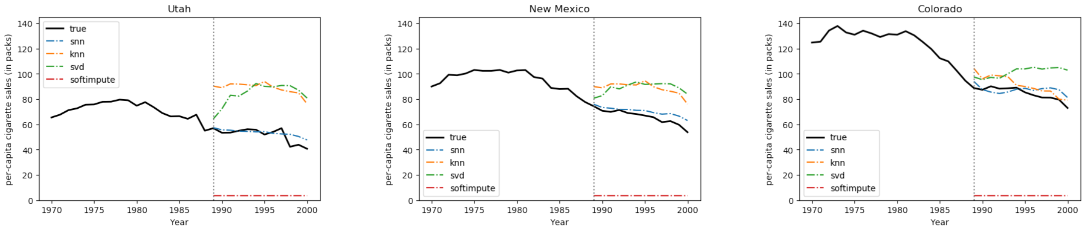
We now compare SNN against the same benchmark matrix completion algorithms using a classic case study of California smoking data of Abadie et al. (2010), which has been widely utilized within the econometrics literature. We do so as this setting has a MNAR sparsity pattern which is quite distinct from what is seen in recommendation systems. We now give a brief overview of the case study. In 1988, California introduced the first modern-time large-scale anti-tobacco legislation in the United States (Proposition 99). There was interest in estimating the effect of this legislation on tobacco sales in California. Towards this, per-capita cigarette sales data was collected across 39 U.S. states from 1970 to 2000. Among the 39 states, there was one "treated" state, California, which implemented the legislation; the remaining 38 states were chosen as "control"

Algorithm	Rec Sys (limited MNAR)		Rec Sys (general MNAR)		Panel Data	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
PMF	0.30 ± 0.02	0.25 ± 0.02	0.64 ± 0.14	0.56 ± 0.11	89.4 ± 92	105 ± 94
1bitMC-PMF	0.42 ± 0.02	0.36 ± 0.02	0.69 ± 0.12	0.61 ± 0.11	69.2 ± 85	84.6 ± 91
LR-PMF	0.28 ± 0.02	0.28 ± 0.02	0.66 ± 0.15	0.57 ± 0.13	32.1 ± 56	47.5 ± 76
SVD	0.14 ± 0.01	0.11 ± 0.00	0.49 ± 0.03	0.39 ± 0.03	14.9 ± 2.3	10.9 ± 1.6
1bitMC-SVD	0.15 ± 0.01	0.12 ± 0.01	0.49 ± 0.03	0.39 ± 0.03	15.0 ± 2.3	10.9 ± 1.6
LR-SVD	0.14 ± 0.01	0.11 ± 0.00	0.49 ± 0.03	0.39 ± 0.03	15.0 ± 2.3	10.9 ± 1.6
SVD++	0.07 ± 0.02	0.06 ± 0.01	0.44 ± 0.03	0.34 ± 0.03	161 ± 76	160 ± 76
1bitMC-SVD++	0.08 ± 0.02	0.08 ± 0.01	0.45 ± 0.03	0.35 ± 0.03	143 ± 86	141 ± 87
LR-SVD++	0.08 ± 0.02	0.08 ± 0.01	0.44 ± 0.03	0.35 ± 0.03	180 ± 57	178 ± 57
softImpute	1.03 ± 0.05	0.89 ± 0.04	1.50 ± 0.06	1.44 ± 0.05	101 ± 4.1	99.1 ± 4.1
1bitMC-softImpute	1.21 ± 0.04	1.06 ± 0.03	1.52 ± 0.09	1.46 ± 0.09	100 ± 4.1	97.7 ± 4.1
LR-softImpute	1.03 ± 0.05	0.89 ± 0.04	1.50 ± 0.06	1.44 ± 0.05	103 ± 3.8	101 ± 3.9
WTN	0.13 ± 0.01	0.10 ± 0.01	0.52 ± 0.13	0.44 ± 0.11	99.9 ± 4.1	97.7 ± 4.1
1bitMC-WTN	0.10 ± 0.01	0.08 ± 0.00	0.55 ± 0.15	0.47 ± 0.14	100 ± 4.1	97.8 ± 4.1
LR-WTN	0.12 ± 0.01	0.10 ± 0.00	0.52 ± 0.16	0.43 ± 0.15	99.9 ± 4.1	97.8 ± 4.1
MaxNorm	0.05 ± 0.01	0.03 ± 0.01	0.29 ± 0.08	0.20 ± 0.06	99.9 ± 4.1	97.7 ± 4.1
1bitMC-MaxNorm	0.05 ± 0.01	0.03 ± 0.01	0.23 ± 0.06	0.17 ± 0.05	100 ± 4.1	97.8 ± 4.1
LR-MaxNorm	0.05 ± 0.01	0.03 ± 0.01	0.31 ± 0.09	0.20 ± 0.06	100 ± 4.1	97.8 ± 4.1
ExpoMF	2.08 ± 0.01	2.00 ± 0.01	1.99 ± 0.05	1.90 ± 0.05	75.0 ± 5.5	64.8 ± 19
KNN	0.51 ± 0.02	0.40 ± 0.02	0.40 ± 0.06	0.30 ± 0.05	15.4 ± 2.5	12.0 ± 1.7
SNN	0.08 ± 0.01	0.03 ± 0.01	0.20 ± 0.07	0.11 ± 0.06	10.3 ± 1.0	8.00 ± 0.7

Table 4.1: RMSEs and MAEs of matrix completion methods on a recommender system experiment and a panel data experiment. The first two columns correspond to Section 4.6.2, the middle two columns correspond to Section 4.6.2, and the final two columns correspond to Section 4.6.3. The results are the averages ± standard deviations across 10 experimental repeats.

states as they neither instituted a tobacco control program nor raised cigarette sales taxes by 50 cents or more. These other 38 control states were then used to build a "synthetic California", i.e., a synthetic trajectory of cigarette sales in California if it had not introduced any tobacco legislation.

Experimental setup. We consider the time horizon of $n = 31$ years and restrict our focus to the $m = 38$ control units in the original dataset. This data is encoded into a 38×31 matrix, $\boldsymbol{Y}$, where the entry Y_{ij} represents the potential outcome of per-capita cigarette sales (in packs) for state i in year j under control, i.e., without any intervention in place. To generate MNAR data, we artificially introduce interventions to a subset of states in 1989, where the probability a state adopts an intervention (e.g., tobacco control program) depends on their change in cigarette sales pre- and post-1989. More specifically, we consider the following treatment adoption protocol: First, we cluster states into three



(a) Mild state: Utah.

(b) Moderate state: New Mexico.

(c) Severe state: Colorado.

Figure 4.8: True observations are represented in black, SNN estimates shown in blue, KNN estimates shown in orange, SVD estimates shown in green, softImpute estimates shown in red.

categories—mild, moderate, or severe—based on their change in average cigarette sales during 1989-2000 compared to that during 1970-1988; we note that in this context, a negative change means that the cigarette sales in the post-intervention period are lower than that in the pre-intervention period. As such, we define (i) mild states as those whose change is at least one standard deviation above the average change across all states; (ii) severe states as those whose change is at least one standard deviation below the average change across all states; (iii) and moderate states as the remaining states whose change is within one standard deviation.

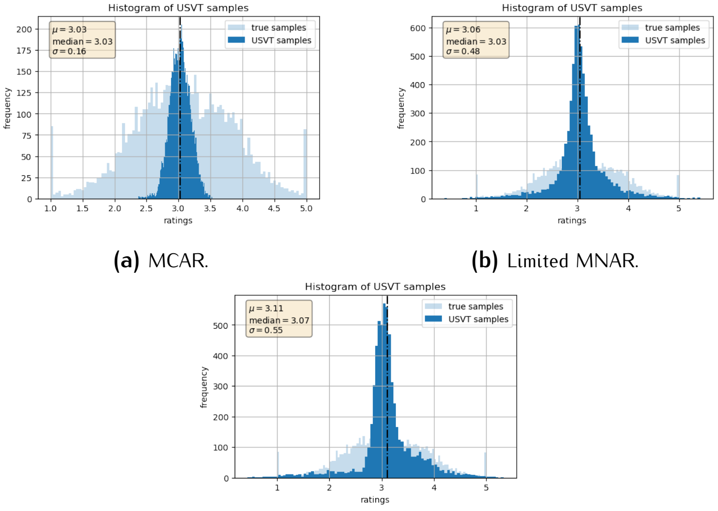
We then designate the probability of intervention for mild, moderate, and severe states as 10%, 30%, and 50%, respectively. In words, this setup reflects the scenario in which a state is more likely to adopt an intervention if their average sales in the post-intervention period is relatively closer to their pre-intervention sales compared to that of their peer states. In the language of causal inference, this is exactly confounding, i.e., there is a correlation between the treatment assignment and the eventual outcome.

For an example of a mild, moderate, and severe state, please see Figure 4.8. Additionally, we remark that once an intervention is adopted, all sales under control during the post-intervention period are, by definition, unobserved, i.e., $\widetilde{Y}_{ij} = \star$ for any intervened on state i and for all $j \geq 19$ (after 1988); this yields the observation pattern shown in Figure 4.6b. Finally, to employ logistic regression, i.e., LR to de-bias the estimates, we use state covariate data from Abadie et al. (2010), which include average retail price of cigarettes, per capita state personal income (logged), the percentage of the population age 15-24, and per capita beer consumption. We note that SNN does not use this auxiliary data.

Results. Using the above setup, we apply the various matrix completion methods to impute the missing counterfactual cigarette sales associated with the artificial intervention states during the post-intervention period. We report the average root mean-squared-errors (RMSEs) and mean absolute errors (MAEs), as well as their respective standard deviations, over 10 experimental runs in Table 4.1. As the table shows, SNN significantly outperforms all baseline algorithms under both error metrics. The only exception is KNN, which performs similarly to SNN; this is interesting as KNN is in essence, the difference-in-differences estimator, a standard method within the panel data econometrics literature. Further, SVD++ and MaxNorm (and its variants), which performed strongly in the recommendation systems example, now incur a significant error. We display a few representative results in Figure 4.8. Collectively across all three studies, we find that SNN is robust under varying missingness mechanisms.

4.7 Original USVT Algorithm Experiments

We run the same experiments in Section 4.1.1 using the original USVT estimator of Chatterjee (2015) rather than the modified version as proposed in Bhattacharya and Chatterjee (2021) for MNAR data. See Figure 4.9 below. Interestingly, we find the original USVT estimator performs better. Compare Figures 4.9a, 4.9b, 4.9c with Figures 4.2b, 4.3b 4.4b, respectively.



(a) MCAR.

(b) Limited MNAR.

(c) General MNAR.

Figure 4.9: Original USVT algorithm under the three different experiments.

■ 4.8 Proof of Theorem 4.3.1

In what follows, the descriptors above the equalities represent the assumption used, e.g., $A1$ represents Assumption 1:

$$\begin{aligned}
A_{ij} &= \mathbb{E}[Y_{ij} | u_i, v_j] \\
&\overset{A1}{=} \mathbb{E}[\langle u_i, v_j \rangle + \varepsilon_{ij} \mid u_i, v_j] \\
&\overset{A2}{=} \langle u_i, v_j \rangle \mid \{u_i, v_j\} \\
&= \langle u_i, v_j \rangle \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D} \\
&\overset{A3}{=} \sum_{\ell \in \mathcal{I}} \beta_\ell \cdot \langle u_\ell, v_j \rangle \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D} \\
&\overset{A2}{=} \sum_{\ell \in \mathcal{I}} \beta_\ell \cdot \mathbb{E}\left[\langle u_\ell, v_j \rangle + \varepsilon_{\ell j} \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}\right] \\
&\overset{A1}{=} \sum_{\ell \in \mathcal{I}} \beta_\ell \cdot \mathbb{E}\left[Y_{\ell j} | \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}\right] \\
&= \sum_{\ell \in \mathcal{I}} \beta_\ell \cdot \mathbb{E}\left[\widetilde{Y}_{\ell j} | \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}\right].
\end{aligned}$$

■ 4.9 Proof of Theorem 4.5.1

For ease of notation, we suppress the conditioning on $\mathcal{E}$ for the remainder of the proof. Further, for every k, let $\varepsilon^{(k)} = [\varepsilon_{\ell j} : \ell \in \text{AR}^{(k)}] \in \mathbb{R}^{|\text{AR}^{(k)}|}$ and $\Delta^{(k)} = \widehat{\beta}^{(k)} - \widetilde{\beta}^{(k)}$. We also recall the definitions provided in Section 4.4.

To begin, recall that $|\text{AR}^{(k)}| \geq \mu$ for each k by assumption. Thus, by Theorem 4.3.1, there exists a $\beta^{(k)} \in \mathbb{R}^{|\text{AR}^{(k)}|}$ such that $A_{ij} = \langle \mathbb{E}[x^{(k)}], \beta^{(k)} \rangle$ for every k, i.e.,

$$A_{ij} = \frac{1}{K} \sum_{k=1}^{K} \langle \mathbb{E}[x^{(k)}], \beta^{(k)} \rangle. \tag{4.7}$$

Additionally, under Assumption 23, we have $\mathbb{E}[x^{(k)}] = \mathcal{P}_{U^{(k)}} \mathbb{E}[x^{(k)}]$. In turn, this implies

$$\langle \mathbb{E}[x^{(k)}], \beta^{(k)} \rangle = \langle \mathbb{E}[x^{(k)}], \widetilde{\beta}^{(k)} \rangle \tag{4.8}$$

$$\langle \mathbb{E}[x^{(k)}], \Delta^{(k)} \rangle = \langle \mathbb{E}[x^{(k)}], \mathcal{P}_{U^{(k)}} \Delta^{(k)} \rangle, \tag{4.9}$$

where $\widetilde{\beta}^{(k)} = \mathcal{P}_{U^{(k)}}\beta^{(k)}$. Together, (4.7), (4.8), and (4.9) yield the following:

$$\begin{aligned}
\widehat{A}_{ij} - A_{ij} &= \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{A}_{ij}^{(k)} - A_{ij}\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\langle x^{(k)}, \widehat{\beta}^{(k)}\rangle - \langle \mathbb{E}[x^{(k)}], \beta^{(k)}\rangle\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\langle x^{(k)}, \widehat{\beta}^{(k)}\rangle - \langle \mathbb{E}[x^{(k)}], \widetilde{\beta}^{(k)}\rangle\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\langle \mathbb{E}[x^{(k)}], \Delta^{(k)}\rangle + \langle \varepsilon^{(k)}, \widetilde{\beta}^{(k)}\rangle + \langle \varepsilon^{(k)}, \Delta^{(k)}\rangle\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\langle \mathbb{E}[x^{(k)}], \mathcal{P}_{U^{(k)}}\Delta^{(k)}\rangle + \langle \varepsilon^{(k)}, \widetilde{\beta}^{(k)}\rangle + \langle \varepsilon^{(k)}, \Delta^{(k)}\rangle\right). \qquad (4.10)
\end{aligned}$$

Below, we bound the three terms on the right-hand side (RHS) of (4.10) separately.

Bounding term 1. By Cauchy-Schwartz inequality, we obtain for every k

$$\langle \mathbb{E}[x^{(k)}], \mathcal{P}_{U^{(k)}}\Delta^{(k)}\rangle \leq \mathbb{E}[x^{(k)}]_2 \cdot \mathcal{P}_{U^{(k)}}\Delta^{(k)}_2.$$

Under Assumption 21, we have $\mathbb{E}[x^{(k)}]_2 \leq |\text{AR}^{(k)}|^{1/2}$. As such,

$$\frac{1}{K}\sum_{k=1}^{K}\langle \mathbb{E}[x^{(k)}], \mathcal{P}_{U^{(k)}}\Delta^{(k)}\rangle \leq \frac{1}{K}\sum_{k=1}^{K}|\text{AR}^{(k)}|^{1/2} \cdot \mathcal{P}_{U^{(k)}}\Delta^{(k)}_2. \qquad (4.11)$$

To bound the expression above, we use the following lemma; its proof is found in Appendix 4.9.1.

Lemma 4.9.1 (Lemma G.1 of Agarwal et al. (2021c)). *Consider the setup of Theorem 4.5.1. Then for any k,*

$$\mathcal{P}_{U^{(k)}}\Delta^{(k)} = O_p\left(\frac{(r^{(k)})^{1/2}}{|AR^{(k)}|^{1/2}|\sim^{(k)}|^{1/4}} + \frac{(r^{(k)})^{3/2}\widetilde{\beta}^{(k)}{}_1 \log^{1/2}(|\sim^{(k)}||AR^{(k)}|)}{|AR^{(k)}|^{1/2}\ \min\{|\sim^{(k)}|^{1/2}, |AR^{(k)}|^{1/2}\}}\right).$$

Plugging Lemma 4.9.1 into (4.11), we conclude

$$\frac{1}{K}\sum_{k=1}^{K}\langle \mathbb{E}[x^{(k)}], \mathcal{P}_{U^{(k)}}\Delta^{(k)}\rangle = O_p\left(\frac{1}{K}\sum_{k=1}^{K}\frac{(r^{(k)})^{1/2}}{|\sim^{(k)}|^{1/4}} + \frac{(r^{(k)})^{3/2}\widetilde{\beta}^{(k)}{}_1 \log^{1/2}(|\sim^{(k)}||\text{AR}^{(k)}|)}{\min\{|\sim^{(k)}|^{1/2}, |\text{AR}^{(k)}|^{1/2}\}}\right) (4.12)$$

Bounding term 2. We begin with a lemma that is an immediate consequence of Hoeffding's Lemma.

Lemma 4.9.2. *Let γ_k be a sequence of mean zero sub-gaussian r.v.s with $\mathbb{E}[\gamma_k^2] = \sigma_k^2$. Then,*

$$\frac{1}{K}\sum_{k=1}^{K}\gamma_k = O_p\left(\frac{1}{K}\left[\sum_{k=1}^{K}\sigma_k^2\right]^{1/2}\right).$$

By Assumption 20, we have for any k,

$$\begin{aligned}\mathbb{E}[\langle\varepsilon^{(k)}, \widetilde{\beta}^{(k)}\rangle] &= 0\\ \mathrm{Var}(\langle\varepsilon^{(k)}, \widetilde{\beta}^{(k)}\rangle) &= \sum_{\ell\in\mathrm{AR}^{(k)}}(\widetilde{\beta}_\ell^{(k)}\sigma_{\ell j})^2.\end{aligned}\tag{4.13}$$

Since $\langle\varepsilon^{(k)}, \widetilde{\beta}^{(k)}\rangle$ are independent across k, noting that $\sum_{\ell\in\mathrm{AR}^{(k)}}(\widetilde{\beta}_\ell^{(k)}\sigma_{\ell j})^2 \leq \sigma^2\widetilde{\beta}^{(k)2}_2$, and applying Lemma 4.9.2 yields

$$\frac{1}{K}\sum_{k=1}^{K}\langle\varepsilon^{(k)}, \widetilde{\beta}^{(k)}\rangle = O_p\left(\frac{\sigma}{K}\left[\sum_{k=1}^{K}\widetilde{\beta}^{(k)2}_2\right]^{1/2}\right).\tag{4.14}$$

Bounding term 3. We begin by stating a helpful lemma below, the proof of which can be found in Section 4.9.2.

Lemma 4.9.3 (Lemma F.2 of Agarwal et al. (2021c))**.** *Let the setup of Theorem 4.5.1 hold. Then for every k, the following holds with probability at least $1 - O((|\sim^{(k)}||AR^{(k)}|)^{-10})$:*

$$\Delta^{(k)2}_2 \leq C(\sigma)\frac{r^{(k)}\widetilde{\beta}^{(k)2}_2\log\left(|\sim^{(k)}||AR^{(k)}|\right)}{\min\{|\sim^{(k)}|, |AR^{(k)}|\}},$$

where $C(\sigma)$ is a constant that only depends only σ.

By Lemma 4.9.3, it immediately follows that

$$\Delta^{(k)} = O_p\left(\frac{(r^{(k)})^{1/2}\widetilde{\beta}^{(k)}{}_2\log^{1/2}(|\sim^{(k)}||\mathrm{AR}^{(k)}|)}{\min\{|\sim^{(k)}|^{1/2}, |\mathrm{AR}^{(k)}|^{1/2}\}}\right).$$

For every k, we define the event $\mathcal{E}_k$ as

$$\mathcal{E}_k = \left\{ \Delta^{(k)2}_2 \leq \frac{r^{(k)}\widetilde{\beta}^{(k)2}_2 \log\left(|\sim^{(k)}||\mathrm{AR}^{(k)}| \right)}{\min\{|\sim^{(k)}|, |\mathrm{AR}^{(k)}|\}} \right\}.$$

For ease of notation, let $\mathcal{E}_\sharp = \cap_{k=1}^K \mathcal{E}_k$. Next, we define the event

$$\mathcal{E}_\flat = \left\{ \frac{1}{K}\sum_{k=1}^{K} \langle \varepsilon^{(k)}, \Delta^{(k)} \rangle = O\left(\frac{\sigma}{K} \left[\sum_{k=1}^{K} \frac{r^{(k)}\widetilde{\beta}^{(k)2}_2 \log\left(|\sim^{(k)}||\mathrm{AR}^{(k)}| \right)}{\min\{|\sim^{(k)}|, |\mathrm{AR}^{(k)}|\}} \right]^{1/2} \right) \right\}.$$

Now, condition on $\mathcal{E}_\sharp$. By Assumption 20, we have for every k,

$$\begin{aligned} \mathbb{E}[\langle \varepsilon^{(k)}, \Delta^{(k)} \rangle] &= 0 \\ \mathrm{Var}(\langle \varepsilon^{(k)}, \Delta^{(k)} \rangle) &= \sum_{\ell \in \mathrm{AR}^{(k)}} \sigma^2_{\ell j} \, (\widehat{\beta}^{(k)}_\ell - \widetilde{\beta}^{(k)}_\ell)^2 \leq \sigma^2 \Delta^{(k)2}_2. \end{aligned} \tag{4.15}$$

Given that $\langle \varepsilon^{(k)}, \Delta^{(k)} \rangle$ are independent across k, Lemmas 4.9.2, 4.9.3, and (4.15) imply $\mathcal{E}_\flat | \mathcal{E}_\sharp$ occurs w.h.p. Further,

$$\mathbb{P}(\mathcal{E}_\flat) = \mathbb{P}(\mathcal{E}_\flat|\mathcal{E}_\sharp)\mathbb{P}(\mathcal{E}_\sharp) + \mathbb{P}(\mathcal{E}_\flat|\mathcal{E}_\sharp^c)\mathbb{P}(\mathcal{E}_\sharp^c) \geq \mathbb{P}(\mathcal{E}_\flat|\mathcal{E}_\sharp)\mathbb{P}(\mathcal{E}_\sharp). \tag{4.16}$$

Applying the union bound and DeMorgan's Law, we obtain

$$\mathbb{P}(\mathcal{E}_\sharp^c) = \mathbb{P}(\cup_{k=1}^K \mathcal{E}_k^c) \leq \sum_{k=1}^{K} \mathbb{P}(\mathcal{E}_k^c) \leq K \max_k \mathbb{P}(\mathcal{E}_k^c) = O\left(\frac{K}{\min_k |\sim^{(k)}|^{10} |\mathrm{AR}^{(k)}|^{10}} \right),$$

where the final equality follows from Lemma 4.9.3. From our condition on $K = o(\min_k |\sim^{(k)}|^{10} |\mathrm{AR}^{(k)}|^{10})$, we have that $\mathcal{E}_\sharp$ occurs w.h.p. Since both $\mathcal{E}_\sharp$ and $\mathcal{E}_\flat|\mathcal{E}_\sharp$ occur w.h.p., it follows from (4.16) that $\mathcal{E}_\flat$ then occurs w.h.p. Therefore,

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K} \langle \varepsilon^{(k)}, \Delta^{(k)} \rangle &= O_p\left(\frac{\sigma}{K} \left[\sum_{k=1}^{K} \frac{r^{(k)}\widetilde{\beta}^{(k)2}_2 \log\left(|\sim^{(k)}||\mathrm{AR}^{(k)}| \right)}{\min\{|\sim^{(k)}|, |\mathrm{AR}^{(k)}|\}} \right]^{1/2} \right), \\ &= O_p\left(\frac{\sigma}{K} \sum_{k=1}^{K} \frac{(r^{(k)})^{1/2} \widetilde{\beta}^{(k)}{}_2 \log^{1/2}(|\sim^{(k)}||\mathrm{AR}^{(k)}|)}{\min\{|\sim^{(k)}|^{1/2}, |\mathrm{AR}^{(k)}|^{1/2}\}} \right). \end{aligned} \tag{4.17}$$

Collecting terms. Incorporating (4.12), (4.14), (4.17) into (4.10), and simplifying yields

$$\widehat{A}_{ij} - A_{ij} = O_p\left(\frac{1}{K}\left\{\sum_{k=1}^{K}\frac{(r^{(k)})^{1/2}}{|\sim^{(k)}|^{1/4}} + \sum_{k=1}^{K}\frac{(r^{(k)})^{3/2}\widetilde{\beta}^{(k)}{}_1\log^{1/2}(|\sim^{(k)}||\text{AR}^{(k)}|)}{\min\{|\sim^{(k)}|^{1/2}, |\text{AR}^{(k)}|^{1/2}\}} + \left[\sum_{k=1}^{K}\widetilde{\beta}^{(k)}{}_2^2\right]^{1/2}\right\}\right).$$

This concludes the proof.

4.9.1 Proof of Lemma 4.9.1

The result is immediate from Lemma G.1 of Agarwal et al. (2021c) after adapting the notation used in Agarwal et al. (2021c) to that used in this paper. For every k, let $Y_{\text{pre},n} = q$, $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] = \mathbb{E}[\boldsymbol{S}^{(k)}]$, $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} = \boldsymbol{S}^{(k)}$, $\boldsymbol{V}_{\text{pre}} = \boldsymbol{U}^{(k)}$, $\widehat{w}^{(n,d)} = \widehat{\beta}^{(k)}$, $\widetilde{w}^{(n,d)} = \widetilde{\beta}^{(k)}$, where $(Y_{\text{pre},n}, \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}], \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}, \boldsymbol{V}_{\text{pre}}, \widehat{w}^{(n,d)}, \widetilde{w}^{(n,d)})$ are the notations used in Agarwal et al. (2021c).

4.9.2 Proof of Lemma 4.9.3

The result is immediate from Lemma F.2 of Agarwal et al. (2021c) after adapting the notation used in Agarwal et al. (2021c) to that used in this paper. For every k, let $Y_{\text{pre},n} = q$, $\mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}] = \mathbb{E}[\boldsymbol{S}^{(k)}]$, $\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}} = \boldsymbol{S}^{(k)}$, $\boldsymbol{V}_{\text{pre}} = \boldsymbol{U}^{(k)}$, $\widehat{w}^{(n,d)} = \widehat{\beta}^{(k)}$, $\widetilde{w}^{(n,d)} = \widetilde{\beta}^{(k)}$, where $(Y_{\text{pre},n}, \mathbb{E}[\boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}], \boldsymbol{Y}_{\text{pre},\mathcal{I}^{(d)}}, \boldsymbol{V}_{\text{pre}}, \widehat{w}^{(n,d)}, \widetilde{w}^{(n,d)})$ are the notations used in Agarwal et al. (2021c).

4.10 Proof of Proposition 2

First, let us consider recovery of entry $(i, j) = (1, 1)$. We let $C > 0$ denote an absolute constant. Define parameter $Q \geq 1$. Excluding row 1 and column 1, partition the remaining $(L - 1)$ rows and $(L - 1)$ columns into $(L - 1)/Q$ mutually exclusive blocks each of size $Q + 1$. In particular, $\boldsymbol{M}_\ell^{(1,1)} \in \mathbb{R}^{Q+1 \times Q+1}$ for $\ell \in [(L - 1)/Q]$ corresponds to the sub-matrix induced by selecting only rows $\{1, (\ell - 1)Q + 2, \dots, \ell Q + 1\}$ and columns $\{1, (\ell - 1)Q + 2, \dots, \ell Q + 1\}$. Let $\mathbb{1}_\ell^{(1,1)}$ be a binary r.v. which is equal to 1 if all entries in the sub-matrix $\boldsymbol{M}_\ell^{(1,1)}$ not including $(1, 1)$ are revealed (i.e., we do not condition on whether $(1, 1)$ is revealed or missing).

Define the event $\mathcal{E}_{(1,1)} := \{\mathbb{1}_{\ell}^{(1,1)} = 0 : \forall\ \ell \in [(L-1)/Q]\}$, i.e., $\mathcal{E}_{(1,1)}$ is the event that none of the $(L-1)/Q$ sub-matrices $\boldsymbol{M}_{\ell}^{(1,1)}$ are fully revealed. Note $\mathbb{1}_{\ell}^{(1,1)}$ is equal to 1 with probability $p^{Q^2+2Q} \geq p^{2Q^2} =: q$. Observe that $\mathbb{1}_{\ell}^{(1,1)}$ and $\mathbb{1}_{\ell'}^{(1,1)}$ for $\ell \neq \ell'$ are independent r.v.s. Then the probability $\mathcal{E}_{(1,1)}$ occurs is at most $(1-q)^{(L-1)/Q} \leq \exp^{-\frac{q(L-1)}{Q}}$. Note,

$$\exp^{-\frac{q(L-1)}{Q}} \leq (L-1)^{-10} \iff q \geq \frac{10\log((L-1))Q}{(L-1)} \iff p \geq C\left(\frac{\log(L)Q}{L}\right)^{\frac{1}{Q^2}}$$

To get an additive error of at most $O_p(\delta)$ for $A_{1,1}$, we require $Q = C^*\delta^{-6}$ by Corollary 4.5.1—this can be seen by noting that Q needs to equal the total number of anchor rows which is $N \times K$, where N and K are defined in Corollary 4.5.1. In summary, we have that $A_{i,j} - \widehat{A}_{i,j} = O_p(\delta)$ if $Q = C^*\delta^{-6}$ and $\mathcal{E}^c_{(1,1)}$ holds, where $\mathcal{E}^c_{(1,1)}$ occurs with probability at least $1 - (L-1)^{-10}$ if $p \geq C\left(\frac{\log(L)Q}{L}\right)^{\frac{1}{Q^2}}$.

Now we generalize to any (i,j) pair. Define $\mathcal{E}_{(i,j)}$ analogously to $\mathcal{E}_{(1,1)}$ The difference being that we replace the fixed row and column from $(1,1)$ to (i,j), and partition the remaining $(L-1)$ rows and $(L-1)$ columns to create the matrices $\boldsymbol{M}_{\ell}^{(i,j)}$ for $\ell \in [(L-1)/Q]$. $\mathbb{1}_{\ell}^{(i,j)}$ is then defined with respect to $\boldsymbol{M}_{\ell}^{(i,j)}$, analogous to the way $\mathbb{1}_{\ell}^{(1,1)}$ is defined with respect to $\boldsymbol{M}_{\ell}^{(1,1)}$. To ensure that $A_{i,j} - \widehat{A}_{i,j} = O(\delta)$ uniformly for all (i,j), we then require the event $\bigcap_{(i,j)\in[L]\times[L]} \mathcal{E}^c_{(i,j)}$ to hold with $Q = C^*\delta^{-6}$ as before. Appealing to the definition of p in statement of Proposition 2, this occurs with probability,

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{(i,j)\in[L]\times[L]} \mathcal{E}^c_{(i,j)}\Big) &= 1 - \mathbb{P}\Big(\bigcup_{(i,j)\in[L]\times[L]} \mathcal{E}_{(i,j)}\Big) \\ &\geq 1 - \sum_{(i,j)\in[L]\times[L]} \mathbb{P}(\mathcal{E}_{(i,j)}) \\ &\geq 1 - \frac{C}{L^8}.\end{aligned}$$

This completes the proof.

■ 4.11 Proof of Theorem 4.5.2

For ease of notation, we suppress the conditioning on $\mathcal{E}$ for the remainder of the proof. To begin, we scale the left-hand side (LHS) of (4.10) by

$$\frac{K}{\sqrt{\sum_{k=1}^{K}\left(\tilde{\sigma}^{(k)}\right)^2}} \tag{4.18}$$

and analyze each of the three resulting terms on the right-hand side (RHS) separately.

Bounding term 1. To address the first term, we scale (4.12) by (4.18) and recall our assumption on K given by (4.5). We then obtain

$$\frac{1}{\sqrt{\sum_{k=1}^{K}\left(\tilde{\sigma}^{(k)}\right)^2}}\sum_{k=1}^{K}\langle\mathbb{E}[x^{(k)}],\mathcal{P}_{U^{(k)}}\Delta^{(k)}\rangle = o_p(1). \tag{4.19}$$

Bounding term 2. Since $\langle\varepsilon^{(k)},\widetilde{\beta}^{(k)}\rangle$ are independent across k, the Lindeberg-Lévy Central Limit Theorem and (4.13) yields

$$\frac{\sum_{k=1}^{K}\langle\varepsilon^{(k)},\widetilde{\beta}^{(k)}\rangle}{\sqrt{\sum_{k=1}^{K}\left(\tilde{\sigma}^{(k)}\right)^2}} \xrightarrow{d} \mathcal{N}(0,1). \tag{4.20}$$

Bounding term 3. Next, we scale (4.17) by (4.18) and recall our assumption on K. This yields

$$\frac{1}{\sqrt{\sum_{k=1}^{K}\left(\tilde{\sigma}^{(k)}\right)^2}}\sum_{k=1}^{K}\langle\varepsilon^{(k)},\Delta^{(k)}\rangle = o_p(1). \tag{4.21}$$

Collecting terms. From (4.19), (4.20), and (4.21), we conclude

$$\frac{K(\widehat{A}_{ij}-A_{ij})}{\sqrt{\sum_{k=1}^{K}\left(\tilde{\sigma}^{(k)}\right)^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

Chapter 5

On Multivariate Singular Spectrum Analysis

■ 5.1 Introduction

Multivariate time series data is of great interest across many application areas, including cyber-physical systems, finance, retail, healthcare to name a few. An important goal across these domains can be summarized as accurate imputation and forecasting of a multivariate time series in the presence of noisy and/or missing data.

Setup. We consider a discrete time setting with time indexed as $t \in \mathbb{Z}$. For $N \in \mathbb{N}$, let the collection $f_n : \mathbb{Z} \to \mathbb{R}, \ n \in [N] := \{1, \dots, N\}$ be the latent time series of interest. For $t \in [T]$ and $n \in [N]$, we observe $X_n(t)$ where for $\rho \in (0, 1]$,

$$X_n(t) = \begin{cases} f_n(t) + \eta_n(t) & \text{with probability } \rho \\ \star & \text{with probability } 1 - \rho. \end{cases} \tag{5.1}$$

Here $\star$ represents a missing observation and $\eta_n(t)$ represents the per-step noise, which we assume to be an independent (across t, n) mean-zero random variable. Though $\eta_n(t)$ is independent, we note that the underlying time series, $f_n(\cdot)$, is of course strongly dependent across t, n. Indeed the presence of per-step noise $\eta_n(t)$ and missing values (denoted by $\star$) represent an additional challenge of measurement error in our setup. The generic spatio-temporal factor model for $f_n(\cdot), n \in [N]$ described in Section 5.3 *without* additional noise $\eta_n(\cdot)$ or missingness already provides an expressive model for a time series including any finite sum of products of harmonics and polynomials, any differentiable periodic function, and any Hölder continuous function.

Goal. Our objective is two-folds, for $n \in [N]$: (i) imputation – estimating $f_n(t)$ for all

$t \in [T]$; (ii) out-of-sample forecasting – predicting $f_n(t)$ for $t > T$.

■ 5.1.1 Multivariate Singular Spectrum Analysis

Multivariate singular spectrum analysis (mSSA) is a known method to impute and forecast a multivariate time series (see Broomhead and King (1986); Plaut and Vautard (1994); Ghil et al. (2002); Oropeza and Sacchi (2011); Hassani and Mahmoudvand (2013); Hassani et al. (2013); Bógalo et al. (2020)). mSSA has been used for both imputation and forecasting, and signal extraction—decomposing a time series into a small number of simpler time series (e.g., periodic, trend, autoregressive component). However, despite its heavy use in practice, the theoretical properties of mSSA are not well understood. Hence, we introduce a variant of mSSA for which we provide a rigorous finite-sample analysis of its imputation and out-of-sample forecasting properties; such a finite-sample analysis of mSSA has been missing from the literature. We note that we do not focus on the task of signal extraction which we leave as important future work. The variant of mSSA we introduce is arguably much simpler to implement than the original mSSA method and we begin by describing it in detail below. In Section 5.2, we compare the original mSSA method with this variant and discuss key differences. See Figure 5.1 for a visual depiction of the key steps in this variant of mSSA.

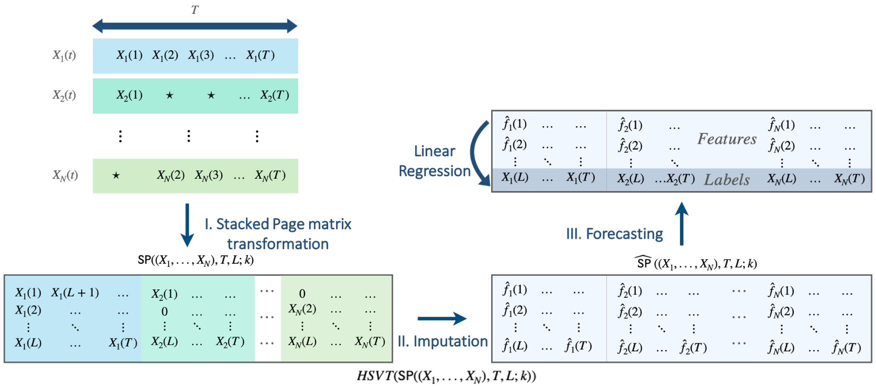

T
$X_1(t)$
$X_1(1)$ $X_1(2)$ $X_1(3)$ … $X_1(T)$
$X_2(t)$
$X_2(1)$ ⋆ ⋆ … $X_2(T)$
$X_N(t)$
⋆ $X_N(2)$ $X_N(3)$ … $X_N(T)$
I. Stacked Page matrix transformation
$\mathrm{SP}((X_1,\dots,X_N),T,L;k)$
II. Imputation
$HSVT(\mathrm{SP}((X_1,\dots,X_N),T,L;k))$
$\widehat{\mathrm{SP}}((X_1,\dots,X_N),T,L;k)$
III. Forecasting
Linear Regression
Features
Labels

Figure 5.1: Key steps of our proposed variant of the mSSA algorithm.

Singular spectrum analysis (SSA). For ease of exposition and to build intuition, we start with $N = 1$, i.e. a univariate time series. There are two algorithmic parameters: $1 \leq L \leq T$ and $k \geq 1$. For simplicity and without loss of generality assume that T is an integer multiple of L, i.e. $T/L \in \mathbb{N}$ and $k \leq \min(L, T/L)$. When $T/L \notin \mathbb{N}$, by applying both the imputation and forecasting algorithms for two ranges, $1, \dots, \lfloor T/L \rfloor \times L$

and $(T \mod L) + 1, \ldots, T$, this condition will be satisfied in each range and will provide imputation and forecasting for all T. Here $\lfloor T/L \rfloor$ refers to the floor of T/L. We give guidance on how to pick L and k when we discuss our theoretical results.

First, transform the time series $X_1(t), \; t \in [T]$ into an $L \times T/L$ matrix where the entry of the matrix in row $i \in [L]$ and column $j \in [T/L]$ is $X_1(i + (j-1) \text{x} L)$. This matrix induced by the time series is called the Page matrix, and we denote it as $\mathrm{P}(X_1, T, L)$.

Imputation. After replacing missing values (i.e. $\star$) in the matrix $\mathrm{P}(X_1, T, L)$ by 0, we compute its singular value decomposition, which we denote as

$$\mathrm{P}(X_1, T, L) = \sum_{\ell=1}^{\min(L, T/L)} s_\ell u_\ell v_\ell^T,$$

where $s_1 \geq s_2 \ldots \geq s_{\min(L,T/L)} \geq 0$ denote its ordered singular values, and $u_\ell \in \mathbb{R}^L, v_\ell \in \mathbb{R}^{T/L}$ denote its left and right singular vectors, respectively, for $\ell \in [\min(L, T/L)]$. Let $\hat{\rho}_1$ be the fraction of observed entries of X_1, precisely defined as $(\max(1, \sum_{t=1}^T \mathbf{1}(X_1(t) \neq \star)))/T$. Let the normalized, truncated version of $\mathrm{P}(X_1, T, L)$ be

$$\widehat{\mathrm{P}}(X_1, T, L; k) = \frac{1}{\hat{\rho}_1} \sum_{\ell=1}^{k} s_\ell u_\ell v_\ell^T, \tag{5.2}$$

i.e., we perform Hard Singular Value Thresholding (HSVT) on $\mathrm{P}(X_1, T, L)$ to obtain $\widehat{\mathrm{P}}(X_1, T, L; k)$. We then define the *de-noised and imputed estimate* of the original time series, denoted by $\widehat{f}_1$, as follows: for $t \in [T]$, $\widehat{f}_1(t)$ equals the entry of $\widehat{\mathrm{P}}(X_1, T, L; k)$ in row $(t - 1 \mod L) + 1$ and column $\lceil t/L \rceil$. Here $\lceil t/L \rceil$ refers to the ceiling of t/L.

Forecasting. To forecast, we learn a linear model $\hat{\beta}(X_1, T, L; k) \in \mathbb{R}^{L-1}$, which is the solution to

$$\text{minimize} \quad \sum_{m=1}^{T/L} (y_m - \beta^T x_m)^2 \quad \text{over} \quad \beta \in \mathbb{R}^{L-1},$$

where $y_m = (1/\hat{\rho}_1) X_1(L \text{x} m)$, $x_m = [\widehat{f}_1(L \text{x} (m-1)+1) \ldots \widehat{f}_1(L \text{x} (m-1)+L-1)]$ for $m \in [T/L]$. [1] Note to define y_m we impute missing values in X_1 by 0. We now describe how to use

[1] To establish theoretical results for the forecasting algorithm, we produce estimates $(\widehat{f}_1(L \text{x} (m-1) + 1) \ldots \widehat{f}_1(L \text{x} (m-1) + L - 1))$ for $m \in [T/L]$ by applying the imputation algorithm on $\mathrm{P}(X_1, T, L)$ *after* setting its Lth row equal to 0. Also, $\hat{\rho}_1$ in the definition of y_m is computed using only the first $L-1$ rows of $\mathrm{P}(X_1, T, L)$.

$\hat{\beta}(X_1, T, L; k)$ to produce both in-sample and out-of-sample forecasts. (i) In-sample forecast: for time $t = L \text{x}\, m$ and $m \in [T/L]$, the forecast is given by $\bar{f}_1(L \text{x}\, m) = \hat{\beta}(X_1, T, L; k)^T x_m$. (ii) Out-of-sample forecast: for $m > T/L$, i.e., for time $t > T$, the forecast is given by $\bar{f}_1(L \text{x}\, m) = \hat{\beta}(X_1, T, L; k)^T x'_m$ where $x'_m = \frac{1}{\hat{\rho}_1}[X_1(L \text{x}\,(m-1)+1) \dots X_1(L \text{x}\,(m-1)+L-1)]$ after imputing missing values in X_1 by 0.

Multivariate singular spectrum analysis (mSSA). Below we describe the variant of mSSA we propose, which is an extension of the SSA algorithm described above, to when we have a multivariate time series, i.e., $N > 1$. The key change is in the first step where we construct the Page matrix—instead of considering the Page matrix of a single time series, we now consider a 'stacked' Page matrix, which is obtained by a column-wise concatenation of the Page matrices induced by each time series separately. Specifically, like SSA, it has two algorithmic parameters, $L \geq 1$ and $k \geq 1$. For each time series, $n \in [N]$, create its $L \times T/L$ Page matrix $\text{P}(X_n, T, L)$, where the entry in row $i \in [L]$ and column $j \in [T/L]$ is $X_n(i + (j-1) \text{x}\, L)$. We then create a stacked Page matrix from these N time series by performing a column wise concatenation of the N matrices, $\text{P}(X_n, T, L)$, $n \in [N]$. We denote this matrix as $\text{SP}((X_1, \dots, X_N), T, L)$, and note that it has L rows and $N \text{x}\, T/L$ columns.

Imputation. We replace missing values (i.e. $\star$s) in $\text{SP}((X_1, \dots, X_N), T, L)$ by 0. Similar to (5.2), we perform HSVT on $\text{SP}((X_1, \dots, X_N), T, L)$ and denote its normalized, truncated version as $\widehat{\text{SP}}((X_1, \dots, X_N), T, L; k)$ (instead of $\hat{\rho}_1$, we now normalize by $\hat{\rho} := (\max(1, \sum_{n=1}^{N} \sum_{t=1}^{T} \mathbf{1}(X_n(t) \neq \star)))/NT)$. From $\widehat{\text{SP}}((X_1, \dots, X_N), T, L; k)$, like in SSA, we can *read off* $\hat{f}_n(t)$ for $n \in [N]$, $t \in [T]$, the *de-noised and imputed estimate* of the N time series over T time steps. In particular, let $\widehat{\text{P}}(X_n, T, L; k)$ refer to sub-matrix of $\widehat{\text{SP}}((X_1, \dots, X_N), T, L; k)$ induced by selecting only its $[(n-1) \text{x}\,(T/L) + 1, \dots, n \text{x}\, T/L]$ columns. Then for $t \in [T]$, $\hat{f}_n(t)$ equals the entry of $\widehat{\text{P}}(X_n, T, L; k)$ in row $(t-1 \mod L)+1$ and column $\lceil t/L \rceil$.

Forecasting. Similar to SSA, to forecast, we learn a linear model $\hat{\beta}((X_1, \dots, X_N), T, L; k) \in \mathbb{R}^{L-1}$, which is the solution to

$$\text{minimize} \quad \sum_{m=1}^{N \text{x}\, T/L} (y_m - \beta^T x_m)^2 \quad \text{over} \quad \beta \in \mathbb{R}^{L-1}, \tag{5.3}$$

This avoids dependencies in the noise between y_m and x_m for $m \in [T/L]$.

where y_m is the mth component of $(1/\hat{\rho})[X_1(L),\ X_1(2 \times L), \ldots, X_1(T),\ X_2(L), \ldots, X_2(T), \ldots, X_N(T)] \in \mathbb{R}^{N \times T/L}$, and $x_m \in \mathbb{R}^{L-1}$ corresponds to the vector formed by the entries of the first $L-1$ rows in the mth column of $\widehat{\mathrm{SP}}((X_1, \ldots, X_N), T, L; k)$[2] for $m \in [N \times T/L]$. Note, to define y_m, we impute missing values in $X_1, \ldots, X_n$ by 0. (i) In-sample forecast: for time step $t = L \times m'$ for $m' \in [T/L]$ and for time series $n \in [N]$, the forecast is given by $\bar{f}_n(L \times m') = \hat{\beta}((X_1, \ldots, X_N), T, L; k)^T x_m$ where $m = m' + (n-1) \times T/L$. (ii) Out-of-sample forecast: for $m' > T/L$, i.e., for time $t > T$, and for time series $n \in [N]$, the forecast is given by $\bar{f}_n(L \times m') = \hat{\beta}((X_1, \ldots, X_N), T, L; k)^T x'_{m'}$, where $x'_{m'} = \frac{1}{\hat{\rho}}[X_n(L \times m' - (L-1)) \ldots X_n(L \times m' - 1)]$ after imputing missing values in X_n by 0. See Figure 5.1 for a visual depiction of the key steps above.

Page vs. Hankel mSSA. See Appendix 5.9 for a detailed discussion of the various benefits and drawbacks of the using the Page matrix representation as we propose in our variant, instead of the Hankel representation used in the original mSSA.

Empirical performance of mSSA. This variant of mSSA we propose is fully described above, with its two major steps consisting of simply singular value thresholding and ordinary least squares. A key question is how well does it perform empirically? In Table 5.1, we provide a summary comparison of mSSA's performance for imputation and forecasting on benchmark datasets with respect to state-of-the-art time series algorithms. We find that by using the stacked Page matrix in mSSA, it greatly improves performance over SSA; indicating that mSSA is effectively utilizing information *across* multiple time series. Surprisingly, our variant of mSSA performs competitively or outperforms popular neural network based methods, such as LSTM and DeepAR—we note that these state-of-the-art neural network based methods have no associated theoretical analysis. Indeed, apart from its use in practice, the empirical performance of (our variant of) mSSA strongly motivates a theoretical analysis of when and why mSSA works.

■ 5.1.2 Our Contributions

As our primary contribution, we provide an answer to the question posed above—under a spatio-temporal factor model that we introduce, the finite-sample analysis we carry

[2] Similar to the SSA forecasting algorithm, when creating a forecasting model in mSSA, we produce $\widehat{\mathrm{SP}}((X_1, \ldots, X_N), T, L; k)$ by first setting the Lth row of $\mathrm{SP}((X_1, \ldots, X_N), T, L; k)$ equal to zero before performing the SVD and the subsequent truncation. Also, $\hat{\rho}$ in the definition of y_m is computed only using the first $L-1$ rows of $\mathrm{SP}((X_1, \ldots, X_N), T, L; k)$.

Table 5.1: mSSA statistically outperforms SSA, other state-of-the-art algorithms, including LSTMs and DeepAR across many datasets. We use the average normalized root mean squared error (NRMSE) as our metric. Details of experiments run to produce results can are in Section 5.6.

	Mean Imputation (NRMSE)					Mean Forecasting (NRMSE)				
	Electricity	Traffic	Synthetic	Financial	M5	Electricity	Traffic	Synthetic	Financial	M5
mSSA	**0.398**	0.508	**0.416**	**0.238**	**0.883**	0.485	0.536	**0.281**	**0.251**	**1.021**
SSA	0.514	0.713	0.675	0.467	0.958	0.632	0.696	0.665	0.303	1.068
LSTM	NA	NA	NA	NA	NA	0.558	0.478	0.559	1.205	1.034
DeepAR	NA	NA	NA	NA	NA	**0.479**	**0.464**	0.415	0.316	1.050
TRMF	0.641	**0.460**	0.564	0.430	0.916	0.495	0.508	0.422	0.291	1.032
Prophet	NA	NA	NA	NA	NA	0.569	0.614	1.010	1.286	1.100

out of mSSA's estimation error for imputation and out-of-sample forecasting establishes consistency, as well as its ability to effectively utilize both the spatial and temporal structure in a multivariate time series. Below, we detail the various aspects of our contribution with respect to the: (a) spatio-temporal factor model; (b) finite sample analysis of mSSA; (c) algorithmic extensions (and associated theoretical analysis) of mSSA to do time-varying variance estimation, and a tensor variant of mSSA which we show has a better imputation error convergence rate compared to mSSA for certain relative scalings of N and T.

Spatio-temporal factor model. Note that the collection of latent multivariate time series $f_n(t)$, for $n \in [N]$, $t \in [T]$ can be collectively viewed as a $N \times T$ matrix. To capture the spatial structure, i.e. the relationship across rows, we model this matrix to be low-rank—there exists a low-dimensional latent factor (or feature) associated with each of N time series; analogously, there exists a low-dimensional latent factor associated with each of the T time steps. To capture the temporal structure, we further assume that each component of the latent temporal factor has an *approximately low-rank Hankel matrix* representation (see Definition 5.3.1 for the Hankel matrix induced by a time series), i.e., the Hankel—and therefore Page—matrix induced by each component of the latent temporal factor is approximately low-rank. This additional structure imposed on the temporal factors is what motivates using the stacked Page matrix representation in mSSA, which is of dimension $L \times (N \text{x} T/L)$, where L is a hyper-parameter. We note that for $N = 1$ this subsumes the model considered to explain the success of SSA in Agarwal et al. (2018) as a special case.

As stated earlier, our model is expressive in that it includes any finite sum of products of harmonics and polynomials, any differentiable periodic function, and any Hölder continuous function. Further, we establish that the set of time series that have an approximately low-rank Hankel representation is closed under component-wise addition and multiplication. Such a model *calculus* helps characterize the representational strength of the spatio-temporal factor model we introduce.

Finite sample analysis of mSSA. Under the spatio-temporal factor model, we establish that mean squared imputation error scales as $1/\sqrt{\min(N,T)T}$ (see Theorem 5.4.1) and the out-of-sample forecasting error scales as $\max(1/\sqrt{NT}, \; N/T^2)$ (see Theorem 5.4.3, and Corollary 5.4.1). When $N < T$, the error rate is $1/\sqrt{NT}$. When $N > T$, one can simply divide the various time series into sets of size $O(T)$; this will result in a mean squared error rate of $1/T$. Hence, effectively the error is of order $1/\sqrt{\min(N,T)T}$. For exact details on the relative scaling of N and T, please refer to Theorem 5.4.3. For $N = 1$, it implies that the SSA algorithm described above has imputation and forecasting error scaling as $1/\sqrt{T}$. That is, mSSA improves performance by a $\sqrt{N}$ factor over SSA by utilizing information across the N time series. This also improves upon the prior work of Agarwal et al. (2018) which established the weaker result that SSA has imputation error scaling as $1/T^{\frac{1}{4}}$ (i.e., when $N = 1$). Further Agarwal et al. (2018) *does not* establish a result for the out-of-sample forecasting error of SSA. We note that the asymmetry in our finite-sample analysis between N and T is to be expected as we impose further structure on the latent temporal factors; they satisfy a low-rank Hankel representation, which is not assumed of the spatial factors.

Further, existing matrix estimation based methods applied to the $N \times T$ matrix of time series observations (i.e, without first performing the Page matrix transformation as done in mSSA) establish that the imputation prediction error scales as $1/\min(N,T)$. This is indeed the primary result of the works Yu et al. (2016); Rao et al. (2015), as seen in Theorem 2 of Rao et al. (2015). [3] That is, while the algorithm stated in Yu et al. (2016); Rao et al. (2015) utilizes the temporal structure in addition to the spatial structure, the theoretical guarantees do not reflect it—the guarantees provided by such methods are weaker (since $1/\min(N,T) \geq 1/\sqrt{\min(N,T)T}$) than that obtained by mSSA. Again, we emphasize that the existing analysis of SSA and matrix estimation based methods (for example Agarwal et al. (2018); Yu et al. (2016); Rao et al. (2015)) *do not establish (finite-sample) bounds for out-of-sample forecasting error.*

[3] There seems to be a typo in Corollary 2 of Yu et al. (2016) in applying Theorem 2: square in Frobenius-norm error is missing.

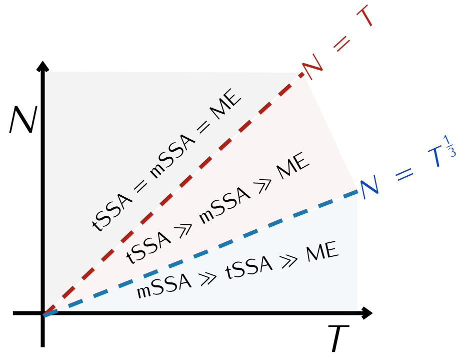


Figure 5.2: Relative effectiveness of tSSA, mSSA, ME for varying N, T.

Algorithmic extensions: variance and tensor SSA (tSSA). First, we extend mSSA to estimate the latent time-varying variance, i.e. $\mathbb{E}[\eta_n^2(t)]$, $n \in [N]$, $t \in [T]$. We establish the efficacy of such an extension when the time-varying variance is also modeled through a spatio-temporal factor model. To the best of our knowledge, this is the first result that provides provable finite-sample performance guarantees for estimating the time-varying variance of a time series. Second, we propose a novel tensor variant of SSA, termed tSSA, which exploits recent developments in the tensor estimation literature. In tSSA, rather than doing a column-wise stacking of the Page Matrices induced by each of the N time series to form a larger matrix, we instead view each Page matrix as a slice of a $L \times T/L \times N$ order-three tensor. In other words, the entry of the tensor with indices $i \in [L], j \in [T/L]$ and $n \in [N]$ equals the entry of $\mathrm{P}(X_n, L, T)$ with indices i, j. In Proposition 13, with respect to imputation error, we characterize the relative performance of tSSA, mSSA, and "vanilla" matrix estimation (ME). We find that when $N = o(T^{1/3})$, mSSA outperforms tSSA; when $T^{1/3} = o(N)$, $N = o(T)$ tSSA outperforms mSSA; when $T = o(N)$, standard matrix estimation methods are equally as effective as mSSA and tSSA. See Figure 5.2 for a graphical depiction;

Summary of contributions. We now briefly summarize our contributions:

(1) A novel spatio-temporal factor model to analyze mSSA. We show that a large family of time series dynamics fall within our factor model.

(2) Finite-sample analysis for imputation and out-of-sample forecasting. The tools we use for imputation borrow from the existing literature on matrix estimation. However, our out-of-sample forecasting requires making novel technical contributions. We believe these

tools might be of interest for online learning with a spatio-temporal factor model.

(3) A novel time-varying variance estimation algorithm with theoretical guarantees. To the best of our knowledge, neither such an algorithm nor an associated theoretical analysis exists.

(4) A novel tensor variant of the mSSA algorithm called tSSA, which exploits recent developments in the tensor estimation literature. We find that when N is large compared to T, tSSA has better sample complexity compared to mSSA. We believe this tensor variant opens a direction to future work to understand the appropriate statistical and computational trade offs for time series analysis.

■ 5.2 Literature Review

Given the ubiquity of multivariate time series analysis, it will not be possible to do justice to the entire literature. We focus on a few techniques most relevant to compare against, either theoretically or empirically.

SSA and mSSA. A good overview of the literature on SSA can be found in Golyandina et al. (2001). As alluded to earlier, the original SSA method differs from the variant discussed in Agarwal et al. (2018) and in this work. The key steps of the original SSA method are: Step 1–create a Hankel matrix from the time series data; Step 2–do a Singular Value Decomposition (SVD) of it; Step 3–group the singular values based on user belief of the model that generated the process; Step 4–perform diagonal averaging to "Hankelize" the grouped rank-1 matrices outputted from the SVD to create a set of time series; and Step 5–learn a linear model for each "Hankelized" time series for the purpose of forecasting. The theoretical analysis of this original SSA method has been focused on proving that many univariate time series have a low-rank Hankel representation, and secondly on defining sufficient *asymptotic* conditions for when the singular values of the various time series components are separable, thereby justifying Step 3 of the method. Step 3 of the original SSA method requires user input and Steps 4 and 5 are not robust to noise and missing values due to the strong dependence across entries of the Hankel representation of the time series. To overcome these limitations, in Agarwal et al. (2018) a simpler and practically useful version as described in Section 5.1.1 was introduced. As discussed earlier, this work improves upon the analysis of Agarwal et al. (2018) by providing stronger bounds for imputation prediction error, and gives new bounds for forecasting prediction error, which were missing in Agarwal et al. (2018). The original mSSA method, like the original SSA method, involves the five steps described above, but first the Hankel matrices induced by each of the N time series are stacked either column-wise (horizontal mSSA) or row-wise (vertical mSSA); see Hassani and Mahmoudvand (2018).

We note given the popularity of mSSA, there are many algorithmic variants of it proposed in the literature motivated by different applications: see Broomhead and King (1986); Plaut and Vautard (1994); Ghil et al. (2002); Oropeza and Sacchi (2011); Hassani and Mahmoudvand (2013); Hassani et al. (2013); Bógalo et al. (2020). A significant focus of these works is signal extraction, i.e., decomposing the observed time series into a small number of simpler time series (e.g., periodic, trend, autoregressive component); these extracted signals are then subsequently utilized for imputation and forecasting as described in the preceding paragraph. As stated earlier, despite the popularity of the mSSA framework, a rigorous finite-sample analysis of its imputation and out-of-sample forecasting properties are missing in the literature; the challenge in such an analysis is exacerbated with missing data and measurement error. In this work, as described in Section 5.1.1, we introduce a simpler variant of mSSA that uses the Page instead of the Hankel matrix representation. This variant is simpler as it focuses only on the task of imputation and forecasting, and not signal extraction. We do a finite-sample analysis of our variant of mSSA and establish its consistency with respect to imputation and forecasting, which so far has been missing from the mSSA literature. In Appendix 5.9, we compare our variant to the original version of mSSA which use the Hankel matrix, both with respect to their theoretical and practical properties.

Matrix factorization based methods for multivariate time series. There is a rich line of work in econometrics and statistics on viewing multiple time series as a matrix, and where some form of matrix factorization is performed to learn the spatial and temporal factors induced by the matrix; such models have also been called dynamic factor models. Some representative papers (and by no means exhaustive) include Stock and Watson (2002); Forni et al. (2000); Hallin and Liška (2007); Doz et al. (2012); Banbura and Modugno (2014); Barigozzi and Luciani (2019). Stock and Watson (2002) consider the estimation by principal components of this $N \times T$ matrix. They use the model for signal extraction and forecasting. Also, they proposed an expectation-maximization (EM) algorithm to handle missing data and imputation. Forni et al. (2000); Hallin and Liška (2007) also estimate principal components and restrict the singular vectors to be related to the Fourier basis.

Doz et al. (2012); Barigozzi and Luciani (2019) consider maximum likelihood estimation based on Kalman filtering and also consider forecasting and signal extraction. Banbura and Modugno (2014) show how to handle missing data and imputation. Similar to the mSSA literature, the general focus of these works is first signal extraction, which can then be subsequently used for imputation and forecasting. The theoretical analysis of these methods has generally been asymptotic in nature, and has focused on recovery of the spatial and temporal factors, i.e., signal extraction. Our work complements this literature as we focus directly on finite-sample analysis for imputation and out-of-sample forecasting (without first needing to signal extraction), and establish consistency for the variant of mSSA we propose. To the best of our knowledge, finite-sample consistency results such as ours are limited in the literature.

Additionally, there is a recent line of work from the machine learning literature which also employs matrix factorization based methods (see Wilson et al. (2008); Yu et al. (2016)). Most such methods make strong prior model assumptions on the underlying time series and the algorithm changes based on the assumptions made on the time series dynamics that generated the data. Further, finite sample analysis, especially with respect to forecasting error, of such methods is usually lacking. We highlight one method, Temporal Regularized Matrix Factorization (TRMF) (see Yu et al. (2016)), which we directly compare against due to its popularity, and as it achieves state-of-the-art empirical imputation and forecasting performance. The authors in Yu et al. (2016) provide finite sample imputation analysis for an instance of the model considered in this work, but forecasting analysis is absent. As discussed earlier, they establish that imputation error scales as $1/\min(N, T)$. This is a direct consequence of the low-rank structure of the original $N \times T$ matrix. But they fail to utilize, at least in the theoretical analysis, the temporal structure. Indeed, our analysis captures such temporal structure and hence our imputation error scales as $1/\sqrt{\min(N, T)T}$ which is a stronger guarantee. For example, for $N = \Theta(1)$, their error bound remains $\Theta(1)$ for any T, suggesting that TRMF Yu et al. (2016) fails to utilize the temporal structure for better estimation, while the error for mSSA would vanish as T grows.

Other relevant literature. We take a brief note of some popular time series methods in the recent literature. In particular, recently neural network (NN) based approaches have been popular and empirically effective. Some industry standard neural network methods include LSTMs, from the Keras library (a standard NN library, see Chollet (2015)) and DeepAR (an industry leading NN library for time series analysis, see Salinas et al.

(2019)). Though they have no theoretical guarantees, which is the focus of our work, we compare with them empirically.

■ 5.3 Model

■ 5.3.1 Spatio-Temporal Factor Model

Below, we introduce the spatio-temporal factor model we use to explain the success of mSSA. In short, the model requires that the underlying latent multivariate time series satisfies Properties 5.3.1 and 5.3.2, which capture the "spatial" and "temporal" structure within it, respectively.

Spatial structure in data. Consider the matrix $\boldsymbol{M} \in \mathbb{R}^{N \times T}$, where its entry in row n and column t, $\boldsymbol{M}_{nt}$ is equal to $f_n(t)$, the value of the latent time series n at time t. We posit that the matrix $\boldsymbol{M}$ is low-rank. Precisely,

Property 5.3.1. *Let rank$(\boldsymbol{M}) = R$. That is, for any $n \in [N], t \in [T]$, $\boldsymbol{M}_{nt} = \sum_{r=1}^{R} U_{nr}\; W_{rt}$, where $|U_{nr}| \leq \Gamma_1$, $|W_{rt}| \leq \Gamma_2$ for constants $\Gamma_1, \Gamma_2 > 0$.*

Property 5.3.1 effectively captures the "spatial" structure amongst the N time series. Similar to the dynamic factor model literature, we can interpret this model as there existing R latent time series $W_{r\cdot}$ for $r \in [R]$, and each time series $f_n(\cdot)$ is a linear combination of these R time series, where the weights are given by $U_{n\cdot}$.

Temporal structure in data. To explicitly capture the temporal structure in the data, we impose additional structure on $W_{r\cdot}$. To that end, we introduce the notion of the Hankel matrix induced by a time series.

Definition 5.3.1 (Hankel Matrix). *Given a time series $g : \mathbb{Z} \to \mathbb{R}$, its Hankel matrix associated with observations over T time steps, $\{1, \dots, T\}$, is given by the matrix $H \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ with $H_{ij} = g(i + j - 1)$ for $i, j \in [\lfloor T/2 \rfloor]$.*

Now, for a given $r \in [R]$, consider the time series W_{rt} for $t \in [T]$. Let $H(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ denote its Hankel matrix restricted to $[T]$, i.e. $H(r)_{ij} = W_{r(i+j-1)}$ for $i, j \in [\lfloor T/2 \rfloor]$.

Property 5.3.2. *For each $r \in [R]$ and for any $T \geq 1$, the Hankel Matrix $H(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ associated with time series $W_{rt},\; t \in [T]$ has rank at most G.*

Property 5.3.2 captures the temporal structure within the latent factors associated with time; indeed, such a low-rank Hankel representation includes a rich family of time series dynamics as noted in Proposition 3 below.

Proposition 3 (Proposition 5.2, Agarwal et al. (2018)). *Consider a time series $f : \mathbb{Z} \to \mathbb{R}$ with its element at time t denoted as*

$$f(t) = \sum_{a=1}^{A} \exp(\alpha_a t) \cdot \cos(2\pi\omega_a t + \phi_a) \cdot P_{m_a}(t), \tag{5.4}$$

where $\alpha_a, \omega_a, \phi_a \in \mathbb{R}$ are parameters, P_{m_a} is a degree $m_a \in \mathbb{N}$ polynomial in t. Then $f(\cdot)$ satisfies Property 5.3.2. In particular, consider the Hankel matrix of f over $[T]$, denoted as $H(f) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ with $H(f)_{ij} = f(i+j-1)$ for $i, j \in [\lfloor T/2 \rfloor]$. For any T, the rank of $H(f)$ is at most $G = A(m_{\max}+1)(m_{\max}+2)$, where $m_{\max} = \max_{a \in A} m_a$.

Proposition 3 states any finite sum of (products of) harmonics, polynomials, and exponentials has a low-rank Hankel representation. Each of these functions are popular to model various aspects of a time series such as periodicity and trend. Further, we note that the spectral representation of generic stationary processes, which includes autoregressive processes, implies that *any* sample-path of a stationary process can be decomposed into a weighted sum (precisely an integral) of harmonics, where the weights in the sum are sample path dependent—see Property 4.1, Chapter 4 of Robert H. Shumway (2015). That is, a finite (weighted) sum of harmonics provides a good model representation for stationary processes with the model becoming more expressive as the number of harmonics grows. In Section 5.5, we extend this model when Property 5.3.2 is only approximately satisfied. In particular, we quantify the approximation error based on the smoothness of the underlying time series and the number of harmonics used in the summation to approximate it.

Spatio-temporal model implies stacked Page matrix is low-rank. Recall that the primary representation utilized by mSSA is the stacked Page matrix (with parameter L). Observe that the Page matrix of a univariate time series for any $L \leq \lfloor T/2 \rfloor$ is simply the sub-matrix of the associated Hankel matrix: precisely, the Page matrix can be obtained by restricting to the top L rows and columns $1, L+1, \dots$ of the Hankel matrix. Therefore, the rank of the Hankel matrix is a bound on the rank of the Page matrix. Under the spatio-temporal factor model satisfying Properties 5.3.1 and 5.3.2, we establish the following low-rank property of the Page matrix of any particular time series as well as that of the stacked

Page matrix.

Proposition 4. *Let Properties 5.3.1 and 5.3.2 hold. Then for any $L \leq \lfloor T/2 \rfloor$ with any $T \geq 1$, the rank of the Page matrix induced by the univariate time series $f_n(\cdot)$ for $n \in [N]$ is at most $R \, \mathrm{x} \, G$. Further, the rank of the stacked Page matrix induced by all N time series $f_1(\cdot), \dots, f_N(\cdot)$ is also at most $R \, \mathrm{x} \, G$.*

The proof is in Appendix 5.12 where a more general result is established in Proposition 6.

■ 5.3.2 A Diagnostic Test for the Spatio-Temporal Model

In Sections 5.4 and 5.5, under the model described above, we theoretically establish the efficacy of mSSA. Beyond this model though, our work does not provide any guarantees for mSSA. Therefore, to utilize the guarantees of this work, it would be useful to have a data-driven diagnostic test that can help identify scenarios when the model of Section 5.3 may or may not hold. We discuss one such test in this section.

In particular, Proposition 4 suggests a "data driven diagnosis test" to verify whether mSSA is likely to succeed as per the results of this work. Specifically, if the (effective) rank—defined as the minimum number of singular values capturing $> 90\%$ of its spectral energy—of the Page matrix associated with any of the univariate components $f_n(\cdot)$ and the (effective) rank of stacked Page matrix associated with the multivariate time series with N component are *very different, then mSSA may not be effective* compared to SSA, but if they are *very similar then mSSA is likely to be more effective* compared to SSA. Our finite-sample results in Sections 5.4 and 5.5 indicate that the optimal value for L is $\sqrt{\min(N, T)T}$. Thus as a further test, if the effective rank of the stacked Page matrix does not scale much slower than L for $L \sim \sqrt{\min(N, T)T}$, then SSA (and mSSA) are unlikely to be effective methods.

Table 5.2 compares the (effective) rank of the stacked Page matrices for different benchmark time series data sets. The value of T equals 3993, 26304, and 10560 for the Financial, Electricity, and Traffic datasets respectively (see Appendix 5.10 for details on the datasets). We set $L = \lfloor \sqrt{\min(N, T)T} \rfloor$ for all datasets. When $N = 1$, this corresponds to L equals 63, 162, and 102 for the Financial, Electricity, and Traffic datasets respectively. Table 5.2 shows the effective rank in each dataset as we vary N. As can be seen, for $N = 1$, the effective rank is much smaller than L (or T) suggesting that SSA is likely to be effective.

For Electricity and Financial datasets, the rank does not change by much as we increase N. However, relatively the rank does increase substantially for the Traffic dataset. This might explain why mSSA is relatively less effective for the Traffic dataset in contrast to the Financial and Electricity datasets as noted in Table 5.1.

Table 5.2: Effective rank of stacked Page matrix across benchmarks as we vary N.

Dataset	N = 1	N =10	N = 100	N = 350
Electricity	19	37	44	31
Financial	1	3	3	6
Traffic	14	32	69	116

■ 5.4 Main Results

We now provide bounds on the imputation and forecasting prediction error for mSSA under the spatio-temporal model introduced in Section 5.3. We start by defining the metric by which we measure prediction error. For imputation, we define prediction error as

$$\text{ImpErr}(N, T) = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \mathbb{E}[(f_n(t) - \hat{f}_n(t))^2]. \tag{5.5}$$

Here, the imputed estimate $\hat{f}_n(\cdot)$, $n \in [N]$ are produced by the imputation algorithm of Section 5.1.1. For forecasting, we define the in-sample prediction error as

$$\text{ForErr}(N, T, L) = \frac{L}{NT} \sum_{n=1}^{N} \sum_{m'=1}^{T/L} \mathbb{E}[(f_n(L \times m') - \bar{f}_n(L \times m'))^2]. \tag{5.6}$$

Further, let $T_1 \in \mathbb{Z}$ such that $T_1 \geq L$. Then, we define the out-of-sample prediction error as

$$\text{TestForErr}(N, T, T_1, L) = \frac{L}{NT_1} \sum_{n=1}^{N} \sum_{m'=1}^{T_1/L} \mathbb{E}[(f_n(T + L \times m') - \bar{f}_n(T + L \times m'))^2]. \tag{5.7}$$

Again, the forecasted estimate $\bar{f}_n(\cdot)$, $n \in [N]$ are produced by the forecasting algorithm of Section 5.1.1. In (5.5), (5.6), and (5.7), the expectation is with respect to the randomness in observations due to noise and missingness.

■ 5.4.1 Assumptions

To state the main results, we make the following assumptions. Recall from (5.1) that for each $n \in [N]$ and $t \in [T]$, we observe $f_n(t)+\eta_n(t)$ with probability $\rho \in (0,1]$ independently. We shall assume that noise $\eta_n(\cdot)$, $n \in [N]$ satisfy the following property.

Property 5.4.1. *For $n \in [N]$, $t \in [T]$, $\eta_n(t)$ are independent sub-gaussian random variables, with $\mathbb{E}[\eta_n(t)] = 0$ and $\eta_n(t)_{\psi_2} \leq \gamma$.*

For definition of $\cdot_{\psi_\alpha}$-norm, see Vershynin (2010), for example.

Property 5.4.2. ***(Balanced spectra)**. Denote the $L \times (NT/L)$ stacked Page matrix associated with all N time series $f_1(\cdot), \dots, f_N(\cdot)$ as $\mathrm{SP}(f) := \mathrm{SP}((f_1,\dots,f_N),T,L)$. Under the setup of Proposition 4, rank$(\mathrm{SP}(f)) = \ell \geq 1$ and $\ell \leq R \times G$. Then, for $L = \sqrt{\min(N,T)T}$, $\mathrm{SP}(f)$ is such that $\sigma_\ell(\mathrm{SP}(f)) \geq c\sqrt{NT}/\sqrt{\ell}$ for some absolute constant $c > 0$, where σ_ℓ is the ℓ-th largest singular value of $\mathrm{SP}(f)$.*

Note that if $\sigma_\ell(\mathrm{SP}(f)) = \Theta(\sigma_1(\mathrm{SP}(f)))$, then one can verify that Property 5.4.2 holds. Indeed, assuming that the non-zero singular values are 'well-balanced' is standard in the matrix/tensor estimation literature. To state our results for out-of-sample forecasting error, let $\mathrm{SP}_1(f)$ be the $L \times (NT_1/L)$ stacked Page matrix associated with all N time series $f_1(t), \dots, f_N(t)$ entries for $t \in [T+1, T+T_1]$. We assume an analogous condition on $\mathrm{SP}_1(f)$ as we do for $\mathrm{SP}(f)$.

Property 5.4.3. ***(Balanced spectra (out-of-sample))**. Under the setup of Proposition 4, we have that rank$(\mathrm{SP}_1(f)) = \ell \geq 1$ and $\ell \leq R \times G$. Then, for $L = \sqrt{\min(N,T)T}$, $\mathrm{SP}_1(f)$ is such that $\sigma_\ell(\mathrm{SP}_1(f)) \geq c\sqrt{NT_1}/\sqrt{\ell}$ for some absolute constant $c > 0$, where σ_ℓ is the ℓ-th largest singular value of $\mathrm{SP}_1(f)$.*

Again, note that if $\sigma_\ell(\mathrm{SP}_1(f)) = \Theta(\sigma_1(\mathrm{SP}_1(f)))$, then one can verify that Property 5.4.3 holds.

Lastly, we shall first impose some restrictions on the complexity of the N time series $f_1(t), \dots, f_N(t)$ for $t > T$. Let $\mathrm{SP}'(f)$ denote the $(L-1) \times (NT_1/L)$ matrix formed using the top $L-1$ rows of $\mathrm{SP}(f)$. Define $\mathrm{SP}'_1(f)$ analogously with respect to $\mathrm{SP}_1(f)$. Let $\mathrm{colspan}(\mathrm{SP}'(f))$ and $\mathrm{colspan}(\mathrm{SP}'_1(f))$ denote the subspace of $\mathbb{R}^{L-1}$ spanned by the columns of $\mathrm{SP}'(f)$ and $\mathrm{SP}'_1(f)$, respectively. We assume the following property.

Property 5.4.4. ***(Subspace inclusion).*** *colspan*$(\mathsf{SP}'_1(f)) \subseteq$ *colspan*$(\mathsf{SP}'(f))$.

Intuitively, this requires that to effectively forecast, the associated stacked Page matrix of the out-of-sample time series colspan$(\mathsf{SP}'_1(f))$ is only as "rich" as that of $\mathsf{SP}'(f)$.

Picking hyper-parameter L. The proof of Theorems 5.4.1, 5.4.2, and 5.4.3 imply the optimal choice of L is to set it to $\sqrt{\min(N,T)T}$. Intuitively, this choice of L leads to the stacked Page matrix $\mathsf{SP}(f)$ to be as square as possible, and our analysis implies that the error rate is inversely proportional to the minimum of the number of the rows and columns of $\mathsf{SP}(f)$. Hence, for the remainder of the paper, we state our results for $L = \sqrt{\min(N,T)T}$.

Picking hyper-parameter k. For our theoretical result, we assume that we pick $k = \ell$, where ℓ is the rank of $\mathsf{SP}(f)$. Empirically, we pick k to equal the "effective rank" of the observed Page matrix as defined in Section 5.3.2.

■ 5.4.2 Finite-sample Analysis for Imputation and Forecasting

Now we state the main results. In what follows, we let $C(c,\Gamma_1,\Gamma_2,\gamma)$ denote a constant thats depends only (polynomially) on model parameters $c,\Gamma_1,\Gamma_2,\gamma$. We also remind the reader that R,Γ_1,Γ_2 are defined in Property 5.3.1, G in 5.3.2, γ in Property 5.4.1 and c in Property 5.4.2.

Imputation. We begin with our imputation result.

Theorem 5.4.1 (Imputation). *Let Properties 5.3.1, 5.3.2, 5.4.1 and 5.4.2 hold. For a large enough absolute constant $C > 0$, let $\rho \geq C\frac{\log NT}{\sqrt{NT}}$. Then with hyper-parameters $L = \sqrt{\min(N,T)T}$ and $k = \ell$,*

$$\mathsf{ImpErr}(N,T) \leq C(c,\Gamma_1,\Gamma_2,\gamma)\left(\frac{R^3 G \log NT}{\rho^4\sqrt{\min(N,T)T}}\right).$$

In-sample forecasting. Recall from (5.3) that in mSSA, we learn a linear model between the last row of $\mathsf{SP}((X_1,\dots,X_N),T,L)$ and the $L-1$ rows above it (after de-noising the sub-matrix induced these $L-1$ rows via HSVT). Hence, we first establish that in the idealized scenario (no noise, no missing values), there does indeed exist a linear model between the last row and the $L-1$ rows above of $\mathsf{SP}(f)$. Let $\mathsf{SP}(f)_{L\cdot}$ denote the L-th row of $\mathsf{SP}(f)$ and recall $\mathsf{SP}'(f) \in \mathbb{R}^{(L-1)\times(NT/L)}$ denotes the sub-matrix of $\mathsf{SP}(f)$ formed by selecting top $L-1$ rows. In the proposition below, we show there exists a linear relationship between $\text{SP}(f)_{L\cdot}$ and $\text{SP}'(f)$.

Proposition 5. *Let Properties 5.3.1 and 5.3.2 hold. Then there exists $\beta^* \in \mathbb{R}^{L-1}$ such that $\text{SP}(f)_{L\cdot}^T = \text{SP}'(f)^T\beta^*$. Further, $\beta^*{}_0 \leq RG$.*

Theorem 5.4.2 (In-sample forecasting)**.** *Let the conditions of Theorem 5.4.1 hold. Then, with β^* defined in Proposition 5, we have*

$$\text{ForErr}(N,T,L) \leq C(c,\gamma,\Gamma_1,\Gamma_2)\max(1,\beta^{*2}_{1})\Big(\frac{R^3 G\log NT}{\rho^4\sqrt{\min(N,T)T}}\Big).$$

Out-of-sample forecasting.

Theorem 5.4.3 (Out-of-sample Forecasting)**.** *Let Properties 5.3.1, 5.3.2, 5.4.1, 5.4.2, 5.4.3, and 5.4.4 hold. Let the hyper-parameters $L = \sqrt{\min(N,T)T}$ and $k = \ell$. Then for a large enough absolute constant $C > 0$, let $\rho \geq C\max\left(\frac{\log NT}{\sqrt{NT}}, (\gamma + R\Gamma_1\Gamma_2)\sqrt{\frac{RG}{L}}\right)$. Then, with β^* defined in Proposition 5, we have*

$$\text{TestForErr}(N,T,T_1,L) \leq C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\beta^{*2}_{1})\left(\frac{R^9G^3\log(N\max(T,T_1))}{\rho^4\sqrt{\min(N,T)T}}\left(\max\left(1,\frac{N}{T}\right)+\frac{T}{T_1}\right)\right).$$

Corollary 5.4.1. *Let the conditions of Theorem 5.4.3 hold. Then, with $T_1 = \Theta(T)$, we have*

$$\text{TestForErr}(N,T,T_1,L) \leq C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\beta^{*2}_{1})\left(\frac{R^9G^3\log(NT)\max\left(1,\frac{N}{T}\right)}{\rho^4\sqrt{\min(N,T)T}}\right).$$

Corollary 5.4.1 implies that when $N = o(T)$, then the error scales as $\sim 1/\sqrt{NT}$. When $T = o(N)$, then one can simply divide the N time series up into sets of size T. Corollary 5.4.1 implies that this will result in error scaling as $\sim 1/T$. Thus effectively, the error rate scales as $\sim 1/\sqrt{\min(N,T)T}$.

We note that Theorems 5.4.1, 5.4.2 and Proposition 5 are special cases of Theorems 5.5.1, 5.5.2 and Proposition 11 stated in the next section, respectively. Their proofs are in Appendices 5.16, 5.17, and 5.17.1, respectively. The proof of Theorem 5.4.3 is in Appendix 5.18.

■ 5.5 Approximate Low-Rank Hankel Representation

In this section, we extend the model presented in Section 5.3 by relaxing Property 5.3.2 to only hold approximately. We establish a 'calculus' for this extended model – the set of time series functions which have this approximate low-rank Hankel representation is closed under component-wise addition and multiplication. We show important examples of time series dynamics studied in the literature have an approximate low-rank Hankel representation. Lastly, we present generalizations of Theorems 5.4.1 and 5.4.2 for this extended model.

■ 5.5.1 Approximate Low-rank Hankel Representation and Hankel Calculus

We first introduce the definition of the approximate rank of a matrix.

Definition 5.5.1 (ϵ-approximate rank). *Given $\epsilon > 0$, a matrix $M \in \mathbb{R}^{a \times b}$ is said to have ϵ-approximate rank at most $s \geq 1$ if there exists a rank s matrix $M_s \in \mathbb{R}^{a \times b}$ such that $M - M_{s\infty} < \epsilon$.*

Definition 5.5.2 ((G, ϵ)-Hankel Time Series). *For a given $\epsilon \geq 0$ and $G \geq 1$, a time series $f : \mathbb{Z} \to \mathbb{R}$ is called a (G, ϵ)-Hankel time series if for any $T \geq 1$, its Hankel matrix has ϵ-approximate rank G.*

We extend the model of Section 5.3 by replacing Property 5.3.2 by the following.

Property 5.5.1. *For each $r \in [R]$ and for any $T \geq 1$, the Hankel Matrix $H(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ associated with time series $W_{rt},\ t \in [T]$ has ϵ-approximate rank at most G for $\epsilon > 0$. That is, for each $r \in [R]$, $W_{r\cdot}$ is a (G, ϵ)-Hankel time series.*

We state an implication of the above stated properties on the stacked Page matrix.

Proposition 6. *Let Properties 5.3.1 and 5.5.1 hold. For any $L \leq \lfloor T/2 \rfloor$ with any $T \geq 1$, the stacked Page matrix induced by the N time series $f_1(\cdot), \dots, f_N(\cdot)$ has ϵ'-rank at most R x G for $\epsilon' = R\Gamma_1\epsilon$.*

Hankel calculus. We present a key property of the model class satisfying Property 5.5.1, i.e. time series that have an approximate low-rank Hankel matrix representation. To that end, we define 'addition' and 'multiplication' for time series. Given two time series $f_1,\ f_2 : \mathbb{Z} \to \mathbb{R}$, define their addition, denoted $f_1 + f_2 : \mathbb{Z} \to \mathbb{R}$ as $(f_1 + f_2)(t) = f_1(t) + f_2(t)$, for all $t \in \mathbb{Z}$. Similarly, their multiplication, denoted $f_1 \circ f_2 : \mathbb{Z} \to \mathbb{R}$ as $(f_1 \circ f_2)(t) = f_1(t) \text{x} f_2(t)$, for all $t \in \mathbb{Z}$. Now, we state a key property for the model class satisfying Property 5.5.1 (proof in Appendix 5.13).

Proposition 7. *For $i \in \{1, 2\}$, let f_i be a (G_i, ϵ_i)-Hankel time series for $G_i \geq 1,\ \epsilon_i \geq 0$. Then, $f_1 + f_2$ is a $(G_1 + G_2, \epsilon_1 + \epsilon_2)$-Hankel time series and $f_1 \circ f_2$ is a $\Big(G_1 G_2, 3\max(\epsilon_1, \epsilon_2) \cdot \max(\left\|f_1\right\|_\infty, \left\|f_2\right\|_\infty)\Big)$-Hankel time series.*

■ 5.5.2 Examples of (G, ϵ)-Hankel Time Series

We establish that many important classes of time series dynamics studied in the literature are instances of (G, ϵ)-Hankel time series, i.e. they satisfy Property 5.5.1. In particular, any differentiable periodic function (Proposition 9), and any time series with a Hölder continuous latent variable representation (Proposition 10). Proofs of Propositions 8, 9, and 10 can be found in Appendix 5.13.

Example 1. **(G, ϵ)-LRF time series.** We start by defining a linear recurrent formula (LRF), which is a standard model for linear time-invariant systems.

Definition 5.5.3 ((G, ϵ)-LRF). *For $G \in \mathbb{N}$ and $\epsilon \geq 0$, a time series f is said to be a (G, ϵ)-Linear Recurrent Formula (LRF) if for all $T \in \mathbb{Z}$ and $t \in [T]$, there exists $g : \mathbb{Z} \to \mathbb{R}$ such that*

$$f(t) = g(t) + h(t),$$

where for all $t \in \mathbb{Z}$, (i) $g(t) = \sum_{l=1}^{G} \alpha_l g(t - l)$ with constants $\alpha_1, \ldots, \alpha_G$, and (ii) $|h(t)| \leq \epsilon$.

Now we establish a time series f that is a (G, ϵ)-LRF is also (G, ϵ)-Hankel.

Proposition 8. *If f is (G, ϵ)-LRF representable, then it is (G, ϵ)-Hankel representable.*

LRF's cover a broad class of time series functions, including any finite sum of products of harmonics, polynomials and exponentials. In particular, it can be easily verified that a time series described by (5.4) is a $(G, 0)$-LRF, where $G \leq A(m_{\max} + 1)(m_{\max} + 2)$ with $m_{\max} = \max_{a \in A} m_a$.

Example 2. "smooth" and periodic time series. We establish that any differentiable periodic function is (G, ϵ)-LRF and hence (G, ϵ)-Hankel for appropriate choices of G and ϵ.

Definition 5.5.4 ($C^k(R, \text{PER})$). *For $k \geq 1$ and $R > 0$, we use $C^k(R, PER)$ to denote the class of all time series $f : \mathbb{R} \to \mathbb{R}$ such that it is R periodic, i.e. $f(t + R) = f(t)$ for all $t \in \mathbb{R}$ and the k-th derivative of f, denoted $f^{(k)}$, exists and is continuous.*

Proposition 9. *Any $f \in C^k(R, PER)$ is*

$$\left(4G, C(k, R)\frac{\left\|f^{(k)}\right\|}{G^{k-0.5}}\right) - \textit{Hankel representable,}$$

for any $G \geq 1$. Here $C(k, R)$ is a term that depends only on k, R and $\left\|f^{(k)}\right\|^2 = \frac{1}{R}\int_0^R (f^{(k)}(t))^2 dt$.

Example 3. time series with latent variable model (LVM) structure. We now show that if a time series has a LVM representation, and the latent function is Hölder continuous, then it has a (G, ϵ)-Hankel representation for appropriate choice of $G \geq 1$ and $\epsilon \geq 0$. We first define the Hölder class of functions; this class of functions is widely adopted in the non-parametric regression literature Wasserman (2006). Given a function $g : [0, 1)^K \to \mathbb{R}$, and a multi-index $\kappa \in \mathbb{N}^K$, let the partial derivate of g at $x \in [0, 1)^K$, if it exists, be denoted as $\nabla_\kappa g(x) = \frac{\partial^{|\kappa|} g(x)}{(\partial x)^\kappa}$ where $|\kappa| = \sum_{j=1}^K \kappa_j$ and $(\partial x)^\kappa = \partial^{\kappa_1} x_1 \cdots \partial^{\kappa_K} x_K$.

Definition 5.5.5 (**$(\alpha, \mathcal{L})$-Hölder Class**). *Given $\alpha, \mathcal{L} > 0$, the Hölder class $\mathcal{H}(\alpha, \mathcal{L})$ on $[0, 1)^K$ is defined as the set of functions $g : [0, 1)^K \to \mathbb{R}$ whose partial derivatives satisfy for all $x, x' \in [0, 1)^K$, $\sum_{\kappa : |\kappa| = \lfloor \alpha \rfloor} \frac{1}{\kappa!} |\nabla_\kappa g(x) - \nabla_\kappa g(x')| \leq \mathcal{L} \left\|x - x'\right\|_\infty^{\alpha - \lfloor \alpha \rfloor}$. Here $\lfloor \alpha \rfloor$ refers to the greatest integer strictly smaller than α and $\kappa! = \prod_{j=1}^K \kappa_j!$.*

Note that if $\alpha \in (0, 1]$, then the definition above is equivalent to the $(\alpha, \mathcal{L})$-Lipschitz condition, i.e., $|g(x) - g(x')| \leq \mathcal{L} \left\|x - x'\right\|_\infty^\alpha$, for $x, x' \in [0, 1)^K$. Given a time series $f : \mathbb{Z} \to \mathbb{R}$, for any $T \geq 1$, recall the Hankel matrix $H \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ is defined such that its entry in row $i \in [\lfloor T/2 \rfloor]$ and column $j \in [\lfloor T/2 \rfloor]$ is given by $H_{ij} = f(i + j - 1)$. We call a time series f to have $(\alpha, \mathcal{L})$-Hölder smooth LVM representation for $\alpha, \mathcal{L} > 0$ if for any given $T \geq 1$, the corresponding Hankel matrix H satisfies: for $i, j \in [\lfloor T/2 \rfloor]$, $H_{ij} = g(\theta_i, \omega_j)$, where $\theta_i, \omega_j \in [0, 1)^K$ are latent parameters and $g(\cdot, \omega) \in \mathcal{H}(\alpha, \mathcal{L})$ for any $\omega \in [0, 1)^K$. It can be verified that a $(G, 0)$-Hankel time series is an instance of such a LVM representation with corresponding $g(x, y) = x^T y$. Thus in a sense, this model is a natural generalization of the $(G, 0)$-Hankel matrix representation. The following proposition connects this LVM representation to the (G, ϵ)-Hankel representation for appropriately defined $G \geq 1, \epsilon > 0$.

Proposition 10. *Given $\alpha, \mathcal{L} > 0$, let f have $(\alpha, \mathcal{L})$-Hölder smooth LVM representation. Then for all $\epsilon > 0$, f is*

$$(C(\alpha, K)\left(\frac{1}{\epsilon}\right)^K, \mathcal{L}\epsilon^\alpha) - \textit{Hankel representable.}$$

Here $C(\alpha, K)$ is a term that depends only on α and K.

■ 5.5.3 Extending Main Results

Below, we provide generalizations of the imputation and in-sample forecasting results stated in Section 5.4. To do so, we utilize Property 5.5.2 which is analogous to Property 5.4.2 but for the approximate low-rank setting.

Property 5.5.2. *(Approximately balanced spectra). Under the setup of Proposition 6, we can represent the $L \times (NT/L)$ stacked Page matrix associated with all N time series $f_1(\cdot), \dots, f_N(\cdot)$ as $\mathrm{SP}(f) = \tilde{\boldsymbol{M}} + \boldsymbol{E}$ with rank$(\tilde{\boldsymbol{M}}) = \ell \geq 1$ and $\ell \leq R \text{x} G$ and $\boldsymbol{E}_\infty \leq R\Gamma_1\epsilon$. Then, for $L = \sqrt{\min(N, T)T}$, $\tilde{\boldsymbol{M}}$ is such that $\sigma_\ell(\tilde{\boldsymbol{M}}) \geq c\sqrt{NT}/\sqrt{\ell}$ for some absolute constant $c > 0$, where σ_ℓ is the ℓ-th largest singular value of $\tilde{\boldsymbol{M}}$.*

Theorem 5.5.1 (Imputation). *Let Properties 5.3.1, 5.5.1, 5.4.1 and 5.5.2 hold. For a large enough absolute constant $C > 0$, let $\rho \geq C\frac{\log NT}{\sqrt{NT}}$. Then, with hyper-parameters $L = \sqrt{\min(N, T)T}$ and $k = \ell$,*

$$\mathrm{ImpErr}(N, T) \leq C(c, \Gamma_1, \Gamma_2, \gamma)\left(\frac{R^3 G \log NT}{\rho^4\sqrt{\min(N, T)T}} + \frac{R^4 G(\epsilon + \epsilon^3)}{\rho^2}\right)$$

where $C(c, \Gamma_1, \Gamma_2, \gamma)$ is a positive constant dependent on model parameters including $\Gamma_1, \Gamma_2, \gamma$.

We remind the reader that R, Γ_1, Γ_2 are defined in Property 5.3.1, G in 5.3.2, γ in Property 5.4.1 and c in Property 5.5.2.

Existence of Linear Model. We now state Proposition 11, which is analogous to Proposition 5, but for the approximate low-rank setting.

Proposition 11. *Let Properties 5.3.1 and 5.5.1 hold. Then, there exists $\beta^* \in \mathbb{R}^{L-1}$, such that* $\mathrm{SP}(f)^T_{L\cdot} - \mathrm{SP}'(f)^T \beta^*{}_\infty \leq R\Gamma_1(1 + \beta^*{}_1)\epsilon.$*, Further* $\beta^*{}_0 \leq RG.$

Theorem 5.5.2 (In-sample forecasting). *Let the conditions of Theorem 5.5.1 hold. Then with β^* defined in Proposition 11, we have*

$$\mathrm{ForErr}(N, T, L) \leq C(c, \gamma, \Gamma_1, \Gamma_2) \max(1, \beta^{*2}_{1}) \Big(\frac{R^3 G \log NT}{\rho^4 \sqrt{\min(N, T) T}} + \frac{R^4 G(\epsilon + \epsilon^3)}{\rho^2} \Big).$$

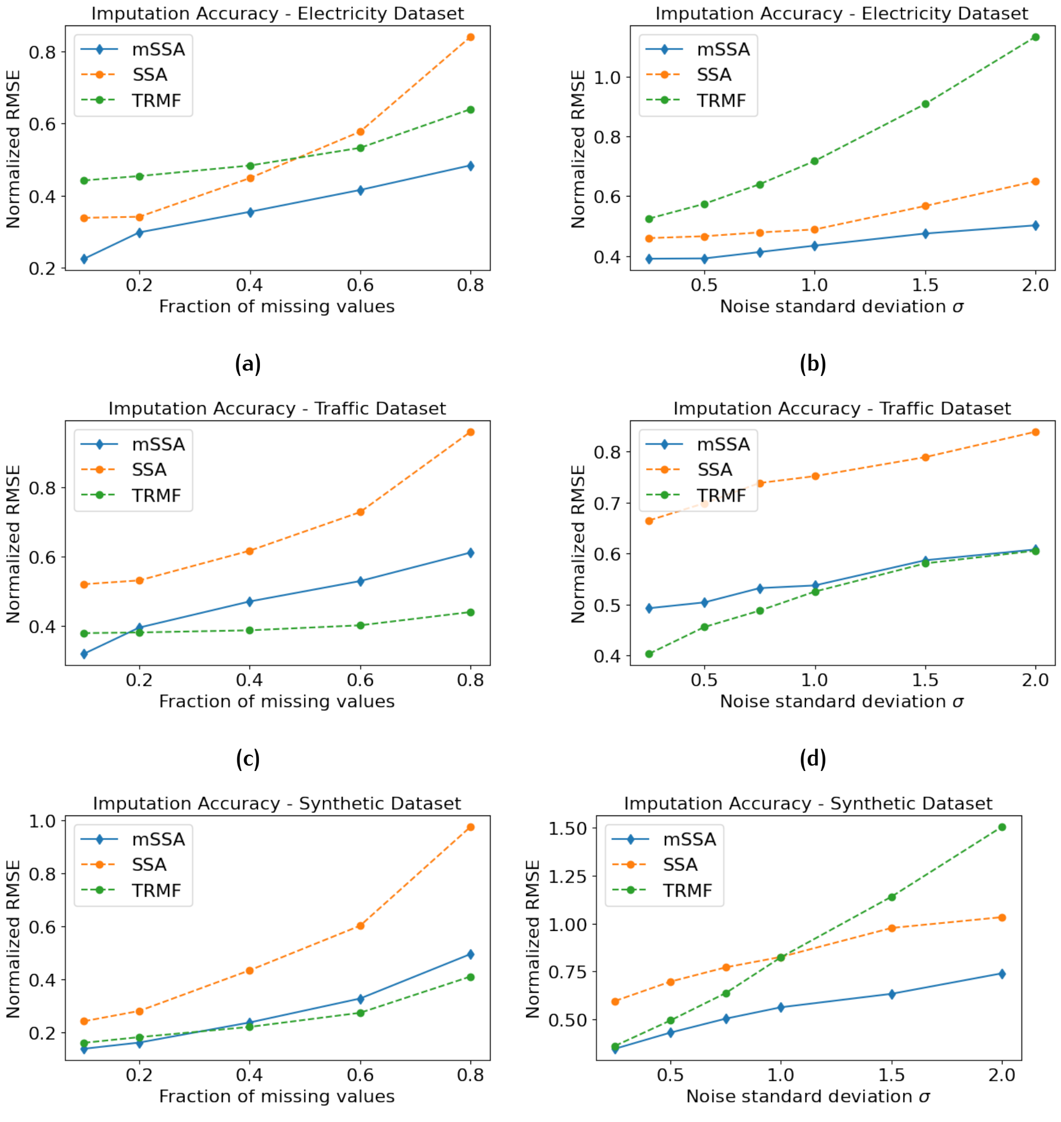
5.6 Experiments

We describe experiments supporting our theoretical results for mSSA. In particular, we provide details of the experiments run to create the summary results described earlier in Table 5.1. In Appendix 5.10, we describe the datasets utilized and the various algorithms we compare with as well as the procedure for selecting the hyper-parameters in each algorithm. In Section 5.6.1 and 5.6.2, we report the imputation and forecasting results. Note that in all experiments, we use the Normalized Root Mean Squared Error (NRMSE) as out accuracy metric. That is, we normalize all the underlying time series to have zero mean and unit variance before calculating the root mean squared error. We use this metric as it weighs the error on each time series equally.

5.6.1 Imputation

Setup. We test the robustness of the imputation performance by adding two sources of corruption to the data - varying the percentage of observed values and varying the amount of noise we perturb the observations by. We test imputation performance by how accurately we recover missing values. We compare the performance of mSSA with TRMF, a method which achieves state-of-the-art imputation performance. Further, to analyze the added benefit of exploiting the spatial structure in a multivariate time series using mSSA, we compare with the SSA variant introduced in Agarwal et al. (2018) .

Results. Figures 5.3a, 5.3c, 5.3e, 5.4a, and 5.4c show the imputation error in the aforementioned datasets as we vary the fraction of missing values, while Figures 5.3b, 5.3d, 5.3f, 5.4b, and 5.4d show the imputation error as we vary σ, the standard deviation of the gaussian noise. We see that as we vary the fraction of missing values and noise levels, mSSA outperforms both TRMF and SSA in ∼ 75% of experiments run. It is noteworthy the large empirical gain in mSSA over SSA, giving credence to the spatio-temporal model we introduce. The average NRMSE across all experiments for each dataset is reported in Table 5.1, where mSSA outperforms every other method across all datasets except for the Traffic dataset.



(a)

(b)

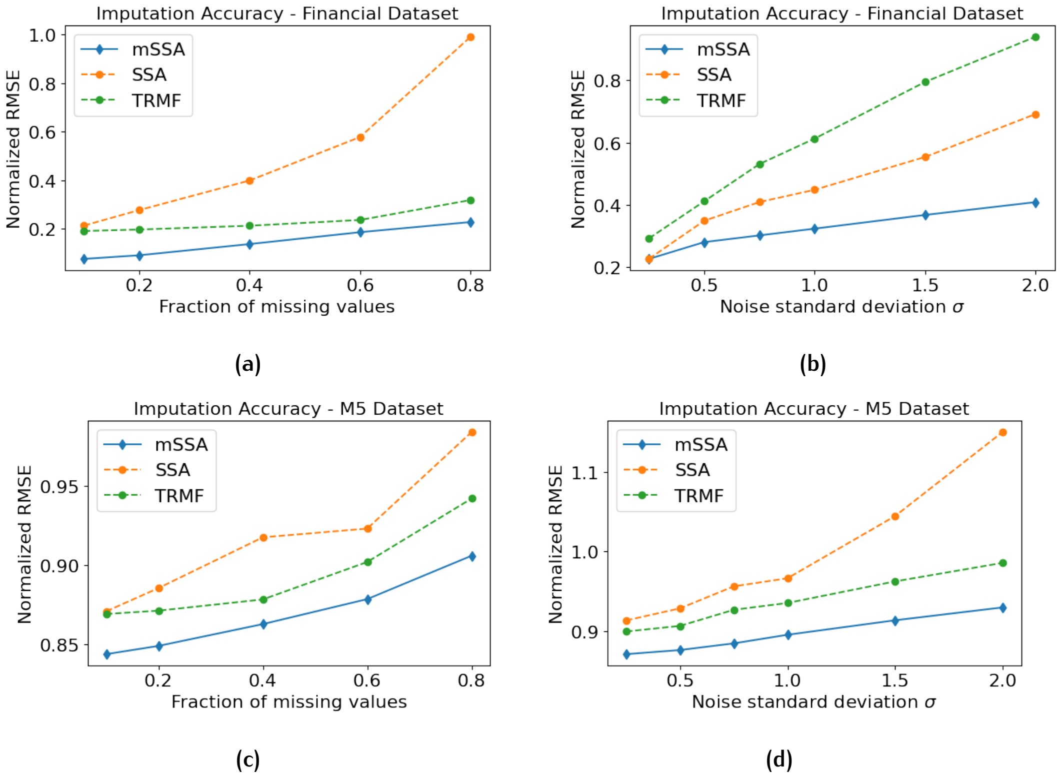
(c)

(d)

(e)

(f)

Figure 5.3: mSSA vs. TRMF vs. SSA – imputation performance on the Electricity, Traffic and Synthetic datasets. Figures 5.3a, 5.3c, and 5.3e, show imputation accuracy of mSSA, TRMF and SSA as we vary the fraction of missing values; Figures 5.3b, 5.3d, and 5.3f show imputation accuracy as we vary the noise level (and with 50% of values missing).



(a)

(b)

(c)

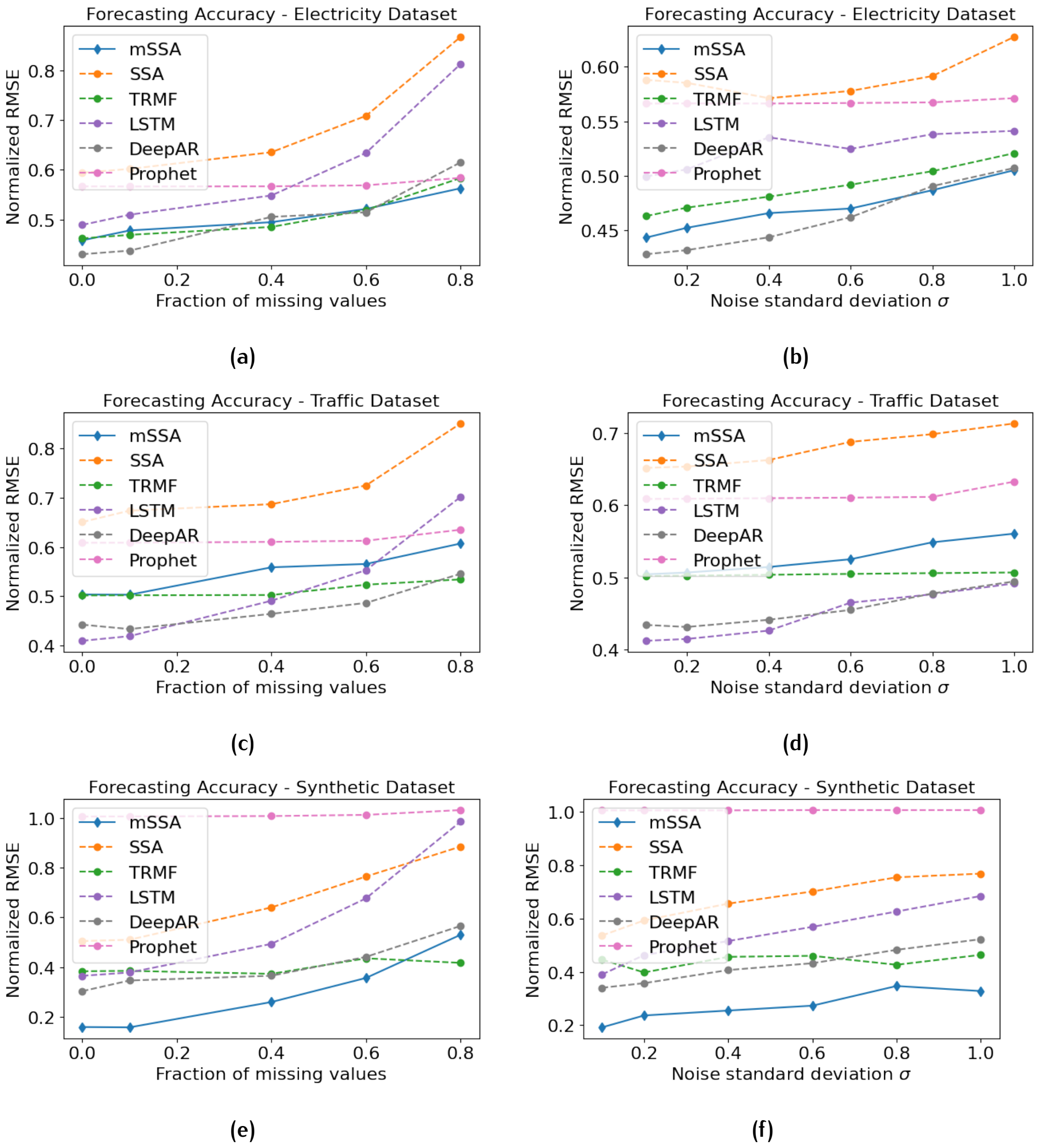
(d)

Figure 5.4: mSSA vs. TRMF vs. SSA – imputation performance on the Financial and M5 datasets. Figures 5.4a, and 5.4c show imputation accuracy of mSSA, TRMF and SSA as we vary the fraction of missing values; Figures 5.4b, and 5.4d show imputation accuracy as we vary the noise level (and with 50% of values missing).

■ 5.6.2 Forecasting

Setup. We test the forecasting accuracy of the proposed mSSA against several state-of-the-art algorithms. For each dataset, we split the data into training, validation, and testing datasets as outlined in Appendix 5.10.1. As was done in the imputation experiments, we vary how much each dataset is corrupted by varying the percentage of observed values and the noise levels.

Results. Figures 5.5a, 5.5c, 5.5e, 5.6a, and 5.6c show the forecasting accuracy of mSSA and other methods in the aforementioned datasets as we vary the fraction of missing values, while Figures 5.5b, 5.5d, 5.5f, 5.6b, and 5.6d show the forecasting accuracy as we vary the standard deviation of the added gaussian noise. We see that as we vary the fraction of missing values and noise level, mSSA is the best or comparable to the best performing method in ∼ 80% of experiments. In terms of the average NRMSE across all experiments, we find that mSSA performs similar to or better than every other method across all datasets except for the traffic dataset as was reported in Table 5.1.



(a)

(b)

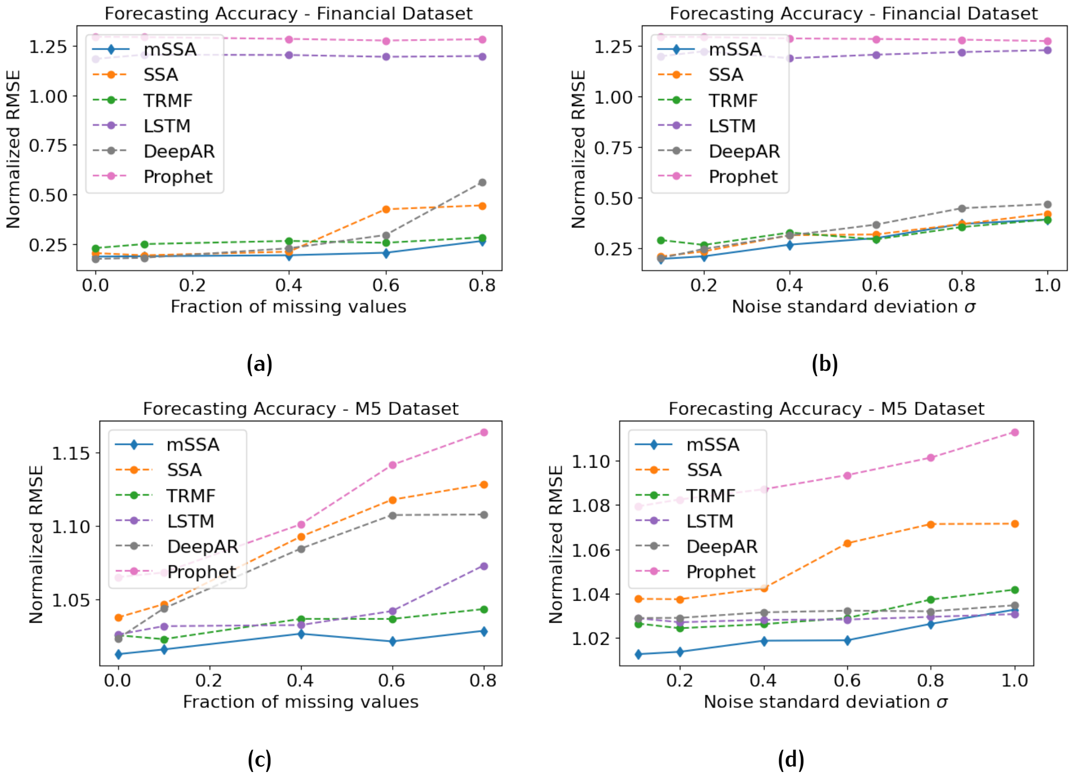
(c)

(d)

(e)

(f)

Figure 5.5: mSSA forecasting performance on standard multivariate time series benchmark is competitive with/outperforming industry standard methods as we vary the number of missing data and noise level. Figures 5.5a, 5.5c, and 5.5e show the forecasting accuracy of all methods on the Electricity, Traffic and Synthetic datasets with varying fraction of missing values; Figures 5.5b, 5.5d, and 5.5f showsthe forecasting accuracy on the same datasets with varying noise level.



(a)

(b)

(c)

(d)

Figure 5.6: Figures 5.6a, and 5.6c show the forecasting accuracy of all methods on the financial and M5 datasets with varying fraction of missing values; Figures 5.6b, and 5.6d show the forecasting accuracy on the same datasets with varying noise levels.

■ 5.7 Algorithmic Extensions of mSSA

■ 5.7.1 Variance Estimation

We extend the mSSA algorithm to estimate the time-varying variance of a time series by making the following simple observation. If we apply mSSA to the squared observations, $X_n^2(t)$, we will recover an estimate of $\mathbb{E}[X_n^2(t)]$ (for $\rho = 1$). However, observe that $\text{Var}[X_n(t)] = \mathbb{E}[X_n^2(t)] - \mathbb{E}[X_n(t)]^2$. Therefore, by applying mSSA twice, once on $X_n(t)$ and once on $X_n^2(t)$ for $n \in [N]$ and $t \in [T]$, and subsequently taking the component-wise difference of the two estimates will lead to an estimate of the variance. This suggests a simple algorithm which we describe next. We note this observation suggests *any* mean estimation algorithm (or imputation) in time series analysis can be converted to estimate the time varying variance – this ought to be of interest in its own right.

Algorithm. As described in Section 5.1.1, let $L \geq 1$ and $k, k' \geq 1$ be algorithm parameters. First, apply mSSA on observations $X_n(t),\ n \in [N],\ t \in [T]$ to produce imputed estimates $\hat{f}_n(t)$ using the hyper-parameters L and k. Next, apply mSSA on observations $X_n^2(t),\ n \in [N],\ t \in [T]$ to produce imputed estimates $\hat{g}_n(t)$ using the hyper-parameters L and k'. Lastly, we denote $\hat{\sigma}_n^2(t) = \max(0, \hat{g}_n(t) - \hat{f}_n(t)^2),\ n \in [N],\ t \in [T]$ as our estimate of the time-varying variance.

Model. For $n \in [N],\ t \in [T]$, let $\sigma_n^2(t) = \mathbb{E}[\eta_n^2(t)]$ be the time-varying variance of the time series observations, i.e., if $\rho = 1$ then $\sigma_n^2(t) = \text{Var}[X_n(t)] = \mathbb{E}[X_n^2(t)] - f_n^2(t)$. Let $\Sigma \in \mathbb{R}^{N \times T}$ be the matrix induced by the latent time-varying variances of the N time series of interest, i.e., the entry in row n at time t in $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}_{nt} = \sigma_n^2(t)$. To capture the "spatial" and "temporal" structure across the N latent time-varying variances, we assume the latent variance matrix $\boldsymbol{\Sigma}$ satisfies Properties 5.7.1 and 5.7.2. These properties are analogous to those assumed about the latent mean matrix $\boldsymbol{M}$ (defined in Section 5.3); in particular, Properties 5.3.1 and 5.3.2. We state them next.

Property 5.7.1. *Let $R' = rank(\boldsymbol{\Sigma})$, i.e, for any $n \in [N], t \in [T]$, $\boldsymbol{\Sigma}_{nt} = \sum_{r=1}^{R'} U'_{nr}\ W'_{rt}$, where the factorization is such that $|U'_{nr}| \leq \Gamma'_1$, $|W'_{rt}| \leq \Gamma'_2$ for $\Gamma'_1, \Gamma'_2 > 0$.*

Like Property 5.3.1, the above property captures the "spatial" structure within N time series of variances. To capture the "temporal" structure, next we introduce an analogue of Property 5.3.2. To that end, for each $r \in [R']$, define the $\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor$ Hankel matrix of each time series $W'_{rt},\ t \in [T]$ as $H'(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$, where $H'(r)_{ij} = W'_{r(i+j-1)}$ for $i, j \in [\lfloor T/2 \rfloor]$.

Property 5.7.2. *For each $r \in [R']$, the Hankel Matrix $H'(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ associated with time series $W'_{rt}, t \in [T]$ has rank at most G'.*

Result. To establish the estimation error for the variance estimation algorithm under the spatio-temporal model above, we need the following additional property (analogous to Property 5.4.2).

Property 5.7.3 (Balanced spectra). *Denote the $L \times (NT/L)$ stacked Page matrix associated with all N time series $\sigma_1^2(\cdot), \dots, \sigma_N^2(\cdot)$ as $\text{SP}(\sigma^2) := \text{SP}((\sigma_1^2, \dots, \sigma_N^2), T, L)$. Due to Properties 5.7.1 and 5.7.2, and a simple variant of Proposition 4, we have $rank(\text{SP}(\sigma^2)) = \ell' \geq 1$ and $\ell' \leq R' \text{x}\, G'$. Then, for $L = \sqrt{\min(N,T)T}$, $\text{SP}(\sigma^2)$ is such that $\sigma'_{\ell}(\boldsymbol{M}) \geq c\sqrt{NT}/\sqrt{\ell'}$ for some absolute constant $c > 0$, where σ'_{ℓ} is the ℓ-th singular value, order by magnitude, of $\text{SP}(\sigma^2)$.*

Theorem 5.7.1 (Variance Estimation). *Let Properties 5.3.1, 5.3.2, 5.4.1, 5.4.2, 5.7.1, 5.7.2, and 5.7.3 hold. Additionally let $|\hat{f}_n(t)| \leq \Gamma_3$ for all $n \in [N], t \in [T]$. Lastly, let hyper-parameters $L = \sqrt{\min(N, T)T}$, $k = \ell$, $k' = \ell'$. Let $\rho = 1$. Then the variance prediction error is bounded above as*

$$\frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \mathbb{E}[(\sigma_n(t)^2 - \hat{\sigma}_n^2(t))^2] \leq \tilde{C} \left(\frac{(G^2 + G') \log^2 NT}{\sqrt{\min(N, T)T}}. \right).$$

where $\tilde{C}$ is a constant dependent (polynomially) on model parameters Γ_1, Γ_2, Γ_3, Γ_1', Γ_2', γ, R, R'.

Proof of Theorem 5.7.1 can be found in Appendix 5.19.

■ 5.7.2 Tensor SSA

Page tensor. We introduce an order-three tensor representation of a multivariate time series which we term the 'Page tensor'. Given N time series, with observations over T time steps and hyper-parameter $L \geq 1$, define $\mathbf{T} \in \mathbb{R}^{N \times T/L \times L}$ such that

$$\mathbf{T}_{n\ell s} = f_n((s-1) \times L + \ell), \quad n \in [N], \ \ell \in [L], \ s \in [T/L].$$

The corresponding observation tensor, $\mathbb{T} \in (\mathbb{R} \cup \{\star\})^{N \times T/L \times L}$, is

$$\mathbb{T}_{n\ell s} = X_n((s-1) \times L + \ell), \quad n \in [N], \ \ell \in [L], \ s \in [T/L]. \tag{5.8}$$

See Figure 5.7 for a visual depiction of $\mathbb{T}$.

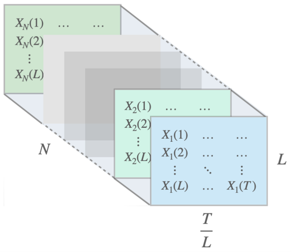


Figure 5.7: The observations Page tensor.

Let the CP-rank of an order-d tensor $\boldsymbol{T} \in \mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$ be the smallest value of $r \in \mathbb{N}$

such that $T_{i_1,\dots,i_d} = \sum_{k=1}^{r} u_{i_1,k} \dots u_{i_d,k}$, where $u_{i_\ell,\cdot}$ are latent factors for $\ell \in [d]$. Under the model described in Section 5.3, we have the following properties.

Proposition 12. *Let Properties 5.3.1, 5.3.2, and 5.4.1 hold. Then, for any $1 \leq L \leq \sqrt{T}$, $\mathbf{T}$ has canonical polyadic (CP)-rank at most R x G. Further, all entries of $\mathbb{T}$ are independent random variables with each entry observed with probability $\rho \in (0, 1]$, and $\mathbb{E}[\mathbb{T}] = \rho\mathbf{T}$.*

tSSA: time series imputation using the Page tensor representation. The Page tensor representation and Proposition 12 collectively suggest that time series imputation can be reduced to low-rank tensor estimation, i.e., recovering a tensor of low CP-rank from its noisy, partial observations. Over the past decade, the field of low-rank tensor (and matrix) estimation has received great empirical and theoretical interest, leading to a large variety of algorithms including spectral, convex optimization, and nearest neighbor based approaches. We list a few works which have explicit finite-sample rates for noisy low-rank tensor completion Barak and Moitra (2016); Xia et al. (2018); Cai et al. (2021); Yu (2020); Shah and Yu (2019)). As a result, we "blackbox" the tensor estimation algorithm used in tSSA as a pivotal subroutine. Doing so allows one the flexibility to use the tensor estimation algorithm of their choosing within tSSA. Consequently, as the tensor estimation literature continues to advance, the "meta-algorithm" of tSSA will continue to improve in parallel. To that end, we give a definition of a tensor estimation algorithm for a generic order-d tensor. Note that when $d = 2$, this reduces to standard matrix estimation (ME).

Definition 5.7.1 (Matrix/Tensor Estimation)**.** *For $d \geq 2$, denote $TE_d : \{\star, \mathbb{R}\}^{n_1 \times n_2 \times \dots n_d} \to \mathbb{R}^{n_1 \times n_2 \times \dots n_d}$ as an order-d tensor estimation algorithm. It takes as input an order-d tensor $\mathbb{G}$ with noisy, missing entries, where $\mathbb{E}[\mathbb{G}] = \rho\boldsymbol{G}$ and $\rho \in (0, 1]$ is the probability of each entry in $\mathbb{G}$ being observed. TE_d then outputs an estimate of $\boldsymbol{G}$ denoted as $\widehat{\boldsymbol{G}} = TE_d(\mathbb{G})$.*

We assume the following 'oracle' error convergence rate for TE_d; for ease of exposition, we restrict our attention to the setting where $\rho = 1$.

Property 5.7.4. *For $d \geq 2$, assume TE_d satisfies the following: the estimate $\widehat{\boldsymbol{G}} \in \mathbb{R}^{n_1 \times n_2 \times \dots n_d}$, which is the output of $TE_d(\mathbb{G})$ with $\mathbb{E}[\mathbb{G}] = \boldsymbol{G}$, satisfies*

$$\frac{1}{n_1 \dots n_d} \widehat{\boldsymbol{G}} - \boldsymbol{G}_F^2 = \tilde{\Theta}\left(1/\min(n_1, \dots, n_d)^{\lceil d/2 \rceil}\right).$$

Here, $\tilde{\Theta}(\cdot)$ suppresses dependence on noise, i.e., $\boldsymbol{E} = \mathbb{G} - \mathbb{E}[\mathbb{G}]$, $\log(\cdot)$ factors, and CP-rank of $\boldsymbol{G}$.

Property 5.7.4 holds for a variety of matrix/tensor estimation algorithms. For $d = 2$, it holds for HSVT as we establish in the proof of Theorem 5.4.1 for mSSA of $\tilde{O}(1/\sqrt{\min(N, T), T})$. It is straightforward to show that this is the best rate achievable for TE_2. For $d \geq 3$, it has recently been shown that Property 5.7.4 provably holds for a spectral gradient descent based algorithm Cai et al. (2021), conditioned on certain standard "incoherence" conditions imposed on the latent factors of $\boldsymbol{G}$; another spectral algorithm that achieved the same rate was furnished in Xia et al. (2018), which the authors also establish is minimax optimal.

tSSA algorithm. We now define the "meta" tSSA algorithm; the two algorithmic hyper-parameters are $L \geq 1$ (defined in (5.8)) and TE_3 (the order-three tensor estimation algorithm one chooses). First, using $X_n(t)$ for $n \in [N], t \in [T]$, construct Page tensor $\mathbb{T}$ as in (5.8). Second, obtain $\widehat{\mathsf{T}}$ as the output of $\text{TE}_3(\mathbb{T})$ and read off $\hat{f}_n(t)$ by selecting appropriate entry in $\widehat{\mathsf{T}}$.

Algorithmic comparison: tSSA vs. mSSA vs. ME. We now provide a unified view of tSSA, mSSA, and "vanilla" ME (which we describe below) to do time series imputation. All three methods have two key steps: (i) data transformation – converting the observations $X_n(t)$ into a particular data representation/structure; (ii) de-noising– applying some form of matrix/tensor estimation to de-noise the constructed data representation.

- tSSA – using $X_n(t)$, create the Page tensor $\mathbb{T} \in \mathbb{R}^{N \times L \times T/L}$ as in (5.8); apply $\text{TE}_3(\mathbb{T})$ to get $\widehat{\mathsf{T}}$ (e.g. using the method in Cai et al. (2021)); read off $\hat{f}_n(t)$ by selecting appropriate entry in $\widehat{\mathsf{T}}$.
- mSSA – using $X_n(t)$, create the stacked Page matrix $\text{SP}((X_1, \ldots, X_N), T, L) \in \mathbb{R}^{L \times (N \times T/L)}$ as detailed in Section 5.1.1; apply $\text{TE}_2(\text{SP}((X_1, \ldots, X_N), T, L))$ to get $\widehat{\text{SP}}((X_1, \ldots, X_N), T, L)$ (where we use HSVT for $\text{TE}_2(\cdot)$); read off $\hat{f}_n(t)$ by selecting appropriate entry in $\widehat{\text{SP}}((X_1, \ldots, X_N), T, L)$.
- ME – using $X_n(t)$, create $\boldsymbol{X} \in \mathbb{R}^{N \times T}$, where $\boldsymbol{X}_{nt}$ is equal to $X_n(t)$; apply $\text{TE}_2(\boldsymbol{X})$ (e.g. using HSVT as in mSSA) to get $\widehat{\boldsymbol{X}}$; read off $\hat{f}_n(t)$ by selecting appropriate entry in $\widehat{\boldsymbol{X}}$.

This perspective also suggests that one can use any "blackbox" matrix estimation routine to de-noise the constructed stacked Page matrix in mSSA; HSVT is one such choice that we analyze.

Theoretical comparison: tSSA vs. mSSA vs. ME. We now do a theoretical comparison of the relative effectiveness of tSSA, mSSA, and ME in imputing a multivariate time series

$X_n(t)$ for $n \in [N], t \in [T]$, as we vary N and T. To that end, let ImpErr(N, T; tSSA), ImpErr(N, T; mSSA), and ImpErr(N, T; ME) denote the imputation error for tSSA, mSSA, and ME, respectively.

Proposition 13. *For tSSA and mSSA, pick hyper-parameter $L = \sqrt{T}, L = \sqrt{\min(N,T)T}$, respectively. Let Property 5.7.4 hold. Then,*

(i) $T = o(N)$: $\text{ImpErr}(N, T; tSSA),\ \text{ImpErr}(N, T; mSSA) = \tilde{\Theta}(\text{ImpErr}(N, T; ME))$;
(ii) $T^{1/3} = o(N),\ N = o(T)$: $\text{ImpErr}(N, T; tSSA) = \tilde{o}(\text{ImpErr}(N, T; mSSA))$, $\text{ImpErr}(N, T; mSSA) = \tilde{o}(\text{ImpErr}(N, T; ME)$;
(iii) $N = o(T^{1/3})$: $\text{ImpErr}(N, T; mSSA) = \tilde{o}(\text{ImpErr}(N, T; tSSA))$, $\text{ImpErr}(N, T; tSSA) = \tilde{o}(\text{ImpErr}(N, T; ME))$,

where $\tilde{o}(\cdot)$, $\tilde{\Theta}(\cdot)$ suppresses dependence on noise parameters, CP-rank, poly-logarithmic factors.

We note given Property 5.7.4, $L = \sqrt{T}$ is optimal for tSSA and $L = \sqrt{\min(N,T)T}$ is optimal for mSSA. See Figure 5.2 in Section 5.1 for a graphical depiction of the different regimes in Proposition 13. Proofs of Proposition 12 and 13 below can be found in Appendix 5.20.

Application to Time-varying Recommendation Systems In Appendix 5.11, we discuss the extension of our spatio-temporal model and tSSA to time-varying recommendation systems.

■ 5.8 Conclusion

We provide theoretical justification of a practical, simple variant of mSSA, a method heavily used in practice but with limited theoretical understanding. We show how to extend mSSA to estimate time-varying variance and introduce a tensor variant, tSSA, which builds upon recent advancements in tensor estimation. We hope this work motivates future inquiry into the connections between the classical field of time series analysis and the modern, growing field of matrix/tensor estimation.

■ 5.9 Page vs. Hankel mSSA

This section discusses the benefits and drawbacks of using the Page matrix representation, as we propose in our variant, instead of the Hankel representation used in the original mSSA. Recall the key steps of the original SSA method in Section 5.2. The extension to mSSA is done by stacking the Hankel matrices induced by each of the N time series either column-wise (horizontal mSSA) or row-wise (vertical mSSA) Hassani and Mahmoudvand (2018). In this section, we will use mSSA to denote our mSSA variant, and hSSA/vSSA to denote the original horizontal/vertical mSSA. In what follows, we will compare our mSSA variant with hSSA/vSSA in terms of their: (i) theoretical analysis; (ii) computational complexity; and (iii) empirical performance.

Theoretical analysis. We re-emphasize that to the best of our knowledge, the theoretical analysis of the mSSA algorithm, both hSSA and vSSA, have been absent from the literature, despite their popularity. We do a comprehensive theoretical analysis of the variant of mSSA we propose. By utilizing the Page matrix, it allows us to invoke results from random matrix theory to prove our imputation and forecasting results. However, extending our analysis to the Hankel matrix representation is challenging as the Hankel matrix has repeated entries of the same time series observation. This leads to correlation in the noise in the observation of the entries of the Hankel matrix, which prevents us from invoking the results from random matrix theory in a straightforward way. The Page matrix representation does not have repeated entries of the same observation, and thus allows us to circumvent this issue in our theoretical analysis.

Computational complexity. Our mSSA variant is computationally far more efficient than both hSSA and vSSA. This is because the Page matrix representation of a multivariate time series with N time series and T time steps is a matrix of dimension $\sqrt{NT} \times \sqrt{NT}$ (with $L = \sqrt{NT}$)., i.e., it has a total of $\mathcal{O}(NT)$ entries. In contrast, the Hankel matrix representation is of dimension $T/4 \times 3NT/4$ for hSSA and $NT/4 \times 3T/4$ for vSSA (we set the parameter L to $T/4$ as recommended in Hassani and Mahmoudvand (2018)), i.e., both variants of the Hankel matrix have $\mathcal{O}(NT^2)$ entries. This makes computing the SVD (the most computationally intensive step of mSSA) prohibitive for hSSA and mSSA even for the standard time series benchmarks we consider in Section 5.6.

To empirically demonstrate the computational efficiency of our variant of mSSA, we compare its training time to that of hSSA and vSSA. Specifically, we measure the training time for mSSA, hSSA, and vSSA as we increase the number of time steps $T \in [400, 10000]$. We perform this experiment on two datasets: (i) the synthetic dataset; (ii) a subset of the electricity dataset, where we choose only 50 of the available 370 time series. Both datasets are described in details in Appendix 5.10. Figure 5.8 shows that in both datasets, the training time of both hSSA and vSSA can be as 600-1000x as high as the training time of our mSSA variant as we increase T.

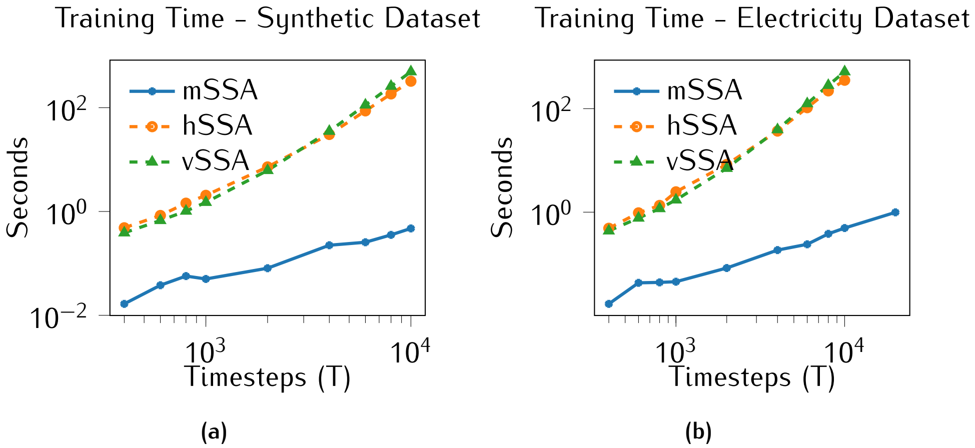
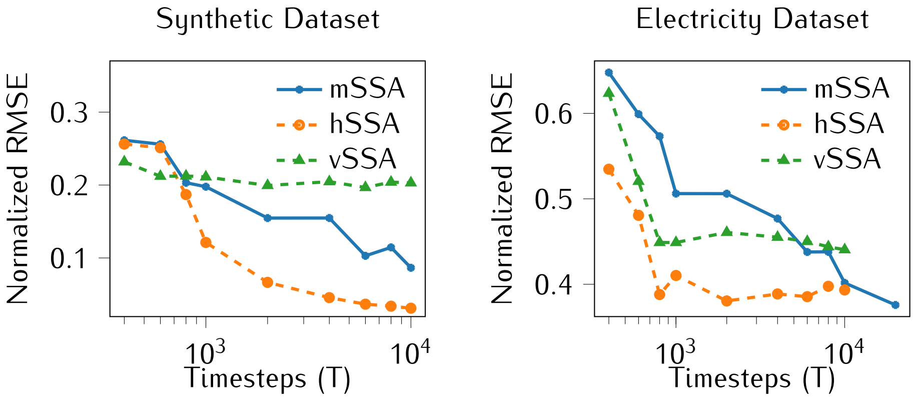


Figure 5.8: The training time of the original mSSA variants (hSSA in the orange dotted line and vSSA in the green dotted line) are orders of magnitude higher than that of the mSSA variant we propose (blue solid line).

Empirical performance. Here, we compare the forecasting performance of mSSA to that of hSSA and vSSA. We report performance in terms of the NRMSE of the three methods as we increase the number of time steps $T \in [400, 10000]$ in the aforementioned synthetic and electricity dataset. The goal in the synthetic dataset is to predict the next 50 time steps using one step ahead forecasts, while the goal in the electricity dataset is to predict the next three days using day-ahead forecasts. For hSSA and vSSA, we choose $L = T/4$ as recommended in Hassani and Mahmoudvand (2018); and for mSSA, we choose $L = \lfloor\sqrt{NT}\rfloor$. For all three methods, we choose the number of retained singular values based on the thresholding procedure outlined in Donoho and Gavish (2013).

Figures 5.9 shows the performance of the three methods in both datasets. We find that initially, with few data points ($T < 600$ in the synthetic data and $T < 4000$ in the electricity data), both hSSA and vSSA outperform mSSA. As we increase T, mSSA performance significantly improves and eventually outperforms vSSA. In the electricity dataset, mSSA performs similar to hSSA for $T = 10000$. These experiments suggest that if only a few observations were available, hSSA and vSSA might provide better performance. However, if the number of observations were relatively large, then the performance of mSSA is superior to vSSA and relatively similar to hSSA.



(a)

(b)

Figure 5.9: The forecasting error of the original mSSA variants (hSSA in the orange dotted line and vSSA in the green dotted line) and the proposed mSSA variant (blue solid line) as we increase T.

Importantly, the electricity dataset experiment illustrates a critical advantage of our mSSA variant. Specifically, when T is large such that running hSSA or vSSA is computationally infeasible, then one can achieve better accuracy using mSSA. For example, while we could not run the hSSA and vSSA on the electricity dataset with $T = 20000$ due to memory constraints, we were able to run mSSA and achieve a lower NRMSE. This suggests that our mSSA variant is the more practical mSSA algorithm when it comes to efficiently utilizing large multivariate time series.

■ 5.10 Experiment Details

In Appendix 5.10.1, we describe the datasets utilized. In Appendix 5.10.2, we describe the various algorithms we compare with as well as the choice of hyper-parameters used for each of them.

■ 5.10.1 Datasets

We use four real-world datasets and one synthetic dataset. The description and preprocessing we do for each of these datasets are as follows.

Electricity Dataset. This is a public dataset obtained from the UCI repository which shows the 15-minutes electricity load of 370 households Trindade (2014). As was done in Yu et al. (2016),Sen et al. (2019),Salinas et al. (2019), we aggregate the data into hourly intervals and use the first 25824 time-points for training, the next 288 points for validation, and the last 168 points for testing in the forecasting experiments. Specifically, in our testing period, we do 24-hour ahead forecasts for the next seven days (i.e. 24-step ahead forecast). See Table 5.3 for more details.

Traffic Dataset. This public dataset obtained from the UCI repository shows the occupancy rate of traffic lanes in San Francisco Trindade (2014). The data is sampled every 15 minutes but to be consistent with previous work in Yu et al. (2016), Sen et al. (2019), we aggregate the data into hourly data and use the first 10248 time-points for training, the next 288 points for validation, and the last 168 points for testing in the forecasting experiments. Specifically, in our testing period, we do 24-hour ahead forecasts for the next seven days (i.e. 24-step ahead forecast). See Table 5.3 for more details.

Financial Dataset. This dataset is obtained from the Wharton Research Data Services (WRDS) and contains the average daily stocks prices of 839 companies from October 2004 till November 2019 WRDS (2021). The dataset was preprocessed to remove stocks with any null values, or those with an average price below 30$ across the aforementioned period. This was simply done to constrain the number of time series for ease of experimentation and we end up with 839 time series (i.e. stock prices of listed companies) each with 3993 readings of daily stock prices. In our forecasting experiments, we train on the first 3693 time points, validate on the next 120 time points, while for testing we consider the task of predicting 180 time-points ahead one point at a time. That is, the goal here is to do one-day ahead forecasts for the next 180 days (i.e. 1-step ahead forecast). We choose to do so as this is a standard goal in finance. See Table 5.3 for more details.

M5 Dataset. This public dataset obtained from Kaggle's M5 Forecasting competition include daily sales data of 30490 items across different Walmart stores for 1941 days Makridakis et al. (2020). The dataset was preprocessed to only include items that has more than zero sales in at least 500 days. For forecasting, as is the goal in the Kaggle competition, we consider the task of predicting the sales for the next 28 days (i.e. 28-step ahead forecast). We use the first 1829 points for training, the next 84 points for cross validation, and the last 28 points for testing.

Synthetic Dataset. We generate the observation tensor $X \in \mathbb{R}^{n\times m\times T}$ by first randomly generating the two matrices $U \in \mathbb{R}^{r\times n} = [u_1, \ldots, u_n]$ and $V \in \mathbb{R}^{r\times m} = [v_1, \ldots, v_m]$; we do so by randomly sampling each coordinate of U, V independently from a standard normal. Then, we generate r mixtures of harmonics where each mixture $g_k(t), k \in [r]$, is generated as: $g_k(t) = \sum_{h=1}^{4} \alpha_h \cos(\omega_h t/T)$ where the parameters α_h, ω_h are selected uniformly at randomly from the ranges $[-1, 10]$ and $[1, 1000]$, respectively. Then each value in the observation tensor is constructed as follows: $X_{i,j}(t) = \sum_{k=1}^{r} u_{ik} v_{jk} g_k(t)$, where r is the tensor rank, $i \in [n]$, $j \in [m]$. In our experiment, we select $n = 5$, $m = 10$, $T = 15000$, and $r = 4$. This gives us $N = n \text{ x } m = 50$ time series each with 15000 observations per time series. In the forecasting experiments, we use the first 13700 points for training, the next 300 points for validation, while for testing, we do 10-step ahead forecasts for the final 1000 points. See Table 5.3 for more details.

Table 5.3: Dataset and training/validation/test split details.

Dataset	No.time series	Observations per time series	Forecast horizon (h)	Training period	No. validation windows W_{val}	Validation period	No. test windows	Test period
Electricity	370	26136	24	1 to 25824	2	25825 to 25968	7	25969 to 26136
Traffic	963	10560	24	1 to 10248	2	10249 to 10392	7	10393 to 10560
Synthetic	50	15000	10	1 to 13700	10	13701 to 14000	100	14001 to 15000
Financial	839	3993	1	1 to 3693	40	3694 to 3813	180	3814 to 3993
M5	15678	1941	28	1 to 1829	1	1830 to 1913	1	1914 to 1941

■ 5.10.2 Algorithms.

In this section, we describe the algorithms used throughout the experiments in more detail and the hyper-parameters/implementation used for each method.

mSSA & SSA. Note that since the SSA's variant described in Agarwal et al. (2018) is a special case of our proposed mSSA algorithm, we use our mSSA's implementation to perform the SSA experiments; key difference in SSA is that we do not "stack" the various Page matrices induced by each time series. For all experiments we choose the parameters through the cross validation process detailed in Appendix 5.10.3, where we perform a grid search for the following parameters:

1. *The number of retained singular values, k.* This parameter is chosen using one of the following data-driven methods: (i) we choose k based on the thresholding procedure outlined in Donoho and Gavish (2013), where the threshold is determined by the median of the singular values and the shape of the matrix; (ii) we choose k as the minimum number of singular values capturing $> 90\%$ of its spectral energy; (iii) we choose a constant low rank, specifically $k = 3$.
2. *The shape of the Page matrix.* For mSSA, we vary the shape of the Page matrix by choosing $L \in \{500, 1000, 2000, 3000\}$ for the electricity and Traffic datasets, $L \in \{500, 700, 800\}$ for the synthetic dataset, $L \in \{250, 500, 1000, 1500\}$ for the financial dataset, and $L \in \{10, 50, 100, 500\}$ for the M5 dataset . For SSA, we choose $L \in \{50, 100, 150\}$ in the electricity and Traffic datasets, $L \in \{30, 50, 100\}$ in the synthetic dataset, $L \in \{20, 30, 50\}$ in the financial dataset, and $L \in \{5, 10, 20, 40\}$ in the M5 dataset.
3. *Missing values initialization.* Initializing the missing values is done according to one of two methods: (i) set the missing values to zero; (ii) perform forward filling where each missing value is replaced by the nearest preceding observation, followed by backward filling to accommodate the situation when the first observation is missing.

DeepAR. We use the “DeepAREstimator” algorithm provided by the GluonTS package. We choose the parameters through a grid search for the following parameters:

1. *Context length.* This parameter determines the number of steps to unroll the RNN for before computing predictions. We choose this from the set $\{h \text{ (default)}, 2h, 3h\}$, where h is the prediction horizon.
2. *Number of Layers.* This parameter determines the number of RNN layers. We choose this from the set $\{2 \text{ (default)}, 3\}$.

TRMF. We use the implementation provided by the authors in the Github repository associated with the paper (Yu et al. (2016)). We choose the parameters through a grid search, as suggested by the authors in their codebase, for the following parameters:

1. *Matrix rank k.* This parameter represents the chosen rank for the $T \times N$ time series matrix, we choose k from the set $\{5, 10, 20, 40, 60\}$.
2. *Regularization parameters $\lambda_f, \lambda_x, \lambda_w$.* We choose these parameters from $\{0.05, 0.5, 5, 50\}$ as suggested in the authors repository.

For the lag indices , we include the last day and the same weekday in the last week for the traffic and electricity data, the last 30 points for the financial and synthetic dataset, and the last 10 points for the M5 dataset.

LSTM. Across all datasets, we use an LSTM network with $H \in \{2, 3, 4\}$ hidden layers each, with 45 neurons per layer, as is done in Sen et al. (2019). We use the Keras implementation of LSTM. As with other methods' parameters, H is chosen via cross validation.

Prophet. We used Prophet's Python library with the parameters selected using a grid search of the following parameters as suggested in Facebook (2020):

1. *Changepoint prior scale.* This parameter determines how much the trend changes at the detected trend changepoints. We choose this parameter from $\{0.001, 0.05, 0.2\}$.
2. *Seasonality prior scale.* This parameter controls the magnitude of the seasonality. We choose this parameter from $\{0.01, 10\}$.
3. *Seasonality Mode.* Which is chosen to be either 'additive' or 'multiplicative'.

■ 5.10.3 Parameters Selection

In all experiments, we choose the hyperparameters for out method and for the baselines by using cross-validation. Below, we detail the procedure for both imputation and forecasting experiments.

Imputation Experiments. To select the parameters in our imputation experiments, we additionally mask 10% of the observed data uniformly at random. Then, we evaluate the performance of each parameter choice in recovering these additionally masked observations. This process is repeated 3 times, and the choice of parameters that achieves the best performance (in NRMSE) across these runs is selected. In our results, we report the accuracy of the selected parameters in recovering the original missing values.

Forecasting Experiments. For parameters selection in the forecasting experiments, we use cross-validation on a rolling basis as typically used in time-series forecasting models Hyndman and Athanasopoulos (2018). In this procedure, there are multiple validation sets. For each validation set, we train the model only on previous observations. That is, no future observations can be used in training the model, which will occur when a typical cross-validation procedure is followed for time series data. In our experiments, we start with a subset of the data used for training, then we forecast the first validation set using h-step ahead forecasts for W_{val} windows , where the horizon h and the number of validation windows W_{val} are detailed in Table 5.3. We do this for three validation sets, each of length $h \times W_{val}$, and select the choice of parameters that achieves the best performance (in NRMSE) for evaluation on the test set. When evaluating on the test set, both the training and validation periods are used for training.

■ 5.11 Time-varying Recommendation Systems

In Section 5.7.2, we considered the setting where the $N \times T$ matrix $\boldsymbol{M}$ induced by the latent time series $f_1(\cdot), \dots, f_N(\cdot)$ is low-rank; in particular, Property 5.3.1 captures this spatial structure across these N time series. However, in many settings there is *additional* spatial structure across the N time series.

Recommendation systems – time-varying matrices/tensors. For example, in recommendation systems, for each $t \in T$, there is a $N_1 \times N_2$ matrix, $\boldsymbol{M}^{(t)} \in \mathbb{R}^{N_1 \times N_2}$ of interest. The n_1-th row and n_2-th column of $\boldsymbol{M}^{(t)}$ denotes the latent rating user n_1 has for product n_2, i.e., $\boldsymbol{M}^{(t)}_{n_1,n_2}$ denotes the value of the latent time series $f_{n_1,n_2}(\cdot)$ at time step t. To capture the latent structure across users and products, one typically assumes that each $\boldsymbol{M}^{(t)}$ is low-rank. More generally, at each time step t, $\boldsymbol{M}^{(t)} \in \mathbb{R}^{N_1 \times N_2, \dots, \times N_d}$ could be an order-d tensor. That is, $\boldsymbol{M}^{(t)}_{n_1,\dots,n_d}$ denotes the value of the latent time series $f_{n_1,\dots,n_d}(\cdot)$ at time step t for $n_1, \dots, n_d \in [N_1] \times \dots \times [N_d]$. For example, if $d = 3$, $\boldsymbol{M}^{(t)}$ might represent the t-th measurement for a collection of (x, y, z)-spatial coordinates. Let $\boldsymbol{N} \in \mathbb{R}^{N_1 \times N_2, \dots, \times N_d \times T}$ denote the $d + 1$ order tensor induced by viewing each order-d tensor $\boldsymbol{M}^{(t)}$ as the t-th 'slice' of $\boldsymbol{N}$, for $t \in [T]$. Again, to capture the spatial and temporal structure of these latent time series, we posit the following spatio-temporal model for $\boldsymbol{N}$, which is a higher-order analog of the model assumed in Property 5.3.1.

Property 5.11.1. *Let $\boldsymbol{N}$ have CP-rank at most R. That is, for any $n_1, \dots, n_d \in [N_1] \times \dots \times [N_d]$*

$$\boldsymbol{N}_{n_1,\dots,n_d,t} = \sum_{r=1}^{R} U_{n_1,r} \dots U_{n_d,r} \; W_{rt},$$

where the factorization is such that $|U_{n_1,r}|, \dots |U_{n_d,r}| \leq \Gamma_1$, $|W_{rt}| \leq \Gamma_2$ for constants $\Gamma_1, \Gamma_2 > 0$.

As before, to explicitly model the temporal structure, we continue to assume Property 5.3.2 holds for the latent time factors $W_{r\cdot}$ for $r \in [R]$.

Order-$d+2$ Page tensor representation. We now consider the following order-$d+2$ Page tensor representation of $\boldsymbol{N}$. In particular, given the hyper-parameter $L \geq 1$, define $\mathbf{HT} \in \mathbb{R}^{N_1 \times \ldots \times N_d \times T/L \times L}$ such that for $n_1, \ldots, n_d \in [N_1] \times \ldots \times [N_d]$, $\ell \in [L]$, $s \in [T/L]$,

$$\mathbf{HT}_{n_1,\ldots,n_d,\ell,s} = f_{n_1,\ldots,n_d}((s-1) \times L + \ell).$$

The corresponding observation tensor, $\mathbb{HT} \in (\mathbb{R} \cup \{\star\})^{N_1 \times \ldots \times N_d \times T/L \times L}$, is

$$\mathbb{HT}_{n_1,\ldots,n_d,\ell,s} = X_{n_1,\ldots,n_d}((s-1) \times L + \ell). \tag{5.9}$$

Recall from (5.1) that $X_{n_1,\ldots,n_d}(t)$ is the noisy, missing observation we get of $f_{n_1,\ldots,n_d}(t)$. $\mathbf{HT}$ and $\mathbb{HT}$ then have the following property:

Proposition 14. *Let Properties 5.11.1, 5.3.2, and 5.4.1 hold. Then, for any $1 \leq L \leq \sqrt{T}$, $\mathbf{HT}$ has CP-rank at most R x G. Further, all entries of $\mathbb{HT}$ are independent random variables with each entry observed with probability $\rho \in (0,1]$, and $\mathbb{E}[\mathbb{HT}] = \rho\mathbf{HT}$.*

Analogous to Proposition 12, Proposition 14 also establishes that order-$d+2$ Page tensor representation of the various latent time series $f_{n_1,\ldots,n_d}(\cdot)$ has CP-rank that continues to be bounded by R x G. Proof of Proposition 14 can be found in Appendix 5.20.

Higher-order tensor singular spectrum analysis (htSSA). Proposition 14 motivates the following algorithm, which exploits the further spatial structure amongst the N time series. We now define the "meta" htSSA algorithm. The two algorithmic hyper-parameters are $L \geq 1$ (defined in (5.8)) and TE_{d+2} (the order-$d+2$ tensor estimation algorithm one chooses). First, using the observations $X_{n_1,\ldots,n_d}(t)$ for $n_1, \ldots, n_d \in [N_1] \times \ldots \times [N_d], t \in [T]$ we construct the higher-order Page tensor $\mathbb{HT}$ as in (5.9). Second, we obtain $\widehat{\mathbf{HT}}$ as the output of $\text{TE}_{d+2}(\mathbb{HT})$, and read off $\hat{f}_{n_1,\ldots,n_d}(t)$ by selecting the appropriate entry in $\widehat{\mathbf{HT}}$.

Relative effectiveness of mSSA, htSSA, and tensor estimation (TE). Again, for ease of exposition, we consider the case where $\rho = 1$. We now briefly discuss the relative effectiveness of htSSA, mSSA,and "vanilla" tensor estimation (TE) in imputing $X_{n_1,\ldots,n_d}(\cdot)$ to estimate $f_{n_1,\ldots,n_d}(\cdot)$. mSSA and htSSA have been previously described. In TE, one directly de-noises the original order-$d+1$ tensor induced by the noisy observations, which we denote $\boldsymbol{X} \in \mathbb{R}^{N_1 \times N_2, \ldots, \times N_d \times T}$, where $\boldsymbol{X}_{n_1,\ldots,n_d,t} = X_{n_1,\ldots,n_d}(t)$. In particular, one produces an estimate of $\widehat{\boldsymbol{N}} = \text{TE}_{d+1}(\boldsymbol{X})$, and then produces the estimates $\hat{f}_{n_1,\dots,n_d}(t)$ by reading off the appropriate entry of $\widehat{\boldsymbol{N}}$. Let $\text{ImpErr}(N, T; \text{htSSA})$, $\text{ImpErr}(N, T; \text{mSSA})$, and $\text{ImpErr}(N, T; \text{TE})$ denote the imputation error for htSSA, mSSA, and TE, respectively. Now if we assume Property 5.7.4 holds, we have

$$\begin{aligned}
\text{ImpErr}(N, T; \text{htSSA}) &= \tilde{\Theta}\left(\frac{1}{\min\left(N_1, \dots, N_d, \sqrt{T}\right)^{\lceil \frac{d+2}{2} \rceil}}\right), \\
\text{ImpErr}(N, T; \text{mSSA}) &= \tilde{\Theta}\left(\frac{1}{\sqrt{\min(N, T)T}}\right), \\
\text{ImpErr}(N, T; \text{TE}) &= \tilde{\Theta}\left(\frac{1}{\min\left(N_1, \dots, N_d, T\right)^{\lceil \frac{d+1}{2} \rceil}}\right).
\end{aligned}$$

Then just as was done in the proof of Proposition 13, for any given d, one can reason about the relative effectiveness of htSSA, mSSA, and TE for different asymptotic regimes of the relative ratio of N and T.

■ 5.12 Proof of Proposition 6

Below, we present the proof of Proposition 6. First we define the stacked Hankel matrix of N time series over T time steps. Precisely, given N latent time series $f_1, \dots, f_N$, consider the stacked Hankel matrix induced by each of them over T time steps, $[T]$, defined as follows. It is $\text{SH} \in \mathbb{R}^{\lfloor T/2 \rfloor \times N \lfloor T/2 \rfloor}$ where its entry in row $i \in [\lfloor T/2 \rfloor]$ and column $j \in [N\lfloor T/2 \rfloor]$, SH_{ij}, is given by

$$\text{SH}_{ij} = f_{n(i,j)}(i + (j \mod \lfloor T/2 \rfloor) - 1), \quad \text{where } n(i,j) = \left\lceil \frac{j}{\lfloor T/2 \rfloor} \right\rceil.$$

We now establish Proposition 15, which immediately implies Proposition 6 – the stacked Page matrix can be viewed as a sub-matrix of SH, by selecting the appropriate columns.

Proposition 15. *Let Properties 5.3.1 and 5.5.1 hold for N latent time series of interest, $f_1, \dots, f_N$. Then for any $T \geq 1$, the stacked Hankel Matrix of these N time series has ϵ'-approximate rank $R \times G$ with $\epsilon' = R\Gamma_1 \epsilon$.*

Proof. We have N latent time series $f_1, \dots, f_n$ satisfying Properties 5.3.1 and 5.5.1.

Consider their stacked Hankel matrix over $[T]$, $\mathrm{SH} \in \mathbb{R}^{\lfloor T/2 \rfloor \times N \lfloor T/2 \rfloor}$. By definition for $i \in [\lfloor T/2 \rfloor]$ and $j = (n-1) \times \lfloor T/2 \rfloor + j'$ for $j' \in [\lfloor T/2 \rfloor]$, we have

$$\mathrm{SH}_{ij'} = f_n(i + j' - 1).$$

That is,

$$\begin{aligned} \mathrm{SH}_{ij} &= f_n(i + j' - 1) \\ &= \sum_{r=1}^{R} U_{nr} W_{r(i+j'-1)}. \end{aligned} \tag{5.10}$$

Let $H(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ be the Hankel matrix associated with $W_{r\cdot}$ over $[T]$. Due to Property 5.5.1, there exists a low-rank matrix $M(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ such that (a) $\text{rank}(M(r)) \leq G$, (b) $H(r) - M(r)_{\infty} \leq \epsilon$. That is, for any $i, j' \in [\lfloor T/2 \rfloor]$, we have that $M(r)_{ij'} = \sum_{g=1}^{G} a^r_{ig} b^r_{j'g}$ for some $a^r_{i\cdot}, b^r_{j'\cdot} \in \mathbb{R}^G$. Therefore, for any $i, j' \in [\lfloor T/2 \rfloor]$, we have that

$$\begin{aligned} W_{r(i+j'-1)} = H(r)_{ij'} &= M(r)_{ij'} + (H(r)_{ij'} - M(r)_{ij'}) \\ &= \sum_{g=1}^{G} a^r_{ig} b^r_{j'g} + (H(r)_{ij'} - M(r)_{ij'}). \end{aligned} \tag{5.11}$$

From (5.10) and (5.11), we conclude that

$$\begin{aligned} \mathrm{SH}_{ij} &= \sum_{r=1}^{R} \sum_{g=1}^{G} U_{nr} a^r_{ig} b^r_{j'g} + \sum_{r=1}^{R} U_{nr}(H(r)_{ij'} - M(r)_{ij'}) \\ &= \sum_{(r,g) \in [R] \times [G]} a^r_{ig} \times (U_{nr} b^r_{j'g}) + \sum_{r=1}^{R} U_{nr}(H(r)_{ij'} - M(r)_{ij'}). \end{aligned}$$

Define matrix $\mathrm{M} \in \mathbb{R}^{\lfloor T/2 \rfloor \times N \lfloor T/2 \rfloor}$ with its entry for row $i \in [\lfloor T/2 \rfloor]$ and column $j = (n-1) \times \lfloor T/2 \rfloor + j'$ for $j' \in [\lfloor T/2 \rfloor]$ given by

$$\begin{aligned} \mathrm{M}_{ij} &= \sum_{(r,g) \in [R] \times [G]} a^r_{ig} \times (U_{nr} b^r_{j'g}) \\ &= \sum_{(r,g) \in [R] \times [G]} \alpha_{i(r,g)} \beta_{j(r,g)}, \end{aligned}$$

where $\alpha_{i(r,g)} = a^r_{ig}$ and $\beta_{j(r,g)} = U_{nr} b^r_{j'g}$. Further,

$$\begin{aligned}|\mathrm{SH}_{ij} - \mathrm{M}_{ij}| &\le \sum_{r=1}^{R} |U_{nr}||(H(r)_{ij'} - M(r)_{ij'})| \\ &\le \sum_{r=1}^{R} \Gamma_1 H(r) - M(r)_\infty \ \le \ R\Gamma_1 \epsilon.\end{aligned}$$

That is, the stacked Hankel matrix SH of N time series of $[T]$ has ϵ'-approximate rank G x R with $\epsilon' = R\Gamma_1\epsilon$. This completes the proof. ■

5.13 Proofs For Section 5.5

5.13.1 Proof of Proposition 7

Proof. Let f_1, f_2 have a (G_1, ϵ_1) and (G_2, ϵ_2)-Hankel representation, respectively. For any $T \ge 1$, let $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ be the Hankel matrices of f_1, f_2, respectively, over the time interval $[T]$. By definition, there exists matrices $M_1, M_2 \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ such that rank$(\boldsymbol{M}_1) \le G_1$, $\boldsymbol{M}_1 - \boldsymbol{H}_{1\infty} \le \epsilon_1$ and rank$(\boldsymbol{M}_2) \le G_2$, $\boldsymbol{M}_2 - \boldsymbol{H}_{2\infty} \le \epsilon_2$.

Component-wise addition. Note the Hankel matrix of $f_1 + f_2$ over $[T]$ is $\boldsymbol{H}_1 + \boldsymbol{H}_2$. Then, matrix $\boldsymbol{M} = \boldsymbol{M}_1 + \boldsymbol{M}_2$ has rank at most $G_1 + G_2$ since for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, it is the case that rank$(\boldsymbol{A} + \boldsymbol{B}) \le$ rank$(\boldsymbol{A})$ + rank$(\boldsymbol{B})$. Further, $\boldsymbol{H}_1 + \boldsymbol{H}_2 - (\boldsymbol{M}_1 + \boldsymbol{M}_2)_\infty \le \epsilon_1 + \epsilon_2$. Therefore it follows that $f_1 + f_2$ has $(G_1 + G_2, \epsilon_1 + \epsilon_2)$-Hankel representation.

Component-wise multiplication. For $f_1 \circ f_2$, its Hankel over $[T]$ is given by $\boldsymbol{H}_1 \circ \boldsymbol{H}_2$ where we abuse notation of $\circ$ in the context of matrices as the Hadamard product of matrices. Let $\boldsymbol{M} = \boldsymbol{M}_1 \circ \boldsymbol{M}_2$. Then rank$(M) \le G_1$ x G_2 since for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, rank$(\boldsymbol{A} \circ \boldsymbol{B}) \le$ rank$(\boldsymbol{A})$rank$(\boldsymbol{B})$. Now

$$\begin{aligned}\boldsymbol{H}_1 \circ \boldsymbol{H}_2 - \boldsymbol{M}_1 \circ \boldsymbol{M}_{2\infty} &\le \boldsymbol{H}_1 \circ \boldsymbol{H}_2 - \boldsymbol{H}_1 \circ \boldsymbol{M}_{2\infty} + \boldsymbol{H}_1 \circ \boldsymbol{M}_2 - \boldsymbol{M}_1 \circ \boldsymbol{M}_{2\infty} \\ &\le \boldsymbol{H}_{1\infty} \boldsymbol{H}_2 - \boldsymbol{M}_{2\infty} + \boldsymbol{M}_{2\infty} \boldsymbol{H}_1 - \boldsymbol{M}_{1\infty} \\ &\le f_{1\infty}\epsilon_2 + (\boldsymbol{M}_2 - \boldsymbol{H}_{2\infty} + \boldsymbol{H}_{2\infty})\epsilon_1 \\ &\le f_{1\infty}\epsilon_2 + (f_{2\infty} + \epsilon_2)\epsilon_1 \\ &= f_{1\infty}\epsilon_2 + f_{2\infty}\epsilon_1 + \epsilon_1\epsilon_2 \ \le \ 3\max(\epsilon_1, \epsilon_2)\max(f_{1\infty}, f_{2\infty}).\end{aligned}$$

This completes the proof of Proposition 7. ■

■ 5.13.2 Proof of Proposition 8

Proof. Proof is immediate from Definitions 5.5.2 and 5.5.3. ■

■ 5.13.3 Proof of Proposition 9

Helper Lemmas for Proposition 9

We begin by stating some classic results from Fourier Analysis. To do so, we introduce some notation. Throughout, we have $R > 0$.

$C[0,R]$ and $L^2[0,R]$ functions. $C[0,R]$ is the set of real-valued, continuous functions defined on $[0,R]$. $L^2[0,R]$ is the set of square integrable functions defined on $[0,R]$, i.e. $\int_0^R f^2(t)dt \leq \infty$

Inner Product of functions in $L^2[0,R]$. $L^2[0,R]$ is a space endowed with inner product defined as $\langle f, g \rangle := \frac{1}{R}\int_0^R f(t)g(t)dt$, and associated norm as $\|f\| := \sqrt{\frac{1}{R}\int_0^R f^2(t)dt}$.

Fourier Representation of functions in $L^2[0,R]$. For $f \in L^2[0,R]$, define its $G \geq 1$-order Fourier representation, $\mathcal{F}(f,G) \in L^2[0,R]$ as

$$\mathcal{F}(f,G)(t) = a_0 + \sum_{g=1}^{G}(a_g \cos(2\pi g t/R) + b_g \cos(2\pi g t/R)), \quad t \in [0,R], \tag{5.12}$$

where a_0, a_g, b_g with $g \in [G]$ are called the Fourier coefficients of f, defined as

$$\begin{aligned}
a_0 &:= \langle f, 1 \rangle = \frac{1}{R}\int_0^R f(t)dt, \\
a_g &:= \langle f, \cos(2\pi g t/R) \rangle = \frac{1}{R}\int_0^R f(t)\cos(2\pi g t/R)dt, \\
b_g &:= \langle f, \sin(2\pi g t/R) \rangle = \frac{1}{R}\int_0^R f(t)\sin(2\pi g t/R)dt.
\end{aligned}$$

We now state a classic result from Fourier analysis.

Theorem 5.13.1 (Grafakos (2008))**.** *Given $k \geq 1, R > 0$, let $f \in C^k(R, PER)$. Then, for any $t \in [0, R]$ (or more generally $t \in \mathbb{R}$),*

$$\lim_{G \to \infty} \mathcal{F}(f, G)(t) \to f(t).$$

We next argue that if $f \in C^k(R, \text{PER})$, then its Fourier coefficients decay rapidly.

Lemma 5.13.1. *Given $k \geq 1, R > 0$, let $f \in C^k(R, PER)$. Then, for $j \in [k]$, the G-order Fourier coefficient of $f^{(j)}$, the j-th derivative of f, recursively satisfy the following relationship: for $g \in [G]$,*

$$a_g^{(j)} = -\Big(\frac{2\pi g}{R}\Big) b_g^{(j-1)}, \qquad b_g^{(j)} = \Big(\frac{2\pi g}{R}\Big) a_g^{(j-1)}. \tag{5.13}$$

Proof. We establish (5.13) for $a_g^{(1)}$, $g \in [G]$. Notice that an identical argument applies to establish (5.13) for any $a_g^{(j)}, b_g^{(j)}$ for $j \in [k]$ and $g \in [G]$.

$$\begin{aligned} a_g^{(1)} = \langle f^{(1)}, \cos(2\pi g t/R) \rangle &= \frac{1}{R} \int_0^R f^{(1)}(t) \cos(2\pi g t/R) dt \\ &\overset{(a)}{=} \frac{1}{R} \Big(\Big[f(t) \cos(2\pi g t/R) \Big]_0^R - \frac{2\pi g}{R} \Big[\frac{1}{R} \int_0^R f(t) \sin(2\pi g t/R) dt \Big] \Big) \\ &= -\Big(\frac{2\pi g}{R}\Big) b_g^{(0)}. \end{aligned}$$

(a) follows by integration by parts. ■

Completing Proof of Proposition 9

Proof. For $G \in \mathbb{N}$, let $\mathcal{F}(f, G)$ be defined as in (5.12). Then for $t \in \mathbb{R}$

$$\begin{aligned} |f(t) - \mathcal{F}(f, G)(t)| &\overset{(a)}{=} \Big| \sum_{g=G+1}^{\infty} (a_g \cos(2\pi g t/R) + b_g \cos(2\pi g t/R)) \Big| \\ &\leq \sum_{g=G+1}^{\infty} |a_g| + |b_g| \\ &\overset{(b)}{\leq} \sum_{g=G+1}^{\infty} \Big(\frac{R}{2\pi g}\Big)^k \Big(|a_g^{(k)}| + |b_g^{(k)}| \Big) \end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \sqrt{2}\Big(\frac{R}{2\pi}\Big)^k \sqrt{\sum_{g=G+1}^{\infty}\Big(\frac{1}{g}\Big)^{2k}} \sqrt{\sum_{g=G+1}^{\infty}\Big(|a_g^{(k)}|^2+|b_g^{(k)}|^2\Big)} \\
&\overset{(d)}{\leq} \sqrt{2}\Big(\frac{R}{2\pi}\Big)^k \frac{1}{G^{k-0.5}} \sqrt{\sum_{g=G+1}^{\infty}\Big(|a_g^{(k)}|^2+|b_g^{(k)}|^2\Big)} \\
&\overset{(e)}{\leq} \sqrt{2}\Big(\frac{R}{2\pi}\Big)^k \frac{\left\|f^{(k)}\right\|}{G^{k-0.5}} \\
&= C(k,R)\frac{\left\|f^{(k)}\right\|}{G^{k-0.5}},
\end{aligned}
$$

where $C(k,R)$ is a constant that depends only on k and R; (a) follows from Theorem 5.13.1; (b) follows from Lemma 5.13.1; (c) follows from Cauchy-Schwarz inequality and fact that $(\alpha+\beta)^2 \leq 2(\alpha^2+\beta^2)$ for any $\alpha,\beta \in \mathbb{R}$; (d) $\sum_{g=G+1}^{\infty} g^{-2k} \leq \int_G^{\infty} x^{-2k}dx$ which can be bounded as $G^{-2k+1}/(2k-1)$ which is at most G^{-2k+1} since $k \geq 1$; (e) follows from Bessel's inequality, i.e. $\left\|f^{(k)}\right\|^2 \geq \sum_{g=0}^{\infty}(|a_g^{(k)}|^2+|b_g^{(k)}|^2)$.

Thus, for any $t \in \mathbb{R}$, we have a uniform error bound for f being approximated by $\mathcal{F}(f,G)$ which is a sum of $2G$ harmonics. Noting $2G$ harmonics can be represented by an order-$4G$ LRF (by Proposition 3),we complete the proof. ■

■ 5.13.4 Proof of Proposition 10

This analysis is adapted from Xu (2017a).

Proof. **Step 1: Partitioning the space $[0,1)^K$.** Consider an equal partition of $[0,1)^K$. Precisely, for any $k \in \mathbb{N}$, we partition the the set $[0,1)$ into $1/k$ half-open intervals of length $1/k$, i.e, $[0,1) = \cup_{i=1}^{k}[(i-1)/k, i/k)$. It follows that $[0,1)^K$ can be partitioned into k^K cubes of forms $\otimes_{j=1}^{K}\big[(i_j-1)/k, i_j/k\big)$ with $i_j \in [k]$. Let $\mathcal{E}_k$ be such a partition with $I_1, I_2, \ldots, I_{k^K}$ denoting all such cubes and $z_1, z_2, \ldots, z_{k^K} \in \mathbb{R}^K$ denoting the centers of those cubes.

Step 2: Taylor Expansion of $g(\cdot,\omega)$. Consider a fixed ω. To reduce notational overload, we suppress dependence of g on ω, and abuse notation by using $g(\cdot) = g(\cdot,\omega)$ in what follows.

For every I_i with $1 \le i \le k^K$, define $P_{I_i,\ell}(x)$ as the degree-ℓ Taylor's series expansion of $g(x)$ at point z_i:

$$P_{I_i,\ell}(x) = \sum_{\kappa:|\kappa|\le\ell} \frac{1}{\kappa!} (x - z_i)^\kappa \, \nabla_\kappa g(z_i), \tag{5.14}$$

where $\kappa = (\kappa_1, \dots, \kappa_d)$ is a multi-index with $\kappa! = \prod_{i=1}^{K} \kappa_i!$, and $\nabla_\kappa g(z_i)$ is the partial derivative defined in Section 5.5.2. Note similar to g, $P_{I_i,\ell}(x)$ really refers to $P_{I_i,\ell}(x, \omega)$.

Now we define a degree-ℓ piecewise polynomial

$$P_{\mathcal{E}_k,\ell}(x) = \sum_{i=1}^{k^K} P_{I_i,\ell}(x) \mathbb{1}(x \in I_i).$$

For the remainder of the proof, let $\ell = \lfloor \alpha \rfloor$ (recall $\lfloor \alpha \rfloor$ refers to the largest integer strictly smaller than α). Since $f \in \mathcal{H}(\alpha, L)$, it follows that

$$\begin{aligned}
\sup_{x\in[0,1)^K} |g(x) - P_{\mathcal{E}_k,\ell}(x)| &= \max_{1\le i\le k^K} \sup_{x\in I_i} |g(x) - P_{I_i,\ell}(x)| \\
&\overset{(a)}{=} \max_{1\le i\le k^K} \sup_{x\in I_i} \left| \sum_{\kappa:|\kappa|\le\ell-1} \frac{\nabla_\kappa g(z_i)}{\kappa!}(x - z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(\tilde{z}_i)}{\kappa!}(x - z_i)^\ell - P_{I_i,\ell}(x) \right| \\
&\overset{(b)}{=} \max_{1\le i\le k^K} \sup_{x\in I_i} \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(\tilde{z}_i)}{\kappa!}(x - z_i)^\ell - \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i)}{\kappa!}(x - z_i)^\ell \right| \\
&= \max_{1\le i\le k^K} \sup_{x\in I_i} \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(\tilde{z}_i) - \nabla_\kappa g(z_i)}{\kappa!}(x - z_i)^\ell \right| \\
&\overset{(c)}{\le} \max_{1\le i\le k^K} \sup_{x\in I_i} x - z_{i}{}_{\infty}^{\ell} \sup_{x\in I_i} \sum_{\kappa:|\kappa|=\ell} \frac{1}{\kappa!} \left|\nabla_\kappa g(\tilde{z}_i) - \nabla_\kappa g(z_i)\right| \\
&\overset{(d)}{\le} \mathcal{L} k^{-\alpha}.
\end{aligned} \tag{5.15}$$

where (a) follows from multivariate version of Taylor's theorem (and using the Lagrange form for the remainder) and $\tilde{z}_i \in [0, 1)^K$ is a vector that can be represented as $z_i + cx$ for $c \in (0, 1)$; (b) follows from (5.14); (c) follows from Holder's inequality; (d) follows from Definition 5.5.5.

Step 3: Construct Low-Rank Approximation of Time Series Hankel Using $P_{\mathcal{E}_k,\ell}$. Recall the Hankel matrix, $\boldsymbol{H} \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ induced by the original time series over $[T]$, where $\boldsymbol{H}_{ts} = g(\theta_t, \omega_s)$, $t, s \in [\lfloor T/2 \rfloor]$ with $g(\cdot, \omega) \in \mathcal{H}(\alpha, \mathcal{L})$ for any ω. We now construct a low-rank approximation of it using $P_{\mathcal{E}_k,\ell} = P_{\mathcal{E}_k,\ell}(\cdot, \omega)$. Define $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$, where $\widetilde{\boldsymbol{H}}_{ts} = P_{\mathcal{E}_k,\ell}(\theta_t, \omega_s)$, $t, s \in [\lfloor T/2 \rfloor]$.

By (5.15), we have that for all $t, s \in [\lfloor T/2 \rfloor]$,

$$\left| \boldsymbol{H}_{ts} - \widetilde{\boldsymbol{H}}_{ts} \right| \le \mathcal{L} k^{-\alpha}.$$

It remains to bound the rank of $\widetilde{\boldsymbol{H}}$. Note that since $P_{\mathcal{E}_k,\ell}(\cdot, \omega)$ is a piecewise polynomial of degree $\ell = \lfloor \alpha \rfloor$ for any given ω, it has the following decomposition: for $t, s \in [\lfloor T/2 \rfloor]$,

$$\widetilde{\boldsymbol{H}}_{ts} = P_{\mathcal{E}_k,\ell}(\theta_t, \omega_s) \; = \sum_{i=1}^{k^K} \langle \Phi(\theta_t), \beta_{I_i,s} \rangle \mathbb{1}(\theta_t \in I_i)$$

where for any $\theta \in \mathbb{R}^K$,

$$\Phi(\theta) = \Big(1, \theta_1, \dots, \theta_K, \dots, \theta_1^\ell, \dots, \theta_K^\ell \Big)^T,$$

the vector of all monomials of degree less than or equal to ℓ, and $\beta_{I_i,s}$ is a vector collecting the corresponding coefficients. The number of such monomials is easily show to be equal to $C(\alpha, K) := \sum_{i=1}^{\lfloor \alpha \rfloor} \binom{i+K-1}{i}$. That is, $\widetilde{\boldsymbol{H}}_{ts} = u_t^T v_s$ where u_t, v_s are of dimension at most $k^K C(\alpha, K)$ for each $t, s \in [\lfloor T/2 \rfloor]$. That is, $\widetilde{\boldsymbol{H}}$ has rank at most $k^K C(\alpha, K)$. Setting $k = \left\lceil \frac{1}{\epsilon} \right\rceil$ completes the proof. ■

■ 5.14 Helper Lemmas

We recall known concentration and perturbation inequalities that will be useful throughout.

Theorem 5.14.1 (Bernstein's Inequality Bernstein (1946)). *Suppose that $X_1, \dots, X_n$ are independent random variables with zero mean, and M is a constant such that $|X_i| \le M$*

with probability one for each i. Let $S := \sum_{i=1}^{n} X_i$ and $v := Var(S)$. Then for any $t \geq 0$,

$$\mathbb{P}(|S| \geq t) \leq 2\exp\left(-\frac{3t^2}{6v + 2Mt}\right).$$

Theorem 5.14.2 (Norm of matrices with sub-gaussian entries Vershynin (2010)). *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$, we have*

$$\left\|A\right\| \leq CK(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp\left(-t^2\right)$. Here, $K = \max_{i,j}\left\|A_{ij}\right\|_{\psi_2}$.

Lemma 5.14.1 (Maximum of sequence of random variables Vershynin (2010)). *Let X_1, $X_2, \ldots, X_n$ be a sequence of random variables, which are not necessarily independent, and satisfy $\mathbb{E}[X_i^{2p}]^{\frac{1}{2p}} \leq Kp^{\frac{\beta}{2}}$ for some $K, \beta > 0$ and all i. Then, for every $n \geq 2$,*

$$\mathbb{E}\max_{i \leq n}|X_i| \leq CK\log^{\frac{\beta}{2}}(n).$$

We note that Lemma 5.14.1 implies that if $X_1, \ldots, X_n$ are ψ_α random variables with $X_{i\psi_\alpha} \leq K_\alpha$ for all $i \in [n]$, then

$$\mathbb{E}\max_{i \leq n}|X_i| \leq CK_\alpha\log^{\frac{1}{\alpha}}(n).$$

Lemma 5.14.2 (Modified Hoeffding Inequality Agarwal et al. (2021d)). *Let $X \in \mathbb{R}^n$ be random vector with independent mean-zero sub-Gaussian random coordinates with $X_{i\psi_2} \leq K$. Let $a \in \mathbb{R}^n$ be another random vector that satisfies $a_2 \leq b$ almost surely for some constant $b \geq 0$. Then for all $t \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^2b^2}\right),$$

where $c > 0$ is a universal constant.

Lemma 5.14.3 (Modified Hanson-Wright Inequality Agarwal et al. (2021d)). *Let $X \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with $X_{i\psi_2} \leq K$. Let $A \in \mathbb{R}^{n \times n}$ be a random matrix satisfying $A_2 \leq a$ and $A_F^2 \leq b$ almost surely for some*

$a, b \geq 0$. Then for any $t \geq 0$,

$$\mathbb{P}\left(\left|X^T \boldsymbol{A} X - \mathbb{E}[X^T \boldsymbol{A} X]\right| \geq t\right) \leq 2 \cdot \exp\left(- c \min\left(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\right)\right).$$

Lemma 5.14.4 (Weyl's inequality). *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, let σ_i and $\widehat{\sigma}_i$ be the i-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then for all $i \in [m \wedge n]$,*

$$|\sigma_i - \widehat{\sigma}_i| \leq \left\|\boldsymbol{A} - \boldsymbol{B}\right\|_2.$$

■ 5.15 Matrix Estimation via HSVT

This section describes and analyzes a well-known matrix estimation method, Hard Singular Value Thresholding (HSVT). While the analysis utilizes known arguments from the literature, we need to adapt it for the setting where the underlying 'signal' is only approximately low-rank.

■ 5.15.1 Setup, Notations

Setup. Given a deterministic matrix $\boldsymbol{M} \in \mathbb{R}^{q \times p}$ with $p, q \in \mathbb{N}$ and $q \leq p$, a random matrix $\boldsymbol{Y} \in \mathbb{R}^{q \times p}$ is such that all of its entries, $Y_{ij},\ i \in [q],\ j \in [p]$ are mutually independent and for any given $i \in [q],\ j \in [p]$,

$$Y_{ij} = \begin{cases} M_{ij} + \boldsymbol{\varepsilon}_{ij} & \text{w.p. } \rho, \text{ (i.e. observed)} \\ 0 & \text{w.p. } 1 - \rho, \text{ (i.e. not observed)} \end{cases}$$

for some $\rho \in (0, 1]$ with $\boldsymbol{\varepsilon}_{ij}$ are independent random variables with $\mathbb{E}[\boldsymbol{\varepsilon}_{ij}] = 0$ and $\left\|\boldsymbol{\varepsilon}_{ij}\right\|_{\psi_2} \leq \sigma$. Given this, we have $\mathbb{E}[\boldsymbol{Y}] = \rho \boldsymbol{M}$. Defineff

$$\hat{\rho} = \max\left(1/(q\ p), (\sum_{i=1}^{q} \sum_{j=1}^{p} \mathbf{1}(Y_{ij} \text{ is obs.}))/(q\ p)\right).$$

Goal of Matrix Estimation. The goal of matrix estimation is to produce an estimate

$\widehat{\boldsymbol{M}}$ from observation $\boldsymbol{Y}$ so that $\widehat{\boldsymbol{M}}$ is close to $\boldsymbol{M}$. In particular, we will be interested in bounding the error between $\widehat{\boldsymbol{M}}$ and $\boldsymbol{M}$ using the following metric: $\widehat{\boldsymbol{M}} - \boldsymbol{M}_{2,\infty}$.

■ 5.15.2 Matrix Estimation using HSVT

Hard Singular Value Thresholding (HSVT) Map. We define the HSVT map. For any $q, p \in \mathbb{N}$, consider a matrix $\boldsymbol{B} \in \mathbb{R}^{q\times p}$ such that $\boldsymbol{B} = \sum_{i=1}^{q\wedge p} \sigma_i(\boldsymbol{B}) x_i y_i^T$. Here for $i \in [q \wedge p]$, $\sigma_i(\boldsymbol{B})$ is the ith largest singular value of $\boldsymbol{B}$ and x_i, y_i are the corresponding left and right singular vectors respectively. Then, for given any $\lambda > 0$, we define the map $\text{HSVT}_\lambda : \mathbb{R}^{q\times p} \to \mathbb{R}^{q\times p}$, which simply shaves off the singular values of the input matrix that are below the threshold λ. Precisely,

$$\text{HSVT}_\lambda(\boldsymbol{B}) = \sum_{i=1}^{q\wedge p} \sigma_i(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) x_i y_i^T.$$

Matrix Estimating using HSVT map. We define a matrix estimation method using the HSVT map that is utilized by mSSA for imputation. Precisely, we estimate $\boldsymbol{M}$ from $\boldsymbol{Y}$ as follows: given parameter $k \geq 1$,

$$\widehat{\boldsymbol{M}} = \frac{1}{\hat{\rho}} \text{HSVT}_{\lambda_k}(\boldsymbol{Y}). \tag{5.16}$$

where $\lambda_k = \sigma_k(\boldsymbol{Y})$, i,e. the kth largest singular value of $\boldsymbol{Y}$.

■ 5.15.3 A Useful Linear Operator

We define a linear map associated to HSVT. For a specific choice of $\lambda \geq 0$, define $\varphi_\lambda^{\boldsymbol{B}} : \mathbb{R}^p \to \mathbb{R}^p$ as follows: for any vector $w \in \mathbb{R}^p$ (i.e. $w \in \mathbb{R}^{p\times 1}$),

$$\varphi_\lambda^{\boldsymbol{B}}(w) = \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^T w. \tag{5.17}$$

Note that $\varphi_\lambda^{\boldsymbol{B}}$ is a linear operator and it depends on the tuple $(\boldsymbol{B}, \lambda)$; more precisely, the singular values and the right singular vectors of $\boldsymbol{B}$, as well as the threshold λ. If $\lambda = 0$, then we will adopt the shorthand notation: $\varphi^{\boldsymbol{B}} = \varphi_0^{\boldsymbol{B}}$. The following is a simple, but curious relationship between $\varphi_\lambda^{\boldsymbol{B}}$ and HSVT_λ that will be useful subsequently.

Lemma 5.15.1 (Lemma 35 of Agarwal et al. (2019b, 2021e)). *Let $\boldsymbol{B} \in \mathbb{R}^{q\times p}$ and $\lambda \geq 0$ be given. Then for any $j \in [q]$,*

$$\varphi_{\lambda}^{\boldsymbol{B}}(\boldsymbol{B}_{j\cdot}^{T}) = \mathrm{HSVT}_{\lambda}(\boldsymbol{B})_{j\cdot}^{T},$$

where $\boldsymbol{B}_{j\cdot} \in \mathbb{R}^{1\times p}$ represents the jth row of $\boldsymbol{B}$, and $\mathrm{HSVT}_{\lambda}(\boldsymbol{B})_{j\cdot} \in \mathbb{R}^{1\times p}$ represents the jth row of the matrix obtained after applying HSVT over $\boldsymbol{B}$ with threshold λ.

Proof. By (5.17), the orthonormality of the right singular vectors and noting $\boldsymbol{B}_{j\cdot}^{T} = \boldsymbol{B}^{T} e_j$ with $e_j \in \mathbb{R}^{p}$ with jth entry 1 and everything else 0, we have

$$\begin{aligned}
\varphi_{\lambda}^{\boldsymbol{B}}(\boldsymbol{B}_{j\cdot}^{T}) &= \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^{T} \boldsymbol{B}_{j\cdot}^{T} = \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^{T} \boldsymbol{B}^{T} e_j \\
&= \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^{T} (\sum_{i'=1}^{q\wedge p} \sigma_{i'}(\boldsymbol{B}) x_{i'} y_{i'}^{T})^{T} e_j = \sum_{i,i'=1}^{q\wedge p} \sigma_{i'}(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^{T} y_{i'} x_{i'}^{T} e_j \\
&= \sum_{i,i'=1}^{q\wedge p} \sigma_{i'}(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i \delta_{ii'} x_{i'}^{T} e_j = \sum_{i=1}^{q\wedge p} \sigma_i(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i x_i^{T} e_j \\
&= HSVT_{\lambda}(\boldsymbol{B})^{T} e_j = HSVT_{\lambda}(\boldsymbol{B})_{j\cdot}^{T}.
\end{aligned}$$

■

■ 5.15.4 HSVT based Matrix Estimation: A Deterministic Bound

We state the following result about property of the estimator.

Lemma 5.15.2. *For $k \geq 1$, let $\boldsymbol{M} = \boldsymbol{M}_k + \boldsymbol{E}_k$ with rank$(\boldsymbol{M}_k) = k$. Let $\varepsilon = \max(\hat{\rho}/\rho, \rho/\hat{\rho}) \geq 1$. Then, the HSVT estimate $\widehat{\boldsymbol{M}}$ with parameter k is such that for all $j \in [q]$,*

$$\begin{aligned}
\widehat{\boldsymbol{M}}_{j\cdot}^{T} - \boldsymbol{M}_{j\cdot\,2}^{T\,2} \leq \frac{2\boldsymbol{Y} - \rho\boldsymbol{M}_2^2 + 2\rho^2\boldsymbol{E}_{k\,2}^{\;2}}{(\sigma_k(\rho\boldsymbol{M}_k))^2}\Big(2\Big\|[\boldsymbol{M}_k]_{j\cdot}^{T}\Big\|_2^2 + \frac{4\varepsilon^2(\boldsymbol{Y}_{j\cdot}^{T} - \rho\boldsymbol{M}_{j\cdot\,2}^{T})^2}{\rho^2}\Big) \\
+ \frac{4\varepsilon^2}{\rho^2}\Big\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^{T} - \rho\boldsymbol{M}_{j\cdot}^{T})\Big\|_2^2 + 2(\varepsilon-1)^2\boldsymbol{M}_{j\cdot\,2}^{T\,2} + 2\Big\|[\boldsymbol{E}_k]_{j\cdot}^{T}\Big\|_2^2. \quad (5.18)
\end{aligned}$$

Proof. We prove our lemma in four steps.

Step 1. Decomposing $\widehat{M}_{j\cdot}^T - M_{j\cdot}^T$ in two terms. Fix a row index $j \in [q]$. Let λ_k be the kth largest singular value of $\boldsymbol{Y}$, as used by HSVT algorithm with parameter $k \geq 1$.

$$\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T = \left(\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\right) + \left(\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) - \boldsymbol{M}_{j\cdot}^T\right).$$

By definition per (5.17), $\varphi_{\lambda_k}^{\boldsymbol{Y}} : \mathbb{R}^p \to \mathbb{R}^p$ is the projection operator onto $\text{span}\{u_1, \ldots, u_k\}$, the span of top k right singular vectors of $\boldsymbol{Y}$, denoted as $u_1, \ldots, u_k$. Therefore,

$$\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) - \boldsymbol{M}_{j\cdot}^T \in \text{span}\{u_1, \ldots, u_k\}^{\perp}.$$

By design, $\text{rank}(\widehat{\boldsymbol{M}}) = k$. Therefore, by Lemma 5.15.1

$$\widehat{\boldsymbol{M}}_{j\cdot} - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) = \frac{1}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T) - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) \in \text{span}\{u_1, \ldots, u_k\}.$$

Therefore, $\langle \widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T), \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) - \boldsymbol{M}_{j\cdot}^T \rangle = 0$, and hence

$$\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\right\|_2^2 = \left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\right\|_2^2 + \left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) - \boldsymbol{M}_{j\cdot}^T\right\|_2^2 \tag{5.19}$$

by the Pythagorean theorem.

Step 2. Bounding Term 1. Term 1 is $\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\right\|_2$ We begin by bounding the first term on the right hand side of (5.19). By Lemma 5.15.1,

$$\begin{aligned}\widehat{\boldsymbol{M}}_{j\cdot} - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) &= \frac{1}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T) - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) = \varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\frac{1}{\hat{\rho}}\boldsymbol{Y}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\right) \\ &= \frac{1}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T) + \frac{\rho - \hat{\rho}}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T).\end{aligned}$$

Using the Parallelogram Law (or, equivalently, combining Cauchy-Schwartz and AM-GM inequalities), we obtain

$$\begin{aligned}\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot})^T\right\|_2^2 &= \left\|\frac{1}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T) + \frac{\rho - \hat{\rho}}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\right\|_2^2 \\ &\leq 2\left\|\frac{1}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T)\right\|_2^2 + 2\left\|\frac{\rho - \hat{\rho}}{\hat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\right\|_2^2 \\ &\leq \frac{2}{\hat{\rho}^2}\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T)\right\|_2^2 + 2\left(\frac{\rho - \hat{\rho}}{\hat{\rho}}\right)^2 \boldsymbol{M}_{j\cdot\,2}^{T\,2}\end{aligned}$$

$$\leq \frac{2\varepsilon^2}{\rho^2} \left\| \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) \right\|_2^2 + 2(\varepsilon - 1)^2 \boldsymbol{M}_{j\cdot}^T{}_2^2. \tag{5.20}$$

From definition of ε, $\frac{1}{\widehat{\rho}} \leq \frac{\varepsilon}{\rho}$ and $\left(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\right)^2 \leq (\varepsilon - 1)^2$. The first term of (5.20) can be decomposed as,

$$\begin{aligned} &\left\| \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) \right\|_2^2 \\ &\leq 2 \left\| \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) - \varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) \right\|_2^2 + 2 \left\| \varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) \right\|_2^2. \end{aligned} \tag{5.21}$$

In above, we have used notation $\varphi^{\boldsymbol{M}_k} = \varphi_0^{\boldsymbol{M}_k}$. Given that $\boldsymbol{M}_k$ is rank k matrix, $\varphi^{\boldsymbol{M}_k} : \mathbb{R}^p \to \mathbb{R}^p$ is the projection operator mapping any element in $\mathbb{R}^p$ to the projection onto the subspace spanned by $\{\mu_1, \dots, \mu_k\}$, where $\mu_1, \dots, \mu_k \in \mathbb{R}^p$ are the k non-trivial right singular vectors of $\boldsymbol{M}_k$. Similarly, by definition $\varphi_{\lambda_k}^{\boldsymbol{Y}}$ is a map $\mathbb{R}^p \to \mathbb{R}^p$ mapping any element in $\mathbb{R}^p$ to its projection onto the subspace spanned by $\{u_1, \dots, u_k\}$, the top k right singular vectors of $\boldsymbol{Y}$–this can be seen by noting $\lambda_k = \sigma_k(\boldsymbol{Y})$ is the k-th top singular value of $\boldsymbol{Y}$. Recall $\sigma_j(\boldsymbol{Y})$, $j \in [q \wedge p]$ is the jth largest singular value of $\boldsymbol{Y}$.

Next, we bound the first term on the right hand side of (5.21). To that end, by Wedin $\sin\Theta$ Theorem (see Davis and Kahan (1970); Wedin (1972)) and recalling rank($\boldsymbol{M}_k$) $= k$,

$$\begin{aligned} \left\| \varphi_{\lambda_k}^{\boldsymbol{Y}} - \varphi^{\boldsymbol{M}_k} \right\|_2 &\leq \frac{\boldsymbol{Y} - \rho \boldsymbol{M}_k{}_2}{\sigma_k(\rho \boldsymbol{M}_k)} \\ &\leq \frac{\boldsymbol{Y} - \rho \boldsymbol{M}_2}{\sigma_k(\rho \boldsymbol{M}_k)} + \frac{\rho \boldsymbol{M} - \boldsymbol{M}_k{}_2}{\sigma_k(\rho \boldsymbol{M}_k)} \\ &\leq \frac{\boldsymbol{Y} - \rho \boldsymbol{M}_2}{\sigma_k(\rho \boldsymbol{M}_k)} + \frac{\rho \boldsymbol{E}_k{}_2}{\sigma_k(\rho \boldsymbol{M}_k)}. \end{aligned} \tag{5.22}$$

Then it follows that

$$\begin{aligned} \left\| \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) - \varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T) \right\|_2 &\leq \varphi_{\lambda_k}^{\boldsymbol{Y}} - \varphi^{\boldsymbol{M}_k}{}_2 \boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T{}_2 \\ &\leq \frac{(\boldsymbol{Y} - \rho \boldsymbol{M}_2 + \rho \boldsymbol{E}_k{}_2)(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T{}_2)}{\sigma_k(\rho \boldsymbol{M}_k)}. \end{aligned} \tag{5.23}$$

Using (5.21) and (5.23) in (5.20),

$$\left\| \widehat{\boldsymbol{M}}_{j\cdot} - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) \right\|_2^2 \leq \frac{4\varepsilon^2}{\rho^2} \frac{(\boldsymbol{Y} - \rho \boldsymbol{M}_2 + \rho \boldsymbol{E}_k{}_2)^2 (\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T{}_2)^2}{(\sigma_k(\rho \boldsymbol{M}_k))^2}$$

$$+\frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T-\rho\boldsymbol{M}_{j\cdot}^T)\right\|_2^2+2(\varepsilon-1)^2\boldsymbol{M}^{T\,2}_{j\cdot\,2}. \tag{5.24}$$

Step 3. Bounding Term 2. Term 2 is $\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)-\boldsymbol{M}_{j\cdot}^T\right\|_2^2$ Recall $\boldsymbol{M}=\boldsymbol{M}_k+\boldsymbol{E}_k$ and using (5.22),

$$\begin{aligned}\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)-\boldsymbol{M}_{j\cdot}^T\right\|_2^2&=\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}([\boldsymbol{M}_k]_{j\cdot}^T+[\boldsymbol{E}_k]_{j\cdot}^T)-[\boldsymbol{M}_k]_{j\cdot}^T-[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2\\&\le2\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}([\boldsymbol{M}_k]_{j\cdot}^T)-[\boldsymbol{M}_k]_{j\cdot}^T\right\|_2^2+2\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}([\boldsymbol{E}_k]_{j\cdot}^T)-[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2\\&=2\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}([\boldsymbol{M}_k]_{j\cdot}^T)-\varphi_{\lambda_k}^{\boldsymbol{M}_k}([\boldsymbol{M}_k]_{j\cdot}^T)\right\|_2^2+2\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}([\boldsymbol{E}_k]_{j\cdot}^T)-[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2\\&\le2\left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}-\varphi_{\lambda_k}^{\boldsymbol{M}_k}\right\|_2^2\left\|[\boldsymbol{M}_k]_{j\cdot}^T\right\|_2^2+2\left\|[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2\\&\le2\frac{(\boldsymbol{Y}-\rho\boldsymbol{M}_2+\rho\boldsymbol{E}_k)^2}{(\sigma_k(\rho\boldsymbol{M}_k))^2}\left\|[\boldsymbol{M}_k]_{j\cdot}^T\right\|_2^2+2\left\|[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2.\end{aligned} \tag{5.25}$$

Step 4. Putting everything together. Inserting (5.24) and (5.25) back to (5.19), we have that for each $j\in[q]$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T-\boldsymbol{M}_{j\cdot}^T\right\|_2^2&\le2\frac{(\boldsymbol{Y}-\rho\boldsymbol{M}_2+\rho\boldsymbol{E}_{k\,2})^2}{(\sigma_k(\rho\boldsymbol{M}_k))^2}\left\|[\boldsymbol{M}_k]_{j\cdot}^T\right\|_2^2+2\left\|[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2\\&\quad+\frac{4\varepsilon^2}{\rho^2}\frac{(\boldsymbol{Y}-\rho\boldsymbol{M}_2+\rho\boldsymbol{E}_{k\,2})^2(\boldsymbol{Y}_{j\cdot}^T-\rho\boldsymbol{M}_{j\cdot\,2}^T)^2}{(\sigma_k(\rho\boldsymbol{M}_k))^2}\\&\quad+\frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T-\rho\boldsymbol{M}_{j\cdot}^T)\right\|_2^2+2(\varepsilon-1)^2\boldsymbol{M}^{T\,2}_{j\cdot\,2}\\&\le\frac{2\boldsymbol{Y}-\rho\boldsymbol{M}_2^2+2\rho^2\boldsymbol{E}_{k\,2}^{\,2}}{(\sigma_k(\rho\boldsymbol{M}_k))^2}\left(2\left\|[\boldsymbol{M}_k]_{j\cdot}^T\right\|_2^2+\frac{4\varepsilon^2(\boldsymbol{Y}_{j\cdot}^T-\rho\boldsymbol{M}_{j\cdot\,2}^T)^2}{\rho^2}\right)\\&\quad+\frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T-\rho\boldsymbol{M}_{j\cdot}^T)\right\|_2^2+2(\varepsilon-1)^2\boldsymbol{M}^{T\,2}_{j\cdot\,2}+2\left\|[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2,\end{aligned}$$

where we used $(a+b)^2\le2a^2+2b^2$. This completes the proof. ■

5.15.5 HSVT based Matrix Estimation: Deterministic To High-Probability

Next, we convert the bound obtained in Lemma 5.15.2 to a bound in expectation (as well as one in high-probability) for our metric of interest: $\widehat{\boldsymbol{M}} - \boldsymbol{M}_{2,\infty}$. In particular, we establish

Theorem 5.15.1. *For $k \geq 1$, let $\boldsymbol{M} = \boldsymbol{M}_k + \boldsymbol{E}_k$ with rank$(\boldsymbol{M}_k) = k$. Let $\epsilon = \boldsymbol{E}_{k\infty}$ and $\Gamma = \boldsymbol{M}_{k\infty}$. Let $\rho \geq C\log(qp)/q$ for C large enough and $q \leq p$. Then, the HSVT estimate $\widehat{\boldsymbol{M}}$ with parameter k is such that*

$$\mathbb{E}[\max_{j\in[q]} \frac{1}{p}\widehat{\boldsymbol{M}}_{j\cdot}^{T} - \boldsymbol{M}_{j\cdot}^{T}{}_{2}^{2}] \leq \frac{p(C\sigma^2 + \rho^2\epsilon q)}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\left(\Gamma^2 + \frac{\sigma^2}{\rho^2}\right) + \frac{C\sigma^2 k\log p}{p\rho^2} + \frac{C(\Gamma+\epsilon)^2}{p} + 2\epsilon^2 + \frac{C}{(pq)^2}.$$

Proof. We start by identifying certain high probability events. Subsequently, using these events and Lemma 5.15.2, we shall conclude the proof.

High Probability Events. For some positive absolute constant $C > 0$, define

$$E_1 := \Big\{|\hat{\rho} - \rho| \leq \rho/20\Big\},$$

$$E_2 := \Big\{\big\|\boldsymbol{Y} - \rho\boldsymbol{M}\big\|_2 \leq C\sigma\sqrt{p}\Big\}, \tag{5.26}$$

$$E_3 := \Big\{\big\|\boldsymbol{Y} - \rho\boldsymbol{M}\big\|_{\infty,2}, \big\|\boldsymbol{Y} - \rho\boldsymbol{M}\big\|_{2,\infty} \leq C\sigma\sqrt{p}\Big\}, \tag{5.27}$$

$$E_4 := \Big\{\max_{j\in[q]}\Big\|\varphi^{\boldsymbol{B}}_{\sigma_k(\boldsymbol{B})}\Big(\boldsymbol{Y}_{j\cdot}^{T} - \rho\boldsymbol{M}_{j\cdot}^{T}\Big)\Big\|_2^2 \leq C\sigma^2 k\log(p)\Big\}, \tag{5.28}$$

$$E_5 := \left\{\left(1 - \sqrt{\frac{20\log(qp)}{\rho qp}}\right)\rho \leq \hat{\rho} \leq \frac{1}{1 - \sqrt{\frac{20\log(qp)}{\rho qp}}}\rho\right\}.$$

In (5.28) above, $\boldsymbol{B} \in \mathbb{R}^{q\times p}$ is a deterministic matrix. Let the singular value decomposition of $\boldsymbol{B}$ be given as $\boldsymbol{B} = \sum_{i=1}^{q}\sigma_i(\boldsymbol{B})x_i y_i^T$, where $\sigma_i(\boldsymbol{B})$ are the singular vectors of $\boldsymbol{B}$ in decreasing order and x_i, y_i are the left and right singular vectors respectively. Recall the definition of $\varphi^{\boldsymbol{B}}_{\lambda}$ in (5.17). In particular, we choose $\lambda = \sigma_k(\boldsymbol{B})$, the kth singular value of $\boldsymbol{B}$ in (5.28). As a result, in effect, we are bounding norm of projection of random vector $\boldsymbol{Y}_{j\cdot} - \rho\boldsymbol{M}_{j\cdot}$ for any given deterministic subspace of $\mathbb{R}^p$ of dimension k.

Lemma 5.15.3. *For some positive constant $c_1 > 0$ and $C > 0$ large enough in definitions of $E_1, \dots, E_5$,*

$$\begin{aligned}
\mathbb{P}(E_1) &\geq 1 - 2e^{-c_1 pq\rho} - (1-\rho)^{pq}, \\
\mathbb{P}(E_2) &\geq 1 - 2e^{-p}, \\
\mathbb{P}(E_3) &\geq 1 - 2e^{-p}, \\
\mathbb{P}(E_4) &\geq 1 - \frac{2}{(qp)^{10}}. \\
\mathbb{P}(E_5) &\geq 1 - \frac{2}{(qp)^{10}}.
\end{aligned} \tag{5.29}$$

Proof. We bound the probability of events $E_1, \dots, E_5$ in that order.

Bounding E_1. Let

$$\hat{\rho}_0 = (\sum_{i=1}^{q} \sum_{j=1}^{p} \mathbf{1}(Y_{ij} \text{ is obs.}))/(q\ p).$$

That is, $\hat{\rho} = \max(\hat{\rho}_0, 1/(pq))$ and $\mathbb{E}[\hat{\rho}_0] = \rho$. We define the event $E_6 := \{\hat{\rho}_0 = \hat{\rho}\}$. Thus, we have that

$$\begin{aligned}
\mathbb{P}(E_1^c) &= \mathbb{P}(E_1^c \cap E_6) + \mathbb{P}(E_1^c \cap E_6^c) \\
&= \mathbb{P}(|\hat{\rho}_0 - \rho| \geq \rho/20) + \mathbb{P}(E_1^c \cap E_6^c) \\
&\leq \mathbb{P}(|\hat{\rho}_0 - \rho| \geq \rho/20) + \mathbb{P}(E_6^c) \\
&= \mathbb{P}(|\hat{\rho}_0 - \rho| \geq \rho/20) + (1-\rho)^{qp},
\end{aligned}$$

where the final equality follows by the independence of observations assumption and the fact that $\hat{\rho}_0 \neq \hat{\rho}$ only if we do not have any observations. By Bernstein's Inequality, we have that

$$\mathbb{P}(|\hat{\rho}_0 - \rho| \geq \rho/20) \geq 1 - 2e^{-c_1 \rho q p}.$$

Bounding E_2. To start with, $\mathbb{E}[\boldsymbol{Y}] = \rho \boldsymbol{M}$. For any $i \in [q], j \in [p]$, the Y_{ij} are independent, 0 with probability $1-\rho$ and with probability ρ equal to $M_{ij} + \boldsymbol{\varepsilon}_{ij}$ with $\left\|\boldsymbol{\varepsilon}_{ij}\right\|_{\psi_2} \leq \sigma$. Therefore, it follows that $Y_{ij} - \rho M_{ij\psi_2} \leq C'\sigma$ for a constant $C' > 0$. Since $q \leq p$, using

Theorem 5.14.2 it follows that for an appropriately large constant $C > 0$,

$$\mathbb{P}(E_2) \geq 1 - 2e^{-p}.$$

Bounding E_3. Recall that we assume $q \leq p$. Observe that for any matrix $A \in \mathbb{R}^{q \times p}$, $A_{\infty,2}$, $A_{2,\infty} \leq A_2$. Thus using the argument to bound $\boldsymbol{E}_2$, we have (5.29).

Bounding E_4. Consider for $j \in [q]$,

$$\left\|\varphi^{\boldsymbol{B}}_{\sigma_k(\boldsymbol{B})}\left(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T\right)\right\|_2^2 = \sum_{i=1}^k \left\|y_i y_i^T(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T)\right\|_2^2 \leq \sum_{i=1}^k \left(y_i^T(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T)\right)_2^2 = \sum_{i=1}^k Z_i^2,$$

where $Z_i = y_i^T(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T)$. By definition of the ψ_2 norm of a random variable and since y_i is unit norm vector that is deterministic (and hence independent the of random vector $\boldsymbol{Y}_{j\cdot}^T - p\boldsymbol{M}_{j\cdot}^T$), it follows that

$$\left\|Z_i\right\|_{\psi_2} = \left\|y_i^T(\boldsymbol{Y}_{j\cdot} - p\boldsymbol{M}_{j\cdot})\right\|_{\psi_2} \leq \left\|(\boldsymbol{Y}_{j\cdot} - p\boldsymbol{M}_{j\cdot})\right\|_{\psi_2}.$$

Since the coordinates of $\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T$ are mean-zero and independent, with ψ_2 norm bounded by $\sqrt{C}\sigma$ for some absolute constant $C > 0$, using arguments from Agarwal et al. (2019b, 2021e), it follows that

$$\mathbb{P}\Big(\sum_{i=1}^k Z_i^2 > t\Big) \leq 2k \exp\Big(-\frac{t}{kC\sigma^2}\Big).$$

Therefore, for choice of $t = C\sigma^2 k \log p$ with large enough constant $C > 0$, $q \leq p$, and taking a union bound over all $j \in [p]$, we have that

$$\mathbb{P}\left(E_4^c\right) \leq \frac{2}{(qp)^{10}}.$$

Bounding E_5. Recall the definition of $\hat{\rho}$. By the binomial Chernoff bound, for $\varepsilon > 1$,

$$\mathbb{P}\Big(\hat{\rho} > \varepsilon\rho\Big) \leq \exp\left(-\frac{(\varepsilon-1)^2}{\varepsilon+1}qp\rho\right), \quad \text{and}$$
$$\mathbb{P}\Big(\hat{\rho} < \frac{1}{\varepsilon}\rho\Big) \leq \exp\left(-\frac{(\varepsilon-1)^2}{2\varepsilon^2}qp\rho\right).$$

By the union bound,

$$\mathbb{P}\Big(\frac{1}{\varepsilon}\rho \leq \hat{\rho} \leq \rho\varepsilon\Big) \geq 1 - \mathbb{P}\Big(\hat{\rho} > \varepsilon\rho\Big) - \mathbb{P}\Big(\hat{\rho} < \frac{1}{\varepsilon}\rho\Big).$$

Noticing $\varepsilon + 1 < 2\varepsilon < 2\varepsilon^2$ for all $\varepsilon > 1$, and substituting $\varepsilon = \Big(1 - \sqrt{\frac{20\log(qp)}{qp\rho}}\Big)^{-1}$ completes the proof. ■

The following are immediate corollaries of the above stated bounds.

Corollary 5.15.1. *Let $E := E_1 \cap E_2$. Then, for $\rho \geq C\log(qp)/q$,*

$$\mathbb{P}(E^c) \leq C_1 e^{-c_2 p},$$

where C_1 and c_2 are positive constants.

Corollary 5.15.2. *Let $E := E_2 \cap E_3 \cap E_4 \cap E_5$. Then,*

$$\mathbb{P}(E^c) \leq \frac{C_1}{(qp)^{10}},$$

where C_1 is an absolute positive constant.

Probabilistic Bound for HSVT based Matrix Estimation. Recall $\epsilon = \boldsymbol{E}_{k\infty}$. Then $\boldsymbol{E}_{k}{}^2_F \leq \epsilon qp$. And $\boldsymbol{E}_{k}{}^2_2 \leq \boldsymbol{E}_{k}{}^2_F \leq \epsilon qp$. Let $\rho \geq C\log(qp)/q$ for C large enough and recall $q \leq p$. Further, recall $\Gamma = \boldsymbol{M}_{k\infty}$; thus, $\boldsymbol{M}_{\infty} \leq \Gamma + \epsilon$. Then $[\boldsymbol{M}_k]^T_{j\cdot}{}_2 \leq \Gamma\sqrt{p}$ and $[\boldsymbol{M}]^T_{j\cdot}{}_2 \leq (\Gamma + \epsilon)\sqrt{p}$.

Define $E = E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5$. Then, from Corollaries 5.15.1 and 5.15.2, we have that $\mathbb{P}(E^c) \leq \frac{C_1}{(qp)^{10}}$ for large enough constant $C_1 > 0$.

Under E_5, we have $\varepsilon = \max(\hat{\rho}/\rho, \rho/\hat{\rho}) \leq \Big(1 - \sqrt{\frac{20\log(qp)}{qp\rho}}\Big)^{-1}$. Under this choice of

ε and using $\rho \geq C \log(qp)/q$, we have that for C large enough, $\varepsilon \leq C$ and $(\varepsilon - 1)^2 \leq C/p$.

Given this setup, under event E, Lemma 5.15.2 leads to the following: for all $j \in [q]$ and with appropriately (re-defined) large enough constant $C > 0$,

$$\begin{aligned}\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot\,2}^{T\,2} \leq C\frac{\sigma^2 p + \rho^2 \epsilon q p}{\rho^2 \sigma_k(\boldsymbol{M}_k)^2}\Big(p\Gamma^2 + \frac{\sigma^2 p}{\rho^2}\Big) \\ + \frac{C\sigma^2 k \log p}{\rho^2} + C(\Gamma + \epsilon)^2 + 2p\epsilon^2. \end{aligned} \tag{5.30}$$

That is, under event E,

$$\max_{j\in[q]} \frac{1}{p}\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot\,2}^{T\,2} \leq C\frac{p(\sigma^2 + \rho^2\epsilon q)}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\Big(\Gamma^2 + \frac{\sigma^2}{\rho^2}\Big) + \frac{C\sigma^2 k \log p}{p\rho^2} + \frac{C(\Gamma+\epsilon)^2}{p} + 2\epsilon^2 \tag{5.31}$$

For any random variable X and event A, such that under event A, $X \leq B$ and $\mathbb{P}(A^c) \leq \delta$, we have

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbb{1}(A)] + \mathbb{E}[X\mathbb{1}(A^c)] \\ &\leq \mathbb{E}[X\mathbb{1}(A)] + \mathbb{E}[X^2]^{\frac{1}{2}}\mathbb{P}(A^c)^{\frac{1}{2}} \\ &\leq B + \mathbb{E}[X^2]^{\frac{1}{2}}\delta^{\frac{1}{2}}. \end{aligned} \tag{5.32}$$

We shall use this reasoning above to bound $\mathbb{E}[\max_{j\in[q]} \frac{1}{p}\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot\,2}^{T\,2}]$: let $X = \max_{j\in[q]} \frac{1}{p}\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot\,2}^{T\,2}$ and $A = E$; B is given by right hand side of (5.31), $\delta = \frac{C_1}{(qp)^{10}}$; the only missing quantity that remains to be bounded is $\mathbb{E}[X^2]$. We do that next.

To begin with, for any $j \in [q]$,

$$\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\right\|_2 \leq \left\|\widehat{\boldsymbol{M}}_{j\cdot}^T\right\|_2 + \left\|\boldsymbol{M}_{j\cdot}^T\right\|_2 \tag{5.33}$$

by triangle inequality. As stated earlier, $[\boldsymbol{M}]_{j\cdot\,2}^T \leq (\Gamma + \epsilon)\sqrt{p}$. Next, we bound $\left\|\widehat{\boldsymbol{M}}_{j\cdot}\right\|_2^T$. From (5.16), the fact that $\hat{\rho} \geq 1/(qp)$, and Lemma 5.15.1, we have

$$\begin{aligned}\widehat{\boldsymbol{M}}_{j\cdot\,2}^T &= \frac{1}{\hat{\rho}}\mathrm{HSVT}_{\lambda_k}(\boldsymbol{Y})_{j\cdot\,2}^T \\ &\leq q\;p\phi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T)_2 \\ &\leq q\;p\phi_{\lambda_k\,2}^{\boldsymbol{Y}}\boldsymbol{Y}_{j\cdot\,2}^T \end{aligned}$$

$$\leq q \;\; p\boldsymbol{Y}_{j\cdot}^{T}{}_{2}, \tag{5.34}$$

where we used the fact that $\phi_{\lambda_k}^{\boldsymbol{Y}}$ is a projection operator and hence $\phi_{\lambda_k}^{\boldsymbol{Y}}{}_2 = 1$. Note that $Y_{ij} = B_{ij} \text{ x } (M_{ij} + \boldsymbol{\varepsilon}_{ij})$, where B_{ij} is an independent Bernoulli variable with $\mathbb{P}(B_{ij} = 1) = \rho$ representing whether $(M_{ij} + \boldsymbol{\varepsilon}_{ij}$ is observed or not. Therefore, $|Y_{ij}| = |B_{ij}| \text{ x } |M_{ij} + \boldsymbol{\varepsilon}_{ij}| \leq (\Gamma + \epsilon) + |\boldsymbol{\varepsilon}_{ij}|$. Therefore, from (5.33) and (5.34),

$$\begin{aligned}
\max_{j\in[q]} \left\| \widehat{\boldsymbol{M}}_{j\cdot}^{T} - \boldsymbol{M}_{j\cdot}^{T} \right\|_2 &\leq (\Gamma + \epsilon)\sqrt{p} + qp(\max_{j\in[q]} \boldsymbol{Y}_{j\cdot}^{T}{}_2) \\
&\leq (\Gamma + \epsilon)\sqrt{p} + qp \text{ x } \sqrt{p}(\max_{i\in[p],j\in[q]} |Y_{ij}|) \\
&\leq 2qp^{\frac{3}{2}}(\Gamma + \epsilon + \max_{i\in[p],j\in[q]} |\boldsymbol{\varepsilon}_{ij}|).
\end{aligned} \tag{5.35}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ twice, we have $(a+b)^4 \leq 8(a^4 + b^4)$. Therefore, from (5.36)

$$\max_{j\in[q]} \left\| \widehat{\boldsymbol{M}}_{j\cdot}^{T} - \boldsymbol{M}_{j\cdot}^{T} \right\|_2^4 \leq 16q^4p^6((\Gamma + \epsilon)^4 + \max_{i\in[p],j\in[q]} |\boldsymbol{\varepsilon}_{ij}|^4). \tag{5.36}$$

Recall $\mathbb{E}[\boldsymbol{\varepsilon}_{ij}] = 0$, $\left\| \boldsymbol{\varepsilon}_{ij} \right\|_{\psi_2} \leq \sigma$ and $\boldsymbol{\varepsilon}_{ij}$ are independent across i, j. A property of ψ_2-random variables is that $|\eta_{ij}|^{\theta}$ is a $\psi_{2/\theta}$-random variable for $\theta \geq 1$. With choice of $\theta = 4$, we have

$$\mathbb{E}[\max_{ij} \left| \boldsymbol{\varepsilon}_{ij} \right|^4] \leq C'\sigma^4 \log^2(qp), \tag{5.37}$$

for some $C' > 0$ by Lemma 5.14.1. From (5.34), (5.36), and (5.37), we have that

$$\left(\mathbb{E}[\max_{j\in[q]} \frac{1}{p^2} \left\| \widehat{\boldsymbol{M}}_{j\cdot}^{T} - \boldsymbol{M}_{j\cdot}^{T} \right\|_2^4] \right)^{\frac{1}{2}} \leq 4q^2p^2((\Gamma + \epsilon)^4 + C'\sigma^4 \log^2(qp))^{\frac{1}{2}}. \tag{5.38}$$

Finally, using (5.31), (5.32) and (5.38), we conclude

$$\mathbb{E}[\max_{j\in[q]} \frac{1}{p} \widehat{\boldsymbol{M}}_{j\cdot}^{T} - \boldsymbol{M}_{j\cdot\;2}^{T\;2}] \leq \frac{p(C\sigma^2 + \rho^2\epsilon q)}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\left(\Gamma^2 + \frac{\sigma^2}{\rho^2}\right) + \frac{C\sigma^2 k \log p}{p\rho^2} + \frac{C(\Gamma + \epsilon)^2}{p} + 2\epsilon^2 + \frac{C}{(pq)^2}.$$

This completes the proof of Theorem 5.15.1. ■

■ 5.16 Proof of Theorem 5.5.1

The proof of Theorem 5.5.1 will utilize Theorem 5.15.1. To begin with, given N time series with observations over $[T]$, the mSSA algorithm as described in Section 5.1.1 constructs the $L \times (NT/L)$ stacked page matrix $\text{SP}((X_1, \dots, X_N), T, L)$ with $L = \sqrt{\min(N,T)T}$, i.e. $L \leq T$.

As per the model described by (5.1) and Section 5.3, it follows that each entry of $\text{SP}((X_1, \dots, X_N), T, L)$ is an independent random variable; it is observed with probability $\rho \in (0, 1]$ independently and when it is observed, its equal to value of the latent time series plus zero-mean sub-Gaussian noise. In particular,

$$\mathbb{E}[\text{SP}((X_1, \dots, X_N), T, L)] = \rho \text{SP}((f_1, \dots, f_N), T, L),$$

where $\text{SP}((f_1, \dots, f_N), T, L) \in \mathbb{R}^{L\times(NT/L)}$ with entry in row $\ell \in [L]$ and column $(n-1) \text{ x } T/L + j$ equal to $f_n(\ell + (j-1) \text{ x } L)$. Further, when entry in row $\ell \in [L]$ and column $(n-1) \text{ x } T/L + j$ in $\text{SP}((X_1, \dots, X_N), T, L)$ is observed, i.e. $X_n(\ell + (j-1) \text{ x } L) \neq \star$, it is equal to $f_n(\ell + (j-1) \text{ x } L) + \eta_n(\ell + (j-1) \text{ x } L)$ where $\eta_n(\cdot)$ are independent, zero-mean sub-Gaussian variables with $\eta_n(\cdot)_{\psi_2} \leq \gamma$ as per the Property 5.4.1.

Under Properties 5.3.1 and 5.5.1, as a direct implication of Proposition 15, $\text{SP}((f_1, \dots, f_N), T, L)$ has ϵ'-rank at most $R \text{ x } G$ with $\epsilon' = R\Gamma_1\epsilon$. That is, there exist rank $k \leq R \text{ x } G$ matrix $\boldsymbol{M}_k \in \mathbb{R}^{L\times(NT/L)}$ so that

$$\text{SP}((f_1, \dots, f_N), T, L) = \boldsymbol{M}_k + \boldsymbol{E}_k,$$

where $\boldsymbol{E}_{k\infty} \leq \epsilon'$. Due to Property 5.3.1, it follows that $\boldsymbol{M}_{k\infty} \leq R\Gamma_1\Gamma_2 + \epsilon'$. Under Property 5.5.2, we have $\sigma_k(\boldsymbol{M}_k) \geq c\sqrt{NT}/\sqrt{k}$ for some constant $c > 0$.

Define

$$\Gamma = R\Gamma_1\Gamma_2 + \epsilon' \ = \ R\Gamma_1(\Gamma_2 + \epsilon).$$

Recall from Section 5.1.1, the elements of the imputed multivariate time series are simply the entries of the matrix $\widehat{\text{SP}}((X_1, \dots, X_N), T, L)$ where $\widehat{\text{SP}}((X_1, \dots, X_N), T, L) = \frac{1}{\hat{\rho}}\text{HSVT}_k(\text{SP}((X_1, \dots, X_N), T, L))$. That is, imputation in mSSA is carried out by applying HSVT to the stacked page matrix $\text{SP}((X_1, \dots, X_N), T, L)$.

All in all, the above description precisely meets the setup of Theorem 5.15.1. To apply Theorem 5.15.1, we require $\rho \geq C \log(NT)/\sqrt{NT}$ for $C > 0$ large enough. Note that the number of columns in $\widehat{\text{SP}}((X_1, \dots, X_N), T, L)$ is equal to NT/L for $L = \sqrt{\min(N, T)T}$ – for this choice of L, note that $NT/L \geq L$. Using $\sigma_k^2(\boldsymbol{M}_k) \geq cNT/k$, for some absolute constant $= c \geq 0$, and using Theorem 5.15.1, we obtain

$$\begin{aligned}
&\mathbb{E}[\frac{1}{(NT/L)}\widehat{\text{SP}}((X_1, \dots, X_N), T, L) - \text{SP}((f_1, \dots, f_N), T, L)^2_{2,\infty}] \qquad (5.39)\\
&\leq \frac{k(NT/L)(C\gamma^2 + \rho^2\epsilon' L)}{\rho^2 c^2 NT}\left(\Gamma^2 + \frac{\gamma^2}{\rho^2}\right) + \frac{C\gamma^2 k \log NT}{(NT/L)\rho^2} + \frac{C(\Gamma + \epsilon')^2}{(NT/L)} + 2(\epsilon')^2 + \frac{C}{(NT)^2}
\end{aligned}$$

Recall that $k \leq R \,\mathrm{x}\, G$, $\epsilon' = R\Gamma_1\epsilon$, and $\Gamma = R\Gamma_1(\Gamma_2 + \epsilon)$. Hence, simplifying (5.39), we obtain that

$$\begin{aligned}
&\mathbb{E}[\frac{1}{(NT/L)}\widehat{\text{SP}}((X_1, \dots, X_N), T, L) - \text{SP}((f_1, \dots, f_N), T, L)^2_{2,\infty}]\\
&\leq \tilde{C}\left(\frac{RG(1 + \rho^2 R\epsilon L)}{\rho^2 L}\left(R^2(1 + \epsilon^2) + \frac{1}{\rho^2}\right) + \frac{RG \log NT}{(NT/L)\rho^2} + \frac{(R(1 + \epsilon))^2}{(NT/L)} + (R\epsilon)^2\right)\\
&\leq \tilde{C}\left(\frac{R^3 G \log NT}{\rho^4 L} + \frac{R^4 G(\epsilon + \epsilon^2 + \epsilon^3)}{\rho^2}\right), \qquad (5.40)
\end{aligned}$$

where $\tilde{C} = C(c, \Gamma_1, \Gamma_2, \gamma)$ is a positive constant dependent on model parameters including $\Gamma_1, \Gamma_2, \gamma$.

It can be easily verified that for any matrix, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,

$$\frac{1}{mn}\|\boldsymbol{A}\|_F^2 \leq \frac{1}{n}\|\boldsymbol{A}\|_{\infty,2}^2 . \qquad (5.41)$$

Further, there is a one-to-one mapping of $\hat{f}_n(\cdot)$ (resp. $f_n(\cdot)$) to the entries of $\widehat{\text{SP}}((X_1, \dots, X_N), T, L)$ (resp. $\text{SP}((f_1, \dots, f_N), T, L)$). Hence,

$$\text{ImpErr}(N, T) = \mathbb{E}[\frac{1}{NT}\widehat{\text{SP}}((X_1, \dots, X_N), T, L) - \text{SP}((f_1, \dots, f_N), T, L)^2_F] \qquad (5.42)$$

Therefore, from (5.40), (5.41), and (5.42) it follows that

$$\text{ImpErr}(N,T) \le C(c,\Gamma_1,\Gamma_2,\gamma)\left(\frac{R^3 G \log NT}{\rho^4 L} + \frac{R^4 G(\epsilon+\epsilon^2+\epsilon^3)}{\rho^2}\right)$$

This completes the proof of Theorem 5.5.1.

■ 5.17 Proof of Theorem 5.5.2

The forecasting algorithm, as described in Section 5.1.1, computes a linear model between the recent past and immediate future to forecast. We shall bound the forecasting error, ForErr(N,T,L) as defined in (5.6). We start with some setup and notations, followed by a key proposition that establishes the existence of a linear model under the setup of Theorem 5.5.2, and then conclude with a detailed analysis of noisy, mis-specified least-squares.

Setup, Notations. For $L \ge 1, k \ge 1$, for ease of notations, we define

- $\text{SP}(X) = \text{SP}((X_1,\dots,X_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\text{SP}(f) = \text{SP}((f_1,\dots,f_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\text{SP}'(X) \in \mathbb{R}^{(L-1)\times(NT/L)}$ as the top $L-1$ rows of $\text{SP}((X_1,\dots,X_N),T,L)$,
- $\text{SP}'(f) \in \mathbb{R}^{(L-1)\times(NT/L)}$ as the top $L-1$ rows of $\text{SP}((f_1,\dots,f_N),T,L)$.

It is worth noting that $\mathbb{E}[\text{SP}(X)] = \rho\text{SP}(f)$ and hence

$$\text{SP}_{L\cdot}(X)^T = \rho\text{SP}_{L\cdot}(f)^T + \eta, \tag{5.43}$$

where $\eta \in \mathbb{R}^{(NT)/L}$ is a random vector with each component being independent, zero-mean with its distribution given as: it is 0 with probability $1-\rho$ and with probability ρ, due to Property 5.4.1, it equals a zero-mean sub-Gaussian random variable with $\cdot_{\psi_2} \le \gamma$. Therefore, using arguments in Agarwal et al. (2019b, 2021e), each component of η is an independent, zero-mean random variable with $\cdot_{\psi_2}$ bounded above by $C'(\gamma^2 + R\Gamma_1\Gamma_2)$ for some absolute constant $C' > 0$. Let $K = C'(\gamma^2 + R\Gamma_1\Gamma_2)$ and hence each component of η has $\cdot_{\psi_2}$ bounded by K.

Now, recall that for forecasting, we first apply the imputation algorithm (i.e. HSVT) to $\mathrm{SP}((X_1, \dots, X_N), T, L)$ by replacing $\star$s, i.e. missing observations by 0 as well as setting all the entries in the last row equal to 0. Equivalently, the imputation algorithm is applied to $\mathrm{SP}'(X)$ after setting all missing values to 0. Let $\widehat{\mathrm{SP}'} \in \mathbb{R}^{L-1\times(NT/L)}$ be the estimate produced from the imputation algorithm applied to $\mathrm{SP}'(X)$. Under the setup of Theorem 5.5.1, by following arguments identical to that of Theorems 5.15.1 and 5.5.1–in particular, refer to (5.40)–it follows that by selecting the right choice of $k \le R \,\mathrm{x}\, G$, we have

$$\mathbb{E}\Big[\frac{1}{(NT/L)}\widehat{\mathrm{SP}'} - \mathrm{SP}'(f)^2_{2,\infty}\Big] \le \tilde{C}\left(\frac{R^3 G \log NT}{\rho^4 L} + \frac{R^4 G(\epsilon + \epsilon^2 + \epsilon^3)}{\rho^2}\right), \tag{5.44}$$

where $\tilde{C} = C(c, \Gamma_1, \Gamma_2, \gamma) > 0$ is a constant dependent on $c, \Gamma_1, \Gamma_2, \gamma$.

Now, the mSSA forecasting algorithm finds $\widehat{\beta} = \widehat{\beta}((X_1, \dots, X_N), TL; k)$, by solving the following Ordinary Least Squares (OLS):

$$\widehat{\beta} \in \text{minimize} \quad \frac{1}{\hat{\rho}}\mathrm{SP}(X)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^2_2 \quad \text{over} \quad \beta \in \mathbb{R}^{L-1}. \tag{5.45}$$

And subsequently, $\widehat{\mathrm{SP}'}^{T}\widehat{\beta}$ is used as the estimate for $\mathrm{SP}(f)_{L\cdot} \in \mathbb{R}^{NT/L}$, the Lth row of the latent $\mathrm{SP}(f)$. The goal is to bound the forecasting error $\mathrm{ForErr}(N, T, L)$, which is given by

$$\mathrm{ForErr}(N, T, L) = \mathbb{E}\Big[\frac{1}{(NT/L)}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^2_2\Big].$$

Therefore, our interest is in bounding $\mathbb{E}[\mathrm{SP}_{L\cdot}(f) - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^2_2]$.

Now, we recall from Proposition 11 that there exists $\beta^* \in \mathbb{R}^{L-1}$, such that

$$\mathrm{SP}(f)^T_{L\cdot} - \mathrm{SP}'(f)^T\beta^*{}_\infty \le C_2\epsilon,$$

where $C_2 := R\Gamma_1(1 + \beta^*{}_1)$.

Bounding $\mathbb{E}[\mathrm{SP}_{L\cdot}(f) - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^2_2]$. By (5.45) and (5.43)

$$\frac{1}{\hat{\rho}}\mathrm{SP}(X)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^2_2 \le \frac{1}{\hat{\rho}}\mathrm{SP}(X)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_2$$

$$
\begin{aligned}
&= \frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} + \eta - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2} \\
&= \frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2} + \eta^{2}_{2} + 2\eta^{T}(\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*}). \quad (5.46)
\end{aligned}
$$

Also,

$$
\begin{aligned}
\frac{1}{\hat{\rho}}\mathrm{SP}(X)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^{2}_{2} &= \frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} + \eta - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^{2}_{2} \\
&= \frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^{2}_{2} + \eta^{2}_{2} + 2\eta^{T}(\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}). \quad (5.47)
\end{aligned}
$$

From (5.46) and (5.47)

$$
\begin{aligned}
&\mathbb{E}[\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^{2}_{2}] \quad (5.48) \\
&\leq \mathbb{E}[\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2}] + 2\mathbb{E}[\eta^{T}\widehat{\mathrm{SP}'}^{T}(\beta^{*} - \widehat{\beta})]
\end{aligned}
$$

η is independent of $\widehat{\mathrm{SP}'}$, β^{*}, and $\hat{\rho}$; $\mathbb{E}[\eta] = \mathbf{0}$; thus, we have that

$$
\mathbb{E}[\eta^{T}\widehat{\mathrm{SP}'}^{T}\beta^{*}] = 0. \quad (5.49)
$$

By (5.45), we have $\widehat{\beta} = \widehat{\mathrm{SP}'}^{T,\dagger}\frac{1}{\hat{\rho}}\mathrm{SP}(X)_{L\cdot}$, where $\widehat{\mathrm{SP}'}^{T,\dagger}$ is pseudo-inverse of $\widehat{\mathrm{SP}'}^{T}$. That is,

$$
\widehat{\beta} = \widehat{\mathrm{SP}'}^{T,\dagger}\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} + \frac{1}{\hat{\rho}}\widehat{\mathrm{SP}'}^{T,\dagger}\eta. \quad (5.50)
$$

Using cyclic and linearity of Trace operator; the independence properties of η; and (5.50); we have

$$
\begin{aligned}
\mathbb{E}[\eta^{T}\widehat{\mathrm{SP}'}^{T}\widehat{\beta}] &= \mathbb{E}[\eta^{T}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot}] + \mathbb{E}[\frac{1}{\hat{\rho}}\eta^{T}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}\eta] \\
&= \mathbb{E}[\eta]^{T}\mathbb{E}[\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}\frac{\rho}{\hat{\rho}}]\mathrm{SP}(f)_{L\cdot} + \mathbb{E}[\frac{1}{\hat{\rho}}\,\mathrm{Tr}\left(\eta^{T}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}\eta\right)] \\
&= \mathbb{E}[\frac{1}{\hat{\rho}}\,\mathrm{Tr}\left(\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}\eta\eta^{T}\right)] \\
&= \mathrm{Tr}\,(\mathbb{E}[\frac{1}{\hat{\rho}}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}]\mathbb{E}[\eta\eta^{T}]) \\
&\leq \; C(\gamma)k/\rho, \quad (5.51)
\end{aligned}
$$

where $C(\gamma)$ is a function only of γ. To see the last inequality, we use various facts. First, by the definition of the HSVT algorithm $\widehat{\mathrm{SP}'}^{T}$ has rank at most k. Second, let $\widehat{\mathrm{SP}'}^{T} = \boldsymbol{USV}^{T}$ be the singular value decomposition of $\widehat{\mathrm{SP}'}^{T}$, we have

$$\begin{aligned}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger} &= \boldsymbol{USV}^{T}\boldsymbol{VS}^{\dagger}\boldsymbol{U}^{T}\\ &= \boldsymbol{U\tilde{I}U}^{T},\end{aligned}$$

That is, $\frac{1}{\hat{\rho}}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}$ is a positive semi-definite matrix and $\mathrm{Tr}\left(\frac{1}{\hat{\rho}}\widehat{\mathrm{SP}'}^{T}\widehat{\mathrm{SP}'}^{T,\dagger}\right) \leq k/\hat{\rho}$. The matrix $\mathbb{E}[\eta\eta^{T}]$ is diagonal with all the non-zero entries on diagonal (variance of components of η) bounded above by a constant that depends on γ. For a positive semi-definite matrix A and positive semi-definite diagonal matrix B, $\mathrm{Tr}(AB) \leq B_2\mathrm{Tr}(A)$. For $\rho \geq C\log(NT)/\sqrt{NT}$ for large enough C, one can verfiy that $\mathbb{E}[1/\hat{\rho}] \leq 2/\rho$. This completes the justification of the last step of (5.51).

Now consider the term $\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2}$. Note,

$$\begin{aligned}\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2} &= (\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*}) + (\frac{\rho-\hat{\rho}}{\hat{\rho}})\mathrm{SP}(f)_{L\cdot}{}^{2}_{2}\\ &\leq 2(\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*})^{2}_{2} + 2\frac{\rho-\hat{\rho}}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot}{}^{2}_{2}.\end{aligned} \tag{5.52}$$

We will bound the two terms on the r.h.s of (5.52) separately. We now consider the first term.

$$\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2} \leq 2\mathrm{SP}(f)_{L\cdot} - \mathrm{SP}'(f)^{T}\beta^{*2}_{2} + 2\mathrm{SP}'(f)^{T}\beta^{*} - \widehat{\mathrm{SP}'}^{T}\beta^{*2}_{2}. \tag{5.53}$$

By Proposition 11

$$\mathrm{SP}(f)_{L\cdot} - \mathrm{SP}'(f)^{T}\beta^{*}{}_{2} \leq \mathrm{SP}(f)_{L\cdot} - \mathrm{SP}'(f)^{T}\beta^{*}{}_{\infty}\sqrt{NT/L} \;\leq\; C_2\epsilon\sqrt{NT/L}, \tag{5.54}$$

where we used the fact that for any $v \in \mathbb{R}^{p}$, $v_2 \leq v_{\infty}\sqrt{p}$. And,

$$\mathrm{SP}'(f)^{T}\beta^{*} - \widehat{\mathrm{SP}'}^{T}\beta^{*}{}_{2} = (\mathrm{SP}'(f) - \widehat{\mathrm{SP}'})^{T}\beta^{*}{}_{2} \;\leq\; \mathrm{SP}'(f) - \widehat{\mathrm{SP}'}_{2,\infty}\beta^{*}{}_{1}, \tag{5.55}$$

where we used the fact that for any $A \in \mathbb{R}^{q\times p}, v \in \mathbb{R}^{p}$, $Av_2 \leq A^{T}{}_{2,\infty}v_1$. Finally, note that

$$\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^{2}_{2} \leq 2\frac{\rho}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}^{2}_{2} + 2\frac{\rho-\hat{\rho}}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot}{}^{2}_{2}. \tag{5.56}$$

Using (5.48), (5.95), (5.51), (5.52),(5.53), (5.54), (5.55), and the bound in (5.56), we obtain

$$\begin{aligned}&\mathbb{E}[\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}_2^2] \\ &\leq 4C(\gamma)k/\rho + 6\mathbb{E}[\frac{\rho-\hat{\rho}}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot}{}_2^2] + 2C_2\epsilon^2(NT/L) + 2\beta^*{}_1^2\mathrm{SP}'(f) - \widehat{\mathrm{SP}'}_{2,\infty}^2.\end{aligned} \tag{5.57}$$

Note that $\mathrm{SP}(f)_{\infty} \leq R\Gamma_1\Gamma_2$. Hence, $\mathrm{SP}(f)_{L\cdot}{}_2^2 \leq C(\Gamma_1, \Gamma_2)R^2(NT/L)$, for large enough constant $C(\Gamma_1, \Gamma_2)$ that may depend on Γ_1, Γ_2. Using the bounds derived in Lemma 5.15.3, one can verify that $\mathbb{E}[(\frac{\rho-\hat{\rho}}{\hat{\rho}})^2] \leq C/(NT/L)$ for large enough positive constant C. Therefore, we have that

$$6\mathbb{E}[\frac{\rho-\hat{\rho}}{\hat{\rho}}\mathrm{SP}(f)_{L\cdot}{}_2^2] \leq C(\Gamma_1, \Gamma_2)R^2 \tag{5.58}$$

Using (5.44), (5.58), and the bound in (5.57); diving by $1/(NT/L)$ on both sides; and noting $k \leq R \,\mathrm{x}\, G$, we obtain

$$\begin{aligned}&\mathbb{E}[\frac{1}{(NT/L)}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}_2^2] \\ &\leq C(c, \gamma, \Gamma_1, \Gamma_2)\left(\frac{RG}{\rho(NT/L)} + \frac{R^2}{(NT/L)} + R(1+\beta^*{}_1)\epsilon^2 + \beta^*{}_1^2\left(\frac{R^3 G \log NT}{\rho^4 L} + \frac{R^4 G(\epsilon + \epsilon^2 + \epsilon^3)}{\rho^2}\right)\right) \\ &\leq C(c, \gamma, \Gamma_1, \Gamma_2)\left(\max(1, \beta^*{}_1, \beta^*{}_1^2)\left(\frac{R^3 G \log NT}{\rho^4 L} + \frac{R^4 G(\epsilon + \epsilon^2 + \epsilon^3)}{\rho^2}\right)\right)\end{aligned} \tag{5.59}$$

Letting $L = \sqrt{\min(N, T)T}$, using (5.59), and noting that

$$\mathrm{ForErr}(N, T, L) = \mathbb{E}[\frac{1}{(NT/L)}\mathrm{SP}(f)_{L\cdot} - \widehat{\mathrm{SP}'}^{T}\widehat{\beta}_2^2]$$

completes the proof of Theorem 5.5.2.

■ 5.17.1 Proof of Proposition 11

For this proof, we utilize a modified version of the stacked Hankel matrix defined in Appendix 5.12. Define the modified Hankel matrix for time series f_n, for $n \in [N]$, as

$\widetilde{\text{H}}(n) \in \mathbb{R}^{T \times 2T}$, where for $i \in [T], j \in [2T]$, we have

$$\widetilde{\text{H}}(n)_{ij} = f_n(i + j - 1 - T).$$

Define $\widetilde{\text{SH}} \in \mathbb{R}^{T \times NT}$ as the column wise concatenation of the matrices $\widetilde{\text{H}}(n)$ for $n \in [N]$, i.e., $\widetilde{\text{SH}} := [\widetilde{\text{H}}(1), \dots, \widetilde{\text{H}}(N)]$. By a straightforward modification of the proof of Proposition 15, we have $\widetilde{\text{SH}}$ has ϵ'-rank bounded by $R \,\text{x}\, G$ with $\epsilon' = R\Gamma_1 \epsilon$. That is, there exists a matrix $\text{M} \in \mathbb{R}^{T \times NT}$ such that,

$$\text{rank}(\text{M}) \le RG, \quad \widetilde{\text{SH}} - \text{M}_{\infty} \le \epsilon'$$

Since $\text{rank}(\text{M}) \le RG$, it must be the case that within the last RG rows of M, there exists at least one row, which we denote as r^*, that can be written as a linear combination of at most RG rows above it, which we denote as $r_1, \dots, r_{RG}$. Specifically there exists a vector $\theta := (\theta_1, \dots, \theta_{RG}) \in \mathbb{R}^{RG}$ such that

$$\text{M}_{r^*,\cdot} = \sum_{\ell=1}^{RG} \theta_\ell \text{M}_{r_\ell,\cdot}$$

Hence for $j \in [2T]$,

$$\begin{aligned}
&\left| \widetilde{\text{SH}}_{r^*,j} - \sum_{\ell=1}^{RG} \theta_\ell \widetilde{\text{SH}}_{r_\ell,j} \right| \\
&= \left| \widetilde{\text{SH}}_{r^*,j} \pm \text{M}_{r^*,j} - \sum_{\ell=1}^{RG} \theta_\ell \widetilde{\text{SH}}_{r_\ell,j} \pm \sum_{\ell=1}^{RG} \theta_\ell \text{M}_{r_\ell,t} \right| \\
&\le \left| \widetilde{\text{SH}}_{r^*,j} - \text{M}_{r^*,j} \right| + \left| \sum_{\ell=1}^{RG} \theta_\ell \widetilde{\text{SH}}_{r_\ell,j} - \sum_{\ell=1}^{RG} \theta_\ell \text{M}_{r_\ell,t} \right| + \left| \text{M}_{r^*,j} - \sum_{\ell=1}^{RG} \theta_\ell \text{M}_{r_\ell,t} \right| \\
&= \left| \widetilde{\text{SH}}_{r^*,j} - \text{M}_{r^*,j} \right| + \left| \sum_{\ell=1}^{RG} \theta_\ell (\widetilde{\text{SH}}_{r_\ell,j} - \text{M}_{r_\ell,t}) \right| \\
&\le \epsilon' + \theta_1 \widetilde{\text{SH}}_{r_\ell,j} - \text{M}_{r_\ell,t\infty} \\
&\le R\Gamma_1 (1 + \theta_1) \epsilon. \qquad (5.60)
\end{aligned}$$

Observe that every entry of $\text{SP}(f)_{L\cdot}$ appears within $\widetilde{\text{SH}}_{r^*,\cdot}$; this can be seen by noting that $\widetilde{\text{SH}}$ is skew-symmetric and thus every entry in the last row of $\widetilde{\text{SH}}$ appears along the appropriate diagonal. Using this skew-symmetric property of $\widetilde{\mathrm{SH}}$ and (5.60), it implies that by appropriately selecting entries in $\widetilde{\mathrm{SH}}$, there exists $\beta^* \in \mathbb{R}^{L-1}$,

$$\mathrm{SP}(f)_{L\cdot}^T - \mathrm{SP}'(f)^T \beta^*{}_{\infty} \leq R\Gamma_1(1+\beta_1)\epsilon,$$

where the non-zero entries in β^* correspond to the entries of θ. Noting that $\theta \in \mathbb{R}^{RG}$ implies $\beta^*{}_0 \leq RG$. This completes the proof.

5.18 Proof of Theorem 5.4.3

Notation. For integers $t_1 < t_2$ where $t_2 - t_1 + 1 \geq L$, let $\mathrm{SP}((X_1, \dots, X_N), t_1 : t_2, L)$ represents the stacked page matrix constructed using the contiguous observations $X_n(t_1), \dots, X_n(t_2)$, $\forall n \in [N]$. Throughout, we use the following notations:

- $\mathrm{SP}_0(X) = \mathrm{SP}((X_1, \dots, X_N), 1 : T, L) \in \mathbb{R}^{L \times (NT/L)}$, with zeros replacing missing values.
- $\mathrm{SP}_1(X) = \mathrm{SP}((X_1, \dots, X_N), T+1 : T+T_1, L) \in \mathbb{R}^{L \times (NT_1/L)}$, with zeros replacing missing values.
- $\mathrm{SP}_0(f) = \mathrm{SP}((f_1, \dots, f_N), 1 : T, L) \in \mathbb{R}^{L \times (NT/L)}$.
- $\mathrm{SP}_1(f) = \mathrm{SP}((f_1, \dots, f_N), T+1 : T+T_1, L) \in \mathbb{R}^{L \times (NT_1/L)}$.
- $\mathrm{SP}_1(\eta) = \mathrm{SP}((\eta_1, \dots, \eta_N), T+1 : T+T_1, L) \in \mathbb{R}^{L \times (NT_1/L)}$.
- $\mathrm{SP}_0'(X) \in \mathbb{R}^{(L-1) \times (NT/L)}$ as the top $L-1$ rows of $\mathrm{SP}_0(X)$. Let $\mathrm{SP}_1'(X), \mathrm{SP}_0'(f), \mathrm{SP}_1'(f)$ and $\mathrm{SP}_1'(\eta)$ be defined analogously.
- $\hat{\rho} := (\max(1, \sum_{i=1}^{L-1} \sum_{j=1}^{NT/L} \mathbf{1}(\mathrm{SP}_0(X)_{ij} \neq \star)))/(NT - NT/L)$

Recall that we are interested in bounding the following out-of-sample prediction error:

$$\mathrm{TestForErr}(N, T, T_1, L) = \frac{L}{NT_1} \sum_{n=1}^{N} \sum_{m'=1}^{T_1/L} \mathbb{E}[(f_n(T + L \times m') - \bar{f}_n(T + L \times m'))^2].$$

Where the forecasted estimate $\bar{f}_n(\cdot)$, $n \in [N]$ are produced by the algorithm detailed in Section 5.1.1.

Based on the algorithm, we can write TestForErr(N, T, T_1, L) as follows:

$$\begin{aligned}\text{TestForErr}(N, T, T_1, L) &= \frac{1}{(NT_1/L)}\mathbb{E}\Big[\frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \text{SP}_1(f)^{T\,2}_{L\cdot 2}\Big] \\ &= \frac{1}{(NT_1/L)}\mathbb{E}\Big[\frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \text{SP}'_1(f)^T\beta^{*2}_{\ 2}\Big].\end{aligned}$$

Before bounding this term, we introduce the following important notation. For $i \in \{0, 1\}$, let $\boldsymbol{U}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^T$ denote the Singular Value Decomposition (SVD) of $\text{SP}'_i(f)$. Also, let $\widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\boldsymbol{V}}_i^T$ denote the top k singular components of the SVD of $\text{SP}'_i(X)$, while $\widetilde{\boldsymbol{U}}_i^{\perp}\widetilde{\boldsymbol{\Sigma}}_i^{\perp}(\widetilde{\boldsymbol{V}}_i^{\perp})^T$ denote the remaining $L-k-1$ components such that $\text{SP}'_i(X) = \widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\boldsymbol{V}}_i^T + \widetilde{\boldsymbol{U}}_i^{\perp}\widetilde{\boldsymbol{\Sigma}}_i^{\perp}(\widetilde{\boldsymbol{V}}_i^{\perp})^T$. Finally, let $\boldsymbol{V}_i^{\perp}$ and $\boldsymbol{U}_i^{\perp}$ be matrices of orthornormal basis vectors that span the null space of $\text{SP}'_i(f)$ and $\text{SP}'_i(f)^T$, respectively. Further, let $\widehat{\text{SP}'_i}$ be the HSVT estimate of $\text{SP}'_i(f)$. That is $\widehat{\text{SP}'_i} = \frac{1}{\hat{\rho}}\widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\boldsymbol{V}}_i^T$. Also, let $\widehat{\text{SP}'_i}^{\perp} = \frac{1}{\hat{\rho}}\widetilde{\boldsymbol{U}}_i^{\perp}\widetilde{\boldsymbol{\Sigma}}_i^{\perp}(\widetilde{\boldsymbol{V}}_i^{\perp})^T$.

We start the proof by providing a deterministic upper bound for out-of-sample error.

Deterministic Bound. Due to triangle inequality, we have

$$\begin{aligned}\frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \text{SP}'_1(f)^T\beta^{*2}_{\ 2} &= \frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \text{SP}'_1(f)^T\beta^{*} + \widehat{\text{SP}'_1}^T\widehat{\beta} - \widehat{\text{SP}'_1}^T\widehat{\beta}^2_2 \\ &\leq 2\frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \widehat{\text{SP}'_1}^T\widehat{\beta}^2_2 + 2\widehat{\text{SP}'_1}^T\widehat{\beta} - \text{SP}'_1(f)^T\beta^{*2}_{\ 2}.\end{aligned}$$

Next, we proceed to bound each of the two terms on the right hand side.

First term: $\frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \widehat{\text{SP}'_1}^T\widehat{\beta}^2_2$.

$$\begin{aligned}\frac{1}{\hat{\rho}}\text{SP}'_1(X)^T\widehat{\beta} - \widehat{\text{SP}'_1}^T\widehat{\beta}^2_2 &= (\widehat{\text{SP}'_1}^{\perp})^T\widehat{\beta}^2_2 \qquad (5.61) \\ &= \frac{1}{\hat{\rho}}\widetilde{\boldsymbol{V}}_1^{\perp}\widetilde{\boldsymbol{\Sigma}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}^2_2 \\ &\leq \frac{1}{\hat{\rho}}\widetilde{\boldsymbol{\Sigma}}_1^{\perp\,2}{}_2(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}^2_2.\end{aligned}$$

Note that $\widetilde{\boldsymbol{\Sigma}}_1^{\perp}{}_2$ equals the $(k+1)$-th singular value of $\mathsf{SP}_1'(X)$. Recall that $\mathbb{E}[\mathsf{SP}_1'(X)] = \rho\mathsf{SP}_1'(f)$ and hence

$$\mathsf{SP}_1'(X) = \rho\mathsf{SP}_1'(f) + \boldsymbol{\iota}_1, \tag{5.62}$$

where $\boldsymbol{\iota}_1 \in \mathbb{R}^{(L-1)\times(NT_1)/L}$ is a random matrix with zero-mean i.i.d. entries where each entry is 0 with probability $1-\rho$ and equals a zero-mean sub-Gaussian random variable with $\cdot_{\psi_2} \leq \gamma$ with probability ρ (due to Property 5.4.1). Next, we show that each component of $\boldsymbol{\iota}_1$ is an independent, zero-mean random variable with $\cdot_{\psi_2}$ bounded above by $C'(\gamma + R\Gamma_1\Gamma_2)$ for some absolute constant $C' > 0$. Let ζ_{ij} for $i \in [L-1]$ and $j \in [NT/L]$ denotes the ij-th entry in $\boldsymbol{\iota}_1$. Further, let $P_{ij} \in \{0,1\}$ denotes the random mask which takes the value 1 with probability ρ such that $\mathsf{SP}_1'(X)_{ij} = P_{ij}(\mathsf{SP}_1'(f)_{ij} + \mathsf{SP}_1'(\eta)_{ij})$. Then, we have

$$\begin{aligned}
\zeta_{ij\psi_2} &= \mathsf{SP}_1'(X)_{ij} - \rho\mathsf{SP}_1'(f)_{ij\psi_2} \\
&= P_{ij}\mathsf{SP}_1'(f)_{ij} + P_{ij}\mathsf{SP}_1'(\eta)_{ij} - \rho\mathsf{SP}_1'(f)_{ij\psi_2} \\
&\leq P_{ij}\mathsf{SP}_1'(\eta)_{ij\psi_2} + P_{ij}\mathsf{SP}_1'(f)_{ij} - \rho\mathsf{SP}_1'(f)_{ij\psi_2} \\
&\leq C\gamma + \mathsf{SP}_1'(f)_{ij}P_{ij} - \rho_{\psi_2} \\
&\leq C'(\gamma + R\Gamma_1\Gamma_2),
\end{aligned}$$

where $C, C' > 0$ are absolute constants. The first inequality is due to triangle inequality, and the last follows since $P_{ij} - \rho$ is a random variable bounded between $[-\rho, 1-\rho]$ and $\mathsf{SP}_1'(f)_{ij}$ is bounded by $R\Gamma_1\Gamma_2$. With a similar argument, we can also write

$$\mathsf{SP}_0'(X) = \rho\mathsf{SP}_0'(f) + \boldsymbol{\iota}_0,$$

where each component of $\boldsymbol{\iota}_0$ is again an independent, zero-mean random variable with $\cdot_{\psi_2}$ bounded above by $C'(\gamma + R\Gamma_1\Gamma_2)$. Now, recalling that $\mathsf{SP}_1'(X) = \rho\mathsf{SP}_1'(f) + \boldsymbol{\iota}_1$ and using Weyl's inequality (see Lemma 5.14.4), we can bound the $(k+1)$-th singular value of $\mathsf{SP}_1'(X)$ by the largest singular value of $\boldsymbol{\iota}_1$. That is,

$$\widetilde{\boldsymbol{\Sigma}}_1^{\perp}{}_2^2 \leq \boldsymbol{\iota}_1{}_2^2. \tag{5.63}$$

Next, we bound the term $(\widetilde{U}_1^{\perp})^T\widehat{\beta}_2^2$.

$$\begin{aligned}
(\widetilde{U}_1^{\perp})^T\widehat{\beta}_2^2 &= \widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T\widehat{\beta}\ {}_2^2 \\
&= \widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T\beta^* + \widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T(\widehat{\beta}-\beta^*)_2^2 \\
&\le 2\widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T\beta^{*2}_2 + 2\widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T(\widehat{\beta}-\beta^*)_2^2 \\
&\le 2\widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T\beta^{*2}_2 + 2\widehat{\beta}-\beta^{*2}_2.
\end{aligned} \tag{5.64}$$

First, consider

$$\begin{aligned}
\widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T\beta^*{}_2 &= \widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T U_1(U_1)^T\beta^*{}_2 \\
&\le \left\| U_1^{\perp}(U_1^{\perp})^T U_1(U_1)^T\beta^* \right\|_2 + \left\| \left(\widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T U_1(U_1)^T - U_1^{\perp}(U_1^{\perp})^T U_1(U_1)^T \right) \beta^* \right\|_2 \\
&\le \left\| \left(\widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T - U_1^{\perp}(U_1^{\perp})^T \right) \beta^* \right\|_2 \\
&\le \left\| \widetilde{U}_1^{\perp}(\widetilde{U}_1^{\perp})^T - U_1^{\perp}(U_1^{\perp})^T \right\|_2 \left\|\beta^*\right\|_2 \\
&= \left\| \widetilde{U}_1\widetilde{U}_1^T - U_1U_1^T \right\|_2 \left\|\beta^*\right\|_2 .
\end{aligned} \tag{5.65}$$

Where in the first equality we use the fact that $\beta^* = U_1(U_1)^T\beta^*$, i.e., β^* lives in the column space of $\mathrm{SP}_1'(f)$ (Property 5.4.4). Next, by Wedin $\sin\Theta$ Theorem (see Davis and Kahan (1970); Wedin (1972)) we bound $\left\| \widetilde{U}_1\widetilde{U}_1^T - U_1U_1^T \right\|_2$ as follows:

$$\begin{aligned}
\left\| \widetilde{U}_1\widetilde{U}_1^T - U_1U_1^T \right\|_2 \left\|\beta^*\right\|_2 &\le \frac{\mathrm{SP}_1'(X) - \rho\mathrm{SP}_1'(f)_2}{\sigma_k(\rho\mathrm{SP}_1'(f))} \left\|\beta^*\right\|_2 \\
&= \frac{\iota_{12}}{\sigma_k(\rho\mathrm{SP}_1'(f))} \left\|\beta^*\right\|_2 .
\end{aligned} \tag{5.66}$$

For $\widehat{\beta}-\beta^*{}_2$, we have:

$$\begin{aligned}
\widehat{\beta}-\beta^{*2}_2 &= \widetilde{U}_0^{\perp}(\widetilde{U}_0^{\perp})^T(\widehat{\beta}-\beta^*) + \widetilde{U}_0(\widetilde{U}_0)^T(\widehat{\beta}-\beta^*)_2^2 \\
&= \widetilde{U}_0^{\perp}(\widetilde{U}_0^{\perp})^T(\widehat{\beta}-\beta^*)_2^2 + \widetilde{U}_0(\widetilde{U}_0)^T(\widehat{\beta}-\beta^*)_2^2 \\
&= \widetilde{U}_0^{\perp}(\widetilde{U}_0^{\perp})^T(\widehat{\beta}-\beta^*)_2^2 + \widetilde{U}_0^T(\widehat{\beta}-\beta^*)_2^2 \\
&= \widetilde{U}_0^{\perp}(\widetilde{U}_0^{\perp})^T(\beta^*)_2^2 + \widetilde{U}_0^T(\widehat{\beta}-\beta^*)_2^2.
\end{aligned} \tag{5.67}$$

Note that the last equality follow from the fact that $\widehat{\beta} = \widehat{\mathrm{SP}'_0}^{T,\dagger} \frac{1}{\hat{\rho}} \mathrm{SP}_0(X)_{L\cdot} = \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^{\dagger} \widetilde{\boldsymbol{V}}^T \mathrm{SP}_0(X)_{L\cdot}$, where $\widehat{\mathrm{SP}'_0}^{T,\dagger}$ is the pseudoinverse of $\widehat{\mathrm{SP}'_0}^{T}$, and thus $(\widetilde{\boldsymbol{U}}_0^{\perp})^T \widehat{\beta} = 0$. The first term in (5.67) can be bounded using the same argument in (5.65) and (5.66), where we utilize the fact that $\beta^* = \boldsymbol{U}_0(\boldsymbol{U}_0)^T \beta^*$ and Wedin $\sin \Theta$ Theorem to get

$$\widetilde{\boldsymbol{U}}_0^{\perp}(\widetilde{\boldsymbol{U}}_0^{\perp})^T \beta^*{}_2 \leq \frac{\boldsymbol{\iota}_0{}_2}{\sigma_k(\rho \mathrm{SP}'_0(f))} \left\| \beta^* \right\|_2 . \tag{5.68}$$

What is left is bounding $\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta} - \beta^*)_2^2$. To that end, first consider

$$\begin{aligned} \widehat{\mathrm{SP}'_0}^{T}(\widehat{\beta} - \beta^*)_2^2 &\leq 2\widehat{\mathrm{SP}'_0}^{T} \widehat{\beta} - \mathrm{SP}'_0(f)^T \beta^*{}_2^2 + 2\mathrm{SP}'_0(f)^T \beta^* - \widehat{\mathrm{SP}'_0}^{T} \beta^*{}_2^2 \\ &\leq 2\widehat{\mathrm{SP}'_0}^{T} \widehat{\beta} - \mathrm{SP}'_0(f)^T \beta^*{}_2^2 + 2\mathrm{SP}'_0(f) - \widehat{\mathrm{SP}'_0}{}_{0 2,\infty}^{2} \beta^*{}_1^2. \end{aligned} \tag{5.69}$$

Also, consider

$$\begin{aligned} \widehat{\mathrm{SP}'_0}^{T}(\widehat{\beta} - \beta^*)_2^2 &= (\widehat{\beta} - \beta^*)^T \frac{1}{\hat{\rho}^2} \widetilde{\boldsymbol{U}}_0 \widetilde{\boldsymbol{\Sigma}}_0^2 \widetilde{\boldsymbol{U}}^T (\widehat{\beta} - \beta^*) \\ &\geq \sigma_k(\widehat{\mathrm{SP}'_0})^2 \widetilde{\boldsymbol{U}}_0^T (\widehat{\beta} - \beta^*)_2^2. \end{aligned} \tag{5.70}$$

From (5.70) and (5.69) we get,

$$\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta} - \beta^*)_2^2 \leq \frac{2}{\sigma_k(\widehat{\mathrm{SP}'_0})^2} (\widehat{\mathrm{SP}'_0}^{T} \widehat{\beta} - \mathrm{SP}'_0(f)^T \beta^*{}_2^2 + \mathrm{SP}'_0(f) - \widehat{\mathrm{SP}'_0}{}_{0 2,\infty}^{2} \beta^*{}_1^2). \tag{5.71}$$

Note that, similar to argument in (5.62), $\mathrm{SP}_0(X)_{L\cdot} = \rho \mathrm{SP}_0(f)_{L\cdot} + \zeta_0^L$, where ζ_0^L is a vector of i.i.d. entries with $\cdot_{\psi_2} \leq C'(\boldsymbol{\gamma} + R\Gamma_1\Gamma_2)$. Then the term $\widehat{\mathrm{SP}'_0}^{T} \widehat{\beta} - \mathrm{SP}'_0(f)^T \beta^*{}_2^2$ can be bounded as follows

$$\begin{aligned} &\widehat{\mathrm{SP}'_0}^{T} \widehat{\beta} - \frac{1}{\rho} \mathrm{SP}_0(X)_{L\cdot}{}_2^2 \\ =&\widehat{\mathrm{SP}'_0}^{T} \widehat{\beta} - \mathrm{SP}_0(f)_{L\cdot} - \frac{1}{\rho} \zeta_0^L{}_2^2 \end{aligned}$$

$$=\widehat{\mathrm{SP}'_0}^T\widehat{\beta}-\mathrm{SP}'_0(f)^T\beta^*{}_2^2+\frac{1}{\rho}\zeta_{0\,2}^{L\,2}-\frac{2}{\rho}(\widehat{\mathrm{SP}'_0}^T\widehat{\beta}-\mathrm{SP}'_0(f)^T\beta^*)^T\zeta_0^L. \tag{5.72}$$

Also, we have

$$\begin{aligned}
&\widehat{\mathrm{SP}'_0}^T\widehat{\beta}-\frac{1}{\rho}\mathrm{SP}_0(X)_{L\cdot}{}_2^2\\
\leq&\widehat{\mathrm{SP}'_0}^T\beta^*-\frac{1}{\rho}\mathrm{SP}_0(X)_{L\cdot}{}_2^2\\
=&(\widehat{\mathrm{SP}'_0}^T-\mathrm{SP}'_0(f)^T)\beta^*-\frac{1}{\rho}\zeta_0^L{}_2^2\\
=&(\widehat{\mathrm{SP}'_0}^T-\mathrm{SP}'_0(f)^T)\beta^*{}_2^2+\frac{1}{\rho}\zeta_{0\,2}^{L\,2}-\frac{2}{\rho}\left((\widehat{\mathrm{SP}'_0}^T-\mathrm{SP}'_0(f)^T)\beta^*\right)^T\zeta_0^L.
\end{aligned} \tag{5.73}$$

From (5.72) and (5.73) we have,

$$\begin{aligned}
\widehat{\mathrm{SP}'_0}^T\widehat{\beta}-\mathrm{SP}'_0(f)^T\beta^*{}_2^2&\leq(\widehat{\mathrm{SP}'_0}^T-\mathrm{SP}'_0(f)^T)\beta^*{}_2^2+\frac{2}{\rho}\left((\widehat{\mathrm{SP}'_0}^T)(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L \qquad (5.74)\\
&\leq\widehat{\mathrm{SP}'_0}-\mathrm{SP}'_0(f)_{2,\infty}^2\beta^*{}_1^2+\frac{2}{\rho}\left((\widehat{\mathrm{SP}'_0}^T)(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L.
\end{aligned}$$

Finally, from (5.71) and (5.74) we get

$$\widetilde{U}_0^T(\widehat{\beta}-\beta^*)_2^2\leq\frac{4}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}\left(\mathrm{SP}'_0(f)-\widehat{\mathrm{SP}'}_{0\,2,\infty}^{\,2}\beta^*{}_1^2+\frac{1}{\rho}\left((\widehat{\mathrm{SP}'_0}^T)(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L\right) \tag{5.75}$$

From (5.67), (5.68), and (5.75) we have

$$\begin{aligned}
\widehat{\beta}-\beta^*{}_2^2\leq&\frac{\iota_0{}_2^2}{\sigma_k(\rho\mathrm{SP}'_0(f))^2}\left\|\beta^*\right\|_2^2 \qquad (5.76)\\
&+\frac{4}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}\left(\mathrm{SP}'_0(f)-\widehat{\mathrm{SP}'}_{0\,2,\infty}^{\,2}\beta^*{}_1^2+\frac{1}{\rho}\left((\widehat{\mathrm{SP}'_0}^T)(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L\right).
\end{aligned}$$

For ease of exposition, let

$$\begin{aligned}\Delta_1 &:= \mathrm{SP}'_0(f) - \widehat{\mathrm{SP}'}_{0}{}^{2}_{2,\infty}\beta^{*}{}^{2}_{1} + \frac{1}{\rho}\left(\widehat{\mathrm{SP}'_0}^{\,T}(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L \\ \Delta_2 &:= \frac{\boldsymbol{\iota}_0{}^2_2}{\sigma_k(\rho \mathrm{SP}'_0(f))^2}\left\|\beta^*\right\|_2^2 + \frac{4}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}(\Delta_1). \end{aligned} \tag{5.77}$$

Using this definition, (5.61), (5.63), (5.64), (5.66), and (5.76), we have

$$\frac{1}{\hat{\rho}}\mathrm{SP}'_1(X)^T\widehat{\beta} - \widehat{\mathrm{SP}'_1}^{\,T}\widehat{\beta}^2_2 \leq \frac{1}{\hat{\rho}}\boldsymbol{\iota}_1{}^2_2\left(\frac{2\boldsymbol{\iota}_1{}^2_2\left\|\beta^*\right\|_2^2}{\sigma_k(\rho \mathrm{SP}'_1(f))^2} + 2\Delta_2\right). \tag{5.78}$$

Second term: $\mathrm{SP}'_1(f)^T\beta^* - \widehat{\mathrm{SP}'_1}^{\,T}\widehat{\beta}^2_2$. To bound the second term, we follow a similar proof to that shown in Agarwal et al. (2021d).

$$\begin{aligned}\mathrm{SP}'_1(f)^T\beta^* - \widehat{\mathrm{SP}'_1}^{\,T}\widehat{\beta}^2_2 &= \mathrm{SP}'_1(f)^T\beta^* + \widehat{\mathrm{SP}'_1}^{\,T}\beta^* - \widehat{\mathrm{SP}'_1}^{\,T}\beta^* - \widehat{\mathrm{SP}'_1}^{\,T}\widehat{\beta}^2_2 \\ &\leq 2(\mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'_1})^T\beta^*{}^2_2 + 2\widehat{\mathrm{SP}'_1}^{\,T}(\beta^* - \widehat{\beta})^2_2. \end{aligned} \tag{5.79}$$

Next, we bound the two terms on the right hand side. First, we bound $(\mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'_1})^T\beta^*{}^2_2$ as follows.

$$(\mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'_1})^T\beta^*{}^2_2 \leq \mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'}_{1}{}^{2}_{2,\infty}\beta^*{}^2_1. \tag{5.80}$$

Next, we bound the second term $\widehat{\mathrm{SP}'_1}^{\,T}(\beta^* - \widehat{\beta})^2_2$.

$$\begin{aligned}\widehat{\mathrm{SP}'_1}^{\,T}(\beta^* - \widehat{\beta})^2_2 &\leq \frac{1}{\hat{\rho}^2}(\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T + \rho \mathrm{SP}'_1(f)^T - \rho \mathrm{SP}'_1(f)^T)(\beta^* - \widehat{\beta})^2_2 \\ &\leq \frac{2}{\hat{\rho}^2}(\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \rho \mathrm{SP}'_1(f)^T)(\beta^* - \widehat{\beta})^2_2 + \frac{2\rho^2}{\hat{\rho}^2}\mathrm{SP}'_1(f)^T(\beta^* - \widehat{\beta})^2_2 \end{aligned}$$

$$\leq \frac{2}{\hat{\rho}^2}\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \rho \mathrm{SP}_1'(f)^{T}{}_2^2(\beta^* - \widehat{\beta})_2^2 + \frac{2\rho^2}{\hat{\rho}^2}\mathrm{SP}_1'(f)^{T}(\beta^* - \widehat{\beta})_2^2.$$

Further, note that

$$\begin{aligned}\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \rho \mathrm{SP}_1'(f)^{T}{}_2^2 &\leq 2\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \mathrm{SP}_1'(X)^{T}{}_2^2 + 2\mathrm{SP}_1'(X)^{T} - \rho \mathrm{SP}_1'(f)^{T}{}_2^2 \\ &\leq 4\mathrm{SP}_1'(X)^{T} - \rho \mathrm{SP}_1'(f)^{T}{}_2^2 = 4\boldsymbol{\iota}_1{}_2^2.\end{aligned}$$

Where the last inequality follows from the fact that $\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \mathrm{SP}_1'(X)^{T}{}_2$ is the $k+1$-th singular value of $\mathrm{SP}_1'(X)$ and hence is bounded by $\mathrm{SP}_1'(X)^{T} - \rho \mathrm{SP}_1'(f)^{T}{}_2$ using Weyl's inequality. Therefore,

$$\widehat{\mathrm{SP}_1'}^{T}(\beta^* - \widehat{\beta})_2^2 \leq \frac{8}{\hat{\rho}^2}\boldsymbol{\iota}_1{}_2^2\beta^* - \widehat{\beta}_2^2 + \frac{2\rho^2}{\hat{\rho}^2}\mathrm{SP}_1'(f)^{T}(\beta^* - \widehat{\beta})_2^2. \tag{5.81}$$

Next, we bound $\mathrm{SP}_1'(f)^{T}(\beta^* - \widehat{\beta})_2^2$. Recall that $\boldsymbol{U}_0$ span the column space of $\mathrm{SP}_1'(f)$. Thus $\mathrm{SP}_1'(f)^{T} = \mathrm{SP}_1'(f)^{T}\boldsymbol{U}_0\boldsymbol{U}_0^T$, therefore,

$$\begin{aligned}\mathrm{SP}_1'(f)^{T}(\beta^* - \widehat{\beta})_2^2 &= \mathrm{SP}_1'(f)^{T}\boldsymbol{U}_0\boldsymbol{U}_0^T(\beta^* - \widehat{\beta})_2^2 \\ &\leq \mathrm{SP}_1'(f)_2^2\boldsymbol{U}_0\boldsymbol{U}_0^T(\beta^* - \widehat{\beta})_2^2.\end{aligned} \tag{5.82}$$

Recall that $\widetilde{\boldsymbol{U}}_0$ denote the top k left singular vectors of $\mathrm{SP}_0'(x)$, and consider

$$\begin{aligned}\boldsymbol{U}_0\boldsymbol{U}_0^T(\beta^* - \widehat{\beta})_2^2 &= (\boldsymbol{U}_0\boldsymbol{U}_0^T + \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^T - \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^T)(\beta^* - \widehat{\beta})_2^2 \\ &\leq 2\boldsymbol{U}_0\boldsymbol{U}_0^T - \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^{T}{}_2^2\beta^* - \widehat{\beta}_2^2 + 2\widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^T(\beta^* - \widehat{\beta})_2^2.\end{aligned} \tag{5.83}$$

Using (5.83), (5.75) and Wedin $\sin\Theta$ Theorem, we obtain,

$$\begin{aligned}\boldsymbol{U}_0\boldsymbol{U}_0^T(\beta^* - \widehat{\beta})_2^2 &\leq \frac{2\boldsymbol{\iota}_0{}_2^2}{\sigma_k(\rho\mathrm{SP}_0'(f))^2}\beta^* - \widehat{\beta}_2^2 \\ &+ \frac{8}{\sigma_k(\widehat{\mathrm{SP}_0'})^2}\left(\mathrm{SP}_0'(f) - \widehat{\mathrm{SP}_0'}{}_{2,\infty}^2\beta^*{}_1^2 + \frac{1}{\rho}\left(\widehat{\mathrm{SP}_0'}^{T}(\widehat{\beta} - \beta^*)\right)^{T}\zeta_0^L\right).\end{aligned} \tag{5.84}$$

Using (5.82) and (5.84), we have

$$
\begin{aligned}
\mathrm{SP}'_1(f)^T(\beta^* - \widehat{\beta})_2^2 \le \mathrm{SP}'_1(f)_2^2 &\frac{2\boldsymbol{\iota}_0{}_2^2}{\sigma_k(\rho\mathrm{SP}'_0(f))^2}\beta^* - \widehat{\beta}_2^2 \\
&+ \frac{8\mathrm{SP}'_1(f)_2^2}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}\left(\mathrm{SP}'_0(f) - \widehat{\mathrm{SP}'}_0{}_{2,\infty}^2\beta^*{}_1^2 + \frac{1}{\rho}\left(\widehat{\mathrm{SP}'_0}^T(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right).
\end{aligned} \tag{5.85}
$$

Finally, using (5.85) and (5.81), we have

$$
\begin{aligned}
\widehat{\mathrm{SP}'_1}^T(\beta^* - \widehat{\beta})_2^2 \le &\frac{8}{\hat{\rho}^2}\boldsymbol{\iota}_1{}_2^2\beta^* - \widehat{\beta}_2^2 \\
&+ \frac{4}{\hat{\rho}^2}\frac{\boldsymbol{\iota}_0{}_2^2\mathrm{SP}'_1(f)_2^2}{\sigma_k(\mathrm{SP}'_0(f))^2}\beta^* - \widehat{\beta}_2^2 \\
&+ \frac{16\rho^2}{\hat{\rho}^2}\frac{\mathrm{SP}'_1(f)_2^2}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}\left(\mathrm{SP}'_0(f) - \widehat{\mathrm{SP}'}_0{}_{2,\infty}^2\beta^*{}_1^2 + \frac{1}{\rho}\left(\widehat{\mathrm{SP}'_0}^T(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right).
\end{aligned} \tag{5.86}
$$

Finally, combining (5.86), (5.80), (5.79), and (5.77) yields,

$$
\begin{aligned}
\mathrm{SP}'_1(f)^T\beta^* - \widehat{\mathrm{SP}'_1}^T\widehat{\beta}_2^2 \le &\, C\mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'}_1{}_{2,\infty}^2\beta^*{}_1^2 + \frac{C}{\hat{\rho}^2}\boldsymbol{\iota}_1{}_2^2\Delta_2 \\
&+ \frac{C}{\hat{\rho}^2}\frac{\boldsymbol{\iota}_0{}_2^2\mathrm{SP}'_1(f)_2^2}{\sigma_k(\mathrm{SP}'_0(f))^2}\Delta_2 + \frac{C\rho^2}{\hat{\rho}^2}\frac{\mathrm{SP}'_1(f)_2^2\Delta_1}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}.
\end{aligned} \tag{5.87}
$$

Combining. Incorporating the two bounds in (5.78) and (5.87) yields,

$$
\begin{aligned}
\frac{1}{\hat{\rho}}\mathrm{SP}'_1(X)^T\widehat{\beta} - \mathrm{SP}'_1(f)^T\beta^*{}_2^2 \le &\, C\frac{1}{\hat{\rho}}\boldsymbol{\iota}_1{}_2^2\left(\frac{\boldsymbol{\iota}_1{}_2^2\|\beta^*\|_2^2}{\sigma_k(\rho\mathrm{SP}'_1(f))^2} + \Delta_2\right) \\
&+ C\mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'}_1{}_{2,\infty}^2\beta^*{}_1^2 \\
&+ \frac{C}{\hat{\rho}^2}\frac{\boldsymbol{\iota}_0{}_2^2\mathrm{SP}'_1(f)_2^2}{\sigma_k(\mathrm{SP}'_0(f))^2}\Delta_2 + \frac{C\rho^2}{\hat{\rho}^2}\frac{\mathrm{SP}'_1(f)_2^2\Delta_1}{\sigma_k(\widehat{\mathrm{SP}'_0})^2}.
\end{aligned} \tag{5.88}
$$

For some absolute constant $C > 0$.

High Probability Bound. We start by defining the following high probability events. Let $C(\Gamma_1, \Gamma_2, \gamma)$ be a positive constant dependent on model parameters $\Gamma_1, \Gamma_2, \gamma$, and let $C > 0$ be some positive absolute constant, define

$$\bar{E}_1 := \left\{ \left\|\iota_0\right\|_2 \le C(\gamma + R\Gamma_1\Gamma_2)\sqrt{NT/L} \right\},$$

$$\bar{E}_2 := \left\{ \left\|\iota_1\right\|_2 \le C(\gamma + R\Gamma_1\Gamma_2)\sqrt{NT_1/L} \right\},$$

$$\bar{E}_3 := \left\{ \left(1 - \sqrt{\frac{20\log(NT)}{\rho NT}}\right)\rho \le \hat{\rho} \le \frac{1}{1 - \sqrt{\frac{20\log(NT)}{\rho NT}}}\rho \right\},$$

$$\bar{E}_4 := \left\{ \text{SP}_0'(f) - \widehat{\text{SP}'}_{02,\infty}^{\,2} \le C(\gamma, \Gamma_1, \Gamma_2)\left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\text{SP}_0'(f))^2L^2} + \frac{kR^2\log NT/L}{\rho^2} \right) \right\}, \tag{5.89}$$

$$\bar{E}_5 := \left\{ \text{SP}_1'(f) - \widehat{\text{SP}'}_{12,\infty}^{\,2} \le C(\gamma, \Gamma_1, \Gamma_2)\left(\frac{(NT_1)^2R^2}{\rho^4\sigma_k(\text{SP}_1'(f))^2L^2} + \frac{kR^2\log NT_1/L}{\rho^2} + \frac{R^2T_1}{T} \right) \right\}. \tag{5.90}$$

Using Theorem 5.14.2, we have the following,

$$\mathbb{P}(\bar{E}_1) \ge 1 - 2\exp\left(\frac{-NT}{L}\right),$$

$$\mathbb{P}(\bar{E}_2) \ge 1 - 2\exp\left(\frac{-NT_1}{L}\right).$$

Further by Lemma 5.15.3, $\mathbb{P}(\bar{E}_3) \ge 1 - \frac{2}{(NT)^{10}}$. Finally, the probabilities of $\bar{E}_4$ and $\bar{E}_5$ are bounded as we show next.

Lemma 5.18.1. *Let $\bar{E}_4$ and $\bar{E}_5$ be defined as in (5.89) and (5.90). Then, for a constant $C > 0$,*

$$\mathbb{P}(\bar{E}_4) \ge 1 - \frac{C}{(NT)^{10}},$$

$$\mathbb{P}(\bar{E}_5) \ge 1 - \frac{C}{(NT_1)^{10}} - \frac{C}{(NT)^{10}}.$$

Proof. **Bounding $\bar{E}_4$ and $\bar{E}_5$.** $\mathbb{P}(\bar{E}_4)$ and $\mathbb{P}(\bar{E}_5)$ can be bounded using a direct utilization of Lemma 5.15.2 and the high probability events defined in Appendix 5.15.5. Starting with $\bar{E}_4$, using (5.30), and recalling that in this theorem setup $\epsilon = 0, \Gamma = R\Gamma_1\Gamma_2$ (Property 5.3.1 and Property 5.3.2) and $\sigma = \gamma$ (Property 5.4.1), we have that with probability

$1 - \frac{C}{(NT)^{10}}$,

$$\mathrm{SP}'_0(f) - \widehat{\mathrm{SP}'}^{\,2}_{02,\infty} \leq C\frac{\gamma^2(NT)^2}{\rho^2\sigma_k(\mathrm{SP}'_0(f))^2L^2}\left((R\Gamma_1\Gamma_2)^2 + \frac{\gamma^2}{\rho^2}\right) + \frac{C\gamma^2 k\log NT/L}{\rho^2} + C(R\Gamma_1\Gamma_2)^2$$

$$\leq C(\gamma,\Gamma_1,\Gamma_2)\left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\mathrm{SP}'_0(f))^2L^2} + \frac{kR^2\log NT/L}{\rho^2}\right).$$

A similar argument can be used for $\bar{E}_5$, while noting that the term $\frac{C}{(NT)^{10}}$ shows up due to utilizing the estimate $\hat{\rho}$, which is estimated from the first T observations. Precisely, we get the following,

$$\mathrm{SP}'_1(f) - \widehat{\mathrm{SP}'}^{\,2}_{12,\infty} \leq C\frac{\gamma^2(NT_1)^2}{\rho^2\sigma_k(\mathrm{SP}'_1(f))^2L^2}\left((R\Gamma_1\Gamma_2)^2 + \frac{\gamma^2}{\rho^2}\right) + \frac{C\gamma^2 k\log(NT_1/L)}{\rho^2} + C\frac{(R\Gamma_1\Gamma_2)^2T_1}{T}$$

$$\leq C(\gamma,\Gamma_1,\Gamma_2)\left(\frac{(NT_1)^2R^2}{\rho^4\sigma_k(\mathrm{SP}'_1(f))^2L^2} + \frac{R^2k\log(NT_1/L)}{\rho^2} + \frac{R^2T_1}{T}\right).$$

■

Now, given these events, we will provide the high probability bound. Let $\bar{E} := \bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3 \cap \bar{E}_4 \cap \bar{E}_5$.

$$\mathbb{P}(\bar{E}^c) \leq \frac{C_0}{(NT)^{10}} + \frac{C_1}{(NT_1)^{10}}, \tag{5.91}$$

for some absolute constants $C_0, C_1 > 0$. Note that under event $\bar{E}_3$, we have that $\hat{\rho} \geq \rho\left(1 - \sqrt{\frac{20\log(NT)}{\rho NT}}\right)$. By further using the assumption $\rho \geq C\log(NT)/\sqrt{NT}$ for a sufficiently large C we have that $\hat{\rho} \geq C'\rho$ and $\frac{(\hat{\rho}-\rho)^2}{\hat{\rho}^2} \leq \frac{C}{\sqrt{NT}}$. Now, recall Δ_1 and Δ_2 definition in (5.77). Under event $\bar{E}$, we can bound Δ_1 as follows,

$$\Delta_1 = \widehat{\mathrm{SP}'_0} - \mathrm{SP}'_0(f)^2_{2,\infty}\beta^{*2}_{1} + \frac{1}{\rho}\left(\widehat{\mathrm{SP}'_0}^{\,T}(\widehat{\beta} - \beta^*)\right)^T\zeta^L_0$$

$$\leq C(\gamma, \Gamma_1, \Gamma_2)\beta^{*2}_1 \left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\text{SP}'_0(f))^2L^2} + \frac{kR^2\log(NT/L)}{\rho^2} \right) + \frac{1}{\rho}\left(\widehat{\text{SP}'_0}^T(\widehat{\beta} - \beta^*) \right)^T \zeta_0^L.$$

Similarly, under event $\bar{E}$, we can bound Δ_2 as follows,

$$\Delta_2 \leq C(\gamma, \Gamma_1, \Gamma_2)\beta^{*2}_1 \left(\frac{NTR^2}{L\sigma_k(\rho\text{SP}'_0(f))^2} + \frac{1}{\sigma_k(\widehat{\text{SP}'_0})^2} \left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\text{SP}'_0(f))^2L^2} + \frac{kR^2\log NT/L}{\rho^2} \right) \right)$$

$$+ \frac{C}{\rho\sigma_k(\widehat{\text{SP}'_0})^2} \left(\left(\widehat{\text{SP}'_0}^T(\widehat{\beta} - \beta^*) \right)^T \zeta_0^L \right).$$

Further, using Weyl's inequality (see Lemma 5.14.4), we can bound $|\sigma_k(\widehat{\text{SP}'_0}) - \sigma_k(\text{SP}'_0(f))|$ as follows,

$$\begin{aligned} |\sigma_k(\widehat{\text{SP}'_0}) - \sigma_k(\rho\text{SP}'_0(f))| &= \frac{1}{\hat{\rho}}|\sigma_k(\widetilde{\mathbf{\Sigma}}_0) - \hat{\rho}\sigma_k(\text{SP}'_0(f))| \\ &\leq \frac{1}{\hat{\rho}}|\sigma_k(\widetilde{\mathbf{\Sigma}}_0) - \rho\sigma_k(\text{SP}'_0(f))| + \frac{|\hat{\rho} - \rho|}{\hat{\rho}}\sigma_k(\text{SP}'_0(f)) \\ &\leq \frac{\boldsymbol{\iota}_{02}}{\hat{\rho}} + \frac{|\hat{\rho} - \rho|}{\hat{\rho}}\sigma_k(\text{SP}'_0(f)) \end{aligned}$$

Under $\bar{E}$, and using property 5.4.2, we have that with probability of at least $1 - \frac{1}{(NT)^{10}}$,

$$\begin{aligned} \frac{|\sigma_k(\widehat{\text{SP}'_0}) - \sigma_k(\text{SP}'_0(f))|}{\sigma_k(\text{SP}'_0(f))} &\leq \frac{C(\gamma + R\Gamma_1\Gamma_2)\sqrt{NT/L}}{\rho\sigma_k(\text{SP}'_0(f))} + \frac{|\hat{\rho} - \rho|}{\hat{\rho}} \\ &\leq \frac{C(\gamma + R\Gamma_1\Gamma_2)\sqrt{k}}{\rho\sqrt{L}} + \frac{C}{\sqrt{NT}}. \end{aligned}$$

Using $\rho \geq C(\gamma + R\Gamma_1\Gamma_2)\sqrt{\frac{k}{L}}$ we get $\frac{1}{\sigma_k(\widehat{\text{SP}'_0})^2} \leq \frac{C}{\sigma_k(\text{SP}'_0(f))^2}$. Using property 5.4.2, we get the following bounds for Δ_1 and Δ_2,

$$\Delta_1 \leq C(\gamma, \Gamma_1, \Gamma_2, c)\beta^{*2}_1 kR^2 \left(\frac{NT}{L^2\rho^4} + \frac{\log(NT/L)}{\rho^2} \right) + \frac{1}{\rho}\left(\widehat{\text{SP}'_0}^T(\widehat{\beta} - \beta^*) \right)^T \zeta_0^L. \tag{5.92}$$

$$
\begin{aligned}
\Delta_2 &\le C(\gamma, \Gamma_1, \Gamma_2, c)\beta^{*2}_1\left(\frac{kR^2}{L\rho^2} + \frac{k^2R^2}{NT}\left(\frac{NT}{L^2\rho^4} + \frac{\log(NT/L)}{\rho^2}\right)\right) \qquad (5.93)\\
&+ \frac{Ck}{\rho NT}\left(\left(\widehat{\mathrm{SP}'_0}^{T}(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right)\\
&\le C(\gamma, \Gamma_1, \Gamma_2, c)\beta^{*2}_1 k^2R^2\left(\frac{1}{L\rho^2} + \frac{\log(NT/L)}{L}\right)\\
&+ \frac{Ck}{\rho NT}\left(\left(\widehat{\mathrm{SP}'_0}^{T}(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right),
\end{aligned}
$$

where $\rho \ge C(\gamma + R\Gamma_1\Gamma_2)\sqrt{\frac{k}{L}}$ is used to obtain the last inequality. Finally, using properties 5.4.2 and 5.4.3, $\hat{\rho} \ge C'\rho$, and (5.88), (5.92), and (5.93), we have under event $\bar{E}$,

$$
\begin{aligned}
\frac{1}{\hat{\rho}}\mathrm{SP}'_1(X)^T\widehat{\beta} - \mathrm{SP}'_1(f)^T\beta^{*2}_2 &\le C(\gamma, \Gamma_1, \Gamma_2, c)\left(\frac{k^3NT_1R^6}{L^2\rho^4} + \frac{RT_1}{T}\right)\beta^{*2}_1 \qquad (5.94)\\
&+ C(\gamma, \Gamma_1, \Gamma_2, c)\left(\frac{k^3R^6\log(NT/L)}{\rho^2}\left(\frac{NT_1}{L^2} + \frac{T_1}{T}\right) + \frac{kR^2\log(NT_1/L)}{\rho^2}\right)\beta^{*2}_1\\
&+ C(\gamma, \Gamma_1, \Gamma_2, c)\frac{R^4k^2T_1}{T\rho^3}\left(\widehat{\mathrm{SP}'_0}^{T}(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L.
\end{aligned}
$$

Expectation Bound. We get the bound in expectation using the high probability bound above, and by assuming that our forecast is bounded such that $|\bar{f}_n(T + L \times m')| \le R\Gamma_1\Gamma_2$ for $m' \in [T_1/L]$. Specifically, we have using (5.94) and (5.91),

$$
\begin{aligned}
\mathrm{TestForErr}(N, T, T_1, L) &= \frac{1}{(NT_1/L)}\mathbb{E}\left[\left\|\frac{1}{\hat{\rho}}\mathrm{SP}'_1(X)^T\widehat{\beta} - \mathrm{SP}'_1(f)^T\beta^*\right\|_2^2\right]\\
&\le \frac{1}{(NT_1/L)}\mathbb{E}\left[\left\|\frac{1}{\hat{\rho}}\mathrm{SP}'_1(X)^T\widehat{\beta} - \mathrm{SP}'_1(f)^T\beta^*\right\|_2^2 \middle| \bar{E}\right] + \frac{CR^2\Gamma_1^2\Gamma_2^2}{(N\min(T, T_1))^{10}}\\
&\le \frac{L}{NT_1}C(\gamma, \Gamma_1, \Gamma_2, c)\left(\left(\frac{k^3NT_1R^6}{L^2\rho^4} + \frac{RT_1}{T}\right)\beta^{*2}_1\right.\\
&+ \left(\frac{k^3R^6\log(NT/L)}{\rho^2}\left(\frac{NT_1}{L^2} + \frac{T_1}{T}\right) + \frac{kR^2\log(NT_1/L)}{\rho^2}\right)\beta^{*2}_1
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{R^4k^2T_1}{T\rho^3}\mathbb{E}\Big[\Big(\widehat{\mathrm{SP}'_0}^{T}(\widehat{\beta}-\beta^*)\Big)^T\zeta_0^L\Big|\bar{E}\Big]\Big)\\
&+ \frac{CR^2\Gamma_1^2\Gamma_2^2}{(N\min(T,T_1))^{10}}.
\end{aligned}
$$

Noting that the $\mathbb{E}[\zeta_0^L|\bar{E}] = \mathbf{0}$, and ζ_0^L is independent of $\widehat{\mathrm{SP}'_0}$, $\hat{\rho}$, β^* and the event $\bar{E}$; we have

$$
\mathbb{E}\Big[\Big(\widehat{\mathrm{SP}'_0}^{T}\beta^*\Big)^T\zeta_0^L\Big] = 0. \tag{5.95}
$$

By (5.3), we have $\widehat{\beta} = \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^{\dagger}\widetilde{\boldsymbol{V}}^T\mathrm{SP}_0(X)_{L\cdot}$. That is,

$$
\widehat{\beta} = \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^{\dagger}\widetilde{\boldsymbol{V}}^T\rho\mathrm{SP}_0(f)_{L\cdot} + \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^{\dagger}\widetilde{\boldsymbol{V}}^T\zeta_0^L. \tag{5.96}
$$

Using cyclic and linearity of Trace operator; the independence properties of ζ_0^L; and (5.96); we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\widehat{\mathrm{SP}'_0}^{T}\widehat{\beta}\Big)^T\zeta_0^L\Big] \qquad\qquad (5.97)\\
&= \mathbb{E}\Big[\Big(\widehat{\mathrm{SP}'_0}^{T}\widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^{\dagger}\widetilde{\boldsymbol{V}}^T\rho\mathrm{SP}_0(f)_{L\cdot}\Big)^T\zeta_0^L\;\Big] + \mathbb{E}\Big[\Big(\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T\zeta_0^L\Big)^T\zeta_0^L\Big]\\
&= \mathbb{E}[\mathrm{Tr}\Big((\zeta_0^L)^T\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T\zeta_0^L\Big)]\\
&= \mathrm{Tr}\,(\mathbb{E}[\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T]\mathbb{E}[\zeta_0^L(\zeta_0^L)^T])\\
&\le C(\gamma+\Gamma_1\Gamma_2R)^2k.
\end{aligned}
$$

Where to obtain the last inequality we use the trace property $\mathrm{Tr}(AB) \le B_2\mathrm{Tr}(A)$ for positive semi-definite matrices A, B, and that rank of $\widehat{\mathrm{SP}'_0}$ is k. Finally, using (5.97), and recalling that $T_1 \ge L$ and $L \le T$ we get,

$$
\begin{aligned}
\mathrm{TestForErr}(N,T,T_1,L) \le \frac{L}{NT_1}C(\gamma,\Gamma_1,\Gamma_2,c)\Bigg(&\Bigg(\frac{R^6k^3NT_1}{L^2\rho^4} + \frac{RT_1}{T}\Bigg)\beta^{*2}_{1}\\
&+\Bigg(\frac{R^6k^3\log(NT/L)}{\rho^2}\Bigg(\frac{NT_1}{L^2}+\frac{T_1}{T}\Bigg) + \frac{R^2k\log(NT_1/L)}{\rho^2}\Bigg)\beta^{*2}_{1} + \frac{R^6k^3T_1}{T\rho^3}\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{CR^2\Gamma_1^2\Gamma_2^2}{(NL)^{10}} \\
&\leq \frac{L}{NT_1} C(\gamma, \Gamma_1, \Gamma_2, c) \max(1, \beta^{*2}_1) \Bigg(\frac{R^6k^3NT_1}{L^2\rho^4} + \frac{R^6k^3T_1}{T\rho^3} \\
&+ \frac{R^6k^3\log(NT)}{\rho^2} \left(\frac{NT_1}{L^2} + \frac{T_1}{T} \right) + \frac{R^2k\log(NT_1)}{\rho^2} + \frac{R^2}{(NL)^{10}} \Bigg) \\
&\leq \frac{L}{NT_1} C(\gamma, \Gamma_1, \Gamma_2, c) \max(1, \beta^{*2}_1) \left(\frac{R^6k^3\log(NT)}{\rho^4} \left(\frac{NT_1}{L^2} + \frac{T_1}{T} \right) + \frac{R^2k\log(NT_1)}{\rho^2} \right).
\end{aligned}
$$

Then, with $L = \sqrt{\min(N, T)T}$, we get,

$$
\begin{aligned}
&\mathrm{TestForErr}(N, T, T_1, L) \\
&\leq \frac{\sqrt{\min(N, T)T}}{NT_1} C(\gamma, \Gamma_1, \Gamma_2, c) \max(1, \beta^{*2}_1) \left(\frac{R^6k^3\log(NT)}{\rho^4} \left(\frac{NT_1}{T\min(N, T)} + \frac{T_1}{T} \right) + \frac{R^2k\log(NT_1)}{\rho^2} \right) \\
&\leq \frac{T}{T_1} \frac{\sqrt{\min(N, T)T}}{NT} C(\gamma, \Gamma_1, \Gamma_2, c) \max(1, \beta^{*2}_1) \left(\frac{R^6k^3\log(NT)}{\rho^4} \left(\frac{NT_1}{T\min(N, T)} + \frac{T_1}{T} \right) + \frac{R^2k\log(NT_1)}{\rho^2} \right) \\
&\leq \frac{\sqrt{\min(N, T)T}}{NT} C(\gamma, \Gamma_1, \Gamma_2, c) \max(1, \beta^{*2}_1) \left(\frac{R^6k^3\log(NT)}{\rho^4} \left(\frac{N}{\min(N, T)} + 1 \right) + \frac{TR^2k\log(NT_1)}{T_1\rho^2} \right) \\
&\leq C(\gamma, \Gamma_1, \Gamma_2, c) \max(1, \beta^{*2}_1) \left(\frac{R^6k^3\log(N\max(T, T_1))}{\rho^4\sqrt{\min(N, T)T}} \left(\max(1, \frac{N}{T}) + \frac{T}{T_1} \right) \right).
\end{aligned}
$$

Choosing $k = RG$ completes the proof.

■ 5.19 Proof of Theorem 5.7.1

Setup, Notations. For $L \geq 1, k \geq 1$, for ease of notations, we define

- $\mathrm{SP}(X) = \mathrm{SP}((X_1, \ldots, X_N), T, L) \in \mathbb{R}^{L \times (NT/L)}$,
- $\mathrm{SP}(X^2) = \mathrm{SP}((X_1^2, \ldots, X_N^2), T, L) \in \mathbb{R}^{L \times (NT/L)}$,
- $\mathrm{SP}(f) = \mathrm{SP}((f_1, \ldots, f_N), T, L) \in \mathbb{R}^{L \times (NT/L)}$,

- $\text{SP}(f^2) = \text{SP}((f_1^2, \dots, f_N^2), T, L) \in \mathbb{R}^{L \times (NT/L)}$,
- $\text{SP}(\sigma^2) = \text{SP}((\sigma_1^2, \dots, \sigma_N^2), T, L) \in \mathbb{R}^{L \times (NT/L)}$,
- $\text{SP}(f^2 + \sigma^2) = \text{SP}(f^2) + \text{SP}(\sigma^2)$.

Recalling that $\rho = 1$, we note that

$$\mathbb{E}[\text{SP}(X)] = \text{SP}(f), \quad \mathbb{E}[\text{SP}(X^2)] = \text{SP}(f^2 + \sigma^2).$$

Further, from the definition of the variance estimation algorithm, we recall

$$\widehat{\text{SP}}(f) := \widehat{\text{SP}}((X_1, \dots, X_N), T, L) = \frac{1}{\hat{\rho}}\text{HSVT}_k(\text{SP}((X_1, \dots, X_N), T, L))$$
$$\widehat{\text{SP}}(f^2 + \sigma^2) := \widehat{\text{SP}}((X_1^2, \dots, X_N^2), T, L) = \frac{1}{\hat{\rho}}\text{HSVT}_k(\text{SP}((X_1^2, \dots, X_N^2), T, L))$$

We denote

- $\widehat{\text{SP}}(f^2) = \widehat{\text{SP}}(f) \circ \widehat{\text{SP}}(f)$
- $\widehat{\text{SP}}(\sigma^2) = \max\left(\widehat{\text{SP}}(f^2 + \sigma^2) - \widehat{\text{SP}}(f^2), \mathbf{0}\right)$,

where $\mathbf{0} \in \mathbb{R}^{L \times (NT/L)}$ is a matrix of all zeroes, and we apply the $\max(\cdot)$ above entry-wise. We remind the reader the output of the variance estimation algorithm is $\widehat{\text{SP}}(\sigma^2)$. Thus, we have

$$\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}(\sigma_n(t)^2 - \hat{\sigma}_n^2(t))^2 = \frac{1}{NT}\text{SP}(\sigma^2) - \widehat{\text{SP}}(\sigma^2)_F^2.$$

Initial Decomposition. Note that since $\sigma_n^2(t) \geq 0$ for $n \in [N]$ and $t \in [T]$, we have that

$$\begin{aligned}
&\frac{1}{NT}\text{SP}(\sigma^2) - \widehat{\text{SP}}(\sigma^2)_F^2 \\
&\leq \frac{1}{NT}\text{SP}(\sigma^2) - (\widehat{\text{SP}}(f^2 + \sigma^2) - \widehat{\text{SP}}(f^2))_F^2 \\
&= \frac{1}{NT}\text{SP}(f^2 + \sigma^2) - \text{SP}(f^2) - (\widehat{\text{SP}}(f^2 + \sigma^2) - \widehat{\text{SP}}(f^2)_F^2 \\
&\leq \frac{2}{NT}\text{SP}(f^2 + \sigma^2) - \widehat{\text{SP}}(f^2 + \sigma^2)_F^2 + \frac{2}{NT}\text{SP}(f^2) - \widehat{\text{SP}}(f^2)_F^2
\end{aligned} \tag{5.98}$$

We bound the two terms on the r.h.s of (5.98) separately.

Bounding $\mathbb{E}[\mathrm{SP}(f^2) - \widehat{\mathrm{SP}}(f^2)_F^2]$.

$$\begin{aligned}
\mathrm{SP}(f^2) - \widehat{\mathrm{SP}}(f^2)_F^2 &= \sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n^2(t) - \hat{f}_n^2(t)\right)^2 \\
&= \sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n(t) - \hat{f}_n(t)\right)^2\left(f_n(t) + \hat{f}_n(t)\right)^2 \\
&\leq \left[\max_{n\in[N],t\in[T]}\left(f_n(t) + \hat{f}_n(t)\right)^2\right]\left[\sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n(t) - \hat{f}_n(t)\right)^2\right] \\
&\overset{(a)}{\leq} C(\Gamma_1,\Gamma_2,\Gamma_3)R^2\left[\sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n(t) - \hat{f}_n(t)\right)^2\right] \\
&= C(\Gamma_1,\Gamma_2,\Gamma_3)R^2\left\|\mathrm{SP}(f) - \widehat{\mathrm{SP}}(f)\right\|_F^2 \qquad (5.99)
\end{aligned}$$

Bounding $\mathrm{SP}(f^2+\sigma^2)-\widehat{\mathrm{SP}}(f^2+\sigma^2)_F^2$. To bound $\mathrm{SP}(f^2+\sigma^2)-\widehat{\mathrm{SP}}(f^2+\sigma^2)_F^2$, we modify the proof of Theorem 5.5.1 in a straightforward manner. The need for the modification is that Theorem 5.5.1 was proven for the case where the coordinate wise noise, $\eta_n(t) = X_n(t)-f_n(t)$ are independent sub-gaussian random variables, and $\eta_{\psi_2}\leq \gamma$. However, one can verify that $X_n^2(t) - f_n^2(t) - \sigma_n^2(t)$ is a sub-exponential random variable with $\cdot_{\psi_1}$ norm bounded as

$$\begin{aligned}
\left\|X_n^2(t) - f_n^2(t) - \sigma_n^2(t)\right\|_{\psi_1} &\leq \left\|X_n^2(t)\right\|_{\psi_1} \\
&= \left\|f_n^2(t) + 2f_n(t)\eta_n(t) + \eta_n^2(t)\right\|_{\psi_1} \\
&\leq 2\left\|f_n^2(t)\right\|_{\psi_1} + 2\left\|\eta_n^2(t)\right\|_{\psi_1} \\
&= 2\left\|f_n(t)\right\|_{\psi_2}^2 + 2\left\|\eta_n(t)\right\|_{\psi_2}^2 \\
&\leq C(\Gamma_1,\Gamma_2)R^2 + 2\gamma^2 \\
&\leq C(\Gamma_1,\Gamma_2,\gamma)R^2,
\end{aligned}$$

where we have use the standard facts that for a random variable A, $A - \mathbb{E}[A]_{\psi_1}\leq A_{\psi_1}$ and $A^2{}_{\psi_1} = A_{\psi_2}^2$.

Further, note that by using Properties 5.3.1, 5.3.2, 5.7.1, and 5.7.2, and a straightforward modification of Proposition 15, we have

$$\begin{aligned}\text{rank}(\text{SP}(f^2+\sigma^2)) &\leq \text{rank}(\text{SP}(f^2)) + \text{rank}(\text{SP}(\sigma^2)) \\ &\leq (RG)^2 + (R'G'),\end{aligned}$$

where we have used that for any two matrices $\boldsymbol{A}, \boldsymbol{B}$, we have $\text{rank}(\boldsymbol{A} \circ \boldsymbol{A}) \leq \text{rank}(\boldsymbol{A})^2$, where $\circ$ denotes Hadamard product, and $\text{rank}(\boldsymbol{A}+\boldsymbol{B}) \leq \text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{B})$. We define $\tilde{k} := (RG)^2 + (R'G')$.

Modified Theorem 5.5.1. Below, we state the modified version of Theorem 5.5.1 to get our desired result.

Lemma 5.19.1 (Imputation Error)**.** *Let the conditions of Theorem 5.7.1 hold. Then,*

$$\begin{aligned}&\mathbb{E}[\max_{j\in[L]} \frac{1}{(NT/L)} \text{SP}(f^2+\sigma^2)^T_{L,\cdot} - \widehat{\text{SP}}(f^2+\sigma^2)^T_{L,\cdot}{}^2_2] \\ &\leq C(\Gamma_1, \Gamma_2, \Gamma'_1, \Gamma'_2, \gamma, R, R') \left(\frac{(G^2+G')\log^2 NT}{L}. \right),\end{aligned}$$

where $C(\Gamma_1, \Gamma_2, \Gamma'_1, \Gamma'_2, \gamma, R, R')$ is a term that depends only polynomially on Γ_1, Γ_2, Γ'_1, Γ'_2, γ, R, R'.

Proof. To reduce redundancy, we provide an overview of the argument needed for this proof, focusing only the parts of the arguments made in Theorem 5.5.1 that need to be modified. For ease of exposition, we let $\tilde{C} = C(\Gamma_1, \Gamma_2, \Gamma'_1, \Gamma'_2, \gamma, R, R')$. We being by matching notation with that used in Theorem 5.5.1; in particular with respect to $\rho, k, \epsilon, \Gamma$. Under the setup of Theorem 5.7.1, we have $\rho = 1$, $k = \tilde{k}$, $\epsilon = 0$, $\Gamma \leq \tilde{C}$ Further, recall the definition of $\boldsymbol{Y}, \boldsymbol{M}, p, q, \sigma$ from Appendix 5.15.1. We will now use $\boldsymbol{Y} = \text{SP}(X^2)$, and $\boldsymbol{M} = \text{SP}(f^2+\sigma^2)), \sigma = \gamma, p = (NT/L), q = L$. One can verify that there is only required change to the proof of Theorem 5.5.1; in particular, in the argument made to prove Theorem 5.15.1, we need to re-define events E_2, E_3, E_4 in (5.26), (5.27), (5.28) for the case where $(\boldsymbol{Y}-\boldsymbol{M})_{ij}$ is mean-zero sub-exponential. Using the result from Agarwal et al. (2019b, 2021e), which bounds the operator norm of a matrix with sub-exponential mean-zero entries, we have with probability at least $1 - 1/((NT)^{10})$

$$\boldsymbol{Y} - \boldsymbol{M}_2 \leq \tilde{C}\sqrt{(NT/L)}\log^2 NT \tag{5.100}$$

As a result (5.100), and standard concentration inequalities for sub-exponential random variables, we have the modified events, $\tilde{E}_2, \tilde{E}_3, \tilde{E}_4$.

$$\begin{aligned}
\tilde{E}_2 &:= \left\{ \left\|\boldsymbol{Y} - \rho \boldsymbol{M}\right\|_2 \leq \tilde{C}\sqrt{(NT/L)} \log^2 NT \right\}, \\
\tilde{E}_3 &:= \left\{ \left\|\boldsymbol{Y} - \rho \boldsymbol{M}\right\|_{\infty,2}, \left\|\boldsymbol{Y} - \rho \boldsymbol{M}\right\|_{2,\infty} \leq \tilde{C}\sqrt{(NT/L)} \log^2 NT \right\}, \\
\tilde{E}_4 &:= \left\{ \max_{j \in [q]} \left\| \varphi^{\boldsymbol{B}}_{\sigma_k(\boldsymbol{B})} \left(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot} \right) \right\|_2^2 \leq \tilde{C}\tilde{k} \log^2 (NT/L) \right\},
\end{aligned}$$

Using these modified events in the proofs of Theorem 5.15.1 and Theorem 5.5.1, and appropriately simplifying leads to the desired result. ■

By Lemma 5.19.1 and (5.41), we have that

$$\begin{aligned}
\frac{1}{NT}\mathbb{E}[\mathrm{SP}(f^2+\sigma^2) - \widehat{\mathrm{SP}}(f^2+\sigma^2)_F^2 &\leq \mathbb{E}[\max_{j \in [L]} \frac{1}{(NT/L)} \mathrm{SP}(f^2+\sigma^2)^T_{L,\cdot} - \widehat{\mathrm{SP}}(f^2+\sigma^2)^{T}_{L,\cdot}{}^{2}_{2}] \\
&\leq C(\Gamma_1, \Gamma_2, \Gamma'_1, \Gamma'_2, \gamma, R, R') \left(\frac{(G^2+G')\log^2 NT}{L}. \right). \qquad (5.101)
\end{aligned}$$

Completing proof. Substituting (5.99) and (5.101) into (5.98) and letting $L = \sqrt{\min(N,T)T}$

$$\frac{1}{NT}\mathrm{SP}(\sigma^2) - \widehat{\mathrm{SP}}(\sigma^2)_F^2 \leq C(\Gamma_1, \Gamma_2, \Gamma_3, \Gamma'_1, \Gamma'_2, \gamma, R, R') \left(\frac{(G^2+G')\log^2 NT}{\sqrt{\min(N,T)T}}. \right).$$

This completes the proof.

■ 5.20 tSSA Proofs

■ 5.20.1 Proof of Proposition 12

Consider $n \in [N]$, $\ell \in [L]$, $s \in [T/L]$. By Property 5.3.1,

$$\begin{aligned}
\mathsf{T}_{n\ell s} &= f_n((s-1) \times L + \ell) \\
&= \sum_{r=1}^{R} U_{nr} W_{r((s-1)\times L + \ell)}. \qquad (5.102)
\end{aligned}$$

The Hankel matrix induced by time series $W_{r\cdot}$ has rank at most G as per Property 5.3.2. The Page matrix associated with it is of dimension $L \times T/L$ with entry in its ℓ-th row and s-th column equal to $W_{r((s-1)\times L+\ell)}$. Since this Page matrix can be viewed as a sub-matrix of the Hankel matrix, it has rank at most G as well. That is, there exists vectors $w^r_{\ell\cdot}, v^r_{s\cdot} \in \mathbb{R}^G$ such that

$$W_{r((s-1)\times L+\ell)} = \sum_{g=1}^{G} w^r_{\ell g} v^r_{sg}. \tag{5.103}$$

From (5.102) and (5.103), it follows that

$$\begin{aligned} \mathbf{T}_{n\ell s} &= \sum_{r=1}^{R} U_{nr} \Big(\sum_{g=1}^{G} w^r_{\ell g} v^r_{sg} \Big) \\ &= \sum_{r\in[R],g\in[G]} U_{nr} w^r_{\ell g} v^r_{sg} \\ &= \sum_{r\in[R],g\in[G]} a_{n\ (r,g)} b_{\ell\ (r,g)} c_{s\ (r,g)}, \end{aligned} \tag{5.104}$$

where $a_{n\ (r,g)} = U_{nr}$, $b_{\ell\ (r,g)} = w^r_{\ell g}$ and $c_{s\ (r,g)} = v^r_{sg}$. Thus (5.104) implies that $\mathbf{T}$ has CP-rank at most R x G.

By the setup and model definition, it follows $\mathbb{T}_{n\ell s} = X_n((s-1) \times L + \ell)$. And $X_n((s-1) \times L + \ell) = \star$ with probability $1-\rho$ and $f_n((s-1) \times L + \ell) + \eta_n((s-1) \times L + \ell)$ with probability ρ, where $\eta_n((s-1) \times L + \ell)$ are independent and zero-mean. Therefore, it follows that the entries of $\mathbb{T}$ are independent and

$$\begin{aligned} \mathbb{E}[\mathbb{T}_{n\ell s}] &= \mathbb{E}[X_n((s-1) \times L + \ell)] \\ &= \rho f_n((s-1) \times L + \ell) \\ &= \rho \mathbf{T}_{n\ell s}. \end{aligned}$$

That is, $\mathbb{E}[\mathbb{T}] = \rho\mathbf{T}$. This concludes the proof.

■ 5.20.2 Proof of Proposition 13

From Property 5.7.4, and our choice of parameter L for mSSA ($L = \sqrt{\min(N, T)T}$) and tSSA ($L = \sqrt{T}$), we have that

$$\text{ImpErr}(N, T; \text{tSSA}) = \tilde{\Theta}\left(\frac{1}{\min\left(N, \sqrt{T}\right)^2}\right) = \tilde{\Theta}\left(\frac{1}{\min\left(N^2, T\right)}\right), \tag{5.105}$$

$$\text{ImpErr}(N, T; \text{mSSA}) = \tilde{\Theta}\left(\frac{1}{\sqrt{\min(N, T)T}}\right), \tag{5.106}$$

$$\text{ImpErr}(N, T; \text{ME}) = \tilde{\Theta}\left(\frac{1}{\min\left(N, T\right)}\right). \tag{5.107}$$

We proceed in cases.

Case 1: $T = o(N)$**.** In this case, from (5.105), (5.106), and (5.107), we have

$$\text{ImpErr}(N, T; \text{tSSA}),\ \text{ImpErr}(N, T; \text{mSSA}),\ \text{ImpErr}(N, T; \text{ME}) = \tilde{\Theta}\left(\frac{1}{T}\right)$$

Case 2: $N = o(T)$**.** In this case, from (5.105), (5.106), and (5.107), we have

$$\text{ImpErr}(N, T; \text{tSSA}) = \tilde{\Theta}\left(\frac{1}{N^2}\right), \tag{5.108}$$

$$\text{ImpErr}(N, T; \text{mSSA}) = \tilde{\Theta}\left(\frac{1}{\sqrt{NT}}\right), \tag{5.109}$$

$$\text{ImpErr}(N, T; \text{ME}) = \tilde{\Theta}\left(\frac{1}{N}\right).$$

In this case, we have

$$\text{ImpErr}(N, T; \text{tSSA}), \text{ImpErr}(N, T; \text{mSSA}) = \tilde{o}(\text{ImpErr}(N, T; \text{ME})).$$

It remains to compare the relative performance of tSSA and mSSA for the regime $N = o(T)$. Towards this, note from (5.108) and (5.109) that

$$\text{ImpErr}(N, T; \text{tSSA}) = \tilde{o}(\text{ImpErr}(N, T; \text{mSSA}))$$

$$\begin{aligned}
&\iff \frac{1}{N^2} = \tilde{o}(\frac{1}{\sqrt{NT}}) \\
&\iff T^{1/3} = o(N)
\end{aligned}$$

This completes the proof.

■ 5.20.3 Proof of Proposition 14

Proposition 16. *Let Properties 5.11.1, 5.3.2, and 5.4.1 hold. Then, for any* $1 \leq L \leq \sqrt{T}$*,* **HT** *has CP-rank at most* $R \times G$*. Further, all entries of* $\mathbb{HT}$ *are independent random variables with each entry observed with probability* $\rho \in (0, 1]$*, and* $\mathbb{E}[\mathbb{HT}] = \rho\mathbf{HT}$*.*

Consider $n_1, \dots, n_d \in [N_1] \times \dots \times [N_d]$, $\ell \in [L]$, $s \in [T/L]$. By Property 5.11.1,

$$\begin{aligned}
\mathbf{HT}_{n_1,\dots,n_d,\ell,s} &= f_{n_1,\dots,n_d}((s-1) \times L + \ell) \\
&= \sum_{r=1}^{R} U_{n_1,r} \dots U_{n_d,r} \; W_{r,((s-1)\times L+\ell)},
\end{aligned}$$

The rest of the proof follows in a similar fashion to that of Proposition 12.

Bibliography

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 2020.

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American Statistical Association*, 2010.

P.-A. Absil, A. Edelman, and P. Koev. On the largest principal angle between random subspaces. *Linear Algebra and its Applications*, 414(1):288 – 294, 2006. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2005.10.004. URL http://www.sciencedirect.com/science/article/pii/S0024379505004878.

Anish Agarwal and Rahul Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*, 2021.

Anish Agarwal, Vishal Misra, Bjoern Schelter, Devavrat Shah, Dennis Shen, Helen Shiells, and Claude Wischik. Applying low-rank tensor completion to an alzheimer's clinical trial: Personalized treatments & dropouts. *Working paper*.

Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Model agnostic time series analysis via matrix estimation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(3):40, 2018.

Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Model agnostic time series analysis via matrix estimation. *ACM Sigmetrics*, 2019a.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32:9893–9903, 2019b.

Anish Agarwal, Abdullah Alomar, and Devavrat Shah. tspdb: Time series predict db. *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, 133:27–56, 2021a.

Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. *arXiv preprint arXiv:2109.15154*, 2021b.

Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2021c.

Anish Agarwal, Devavrat Shah, and Dennis Shen. On principal component regression in a high-dimensional error-in-variables setting, 2021d.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Journal of the American Statistical Association*, pages 1–34, 2021e.

Anish Agarwal, Abdullah Alomar, and Devavrat Shah. On multivariate singular spectrum analysis and its variants. *ACM Sigmetrics*, 2022.

Gabriela Alexe, Sorin Alexe, Yves Crama, Stephan Foldes, Peter Hammer, and Bruno Simeone. Consensus algorithms for the generation of all maximal bicliques. 09 2003. doi: 10.1016/jdam.2003.09.004.

Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19:1–51, 2018.

Muhummad Amjad, Vishal Mishra, Devavrat Shah, and Dennis Shen. mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2), 2019.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009. ISBN 9780691120348.

Manuel Arellano and Bo Honore. Panel data models: Some recent developments. *Handbook of Econometrics*, 02 2000.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference in differences, 2020.

Orley C Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs, 1984.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, pages 1–41, 2021.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3082043.

Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4): 1229–1279, 2009. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/40263859.

Jushan Bai and Serena Ng. Matrix completion, counterfactuals, and factor analysis of missing data, 2020.

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

Marta Banbura and Michele Modugno. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29(1):133–160, 2014. URL https://EconPapers.repec.org/RePEc:wly:japmet:v:29:y:2014:i:1:p:133-160.

Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 417–445. JMLR.org, 2016. URL http://proceedings.mlr.press/v49/barak16.html.

Matteo Barigozzi and Matteo Luciani. Quasi maximum likelihood estimation of non-stationary large approximate dynamic factor models. *arXiv preprint arXiv:1910.09841*, 2019.

Alexandre Belloni, Victor Chernozhukov, Abhishek Kaul, Mathieu Rosenbaum, and Alexandre B. Tsybakov. Pivotal estimation via self-normalization for high-dimensional linear models with errors in variables. *arXiv:1708.08353*, 2017a.

Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B Tsybakov. Linear and conic programming approaches to high-dimensional errors-in-variables models. *Journal of the Royal Statistical Society*, 79:939–956, 2017b.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method, 2020.

Sergei Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, 1946.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1): 249–275, 2004.

Sohom Bhattacharya and Sourav Chatterjee. Matrix completion with data-dependent missingness probabilities. *arXiv preprint arXiv:2106.02290*, 2021.

Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, second edition, 1986.

Christopher M Bishop. Bayesian pca. In *Advances in neural information processing systems*, pages 382–388, 1999.

Juan Bógalo, Pilar Poncela, and Eva Senra. Understanding fluctuations through multivariate circulant singular spectrum analysis. *arXiv preprint arXiv:2007.07561*, 2020.

David Broomhead and Gregory King. *On the Qualitative Analysis of Experimental Dynamical Systems*, volume 11. 01 1986.

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank symmetric tensor completion from noisy data. *arXiv preprint arXiv:1911.04436*, 2019.

Changxiao Cai, H Vincent Poor, and Yuxin Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *International Conference on Machine Learning*, pages 1271–1282. PMLR, 2020.

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. *Operations Research*, 2021.

T. Tony Cai and Peter Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159 – 2179, 2006. doi: 10.1214/009053606000000830. URL https://doi.org/10.1214/009053606000000830.

T Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525, 2016.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

R.B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, pages 245–276, 1966.

Gary Chamberlain. Panel data. In Z. Griliches† and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 2, chapter 22, pages 1247–1318. Elsevier, 1 edition, 1984. URL https://EconPapers.repec.org/RePEc:eee:ecochp:2-22.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912275.

Mark K. Chan and Simon Kwok. The PCDID Approach: Difference-in-Differences when Trends are Potentially Unparallel and Stochastic. Working Papers 2020-03, University of Sydney, School of Economics, March 2020. URL https://ideas.repec.org/p/syd/wpaper/2020-03.html.

Guoqing Chao, Yuan Luo, and Weiping Ding. Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction*, 1(1):341–358, 2019. ISSN 2504-4990. doi: 10.3390/make1010020. URL http://www.mdpi.com/2504-4990/1/1/20.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43:177–214, 2015.

George H Chen, Devavrat Shah, et al. *Explaining the success of nearest neighbor methods in prediction*. Now Publishers, 2018.

Yudong Chen and Constantine Caramanis. Orthogonal matching pursuit with noisy and missing data: Low and high dimensional results. *arXiv preprint arXiv:1206.0823*, 2012.

Yudong Chen and Constantine Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *International Conference on Machine Learning*, pages 383–391, 2013.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls, 2020a.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. Practical and robust t-test based inference for synthetic control and related methods, 2020b.

Francois Chollet. keras. https://github.com/fchollet/keras, 2015.

Abhirup Datta and Hui Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426, 2017.

Mark A. Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016. doi: 10.1109/JSTSP.2016.2539100.

Mark A. Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion, 2014.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Marie-Hélène Descary, Victor M Panaretos, et al. Functional data analysis by matrix completion. *Annals of Statistics*, 47(1):1–38, 2019.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

David Donoho and Matan Gavish. The optimal hard threshold for singular values is. *IEEE Transactions on Information Theory*, 60, 05 2013. doi: 10.1109/TIT.2014.2323359.

N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Paper No. 22791*, 2016a.

N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Paper No. 22791*, 2016b.

Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A Quasi–Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics*, 94(4):1014–1024, 11 2012. ISSN 0034-6535. doi: 10.1162/REST_a_00225. URL https://doi.org/10.1162/REST_a_00225.

Facebook. Prophet. https://facebook.github.io/prophet/, 2020. Online; accessed 25 February 2020.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

Vivek Farias and Andrew Li. Learning preferences with side information. *Management Science*, 65, 05 2019. doi: 10.1287/mnsc.2018.3092.

Iván Fernández-Val, Hugo Freeman, and Martin Weidner. Low-rank approximations of nonseparable panel models. *arXiv preprint arXiv:2010.12439*, 2020.

Iván Fernández-Val, Hugo Freeman, and Martin Weidner. Low-rank approximations of nonseparable panel models, 2020.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82 (4):540–554, 2000. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/2646650.

Simon Funk. Netflix update: Try this at home, 2006.

Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. doi: 10.1109/TIT.2014.2323359.

M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40(1):3–1–3–41, 2002. doi: https://doi.org/10.

1029/2000RG000092. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000RG000092.

David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

Nina Golyandina, Vladimir Nekrutkin, and Anatoly A Zhigljavsky. *Analysis of time series structure: SSA and related techniques*. Chapman and Hall/CRC, 2001.

Google. Google LLC google covid-19 community mobility reports. https://www.google.com/covid19/mobility/, 2020. Accessed: 2020-06-01.

Loukas Grafakos. *Classical fourier analysis*, volume 2. Springer, 2008.

Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. *arXiv preprint arXiv:2004.03758*, 2020.

Peter Hall and Joel L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70 – 91, 2007. doi: 10.1214/009053606000000957. URL https://doi.org/10.1214/009053606000000957.

Peter Hall, Hans-Georg Müller, and Jane-Ling Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493 – 1517, 2006. doi: 10.1214/009053606000000272. URL https://doi.org/10.1214/009053606000000272.

Marc Hallin and Roman Liška. Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617, 2007. ISSN 01621459. URL http://www.jstor.org/stable/27639890.

Hossein Hassani and Rahim Mahmoudvand. Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics*, 1(01):55–83, 2013.

Hossein Hassani and Rahim Mahmoudvand. *Singular spectrum analysis: Using R*. Springer, 2018.

Hossein Hassani, Saeed Heravi, and Anatoly Zhigljavsky. Forecasting uk industrial production with multivariate singular spectrum analysis. *Journal of Forecasting*, 32(5):395–408, 2013.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015. URL http://jmlr.org/papers/v16/hastie15a.html.

Peter Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102:674–685, 02 2007. doi: 10.2307/27639896.

Cheng Hsiao, H. Steve Ching, and Shui Ki Wan. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740, 2012. doi: https://doi.org/10.1002/jae.1230.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society*, 31(3):300–303, 1982.

Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization, 2018.

Abhishek Kaul and Hira L Koul. Weighted ℓ_1-penalized corrected quantile regression for high dimensional measurement error models. *Journal of Multivariate Analysis*, 140:72–91, 2015.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010a.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010b.

Jon Kleinberg and Mark Sandler. Using mixture models for collaborative filtering. *Journal of Computer and System Sciences*, 74(1):49–69, 2008.

Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 426–434, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401944. URL https://doi.org/10.1145/1401890.1401944.

Yehuda Koren and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 77–118, 2015.

Christina E. Lee, Yihua Li, Devavrat Shah, and Dogyoon Song. Blind regression: Non-parametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems 29*, pages 2155–2163, 2016.

Kathleen T. Li. Inference for factor model based average treatment effects. *Available at SSRN 3112775*, 2018.

Kathleen T. Li and David R. Bell. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197(1):65 – 75, 2017. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2016.01.011.

Yehua Li and Tailen Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321 – 3351, 2010. doi: 10.1214/10-AOS813. URL https://doi.org/10.1214/10-AOS813.

Dawen Liang, Laurent Charlin, James McInerney, and David M. Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 951–961, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883090. URL https://doi.org/10.1145/2872427.2883090.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 04 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.1.13. URL https://doi.org/10.1093/biomet/73.1.13.

Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Po-ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3): 1637–1664, 2012.

Yuping Lu, Charles Phillips, and Michael Langston. Biclique: an r package for maximal biclique enumeration in bipartite graphs. *BMC Research Notes*, 13, 12 2020. doi: 10.1186/s13104-020-04955-0.

Bingqing Lyu, Lu Qin, Xuemin Lin, Ying Zhang, Zhengping Qian, and Jingren Zhou. Maximum biclique search at billion scale. *Proc. VLDB Endow.*, 13(9):1359–1372, May 2020. ISSN 2150-8097. doi: 10.14778/3397230.3397234. URL https://doi.org/10.14778/3397230.3397234.

Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *arXiv preprint arXiv:1910.12774*, 2019.

S Makridakis, E Spiliotis, and V Assimakopoulos. The m5 accuracy competition: Results, findings and conclusions. *Int J Forecast*, 2020.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80): 2287–2322, 2010. URL http://jmlr.org/papers/v11/mazumder10a.html.

Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2007/file/d7322ed717dedf1eb4e6e52a37ea7bcd-Paper.pdf.

Hyungsik Roger Moon and Martin Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579, 2015. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/43616977.

Hyungsik Roger Moon and Martin Weidner. Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195, 2017. doi: 10.1017/S0266466615000328.

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Master's Thesis*, 1923.

Vicente Oropeza and Mauricio Sacchi. Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. *Geophysics*, 76(3):V25–V32, 2011.

Art B. Owen and Patrick O. Perry. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564 – 594, 2009. doi: 10.1214/08-AOAS227. URL https://doi.org/10.1214/08-AOAS227.

M. Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3805914.

Guy Plaut and Robert Vautard. Spells of low-frequency oscillations and weather regimes in the northern hemisphere. *Journal of Atmospheric Sciences*, 51(2): 210 – 236, 1994. doi: 10.1175/1520-0469(1994)051<0210:SOLFOA>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/51/2/1520-0469_1994_051_0210_solfoa_2_0_co_2.xml.

Nikhil Rao, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. Collaborative filtering with graph information: Consistency and scalable methods. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2107–2115. Curran Associates, Inc., 2015.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011. doi: 10.1109/TIT.2011.2165799.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

David S. Stoffer Robert H. Shumway. *Time Series Analysis and It's Applications*. Blue Printing, 3rd edition, 2015.

Steven Roman. Graduate texts in mathematics: Advanced linear algebra. *Springer*, 2008.

Mathieu Rosenbaum and Alexandre B. Tsybakov. Sparse recovery under matrix estimation. *The Annals of Statistics*, 38(5):2620–2651, 2010.

Mathieu Rosenbaum and Alexandre B. Tsybakov. Improved matrix uncertainty selector. *From Probability to Statistics and Back: High-Dimensional Models and Processes*, 9: 276–290, 2013.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 2019.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1670–1679, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/schnabel16.html.

Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems*, pages 4838–4847, 2019.

D. Shah and C. L. Yu. Iterative collaborative filtering for sparse noisy tensor estimation. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 41–45, 2019. doi: 10.1109/ISIT.2019.8849683.

Devavrat Shah and Christina Lee Yu. Iterative collaborative filtering for sparse noisy tensor estimation. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 41–45. IEEE, 2019.

Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020a.

Aude Sportisse, Claire Boyer, and Julie Josses. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33, 2020b.

Chandler Squires, Dennis Shen, Anish Agarwal, Devavrat Shah, and Caroline Uhler. Causal imputation via synthetic interventions, 2021.

Nathan Srebro and Russ R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In J. Lafferty, C. Williams, J. Shawe-Taylor,

R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. volume 17, 11 2004.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460): 1167–1179, 2002. ISSN 01621459. URL http://www.jstor.org/stable/3085839.

Gilbert Strang. Linear algebra and its applications. *Brooks/Cole Cengage Learning*, 2006.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3): 611–622, 1999.

Artur Trindade. UCI machine learning repository - individual household electric power consumption data set. 2014. URL https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019. doi: 10.1137/18M1183480.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Menghan Wang, Mingming Gong, Xiaolin Zheng, and Kun Zhang. Modeling dynamic missingness of implicit feedback for recommendation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper/2018/file/8d9766a69b764fefc12f56739424d136-Paper.pdf.

Menghan Wang, Xiaolin Zheng, Yang Yang, and Kun Zhang. Collaborative filtering with social exposure: A modular approach to social recommendation, 2018b. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16058.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6638–6647. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/wang19n.html.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. Causal inference for recommender systems. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, page 426–431, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3412225. URL https://doi.org/10.1145/3383313.3412225.

Larry Wasserman. *All of nonparametric statistics*. Springer, 2006.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Kevin W Wilson, Bhiksha Raj, and Paris Smaragdis. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

Nick Wilson, Amanda Kvalsvig, Lucy Telfar Barnard, and Michael G Baker. Case-fatality risk estimates for covid-19 calculated by using a lag time for fatality. *Emerging infectious diseases*, 26(6):1339, 2020.

Svante Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978. ISSN 00401706.

WRDS. The trade and quote (taq) database. 2021. URL https://wrds-www.wharton.upenn.edu/pages/support/data-overview/wrds-overview-taq/.

Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics, January 2020.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries, 2018.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. *arXiv preprint arXiv:1709.03183*, 2017a.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017b. doi: 10.1017/pan.2016.2.

Chengrun Yang, Lijun Ding, Ziyang Wu, and Madeleine Udell. Tenips: Inverse propensity sampling for tensor completion. *arXiv preprint arXiv:2101.00323*, 2021.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873 – 2903, 2005. doi: 10.1214/009053605000000660. URL https://doi.org/10.1214/009053605000000660.

Christina Lee Yu. Tensor estimation with nearly linear samples. *arXiv preprint arXiv:2007.00736*, 2020.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in neural information processing systems*, pages 847–855, 2016.

Yun Zhang, Charles Phillips, Gary Rogers, Erich Baker, Elissa Chesler, and Michael Langston. On finding bicliques in bipartite graphs: A novel algorithm and its application to the integration of diverse biological data types. *BMC bioinformatics*, 15:110, 04 2014. doi: 10.1186/1471-2105-15-110.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020. URL http://jmlr.org/papers/v21/19-545.html.