Causal Inference: a Tensor's Perspective

by

Dennis Shen

B.S. in Electrical Engineering, University of California San Diego, June 2015

S.M. in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, February 2018

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

September 2020

Signature of Author: ____

Dennis Shen
Department of Electrical Engineering and Computer Science
August 28, 2020

Certified by: ____

Devavrat Shah
Professor of Electrical Engineering and Computer Science
Director of Statistics and Data Science Center
Thesis Supervisor

Certified by: ____

Mark Abramson
Principal Member Technical Staff Draper Laboratory, Inc.
Thesis Co-Supervisor

Accepted by: ____

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

Causal Inference: a Tensor's Perspective

by Dennis Shen

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

Abstract

Quantifying the causal effect of an intervention is a ubiquitous problem that spans a wide net of applications. Typically, this quantity is measured through the difference in outcomes under treatment (e.g., novel drug) and control (e.g., placebo). However, only one outcome ever be revealed – this is the fundamental challenge in causal inference. In order to overcome this obstacle, there have been two main types of studies: experimental (ES) and observational (OS). While the former conducts carefully designed experiments, the latter utilizes observed data.

In this thesis, we reinterpret the classical potential outcomes framework of Rubin through the lens of tensors. Formally, each entry of the potential outcomes tensor is associated with a tuple of entities; namely, the measurement (e.g., time), unit (e.g., patient type), and intervention (e.g., drug). Subsequently, each study can be characterized by a unique sparsity pattern, which allows us to translate the age old problem of estimating counterfactuals into one of tensor estimation. As an added benefit, our tensor formulation also opens the door to discussions about the computational and statistical trade-offs of causal inference methods, a conversation (to the best of our knowledge) that has largely not yet been had.

Ultimately, this novel perspective, coupled with basic principles of the popular synthetic control method for OSs, enables us to provably estimate counterfactual potential outcomes for every unit under all treatments and control with low sample and computational complexity. As a result, we can customize treatment plans for every unit in a computationally tractable and data-efficient manner. Pleasingly, we show that this result bears implications towards what-if scenario planning, drug discovery, and personalized, data-efficient randomized control trials. Methodically, we furnish a data-driven hypothesis test to check when our algorithm can reliably recover the underlying tensor.

The key technical contribution of this thesis advances the state-of-the art analysis for principal component regression.

Thesis Supervisor: Devavrat Shah
Title: Professor of Electrical Engineering and Computer Science
Director of Statistics and Data Science Center

Thesis Co-Supervisor: Mark Abramson
Title: Principal Member Technical Staff Draper Laboratory, Inc.

Acknowledgments

Five years ago, I stumbled upon one of those "good" problems to have: I was fortunate to have the privilege of deciding where to attend graduate school. Initially, I struggled between staying in California and venturing out east to Boston, but I was lucky to have met a friend during my school visits who avidly insisted I come here. In particular, he urged me to reach out to this *young superstar professor* in LIDS. Thankfully, I took his advice and emailed Devavrat, and even more thankfully, Devavrat took a gamble on me. It's been an absolute honor and blessing to have been advised by Devavrat, who seems to never run out of energy or ideas when it comes to research. To this day, he still goes through every line of proofs, manages to remember all of the little details of my projects despite juggling so many responsibilities, and is always happy to have impromptu meetings (particularly, when coffee is involved). Devavrat has allowed me to explore my interests at my own pace, and has counterweighted with guidance and nudges when appropriate. I really can't thank Devavrat enough, and if anything good comes of me in my future, be it academically or personally, it will be largely be because of his influence.

I am also grateful to my committee members, Anette Peko Hosoi and Caroline Uhler, for their mentorship during my graduate studies and guidance for my future career directions. Anyone who knows me knows that I am a massive sports fanatic. Fortunately, Peko has been extremely supportive in giving me opportunities to explore the intersection of sports and technology, ranging from one-on-one meetings and group lunch discussions to participating at the annual MIT Sports Summit – including a dinner with the Spurs CEO, R.C. Buford! At the same time, I am so happy that I had the opportunity to collaborate with and learn from Caroline and her lab through our joint project working with biological data. Previously, I could never quite get myself interested in the life sciences, but this project has flipped the script and motivated me to pursue future endeavors in applying statistics and machine learning to in-vitro and clinical research.

Further, I greatly appreciate Mark Abramson and John Irvine for their guidance and support through the years. Both Mark and John have always made time for me out of their busy schedules to discuss new applications for my research, and given me fresh perspectives. More broadly, I am thankful to Draper for supporting me; in particular, for paying my bills and feeding me for the past several years. I am also thankful towards NPSC, NSA, and MIT for also funding my research.

Of course, none of this would have been possible without my amazing collaborators: Jehangir Amjad, Anish Agarwal, Abdullah Alomar, Romain Cosson, Vishal Misra, Dogyoon Song, and Chandler Squires. I am beyond thankful that I have had the pleasure to work with each of you; you have all taught me so much, many times even beyond research, and helped me grow as both a researcher and individual. I will always cherish our fun conversations, and I hope to work on more projects together in the future. I also want to send a big thank you to Alberto Abadie, Rahul Singh, Zhi Xu (the mystery friend who pushed me to come to Boston), and Christina Lee Yu. Although I haven't (yet!) had the opportunity to work on a project with any of you, each of you has still taken the time to diligently discuss my work and provide immensely helpful feedback. For obvious reasons, I want to give an extra thanks to Alberto who is the mastermind behind synthetic control, since I might not have had a thesis without his brilliance.

More broadly, I am thankful to SPPIN and the LIDS/IDSS/MIT community at large for making my time here so enjoyable. I really appreciate all of the staff members – with special thanks Rachel Cohen, Lynne Dell (congratulations on your retirement!), Gracie Gao, Jeannille Hiciano, Francisco Jaimes, Brian Jones, and Richard Lay – for keeping everything running so smoothly. I am also so grateful for the friendships that I have built during my time here, which have certainly greatly enhanced my student life and mental (and even physical) health.

I am deeply indebted to my amazing mentors and friends through the years, here in Boston and back home in California. I want to first thank my professors at UCSD – particularly, professors Shadi Dayeh, Sujit Dey, Paul Siegel, and Mohan Trivedi – who not only incited and sustained my interest in engineering and research, but also supported all of my goals. This thesis would have never seen the light of day if it wasn't for their guidance and encouragement. At the same time, I am grateful for all of my buddies, especially Behnam Hedayatnia and Dillon Quan, from UCSD – HKN4lyfe! I also want to send a shout-out to my intramural basketball team for all of the super fun games and weekend get-togethers; regardless of the score, we always won in my book. Further, I am thankful for my awesome roommates, and want to highlight Zach Ferris for feeding me during the week leading up to my defense and Anton Nguyen and Stephanie McCoy for always playing Nintendo Switch with me. Finally, I am especially thankful for my closest group of friends who've been with me from the very beginning, the Dunedaine clain: Robert Friedland, Chris Huh, Alan Hung, Logan Ma, Alex Tsai, and Jonathan Yeh. Our group name has gone through several variations through the years, but I am so blessed that our friendship has never wavered.

Everything I have is because of my family. I owe my blessings to their love and sacrifices. As such, I want to dedicate this work to them; particularly, my grandparents who are no longer with us today. I hope I have made you proud.

Ten years ago, my parents brought me to Boston for the first time. At that stage of my life, my world solely revolved around art, music, and, of course, the Lakers (rest in peace, Kobe Bryant). Unsurprisingly, I was quite unhappy and felt out-of-place to be spending an afternoon in Killian Court. Fast forward eight years, and we're in the exact same spot but under entirely different circumstances – celebrating my master's graduation. And now, I'm about to wrap up my graduate studies. Every step of this transformation was because of my mom and dad's sacrifices, and unconditional love and encouragement. Like a kite floating with the wind, I had the freedom to happily explore my various interests and find my own path, with the secure knowledge that both my mom and dad were always holding onto me for support. I am so amazed by my parent's love for me, and I am beyond thankful that my mom and dad are not only the best parents I could have ever asked for, but also my best friends.

I am blessed to have met so many devoted supporters throughout my life. In accordance with the spirit of synthetic control, a central aspect of this thesis, I am a convex combination of every individual above. Without them, I certainly would not be where I am today. Thank you all so much.

Contents

List of Figures

Chapter 1

Introduction

Quantifying the causal effect of an intervention is a problem of interest across a wide array of domains. From policy making to engineering and medicine, estimating the treatment effect is critical to understanding existing systems and moving towards innovation. Typically, this quantity is measured by the difference in outcomes under treatment (e.g., novel drug) and control (e.g., placebo or standard therapy). However, only one outcome can ever be revealed – this is the fundamental challenge of causal inference.

Traditionally, there have been two primary, distinct types of studies: experimental studies (ESs) and observational studies (OSs). While the former conducts carefully designed experiments, the latter utilizes observed data. One canonical ES is a randomized control trial (RCT); as its name suggests, RCTs randomly assign eligible participants to either a treatment or control group. Because the assignments are random (thus reducing biases and confoundedness), the differences between the groups can often be attributed to the treatment, i.e., the treatment is the cause. For this reason, RCTs are considered the gold standard mechanism to draw causal conclusions. However, due to practical and ethical concerns, ESs are not always feasible. This gives rise to OSs, which provide an alternate mechanism to enable causal inferences and may be the only way to explore certain questions.

In this thesis, we follow the classical potential outcomes framework of Rubin, and reinterpret it through the lens of tensors. More formally, we encode our data into a tensor, where each entry is the potential outcome associated with a tuple of entities; namely, the measurement (e.g., time), unit (e.g., patient type), and intervention (e.g., drug). Through this lens, we associate the observations of each study with a particular sparsity pattern, and recast its aim as recovering aspects of the tensor. In general, we translate the age old problem of estimating unobservable counterfactuals into one of tensor estimation. This perspective begs the following questions: (1) *modeling*: which sparsity patterns allow recovery? (2) *algorithmic*: if recovery is possible, what are the computational and statistical trade-offs? Indeed, if such an approach exists, then customized treatment plans for every unit can be achieved in computationally tractable and data-efficient manner.

■ 1.1 Motivating Applications

The problem of estimating treatment effects is ubiquitous, spanning a wide variety of fields. As such, we consider several impactful problems, which will serve as motivation for the rest of this thesis.

Example 1.1.1 (Importance of Controls in OSs: Evaluating the Impact of Gun Restrictions on Violence)**.** On November 8, 2016, in the aftermath of several high profile mass-shootings, voters in California passed Proposition (Prop.) 63 into law BallotPedia (2016). Prop. 63 "outlaw[ed] the possession of ammunition magazines that [held] more than 10 rounds, requir[ed] background checks for people buying bullets," and was proclaimed as an initiative for "historic progress to reduce gun violence" McGreevy (2016). Imagine that we wanted to study the impact of Prop. 63 on the rates of violent crime in California. Although RCTs are ideal mechanisms to draw causal conclusions, they are not applicable here since only one California exists. Instead, a statistical comparative study could be conducted where the rates of violent crime in California are compared to a "control" state after November 2016, which we refer to as the post-intervention period. To reach a statistically valid conclusion, however, the control state must be demonstrably similar to California sans the passage of a Prop. 63 style legislation. In general, there may not exist a natural control state for California, and subject-matter experts tend to disagree on the most appropriate state for comparison.

Example 1.1.2 (Importance of "Synthetic" Interventions: What-if Scenario Planning for COVID-19)**.** It is clear that the COVID-19 pandemic has led to an unprecedented disruption of modern society at a global scale. What is much less clear, however, is the effect that various interventions that have been put into place have had on health and economic outcomes. For example, perhaps a 30% and 60% clampdown in mobility have similar societal health outcomes, yet vastly different implications for the number of people who cannot go to work or file for unemployment. Having a clear understanding of the trade-offs between these interventions is crucial in charting a path forward on how to open up various sectors of society. A key challenge is that policy makers do not have the luxury of actually enacting a variety of interventions and seeing which has the optimal outcome (a la RCTs). In fact, at a societal level, this is simply infeasible and socially irresponsible. Arguably, an even bigger challenge is that the COVID-19 pandemic, and the resulting policy choices ahead of us, are unprecedented in scale. Thus, it is difficult to reliably apply lessons from previous pandemics (e.g., SARS, H1N1). This is only further exacerbated when taking into the account the vastly different economic, cultural, and societal factors that make each town/city/state/country unique. Although epidemiological models (e.g., SIR, SIRS) can shed some insight, they often make strong parametric assumptions. Therefore, in order to understand the unique trade-offs between different policies for every region (while avoiding a heavy reliance on parametric modeling), we need a data-driven, statistically principled way to estimate their potential outcomes under these *policies* before having to actually enact them.

Example 1.1.3 (Towards Personalized RCTs: Development Economics)**.** In the study of Banerjee et al. (2018), the authors collaborated with the Haryana state government in implementing the first large scale evaluation of the effects of different types of interventions on childhood immunization rates. The Haryana immunization trials were conducted with 2523 villages, with data collected monthly over 13 months, and included a total of 74 different interventions. As is standard in RCTs, the authors in Banerjee et al. (2018) randomly partitioned the 2523 villages into 74 groups, corresponding to the 74 different interventions they aimed to study. They then measured the average increase in immunization rates for each of these 74 groups over the 13 month trial period. Subsequently, they made a single policy recommendation to the Haryana state government, corresponding to the intervention that yielded the highest average increase in immunization rates. A core assumption in such RCTs is that villages are homogenous (i.e., all interventions have essentially the same effect on all units) and thus a blanket policy works well. However, this assumption is often violated, and the inherent diversity between different groups of people should be increasingly taken into consideration.

Example 1.1.4 (Towards Data-Efficient, Personalized RCTs: Drug Discovery & Precision Medicine)**.** Consider an FDA approved clinical trial with D new candidate drugs and N patient types. The goal is to prescribe the optimal drug for each of the N patient types. In an ideal world, this can be achieved by administering every drug to each patient type, amounting to $N \times D$ RCTs. Although this framework enables our desired level of personalization, in almost all scenarios, the number of required trials is simply infeasible. Within clinical trials, patient recruitment and compliance is especially costly due to monetary expenses and ethical considerations (e.g., placebo trials). Therefore, the name of the game is to design an experimental protocol and inferencing scheme that can achieve the personalization of the ideal setting yet retain the feasibility of standard RCTs, which are the bread and butter of clinical research. Indeed, the potential application of such a framework, especially in the context of personalized drug design or clinical trials, can have a large impact.

■ 1.2 Problem Statement

Throughout, we follow the potential outcomes framework of Neyman (1923) and Rubin (1974a). More formally, we are interested in outcomes associated with $N \geq 1$ units, across $T \geq 1$ measurements, $P \geq 1$ metrics, and $D \geq 1$ possible interventions. Unless stated otherwise, we index units with $n \in [N]$, measurements with $t \in [T]$, metrics with $p \in [P]$, and interventions with $d \in [D]$. Throughout, let $M_{tn}^{(d,p)}$ denote the potential outcome of the t-th measurement for unit n under intervention d and metric p; similarly, we define $Y_{tn}^{(d,p)}$ as the associated observed outcome, where $\mathbb{E}[Y_{tn}^{(d,p)}] = M_{tn}^{(d,p)}$. Without loss of generality, we denote the control or "null-intervention" with $d = 1$.

Tensor Framework

Crucially, we encode the universe of potential outcomes across measurements, units, interventions, and metrics into an order-four tensor $\boldsymbol{M} \in \mathbb{R}^{T \times N \times D \times P}$ (see Figure 1.1 for a graphical depiction). As we will see in Chapter 4, the tensor structure is a convenient representation to capture inter-dependencies along multiple dimensions.

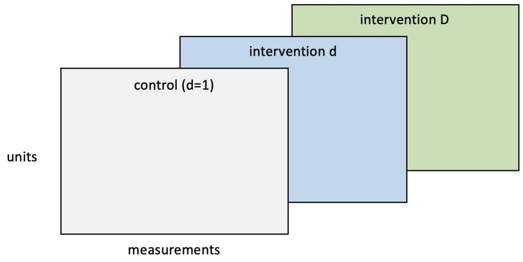


Figure 1.1: Tensor of potential outcomes $\boldsymbol{M}$ for a particular metric p.

Pre- and Post-Intervention Periods

Throughout, we will assume that all units are under a common intervention (e.g., control) for some number of measurements $T_0 \leq T$; this will partition our T measurements into two distinct segments: (i) the "pre-intervention" period, $t \leq T_0$, when all units are assumed, without loss of generality, to be in the no intervention state; and (ii) the "post-intervention" period, $t > T_0$, when each unit receives some intervention or remains unaffected. We group the units by the intervention they receive during the post-intervention period. That is, we denote

$$\mathcal{I}^{(d)} = \{n : \text{unit } n \text{ experiences intervention } d \text{ for all } t > T_0\} \tag{1.1}$$

as the subgroup of units that experience intervention d, and $N^{(d)} = |\mathcal{I}^{(d)}| \geq 1$ as its size.

Observations

As previously mentioned, every study will inevitably suffer from an inherent "missing data" problem, i.e., unobservable counterfactuals. Thus, each study can be represented by a unique observation pattern. Formally, we encode our observations into a sparse, noisy tensor $\boldsymbol{Z} = [Z_{tn}^{(d,p)}] \in \mathbb{R}^{T \times N \times D \times P}$ (see Figure 1.2), where

$$Z_{tn}^{(d,p)} = \begin{cases} Y_{tn}^{(1,p)} \cdot \pi_{tn}^{(1,p)}, & \text{for all } t \leq T_0, n, d = 1, p \in [P] \\ Y_{tn}^{(d,p)} \cdot \pi_{tn}^{(d,p)}, & \text{for all } t > T_0, n \in \mathcal{I}^{(d)}, d \in [D], p \in [P] \\ \star, & \text{otherwise;} \end{cases} \tag{1.2}$$

here, $\pi_{tn}^{(d,p)} \sim \text{Bernoulli}(\rho)$ with $\rho \in (0, 1]$ and $\star$ denotes an unobservable counterfactual. Note, this observation model includes the standard consistency assumption made in the potential outcomes literature (see Hernán and Robins (2020)) with the generalization that there may still be randomly missing data.

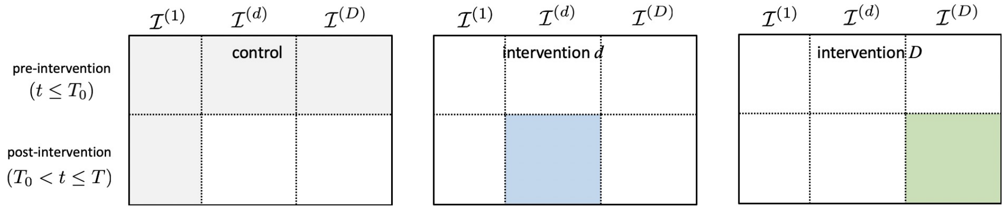


Figure 1.2: Observation tensor $\boldsymbol{Z}$ for a particular metric p, where each slice represents the observation patterns associated with the corresponding control or intervention. The latent counterfactuals are given in white, while the observations are given in gray, blue, and green.

Aim

Given $\boldsymbol{Z}$, we aim to infer potential outcomes for *every* unit under *all* interventions and metrics during the post-intervention period, i.e., $M_{tn}^{(d,p)}$ for all n, p, d, and $t > T_0$.

1.3 Synthetic Control

We begin by introducing synthetic control (SC) Abadie and Gardeazabal (2003); Abadie et al. (2010, 2014); Abadie (2019), whose principles will serve as the bedrock of our analytical framework and algorithmic developments. Over the years, SC has emerged as a standard tool to estimate the outcomes in the *absence* of an intervention using only OS data. Indeed, it has been regarded as "arguably the most important innovation in the policy evaluation literature in the last 15 years" Athey and Imbens (2016). In our context, it provides a solution for a restricted setting: $D = 2$ with $\mathcal{I}^{(1)} = [N] \setminus \{1\}$, $\mathcal{I}^{(2)} = \{1\}$, i.e., only unit 1 (often referred to in the SC literature as the "target" unit) experiences intervention 2 after T_0 while all other $N - 1$ "donor" units, i.e., $\mathcal{I}^{(1)}$, remain under the no-intervention state. The goal in SC is to infer the potential outcomes for unit 1 under the no-intervention state only, i.e., $M_{t1}^{(1,p)}$ for $t > T_0$ and some metric p.

Algorithm

To produce the counterfactuals, SC, as its name suggests, constructs a "synthetic" control for the target unit from donor unit data. Specifically, using pre-intervention data, a synthetic version of unit 1 is created as a weighted combination of the remaining $N - 1$ units. The learnt model is then used to produce counterfactual predictions for the target unit under the no-intervention state during the post-intervention period.

For simplicity, consider the single metric case ($P = 1$). Here, SC learns $\beta^{\text{sc}} \in \mathbb{R}^{N-1}$ as

$$\beta^{\text{sc}} \in \underset{w \in \text{SC}}{\arg\min} \sum_{t=1}^{T_0} \Big(Y_{t1}^{(1)} - \sum_{n \in \mathcal{I}^{(1)}} w_n Y_{tn}^{(1)} \Big)^2,$$

where the constraint set $\text{SC} \subseteq \mathbb{R}^{N-1}$ differs across variants of the method, but is classically taken to be over the probability simplex, i.e., $\beta_n^{\text{sc}} \geq 0$ and $\sum_n \beta_n^{\text{sc}} = 1$, (cf. Abadie and Gardeazabal (2003); Abadie et al. (2010)). Subsequently, $\widehat{Y}_{t1}^{(1),\text{sc}} = \sum_{n \in \mathcal{I}^{(1)}} \beta_n^{\text{sc}} Y_{tn}^{(1)}$ is the estimate for the target unit under no-intervention for $t > T_0$. Comparing $\widehat{Y}_{t1}^{(1),\text{sc}}$ with $Y_{t1}^{(2)}$ for $t > T_0$ allows us to evaluate the impact of intervention 2 on the target unit compared to the control.

Limitations

The last few decades have seen an unprecedented explosion in the availability of data across a myriad of domains. In many applications, however, datasets are plagued by high-dimensional, noisy, sparse, and mixed valued observations. Meanwhile, despite its widespread applicability and popularity, the original SC method was not designed to handle such scenarios.

Theoretically, the finite-sample properties of SC estimators is also surprisingly sparse. Abadie et al. (2010) proved that the classical SC estimator (i.e., where the weights are enforced to form a convex combination) is asymptotically unbiased; however, the authors not only assume the existence of a convex model, but also analyze the estimator using these latent convex weights (as opposed to the model estimates outputted from the algorithm) under a classical regression setting where the dimension of the covariates is fixed. In fact, across the different SC variants, it is typically assumed that a synthetic control for the target unit exists within the universe of donors (this is the fundamental hypothesis that drives SC-like methods); however, it is not clear when such a hypothesis holds, and meaningful finite-sample analysis that captures the behavior of the post-intervention prediction error with respect to the potential outcomes (rather than the observed outcomes) has remained elusive.

Additionally, given SC's widespread use across a plethora of domains, there are excellent heuristic hypothesis tests proposed in the literature (see Abadie (2019)) that serve as robustness checks for whether SC is valid to use; however, to the best of our knowledge, none of them are quantitative nor come with rigorous theoretical guarantees.

In summary, SC, though a powerful method, provides an incomplete answer to our objective laid out above – it only allows one to produce counterfactual estimates in the *absence* of an intervention, while we are also interested in making counterfactual estimates in the *presence* of an intervention. Indeed, extending SC to handle multiple interventions, as required in our setting, is an important open problem (see Abadie (2019)). Empirically, the classical SC estimator is ill-equipped to handle sparse, noisy, and high-dimensional datasets; and theoretically, the SC literature is missing an analytically motivated hypothesis test and a tighter finite-sample analysis with post-intervention prediction error rates.

■ 1.4 Summary of Results

This thesis provides a solution to the objective described above; that is, producing counterfactual estimates of potential outcomes under *all* interventions and metrics for *every* unit. To do so, we build upon both the SC estimator and framework, addressing the limitations detailed in Section 1.3. Along the way, we advance the theoretical analysis of Principal Component Regression (see Jolliffe (1982)), a key subroutine of our proposed SC variants, by viewing it through the lens of the Hard Singular Value Thresholding (HSVT) method (e.g., Chatterjee (2015b), Gavish and Donoho (2014)), which we also establish stronger guarantees for with respect to the $\ell_{2,\infty}$-norm rather than the Frobenius or spectral norms as is commonly done in the matrix estimation literature. Below, we provide an overview of our contributions.

■ 1.4.1 Principal Component Regression (Chapter 5)

We begin our journey by analyzing Principal Component Regression (PCR), which will serve as a key subroutine in our algorithms in producing counterfactual potential outcomes (see Algorithms 4, 5, 6). Therefore, in order to bound the post-intervention counterfactual prediction errors of our proposed algorithms, we first establish that PCR generalizes from training (pre-intervention) data to testing (post-intervention) data in a high-dimensional error-in-variable regression setting.

Problem Setup

In Linear Regression, the data is believed to be generated as per a latent linear model and the goal is to learn the linear predictor. More precisely, for each sample $i \leq n$, the response $y_i \in \mathbb{R}$ is linked to the underlying covariate $\boldsymbol{M}_i \in \mathbb{R}^p$ via the following model: $y_i = \langle \boldsymbol{M}_i, \beta^* \rangle + \epsilon_i$, where $\beta^* \in \mathbb{R}^p$ is the latent model parameter and ϵ_i denotes idiosyncratic noise. In an error-in-variable regression setting, we are given access to a labeled dataset $\{(y_i, \boldsymbol{Z}_i) : i \leq n\}$, where $\boldsymbol{Z}_i \in \mathbb{R}^p$ represents the observed, contaminated version of $\boldsymbol{M}_i$ that is to be utilized in the learning process.

It is well established that PCR is an effective prediction algorithm when the covariates exhibit low-rank structure. However, its ability to handle settings with noisy, missing, and mixed (discrete and continuous) valued covariates (i.e., learning with $\boldsymbol{Z}_i$ as opposed to $\boldsymbol{M}_i$) is not understood and remains an important open challenge, cf. Chao et al. (2019). As a contribution of this thesis, we establish the robustness of PCR in this respect, and provide meaningful finite-sample analysis with a theoretically motivated, data-driven hypothesis test to verify when PCR generalizes to unseen data.

Connection to HSVT

In order to prove the robustness and generalization properties of PCR, we first establish its connection to HSVT via Proposition 5.3.1. This allows us to analyze PCR through the lens of HSVT, which will prove to be a fruitful approach. Along the way, we prove that HSVT is a consistent estimator of the underlying mean matrix with respect to the $\ell_{2,\infty}$-norm (Lemma 5.9.6), which is a stronger guarantee than the standard Frobenius norm, i.e., a bound on the $\ell_{2,\infty}$-norm immediately implies a bound for the Frobenius norm.

Parameter Estimation

Formally, under the high-dimensional error-in-variable setting, we prove that PCR consistently learns the model parameter β^* with the error rate scaling as $1/n$, where n represents the number of training samples (Theorem 5.3.1). Compared to the rich literature of high-dimensional error-in-variable regression (cf. Loh and Wainwright (2012); Datta and Zou (2017); Rosenbaum and Tsybakov (2013)), our method achieves a similar error rate (with respect to n) for model identification *without* explicit knowledge of the underlying covariate noise model or a sparsity assumption on the model parameter; instead, we require the covariates to be low-rank. Moreover, the literature in error-in-variable regression often assumes a restricted eigenvalue condition on the covariate matrix, while we require the non-zero singular values of the covariate matrix to be well-balanced.

Test Prediction Error

Using our model identification result, we prove that PCR also achieves a test (out-of-sample) error rate of $1/n$ in expectation (see Corollary 5.3.2); we state its high probability version in Theorem 5.3.2. Of particular note, we underscore that Theorem 5.3.2 and Corollary 5.3.2 *do not* make any distributional assumptions. That is, while typical generalization error analyses (e.g., Rademacher complexity techniques) adopt an independent and identically distributed (i.i.d.) data generating assumption, our analysis relies on a purely linear algebraic "subspace inclusion" condition. This distinction is pivotal as i.i.d. assumptions can be unrealistic in our setting since potential outcomes from different interventions are likely to come from different distributions; more generally, the data generating process pre- and post-intervention may not be identically distributed.

Importantly, we highlight that the error-in-variable regression literature does not provide test prediction error bounds since the existing algorithms do not provide a method to "de-noise" corrupted test covariates; PCR, on the other hand, does provide a de-noising approach. We provide a summary of comparisons with notable works from the error-in-variable regression literature in Table 1.1.

Literature	Assumptions	Knowledge of Noise Distribution	Parameter Estimation	Test (Out-of-Sample) Prediction Error
Lasso-based	sparsity restricted eigenvalue cond.	Yes	$1/n$	–
This thesis	low-rank well-balanced spectra	No	$1/n$ (Thm. 5.3.1)	$1/n$ (Cor. 5.3.1)

Table 1.1: Comparison with some notable Lasso-based works Loh and Wainwright (2012); Datta and Zou (2017); Rosenbaum and Tsybakov (2013) in the high-dimensional error-in-variable regression literature.

Hypothesis Test

As aforementioned, our test error relies on a "subspace inclusion" property, which enables PCR to generalize to unseen data. Consequently, we furnish a simple, data-driven hypothesis test with provable guarantees to check for this condition in practice. Indeed, we argue that for any given significance level $\alpha \in (0, 1)$, the test statistic is smaller than an explicit critical value τ_α with probability at least $1 - \alpha$ (see Theorem 5.4.1). In the context of our causal inference framework, this serves as a quantitative test to validate when we can reliably extrapolate from our observed outcomes to estimate unobservable, counterfactual potential outcomes.

■ 1.4.2 Robust Synthetic Control (Chapter 6)

We begin our discussion of counterfactual estimation in the context of OSs by addressing the limitations of the SC method described in Section 1.3. As the primary contributions of Chapter 6, we prove the existence of a linear "synthetic" control for a target unit of interest, and propose robust synthetic control (RSC), a robust variant of the classical SC method to estimate potential outcomes under *control* given noisy and sparse observations, with provable statistical guarantees.

Problem Setup

We consider a standard OS (and SC) setting, where there is a single metric of interest (for simplicity) and target unit 1, which receives some intervention after $t > T_0$, with all other donors, i.e., $\mathcal{I}^{(1)}$, remaining unaffected during the entire time horizon T. Using our earlier notation, this translates to $P = 1$, and $D = 2$ with $\mathcal{I}^{(1)} = [N] \setminus \{1\}$, $\mathcal{I}^{(2)} = \{1\}$. Our goal is to infer the potential outcomes for the target unit in the *absence* of any intervention, i.e., $M_{t1}^{(1)}$ for all $t > T_0$. For a graphical depiction of the input and output of RSC, see Figure 1.3.

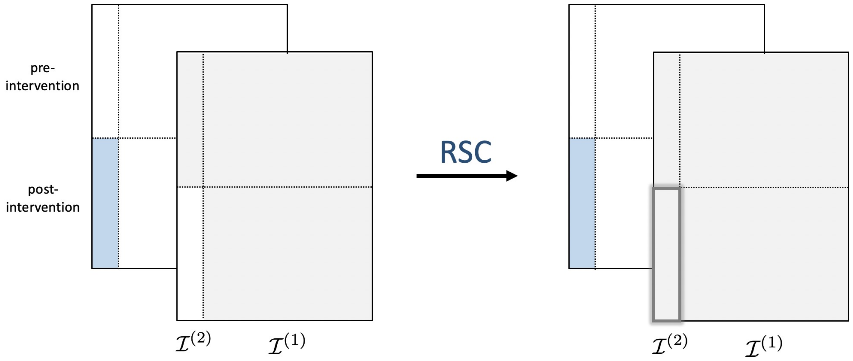


Figure 1.3: RSC predicts counterfactual potential outcomes under control (light gray) for the target unit, which is exposed to some intervention (blue) after T_0.

Existence of a Synthetic Control

In the SC literature, two standard assumptions are made: (i) First, the potential outcomes under control ($d = 1$) follow a linear factor model. Specifically, this model (also considered in Abadie et al. (2010)) states potential outcomes under control are $M_{tn}^{(1)} = \langle u_t, v_n \rangle$, where $u_t \in \mathbb{R}^r$ and $v_n \in \mathbb{R}^r$ are latent factors associated with measurement and unit, respectively. (ii) Second, there exists a linear (or even convex) relationship between the target and donor units.

As a contribution to the SC literature, we establish that a linear relationship between the target and donor units is actually *implied* with high probability under the factor model described above; thus, the existence of a linear synthetic control does not have to be separately assumed as an axiom as is traditionally done, cf. Abadie and Gardeazabal (2003); Abadie et al. (2010).

Robustness to Noise and Sparsity in a High-Dimensional Framework

Having established the robustness properties of PCR in the presence of noisy and sparse covariates in a high-dimensional setting, we introduce RSC, which utilizes PCR as a key subroutine (see Algorithm 4). Consequently, PCR's test error bounds (Theorem 5.3.2 and Corollary 5.3.2) immediately provide meaningful finite-sample post-intervention prediction error bounds for RSC (Theorem 6.2.1 and Corollary 6.2.1), which have been (to the best of our knowledge) absent in the literature. Specifically, we establish that, in expectation, the error scales as $\mathcal{O}(r/T_0)$; recall that T_0 denotes the length of the pre-intervention period and can thus be interpreted as the number of training samples, and r is the inherent model complexity (dimension of the latent spaces).

Empirical Studies: Importance of De-noising

We highlight the robustness properties of RSC through two canonical case studies: the economic impact of terrorism in Basque Country Abadie and Gardeazabal (2003) and the effect of Proposition 99 in the state of California Abadie et al. (2010). In both studies, we also utilize the subspace inclusion hypothesis test since it serves as a natural quantitative test for the validity of when to apply SC-like methods. Interestingly, our hypothesis test suggests that RSC (and possibly the classical SC method) should *not* be applied towards the Proposition 99 case study.

■ 1.4.3 Multi-dimensional RSC (Chapter 7)

We continue the discussion of estimating potential outcomes under control using observational studies via SC-like principles. In particular, we present multi-dimensional RSC (MRSC), a natural extension of RSC, to incorporate auxiliary metrics in a statistically principled manner, and provide theoretical guarantees to highlight its ability to overcome high levels of sparsity.

Problem Setup

We consider an extension to the setup in Section 1.4.2 with $P \geq 1$ metrics. Our interest remains in estimating the potential outcomes for the target unit ($n = 1$) in the absence of any intervention ($d = 1$) for some primary metric of interest p^* during the post-intervention period, i.e., $M_{t1}^{(1,p^*)}$ for all $t > T_0$ and some p^*. However, we now have access to auxiliary metrics of conforming dimension, i.e., $\boldsymbol{Z}^{(p)}$ for all p, which can be utilized to learn a model.

Incorporating Auxiliary Metrics

We extend the standard matrix factor model to a tensor factor model. Formally speaking, we assume the potential outcomes under control follow $M_{tn}^{(1,p)} = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell} w_{p\ell}$, where $u_t, v_n \in \mathbb{R}^r$ are defined as before and $w_p \in \mathbb{R}^r$ is the latent factor associated with metric p. Under this factor model, we again establish that a linear synthetic control exists within the reservoir of donors, which holds across all time and metrics. This suggests a natural extension of RSC (stated in Algorithm 5), which we refer to as multi-dimensional RSC (MRSC), that concatenates the pre-intervention data across all metrics in learning a single linear model, thereby augmenting the number of training samples by a factor of P.

Benefits of Auxiliary Metrics

As before, we utilize the PCR test error results to establish that the overall post-intervention prediction error scales as $\mathcal{O}(r/T_0)$ (see Theorem 7.2.1 and Corollary 7.2.1). However, the generalization error now decays as $\mathcal{O}(r/PT_0)$. Since the training (in-sample) error grows as $\mathcal{O}(r/T_0)$, the benefit of auxiliary metrics can only reduce the overall testing prediction error up to a certain point, irrespective of the amount of additional information. Therefore, the impact of auxiliary metrics is to help alleviate the problem of sparsity. More specifically, as opposed to requiring on the order of r entries per sample in our training set, we may now only need to observe r/P entries per sample.

Empirical Study: Overcoming Limited Training Data & Time Series Forecasting

MRSC's ability to overcome sparsity and limited training samples is further elucidated in an empirical retail case study of forecasting weekly sales at Walmart stores. Across all our experiments, we consistently find that MRSC significantly outperforms RSC when the pre-intervention data is small; however, the two methods perform comparably in the presence of substantial pre-intervention data. These empirical findings are in line with our theoretical results, i.e., in the presence of sparse training data, MRSC provides significant gains over RSC by utilizing information from auxiliary metrics.

Additionally, our mechanism for validating MRSC's performance is also an important and related contribution of this work: episodic time series prediction. Specifically, we propose a method to predict the future evolution of a time series based on limited data when the notion of time is relative and not absolute, i.e., where we have access to a donor pool that has already undergone the desired future evolution.

■ 1.4.4 Synthetic Interventions (Chapter 8)

Our journey culminates with the presentation of synthetic interventions (SI), a method that provides counterfactual estimates of potential outcomes for *each* unit under *all treatments* and *control*. We establish its theoretical performance and discuss its implications towards what-if analysis, drug repurposing, and personalized, data efficient RCTs.

Problem Setup

Here, we consider a significant extension to the setup described in Section 1.4.2 with $D \geq 2$ interventions of interest; for simplicity, we consider $P = 1$. Our objective is equivalent to that described in Section 1.2: to infer the potential outcomes for *every* unit under *all* interventions (including control), i.e., $M_{tn}^{(d)}$ for all n, d, and $t > T_0$. For a

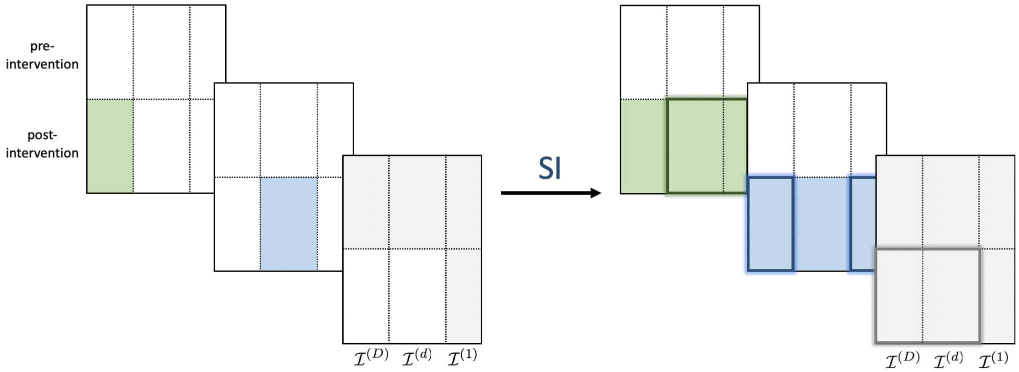

pre-
intervention
post-
intervention
SI
$\mathcal{I}^{(D)}$ $\mathcal{I}^{(d)}$ $\mathcal{I}^{(1)}$
$\mathcal{I}^{(D)}$ $\mathcal{I}^{(d)}$ $\mathcal{I}^{(1)}$

Figure 1.4: SI predicts counterfactual potential outcomes under control (light gray) and all treatments of interest (blue and green) for all units.

graphical depiction of the input and output of SI, see Figure 1.4; we highlight that the input matrix sparsity patterns reflect standard ES and OS data.

Estimating Counterfactuals Under Treatments

Methodologically, SI pleasingly turns out to be straightforward extension of SC, making it easy to implement. Specifically, as in the variants of SC, the model in SI is learnt using pre-intervention data under the no-intervention ($d = 1$) setting; however, to produce post-intervention counterfactual estimates, SI now applies the learnt model to *any* intervention d, including control (thus, SC can be viewed as a special instance of SI).

Transferring Between Interventional Frameworks

Although SI is methodologically similar to SC in terms of learning a model to estimate counterfactual outcomes, it is conceptually significantly different. In particular, it is not clear a priori why the model can be *transferred* between interventions. To establish the validity of SI, we consider a tensor factor model. Specifically, the potential outcomes follow $M_{tn}^{(d)} = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell} w_{d\ell}$, where u_t, v_n are defined as before, and $w_d \in \mathbb{R}^r$ now represents the latent factor associated with intervention d. Under this setting, we establish that there exists an invariant linear model that persists across measurements and interventions.

Moreover, we show SI produces consistent post-intervention counterfactual estimates for *all* units under *all* interventions. Formally, SI's post-intervention prediction error scales as $\mathcal{O}(r/T_0)$ in expectation (Corollary 8.2.1). The statement in high-probability, with explicit dependence on the noise parameters and model complexity, is given in Theorem 8.2.1.

Empirical Studies: Toward Personalized, Data-Efficient Treatments

Given that SI can estimate potential outcomes under *treatment* (as well as control) across all units, SI can effectively simulate treatment groups. As a result, we apply SI to several case studies to highlight its ability to enhance what-if analysis and improve RCTs, the gold standard mechanism in drawing causal conclusions. Most notably, we use real-world observational data to quantify the trade-offs between different policies to combat COVID-19 via SI. While standard OS methods (a la SC variants) can only infer the counterfactual death trajectories if countries did *nothing* to combat COVID-19, SI can additionally, and arguably more importantly, infer counterfactual trajectories if countries implemented different *policies* than what was actually enacted. Indeed, understanding the impact of various policies *before* having to actually enact them may provide guidance to policy makers in making statistically informed decisions as they weigh the difficult choices ahead of them. Furthermore, we use real-world experimental data from a large development economics and e-commerce website to perform data-efficient, personalized RCTs and A/B tests. Finally, we finish our whirlwind tour of case studies with an in-vitro cell-therapy study (with experimental data) that bears implications towards data-efficient drug discovery, thereby establishing SI's widespread applicability.

■ 1.4.5 Comparison with SC Literature

In Table 1.2, we list some of the key comparisons between SC and the extensions (i.e., RSC, MRSC, SI) presented in this thesis. More specifically, we highlight that (M)RSC addresses the limitations of the classical SC work by providing finite-sample guarantees for the post-intervention counterfactual prediction error and a quantitative hypothesis test to check when SC-like methods are appropriate for use, both of which have been missing in the literature. Finally, while the above methods can only produce counterfactual estimates for a target unit under control, SI can also provably estimate the counterfactual potential outcomes under treatment and for all units.

Literature	Intervention Framework			Theoretical Guarantees	
	# Interventions	Recipient	Counterfactual Estimates	Finite Sample	Hypo. Test
SC	1	Target	control	–	–
(M)RSC	1	Target	control	$1/T_0$	Yes
SI	≥ 1	Target, Donors	control, treatment	$1/T_0$	Yes

Table 1.2: Comparison between classical SC literature Abadie and Gardeazabal (2003); Abadie et al. (2010) and extensions presented in this thesis.

■ 1.5 Bibliographic Note

Preliminary versions of the results on PCR (Chapter 5) appeared in Agarwal et al. (2019). Similarly, preliminary versions on the generalizations to the synthetic control method to being robust to noise and missing data (Chapter 6) and incorporating multiple metrics (Chapter 7), i.e., analyzing the utility of side information, appeared in Amjad et al. (2018) and Amjad et al. (2019), respectively. Finally, synthetic interventions (Chapter 8) is currently under submission; a memo version that details an application of synthetic interventions towards COVID-19 and a full preprint version can be found at Agarwal et al. (2020b) and Agarwal et al. (2020a), respectively.

Chapter 2

Literature Survey

■ 2.1 Causal Inference

Causal inference has long been an interest for researchers from a wide array of communities, ranging from economics to machine learning (see Pearl (2009); Rubin (1974b, 1973); Paul R. Rosenbaum (1983) and the references therein). The focus of this work, however, will be on extending the synthetic control literature.

■ 2.1.1 Synthetic Control

Synthetic control (SC) has received widespread attention since its conception by Abadie and Gardeazabal in their pioneering work Abadie et al. (2010); Abadie and Gardeazabal (2003). It has been employed in numerous case studies, ranging from criminology Saunders et al. (2014) to health policy Kreif et al. (2015) to online advertisement to retail; other notable studies include Abadie et al. (2014); Billmeier and Nannicini (2013); Adhikari and Alm (2016); Aytug et al. (2016). Within the clinical realm, a growing trend is to apply SC-like methods to construct synthetic control arms in placebo studies syn (2019). In their paper on the state of applied econometrics for causality and policy evaluation, Athey and Imbens assert that synthetic control is "one of the most important development[s] in program evaluation in the past decade" and "arguably the most important innovation in the evaluation literature in the last fifteen years" Athey and Imbens (2016).

Among the many variants of SC, we note two of particular relevance here. As in our framework, Arkhangelsky et al. (2018) assumes that the observed data is a corrupted (additive noise model) of the true potential outcomes, which follows a factor model, i.e., a low-rank matrix; however, they do not allow for missing data beyond the unobservable counterfactuals. Here, the authors perform convex regression (with ℓ_2-norm constraints) along both the unit and measurement (time) axes (unlike standard SC methods, which only consider regression along the unit axis) to estimate the causal average treatment effect. As is standard, however, the authors of Arkhangelsky et al. (2018) assume that convex weights exist rather than prove its existence. Further, the results of Arkhangelsky et al. (2018) are asymptotic, i.e., they hold when $N, T \to \infty$, while our results are non-asymptotic. Another work that is less related, but is worth commenting on, as it also heavily relies on matrix estimation techniques for SC, is Athey et al. (2017). Here, the authors consider an underlying low-rank matrix of N units and T measurements per unit, and the entries of the observed matrix are considered "missing" once that unit has been exposed to a treatment. To estimate the counterfactuals, Athey et al. (2017) applies a nuclear norm regularized matrix estimation procedure. Some key points of difference are that the performance bounds are with respect to the Frobenius norm over all entries (i.e., units and measurements) in the matrix; meanwhile, we provide a stronger bound that is specific to the single treated unit and only during the post-intervention period.

We refer the reader to Abadie (2019) and references therein for a detailed overview of SC-like methods. Please refer to Table 1.2 and the related discussion in Section 1.4 for a comparison of our results with previous work in the SC literature.

■ 2.1.2 Heterogeneous Treatment Effects.

Randomized control trials (RCTs) are popular methods to study the average treatment effects (ATEs) when the units under consideration are approximately homogeneous. However, RCTs suffer when the units are highly heterogeneous, i.e., when each unit of interest might react very differently to each intervention. A complementary and exciting line of work to tackle this problem has been on estimating heterogeneous treatment effects (see Imbens and Rubin (2015) for a textbook style reference); here, the goal is to estimate the effect of a single intervention (or treatment), conditioned on a sufficiently rich set of covariates about a unit. The setting of SI differs from these works in two important ways: (i) it does not require covariate information regarding the units (this is in line with the work of Athey et al. (2017)), yet can estimate the heterogeneous treatment effect; (ii) it leverages the latent structure across interventions (via a tensor factor model) to estimate the optimal intervention per unit. An interesting line of future work would be to combine the literature on heterogeneous treatment effects with SI to exploit covariate information about units.

■ 2.2 Error-in-variable Regression

There exists a rich body of work regarding high-dimensional error-in-variable regression (see Loh and Wainwright (2012), Datta and Zou (2017), Rosenbaum and Tsybakov (2010), Rosenbaum and Tsybakov (2013), Belloni et al. (2017b), Belloni et al. (2017a), Chen and Caramanis (2012), Chen and Caramanis (2013), Kaul and Koul (2015)). Three common threads of these works include: (1) a sparsity assumption on the model parameter, β^*; and (2) an "incoherence"-like structure, such as the restricted eigenvalue condition (cf. Loh and Wainwright (2012) and references therein), on the underlying covariate matrix. In all of these works, the goal is to recover the underlying model, β^*. Some notable works include Loh and Wainwright (2012), Datta and Zou (2017), Rosenbaum and Tsybakov (2013), which are described in some more detail next.

In Loh and Wainwright (2012), a non-convex ℓ_1-penalization algorithm is proposed based on the plug-in principle to handle covariate measurement errors. This approach requires explicit knowledge of the unobserved noise covariance matrix Σ_H and the estimator *changes* based on their assumption of Σ_H. They also require explicit knowledge of a bound on $\|\beta^*\|_2$, the object they aim to estimate. In contrast, PCR does not require any such knowledge about the distribution of the noise matrix (i.e., PCR does not explicitly use this information to recover the model parameter or make predictions).

The work of Datta and Zou (2017) builds upon Loh and Wainwright (2012) by proposing a convex formulation of Lasso. Although the algorithm introduced does not require knowledge of $\|\beta^*\|_2$, it does also require access to Σ_H; in other words, their algorithm is also not noise-model agnostic. In fact, many works (e.g., Rosenbaum and Tsybakov (2010), Rosenbaum and Tsybakov (2013), Belloni et al. (2017b)) require either Σ_H to be known or the structure of the covariance noise is such that it admits a data-driven estimator for its covariance matrix.

It is worth noting that all of these works only focus on parameter estimation (i.e., learning β^*) and not explicitly de-noising the observed covariates. Thus, even with the knowledge of β^*, it is not clear how these methods can be used to produce predictions of the response variables associated with unseen, noisy covariates.

■ 2.3 Principal Component Regression

The effectiveness of Principal Component Regression (PCR) is well established when the covariates exhibit low-rank structure. Additionally, the regularization property of PCR is well known, at least empirically, due to its ability to reduce the variance. However, its ability to handle settings with noisy, missing, and mixed (discrete and continuous) valued covariates is not understood and remains an important open challenge, cf. Chao et al. (2019). In fact, the formal literature providing an analysis of PCR is surprisingly sparse, especially given its ubiquity in practice. A notable work is that of Bair et al. (2006), which suggests a variation of PCR to infer the direction of the principal components. However, it stops short of providing meaningful finite sample analysis beyond what is naturally implied by that of standard Linear Regression.

■ 2.4 Matrix & Tensor Estimation

Matrix estimation has spurred tremendous theoretical and empirical research across numerous fields, including recommendation systems (see Keshavan et al. (2010a,b); Negahban and Wainwright (2011); Chen and Wainwright (2015); Chatterjee (2015a); Lee et al. (2016); Candès and Tao (2010); Recht (2011); Davenport et al. (2014)), social network analysis (see Abbe and Sandon (2015a,b, 2016); Anandkumar et al. (2013); Hopkins and Steurer (2017)), graph learning (graphon estimation) (see Airoldi et al. (2013); Zhang et al. (2015); Borgs et al. (2015, 2017)), time series analysis (see Agarwal et al. (2018, 2020c)), reinforcement learning (see Shah et al. (2020)), and adversarial learning (see Yang et al. (2019a,b)). Traditionally, the end goal is to recover the underlying mean matrix from an incomplete, noisy sampling of its entries, and possibly with side information (see Farias and Li (2019)). In general, the quality of the estimate is often measured through the spectral or Frobenius norms. Further, entry-wise independence and sub-gaussian noise is typically assumed. A key property of many matrix estimation methods is they are noise-model agnostic (i.e., the de-noising procedure does not change with the noise assumptions). We advance state-of-art for hard singular value thresholding (HSVT), a specific (arguably the most ubiquitous) matrix estimation method by analyzing its error with respect to the $\ell_{2,\infty}$-norm, which is a stronger measure than its Frobenius counterpart. This generalization enables us to prove that PCR, which can be viewed through the lens of HSVT (see Proposition 5.3.1 of Chapter 5), recovers the underlying model parameter and achieves consistent out-of-sample prediction error.

Recently, there has also been exciting developments in tensor estimation. In particular, much theoretical work has focused on convex optimization approaches a la Ji Liu et al. (2009), Gandy et al. (2011). Additionally, there has also been advancements with new algorithms and provable guarantees (see Jain and Oh (2014), Huang et al. (2015), Zhang and Barzilay (2015), Barak and Moitra (2016)). For a thorough review of tensors, we refer the interested reader to Kolda and Bader (2009). We take this opportunity to note that while we take a tensor perspective on counterfactual estimation, we cannot directly apply these standard methods. This is due to a stark difference in modeling assumptions (block sparsities of our setting versus uniform sparsity in standard setups) and objectives (i.e., we require guarantees for every unit-intervention tuple, while standard methods provide guarantees on average across entire tensor). For a more detailed discussion, please see Chapter 8.4.1.

Chapter 3

Preliminaries

A common thread of modern datasets is that observations are often contaminated, both by measurement noise and missing data, and may even be high-dimensional. In other words, we only have access to a sparse, noisy representation of certain phenomena of interest, which may live in a high-dimensional ambient space. To understand how these perturbations affect our ability to recover the underlying signal, we rely on techniques ranging from random matrix theory to non-asymptotic probability theory. As a result, we take this opportunity to review relevant concepts from linear algebra and probability theory in this chapter, which are central in analyzing our problems of interest.

In particular, we will study the geometry of linear operators through the prism of singular values and subspaces, and how these objects change under perturbations. Additionally, we present concentration inequalities, which quantify how random empirical quantities deviate around their deterministic population counterparts, and the speed to which they converge to these quantities. We will anchor on these results to then derive non-asymptotic rates at which the probabilities of "bad" events vanish to zero.

■ 3.1 Linear Algebra

Consider a real-valued $m \times n$ matrix $\boldsymbol{A}$. Recall that $\boldsymbol{A}$ can always be represented via its *singular value decomposition* (SVD), which we write as

$$\boldsymbol{A} = \sum_{i=1}^{r} s_i u_i v_i^T, \tag{3.1}$$

where $r = \text{rank}(\boldsymbol{A})$. Here, s_i denote the *singular values* of $\boldsymbol{A}$ (typically arranged in non-increasing order with $s_i = 0$ for all $i > r$), while the vectors $u_i \in \mathbb{R}^m$ and $v_i \in \mathbb{R}^n$ denote the corresponding *left* and *right singular vectors*, respectively, of $\boldsymbol{A}$. Equivalently, in matrix notation, we can write

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T,$$

where $\boldsymbol{U} = [u_1, \ldots, u_r] \in \mathbb{R}^{m\times r}$, $\boldsymbol{V} = [v_1, \ldots, v_r] \in \mathbb{R}^{n\times r}$, and $\boldsymbol{S} = \operatorname{diag}(s_1, \ldots, s_r) \in \mathbb{R}^{r\times r}$.

Importantly, the left singular vectors u_i are also the orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$; similarly, the right singular vectors v_i are the orthonormal eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$. The singular values s_i are thus the square roots of the eigenvalues λ_i of both $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$.

The *Moore-Penrose pseudoinverse* of $\boldsymbol{A}$, denoted as $\boldsymbol{A}^\dagger$, inverts $\boldsymbol{A}$ where $\boldsymbol{A}$ is invertible, i.e., between the row space and column space of $\boldsymbol{A}$. We write the pseudoinverse as

$$\boldsymbol{A}^\dagger = \boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^T = \sum_{i=1}^{r} (1/s_i)\, v_i u_i^T .$$

In general, we can express $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U}_\perp \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{S} & 0 \\ 0 & \boldsymbol{S}_\perp \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{V}^T \\ \boldsymbol{V}_\perp^T \end{bmatrix}, \tag{3.2}$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times r}$, $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{S} \in \mathbb{R}^{r\times r}$ are defined as before, and $\boldsymbol{U}_\perp \in \mathbb{R}^{m\times(m-r)}$, $\boldsymbol{V}_\perp \in \mathbb{R}^{n\times(n-r)}$, and $\boldsymbol{S}_\perp \in \mathbb{R}^{(m-r)\times(n-r)}$. Observe that if $\operatorname{rank}(\boldsymbol{A}) = r$, then $\boldsymbol{S}_\perp = 0$. Further, we note that the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ span the column and row spaces, respectively, of $\boldsymbol{A}$. As a result, $\boldsymbol{U}_\perp$ denotes the left null space of $\boldsymbol{A}$ and $\boldsymbol{V}_\perp$ denotes the null space of $\boldsymbol{A}$. Since the columns of $\boldsymbol{U}, \boldsymbol{U}_\perp, \boldsymbol{V}$, and $\boldsymbol{V}_\perp$ are orthonormal, we will often refer to them as matrices and subspaces interchangeably, where the subspaces are spanned by their columns. Hence, we will denote $\mathcal{P}_U = \boldsymbol{U}\boldsymbol{U}^T$ and $\mathcal{P}_V = \boldsymbol{V}\boldsymbol{V}^T$ as the orthogonal projection operators onto the column and row spaces of $\boldsymbol{A}$, respectively; similarly, we define $\mathcal{P}_{U_\perp} = \boldsymbol{U}_\perp\boldsymbol{U}_\perp^T$ and $\mathcal{P}_{V_\perp} = \boldsymbol{V}_\perp\boldsymbol{V}_\perp^T$ as the projections onto the left null space and null space, respectively.

Moving forward, for any matrix $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$ with orthonormal columns, we denote $\mathcal{P}_Q = \boldsymbol{Q}\boldsymbol{Q}^T \in \mathbb{R}^{m\times m}$ as the orthogonal projection operator onto the n-dimensional subspace of $\mathbb{R}^m$ spanned by the columns of $\boldsymbol{Q}$.

■ 3.1.1 Matrix Norms

There are several ways to measure the *size* of a matrix. We will mention three of them – operator (or spectral), Frobenius, and $\ell_{2,\infty}$-norms.

The matrix $\boldsymbol{A}$ is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Its *operator* (or *spectral*) norm is defined

$$\|\boldsymbol{A}\| = \max_{x \in S^{n-1}} \|\boldsymbol{A}x\|_2 = \max_{x \in S^{n-1}, y \in S^{m-1}} \langle \boldsymbol{A}x, y \rangle,$$

where S^{n-1} and S^{m-1} are the unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Equivalently, the spectral view of the operator norm states that

$$\|\boldsymbol{A}\| = s_1,$$

i.e., the operator norm of $\boldsymbol{A}$ is the largest singular value of $\boldsymbol{A}$.

From the perspective of the entries of $\boldsymbol{A}$, the *Frobenius* norm of a matrix is the extension of the standard Euclidean ℓ_2-norm on vectors:

$$\|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A}) = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2,$$

where tr denotes the trace operator. In terms of its singular values, the Frobenius norm can also be represented as

$$\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{r} s_i^2.$$

Thus, if $s \in \mathbb{R}^r$ denotes the vector of singular values, then

$$\|\boldsymbol{A}\| = ||s||_\infty \quad \text{and} \quad \|\boldsymbol{A}\|_F = ||s||_2,$$

which yields the inequality $\|\boldsymbol{A}\|_F \leq \sqrt{r} \cdot \|\boldsymbol{A}\|$, where $\mathrm{rank}(\boldsymbol{A}) = r$.

Finally, we introduce the $\ell_{2,\infty}$-norm, which is a mixed-norm on $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_{2,\infty}^2 = \max_{j \in [n]} \|A_j\|_2 = \max_{j \in [n]} \sum_{i=1}^{m} A_{ij}^2,$$

where $A_j \in \mathbb{R}^m$ denotes the j-th column of $\boldsymbol{A}$. Thus, the $\ell_{2,\infty}$-norm measures the maximum

ℓ_2-norm on the columns of $\boldsymbol{A}$. As an important side note, observe that

$$\frac{1}{mn}\left\|\boldsymbol{A}\right\|_F^2 = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2 \leq \frac{1}{m}\max_{j\in[n]}\sum_{i=1}^{m} A_{ij}^2 = \frac{1}{m}\left\|\boldsymbol{A}\right\|_{2,\infty}^2.$$

This formalizes how the $\ell_{2,\infty}$-metric is a stronger guarantee than the Frobenius norm.

■ 3.1.2 Isometries and Projections

The singular values of $\boldsymbol{A}$ have an important geometric interpretation. Specifically, the singular values of $\boldsymbol{A}$ satisfy

$$s_n \cdot \|x\|_2 \leq \left\|\boldsymbol{A}x\right\|_2 \leq s_1 \cdot \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

Because $\boldsymbol{A}$ acts as an operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, the singular values of $\boldsymbol{A}$ characterize the amount of distortion in the size of x after its transformation. In particular, the operator norm of $\boldsymbol{A}$ measures the maximum distortion of the geometry of $\mathbb{R}^n$ under the action of $\boldsymbol{A}$.

Operators that preserve distances exactly, called *isometries*, are of particular interest here. We say that a matrix $\boldsymbol{A}$ (with $m \geq n$) is an isometry if

$$\left\|\boldsymbol{A}x\right\|_2 = \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

Clearly, this implies that the singular values of $\boldsymbol{A}$ are all equal to 1, i.e., $s_1 = s_n = 1$. Additionally, it follows that

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I},$$

and $\mathcal{P}_A = \boldsymbol{A}\boldsymbol{A}^T$ is an orthogonal projection onto an n-dimensional subspace of $\mathbb{R}^m$. A useful consequence is that any subset of the columns of an orthogonal (unitary) matrix is immediately an isometry.

Projection matrices, and particularly orthogonal projection operators, will play an important role in this exposition. Thus, we review some their distinguishing properties. To begin, recall that any projection operator (not necessarily orthogonal), $\mathcal{P} : \mathbb{R}^n \to \mathbb{R}^n$, satisfies $\mathcal{P}^2 = \mathcal{P}$. Intuitively, this means if we start with any vector $x \in \mathbb{R}^n$, then $\mathcal{P}x$ lies in the subspace $\mathcal{P}$ projects onto, and applying the projection again does nothing to the resulting vector. An orthogonal projection matrix further satisfies $\mathcal{P} = \mathcal{P}^T$. An important spectral property of $\mathcal{P}$ is that its eigenvalues are either 1 or 0. As a result, it follows that

$$\left\|\mathcal{P}x\right\|_2 \leq \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

Thus, $\mathcal{P}$ is a bounded operator.

If we let $\boldsymbol{A}$ be defined as in (3.2) with rank r, then we can decompose any $x \in \mathbb{R}^n$ as

$$x = \mathcal{P}_V x + \mathcal{P}_{V_\perp} x,$$

where $\mathcal{P}_V$ and $\mathcal{P}_{V_\perp}$ refer to the orthogonal projections onto the row space and null space of $\boldsymbol{A}$, respectively. A useful representation for the $\mathcal{P}_{V_\perp}$ is then $\mathcal{P}_{V_\perp} = \boldsymbol{I} - \mathcal{P}_V$; similarly, $\mathcal{P}_{U_\perp} = \boldsymbol{I} - \mathcal{P}_U$. We take this opportunity to remind the reader that all the action under $\boldsymbol{A}$ occurs between $\boldsymbol{V}$ and $\boldsymbol{U}$. To see this, observe that applying $\boldsymbol{A}$ to x yields

$$\boldsymbol{A}x = \boldsymbol{A} \cdot (\mathcal{P}_V x + \mathcal{P}_{V_\perp} x) = \boldsymbol{A} \cdot \mathcal{P}_V x.$$

Since $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, it follows that $\boldsymbol{A} \cdot \mathcal{P}_V x = \boldsymbol{A}x$ while $\boldsymbol{A} \cdot \mathcal{P}_{V_\perp} x = 0$. In words, only the component of x that lives within the row space of $\boldsymbol{A}$ gets mapped to the column space while any component of x within the null space is mapped to 0.

■ 3.1.3 Perturbation Theory

Often, our observed data is a perturbed version of our true underlying signal. To recover the latent signal, it is important to understand the effects of the perturbation. Perturbation theory describes how the spectrum of our signal changes under "small" matrix perturbations, and they play a critical role in analyzing spectral methods (e.g., SVDs).

Let $\boldsymbol{A}$, as defined in (3.2), describe our signal matrix, which is approximately rank r (i.e., $s_r \gg s_{r+1}$). We denote $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ as the perturbation matrix (i.e., noise). We partition the SVD of our observation matrix

$$\boldsymbol{Z} = \boldsymbol{A} + \boldsymbol{H}$$

as follows:

$$\boldsymbol{Z} = \begin{bmatrix} \widehat{\boldsymbol{U}} & \widehat{\boldsymbol{U}}_\perp \end{bmatrix} \cdot \begin{bmatrix} \widehat{\boldsymbol{S}} & 0 \\ 0 & \widehat{\boldsymbol{S}}_\perp \end{bmatrix} \cdot \begin{bmatrix} \widehat{\boldsymbol{V}}^T \\ \widehat{\boldsymbol{V}}_\perp^T \end{bmatrix}, \tag{3.3}$$

where $\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{U}}_\perp, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}_\perp, \widehat{\boldsymbol{V}}$, and $\widehat{\boldsymbol{V}}_\perp$ have the same structures as $\boldsymbol{U}, \boldsymbol{U}_\perp, \boldsymbol{S}, \boldsymbol{S}_\perp, \boldsymbol{V}$, and $\boldsymbol{V}_\perp$, respectively.

Perturbation of Singular Subspaces

There are several characterizations to measure the distance between two subspaces, say V and $\widehat{V}$. One viewpoint is through their orthogonal projection operators:

$$d_1(\boldsymbol{V}, \widehat{\boldsymbol{V}}) = \left\| \mathcal{P}_V - \mathcal{P}_{\widehat{V}} \right\|.$$

Another perspective is through the prism of principal angles:

$$d_2(\boldsymbol{V}, \widehat{\boldsymbol{V}}) = \left\| \sin \Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}}) \right\|.$$

The following proposition states that the two definitions, which are both unaffected by global orthonormal transformations, are equivalent, i.e., d_1 and d_2 are equivalent metrics.

Proposition 3.1.1. *Suppose $[\boldsymbol{V}, \boldsymbol{V}_\perp]$ and $[\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{V}}_\perp]$ are orthogonal matrices, where $\boldsymbol{V}_\perp$ and $\widehat{\boldsymbol{V}}_\perp$ are orthogonal complements to $\boldsymbol{V}$ and $\widehat{\boldsymbol{V}}$, respectively. Then,*

$$d(\boldsymbol{V}, \widehat{\boldsymbol{V}}) = \left\| \mathcal{P}_V - \mathcal{P}_{\widehat{V}} \right\| = \left\| \sin \Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}}) \right\|.$$

There are two primary, canonical results from the field of perturbation theory that have spurred tremendous research in recent years: the Davis-Kahan $\sin \Theta$ Theorem (Davis and Kahan (1970)) for eigenspaces and Wedin's modified version for singular subspaces (Wedin (1972)). Of the many exciting works to come out of this field, we highlight that Yu et al. (2015) extends the analyses to provide a useful variant of the results in terms of the population parameters (i.e., the singular values associated with $\boldsymbol{A}$).

Theorem 3.1.1 (Wedin's generalized $\sin \Theta$ theorem; Corollary 1.4.10 in Stratos (2016)). *Let $\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{A} + \boldsymbol{H}$ be defined as in* (3.2) *and* (3.3)*, respectively. Then,*

$$\left\| \sin \Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}}) \right\| \vee \left\| \sin \Theta(\boldsymbol{U}, \widehat{\boldsymbol{U}}) \right\| \leq \frac{2\|\boldsymbol{H}\|}{s_r - s_{r+1}},$$

where s_i denotes the i-th singular value of $\boldsymbol{A}$.

Despite its wide applicability, Wedin's perturbation bound (Wedin (1972)) may be suboptimal in certain settings. Specifically, since Wedin's bound is uniform for both the left and right singular spaces, it may not be useful to apply Wedin's bound when the row and column dimensions of the matrix differ significantly. To that end, Cai and Zhang (2018) resolves this gap by establishing separate optimal rates for the left and right singular subspaces under the same perturbation.

Before we present their results, we make the convenient decomposition of $\boldsymbol{H}$:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U}_\perp \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{V}^T \\ \boldsymbol{V}_\perp^T \end{bmatrix}, \tag{3.4}$$

where

$$\boldsymbol{H}_{11} = \boldsymbol{U}^T \boldsymbol{H} \boldsymbol{V}, \quad \boldsymbol{H}_{12} = \boldsymbol{U}^T \boldsymbol{H} \boldsymbol{V}_\perp, \quad \boldsymbol{H}_{21} = \boldsymbol{U}_\perp^T \boldsymbol{H} \boldsymbol{V}, \quad \text{and} \quad \boldsymbol{H}_{22} = \boldsymbol{U}_\perp^T \boldsymbol{H} \boldsymbol{V}_\perp.$$

Theorem 3.1.2 (Perturbation bounds for singular subspaces Cai and Zhang (2018)). *Let $\boldsymbol{A}, \boldsymbol{Z}$, and $\boldsymbol{H}$ be given as* (3.2), (3.3), *and* (3.4), *respectively. Further, let $h_{ij} := \left\| \boldsymbol{H}_{ij} \right\|$ for $i, j = 1, 2$. Denote*

$$\alpha := \sigma_{\min}(\boldsymbol{U}^T \boldsymbol{Z} \boldsymbol{V}) \quad \textit{and} \quad \beta := \boldsymbol{U}_\perp^T \boldsymbol{Z} \boldsymbol{V}_\perp. \tag{3.5}$$

If $\alpha^2 > \beta^2 + h_{12}^2 \wedge h_{21}^2$, then

$$\left\| \sin \Theta(\boldsymbol{V}, \widehat{\boldsymbol{V}}) \right\| \le \frac{\alpha h_{12} + \beta h_{21}}{\alpha^2 - \beta^2 - (h_{21}^2 \wedge h_{12}^2)} \wedge 1$$
$$\left\| \sin \Theta(\boldsymbol{U}, \widehat{\boldsymbol{U}}) \right\| \le \frac{\alpha h_{21} + \beta h_{12}}{\alpha^2 - \beta^2 - (h_{21}^2 \wedge h_{12}^2)} \wedge 1.$$

In Chapter 5, we will invoke Theorems 3.1.1 and 3.1.2 to show that Principal Component Analysis (PCA) and hard singular value thresholding (HSVT) can accurately recover the singular subspaces of the signal matrix.

Perturbation of Singular Values

Analogous to the bounds for singular subspaces are perturbation bounds for singular values. These results study how the singular values of the signal matrix $\boldsymbol{A}$ change under perturbations. The most well known results are attributed to Weyl, which we state below.

Lemma 3.1.1 (Perturbation of singular values (Weyl's inequality)). *Let s_i and τ_i denote the singular values of $\boldsymbol{A}$ and $\boldsymbol{Z}$, respectively, in decreasing order and repeated by multiplicities. Suppose $\boldsymbol{Z} = \boldsymbol{A} + \boldsymbol{H}$. Then,*

$$\max_{i \in [m \wedge n]} |s_i - \tau_i| \le \left\| \boldsymbol{H} \right\|.$$

■ 3.1.4 Tensors

Here, we briefly overview a tensors, a convenient representation for multi-dimensional data (a la spatio-temporal models; see Bahadori et al. (2014), Agarwal et al. (2020c)). In particular, the structure of tensors naturally lends itself to capturing inter-dependencies along the multiple dimensions.

For convenience, we focus on order-three tensors (though this can be easily extended to higher dimensions). More formally, we denote $\boldsymbol{T}$ as a $m \times n \times p$ order-three tensor. Since $\boldsymbol{T}$ can be viewed as a stack of p matrices of size $m \times n$, it will be convenient for us to represent multi-dimensional data in the form of a tensor. As such, we will provide a very review of an important tensor structure – its rank. Unlike matrices, however, tensors have several notions of rank. For our purposes, we will discuss the canonical polyadic (CP) rank, which is the tensor analogue to the traditional notion of matrix rank.

CP Rank

The canonical polyadic (CP) rank of a tensor $\boldsymbol{T}$ is related to its orthogonal decompositions, and can be regarded as the natural generalization of the matrix SVD (see (3.1)). We say that $\boldsymbol{T}$ is a rank-one tensor if it is expressed as the outer product of three vectors, say u, v, and w for $u \in \mathbb{R}^m, v \in \mathbb{R}^n$, and $w \in \mathbb{R}^p$, i.e.,

$$\boldsymbol{T} = u \otimes v \otimes w,$$

where the (i, j, k)-th entry of $\boldsymbol{T}$ can be written as $T_{ijk} = u_i v_j w_k$. More generally, we say that $\boldsymbol{T}$ has CP rank r if $\boldsymbol{T}$ can be expressed as

$$\boldsymbol{T} = \sum_{\ell=1}^{r} u_\ell \otimes v_\ell \otimes w_\ell,$$

i.e., r is the minimum number such that $\boldsymbol{T}$ can be expressed as a sum of r rank-one tensors.

Remark 3.1.1. *For a more thorough treatment of tensors, we refer the interested reader to Kolda and Bader (2009).*

■ 3.2 Concentration Inequalities

Let X be a random variable. Recall that the quantities $\mathbb{E}[X^k]$ and $\mathbb{E}[(X - \mathbb{E}X)^k]$ for $k \in \mathbb{N}$ represent the k-th moment and k-th central moment of X, respectively (assuming

$\mathbb{E}[|X|^k] < \infty$). The first moment and second central moment are well known as the *mean* and *variance* of X. Further, we denote the *moment generating function* (MGF) of X as

$$M_X(s) = \mathbb{E}[e^{sX}], \tag{3.6}$$

which is well defined for all $s \in \mathbb{R}$ for which (3.6) is finite. We will see that moments of a random variable capture useful information about its tail, i.e., $\mathbb{P}(X \geq x)$ and $\mathbb{P}(X \leq x)$, which denote the upper (right) and lower (left) tails, respectively. In essence, MGFs characterize the rate at which random variables converge to their population quantities (e.g., mean).

For instance, one notable tool, known as Markov's inequality, bounds the tail of a non-negative random variable in terms of its expectation.

Lemma 3.2.1 (Markov's inequality)**.** *For any non-negative random variable X and scalar $t > 0$, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. We begin by fixing any $t > 0$. Observing that $X \geq t \cdot \mathbb{1}(X \geq t)$ and taking expectations completes the proof, i.e., $\mathbb{E}[X] \geq \mathbb{E}[t \cdot \mathbb{1}(X \geq t)] = t \cdot \mathbb{P}(X \geq t)$. ■

To entertain the interested reader, we provide the proofs for Lemmas 3.2.2, 3.2.4, and 3.2.5, which utilize common proof techniques (e.g., Chernoff's exponentiation trick, epsilon nets) to establish concentration. Additionally, we will use these lemmas to prove our results in Chapter 5. Finally, we refer the reader to Vershynin (2018) for a detailed review of high-dimensional probability.

■ 3.2.1 Sub-gaussian Distributions

We introduce an important class of probability distributions known as *sub-gaussian distributions*. As the name suggests, these distributions are an extension of the famous Gaussian (Normal) distribution, and thus exhibit similar desirable properties; namely, tails that decay at least as fast as that of a Gaussian. Belonging to this rich class include distributions such as the Bernoulli, truncated Poisson, and all bounded distributions. As such, many results in probability, data science, and machine learning are proved under a sub-gaussian setting.

Sub-gaussian random variables satisfy many properties – we highlight a few that will be of particular importance. For a sub-gaussian random variable X, we denote its *sub-gaussian* *norm* as

$$\left\|X\right\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(X^2/t^2\right)\right] \leq 2\right\}.$$

Property 3.2.1. *If* $\mathbb{E}X = 0$*, then*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(c\lambda^2\left\|X\right\|_{\psi_2}^2\right) \quad \textit{for all } \lambda \in \mathbb{R},$$

where c *is an absolute constant.*

Property 3.2.2. *Let* $X_1, \ldots, X_n$ *be a sequence of independent, mean zero sub-gaussian random variables. Then* $\sum_{i=1}^n X_i$ *is also a sub-gaussian random variable, and*

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^n \left\|X_i\right\|_{\psi_2}^2,$$

where C *is an absolute constant.*

We now state a modified version of Hoeffding's inequality in Lemma 3.2.2. Effectively, Hoeffding's inequality, which is a more useful restatement of Property 3.2.2, establishes the concentration of sums of independent sub-gaussian random variables.

Lemma 3.2.2 (Modified General Hoeffding's inequality)**.** *Let* $X = (X_1, \ldots, X_n)$ *be a random vector whose entries are independent, mean zero sub-gaussian random variables. Let* $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ *be a random vector, independent of* X*, satisfying* $\|a\|_2 \leq b$ *for some constant* $b \geq 0$*. Then, for every* $t \geq 0$*, we have*

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

where $K = \max_i \left\|X_i\right\|_{\psi_2}$*.*

Proof. Let $S_n = \sum_{i=1}^n a_i X_i$. Then applying Markov's inequality (Lemma 3.2.1) for any $\lambda > 0$, we have

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

We note that the above is an example of Chernoff's exponentiation trick, i.e., applying Markov's inequality to $\exp(\lambda S_n)$ as opposed to simply S_n.

Now, conditioned on the random vector a, observe that

$$\mathbb{E}\left[\exp(\lambda S_n)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\left(CK^2\lambda^2 \|a\|_2^2\right) \leq \exp\left(CK^2\lambda^2 b^2\right),$$

where the equality follows from conditional independence, the first inequality by Property 3.2.1, and the final inequality by assumption. Therefore,

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

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. ■

Sub-gaussian Random Vectors

The concept of sub-gaussian distributions extends to higher dimensions. In particular, we say that a random vector $X \in \mathbb{R}^n$ is sub-gaussian if all one-dimensional marginals, i.e., $\langle X, u\rangle$ of X for $u \in \mathbb{R}^n$, are sub-gaussian random variables. The corresponding sub-gaussian norm is then defined as

$$\|X\|_{\psi_2} = \sup_{u \in \mathbb{S}^{n-1}} \|\langle X, u\rangle\|_{\psi_2}.$$

■ 3.2.2 Sub-exponential Distributions

While sub-gaussian distributions form a wide class of distributions, there are several natural random distributions (e.g., Laplacian, χ^2), which are not sub-gaussian but rather *sub-exponential*. Although these distributions have heavier tails, we will see that they have a close connection with our friendly sub-gaussian distributions.

Lemma 3.2.3 (Sub-exponential is sub-gaussian squared Vershynin (2018))**.** *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

The following statement can be seen as a version of Hoeffding's inequality (Lemma 3.2.2) for sub-exponential distributions.

Theorem 3.2.1 (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*

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

where $c > 0$ is an absolute constant.

Quadratic Forms

Thus far, our focus has been studying sums of independent random variables. However, quadratic forms of the type $\langle X, \boldsymbol{A}X\rangle$ for a random variable $X \in \mathbb{R}^n$ and matrix of coefficients $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, find their way in several important applications. Unfortunately, these terms, often called *chaos* in probability theory, are harder to establish concentration due to the dependence of the terms in the sum.

To that end, we state Lemma 3.2.4, which is a modified version of the Hanson-Wright inequality where $\boldsymbol{A}$ is also a random object.

Lemma 3.2.4 (Modified Hanson-Wright inequality). *Let $X \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with $\left\|X_i\right\|_{\psi_2} \leq K$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a random matrix, independent of X, satisfying $\left\|\boldsymbol{A}\right\| \leq a$ and $\left\|\boldsymbol{A}\right\|_F^2 \leq b$ 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:

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

Therefore, letting

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

we can express

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

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

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

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

$$\left\|X_i^2 - \mathbb{E}[X_i^2]\right\|_{\psi_1} \leq C_1\left\|X_i^2\right\|_{\psi_1} \leq C_2\|X_i\|_{\psi_2}^2 \leq C_2K^2.$$

Conditioned on $\boldsymbol{A}$, we have that

$$\begin{aligned} \mathbb{E}\left[\exp(\lambda S_n)\right] &= \mathbb{E}\left[\exp\left(\sum_{i=1}^{n}\lambda A_{ii}(X_i^2 - \mathbb{E}[X_i^2])\right)\right] \\ &= \prod_{i=1}^{n}\mathbb{E}\left[\exp\Big(\lambda A_{ii}(X_i^2 - \mathbb{E}[X_i^2])\Big)\right] \\ &\leq \prod_{i=1}^{n}\exp\Big(CK^4\lambda^2A_{ii}^2\Big) \\ &\leq \exp\Big(CK^4\lambda^2\|\boldsymbol{A}\|_F^2\Big) \\ &\leq \exp\Big(CK^4\lambda^2 b\Big), \end{aligned}$$

where $|\lambda| \leq c/(aK^2)$. Therefore, optimizing over λ yields

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

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

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

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. ■

■ 3.2.3 Random Matrices

Here, we consider $m \times n$ matrices $\boldsymbol{A}$ with random entries. For a detailed overview, please refer to Vershynin (2010).

Independent Entries

We begin with the simplest, classical model of random matrices where its entries are independent standard Gaussian random variables. Below, we state the behavior of the extreme singular values of such a random matrix.

Theorem 3.2.2 (Gaussian matrices; Corollary 5.35 of Vershynin (2010))**.** *Let $\boldsymbol{A}$ be a $m \times n$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, the following holds with probability at least $1 - 2\exp\left(-t^2/2\right)$:*

$$\|\boldsymbol{A}\| \leq \sqrt{m} + \sqrt{n} + t.$$

Independent Rows

We now consider settings where the rows of our matrix are independent random vectors in $\mathbb{R}^n$. Such settings are important in data science and machine learning since the rows of $\boldsymbol{A}$ denote samples from a potentially high-dimensional distribution. As such, there is no reason to suspect that the columns of $\boldsymbol{A}$, corresponding to features, are not correlated.

As aforementioned, it is important for us to understand the effects of random perturbations to our underlying signal. Luckily for us, as the dimensions of the matrices grow, the spectrum of $\boldsymbol{A}$ tends to "stabilize". This formalized in Lemma 3.2.5, which describes the distribution of the singular values of $\boldsymbol{A}$ under the setting described above.

Lemma 3.2.5. *Let A be an $m \times n$ matrix whose rows A_i are independent, mean zero, sub-gaussian random vectors in $\mathbb{R}^n$ with second moment matrix $\boldsymbol{\Sigma} = (1/m) \cdot \mathbb{E}[\boldsymbol{A}^T\boldsymbol{A}]$. Then for any $t \geq 0$, the following inequality holds with probability at least $1 - \exp\left(-t^2\right)$:*

$$\left\| \frac{1}{m} A^T A - \boldsymbol{\Sigma} \right\| \leq K^2 \max(\delta, \delta^2), \quad \textit{where } \delta = C\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}};$$

here, $K = \max_i \left\|A_i\right\|_{\psi_2}$ and $C > 0$ is an absolute constant.

Remark 3.2.1. *Observe that Lemma 3.2.5 implies that for any $t \geq 0$,*

$$\sqrt{m} \cdot \left\|\boldsymbol{\Sigma}\right\|^{1/2} - CK^2(\sqrt{n} + t) \leq s_{\min}(A) \leq s_1(A) \leq \sqrt{m} \cdot \left\|\boldsymbol{\Sigma}\right\|^{1/2} + CK^2(\sqrt{n} + t)$$

with probability at least $1 - 2\exp\left(-t^2\right)$.

Proof. The following proof extends the proof of Theorem 4.6.1 of Vershynin (2018) for the non-isotropic setting; we present it here for completeness. Recall that the operator norm of $\boldsymbol{A}$ can computed by maximizing the following quadratic form:

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

where S^{n-1}, S^{m-1} denote the unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Rather than searching through the entire unit spheres, we will discretize the spheres using an ϵ-net argument to establish a tight control of the quadratic term $\langle \boldsymbol{A}x, y \rangle$ for any pair of fixed unit vectors x, y. Then, we will take a union bound over all x, y in the net.

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^{n-1} with cardinality $|\mathcal{N}| \leq 9^n$. Applying Lemma 4.4.1 of

Vershynin (2018), we obtain

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

To achieve our desired result, it remains to show that

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

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

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

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

Since the rows of $\boldsymbol{A}$ 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

$$\left\| Y_i^2 - \mathbb{E}[Y_i^2] \right\|_{\psi_1} \leq C_1 \left\| Y_i^2 \right\|_{\psi_1} \leq C_2 \|Y_i\|_{\psi_2}^2 \leq C_2 K^2.$$

As a result, we can apply Bernstein's inequality to obtain

$$\begin{aligned} \mathbb{P}\left(\left| \frac{1}{m} \|\boldsymbol{A}x\|_2^2 - x^T \boldsymbol{\Sigma} x \right| \geq \frac{\epsilon}{2} \right) &= \mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} (Y_i^2 - \mathbb{E}[Y_i^2]) \right| \geq \frac{\epsilon}{2} \right) \\ &\leq 2 \exp\left[-c_1 \min\left(\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2} \right) m \right] \\ &= 2 \exp\left[-c_1 \delta^2 m \right] \\ &\leq 2 \exp\left[-c_1 C^2 (n + t^2) \right], \end{aligned}$$

where the last inequality follows from the definition of δ 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,

$$\mathbb{P}\left(\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|\boldsymbol{A}x\|_2^2 - x^T\boldsymbol{\Sigma}x\right| \geq \frac{\epsilon}{2}\right) \leq 9^n \cdot 2\exp\Big[-c_1 C^2(n+t^2)\Big] \leq 2\exp\Big(-t^2\Big),$$

for large enough C. This concludes the proof. ■

Chapter 4

Potential Outcomes Tensor

In this chapter, we formally present our potential outcomes tensor. This framework will enable us to characterize both ESs and OSs by unique block sparsity patterns, and relate counterfactual estimation to tensor estimation. The following chapters (namely, Chapters 6, 7, and 8) will analyze instances of this general setting.

■ 4.1 Tensor Factor Model

Throughout, we adopt the potential outcomes framework of Neyman (1923) and Rubin (1974a). As an important contribution, we represent the universe of potential outcomes through a tensor object. This structural representation will then allow us to view the estimation of counterfactuals as equivalent to recovering aspects of this tensor, and will allow us to not only establish a relationship between the units, interventions, and metrics, but also prove the existence of synthetic controls and interventions.

■ 4.1.1 Potential Outcomes

Let $\boldsymbol{M} \in \mathbb{R}^{T \times N \times D \times P}$ be an order-four tensor where its (t, n, d, p)-th element, $M_{tn}^{(d,p)}$, represents the potential outcome of the t-th measurement for unit n under intervention d and metric p. It is convenient to think of $\boldsymbol{M}$ as a collection of P order-three tensors, where each tensor represents the collection of potential outcomes across all units, time, and interventions for a particular metric. In particular, for any metric p, let $\boldsymbol{M}^{(d,p)} \in \mathbb{R}^{T \times N}$ denote the d-th frontal slice of the p-th tensor, which represents the matrix of potential outcomes across all measurements and units under intervention d. Further, we denote

$$\begin{aligned} \boldsymbol{M}_{\text{pre}}^{(d,p)} &= [M_{tn}^{(d,p)} : t \leq T_0, n \in \mathcal{I}^{(d)}] \in \mathbb{R}^{T_0 \times N^{(d)}} \\ \boldsymbol{M}_{\text{post}}^{(d,p)} &= [M_{tn}^{(d,p)} : t > T_0, n \in \mathcal{I}^{(d)}] \in \mathbb{R}^{(T-T_0) \times N^{(d)}} \end{aligned} \tag{4.1}$$

as the pre- and post-intervention sub-matrices of $\boldsymbol{M}^{(d,p)}$ restricted to those donors in $\mathcal{I}^{(d)}$ (defined as in (1.1)). For ease of notation, we note that if unit i is the target and receives intervention d, then $\mathcal{I}^{(d)} := \mathcal{I}^{(d)} \setminus \{i\}$ and $N^{(d)} := \left|\mathcal{I}^{(d)} \setminus \{i\}\right|$.

In consistency with the standard econometrics literature, we consider a factor model. Specifically, we extend the standard matrix factor model to a tensor factor model. The low-rank assumption reflects the belief that correlations exist within the different dimensions (i.e., time, units, interventions, and metrics), and thus the potential outcomes can be described by a few latent factors. This is formalized by the following property.

Property 4.1.1 (Low-rank). *The canonical polyadic tensor rank of $\boldsymbol{M}$ is r. That is, there exists vectors $\{u_i\} \in \mathbb{R}^T, \{v_i\} \in \mathbb{R}^N, \{w_i\} \in \mathbb{R}^D$, and $\{q_i\} \in \mathbb{R}^P$ for all $i \in [r]$, such that*

$$\boldsymbol{M} = \sum_{\ell=1}^{r} u_\ell \otimes v_\ell \otimes w_\ell \otimes q_\ell.$$

Interpretation. Property 4.1.1 implies that every frontal slice $\boldsymbol{M}^{(d,p)}$ can be written as

$$\boldsymbol{M}^{(d,p)} = \sum_{\ell=1}^{r} (w_{d\ell} q_{k\ell} \cdot u_\ell) \otimes v_\ell = \boldsymbol{U}^{(d,p)} \boldsymbol{V}^T, \tag{4.2}$$

where $\boldsymbol{U}^{(d,p)} \in \mathbb{R}^{T \times r}$, and $\boldsymbol{V} \in \mathbb{R}^{N \times r}$ has (without loss of generality) orthonormal columns. Hence, the low rank tensor model implies there exists an r-dimensional linear subspace of $\mathbb{R}^N$, denoted by $\boldsymbol{V}$, that describes a latent relationship between units that is invariant across interventions. Under every metric p, each intervention can then be interpreted as some linear transformation, denoted by $\boldsymbol{U}^{(d,p)}$, applied to this subspace.

Property 4.1.2 (Bounded). *The entries of $\boldsymbol{M}$ are bounded by one in absolute value.*

Property 4.1.3 (Well-balanced spectra). *For every intervention d and metric p, the non-zero singular values s_i of $\boldsymbol{M}_{\text{pre}}^{(d,p)}$ are well-balanced, i.e., $s_i^2 = \Theta(N^{(d)} T_0 / r_{\text{pre}}^{(d,p)})$, where* $\text{rank}(\boldsymbol{M}_{\text{pre}}^{(d,p)}) = r_{\text{pre}}^{(d,p)}$. *Similarly, the non-zero singular values τ_i of $\boldsymbol{M}_{\text{post}}^{(d,p)}$ satisfy $\tau_i^2 = \Theta(N^{(d)}(T - T_0)/r_{\text{post}}^{(d,p)})$, where* $\text{rank}(\boldsymbol{M}_{\text{post}}^{(d,p)}) = r_{\text{post}}^{(d,p)}$.

Example 4.1.1 (Embedded Gaussian Features). One classical example in which Property 4.1.3 holds is the probabilistic model for PCA, cf. Bishop (1999); Tipping and Bishop (1999).

Example 4.1.2 (Well-balanced Entries). Another natural setting in which Property 4.1.3 holds is if $M_{tn}^{(d,p)} = \Theta(1)$ and the non-zero singular values of $\boldsymbol{M}_{\text{pre}}^{(d,p)}$ satisfy $s_i^2 = \Theta(\zeta)$ for some ζ. Then, $C\zeta r_{\text{pre}}^{(d,p)} = \left\| \boldsymbol{M}_{\text{pre}}^{(d,p)} \right\|_F^2 = \Theta(N^{(d)} T_0)$ for some $C > 0$. An identical argument applies to $\boldsymbol{M}_{\text{post}}^{(d,p)}$.

Existence of Synthetic Control & Synthetic Interventions

We begin by stating a natural property that theoretically justifies SC-like methods and SI, i.e., artificially constructing control and treatment groups. Let v_n denote the n-th row of $\boldsymbol{V}$, given in (4.2), which is the latent factor associated with unit n.

Property 4.1.4. *Given intervention d and unit i, let v_i lie within* $\text{span}(\{v_n\}_{n \in \mathcal{I}^{(d)}})$. *That is, there exists a $\beta^{(d,i)} \in \mathbb{R}^{N^{(d)}}$ such that $v_i = \sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot v_n$.*

Proposition 4.1.1 (Existence of SC & SI). *Suppose Property 4.1.1 holds. For a given intervention d and unit i, suppose Property 4.1.4 holds. Then for all $(t, d', p) \in [T] \times [D] \times [P]$, we have*

$$M_{ti}^{(d',p)} = \sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot M_{tn}^{(d',p)}.$$

Interpretation. Under a low-rank tensor factor model, Proposition 4.1.1 states that the target unit can be expressed as a linear combination of every donor subgroup *across all measurements, interventions, and metrics*. Indeed, this is the key result that enables both (M)RSC and SI to "transfer" the learned linear model from the pre- to post-intervention period, even if the interventional frameworks differ between the two periods. In Proposition 4.1.2 below, we show that Property 4.1.4 holds with high-probability.

Proposition 4.1.2 (SC & SI Exist whp). *Suppose Property 4.1.1 holds. Let the N units be re-indexed as per some permutation chosen uniformly at random. Then for any unit i, Property 4.1.4 holds w.p. at least $1 - r/N^{(d)}$.*

Interpretation. By the union bound, $\beta^{(d,i)}$ exists for all d simultaneously w.p. at least $1 - \sum_{d=1}^{D} r/N^{(d)}$. Proposition 4.1.2 circumvents the "pathological" case where v_i is orthogonal to all other rows in $\boldsymbol{V}$. Since there are at most $r - 1$ such rows in any rank r matrix, Proposition 4.1.2 establishes that, with respect to the unit indexing randomness, this pathological case will not occur w.h.p. Importantly, we highlight that Proposition 4.1.2 *does not* imply that the units need to be randomly administered interventions, i.e., the random re-indexing is purely an analytical tool and makes no experimental statements.

Remark 4.1.1 (Linearity with respect to metric). By symmetry, we note that a linear relationship between interventions holds across units and time. Therefore, analogous statements such as Property 4.1.4 and Propositions 4.1.1 and 4.1.2 hold for interventions. However, as with units, it is important to note that this does not imply that interventions (or units) are linear combinations of one another, e.g., the chemical structure of drug A is not necessarily a linear combination of the chemical structures of other drugs. Rather, these results simply state, under a low-rank tensor factor model (Property 4.1.1), there is a linear relationship between interventions (or units) with respect to the outcome variable of interest (metric). In general, the low-rank notion simply materializes the ideal that structure is shared across dimensions of the tensor; hence, even though there are N units (and D interventions), there are only a few canonical profiles for which all units (and interventions) are linear combinations of these profiles under the metric of interest.

■ 4.1.2 Observed Outcomes

We assume every observed outcome is corrupted by noise and satisfies

$$Y_{tn}^{(d,p)} = M_{tn}^{(d,p)} + \varepsilon_{tn}^{(d,p)},$$

where $\varepsilon_{tn}^{(d,p)}$ represents measurement noise. We make the following assumptions.

Property 4.1.5 (Sub-gaussian noise). *Let $\varepsilon_{tn}^{(d,p)}$ be a sequence of independent mean zero sub-gaussian random variables with variance bounded by σ^2, and $\left\|\varepsilon_{t}^{(d,p)}\right\|_{\psi_2} \le K$, $\left\|\mathbb{E}[\varepsilon_{t}^{(d,p)} \otimes \varepsilon_{t}^{(d,p)}]\right\| \le \gamma^2$, where $\varepsilon_{t}^{(d,p)} = [\varepsilon_{tn}^{(d,p)} : n \in \mathcal{I}^{(d)}] \in \mathbb{R}^{N^{(d)}}$.*

Interpretation. Since $\varepsilon_{tn}^{(d,p)}$ are independent, K and γ^2 are constants. However, our analysis goes through for the more general case where the noise is *dependent across the donor units* for a given t; here, K and γ^2 quantify the level of dependence in the noise between the donors at a given time.

Property 4.1.6 (Missing at random). *The non-counterfactual entries of Z are independently observed w.p. $\rho \in (0, 1]$, i.e., $\pi_{tn}^{(p)}$, given by (1.2), are a sequence of i.i.d.* Bernoulli*(ρ) random variables.*

Interpretation. Beyond the unobservable counterfactuals, we allow the observed outcomes themselves to be missing at random.

Connection to Error-in-variable Regression

Without loss of generality, consider unit i in Proposition 4.1.1. Then, the observed outcomes for the target unit i during the pre-intervention period (response vector) follow a linear model. That is, for every $\mathcal{I}^{(d)}$ and any metric p,

$$Y_{ti}^{(1,p)} = \sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot M_{tn}^{(d,p)} + \varepsilon_{ti}^{(1,p)} \quad \text{for all } t \leq T_0. \tag{4.3}$$

Since we only ever observe a noisy instantiation of $M_{tn}^{(d,p)}$, namely $Y_{tn}^{(d,p)}$, this is exactly the setting of error-in-variable regression.

Parameter Estimation

To estimate the post-intervention counterfactual potential outcomes, we require a good estimate of $\beta^{(d,i)}$, given in (5.1). It is well known, however, that recovering the latent model parameter without any additional assumptions is ill-defined since infinitely many solutions to (5.1) exist. Thus, for the purposes of model identification, it is standard within the error-in-variable regression literature to assume, for instance, $\beta^{(d,i)}$ is sparse and $\boldsymbol{M}_{\text{pre}}^{(d,p)}$ satisfies the restricted eigenvalue condition; see Loh and Wainwright (2012) and references therein. However, for the purposes of prediction, we argue only the component of $\beta^{(d,i)}$ within the row space of $\boldsymbol{M}_{\text{pre}}^{(d,p)}$ matters since any component within the null space is mapped to zero. This particular $\beta^{(d,i)}$ is unique and has minimum ℓ_2-norm; we show PCR accurately estimates this vector, which may be of independent interest in the error-in-variable regression literature.

■ 4.2 Proof of Existence of SC and SI

■ 4.2.1 Proof of Proposition 4.1.1

Proof. For all $(t, d', k) \in [T] \times [D] \times [K]$, we have that

$$\begin{aligned} M_{ti}^{(d',k)} &= \sum_{\ell=1}^{r} u_{t\ell} \cdot v_{i\ell} \cdot w_{d'\ell} \cdot q_{k\ell} \\ &= \sum_{\ell=1}^{r} u_{t\ell} \cdot \Big(\sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot v_{n\ell} \Big) \cdot w_{d'\ell} \cdot q_{k\ell} \\ &= \sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot \Big(\sum_{\ell=1}^{r} u_{t\ell} \cdot v_{n\ell} \cdot w_{d'\ell} \cdot q_{k\ell} \Big) \\ &= \sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot M_{tn}^{(d',k)}. \end{aligned}$$

This completes the proof. ■

■ 4.2.2 Proof of Proposition 4.1.2

Proof. Fix any d, and recall $\mathcal{I}^{(d)}$ is the corresponding, randomly sampled donor group. Under Property 4.1.1, $\text{rank}(\boldsymbol{V}) = r$; hence, it must be that $\dim(\text{span}\{(v_n)_{n \in \mathcal{I}^{(d)} \cup \{1\}}\}) \leq r$. Since $\mathcal{I}^{(d)}$ and the target unit i are sampled randomly from $[N]$, the probability that $v_1 \notin \text{span}\{(v_n)_{n \in \mathcal{I}^{(d)}}\}$ is at most $r/N^{(d)}$ (since among the $\mathcal{I}^{(d)} \cup \{i\}$ units, there can be at most r linearly independent vectors). Thus, $\mathbb{P}(\mathcal{E}_d) \geq 1 - r/N^{(d)}$, where

$$\mathcal{E}_d := \left\{ \exists \beta^{(d,i)} \in \mathbb{R}^{N^{(d)}} \text{ s.t. } v_1 = \sum_{n \in \mathcal{I}^{(d)}} \beta_n^{(d,i)} \cdot v_n \right\}.$$

■

Chapter 5

Principal Component Regression

As stated in Chapter 4, we assume the potential outcomes follow a low-rank factor model (Property 4.1.1). In a high-dimensional framework where the ambient dimension of the feature space is large, it is well known (particularly, empirically) that Principal Component Regression (PCR), see Jolliffe (1982), is an effective prediction algorithm if the covariates exhibit low-dimensional structure; this motivates the usage of PCR in our context to predict counterfactual potential outcomes.

However, despite PCR's tremendous success in a variety of applications, its ability to handle settings with noisy, sparse, and mixed (discrete and continuous) valued covariates is not understood and remains an important challenge, cf. Chao et al. (2019). As a contribution of this thesis, we establish the robustness of PCR to these scenarios and provide finite-sample guarantees with respect to its parameter estimation and test (out-of-sample) prediction errors.

■ 5.1 Setup

We begin by describing our setup and formalizing our model, observations, and objective.

Training Data

Throughout this chapter, let M_{train} denote a $n \times p$ training covariate matrix of rank $r < n \wedge p$:

$$M_{\text{train}} = \sum_{i=1}^{r} s_i u_i \otimes v_i.$$

Rather than directly observing $\boldsymbol{M}_{\text{train}}$, we only have access to its corrupted version, $\boldsymbol{Z}_{\text{train}}$:

$$\boldsymbol{Z}_{\text{train}} = \sum_{i=1}^{n \wedge p} \widehat{s}_i \widehat{u}_i \otimes \widehat{v}_i.$$

We assume the observed covariates are generated via the model:

$$\boldsymbol{Z}_{\text{train}} = (\boldsymbol{M}_{\text{train}} + \boldsymbol{H}_{\text{train}}) \circ \boldsymbol{P}_{\text{train}},$$

where $\boldsymbol{H}_{\text{train}} \in \mathbb{R}^{n \times p}$ is the measurement noise and $\boldsymbol{P}_{\text{train}} \in \{0, 1\}^{n \times p}$ is a masking matrix. Finally, let $y \in \mathbb{R}^n$ denote the response vector (training labels), which is related to the underlying covariates $\boldsymbol{M}_{\text{train}}$ via the following linear model:

$$y = \boldsymbol{M}_{\text{train}} \beta^* + \varepsilon, \tag{5.1}$$

where $\beta^* \in \mathbb{R}^p$ is the underlying model parameter and $\varepsilon \in \mathbb{R}^n$ is the response noise.

Testing Data

We denote the SVD of the $m \times p$ rank $r' < m \wedge p$ testing covariate matrix $\boldsymbol{M}_{\text{test}}$ as

$$\boldsymbol{M}_{\text{test}} = \sum_{i=1}^{r'} = \tau_i \mu_i \otimes \nu_i.$$

Similar to the above setup, we denote the perturbed version of $\boldsymbol{M}_{\text{test}}$ as $\boldsymbol{Z}_{\text{test}}$, which is generated as

$$\boldsymbol{Z}_{\text{test}} = (\boldsymbol{M}_{\text{test}} + \boldsymbol{H}_{\text{test}}) \circ \boldsymbol{P}_{\text{test}},$$

and admits the following SVD:

$$\boldsymbol{Z}_{\text{test}} = \sum_{i=1}^{m \wedge p} \widehat{\tau}_i \widehat{\mu}_i \otimes \widehat{\nu}_i.$$

We note that $\boldsymbol{H}_{\text{test}} \in \mathbb{R}^{m \times p}$ and $\boldsymbol{P}_{\text{test}} \in \{0, 1\}^{m \times p}$ are defined analogously as above.

Objective

Given $(y, \boldsymbol{Z}_{\text{train}}, \boldsymbol{Z}_{\text{test}})$, our aim is to recover the true model parameter β^* and predict the underlying out-of-sample response vector $\boldsymbol{M}_{\text{test}} \beta^*$.

■ 5.2 Hard Singular Value Thresholding (HSVT)

Before we analyze PCR, we briefly digress to examine HSVT, a powerful matrix estimation technique to de-noise observations and impute missing values. As we will show in Proposition 5.3.1, HSVT is closely related to PCR.

■ 5.2.1 Algorithm

It is well known that the principal components of the data often capture most of the signal. Thus, much like the spirit of Principal Component Analysis (PCA) – a key subroutine of PCR – HSVT retains the top singular components of the data and shaves off the remaining singular components, which often represent noise. We state the HSVT algorithm below.

```
Algorithm 1: HSVT
  Data: $Z = \sum_i \widehat{s}_i \widehat{u}_i \otimes \widehat{v}_i \in \mathbb{R}^{n \times p}, k \in [n \wedge p]$
  Result: $\widehat{M} \in \mathbb{R}^{n \times p}$
    1. $\widehat{M} \leftarrow (1/\widehat{\rho}) \sum_{i=1}^{k} \widehat{s}_i \widehat{u}_i \otimes \widehat{v}_i$, where $\widehat{\rho}$ is the fraction of observed entries in $Z$
```

■ 5.2.2 Results

Due to its immense popularity, HSVT has been widely studied in the literature; however, much of the analyses has been devoted to bounding its estimation error with respect to the Frobenius and operator norms. For our purposes, however, we require a bound on the $\ell_{2,\infty}$-norm. To that end, we denote the HSVT estimation error of $\widehat{M}_{\text{train}}$ as

$$\mathcal{E}_{\text{HSVT}}(\widehat{M}_{\text{train}}) = \frac{1}{n} \left\| \widehat{M}_{\text{train}} - M_{\text{train}} \right\|_{2,\infty}^{2}.$$

We note that the HSVT error of $\widehat{M}_{\text{test}}$ is defined similarly. The following statement bounds the HSVT estimation error when the singular values are chosen correctly under the setting described in Chapter 4.

Lemma 5.2.1 (HSVT $\ell_{2,\infty}$-norm Error). *Suppose M_{train} satisfies Properties 4.1.2, 4.1.3, H_{train} satisfies Property 4.1.5, and P_{train} satisfies Property 4.1.6. Consider $\widehat{M}_{\text{train}} =$ HSVT(Z_{train}, r). For any $\delta > 0$ and some $C > 0$, if $\rho \geq \frac{C \log(1/\delta)}{np}$, then the following holds w.p. at least $1 - \delta$:*

$$\mathcal{E}_{\text{HSVT}}(\widehat{M}_{\text{train}}) \leq \frac{C_1}{\rho^4} \frac{r}{n \wedge p} + \Delta,$$

where

$$\Delta = C_2 \left(\frac{1}{\rho^2} \frac{\sqrt{r}}{n} + \frac{1}{\rho^4} \frac{r}{\sqrt{n}(n \wedge p)} \right) \sqrt{\log p}, \tag{5.2}$$

$C_1 = C(1+\sigma^4)(1+\gamma^2)(1+K^2)$, *and* $C_2 = C_1 K^2 (1 + \log^{3/2}(1/\delta))$.

Interpretation. Lemma 5.2.1 states that the HSVT estimation error vanishes as n, p grow. Further, we note that the above bound is stated in the more general noise setting where the rows of $\boldsymbol{H}_{\text{train}}$ are independent sub-gaussian random vectors rather than the entry-wise independence setting given by Property 4.1.5 (see Chapter 4.1.2 for details). Additionally, the bound holds between $\widehat{\boldsymbol{M}}_{\text{test}}$ and $\boldsymbol{M}_{\text{test}}$ with (n, r) replaced by (m, r'). Given the ubiquity of HSVT, Lemma 5.2.1 may be of interest in its own right.

■ 5.3 Principal Component Regression (PCR)

■ 5.3.1 Algorithm

Often, our underlying signal of interest has low-dimensional structure, but is masked by our perturbed observations that live in a high-dimensional ambient space. It is widely known, however, that the principal components (top singular components) of the data capture most of its information, and can be uncovered via Principal Component Analysis (PCA). As a result, PCR, which uses PCA is a key subroutine, is a widely used technique to extract the latent factors that drive the correlation structure of the data prior to learning a linear model. We state the PCR algorithm below.

```
Algorithm 2: PCR
  Data: $\boldsymbol{Z} = \sum_i \widehat{s}_i \widehat{u}_i \otimes \widehat{v}_i \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n, k \in [n \wedge p]$
  Result: $\widehat{\beta} \in \mathbb{R}^p$
    1. $\widehat{w} \leftarrow \arg\min_{w \in \mathbb{R}^k} \left\| y - (1/\widehat{\rho}) \boldsymbol{Z} \widehat{\boldsymbol{V}}_k w \right\|_2^2$, where $\widehat{\boldsymbol{V}}_k = [\widehat{v}_1, \dots, \widehat{v}_k] \in \mathbb{R}^{p \times k}$ and $\widehat{\rho}$ is the
       fraction of observed entries in $\boldsymbol{Z}$
    2. $\widehat{\beta} \leftarrow \widehat{\boldsymbol{V}}_k \widehat{w}$
```

PCR Intuition

Using the observed covariates, PCR first finds a $k \ll p$ dimensional representation for each feature via PCA; specifically, PCA projects every covariate Z_i onto the subspace spanned by the top k right singular vectors of the covariate matrix. PCR then uses the k-dimensional features to perform linear regression.

Choosing k

In general, the correct number of principal components $k = r$ to use is not known a priori. However, under reasonable signal-to-noise scenarios, Weyl's inequality (Lemma 3.1.1) implies that a "sharp" threshold or gap should exist between the top r singular values and remaining singular values of the observed data $\boldsymbol{Z}$. This gives rise to a natural "elbow" point and suggests choosing a threshold within this gap. Another standard approach is to use a "universal" thresholding scheme that preserves the singular values above a precomputed threshold (see Chatterjee (2015b) and Gavish and Donoho (2014)). Finally, data-driven approaches developed around cross-validation can also be employed.

■ 5.3.2 Equivalence

Since our observed covariates are contaminated, a natural algorithm to recovering a solution to (5.1) is to first de-noise our data via HSVT and then perform Ordinary Least Squares (OLS) to learn a linear model. We note that OLS can be equivalently viewed as PCR with hyper-parameter $k = n \wedge p$. Pleasingly, it turns out that HSVT with OLS is equivalent to PCR. This is formalized in Proposition 5.3.1.

Proposition 5.3.1 (Equivalence). *Let $k \in [n \wedge p]$. Suppose $\widehat{\beta}^{\text{PCR}} = \text{PCR}(\boldsymbol{Z}_{\text{train}}, y, k)$. Further, consider $\widehat{\boldsymbol{M}}_{\text{train}} = \text{HSVT}(\boldsymbol{Z}_{\text{train}}, k)$ and $\widehat{\beta}^{\text{HSVT}} = \text{PCR}(\widehat{\boldsymbol{M}}_{\text{train}}, y, n \wedge p)$. Then,*

$$\widehat{\beta}^{\text{PCR}} = \widehat{\beta}^{\text{HSVT}}$$
$$\boldsymbol{Z}_{\text{train}}\widehat{\beta}^{\text{PCR}} = \widehat{\boldsymbol{M}}_{\text{train}}\widehat{\beta}^{\text{HSVT}}.$$

Interpretation. Given the above equivalence, we can analyze properties of PCR through the lens of the HSVT estimator. In particular, we will utilize Lemma 5.2.1 to prove Lemma 5.3.1 and Theorems 5.3.1, 5.3.2.

■ 5.3.3 PCR Training Prediction Error

Evaluation Metric

For any estimate $\widehat{M}_{\text{train}} \in \mathbb{R}^n$ of $\boldsymbol{M}_{\text{train}}\beta^*$, we define the corresponding training (in-sample) prediction error as

$$\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}_{\text{train}}) = \frac{1}{n}\left\|\widehat{\boldsymbol{M}}_{\text{train}} - \boldsymbol{M}_{\text{train}}\beta^*\right\|_2^2.$$

Estimator

We consider $\widehat{\boldsymbol{M}}_{\text{train}} = \boldsymbol{Z}_{\text{train}}\widehat{\beta}$ with $\widehat{\beta} = \text{PCR}(\boldsymbol{Z}_{\text{train}}, y, r)$ as the estimate of $\boldsymbol{M}_{\text{train}}\beta^*$.

Results

We state the training error bound of $\widehat{\boldsymbol{M}}_{\text{train}}$ in high probability below.

Lemma 5.3.1 (PCR Train Error). *Suppose $(\boldsymbol{M}_{\text{train}}, \boldsymbol{M}_{\text{train}}\beta^*)$ satisfy Properties 4.1.2, $\boldsymbol{M}_{\text{train}}$ satisfies Property 4.1.3, $(\boldsymbol{H}_{\text{train}}, \varepsilon)$ satisfy Property 4.1.5, and $\boldsymbol{P}_{\text{train}}$ satisfies Property 4.1.6. Let β^* be any solution to (5.1). For any $\delta > 0$ and some $C > 0$, if $\rho \geq \frac{C\log(1/\delta)}{np}$, then the following holds w.p. at least $1-\delta$:*

$$\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}_{\text{train}}) \leq \frac{2\sigma^2 r}{n} + \frac{C_2}{\rho^4}\frac{r\sqrt{\log p}}{n \wedge p}\left\|\beta^*\right\|_1^2 + \Delta_1,$$

where

$$\Delta_1 = \frac{C_2}{\sqrt{n}}\left\|\beta^*\right\|_1, \; C_1 = C(1+\sigma^4)(1+\gamma^2)(1+K^2), \; C_2 = C_1K^2(1+\log^2(1/\delta)). \quad (5.3)$$

Interpretation. The first term in the result of Lemma 5.3.1 above represents the minmax error rate from low-dimensional ordinary least squares regression with noiseless covariates; the second term corresponds to the HSVT estimation error as the training covariate matrix is noisily observed; the third term corresponds to the error due to providing a high-probability bound (which will be absent if we choose the expected error to be our metric of choice).

■ 5.3.4 PCR Parameter Estimation Error

In a high-dimensional framework, there may be an infinite number of models that satisfy (5.1). However, as described in Chapter 4.1.2, only the component of β^* that lives within the rowspace of $\boldsymbol{M}_{\text{train}}$ is important for the purposes of prediction. For that reason, we consider, without loss of generality, the unique β^* of minimum ℓ_2-norm that satisfies (5.1). The following result states that PCR can recover this particular model parameter even in the presence of contaminated covariates.

Theorem 5.3.1 (PCR Parameter Estimation Error)**.** *Let the conditions of Lemma 5.3.1 hold. Further, let β^* be the unique vector of minimum ℓ_2-norm that satisfies (5.1). For any $\delta > 0$ and some $C > 0$, if $\rho \geq \sqrt{\frac{C_2 r}{n \wedge p}}$, then the following holds w.p. at least $1 - \delta$:*

$$\left\| \widehat{\beta} - \beta^* \right\|_2^2 \leq \frac{r}{p} \left(\frac{C\sigma^2 r}{n} + \frac{C_2}{\rho^4} \frac{r\sqrt{\log p}}{n \wedge p} \left\| \beta^* \right\|_1^2 + \Delta \right) + \frac{C_1}{\rho^2} \frac{r}{n \wedge p} \left\| \beta^* \right\|_2^2,$$

where C_1, C_2, Δ are given by (5.3).

Interpretation. The first term within the parentheses on the RHS is exactly the in-sample prediction error for PCR (given in Lemma 5.3.1). The second term is the additional cost paid for directly estimating β^*. Now, consider the special case where the following two conditions hold: (i) $\boldsymbol{M}_{\text{train}}$ is directly observed, i.e., there are no missing values or measurement errors in the donor data; and (ii) the columns of $\boldsymbol{M}_{\text{train}}$ are linearly independent (i.e., β^* lies within its row space). Then, the bound above agrees with classical parameter estimation results for ordinary least squares (e.g., Remark 2.3 of Rigollet and Hütter (2017) under Property 4.1.3.)

Comparison with Literature. Suppose $n \leq p$. Since $\left\| \beta^* \right\|_1 \leq \sqrt{p} \left\| \beta^* \right\|_2$, the parameter estimation error scales as $\mathcal{O}(\left\| \beta^* \right\|_2^2 / n)$ with high probability; this is with respect to the minimum ℓ_2-norm β^*. We note that this rate is in line with previous works (c.f. Loh and Wainwright (2012); Datta and Zou (2017); Rosenbaum and Tsybakov (2013)), where the error also grows as $\mathcal{O}(\left\| \beta^* \right\|_2^2 / n)$. Note, however, these previous works make a key sparsity assumption that $\left\| \beta^* \right\|_0 \leq r$, which we do not require for our results. Instead, we make a low-rank assumption on the covariate (donor) matrix. Further, the estimators proposed in Loh and Wainwright (2012); Datta and Zou (2017); Rosenbaum and Tsybakov (2013) explicitly require knowledge of the noise distribution (i.e., its second moment matrix). PCR, on the other hand, does not require this.

Synthetic Simulation

To study the scaling of the parameter estimation error, we performed simulations under an additive noise model. We detail the setup and results below.

Model. We construct our underlying training covariates $\boldsymbol{M} \in \mathbb{R}^{n \times p}$ via the probabilistic PCA model as described in Chapter 4.1.1. That is, we first generate $\boldsymbol{M}_r \in \mathbb{R}^{n \times r}$ by sampling each entry from a standard normal distribution, independently of other entries. 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

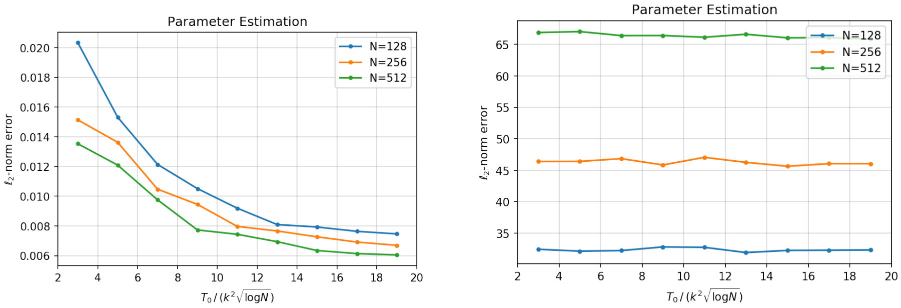
$\boldsymbol{M} = \boldsymbol{M}_r \boldsymbol{Q}$. We define rank$(\boldsymbol{M}) = r = p^{\frac{1}{4}}$, where $p \in \{128, 256, 512\}$.

Next, we generate $\beta \in \mathbb{R}^p$ by first sampling from a multivariate standard normal vector with independent entries and then arbitrarily scaling the resulting values by 5. The underlying response vector $a \in \mathbb{R}^n$ is then defined to be the product $a = \boldsymbol{M}\beta$. Finally, the model parameter of interest, β^*, is then computed as

$$\text{minimize} \quad \|w\|_2^2 \ \text{ subject to } \boldsymbol{M}w = a.$$

Observations. 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 zero and variance $\sigma^2 = 0.2$. The entries of $\boldsymbol{H} \in \mathbb{R}^{n \times p}$ are sampled in an identical fashion. We then define our observed response vector as $y = a + \varepsilon$ and corrupted covariate matrix as $\boldsymbol{Z} = \boldsymbol{M} + \boldsymbol{H}$.

Results. Using $(y, \boldsymbol{Z})$, we perform PCR to yield $\widehat{\beta}$. To demonstrate that PCR can accurately recover β^*, the minimum ℓ_2-norm solution, we compute the ℓ_2-norm parameter estimation error, or root-mean-squared-error (RMSE), with respect to β^* and β in Figures 5.1a and 5.1b, respectively. As suggested by Figure 5.1a, the RMSE with respect to β^* roughly aligns for different values of n, after rescaling the sample size as $n/(r^2\sqrt{\log p})$, and decays to zero as the sample size increases; this is predicted by Theorem 5.3.1. On the other hand, Figure 5.1b shows that the RMSE with respect to β stays roughly constant across different values of p. Therefore, as established in Theorem 5.3.1, PCR performs recovers the minimum-norm solution.



(a) ℓ_2-norm error of $\widehat{\beta}$ with respect to the minimum ℓ_2-norm solution of (5.1), i.e., β^*.

(b) ℓ_2-norm error of $\widehat{\beta}$ with respect to a random solution to (5.1).

Figure 5.1: Plots of ℓ_2-norm error against β^* in 5.1a and β in 5.1b, versus the rescaled sample size $n/(r^2\sqrt{\log p})$ after running PCR with rank $r = p^{\frac{1}{4}}$. As predicted by Theorem 5.3.1, the curves for different values of p under 5.1a roughly align and decay to zero as n increases.

Remark. Recall the discussion of Theorem 5.3.1. That is, if we apply the inequality $\|\beta^*\|_1 \leq \sqrt{p}\|\beta^*\|_2$, then for any fixed δ, the parameter estimation error scales as $\mathcal{O}(r^2\sqrt{\log p}\|\beta^*\|_2^2/(n \wedge p))$. Thus, we choose our rescaled sample size to be $n/(r^2\sqrt{\log p})$.

■ 5.3.5 PCR Testing Prediction Error

Evaluation Metric

For any estimate $\widehat{M}_{\text{test}} \in \mathbb{R}^m$ of $\boldsymbol{M}_{\text{test}}\beta^*$, we define the corresponding testing (out-of-sample) prediction error as

$$\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) = \frac{1}{m}\left\|\widehat{M}_{\text{test}} - \boldsymbol{M}_{\text{test}}\beta^*\right\|_2^2.$$

Estimator

Assume $\boldsymbol{M}_{\text{test}}\beta^*$ satisfies Property 4.1.2. We consider the following prediction estimate of $\boldsymbol{M}_{\text{test}}\beta^*$. Let $\widehat{M}_{\text{test}} = \text{Truncate}(\widehat{\boldsymbol{M}}_{\text{test}}\widehat{\beta})$ be a truncated version of $\widehat{\boldsymbol{M}}_{\text{test}}\widehat{\beta}$, where $\widehat{\boldsymbol{M}}_{\text{test}} = \text{HSVT}(\boldsymbol{Z}_{\text{test}}, r')$, $\widehat{\beta} = \text{PCR}(\boldsymbol{Z}_{\text{train}}, y, r)$, and $\text{Truncate}(\cdot)$ is defined below.

```
Algorithm 3: Truncate
  Data: $X = [X_i : i \leq n]$
  Result: $X^{(\text{trunc})} = [X^{(\text{trunc})} : i \leq n]$
  1. For every $i \leq n$:
        $X_i^{(\text{trunc})} = \begin{cases} X_i, & \text{if } X_i \in [-1, 1] \\ 1, & \text{if } X_i > 1 \\ -1, & \text{if } X_i < -1. \end{cases}$
```

Interpretation. In words, since the underlying test responses are assumed to be bounded under Property 4.1.2, we also restrict our estimates to lie within the unit interval.

Results

In what follows, we denote the right singular vectors of the underlying training ($\boldsymbol{M}_{\text{train}}$) and testing ($\boldsymbol{M}_{\text{test}}$) covariate matrices as $\boldsymbol{V}_{\text{train}} \in \mathbb{R}^{p \times r}$ and $\boldsymbol{V}_{\text{test}} \in \mathbb{R}^{p \times r'}$, respectively. We are now ready to state PCR's test error bound in both high probability and expectation.

Theorem 5.3.2 (PCR Test Error in High-Probability). *Let the conditions of Theorem 5.3.1 hold. Suppose* $(\boldsymbol{M}_{\text{test}}, \boldsymbol{M}_{\text{test}}\beta^*)$ *satisfy Property 4.1.2,* $\boldsymbol{M}_{\text{test}}$ *satisfies Property 4.1.3,* $\boldsymbol{H}_{\text{test}}$ *satisfies Property 4.1.5, and* $\boldsymbol{P}_{\text{test}}$ *satisfies Property 4.1.6. Let* $\text{span}(\boldsymbol{V}_{\text{test}}) \subseteq \text{span}(\boldsymbol{V}_{\text{train}})$*. For any* $\delta > 0$*, if* $\rho \geq \sqrt{\frac{C_1 C_2 r}{n \wedge m \wedge p}}$*, then the following holds w.p. at least* $1 - \delta$*:*

$$\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \leq \frac{r}{r'}\left(\frac{C\sigma^2 r}{n} + \frac{C_1 C_2}{\rho^4}\frac{r\sqrt{\log p}}{n \wedge m \wedge p}\|\beta^*\|_1^2 + \frac{C_1^2}{\rho^4}\frac{rp}{(n \wedge m \wedge p)^2}\|\beta^*\|_2^2 + \Delta\right),$$

where C_1, C_2, Δ *are given by* (5.3)*.*

Corollary 5.3.1 (PCR Test Error in Expectation). *Let the conditions of Theorem 5.3.2 hold. Then for any* $\delta > 0$*,*

$$\mathbb{E}[\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}})] \leq \frac{r}{r'}\left(\frac{2\sigma^2 r}{n} + \frac{C_3 C_4}{\rho^4}\frac{r\log^2(p/\delta)}{n \wedge m \wedge p}\|\beta^*\|_1^2 + \frac{C_4^2}{\rho^4}\frac{rp}{(n \wedge m \wedge p)^2}\|\beta^*\|_2^2\right) + 4\delta,$$

where $C_3 = CK^2(1+\sigma^4)(1+\gamma^2)(1+K^2)$ *and* $C_4 = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$*.*

Interpretation. For simplicity, we let $n = \Theta(m) = \Theta(p)$, and suppress dependencies on β^* and $\log p$. The error bound in Theorem 5.3.2 is quantified by four terms: (i) the first term, scaling as $\mathcal{O}(r/n)$, corresponds to the minmax in-sample prediction error for low-dimensional ordinary least squares with noiseless covariates; (ii) the second term, scaling as $\mathcal{O}(r/(\rho^4 n))$ is the additional error due to the sparsity and noise in the covariates; (iii) the third term, scaling as $\mathcal{O}(r/(\rho^4 n))$ is the generalization error; (iv) the fourth term, $\Delta = \mathcal{O}(1/\sqrt{n})$, disappears if the error is taken in expectation (Corollary 5.3.1); finally, the scaling, r/r', is the ratio of the training and testing covariate matrix ranks, which may be a remnant of our proof technique.

Remarks. It is worth mentioning that Theorem 5.3.2 *does not make any distributional assumptions of having i.i.d. covariates*. Instead, the key assumption that enables Theorem 5.3.2 is a linear algebraic condition: $\text{span}(\boldsymbol{V}_{\text{test}}) \subseteq \text{span}(\boldsymbol{V}_{\text{train}})$. This allows PCR to "generalize" to unseen data. A more general test error result when this condition fails to hold can be found in Lemma 5.12.3 of Section 5.12.

■ 5.4 A Subspace Inclusion Hypothesis Test

As shown in Theorem 5.3.2, a key assumption that enables PCR to generalize is $\text{span}(\boldsymbol{V}_{\text{test}}) \subseteq \text{span}(\boldsymbol{V}_{\text{train}})$. This condition gives rise to a natural hypothesis test:

$$H_0 : \text{span}(\boldsymbol{V}_{\text{test}}) \subseteq \text{span}(\boldsymbol{V}_{\text{train}})$$
$$H_1 : \text{span}(\boldsymbol{V}_{\text{test}}) \nsubseteq \text{span}(\boldsymbol{V}_{\text{train}}).$$

Under H_0, we have $\boldsymbol{V}_{\text{test}} = \mathcal{P}_{V_{\text{train}}} \boldsymbol{V}_{\text{test}}$, where $\mathcal{P}_{V_{\text{train}}} = \boldsymbol{V}_{\text{train}} \boldsymbol{V}_{\text{train}}^T$. However, given that the right singular vectors are never observable, we use the top right singular vectors of $\boldsymbol{Z}_{\text{train}}, \boldsymbol{Z}_{\text{test}}$, denoted as $\widehat{\boldsymbol{V}}_{\text{train}}, \widehat{\boldsymbol{V}}_{\text{test}}$, as proxies. Hence, a natural test statistic is the gap between $\widehat{\boldsymbol{V}}_{\text{test}}$ and $\mathcal{P}_{\widehat{V}_{\text{train}}} \widehat{\boldsymbol{V}}_{\text{test}}$. In particular,

$$\widehat{\tau} = \left\| \widehat{\boldsymbol{V}}_{\text{test}} - \mathcal{P}_{\widehat{V}_{\text{train}}} \widehat{\boldsymbol{V}}_{\text{test}} \right\|_F^2 \underset{H_1}{\overset{H_0}{\lesseqgtr}} \tau_\alpha, \tag{5.4}$$

where $\widehat{\tau}$ is our test statistic, τ_α is the critical value, and α is the significance level.

Theorem 5.4.1 (Subspace Inclusion Type I Error). *Suppose* $(\boldsymbol{M}_{\text{train}}, \boldsymbol{M}_{\text{test}})$ *satisfy Properties 4.1.2, 4.1.3,* $(\boldsymbol{H}_{\text{train}}, \boldsymbol{H}_{\text{test}})$ *satisfy Property 4.1.5, and* $(\boldsymbol{P}_{\text{train}}, \boldsymbol{P}_{\text{test}})$ *satisfy Property 4.1.6. Consider* $\widehat{\boldsymbol{M}}_{\text{train}} = \text{HSVT}(\boldsymbol{Z}_{\text{train}}, r)$ *and* $\widehat{\boldsymbol{M}}_{\text{test}} = \text{HSVT}(\boldsymbol{Z}_{\text{test}}, r')$*. For any* $\alpha \in (0,1)$ *and some* $C > 0$*, if* $\rho \geq \frac{C \log(1/\alpha)}{(n \wedge m)p}$*, then* $\mathbb{P}\left(\widehat{\tau} \geq \tau_\alpha | H_0\right) \leq \alpha$ *with*

$$\tau_\alpha = \frac{C'}{\rho^2} \left(\frac{rr'}{n \wedge m \wedge p} + \frac{rr' \log(1/\alpha)}{(n \wedge m)p} \right),$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$.

Interpretation. Returning to our problem of causal inference, we note that (5.4) also functions as a quantitative hypothesis test to check when we can apply SI and thus SC, something which, to the best of our knowledge, is missing from the literature. Roughly speaking, if our test statistic is smaller than our critical value, then SI can extrapolate from our observed outcomes to estimate the unobservable potential outcomes of interest. As we will see in our empirical studies, the post-intervention (cross-validation) prediction error also corresponds closely to whether this hypothesis test passes or fails, as desired.

■ 5.5 Discussion

■ 5.5.1 New Perspective on High-Dimensional Regression

It is well established that lasso (ℓ_1-regularization) methods have been the de facto algorithmic approach in the context of high-dimensional regression problems. This is largely motivated because identifying model parameters (e.g., β^*) is of fundamental interest in statistics. As such, the sparsity constraint provides one notion of identifiability that is achievable; in turn, this perspective has spurred tremendous theoretical work and has had a profound impact in many important real-world applications (e.g., magnetic resonance imaging Lustig et al. (2007); Lustig et al. (2008)). Meanwhile, we hope that our PCR results offer a new, complementary perspective, where the notion of identifiability is associated with the minimum ℓ_2-norm model parameter – in the context of prediction, this also pleasingly corresponds to the only model of importance. Additionally, it is worth remarking that an added benefit of PCR is that its conditions can be verified in practice by simply observing the spectrum of the covariates. That is, if they exhibit low-dimensional structure, then our results suggest that PCR should achieve desirable prediction performance. This is contrast with standard sparsity assumptions, which are arduous to verify in general.

■ 5.5.2 PCR Robustness Properties

Implicit De-noising

By Theorem 5.3.1, we argue that PCR, without any change, is robust to noisy and sparse covariates. In particular, despite only having access to $\boldsymbol{Z}_{\text{train}}$ – the corrupted version of $\boldsymbol{M}_{\text{train}}$ with noisy, missing entries – we show that PCR recovers the underlying model parameter β^* with high probability. In fact, PCR's parameter estimation error rate matches the minmax rate achieved by OLS (up to logarithmic factors) if one had perfectly observed the true covariates.

Noise Agnostic

Importantly, we note that PCR does not require *any* knowledge about the underlying noise model that corrupts the covariates to achieve consistency with respect to both parameter estimation (Theorem 5.3.1) and prediction (Theorem 5.3.2 and Corollary 5.3.1). This can be also be seen through PCR's connection to HSVT, as stated by Proposition 5.3.1, since it is well-known that HSVT can recover the ground-truth matrix from its noisy observations without knowledge of the underlying noise distribution (e.g., see Chatterjee (2015a)). In contrast, despite the exciting recent advancements in the high-dimensional error-in-variable regression literature (e.g., Loh and Wainwright (2012); Datta and Zou (2017); Rosenbaum and Tsybakov (2013)), the current inventory of methods falls short as they require knowledge of the underlying covariate noise model (i.e., its second moment matrix) to recover a sparse model parameter.

Implicit Regularization

It is well-known that PCR can be viewed as a regularized estimator. To see this, consider any $k \in [n \wedge p]$. Then $\widehat{\beta} = \text{PCR}(\boldsymbol{Z}, y, k)$ can equivalently be expressed as

$$\widehat{\beta} = \arg\min_{w \in \mathbb{R}^p} \left\| y - (1/\widehat{\rho}) \boldsymbol{Z} w \right\|_2^2 \quad \text{subject to} \quad \widehat{\boldsymbol{V}}_{p-k}^T w = 0,$$

where $\widehat{\boldsymbol{V}}_{p-k} = [\widehat{v}_{k+1}, \dots, \widehat{v}_p] \in \mathbb{R}^{p \times (p-k)}$ denotes an orthonormal basis that is orthogonal to the subspace spanned by the top k principal components of $\boldsymbol{Z}$. Thus, if only a proper subset of principal components are chosen (i.e., $k < n \wedge p$), then PCR enforces a hard constraint that the resulting estimator must lie within the subspace spanned by the selected principal components.

■ 5.5.3 Regression with Mixed Valued Covariates

Setup and Question

Regression models with mixed discrete and continuous covariates are ubiquitous in practice. With respect to discrete covariates, a standard generative model assumes the covariates are generated from a categorical distribution (i.e., a multinomial distribution). Formally, a categorical distribution for a random variable X is such that X has support in $[G]$ and the probability mass function (pmf) is given by $\mathbb{P}(X = g) = \rho_g$ for $g \in [G]$ with $\sum_{g=1}^{G} \rho_g = 1$.

For simplicity, we focus on the case where the regression is being done with a collection of Bernoulli random variables (i.e., each X has support in $\{0, 1\}$). The extension to general categorical random variables, in addition to continuous covariates, is straightforward and discussed below.

A standard model in regression with Bernoulli random variables assumes that the response variable is a linear function of the latent parameters of the observed discrete outcomes. Formally, $\boldsymbol{M}_i = [\rho_1^{(i)}, \rho_2^{(i)}, \dots, \rho_p^{(i)}]^T \in \mathbb{R}^p$, where $\rho_j^{(i)}$ for $j \in [p]$ is the latent Bernoulli parameter for the j-th feature and i-th measurement. Further, the mean of the response variable satisfies $\mathbb{E}[y_i] = \sum_{j=1}^{p} \rho_j^{(i)} \beta_j$. However, for each feature, we only get binary observations, i.e., $\boldsymbol{Z} = [Z_{ij}] \in \{0,1\}^{n \times p}$.

As an example, consider $\mathbb{E}[Z_i]$ to be the expected health outcome of patient i. Let there be a total of p possible observable binary symptoms (e.g., cold, fever, headache, etc.). Then $\boldsymbol{M}_i$ denotes the vector of (unobserved) probabilities that patient i has some collection of symptoms (e.g., $M_{i1} = \mathbb{P}(\text{patient } i \text{ has a cold}), M_{i2} = \mathbb{P}(\text{patient } i \text{ has a fever}), \dots)$. However, for each patient, we only observe the binary outcome of these symptoms (i.e., $Z_{i1} = \mathbb{1}(\text{patient } i \text{ has a cold}), Z_{i2} = \mathbb{1}(\text{patient } i \text{ has a fever})$), even though the response is linearly related with the underlying probabilities of the symptoms. The objective in such a setting is to accurately recover $\boldsymbol{M}\beta^*$ given $(y, \boldsymbol{Z})$.

Current Practice

A common practice for regression with categorical variables is to build a separate regression model for every possible combination of the categorical outcomes (i.e., to build a separate regression model conditioned on each outcome). In the healthcare example above, this would amount to building 2^p separate regression models corresponding to each combination of the observed p binary symptoms. This is clearly not ideal for the following two major reasons: (i) the sample complexity is exponential in p; (ii) we do not have access to the underlying probabilities $\boldsymbol{M}_i$ (recall $\boldsymbol{Z}_i \in \{0,1\}^p$), which is what we actually want to regress y against.

Returning to our Framework

Recall from Property 4.1.5 that the key structure we require of the covariate noise is that $\mathbb{E}[\boldsymbol{H}] = 0$. Now even though $Z_{ij} \in \{0,1\}$, it still holds that $\mathbb{E}[Z_{ij}] = \rho_j^{(i)} = M_{ij}$, which immediately implies $\mathbb{E}[\boldsymbol{H}] = \mathbb{E}[\boldsymbol{Z} - \boldsymbol{M}] = 0$. Further, the entries of $\boldsymbol{H}$ are sub-Gaussian as they are bounded by one in absolute value. Thus, the key conditions on the noise are satisfied for PCR to effectively (in the $\ell_{2,\infty}$-norm) de-noise $\boldsymbol{Z}$ to recover the underlying probability matrix $\boldsymbol{M}$; this, in turn, allows PCR to produce accurate estimates $\widehat{\boldsymbol{M}}\widehat{\beta}$ through regression, as seen by Theorem 5.3.2. Pleasingly, the required sample complexity grows with the rank of $\boldsymbol{M}$ (the inherent model complexity), rather than exponentially in p. Further, the de-noising step allows us to regress against the estimated latent probabilities rather than their “noisy”, binary outcomes.

Extension from Bernoulli to General Categorical Distributions

Recall from above that a categorical random variable has support in $[G]$ for $G \in \mathbb{N}$. In this case, one can translate a categorical random variable to a a collection of binary random variables using the standard one-hot encoding method. It is worth highlighting that by using one-hot encoding, the entries within any row of $\boldsymbol{H}$ are not independent as they encodes the same categorical variable. However, from Property 4.1.5, we only require independence of the noise across rows, not within them. Thus, this lack of independence is not an issue. Further, the generalization to multiple categorical variables, in addition to continuous covariates, is achieved by simply appending these features to each row and collectively de-noising the entire matrix before the regression step.

5.6 Proofs: Equivalence

Proof of Proposition 5.3.1. To prove the equivalence of the model parameter estimators, observe that

$$\widehat{w} = (\boldsymbol{Z}_{\text{train}} \widehat{\boldsymbol{V}}_k)^{\dagger} y = (\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k)^{\dagger} y = \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y,$$

where $\widehat{w}$ is defined as in Algorithm 2. As a result, it follows that $\widehat{\beta}^{\text{PCR}} = \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y$. Since $\widehat{\boldsymbol{M}}_{\text{train}} = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^T$, we have that

$$\widehat{\beta}^{\text{HSVT}} = \widehat{\boldsymbol{M}}_{\text{train}}^{\dagger} y = \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y.$$

This establishes the first equivalence.

Using the above, observe that

$$\boldsymbol{Z}_{\text{train}} \widehat{\beta}^{\text{PCR}} = \widehat{\boldsymbol{U}} \widehat{\boldsymbol{S}} \widehat{\boldsymbol{V}}^T \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T y$$

and

$$\widehat{\boldsymbol{M}}_{\text{train}} \widehat{\beta}^{\text{HSVT}} = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^T \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^T y = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^T y.$$

This concludes the proof. ■

5.7 Proof Notations

For ease of notation, we adopt the following notation throughout the rest of the proofs in this chapter:

Underlying Covariates. We denote the latent training and testing covariate matrices as

$$\boldsymbol{M} = \boldsymbol{M}_{\text{train}} \quad \text{and} \quad \boldsymbol{M}' = \boldsymbol{M}_{\text{test}},$$

respectively. Further, we assume their SVDs take the following form:

$$\boldsymbol{M} = \boldsymbol{U}_M \boldsymbol{S}_M \boldsymbol{V}_M^T = \sum_{\ell=1}^{r} s_\ell u_\ell \otimes v_\ell \quad \text{and} \quad \boldsymbol{M}' = \boldsymbol{U}_{M'} \boldsymbol{S}_{M'} \boldsymbol{V}_{M'}^T = \sum_{\ell=1}^{r'} s'_\ell u'_\ell \otimes v'_\ell.$$

Perturbations. We denote the training and testing covariate perturbations as

$$\boldsymbol{H} = \boldsymbol{H}_{\text{train}} \quad \text{and} \quad \boldsymbol{H}' = \boldsymbol{H}_{\text{test}}.$$

Observed Covariates. We denote the observed training and testing covariate matrices as

$$\boldsymbol{Z} = \boldsymbol{Z}_{\text{train}} \quad \text{and} \quad \boldsymbol{Z}' = \boldsymbol{Z}_{\text{test}},$$

respectively, which admit the following SVDs:

$$\boldsymbol{Z} = \widehat{\boldsymbol{U}}_Z \widehat{\boldsymbol{S}}_Z \widehat{\boldsymbol{V}}_Z^T = \sum_{\ell \geq 1} \widehat{s}_\ell \widehat{u}_\ell \otimes \widehat{v}_\ell \quad \text{and} \quad \boldsymbol{Z}' = \widehat{\boldsymbol{U}}_{Z'} \widehat{\boldsymbol{S}}_{Z'} \widehat{\boldsymbol{V}}_{Z'}^T = \sum_{\ell \geq 1} \widehat{s}'_\ell \widehat{u}'_\ell \otimes \widehat{v}'_\ell.$$

Due to the perturbations of $\boldsymbol{H}$ and $\boldsymbol{H}'$, $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ may be full-rank.

Estimators. We denote the estimates of the latent training and testing covariate matrices via HSVT as

$$\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r) \quad \text{and} \quad \widehat{\boldsymbol{M}'} = \text{HSVT}(\boldsymbol{Z}', r'),$$

respectively, with the following SVDs:

$$\widehat{\boldsymbol{M}} = \frac{1}{\widehat{\rho}} \widehat{\boldsymbol{U}}_M \widehat{\boldsymbol{S}}_M \widehat{\boldsymbol{V}}_M^T = \frac{1}{\widehat{\rho}} \sum_{\ell=1}^{r} \widehat{s}_\ell \widehat{u}_\ell \otimes \widehat{v}_\ell \quad \text{and} \quad \widehat{\boldsymbol{M}'} = \frac{1}{\widehat{\rho}'} \widehat{\boldsymbol{U}}_{M'} \widehat{\boldsymbol{S}}_{M'} \widehat{\boldsymbol{V}}_{M'}^T = \frac{1}{\widehat{\rho}'} \sum_{\ell=1}^{r'} \widehat{s}'_\ell \widehat{u}'_\ell \otimes \widehat{v}'_\ell,$$

where $(\widehat{\rho}, \widehat{\rho}')$ are the proportion of observed entries in $(\boldsymbol{Z}, \boldsymbol{Z}')$.

Projection Matrices. For any matrix $\boldsymbol{Q}$ with orthonormal columns, let $\mathcal{P}_Q := \boldsymbol{Q}\boldsymbol{Q}^T$ denote the projection matrix onto the subspace spanned by the columns of $\boldsymbol{Q}$.

Constants & Model Parameters. Throughout these proofs, we will let $C > 0$ denote an absolute constant that is independent of any model parameters. For ease of notation, we will allow the value of C to change from line to line. The dependencies on the model parameters (e.g., σ, γ, K) will be made explicit.

Remark 5.7.1. *Although Property 4.1.5 states that the entries of $\boldsymbol{H}$ and $\boldsymbol{H}'$ are independent, our analyses allow for independent rows (as opposed to entries). For the proofs to follow through, we only need the response (target unit) noise to be independent from the covariate (donor) noise. Thus, for the remainder of these proofs, we operate in the more general setting where we have row-wise independence of $\boldsymbol{H}$ and $\boldsymbol{H}'$, instead of just restricted entry-wise independence.*

■ 5.8 Proofs: Impact of Measurement Noise and Sparsity

In this section, we study the impact of measurement noise and sparsity in the covariate observations through the matrix $\boldsymbol{Z} - \rho \boldsymbol{M}$. Specifically, we analyze the impact of perturbations through the operator (spectral) norm and $\ell_{2,\infty}$-norm, and state the primary results in Lemmas 5.8.4 and 5.8.6, respectively. These results will be critical as we bound the prediction and parameter estimation errors in high probability.

Importantly, we highlight that the following results hold for any $\boldsymbol{M}$ that satisfies Property 4.1.2, $\boldsymbol{H}$ that satisfies Property 4.1.5, and $\boldsymbol{Z}$ that satisfies Property 4.1.6. For any $n \times p$ matrix $\boldsymbol{Q}$, let $Q_i \in \mathbb{R}^p$ and $Q_j \in \mathbb{R}^n$ denote the i-th row and j-th column of $\boldsymbol{Q}$, respectively.

■ 5.8.1 Operator Norm

Lemma 5.8.1. *Suppose that $Y \in \mathbb{R}^n$ and $P \in \{0,1\}^n$ are random vectors. Then,*

$$\left\| Y \circ P \right\|_{\psi_2} \leq \left\| Y \right\|_{\psi_2}.$$

Proof. Given a deterministic binary vector $P_0 \in \{0,1\}^n$, let $I_{P_0} = \{i \in [n] : P_{0i} = 1\}$. Observe that

$$Y \circ P_0 = \sum_{i \in I_{P_0}} e_i e_i^T Y.$$

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

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

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

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

For any $u \in \mathbb{S}^{n-1}$ and $P_0 \in \{0,1\}^n$, observe that

$$\mathbb{E}_Y\left[\exp\left(\left|\left(\sum_{i \in I_P} e_i e_i^T u\right)^T Y\right|^2/t^2\right) \;\middle|\; P = P_0\right] \le \mathbb{E}_Y\left[\exp\left(|u_0^T Y|^2/t^2\right)\right].$$

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

$$\left\|Y \circ P\right\|_{\psi_2} \le \left\|Y\right\|_{\psi_2}.$$

■

Lemma 5.8.2. *Assume Properties 4.1.2, 4.1.5, 4.1.6 hold. Then for all $i \in [m]$,*

$$\left\|Z_i - \rho M_i\right\|_{\psi_2} \le C(1 + K).$$

Proof. Let $\boldsymbol{P} \in \{0,1\}^{m \times n}$ denote a random matrix of independent Bernoulli random variables with parameter ρ. Further, let $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{H}$. Note that $Z_i = Y_i \circ P_i$ when Property 4.1.6 is assumed and $\star$ is identified with 0. By triangle inequality,

$$\begin{aligned}
\left\|Z_i - \rho M_i\right\|_{\psi_2} &= \left\|(Y_i \circ P_i) - \rho M_i\right\|_{\psi_2} \\
&= \left\|(Y_i \circ P_i) - (M_i \circ P_i) - \rho M_i + (M_i \circ P_i)\right\|_{\psi_2} \\
&\le \left\|(Y_i - M_i) \circ P_i\right\|_{\psi_2} + \left\|(M_i \circ P_i) - \rho M_i\right\|_{\psi_2}.
\end{aligned}$$

By the definition of $\boldsymbol{Y}$, Property 4.1.5, and Lemma 5.8.1, we have that

$$\left\|(Y_i - M_i) \circ P_i\right\|_{\psi_2} \leq \left\|Y_i - M_i\right\|_{\psi_2} = \left\|H_i\right\|_{\psi_2} \leq K.$$

Moreover, Property 4.1.2 and the i.i.d. property of $\boldsymbol{P}_{ij}$ for different j gives

$$\begin{aligned}
\left\|(M_i \circ P_i) - \rho M_i\right\|_{\psi_2}^2 &= \sup_{u \in \mathbb{S}^{n-1}} \left\| \sum_{j=1}^{n} u_j M_{ij} \left(P_{ij} - \rho\right) \right\|_{\psi_2}^2 \\
&\leq \sup_{u \in \mathbb{S}^{n-1}} \sum_{j=1}^{n} u_j^2 \left\|M_{ij}(P_{ij} - \rho)\right\|_{\psi_2}^2 \\
&\leq \left(\sup_{u \in \mathbb{S}^{n-1}} \sum_{j=1}^{n} u_j^2 \max_{i \in [m]} \left|M_{ij}\right|^2 \right) \cdot \left\|P_{11} - \rho\right\|_{\psi_2}^2 \\
&\leq \left\|P_{11} - \rho\right\|_{\psi_2}^2.
\end{aligned}$$

The first inequality follows from Property 3.2.2, the second inequality is immediate, and the last inequality follows from Property 4.1.2. Lastly, $\left\|P_{11} - \rho\right\|_{\psi_2} \leq C$ because $P_{11} - \rho$ is a bounded random variable in $[-\rho, 1-\rho]$. ■

Lemma 5.8.3. *Suppose Property 4.1.6 holds. Then,*

$$\left\|\mathbb{E}(\boldsymbol{Z} - \rho \boldsymbol{M})^T(\boldsymbol{Z} - \rho \boldsymbol{M})\right\| \leq \rho(1-\rho) \left(\left\|\text{diag}(\boldsymbol{M}^T \boldsymbol{M})\right\| + \left\|\text{diag}(\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}])\right\| \right) + \rho^2 \left\|\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}]\right\|.$$

Proof. We follow the proof of Lemma A.2 of Shah and Song (2018) and state it here for completeness. Throughout, for any matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times p}$, let $Q_\ell \in \mathbb{R}^p$ denote its ℓ-th row.

To begin, let $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{H}$. Further, observe that

$$\mathbb{E}[(\boldsymbol{Z} - \rho \boldsymbol{M})^T(\boldsymbol{Z} - \rho \boldsymbol{M})] = \sum_{\ell=1}^{n} \mathbb{E}[(Z_\ell - \rho M_\ell) \otimes (Z_\ell - \rho M_\ell)].$$

Importantly, we highlight the following relations: for any $(\ell, i) \in [n] \times [p]$,

$$\begin{aligned}
&\mathbb{E}[Z_{\ell i}] = \rho M_{\ell i} \\
&\mathbb{E}[Z_{\ell i}^2] = \rho^2 \mathbb{E}[Y_{\ell i}^2].
\end{aligned}$$

Now, let us fix a row $\ell \in [n]$ and denote

$$\boldsymbol{W}^{(\ell)} = (Z_\ell - \rho M_\ell) \otimes (Z_\ell - \rho M_\ell).$$

Using the linearity of expectations, the expected value of the (i, j)-th entry of $\boldsymbol{W}^{(\ell)}$ can be written as

$$\mathbb{E}[W_{ij}^{(\ell)}] = \mathbb{E}[Z_{\ell i} Z_{\ell j}] - \rho \mathbb{E}[Z_{\ell i} M_{\ell j}] - \rho \mathbb{E}[Z_{\ell j} M_{\ell i}] + \rho^2 \mathbb{E}[M_{\ell i} M_{\ell j}].$$

Suppose $i = j$, then

$$\mathbb{E}[W_{ii}^{(\ell)}] = \rho \mathbb{E}[Y_{\ell i}^2] - \rho^2 M_{\ell i}^2 = \rho(1-\rho)\mathbb{E}[Y_{\ell i}^2] + \rho^2 \mathbb{E}[(Y_{\ell i} - M_{\ell i})^2]. \tag{5.5}$$

On the other hand, if $i \neq j$,

$$\mathbb{E}[W_{ij}^{(\ell)}] = \rho^2 \mathbb{E}[(Y_{\ell i} - M_{\ell i})(Y_{\ell j} - M_{\ell j})]. \tag{5.6}$$

Therefore, we can express $\boldsymbol{W}^{(\ell)}$ as the sum of two matrices where the diagonal components are generated from (5.5) and the off-diagonal components are generated from (5.6), i.e.,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{W}^{(\ell)}] &= \mathbb{E}\Big(\rho(1-\rho)\text{diag}(Y_\ell \otimes Y_\ell) + \rho^2 \text{diag}(H_\ell \otimes H_\ell)\Big) + \mathbb{E}\Big(\rho^2 (H_\ell \otimes H_\ell) - \rho^2 \text{diag}(H_\ell \otimes H_\ell)\Big) \\
&= \rho(1-\rho)\mathbb{E}[\text{diag}(Y_\ell \otimes Y_\ell)] + \rho^2 \mathbb{E}[H_\ell \otimes H_\ell].
\end{aligned}$$

Taking the sum over all rows $\ell \in [n]$ yields

$$\mathbb{E}[(\boldsymbol{Z} - \rho \boldsymbol{M})^T (\boldsymbol{Z} - \rho \boldsymbol{M})] = \rho(1-\rho)\text{diag}(\mathbb{E}[\boldsymbol{Y}^T \boldsymbol{Y}]) + \rho^2 \mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}]. \tag{5.7}$$

To complete the proof, we apply triangle inequality to (5.7) to obtain

$$\Big\|\mathbb{E}[(\boldsymbol{Z} - \rho \boldsymbol{M})^T (\boldsymbol{Z} - \rho \boldsymbol{M})]\Big\| \leq \rho(1-\rho)\Big\|\text{diag}(\mathbb{E}[\boldsymbol{Y}^T \boldsymbol{Y}])\Big\| + \rho^2 \Big\|\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}]\Big\|.$$

Since $\boldsymbol{H}$ is zero mean, we have

$$\begin{aligned}
\Big\|\text{diag}(\mathbb{E}[\boldsymbol{Y}^T \boldsymbol{Y}])\Big\| &= \Big\|\text{diag}(\boldsymbol{M}^T \boldsymbol{M}) + \text{diag}(\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}])\Big\| \\
&\leq \Big\|\text{diag}(\boldsymbol{M}^T \boldsymbol{M})\Big\| + \Big\|\text{diag}(\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}])\Big\|.
\end{aligned}$$

Collecting terms completes the proof. ■

Lemma 5.8.4. *Assume Properties 4.1.2, 4.1.5, 4.1.6 hold. Then for any $s \geq 0$, the following holds w.p. at least $1 - 2\exp(-s^2)$:*

$$\|\boldsymbol{Z} - \rho \boldsymbol{M}\| \leq \sqrt{C'}\left(\sqrt{n} + \sqrt{p} + s\right),$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$.

Proof. Let $\tilde{\boldsymbol{H}} = \boldsymbol{Z} - \rho \boldsymbol{M}$ with second moment matrix $\boldsymbol{\Sigma} = (1/n)\mathbb{E}[\tilde{\boldsymbol{H}}^T \tilde{\boldsymbol{H}}]$. Applying triangle inequality gives

$$\frac{1}{n}\left\|\tilde{\boldsymbol{H}}\right\|^2 = \left\|\frac{1}{n}\tilde{\boldsymbol{H}}^T\tilde{\boldsymbol{H}}\right\| \leq \|\boldsymbol{\Sigma}\| + \left\|\frac{1}{n}\tilde{\boldsymbol{H}}^T\tilde{\boldsymbol{H}} - \boldsymbol{\Sigma}\right\|.$$

By Lemma 5.8.2, we establish that the rows of $\tilde{\boldsymbol{H}}$ are sub-gaussian with

$$\left\|\tilde{H}_i\right\|_{\psi_2} \leq CK,$$

which are also independent by assumption; hence, we can apply Lemma 3.2.5 to obtain

$$\left\|\frac{1}{n}\tilde{\boldsymbol{H}}^T\tilde{\boldsymbol{H}} - \boldsymbol{\Sigma}\right\| \leq CK^2\max(\delta,\delta^2), \quad \text{where } \delta = C\sqrt{\frac{p}{n}} + \frac{s}{\sqrt{n}}$$

with probability at least $1 - \exp(-s^2)$. Next, we apply Lemma 5.8.3, which gives

$$\begin{aligned}\|\boldsymbol{\Sigma}\| &\leq \frac{1}{n}\left(\rho(1-\rho)\left(\left\|\text{diag}(\boldsymbol{M}^T\boldsymbol{M})\right\| + \left\|\text{diag}(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\right\|\right) + \rho^2\left\|\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]\right\|\right)\\ &\leq \rho(1-\rho)(1+\sigma^2) + \rho^2\gamma^2.\end{aligned}$$

Let $C' = C(1+\gamma^2)(1+K^2)(1+\sigma^2)$. Combining the above results yields,

$$\begin{aligned}\frac{1}{n}\left\|\tilde{\boldsymbol{H}}\right\|^2 &\leq \rho(1-\rho)(1+\sigma^2) + \rho^2\gamma^2 + CK^2\max(\delta,\delta^2)\\ &\leq C'(1+\max(\delta,\delta^2))\\ &\leq C'(1+\delta)^2\\ &\leq C'\left(1 + \frac{p}{n} + \frac{s^2}{n}\right).\end{aligned}$$

Putting everything together, we conclude

$$\|\boldsymbol{Z} - \rho \boldsymbol{M}\| \leq \sqrt{C'}\left(\sqrt{n} + \sqrt{p} + s\right).$$

■

■ 5.8.2 $\ell_{2,\infty}$-norm

To prove Lemma 5.8.6, we will establish that the columns of $\boldsymbol{Z}-\rho\boldsymbol{M}$ are also sub-gaussian.

Lemma 5.8.5. *Assume Properties 4.1.2, 4.1.5, 4.1.6 hold. Then for every $j \in [p]$,*

$$\left\|Z_j - \rho M_j\right\|_{\psi_2} \leq CK.$$

Proof. Let e_j denote the j-th canonical vector. Observe that

$$\begin{aligned}
\left\|Z_j - \rho M_j\right\|_{\psi_2}^2 &= \sup_{u \in \mathbb{S}^{n-1}} \left\|u^T \left(Z_j - \rho M_j\right)\right\|_{\psi_2}^2 \\
&= \sup_{u \in \mathbb{S}^{n-1}} \left\|u^T \left(\boldsymbol{Z} - \rho\boldsymbol{M}\right) e_j\right\|_{\psi_2}^2 \\
&= \sup_{u \in \mathbb{S}^{n-1}} \left\|\sum_{i=1}^n u_i (Z_i - \rho M_i) e_j\right\|_{\psi_2}^2 \\
&\overset{(a)}{\leq} C \sup_{u \in \mathbb{S}^{n-1}} \sum_{i=1}^n u_i^2 \left\|(Z_i - \rho M_i) e_j\right\|_{\psi_2}^2 \\
&\leq C \max_{i \in [n]} \left\|Z_i - \rho M_i\right\|_{\psi_2}^2,
\end{aligned}$$

where (a) follows from Property 3.2.2. The conclusion then follows from Lemma 5.8.2. ■

Lemma 5.8.6. *Assume Properties 4.1.2, 4.1.5, 4.1.6 hold. For any $s \geq 0$, the following inequality holds w.p. at least $1 - p \cdot \exp\left[-c \min\left(\frac{s^2}{K^4 n}, \frac{s}{K^2}\right)\right]$:*

$$\max_{j \in [p]} \left\|Z_j - \rho M_j\right\|_2^2 \leq n(\sigma^2 \rho + \rho(1-\rho)) + s.$$

Proof. By Lemma 5.8.5, we have that the columns of $\boldsymbol{Z} - \rho\boldsymbol{M}$ are sub-gaussian random vectors in $\mathbb{R}^n$ satisfying

$$\left\|Z_j - \rho M_j\right\|_{\psi_2} \leq CK$$

for all $j \in [p]$. Further, since the rows of $\boldsymbol{Z} - \rho\boldsymbol{M}$ are assumed to be independent, it follows that for every column $j \in [p]$, the coordinates of $Z_j - \rho M_j$ are independent, mean zero, sub-gaussian random variables.

To that end, let us fix $j \in [p]$ and define $X = Z_j - \rho M_j \in \mathbb{R}^n$ where $X_i = Z_{ij} - \rho M_{ij}$.

Observe that

$$\|X\|_2^2 - \mathbb{E}\|X\|_2^2 = \sum_{i=1}^{n} (X_i^2 - \mathbb{E}[X_i^2])$$

is a sum of independent, mean zero, sub-exponential random variables with

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

Moreover, observe that

$$\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) = \mathrm{Var}(Z_{ij}) \leq \sigma^2 \rho + \rho(1-\rho).$$

As a result, using Bernstein's inequality (Theorem 3.2.1), we have that

$$\|Z_j - \rho M_j\|_2^2 \leq n(\sigma^2\rho + \rho(1-\rho)) + s$$

with probability at least $1 - \exp\left[-c \min\left(\frac{s^2}{K^4 n}, \frac{s}{K^2}\right)\right]$.

We now unfix j by applying a union bound. Thus, for any $s \geq 0$

$$\mathbb{P}\left(\max_{j\in[p]} \|Z_j - \rho M_j\|_2^2 \geq n\tilde{\sigma}^2 + s\right) \leq p \cdot \exp\left[-c \min\left(\frac{s^2}{K^4 n}, \frac{s}{K^2}\right)\right],$$

where $\tilde{\sigma}^2 = \sigma^2\rho + \rho(1-\rho)$. This completes the proof. ■

■ 5.9 Proofs: HSVT Estimation Error

In order to establish Lemma 5.2.1, we state its deterministic counterpart in Lemma 5.9.6, which expresses the estimation error in terms of the operator and $\ell_{2,\infty}$-norms of $\boldsymbol{Z} - \rho\boldsymbol{M}$. Lemmas 5.8.4 and 5.8.6 of Appendix 5.8 are then utilized to analyze our particular setting of interest, i.e., when Properties 4.1.2, 4.1.5, and 4.1.6 hold. The remainder of the subsection is dedicated to proving the helper lemmas of these results. We begin, however, with notation and a useful observation of the HSVT operation.

Notation. Throughout this section, let

$$v_1 = \|\boldsymbol{Z} - \rho\boldsymbol{M}\|. \tag{5.8}$$

Further, for any $n \times p$ matrix $\boldsymbol{Q}$, let $Q_j \in \mathbb{R}^p$ denote the j-th column of $\boldsymbol{Q}$.

■ 5.9.1 A Column Representation for the HSVT Operator

Consider a matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times p}$ with the following SVD

$$\boldsymbol{Q} = \sum_{i=1}^{n \wedge p} \sigma_i u_i \otimes v_i = \boldsymbol{U \Sigma V}^T.$$

We say $\widehat{\boldsymbol{Q}} = \text{HSVT}(\boldsymbol{Q}, k)$ if

$$\widehat{\boldsymbol{Q}} = \sum_{i=1}^{r} \sigma_i u_i \otimes v_i = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^T;$$

here, $\boldsymbol{U}_r \in \mathbb{R}^{m \times k}$ and $\boldsymbol{V}_r \in \mathbb{R}^{n \times k}$ denote the matrices consisting of the top k left and right singular vectors of $\boldsymbol{Q}$, respectively, and $\boldsymbol{S}_r = \text{diag}(s_1, \dots, s_r) \in \mathbb{R}^{k \times k}$. We now show how $\mathcal{P}_{U_r} \in \mathbb{R}^{m \times m}$ relates to the HSVT operation that retains the top k singular components.

Lemma 5.9.1. *Let* $\widehat{\boldsymbol{Q}} = \text{HSVT}(\boldsymbol{Q}, k)$*. Then for any* $j \in [p]$*,*

$$\mathcal{P}_{U_r} Q_j = \widehat{Q}_j.$$

Proof. Let $e_j \in \mathbb{R}^p$ denote the canonical basis vector in $\mathbb{R}^p$. Then using the orthonormality property of $\boldsymbol{U}$, it follows that

$$\begin{aligned}
\mathcal{P}_{U_r} Q_j &= \sum_{i=1}^{r} u_i u_i^T Q_j \\
&= \sum_{i=1}^{r} u_i u_i^T \left(\sum_{\ell=1}^{n \wedge p} \sigma_\ell u_\ell v_\ell^T \right) e_j \\
&= \left(\sum_{i=1}^{r} \sum_{\ell=1}^{n \wedge p} \sigma_\ell u_i u_i^T u_\ell v_\ell^T \right) e_j \\
&= \left(\sum_{i=1}^{r} \sigma_i u_i v_i^T \right) e_j = \widehat{Q}_j.
\end{aligned}$$

This completes the proof. ■

Remark 5.9.1. *Suppose we have randomly missing data. By Lemma 5.9.1, and linearity* *of the projection operator, we note that*

$$\widehat{M}_j = (1/\widehat{\rho})\ \mathcal{P}_{\widehat{U}_M}(Z_j). \tag{5.9}$$

■ 5.9.2 High Probability Bounds on Noise Deviation

High-Probability Events. We define the following events: for any $\delta_1, \delta_2, \delta_3, \delta_4 > 0$,

$$\begin{aligned}
E_1 &= \left\{ \nu_1 \le C\sqrt{(1+\sigma^2)(1+\gamma^2)(1+K^2)} \left(\sqrt{n} + \sqrt{p} + \sqrt{\log(1/\delta_1)}\right) \right\} \\
E_2 &= \left\{ \left(1 - \sqrt{\frac{C\log(1/\delta_2)}{np\rho}}\right)\rho \le \widehat{\rho} \le \frac{1}{1-\sqrt{\frac{C_2\log(1/\delta_2)}{np\rho}}}\rho \right\} \\
E_3 &= \left\{ \max_{j\in[p]} \left\| Z_j - \rho M_j \right\|_2^2 \le n(\sigma^2\rho + \rho(1-\rho)) + CK^2\sqrt{n\log(p/\delta_3)} \right\} \\
E_4 &= \left\{ \max_{j\in[p]} \left\| \mathcal{P}_{U_M}(Z_j - \rho M_j) \right\|_2^2 \le r(\sigma^2\rho + \rho(1-\rho)) + CK^2\sqrt{r\log(p/\delta_4)} \right\}.
\end{aligned}$$

Finally, we denote

$$E = E_1 \cap E_2 \cap E_3 \cap E_4. \tag{5.10}$$

E_1 occurs with high probability.

Lemma 5.9.2. *Assume Properties 4.1.2, 4.1.5, and 4.1.6 hold. Then for any $\delta_1 > 0$, it follows that $\mathbb{P}(E_1^c) \le \delta_1$.*

Proof. The proof follows immediately from Lemma 5.8.4 for any $s \ge C\sqrt{\log(1/\delta_1)}$. ■

E_2 occurs with high probability.

Lemma 5.9.3. *Assume Property 4.1.6 holds. Then for any $\delta_2 > 0$, it follows that $\mathbb{P}(E_2^c) \le \delta_2$.*

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

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

By the union bound,

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

Noticing $\alpha + 1 < 2\alpha < 2\alpha^2$ for all $\alpha > 1$, and setting the above probability to be at least $1 - \delta_2$ and solving for δ_2 completes the proof. ■

E_3 and E_4 occur with high probability.

Lemma 5.9.4. *Assume Properties 4.1.2, 4.1.5, and 4.1.6 hold. Then for any $\delta_3 > 0$, it follows that $\mathbb{P}(E_3^c) \le \delta_3$.*

Proof. The proof follows immediately from Lemma 5.8.6 for any $s \ge CK^2\sqrt{n\log(p/\delta_3)}$. ■

Lemma 5.9.5. *Assume Properties 4.1.2, 4.1.5, and 4.1.6 hold. Then for any $\delta_4 > 0$, it follows that $\mathbb{P}(E_4^c) \le \delta_4$.*

Proof. Using the arguments made in the proof of Lemma 5.8.6, we see that the columns of $\boldsymbol{Z} - \rho\boldsymbol{M}$ are sub-gaussian random vectors with independent, mean zero, sub-gaussian coordinates. Additionally, its sub-gaussian norms are bounded by CK.

Now, let us fix a column $j \in [p]$, and let $X = Z_j - \rho M_j$ such that $X_t = Z_{tj} - \rho M_{tj}$. Then, we can express

$$\left\|\mathcal{P}_{U_M}(Z_j - \rho M_j)\right\|_2^2 = \left\|\mathcal{P}_{U_M} X\right\|_2^2.$$

By Hanson-Wright's inequality (Theorem 3.2.4), we obtain

$$\left\|\mathcal{P}_{U_M} X\right\|_2^2 \le \mathbb{E}\left\|\mathcal{P}_{U_M} X\right\|_2^2 + s$$

with probability at least $1 - \exp\left[-c\min\left(\frac{s^2}{K^4 r}, \frac{s}{K^2}\right)\right]$. Note that we have made use of the following facts: $\left\|\mathcal{P}_{U_M}\right\| = 1$ and $\left\|\mathcal{P}_{U_M}\right\|_F^2 = r$. To bound the expected value, observe that

$$\mathbb{E}\left\|\mathcal{P}_{U_M} X\right\|_2^2 = \sum_{i=1}^{r} \mathbb{E}[\langle X, u_i\rangle^2] = \sum_{i=1}^{r} \mathrm{Var}(\langle X, u_i\rangle),$$

where u_i denotes the i-th column of $\boldsymbol{U}_M$ (the i-th left singular vector of $\boldsymbol{M}$). By the independence of the entries of X and the orthonormality of $\boldsymbol{U}_M$,

$$\mathrm{Var}(\langle X, u_i\rangle) = \sum_{t=1}^{n} u_{it}^2 \mathrm{Var}(X_t) = \sum_{t=1}^{n} u_{it}^2 \mathrm{Var}(Z_{tj}) \le \sigma^2\rho + \rho(1-\rho).$$

We now unfix j by applying a union bound, which yields for any $s \ge 0$,

$$\mathbb{P}\left(\max_{j\in[p]} \left\|\mathcal{P}_{U_M}(Z_j - \rho M_j)\right\|_2^2 \ge r\tilde{\sigma}^2 + s\right) \le p \cdot \exp\left[-c\min\left(\frac{s^2}{K^4 r}, \frac{s}{K^2}\right)\right],$$

where $\tilde{\sigma}^2 = \sigma^2\rho + \rho(1-\rho)$. Setting the above probability to be less than δ_4 and solving for s, we establish that the probability is bounded above by δ_4 if and only if $s \geq CK^2\sqrt{r\log(p/\delta_4)}$. ∎

∎ 5.9.3 Proof of Lemma 5.2.1

We begin by stating the key lemma used to prove Lemma 5.2.1.

Lemma 5.9.6. *Suppose $\rho/\alpha \leq \widehat{\rho} \leq \alpha\rho$ for some $\alpha \geq 1$. Then,*

$$\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \leq \frac{4\alpha^2 v_1^2}{\rho^4 n s_r^2}\left(\|\boldsymbol{Z}-\rho\boldsymbol{M}\|_{2,\infty}^2 + \|\boldsymbol{M}\|_{2,\infty}^2\right) + \frac{4\alpha^2}{\rho^2 n}\|\mathcal{P}_{U_M}(\boldsymbol{Z}-\rho\boldsymbol{M})\|_{2,\infty}^2 + \frac{2(\alpha-1)^2}{n}\|\boldsymbol{M}\|_{2,\infty}^2.$$

Proof. We prove our key lemma in three steps.

Step 1. Fix a column index $j \in [p]$. Observe that

$$\widehat{M}_j - M_j = \left(\widehat{M}_j - \mathcal{P}_{\widehat{U}_M}M_j\right) + \left(\mathcal{P}_{\widehat{U}_M}M_j - M_j\right).$$

Since $\text{rank}(\widehat{\boldsymbol{M}}) = r$, it follows that $\mathcal{P}_{\widehat{U}_M}$ is an orthogonal projection operator onto the span of the top r left singular vectors of $\boldsymbol{Z}$, namely, span $\{\widehat{u}_1, \dots, \widehat{u}_r\}$. Therefore,

$$\mathcal{P}_{\widehat{U}_M}M_j - M_j \in \text{span}\{\widehat{u}_1, \dots, \widehat{u}_r\}^{\perp}.$$

Additionally, by (5.9), we have that

$$\widehat{M}_j - \mathcal{P}_{\widehat{U}_M}M_j = \frac{1}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}Z_j - \mathcal{P}_{\widehat{U}_M}M_j \in \text{span}\{\widehat{u}_1, \dots, \widehat{u}_r\}.$$

Hence, $\langle \widehat{M}_j - \mathcal{P}_{\widehat{U}_M}M_j, \mathcal{P}_{\widehat{U}_M}M_j - M_j\rangle = 0$, and

$$\left\|\widehat{M}_j - M_j\right\|_2^2 = \left\|\widehat{M}_j - \mathcal{P}_{\widehat{U}_M}M_j\right\|_2^2 + \left\|\mathcal{P}_{\widehat{U}_M}M_j - M_j\right\|_2^2 \tag{5.11}$$

by the Pythagorean theorem. It remains to bound the terms on the right hand side of (5.11).

Step 2. We begin by bounding the first term on the right hand side of (5.11). Again, applying Lemma 5.9.1, we can rewrite

$$\widehat{M}_j - \mathcal{P}_{\widehat{U}_M}M_j = \frac{1}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}Z_j - \mathcal{P}_{\widehat{U}_M}M_j = \mathcal{P}_{\widehat{U}_M}\left((1/\widehat{\rho})\,Z_j - M_j\right)$$

$$= \frac{1}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}(M_j).$$

Using the Parallelogram Law (or, equivalently, combining Cauchy-Schwartz and AM-GM inequalities), we obtain

$$\begin{aligned}\left\|\widehat{M}_j - \mathcal{P}_{\widehat{U}_M}M_j\right\|_2^2 &= \left\|\frac{1}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}(M_j)\right\|_2^2 \\ &\le 2\left\|\frac{1}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j)\right\|_2^2 + 2\left\|\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\mathcal{P}_{\widehat{U}_M}(M_j)\right\|_2^2 \\ &\le \frac{2}{\widehat{\rho}^2}\left\|\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j)\right\|_2^2 + 2\left(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\right)^2 \|M_j\|_2^2 \\ &\le \frac{2\alpha^2}{\rho^2}\left\|\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j)\right\|_2^2 + 2(\alpha - 1)^2\|M_j\|_2^2. \end{aligned} \tag{5.12}$$

To arrive at the above inequality, note that Condition 2 implies $1/\widehat{\rho} \le \alpha/\rho$ and $(\rho - \widehat{\rho})/\widehat{\rho}^2 \le (\alpha - 1)^2$. Further, using the Parallelogram Law, observe that the first term of (5.12) can be decomposed as

$$\left\|\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j)\right\|_2^2 \le 2\left\|\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j) - \mathcal{P}_{U_M}(Z_j - \rho M_j)\right\|_2^2 + 2\left\|\mathcal{P}_{U_M}(Z_j - \rho M_j)\right\|_2^2. \tag{5.13}$$

We now bound the first term on the right hand side of (5.13) separately. First, we apply Theorem 3.1.1 to arrive at the following inequality:

$$\left\|\sin\Theta(\widehat{\boldsymbol{U}}_M, \boldsymbol{U}_M)\right\| \le \frac{2\|\boldsymbol{Z} - \rho\boldsymbol{M}\|}{\rho s_r} = \frac{2\nu_1}{\rho s_r}, \tag{5.14}$$

where $\widehat{\boldsymbol{U}}_M$ and $\boldsymbol{U}_M$ denote the top r left singular vectors of $\boldsymbol{Z}$ and $\boldsymbol{M}$, respectively. Then, it follows that

$$\begin{aligned}\left\|\mathcal{P}_{\widehat{U}_M}(Z_j - \rho M_j) - \mathcal{P}_{U_M}(Z_j - \rho M_j)\right\|_2^2 &\le \left\|\sin\Theta(\widehat{\boldsymbol{U}}_M, \boldsymbol{U}_M)\right\|^2 \cdot \|Z_j - \rho M_j\|_2^2 \\ &\le \frac{2\nu_1^2}{\rho^2 s_r^2}\|Z_j - \rho M_j\|_2^2.\end{aligned}$$

Combining the inequalities together, we have

$$\left\|\widehat{M}_j - \mathcal{P}_{\widehat{U}_M} M_j\right\|_2^2 \leq \frac{4\alpha^2 \nu_1^2}{\rho^4 s_r^2} \|Z_j - \rho M_j\|_2^2 + \frac{4\alpha^2}{\rho^2} \|\mathcal{P}_{U_M}(Z_j - \rho M_j)\|_2^2 + 2(\alpha-1)^2 \|M_j\|_2^2. \tag{5.15}$$

Step 3. We now bound the second term of (5.11). Using (5.23), we obtain

$$\begin{aligned} \left\|\mathcal{P}_{\widehat{U}_M} M_j - M_j\right\|_2^2 &= \left\|\mathcal{P}_{\widehat{U}_M} M_j - \mathcal{P}_{U_M} M_j\right\|_2^2 \\ &\leq \left\|\sin \Theta(\widehat{U}_M, U_M)\right\|^2 \cdot \|M_j\|_2^2 \\ &\leq \frac{2\nu_1^2}{\rho^2 s_r^2} \|M_j\|_2^2. \end{aligned} \tag{5.16}$$

Inserting (5.15) and (5.16) back into (5.11), and observing that this inequality holds for every column $j \in [p]$ completes the proof. ■

Completing Proof of Lemma 5.2.1

Proof. In Lemmas 5.9.2, 5.9.3, 5.9.4, and 5.9.5, set $\delta_i = \delta/4$ for any $\delta > 0$. Then,

$$\mathbb{P}(E^c) \leq \sum_{i=1}^{4} \mathbb{P}(E_i^c) \leq \delta, \tag{5.17}$$

where E is given by (5.10).

Let $C' = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$. We then have the following bounds:

$$\begin{aligned} \nu_1^2 &\leq C'\left(n + p + \log(1/\delta)\right) \\ \frac{\nu_1^2}{s_r^2} &\leq C'\left(\frac{r}{n \wedge p} + \frac{k \log(1/\delta)}{np}\right) \\ \frac{1}{n}\|\boldsymbol{Z} - \rho \boldsymbol{M}\|_{2,\infty}^2 &\leq C(1+\sigma^2) + CK^2 \sqrt{\frac{\log(p/\delta)}{n}} \\ \frac{1}{n}\|\mathcal{P}_{U_M}(\boldsymbol{Z} - \rho \boldsymbol{M})\|_{2,\infty}^2 &\leq C(1+\sigma^2)\frac{r}{n} + CK^2 \frac{\sqrt{r \log(p/\delta)}}{n} \\ \frac{1}{n}\|\boldsymbol{M}\|_{2,\infty}^2 &\leq 1. \end{aligned}$$

Further since, $\rho \geq \frac{C\log(1/\delta)}{np}$, for sufficiently large absolute constant C, we have that

$$\begin{aligned} \alpha &= \left(1 - \sqrt{\frac{C\log(1/\delta)}{np}}\right)^{-1} \leq C \\ (\alpha - 1)^2 &\leq \frac{C\log(1/\delta)}{np}. \end{aligned}$$

Collecting and simplifying the above bounds, we apply Lemma 5.9.6 to obtain

$$\begin{aligned} \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) &\leq \frac{C'}{\rho^4}\left(\frac{r}{n \wedge p} + \frac{r\log(1/\delta)}{np}\right)\left((1+\sigma^2) + K^2\sqrt{\frac{\log(p/\delta)}{n}}\right) \\ &\quad + \frac{C}{\rho^2}\left(\frac{(1+\sigma^2)r}{n} + \frac{K^2\sqrt{k\log(p/\delta)}}{n}\right) + 2(\alpha-1)^2 \\ &\leq \frac{C_1}{\rho^4}\frac{r}{n \wedge p} + \Delta, \end{aligned}$$

where

$$\Delta = \frac{C_1K^2}{\rho^4}\left(\frac{r}{n \wedge p}\sqrt{\frac{\log(p/\delta)}{n}} + \frac{r\log(1/\delta)}{np} + \frac{r\log(1/\delta)\sqrt{\log(p/\delta)}}{np^{3/2}}\right) + \frac{CK^2\sqrt{r\log(p/\delta)}}{\rho^2 n}$$

and $C_1 = C(1+\sigma^4)(1+\gamma^2)(1+K^2)$. Letting $C_2 = C_1K^2(1+\log^{3/2}(1/\delta))$, we bound Δ as follows:

$$\Delta \leq C_2\left(\frac{1}{\rho^4}\frac{r}{\sqrt{n}(n \wedge p)} + \frac{1}{\rho^2}\frac{\sqrt{r}}{n}\right)\sqrt{\log n}.$$

Relabeling the above bound as Δ completes the proof. ■

■ 5.9.4 Corollaries: Bounds in Expectation

Corollary 5.9.1. *Suppose the conditions of Lemma 5.2.1 hold. Then for any $\delta > 0$,*

$$\mathbb{E}[\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \mid E] \leq \frac{C_1'}{\rho^4}\frac{r}{n \wedge p} + \Gamma,$$

where and

$$\Gamma = C_2'\left(\frac{r\log(1/\delta)}{\rho^4\sqrt{n}(n \wedge p)} + \frac{\sqrt{r}}{\rho^2 n}\right)\sqrt{\log(p/\delta)},$$

$C_1' = C(1+\sigma^4)(1+\gamma^2)(1+K^2)$, *and* $C_2' = C_1' K^2$.

Proof. The proof follows that of Lemma 5.2.1 under the event E, which is given by (5.10). ■

■ 5.10 Proofs: Training Prediction Error

Throughout, we will make use of Proposition 5.3.1, i.e., $\widehat{M}_{\text{train}} = \boldsymbol{Z}\widehat{\beta} = \widehat{\boldsymbol{M}}\widehat{\beta}$. This allows us to analyze $\widehat{M}_{\text{train}}$ through the lens of the HSVT estimator.

■ 5.10.1 Proof of Lemma 5.3.1

Lemma 5.10.1. *Suppose Property 4.1.4 holds. Consider* $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r)$. *Then,*

$$\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) \leq \frac{2}{n}\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)\rangle + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \left\|\beta^*\right\|_1^2.$$

Proof. By (5.1), we have that

$$\left\|\widehat{\boldsymbol{M}}\widehat{\beta} - y\right\|_2^2 = \left\|\widehat{\boldsymbol{M}}\widehat{\beta} - \boldsymbol{M}\beta^*\right\|_2^2 + \|\varepsilon\|_2^2 - 2\langle \varepsilon, (\widehat{\boldsymbol{M}}\widehat{\beta} - \boldsymbol{M}\beta^*)\rangle. \tag{5.18}$$

On the other hand, the optimality of $\widehat{\beta}$ yields

$$\left\|\widehat{\boldsymbol{M}}\widehat{\beta} - y\right\|_2^2 \leq \left\|\widehat{\boldsymbol{M}}\beta^* - y\right\|_2^2 = \left\|(\widehat{\boldsymbol{M}} - \boldsymbol{M})\beta^*\right\|_2^2 + \|\varepsilon\|_2^2 - 2\langle \varepsilon, (\widehat{\boldsymbol{M}} - \boldsymbol{M})\beta^*\rangle. \tag{5.19}$$

Combining (5.18) and (5.19), we have

$$\left\|\widehat{\boldsymbol{M}}\widehat{\beta} - \boldsymbol{M}\beta^*\right\|_2^2 \leq \left\|(\widehat{\boldsymbol{M}} - \boldsymbol{M})\beta^*\right\|_2^2 + 2\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)\rangle.$$

We now apply (generalized) Hölder's inequality with $q_1 = 1$ and $q_2 = \infty$ to obtain

$$\left\|(\widehat{\boldsymbol{M}} - \boldsymbol{M})\beta^*\right\|_2^2 \leq \left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_{2,\infty}^2 \cdot \left\|\beta^*\right\|_1^2.$$

Normalizing by n gives the desired result. ■

Lemma 5.10.2. *Suppose Property 4.1.4 holds. Consider* $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r)$. *Then for any* $\delta > 0$, *the following holds w.p. at least* $1 - \delta$:

$$\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)\rangle \leq \sigma^2 r + CK\left(K\sqrt{r} + \sqrt{n}\left\|\beta^*\right\|_1 + \sqrt{n}\cdot \mathcal{E}_{\text{HSVT}}^{1/2}(\widehat{\boldsymbol{M}}) \left\|\beta^*\right\|_1\right)\log(1/\delta).$$

Proof. Let $\boldsymbol{Q} = \widehat{\boldsymbol{M}}\widehat{\boldsymbol{M}}^{\dagger} \in \mathbb{R}^{n \times n}$. Since $\boldsymbol{Q}$ is an orthogonal projection operator, it follows that $\left\|\boldsymbol{Q}\right\|_F^2 = r$, $\left\|\boldsymbol{Q}\right\| = 1$, and $\left\|\boldsymbol{Q}u\right\|_2 \leq \|u\|_2$ for any $u \in \mathbb{R}^n$. Now, observe that

$$\begin{aligned}\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)\rangle &= \left\langle \varepsilon, \widehat{\boldsymbol{M}}\widehat{\beta}\right\rangle - \langle \varepsilon, \widehat{\boldsymbol{M}}\beta^*\rangle \\ &= \langle \varepsilon, \boldsymbol{Q}\boldsymbol{M}\beta^*\rangle + \langle \varepsilon, \boldsymbol{Q}\varepsilon\rangle - \langle \varepsilon, \widehat{\boldsymbol{M}}\beta^*\rangle. \end{aligned} \tag{5.20}$$

It remains to bound each term independently. To begin, for any $s_1 \geq 0$, Lemma 3.2.4 gives

$$\mathbb{P}\left(\langle \varepsilon, \boldsymbol{Q}\varepsilon\rangle - \mathbb{E}[\langle \varepsilon, \boldsymbol{Q}\varepsilon\rangle] \geq s_1\right) \leq \exp\left[-c \min\left(\frac{s_1^2}{K^4 r}, \frac{s_1}{K^2}\right)\right].$$

Using the law of total expectations and the independence within the entries of ε, we bound the expectation as

$$\mathbb{E}[\langle \varepsilon, \boldsymbol{Q}\varepsilon\rangle] = \sum_{i,j=1}^{n} \mathbb{E}\left[\mathbb{E}[Q_{ij}\varepsilon_i\varepsilon_j \mid \boldsymbol{Q}]\right] \leq \sigma^2 \sum_{i=1}^{n} \mathbb{E}[Q_{ii}] = \sigma^2 \mathbb{E}[\mathrm{tr}(\boldsymbol{Q})] = \sigma^2 r.$$

Further, for any $s_2 \geq 0$, Lemma 3.2.2 gives

$$\mathbb{P}\left(\langle \varepsilon, \boldsymbol{Q}\boldsymbol{M}\beta^*\rangle \geq s_2\right) \leq \exp\left(-\frac{c s_2^2}{K^2 \left\|\boldsymbol{M}\beta^*\right\|_2^2}\right).$$

At the same time, if we let

$$v = \frac{\widehat{\boldsymbol{M}}\beta^*}{\left\|\widehat{\boldsymbol{M}}\beta^*\right\|_2},$$

then for any $s_3 \geq 0$, Lemma 3.2.2 yields

$$\mathbb{P}\left(-\langle \varepsilon, v\rangle \geq s_3\right) \leq \exp\left(-\frac{c s_3^2}{K^2}\right),$$

which implies that, with probability at least $1 - \exp\left(-c s_3^2/K^2\right)$,

$$-\langle \varepsilon, \widehat{\boldsymbol{M}}\beta^*\rangle \leq \left\|\widehat{\boldsymbol{M}}\beta^*\right\|_2 \cdot s_3.$$

By triangle inequality, it follows that

$$\left\|\widehat{\boldsymbol{M}}\beta^*\right\|_2 \leq \left\|(\widehat{\boldsymbol{M}} - \boldsymbol{M})\beta^*\right\|_2 + \left\|\boldsymbol{M}\beta^*\right\|_2 \leq \sqrt{n} \cdot \mathcal{E}_{\mathrm{HSVT}}^{1/2}(\widehat{\boldsymbol{M}}) \left\|\beta^*\right\|_1 + \left\|\boldsymbol{M}\beta^*\right\|_2.$$

Further, we have

$$\|M\beta^*\|_2 \le \|M\|_{2,\infty}\|\beta^*\|_1 \le \sqrt{n}\|\beta^*\|_1,$$

where the final inequality follows from Property 4.1.2. To complete the proof, we fix any $\delta > 0$ and set the above probabilities to be less than $\delta/3$ to solve for s_1, s_2, and s_3. ■

Completing Proof of Lemma 5.3.1

Proof. Let us fix some $\delta > 0$. We define the event

$$E_{\text{train}} = \left\{ \langle \varepsilon, \widehat{M}(\widehat{\beta} - \beta^*) \rangle \le \sigma^2 r + CK \left(K\sqrt{r} + \sqrt{n}\|\beta^*\|_1 + \sqrt{n} \cdot \mathcal{E}^{1/2}_{\text{HSVT}}(\widehat{M}) \, \|\beta^*\|_1 \right) \log(1/\delta) \right\},$$

which occurs with probability at least $1 - \delta$ by Lemma 5.10.2. Recall Lemma 5.10.1:

$$\mathcal{E}_{\text{train}}(\widehat{M}\widehat{\beta}) \le \frac{2}{n} \langle \varepsilon, \widehat{M}(\widehat{\beta} - \beta^*) \rangle + \mathcal{E}_{\text{HSVT}}(\widehat{M}) \, \|\beta^*\|_1^2.$$

We note the following simplification:

$$\begin{aligned}
&\mathcal{E}_{\text{HSVT}}(\widehat{M}) \, \|\beta^*\|_1^2 + \frac{1}{\sqrt{n}} \mathcal{E}^{1/2}_{\text{HSVT}}(\widehat{M}) \cdot \log(1/\delta) \|\beta^*\|_1 \\
&\quad \le \frac{C_1 r}{\rho^4 (n \wedge p)} \|\beta^*\|_1^2 + \frac{1}{\sqrt{n}} \sqrt{\frac{C_1 r}{\rho^4 (n \wedge p)}} \log(1/\delta) \|\beta^*\|_1 + \Delta \|\beta^*\|_1^2 + \sqrt{\frac{\Delta}{n}} \log(1/\delta) \|\beta^*\|_1 \\
&\quad \le \frac{C_3 r}{\rho^4 (n \wedge p)} \|\beta^*\|_1^2 + \Delta \|\beta^*\|_1^2 + \sqrt{\frac{\Delta}{n}} \log(1/\delta) \|\beta^*\|_1,
\end{aligned}$$

where C_1, Δ are given by (5.2), and $C_3 = C_1(1 + \log(1/\delta))$. Additionally,

$$\begin{aligned}
&\Delta \|\beta^*\|_1^2 + \sqrt{\frac{\Delta}{n}} \log(1/\delta) \|\beta^*\|_1 \\
&\le \left(\frac{C_2 r}{\rho^4 \sqrt{n} (n \wedge p)} + \frac{C_2 \sqrt{r}}{\rho^2 n} \right) \sqrt{\log p} \|\beta^*\|_1^2 + \left(\frac{\sqrt{C_2 r}}{\rho^2 n^{\frac{1}{4}} (n \wedge p)} + \frac{\sqrt{C_2} r^{\frac{1}{4}}}{\rho n} \right) \log^{\frac{1}{4}} n \cdot \log(1/\delta) \|\beta^*\|_1 \\
&\le C_4 \left(\frac{r}{\rho^4 n^{\frac{1}{4}} (n \wedge p)} + \frac{\sqrt{r}}{\rho^2 n} \right) \sqrt{\log p} \|\beta^*\|_1^2,
\end{aligned}$$

where C_2 is given by (5.2) and $C_4 = C_1 K^2 (1 + \log^{\frac{7}{4}}(1/\delta))$. Thus, under E_{train}, we use the above results to conclude

$$\mathcal{E}_{\text{train}}(\widehat{M}\widehat{\beta}) \le \frac{2\sigma^2 r}{n} + \frac{C_3 r}{\rho^4 (n \wedge p)} \|\beta^*\|_1^2 + \left(\frac{C_4 r \sqrt{\log p}}{\rho^4 n^{\frac{1}{4}} (n \wedge p)} + \frac{C_4 \sqrt{r \log p}}{\rho^2 n} \right) \|\beta^*\|_1^2 + \frac{C_4 \sqrt{r}}{n} \|\beta^*\|_1 + \frac{C_4}{\sqrt{n}} \|\beta^*\|_1$$

$$\begin{aligned}
&\leq \frac{2\sigma^2 r}{n} + \frac{C_3 r}{\rho^4 (n \wedge p)} \|\beta^*\|_1^2 + \left(\frac{C_4 r \sqrt{\log p}}{\rho^4 n^{\frac{1}{4}} (n \wedge p)} + \frac{C_4 \sqrt{r \log p}}{\rho^2 n} \right) \|\beta^*\|_1^2 + \frac{C_4}{\sqrt{n}} \|\beta^*\|_1 \\
&\leq \frac{2\sigma^2 r}{n} + \frac{C_4 r \sqrt{\log p}}{\rho^4 (n \wedge p)} \|\beta^*\|_1^2 + \frac{C_4}{\sqrt{n}} \|\beta^*\|_1.
\end{aligned}$$

Relabeling C_4 completes the proof. ■

■ 5.10.2 Corollaries: Bounds in Expectation

Lemma 5.10.3. *Suppose Property 4.1.4 holds. Consider* $\widehat{\boldsymbol{M}} = \mathrm{HSVT}(\boldsymbol{Z}, r)$. *Then,*

$$\mathbb{E}\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*) \rangle \leq \sigma^2 r.$$

Proof. Let $\boldsymbol{Q} = \widehat{\boldsymbol{M}}\widehat{\boldsymbol{M}}^{\dagger} \in \mathbb{R}^{n \times n}$. Using the arguments that led to (5.20) and linearity of expectations, we obtain

$$\mathbb{E}\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*) \rangle = \mathbb{E}\langle \varepsilon, \boldsymbol{Q}\boldsymbol{M}\beta^* \rangle + \mathbb{E}\langle \varepsilon, \boldsymbol{Q}\varepsilon \rangle - \mathbb{E}\langle \varepsilon, \widehat{\boldsymbol{M}}\beta^* \rangle.$$

Under the independence assumptions, we have the following equalities:

$$\begin{aligned}
\mathbb{E}\langle \varepsilon, \boldsymbol{Q}\boldsymbol{M}\beta^* \rangle &= 0 \\
\mathbb{E}\langle \varepsilon, \widehat{\boldsymbol{M}}\beta^* \rangle &= 0.
\end{aligned}$$

Using the cyclic and linearity properties of the trace operator, we further have

$$\begin{aligned}
\mathbb{E}\langle \varepsilon, \boldsymbol{Q}\varepsilon \rangle &= \mathbb{E}[\mathrm{tr}(\boldsymbol{Q}\varepsilon\varepsilon^T)] \\
&= \mathrm{tr}(\mathbb{E}[\boldsymbol{Q}] \cdot \mathbb{E}[\varepsilon\varepsilon^T]) \\
&\leq \sigma^2 \mathbb{E}[\mathrm{tr}(\boldsymbol{Q})] = \sigma^2 r.
\end{aligned}$$

This completes the proof. ■

Corollary 5.10.1. *Suppose the conditions of Lemma 5.3.1 hold. Then for any* $\delta > 0$,

$$\mathbb{E}[\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) \mid E] \leq \frac{2\sigma^2 r}{n} + \frac{C_1'}{\rho^4} \frac{r}{n \wedge p} \|\beta^*\|_1^2 + \Gamma_1,$$

where

$$\Gamma_1 = C_2' \left(\frac{r \log(1/\delta)}{\rho^4 \sqrt{n} (n \wedge p)} + \frac{\sqrt{r}}{\rho^2 n} \right) \sqrt{\log(p/\delta)}\, \|\beta^*\|_1^2, \tag{5.21}$$

$C_1' = C(1+\sigma^4)(1+\gamma^2)(1+K^2)$, *and* $C_2' = C_1' K^2$.

Proof. We follow the proof of Lemma 5.3.1. In particular, under the event E (given by (5.10)), we apply Lemma 5.10.1 to obtain

$$\mathbb{E}[\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) \mid E] \leq \frac{2}{n}\mathbb{E}[\langle \varepsilon, \widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)\rangle \mid E] + \mathbb{E}[\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \mid E] \cdot \left\|\beta^*\right\|_1^2.$$

We complete the proof by applying Lemmas 5.10.3 and 5.2.1. ■

■ 5.11 Proofs: Parameter Estimation

Proof Sketch. In order to provide a bound on the parameter estimation error (Theorem 5.3.1), we will first show that $\widehat{\boldsymbol{V}}_M$ is a good approximation to the latent feature space spanned by the columns of $\boldsymbol{V}_M$, provided that $\boldsymbol{Z}$ is thresholded appropriately (Lemma 5.11.1).

Next, we state Lemma 5.11.3, which bounds the error between $\widehat{\beta}$ and *any* β^* that is a solution to (5.1), when projected onto the subspace spanned by the columns of $\widehat{\boldsymbol{V}}_M$. Since our estimator $\widehat{\beta}$ lies within $\widehat{\boldsymbol{V}}_M$, which is shown to be close to $\boldsymbol{V}_M$ as per Lemma 5.11.1, it follows that $\widehat{\beta}$ is a good approximation of the component of β^* that lives within $\boldsymbol{V}_M$. This is formalized in Lemma 5.11.4. For a geometric picture of the proof sketch, see Figure 5.2.

Notation. Throughout, we will denote $\boldsymbol{U}_{M\perp} \in \mathbb{R}^{n \times (n-r)}$ and $\boldsymbol{V}_{M\perp} \in \mathbb{R}^{p \times (p-r)}$ as the orthogonal complements to $\boldsymbol{U}_M$ and $\boldsymbol{V}_M$. We continue to define $\nu_1 = \left\|\boldsymbol{Z} - \rho \boldsymbol{M}\right\|$, as in (5.8), and also define

$$\Lambda_M = \frac{2\nu_1}{\rho s_r}. \tag{5.22}$$

■ 5.11.1 Learning Subspaces

We first state Lemma 5.11.1, which bounds the misalignment between the subspaces spanned by $\widehat{\boldsymbol{V}}_M$ and $\boldsymbol{V}_M$.

Lemma 5.11.1. *Consider* $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r)$. *Then,*

$$\left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\| \leq \frac{2\nu_1}{\rho s_r}.$$

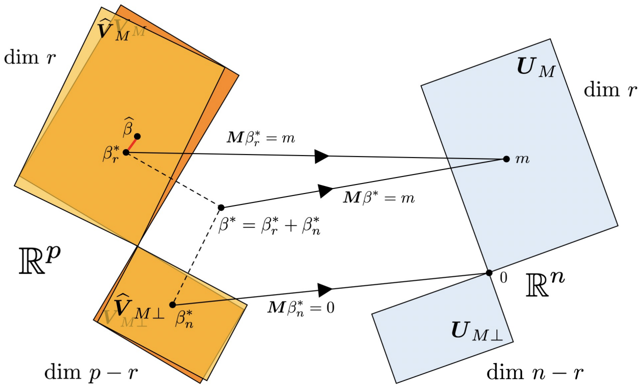

$\widehat{\boldsymbol{V}}_M$
$\boldsymbol{V}_M$
dim r
$\widehat{\beta}$
β_r^*
$\boldsymbol{M}\beta_r^* = m$
$\boldsymbol{U}_M$
dim r
m
$\boldsymbol{M}\beta^* = m$
$\beta^* = \beta_r^* + \beta_n^*$
$\mathbb{R}^p$
0
$\mathbb{R}^n$
$\widehat{\boldsymbol{V}}_{M\perp}$
$\boldsymbol{V}_{M\perp}$
β_n^*
$\boldsymbol{M}\beta_n^* = 0$
$\boldsymbol{U}_{M\perp}$
dim $p-r$
dim $n-r$

Figure 5.2: Interaction between the row and column space of $\boldsymbol{M}$ on any β^*, and the effect of misaligned subspaces between $\widehat{\boldsymbol{V}}_M$ and $\boldsymbol{V}$ on the gap between $\widehat{\beta}$ (which lives in $\widehat{\boldsymbol{V}}_M$) and $\mathcal{P}_{V_M}\beta^*$.

Proof. From Theorem 3.1.1 we have:

$$\left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\| \leq \frac{\left\|\boldsymbol{Z} - \rho\boldsymbol{M}\right\|}{\rho s_r} = \frac{2\nu_1}{\rho s_r}, \tag{5.23}$$

where $\widehat{\boldsymbol{V}}_M$ and $\boldsymbol{V}_M$ denote the top r right singular vectors of $\boldsymbol{Z}$ and $\boldsymbol{M}$, respectively. ■

■ 5.11.2 Bounding the Projected Parameter Estimation Error

Having shown that $\widehat{\boldsymbol{V}}_M$ is close to $\boldsymbol{V}_M$ in Lemma 5.11.1, we now bound the gap between $\widehat{\beta}$ and any β^* that satisfies (5.1) in the subspace spanned by $\widehat{\boldsymbol{V}}_M$; this is formalized in Lemma 5.11.3.

Lemma 5.11.2. *Recall that $\widehat{s}_i$ denotes the i-th singular value of $\boldsymbol{Z}$. Then,*

$$\rho s_i - \nu_1 \leq \widehat{s}_i \leq \rho s_i + \nu_1.$$

Proof. Observing that $\boldsymbol{Z} = (\boldsymbol{Z} - \rho\boldsymbol{M}) + \rho\boldsymbol{M}$ and applying Weyl's Inequality (Lemma 3.1.1) completes the proof. ■

Lemma 5.11.3. *Suppose Property 4.1.4 holds. Consider $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r)$. Then, for any β^* that is a solution to (5.1),*

$$\left\|\mathcal{P}_{\widehat{V}_M}(\widehat{\beta} - \beta^*)\right\|_2^2 \leq \frac{2\widehat{\rho}^2 n}{(\rho s_r - \nu_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\left\|\beta^*\right\|_1^2\right).$$

Proof. Recall that $\widehat{s}_i$ denotes the i-th singular value of $\boldsymbol{Z}$. To achieve our desired result, we will upper and lower bound the ℓ_2-norm of $\widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)$. To begin, observe that

$$\left\|\widehat{\boldsymbol{M}}(\widehat{\beta} - \beta^*)\right\|_2^2 \leq 2\left\|\widehat{\boldsymbol{M}}\widehat{\beta} - \boldsymbol{M}\beta^*\right\|_2^2 + 2\left\|(\widehat{\boldsymbol{M}} - \boldsymbol{M})\beta^*\right\|_2^2.$$

Now, recall that $\widehat{\rho}\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{U}}_M \widehat{\boldsymbol{S}}_M \widehat{\boldsymbol{V}}_M^T$. Letting $x = \widehat{\beta} - \beta^*$ and $y = \widehat{\boldsymbol{V}}_M^T x$, it follows that

$$\begin{aligned}
\widehat{\rho}^2 \left\|\widehat{\boldsymbol{M}}x\right\|_2^2 &= x^T \widehat{\boldsymbol{V}}_M \widehat{\boldsymbol{S}}_M \widehat{\boldsymbol{U}}_M^T \widehat{\boldsymbol{U}}_M \widehat{\boldsymbol{S}}_M \widehat{\boldsymbol{V}}_M^T x \\
&= x^T \widehat{\boldsymbol{V}}_M \widehat{\boldsymbol{S}}_M^2 \widehat{\boldsymbol{V}}_M^T x. \\
&= \sum_{i=1}^r \widehat{s}_i^2 y_i^2 \geq \widehat{s}_r^2 \sum_{i=1}^r y_i^2 = \widehat{s}_r^2 \cdot \left\|y\right\|_2^2.
\end{aligned}$$

Applying Lemma 5.11.2 and combining the above results completes the proof. ■

Below, we state Lemma 5.11.4, which provides a deterministic bound between $\widehat{\beta}$ and the unique β^* satisfying (5.1) with minimum ℓ_2-norm.

Lemma 5.11.4. *Suppose Property 4.1.4 holds. Consider* $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r)$ *and the unique* β^* *that satisfies* (5.1) *with minimum* ℓ_2*-norm. Then,*

$$\left\|\widehat{\beta} - \beta^*\right\|_2^2 \leq \frac{2\widehat{\rho}^2 n}{(\rho s_r - \nu_1)^2} \left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \left\|\beta^*\right\|_1^2 \right) + \Lambda_M^2 \left\|\beta^*\right\|_2^2, \qquad (5.24)$$

where Λ_M *is given by* (5.22).

Proof. To begin, observe that

$$\left\|\widehat{\beta} - \beta^*\right\|_2^2 = \left\|\mathcal{P}_{\widehat{V}_M}(\widehat{\beta} - \beta^*)\right\|_2^2 + \left\|\mathcal{P}_{\widehat{V}_{M\perp}}(\widehat{\beta} - \beta^*)\right\|_2^2.$$

To bound the first term of the above inequality, we can appeal to Lemma 5.11.3. Thus, it remains to bound the second expression.

Observing that $\widehat{\boldsymbol{V}}_{M\perp}^T \widehat{\beta} = 0$ yields

$$\left\|\mathcal{P}_{\widehat{V}_{M\perp}}(\widehat{\beta} - \beta^*)\right\|_2^2 = \left\|\mathcal{P}_{\widehat{V}_{M\perp}} \beta^*\right\|_2^2.$$

Let $\boldsymbol{V}_{M\perp} \in \mathbb{R}^{p \times (p-r)}$ denote the orthogonal complement of $\boldsymbol{V}_M$. Since $\boldsymbol{V}_{M\perp}^T \beta^* = 0$ by assumption, we apply Lemma 5.11.1 to obtain

$$\begin{aligned}\left\|\mathcal{P}_{\widehat{V}_{M\perp}}\beta^*\right\|_2^2 &= \left\|(\mathcal{P}_{\widehat{V}_{M\perp}} - \mathcal{P}_{V_{M\perp}})\beta^* + \mathcal{P}_{V_{M\perp}}\beta^*\right\|_2^2 \\ &= \left\|(\mathcal{P}_{\widehat{V}_{M\perp}} - \mathcal{P}_{V_{M\perp}})\beta^*\right\|_2^2 \\ &= \left\|(\mathcal{P}_{V_M} - \mathcal{P}_{\widehat{V}_M})\beta^*\right\|_2^2 \\ &\le \left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\|^2 \cdot \|\beta^*\|_2^2 \le \Lambda_M^2 \cdot \|\beta^*\|_2^2,\end{aligned}$$

where Λ_M is given by (5.24). Putting everything together and applying Lemma 5.11.3, we arrive at the following inequality:

$$\begin{aligned}\left\|\widehat{\beta} - \beta^*\right\|_2^2 &\le \left\|\mathcal{P}_{\widehat{V}_M}(\widehat{\beta} - \beta^*)\right\|_2^2 + \Lambda_M^2 \|\beta^*\|_2^2 \\ &\le \frac{2\widehat{\rho}^2 n}{(\rho s_r - \nu_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\|\beta^*\|_1^2\right) + \Lambda_M^2\|\beta^*\|_2^2. \qquad (5.25)\end{aligned}$$

This completes the proof. ■

5.11.3 Proof of Theorem 5.3.1

Proof. From Lemma 5.11.4, we have

$$\left\|\widehat{\beta} - \beta^*\right\|_2^2 \le \frac{2\widehat{\rho}^2 n}{(\rho s_r - \nu_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\|\beta^*\|_1^2\right) + \Lambda_M^2\|\beta^*\|_2^2.$$

Now, suppose E, given by (5.10), occurs. Then by Lemmas 5.2.1 and 5.3.1, we note that

$$\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}),\ \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \le \frac{2\sigma^2 r}{n} + \frac{C_2 r\sqrt{\log p}}{\rho^4(n \wedge p)}\|\beta^*\|_1^2 + \Delta_1, \qquad (5.26)$$

where C_2, Δ_1 are given by (5.3). Further, under E and Property 4.1.3,

$$\begin{aligned}\nu_1^2 &\le C_3(n + p + \log(1/\delta)) \\ \Lambda_M^2 &\le \frac{C_3}{\rho^2}\left(\frac{r}{n \wedge p} + \frac{r\log(1/\delta)}{np}\right),\end{aligned}$$

where $C_3 = C(1 + \sigma^2)(1 + \gamma^2)(1 + K^2)$. Our conditions on ρ additionally yield

$$\frac{2\widehat{\rho}^2 n}{(\rho s_r - \nu_1)^2} \le C\frac{r}{p}. \qquad (5.27)$$

Collecting terms and simplifying gives us the desired bound:

$$\left\|\widehat{\beta}-\beta^*\right\|_2^2 \leq \frac{r}{p}\left(\frac{C\sigma^2 r}{n}+\frac{C_2 r\sqrt{\log p}}{\rho^4(n\wedge p)}\|\beta^*\|_1^2+\Delta_1\right)+\frac{C_3 r}{\rho^2(n\wedge p)}\|\beta^*\|_2^2.$$

The proof is complete after relabeling constants. ■

■ 5.12 Proofs: Testing Prediction Error

Having shown that $\widehat{\boldsymbol{V}}_M$ is close to $\boldsymbol{V}_M$ (Lemma 5.11.1), and $\widehat{\beta}$ is close to the unique β^* that lives within $\boldsymbol{V}_M$ (Theorem 5.3.1), we are now ready to complete the proof for our post-intervention (test) counterfactual prediction error.

Notation. As before, we define $\nu_1 = \left\|\boldsymbol{Z}-\rho\boldsymbol{M}\right\|$ and $\Lambda_M = 2\nu_1/(\rho s_r)$, given by (5.8) and (5.22), respectively. Additionally, we define $\nu_1' = \left\|\boldsymbol{Z}'-\rho\boldsymbol{M}'\right\|$.

We also define the following events: for any $\delta > 0$,

$$\begin{aligned}
E_1' &= \left\{\nu_1' \leq C\sqrt{(1+\sigma^2)(1+\gamma^2)(1+K^2)}\left(\sqrt{m}+\sqrt{p}+\sqrt{\log(1/\delta)}\right)\right\}\\
E_2' &= \left\{\left(1-\sqrt{\frac{C\log(1/\delta)}{mp\rho}}\right)\rho \leq \widehat{\rho}' \leq \frac{1}{1-\sqrt{\frac{C\log(1/\delta)}{mp\rho}}}\rho\right\}\\
E_3' &= \left\{\max_{j\in[p]}\left\|Z_j'-\rho M_j'\right\|_2^2 \leq m(\sigma^2\rho+\rho(1-\rho))+CK^2\sqrt{m\log(p/\delta)}\right\}\\
E_4' &= \left\{\max_{j\in[p]}\left\|\mathcal{P}_{U_M}(Z_j'-\rho M_j')\right\|_2^2 \leq r'(\sigma^2\rho+\rho(1-\rho))+CK^2\sqrt{r'\log(p/\delta)}\right\}\\
E' &= E \cap E_1' \cap E_2' \cap E_3' \cap E_4',
\end{aligned} \tag{5.28}$$

where E is given by (5.10).

■ 5.12.1 Helper Lemmas

Lemma 5.12.1. *Let Property 4.1.4 hold. Consider $\widehat{\boldsymbol{M}} = \mathrm{HSVT}(\boldsymbol{Z}, r)$, $\widehat{\boldsymbol{M}'} = \mathrm{HSVT}(\boldsymbol{Z}', r')$, and the unique β^* that satisfies (5.1) with minimum ℓ_2-norm. Then,*

$$\begin{aligned}
\left\|\boldsymbol{M}'(\widehat{\beta}-\beta^*)\right\|_2^2 \leq 2(s_1')^2&\left(\frac{2(\widehat{\rho}')^2(1+\Lambda_M^2)n}{(\rho s_r-\nu_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta})+\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\left\|\beta^*\right\|_1^2\right)+\Lambda_M^4\|\beta^*\|_2^2\right)\\
&+2(s_1')^2\Lambda_M^2\cdot\left\|\boldsymbol{V}_{M'}^T\boldsymbol{V}_{M\perp}\right\|^2\cdot\left\|\widehat{\beta}\right\|_2^2.
\end{aligned}$$

Proof. Let $x = \widehat{\beta} - \beta^*$. Further, it is convenient to express $\boldsymbol{M}'$ as

$$\boldsymbol{M}' = \boldsymbol{M}'\mathcal{P}_{V_M} + \boldsymbol{M}'\mathcal{P}_{V_{M\perp}}.$$

This yields

$$\begin{aligned}\left\|\boldsymbol{M}'x\right\|_2^2 &= \left\|\boldsymbol{M}'\mathcal{P}_{V_M}x + \boldsymbol{M}'\mathcal{P}_{V_{M\perp}}x\right\|_2^2 \\ &= \left\|\boldsymbol{M}'\mathcal{P}_{V_M}x\right\|_2^2 + \left\|\boldsymbol{M}'\mathcal{P}_{V_{M\perp}}x\right\|_2^2. \qquad (5.29)\end{aligned}$$

Term 1. To bound the first term of (5.29), observe that

$$\begin{aligned}\left\|\boldsymbol{M}'\mathcal{P}_{V_M}x\right\|_2^2 &= \left\|\boldsymbol{M}'(\mathcal{P}_{V_M} - \mathcal{P}_{\widehat{V}_M})x + \boldsymbol{M}'\mathcal{P}_{\widehat{V}_M}x\right\|_2^2 \\ &\le 2\left\|\boldsymbol{M}'\right\|^2 \left(\left\|(\mathcal{P}_{V_M} - \mathcal{P}_{\widehat{V}_M})x\right\|_2^2 + \left\|\mathcal{P}_{\widehat{V}_M}x\right\|_2^2 \right). \qquad (5.30)\end{aligned}$$

Recalling (5.25) and applying Lemma 5.11.1 yields

$$\left\|(\mathcal{P}_{\widehat{V}_M} - \mathcal{P}_{V_M})x\right\|_2^2 \le \left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\|^2 \cdot \|x\|_2^2 \le \Lambda_M^2\left\|\mathcal{P}_{\widehat{V}_M}x\right\|_2^2 + \Lambda_M^4\left\|\beta^*\right\|_2^2.$$

Plugging the above result into (5.30) and applying Lemma 5.11.4 gives

$$\begin{aligned}\left\|\boldsymbol{M}'\mathcal{P}_{V_M}x\right\|_2^2 &\le 2(s_1')^2\left((1 + \Lambda_M^2)\left\|\mathcal{P}_{\widehat{V}_M}x\right\|_2^2 + \Lambda_M^4\left\|\beta^*\right\|_2^2 \right) \\ &\le 2(s_1')^2\left(\frac{2(\widehat{\rho}')^2(1 + \Lambda_M^2)n}{(\rho s_r - \nu_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\left\|\beta^*\right\|_1^2 \right) + \Lambda_M^4\left\|\beta^*\right\|_2^2 \right).\end{aligned}$$

Term 2. Now, to bound the second term of (5.29), which yields

$$\begin{aligned}\left\|\boldsymbol{M}'\mathcal{P}_{V_{M\perp}}x\right\|_2^2 &= \left\|\boldsymbol{U}_{M'}\boldsymbol{S}_{M'}\boldsymbol{V}_{M'}^T\mathcal{P}_{V_{M\perp}}x\right\|_2^2 \\ &\le \left\|\boldsymbol{M}'\right\|^2 \cdot \left\|\boldsymbol{V}_{M'}^T\boldsymbol{V}_{M\perp}\right\|^2 \cdot \left\|\mathcal{P}_{V_{M\perp}}x\right\|_2^2.\end{aligned}$$

Recall that $\mathcal{P}_{V_{M\perp}}\beta^* = 0$ and $\mathcal{P}_{\widehat{V}_{M\perp}}\widehat{\beta} = 0$; hence,

$$\begin{aligned}\left\|\mathcal{P}_{V_{M\perp}}x\right\|_2^2 = \left\|\mathcal{P}_{V_{M\perp}}\widehat{\beta}\right\|_2^2 &= \left\|(\mathcal{P}_{V_{M\perp}} - \mathcal{P}_{\widehat{V}_{M\perp}})\widehat{\beta}\right\|_2^2 \\ &\le \left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\|^2 \cdot \left\|\widehat{\beta}\right\|_2^2 \le \Lambda_M^2 \cdot \left\|\widehat{\beta}\right\|_2^2,\end{aligned}$$

where the final inequality follows from Lemma 5.11.1.

Conclusion. Collecting the above terms completes the proof. ■

■ 5.12.2 Proof of Theorem 5.3.2

Lemma 5.12.2. *Suppose Properties 4.1.1, 4.1.2, 4.1.4, 4.1.6 hold. Consider* $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Z}, r)$, $\widehat{\boldsymbol{M}}' = \text{HSVT}(\boldsymbol{Z}', r')$, *and the unique* β^* *that satisfies* (5.1) *with minimum* ℓ_2*-norm. Then,*

$$\begin{aligned}
\mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}}'\widehat{\beta}) \leq{}& \frac{16\rho^2(s_1')^2(1+\Lambda_M^2)n}{m(\rho s_r - v_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\left\|\beta^*\right\|_1^2\right) \\
&+ \frac{32(v_1')^2 n}{m(\rho s_r - v_1)^2}\left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}})\left\|\beta^*\right\|_1^2\right) \\
&+ 2\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}')\left\|\beta^*\right\|_1^2 + \frac{16(v_1')^2\Lambda_M^2}{(\widehat{\rho}')^2 m}\left\|\beta^*\right\|_2^2 + \frac{8\rho^2(s_1')^2\Lambda_M^4}{(\widehat{\rho}')^2 m}\left\|\beta^*\right\|_2^2 \\
&+ \frac{2(s_1')^2\Lambda_M^2}{m}\left\|\boldsymbol{V}_{M'}^T\boldsymbol{V}_{M\perp}\right\|^2 \cdot \left\|\widehat{\beta}\right\|_2^2.
\end{aligned}$$

Proof. By construction,

$$\boldsymbol{Z}' = \sum_{i=1}^{r'}\widehat{s}_i'\widehat{u}_i' \otimes \widehat{v}_i' + \sum_{i>r'}\widehat{s}_i'\widehat{u}_i' \otimes \widehat{v}_i' = \widehat{\rho}'\widehat{\boldsymbol{M}}' + \boldsymbol{E}'.$$

Letting $x = \widehat{\beta} - \beta^*$, we have that

$$\begin{aligned}
\left\|\widehat{\boldsymbol{M}}'x\right\|_2^2 &= \frac{1}{(\widehat{\rho}')^2}\left\|(\boldsymbol{Z}' - \boldsymbol{E}')x\right\|_2^2 \\
&= \frac{1}{(\widehat{\rho}')^2}\left\|(\rho\boldsymbol{M}' + (\boldsymbol{Z}' - \rho\boldsymbol{M}') - \boldsymbol{E}')x\right\|_2^2 \\
&\leq \frac{2\rho^2}{(\widehat{\rho}')^2}\left\|\boldsymbol{M}'x\right\|_2^2 + \frac{2}{(\widehat{\rho}')^2}\left\|(\boldsymbol{Z}' - \rho\boldsymbol{M}' - \boldsymbol{E}')x\right\|_2^2.
\end{aligned}$$

Using the above result, we then obtain

$$\begin{aligned}
\left\|\widehat{\boldsymbol{M}}'\widehat{\beta} - \boldsymbol{M}'\beta^*\right\|_2^2 &\leq 2\left\|\widehat{\boldsymbol{M}}'x\right\|_2^2 + 2\left\|(\widehat{\boldsymbol{M}}' - \boldsymbol{M}')\beta^*\right\|_2^2 \\
&\leq \frac{4\rho^2}{(\widehat{\rho}')^2}\left\|\boldsymbol{M}'x\right\|_2^2 + \frac{4}{(\widehat{\rho}')^2}\left\|(\boldsymbol{Z}' - \rho\boldsymbol{M}' - \boldsymbol{E}')x\right\|_2^2 + 2\left\|(\widehat{\boldsymbol{M}}' - \boldsymbol{M}')\beta^*\right\|_2^2.
\end{aligned} \tag{5.31}$$

We will now proceed to bound each term independently.

Term 1. We apply Lemma 5.12.1 to obtain

$$\begin{aligned}\left\|\boldsymbol{M}'x\right\|_2^2 &\leq 2(s_1')^2 \left(\frac{2(\widehat{\rho}')^2(1+\Lambda_M^2)n}{(\rho s_r - \nu_1)^2} \left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \left\|\beta^*\right\|_1^2 \right) + \Lambda_M^4 \left\|\beta^*\right\|_2^2 \right) \\ &+ 2(s_1')^2\Lambda_M^2 \cdot \left\| \boldsymbol{V}_{M'}^T \boldsymbol{V}_{M\perp} \right\|^2 \cdot \left\|\widehat{\beta}\right\|_2^2. \end{aligned} \tag{5.32}$$

Term 2. To begin, since $\operatorname{rank}(\boldsymbol{M}') = r'$, Weyl's Inequality (Lemma 3.1.1) gives

$$\left\|\boldsymbol{E}'\right\| = \widehat{s}'_{r'+1} \leq \left\|\boldsymbol{Z}' - \rho \boldsymbol{M}'\right\|.$$

As a result, it follows that

$$\begin{aligned}\left\|(\boldsymbol{Z}' - \rho\boldsymbol{M}' - \boldsymbol{E}')x\right\|_2^2 &\leq 2\left\|\boldsymbol{Z}' - \rho\boldsymbol{M}'\right\|^2 \cdot \|x\|_2^2 + 2\left\|\boldsymbol{E}'\right\|^2 \cdot \|x\|_2^2 \\ &\leq 4\left\|\boldsymbol{Z}' - \rho\boldsymbol{M}'\right\|^2 \cdot \|x\|_2^2 \\ &= 4(\nu_1')^2 \left\| \mathcal{P}_{\widehat{V}_M} x \right\|_2^2 + 4(\nu_1')^2 \Lambda_M^2 \left\|\beta^*\right\|_2^2. \end{aligned}$$

Applying Lemma 5.11.4 gives

$$\begin{aligned}\left\|(\boldsymbol{Z}' - \rho\boldsymbol{M}' - \boldsymbol{E}')x\right\|_2^2 &\leq \frac{8(\widehat{\rho}')^2(\nu_1')^2 n}{(\rho s_r - \nu_1)^2} \left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \left\|\beta^*\right\|_1^2 \right) \\ &+ 4(\nu_1')^2\Lambda_M^2 \left\|\beta^*\right\|_2^2. \end{aligned} \tag{5.33}$$

Term 3. Since $\operatorname{rank}(\widehat{\boldsymbol{M}}') = \operatorname{rank}(\boldsymbol{M}')$, we have

$$\left\|(\widehat{\boldsymbol{M}}' - \boldsymbol{M}')\beta^*\right\|_2^2 \leq m \cdot \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}') \left\|\beta^*\right\|_1^2. \tag{5.34}$$

Conclusion. Plugging in (5.32), (5.33), (5.34) into (5.31) and normalizing completes the proof. ■

Lemma 5.12.3. *Assume the conditions of Theorem 5.3.1 hold. Then,*

$$\mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}}'\widehat{\beta}) \leq \frac{r}{r'}\Delta_{\text{train}} + \Delta_{\text{HSVT}'} + \Delta_{\text{gen}} + \Delta_{\text{model}},$$

where

$$\begin{aligned}\Delta_{\text{train}} &= \frac{C\sigma^2 r}{n} + \frac{C_3 r\sqrt{\log p}}{\rho^4(n \wedge p)} \left\|\beta^*\right\|_1^2 + \Delta_1 \\ \Delta_{\text{HSVT}'} &= \frac{C_3 r'\sqrt{\log p}}{\rho^4(m \wedge p)} \left\|\beta^*\right\|_1^2 \end{aligned}$$

$$\begin{aligned}
\Delta_{\text{gen}} &= \frac{C_4^2}{\rho^4} \frac{p}{n \wedge m \wedge p} \left(\frac{r}{n \wedge p} \left(1 + \frac{r}{r'}\right) + \frac{r \log^2(1/\delta)}{np} \left(1 + \frac{r}{r'p}\right) \right) \|\beta^*\|_2^2 \\
\Delta_{\text{model}} &= \frac{C_5^2 rp}{\rho^2 r'(n \wedge p)} \left\| \boldsymbol{V}_{M'}^T \boldsymbol{V}_{M\perp} \right\|^2 \cdot \left(\left\| \widehat{\beta} - \beta^* \right\|_2^2 + \|\beta^*\|_2^2 \right); \qquad (5.35)
\end{aligned}$$

C_3, Δ_1 are given by (5.3)*; $C_4 = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$; and $C_5 = C_4(1+\log(1/\delta))$.*

Proof. By Lemma 5.12.3, we have

$$\begin{aligned}
\mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}}'\widehat{\beta}) \le{}& \frac{16\rho^2(s_1')^2(1+\Lambda_M^2)n}{m(\rho s_r - v_1)^2} \left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \|\beta^*\|_1^2 \right) \\
&+ \frac{32(v_1')^2 n}{m(\rho s_r - v_1)^2} \left(\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}) + \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \|\beta^*\|_1^2 \right) \\
&+ 2\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}') \|\beta^*\|_1^2 + \frac{16(v_1')^2 \Lambda_M^2}{(\widehat{\rho}')^2 m} \|\beta^*\|_2^2 + \frac{8\rho^2(s_1')^2\Lambda_M^4}{(\widehat{\rho}')^2 m} \|\beta^*\|_2^2 \\
&+ \frac{2(s_1')^2\Lambda_M^2}{m} \left\| \boldsymbol{V}_{M'}^T \boldsymbol{V}_{M\perp} \right\|^2 \cdot \left\| \widehat{\beta} \right\|_2^2.
\end{aligned}$$

We will bound each term independently. However, we first use the arguments that led to (5.17) to establish $\mathbb{P}(E') \ge 1 - \delta$, where E' is given by (5.28). Throughout, we suppose E' occurs. Importantly, we highlight that under E' and Property 4.1.3,

$$\Lambda_M^2 \le \frac{C_4}{\rho^2} \left(\frac{r}{n \wedge p} + \frac{r \log(1/\delta)}{np} \right), \qquad (5.36)$$

where $C_4 = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$.

Term 1. Recall from (5.26),

$$\mathcal{E}_{\text{train}}(\widehat{\boldsymbol{M}}\widehat{\beta}),\ \mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}}) \le \frac{2\sigma^2 r}{n} + \frac{C_3 r\sqrt{\log p}}{\rho^4(n \wedge p)} \|\beta^*\|_1^2 + \Delta_1,$$

where C_3, Δ_1 are given by (5.3). Further, Property 4.1.3 and (5.27) yield

$$\frac{\rho^2(s_1')^2 n}{m(\rho s_r - v_1)^2} \le C\frac{r}{r'}.$$

Given that $\Lambda_M^2 = o(1)$, we conclude

$$\{\text{term } 1\} \le \frac{r}{r'} \left(\frac{C\sigma^2 r}{n} + \frac{C_3 r\sqrt{\log p}}{\rho^4(n \wedge p)} \|\beta^*\|_1^2 + \Delta_1 \right).$$

Term 2. We follow the proof of term 1. By Property 4.1.3, we have $v_1' \leq s_1'$. As a result, we conclude that the second term is bounded above by the first term.

Term 3. We apply Lemma 5.2.1 to obtain

$$\{\text{term } 3\} = 2\mathcal{E}_{\text{HSVT}}(\widehat{\boldsymbol{M}'}) \|\beta^*\|_1^2 \leq \frac{C_3 r' \sqrt{\log p}}{\rho^4 (m \wedge p)} \|\beta^*\|_1^2,$$

where C_3 is given by (5.3).

Term 4. We use (5.36) to obtain

$$\frac{(v_1')^2 \Lambda_M^2}{(\widehat{\rho}')^2 m} \leq \frac{C_4^2}{\rho^4} \left(1 + \frac{p}{m}\right) \left(\frac{r}{n \wedge p} + \frac{r \log^2(1/\delta)}{np} \right).$$

Term 5. By (5.36), observe that

$$\Lambda_M^4 \leq \frac{C_4^2}{\rho^4} \left(\frac{r^2}{(n \wedge p)^2} + \frac{r^2 \log^2(1/\delta)}{(np)^2} \right).$$

This yields

$$\frac{\rho^2 (s_1')^2 \Lambda_M^4}{(\widehat{\rho}')^2 m} \leq \frac{C_4^2 p}{\rho^4 r'} \left(\frac{r^2}{(n \wedge p)^2} + \frac{r^2 \log^2(1/\delta)}{(np)^2} \right).$$

Combining the bounds for terms 4 and 5 gives us the following upper bound:

$$\{\text{term } 4 + \text{term } 5\} \leq \frac{C_4^2}{\rho^4} \frac{p}{n \wedge m \wedge p} \left(\frac{r}{n \wedge p} \left(1 + \frac{r}{r'}\right) + \frac{r \log^2(1/\delta)}{np} \left(1 + \frac{r}{r' p}\right) \right) \|\beta^*\|_2^2. \tag{5.37}$$

Term 6. Using the arguments from above, we obtain

$$\frac{(s_1')^2 \Lambda_M^2}{m} \leq \frac{C_5}{\rho^2} \frac{rp}{r'(n \wedge p)},$$

where $C_5 = C_4(1 + \log(1/\delta))$. Further, we note that

$$\left\|\widehat{\beta}\right\|_2^2 \leq 2\left\|\widehat{\beta} - \beta^*\right\|_2^2 + 2\|\beta^*\|_2^2$$

Combining the above yields

$$\{\text{term 6}\} \leq \frac{C_5^2}{\rho^2} \frac{rp}{r'(n \wedge p)} \left\| \boldsymbol{V}_{M'}^T \boldsymbol{V}_{M\perp} \right\|^2 \cdot \left(\left\| \widehat{\beta} - \beta^* \right\|_2^2 + \|\beta^*\|_2^2 \right).$$

Conclusion. Putting everything together, we conclude

$$\mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}'}\widehat{\beta}) \leq \frac{r}{r'} \Delta_{\text{train}} + \Delta_{\text{HSVT}'} + \Delta_{\text{gen}} + \Delta_{\text{model}},$$

where $\Delta_{\text{train}}, \Delta_{\text{HSVT}'}, \Delta_{\text{gen}}, \Delta_{\text{model}}$ are given in (5.35). ■

Completing Proof of Theorem 5.3.2.

Proof. We will simplify the terms in Lemma 5.12.3. To begin, since $\text{span}(\boldsymbol{V}_{M'}) \subseteq \text{span}(\boldsymbol{V}_M)$, it follows that $\Delta_{\text{model}} = 0$, $r \geq r'$, and

$$\frac{r}{r'} \Delta_{\text{train}} + \Delta_{\text{HSVT}'} \leq \frac{r}{r'} \left(\frac{C\sigma^2 r}{n} + \frac{C_3 r \sqrt{\log p}}{\rho^4 (n \wedge m \wedge p)} \|\beta^*\|_1^2 + \Delta_1 \right), \tag{5.38}$$

where C_3, Δ_1 are given by (5.3). Further,

$$\Delta_{\text{gen}} \leq \frac{C_4^2}{\rho^4} \frac{r}{r'} \left(\frac{rp}{(n \wedge m \wedge p)^2} + \frac{r \log^2(1/\delta)}{n(n \wedge m \wedge p)} \right) \|\beta^*\|_2^2, \tag{5.39}$$

where C_4 is given by (5.35). Collecting and simplifying the above results gives the following:

$$\mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}'}\widehat{\beta}) \leq \frac{r}{r'} \left(\frac{C\sigma^2 r}{n} + \frac{C_1 C_6}{\rho^4} \frac{r\sqrt{\log p}}{n \wedge m \wedge p} \|\beta^*\|_1^2 + \frac{C_4^2}{\rho^4} \frac{rp}{(n \wedge m \wedge p)^2} \|\beta^*\|_2^2 + \Delta_1 \right),$$

where C_1 is given by (5.3) and $C_6 = C_1 K^2 (1 + \log^2(1/\delta))$. The proof is complete after relabeling constants and observing that $\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \leq \mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}'}\widehat{\beta})$. ■

■ 5.12.3 Corollaries: Bounds in Expectation

Corollary 5.12.1. *Suppose the conditions of Theorem 5.3.2 hold. Then for any $\delta > 0$,*

$$\mathbb{E}[\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}})] \leq \frac{r}{r'} \left(\frac{2\sigma^2 r}{n} + \frac{C_2' C_3' r \log^2(p/\delta)}{\rho^4 (n \wedge m \wedge p)} \|\beta^*\|_1^2 + \frac{(C_3')^2 rp}{\rho^4 (n \wedge m \wedge p)^2} \|\beta^*\|_2^2 \right) + 4\delta,$$

where C_2' is given by (5.21); and $C_3' = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$.

Proof. The proof follows that of Theorem 5.3.2. As shown in the proof of Lemma 5.12.3, $\mathbb{P}(E') \geq 1 - \delta$, where E' is given by (5.28). Now, observe that

$$\mathbb{E}[\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}})] \leq \mathbb{E}[\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \mid E'] + \delta \cdot \mathbb{E}[\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \mid (E')^c].$$

We proceed to bound each term separately.

Term 1. Suppose E occurs. Using Lemma 5.12.3, we arrive at the following inequality:

$$\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \leq \mathcal{E}_{\text{test}}(\widehat{\boldsymbol{M}'}\widehat{\beta}) \leq \frac{r}{r'}\Delta_{\text{pre}} + \Delta_{\text{HSVT}'} + \Delta_{\text{gen}} + \Delta_{\text{model}}.$$

First, note that our assumptions give $\Delta_{\text{model}} = 0$ and $r \geq r'$. We then use (5.38), coupled with Corollaries 5.9.1 and 5.10.1, to establish

$$\begin{aligned}\frac{r}{r'}\Delta_{\text{pre}} + \Delta_{\text{HSVT}'} \leq \frac{r}{r'}\bigg(&\frac{2\sigma^2 r}{n} + \frac{C_1' r}{\rho^4 (n \wedge m \wedge p)}\|\beta^*\|_1^2 \\ &+ \bigg(\frac{C_2' r \log(1/\delta)}{\rho^4 \sqrt{n \wedge m}(n \wedge m \wedge p)} + \frac{C_2'\sqrt{r}}{\rho^2 (n \wedge m)}\bigg)\sqrt{\log(p/\delta)}\|\beta^*\|_1^2\bigg),\end{aligned}$$

where C_1', C_2' are given by (5.21). At the same time, following the arguments that led to (5.39), we obtain

$$\Delta_{\text{gen}} \leq \frac{(C_3')^2}{\rho^4}\frac{r}{r'}\left(\frac{rp}{(n \wedge m \wedge p)^2} + \frac{r \log^2(1/\delta)}{n(n \wedge m \wedge p)}\right)\|\beta^*\|_2^2,$$

where $C_3' = C(1+\gamma^2)(1+K^2)$. Therefore, combining and simplifying the above results yield

$$\mathbb{E}[\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}})] \leq \frac{r}{r'}\left(\frac{2\sigma^2 r}{n} + \frac{C_2' C_3' r \log^2(p/\delta)}{\rho^4 (n \wedge m \wedge p)}\|\beta^*\|_1^2 + \frac{(C_3')^2 rp}{\rho^4 (n \wedge m \wedge p)^2}\|\beta^*\|_2^2\right).$$

Term 2. By Property 4.1.2 and Algorithm 3, it immediately follows that

$$\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \leq 4.$$

Conclusion. Collecting terms completes the proof. ∎

Completing Proof of Corollary 5.3.1

Proof. The result follows immediately after applying Corollary 5.12.1 with $n = \Theta(m) = \Theta(p)$. ■

■ 5.13 Proofs: A Subspace Inclusion Hypothesis Test

Lemma 5.13.1. *Suppose Properties 4.1.1, 4.1.2, 4.1.3, 4.1.5, 4.1.6 hold. Consider* $\operatorname{rank}(\widehat{\boldsymbol{M}}) = \operatorname{rank}(\boldsymbol{M}) = r$ *and* $\operatorname{rank}(\widehat{\boldsymbol{M}'}) = \operatorname{rank}(\boldsymbol{M}') = r'$. *Then under* H_0,

$$\left\| \widehat{\boldsymbol{V}}_{M'} - \mathcal{P}_{\widehat{V}_M} \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 \leq 2r' \left(\left\| \sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M) \right\|^2 + \left\| \sin\Theta(\widehat{\boldsymbol{V}}_{M'}, \boldsymbol{V}_{M'}) \right\|^2 \right).$$

Proof. Observe that

$$\begin{aligned}
\left\| \widehat{\boldsymbol{V}}_{M'} - \mathcal{P}_{\widehat{V}_M} \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 &= \left\| \mathcal{P}_{\widehat{V}_{M\perp}} \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 \\
&= \left\| (\mathcal{P}_{\widehat{V}_{M\perp}} - \mathcal{P}_{V_{M\perp}}) \widehat{\boldsymbol{V}}_{M'} + \mathcal{P}_{V_{M\perp}} \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 \\
&\leq 2 \left\| (\mathcal{P}_{\widehat{V}_{M\perp}} - \mathcal{P}_{V_{M\perp}}) \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 + 2 \left\| \mathcal{P}_{V_{M\perp}} \widehat{\boldsymbol{V}}_{M'} \right\|_F^2. \qquad (5.40)
\end{aligned}$$

We will now bound each term independently.

For the first term, we have

$$\left\| (\mathcal{P}_{\widehat{V}_{M\perp}} - \mathcal{P}_{V_{M\perp}}) \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 \leq \left\| \sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M) \right\|^2 \cdot \left\| \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 = r' \cdot \left\| \sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M) \right\|^2 \qquad (5.41)$$

Further, under H_0, recall that $\mathcal{P}_{V_{M\perp}} \boldsymbol{V}_{M'} = 0$. Therefore, using the isometric property of $\widehat{\boldsymbol{V}}_{M'}$, we obtain

$$\begin{aligned}
\left\| \mathcal{P}_{V_{M\perp}} \widehat{\boldsymbol{V}}_{M'} \right\|_F^2 &= \left\| \mathcal{P}_{V_{M\perp}} \mathcal{P}_{\widehat{V}'_M} \right\|_F^2 \\
&= \left\| \mathcal{P}_{V_{M\perp}} (\mathcal{P}_{\widehat{V}'_M} - \mathcal{P}_{V'_M}) \right\|_F^2 \\
&\leq \left\| \mathcal{P}_{V_{M\perp}} \right\|^2 \cdot \left\| \sin\Theta(\widehat{\boldsymbol{V}}_{M'}, \boldsymbol{V}_{M'}) \right\|_F^2 \\
&\leq r' \cdot \left\| \sin\Theta(\widehat{\boldsymbol{V}}_{M'}, \boldsymbol{V}_{M'}) \right\|^2. \qquad (5.42)
\end{aligned}$$

Plugging in (5.41) and (5.42) into (5.40) completes the proof. ■

■ 5.13.1 Proof of Theorem 5.4.1

Proof. Let us first fix some $\alpha > 0$. Let $C' = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$. By Lemma 5.8.4, it follows that with probability at least $1-\alpha$,

$$\begin{aligned}\left\|\boldsymbol{Z} - \rho\boldsymbol{M}\right\|^2 &\leq C'\left(n + p + \log(1/\alpha)\right)\\ \left\|\boldsymbol{Z}' - \rho\boldsymbol{M}'\right\|^2 &\leq C'\left(m + p + \log(1/\alpha)\right).\end{aligned}$$

Combining the above with Lemma 5.11.1 then yields

$$\begin{aligned}\left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\|^2 &\leq \frac{C'r}{\rho^2}\left(\frac{1}{n\wedge p} + \frac{\log(1/\alpha)}{np}\right)\\ \left\|\sin\Theta(\widehat{\boldsymbol{V}}_{M'}, \boldsymbol{V}_{M'})\right\|^2 &\leq \frac{C'r'}{\rho^2}\left(\frac{1}{m\wedge p} + \frac{\log(1/\alpha)}{mp}\right).\end{aligned}$$

Plugging the above into Lemma 5.13.1 concludes the proof. ■

Chapter 6

Robust Synthetic Control

6.1 Introduction

During the early 1970's, Basque Country, one of the wealthiest regions in Spain, began to experience terrorist activity. Intuitively, this political and social unrest should have had adverse effects on the region's economic wealth. However, evidence of this belief is difficult to establish since it is impossible to simultaneously observe (and thus compare) Basque Country's economic health in the presence *and* absence of terrorist conflict. This is the fundamental missing data problem of causal inference.

Given that experimental studies (ESs) are simply infeasible in such a setting, a first order attempt may be to identify a control region for comparison. Unfortunately, a simple juxtaposition of Basque's economic trajectory with that of a nearby Spanish region would not provide a statistically valid conclusion unless that region proved to be demonstrably similar to Basque sans the political and societal instability. In general, there may not ever exist a natural control state, and subject-matter experts tend to disagree on the most appropriate control for comparison. At this point, one may opt to apply a pure time series analysis instead. However, this approach will also often prove to be futile. To see this, consider the situation where the economy was already in decline prior to the start of the decade. If the economy continues to fall thereafter, then it would be difficult to attribute the source of the downturn to the underlying trend or the terrorist activity. This summarizes a major bottleneck of observational studies (OSs).

6.1.1 Problem Statement

More formally, we are interested in outcomes (e.g., per-capita GDP) associated with N units (e.g., Spanish regions) across T measurements (e.g., time points). In the context of standard OS settings, it follows that $D = 2$, where (without loss of generality) we denote the "no-intervention" state (or control) with $d = 1$ and let $d = 2$ represent an actual intervention (e.g., terrorism). For simplicity, we consider the single metric case ($P = 1$) and suppress all dependencies on the metric.

Observations

Throughout this chapter, we let unit $n = 1$ represent our target unit of interest, which is assumed to be exposed to intervention $d = 2$ during the post-intervention period; thus, $\mathcal{I}^{(2)} = \{1\}$. Meanwhile, all other units form our universe of donors, yielding $\mathcal{I}^{(1)} = [N] \backslash \{1\}$ and $N^{(1)} = N - 1$.

We encode our observations into a $T \times N \times 2$ tensor $\boldsymbol{Z} = [Z_{tn}^{(d)}]$, where

$$Z_{tn}^{(d)} = \begin{cases} Y_{tn}^{(1)} \cdot \pi_{tn}^{(1)}, & \text{for all } t \in [T], n > 1, d = 1 \\ Y_{t1}^{(1)} \cdot \pi_{t1}^{(1)}, & \text{for all } t \leq T_0, n = 1, d = 1 \\ Y_{t1}^{(2)} \cdot \pi_{t1}^{(2)}, & \text{for all } t > T_0, n = 1, d = 2 \\ \star, & \text{otherwise.} \end{cases}$$

In words, our donor units remain in the no-intervention state across all measurements. Our target unit, on the other hand, is also unaffected during the pre-intervention period, but receives intervention $d = 2$ during the post-intervention period.

Aim

Given $\boldsymbol{Z}$, our goal is to infer the potential outcomes in the *absence* of any intervention for the *target unit* during the post-intervention period, i.e., $M_{t1}^{(1)}$ for all $t > T_0$.

■ 6.1.2 Classical Synthetic Control

As a suggested remedy to overcome the limitations of OSs, Abadie and Gardeazabal (2003) proposed a powerful, data-driven approach known as synthetic control (SC). Indeed, SC has grown to become a standard method in econometrics for comparative case studies and policy evaluation, and since its conception, it has been analyzed in Abadie et al. (2010), Doudchenko and Imbens (2016), Athey and Imbens (2016), Athey et al. (2017), Hsiao et al. (2018).

Returning to our example, SC predicts Basque Country's counterfactual economic evolution without terrorism by constructing a "synthetic" control region, which is represented by the combination of donor units (assumed to be unaffected) that best resembles Basque

Country prior to the outset of political unrest. More formally, SC learns $\beta^{\text{sc}} \in \mathbb{R}^{N-1}$ as

$$\beta^{\text{sc}} \in \arg\min_{w \in \text{SC}} \sum_{t=1}^{T_0} \left(Y_{t1}^{(1)} - \sum_{n=2}^{N} w_n Y_{tn}^{(1)} \right)^2,$$

where the constraint set $\text{SC} \subseteq \mathbb{R}^{N-1}$ differs across variants of the method, but is classically taken to be over the probability simplex, i.e., $\beta_n^{\text{sc}} \geq 0$ and $\sum_n \beta_n^{\text{sc}} = 1$, (cf. Abadie and Gardeazabal (2003); Abadie et al. (2010)). Subsequently, $\widehat{Y}_{t1}^{(1),\text{sc}} = \sum_{n=2}^{N} \beta_{n-1}^{\text{sc}} Y_{tn}^{(1)}$ is the estimate for the target unit under no-intervention for $t > T_0$. Comparing $\widehat{Y}_{t1}^{(1),\text{sc}}$ with $Y_{t1}^{(2)}$ for $t > T_0$ evaluates the impact of intervention 2 on the target unit compared to the no-intervention effect.

Limitations

Within the SC literature, two standard assumptions are made: (i) The potential outcomes under the null-intervention follow a factor model; that is, $M_{tn}^{(1)} = \langle u_t, v_n \rangle$ for all (t, n), where $u_t, v_n \in \mathbb{R}^r$ are the latent factors associated with measurement and unit, respectively. We note that this a special case of the proposed tensor factor model given by Property 4.1.1 which only considers the frontal slice of the order-three potential outcomes tensors corresponding to metric $p = 1$. (ii) There exists a synthetic control for the target unit that is formed as a weighted combination of the donor units. Indeed, the latter assumption is the fundamental hypothesis that drives all SC-related works, but it is not clear when such a hypothesis holds.

Algorithmically, despite its widespread applicability, the classical SC method is unable to handle settings with noisy and sparse covariates (donor data), typical characteristics of modern datasets. Theoretically, a quantitative hypothesis test to check the appropriateness of applying SC-like methods and meaningful non-asymptotic analysis (under a high-dimensional framework), particularly that which captures the behavior of the post-intervention error, have also remained elusive.

■ 6.2 Robust Synthetic Control

As the primary contribution of this chapter, we present the robust synthetic control (RSC) algorithm, which overcomes the limitations of the classical SC method described above.

■ 6.2.1 Algorithm

Having established the robustness and generalization properties of PCR in Chapter 5, we utilize it as the key subroutine within RSC (Algorithm 4) to learn a synthetic control. For a graphical depiction of the input and output of RSC, please refer to Figure 1.3.

Notation. Recall that unit 1 is our target unit. In consistency with (4.1), we represent the pre- and post-intervention observation matrices associated with the donors, which remain in the no-intervention state ($d = 1$) across all T, as

$$\begin{aligned} \boldsymbol{Z}^{(1)}_{\text{pre}} &= [Z^{(1)}_{tn} : t \leq T_0, n > 1] \in \mathbb{R}^{T_0 \times (N-1)} \\ \boldsymbol{Z}^{(1)}_{\text{post}} &= [Z^{(1)}_{tn} : t > T_0, n > 1] \in \mathbb{R}^{(T-T_0) \times (N-1)}. \end{aligned}$$

Further, we denote $y^{(1)}_1 = [Y^{(1)}_{t1} : t \leq T_0]$ as the pre-intervention observations for our target unit, which is also observed under $d = 1$. We are now ready to state the RSC algorithm.

```
Algorithm 4: RSC
  Data: $y^{(1)}_1, \boldsymbol{Z}^{(1)}_{\text{pre}}, \boldsymbol{Z}^{(1)}_{\text{post}}, k, k'$
  Result: $\widehat{M}^{(1)}_1 = [\widehat{M}^{(1)}_{t1} : t > T_0]$
    1. Learn synthetic target model:
       (a) $\widehat{\beta} \leftarrow \text{PCR}(\boldsymbol{Z}^{(1)}_{\text{pre}}, y^{(1)}_1, k)$
    2. Predict counterfactual prediction outcomes:
       (a) $\widehat{\boldsymbol{M}}^{(1)}_{\text{post}} \leftarrow \text{HSVT}(\boldsymbol{Z}^{(1)}_{\text{post}}, k')$
       (b) $\widehat{M}^{(1)}_1 \leftarrow \text{Truncate}(\widehat{\boldsymbol{M}}^{(1)}_{\text{post}} \widehat{\beta})$
```

Algorithmic Intuition

In words, RSC first builds a synthetic control of the target unit using the entire collection of donor units; that is, RSC finds the set of weights, defined by $\widehat{\beta}$, that best approximates the outcome variables of the target unit during the pre-intervention period. Crucially, RSC uses PCR to protect against over-fitting to the idiosyncrasies of the data beyond the inherent model complexity in the target and donor trajectories (recall PCR learns a linear model in the reduced subspace spanned by the top principal components of $\boldsymbol{Z}^{(1)}_{\text{pre}}$). Once the model is learned, RSC rescales the observed outcome variables associated with the donor units during the post-intervention period according to the set of weights $\widehat{\beta}$. As with the first step (pre-intervention model learning phase), RSC performs HSVT on the donor post-intervention data for de-noising and regularization purposes.

Importantly, we underscore that both the model learning and prediction processes only utilize data under the no-intervention state. This distinction will be made clear when we discuss synthetic interventions in Chapter 8.

■ 6.2.2 Existence of SC

Up until this point, we have not justified the regression step of RSC. In fact, across the many SC variants, (regularized) regression is consistently employed as a key subroutine in learning a synthetic control without any justification. As described in Section 6.1.2, it is a standard, fundamental assumption within the SC literature that a synthetic control for the target unit exists within the reservoir of donors; for instance, the classical works of Abadie and Gardeazabal (2003); Abadie et al. (2010) assume that the target unit can be expressed as a convex combination of the donors. To the best of our knowledge, despite the ubiquity of this assumption, it is not clear when such a hypothesis holds.

However, as stated by Proposition 4.1.1 of Chapter 4, the standard matrix factor model within the SC literature (a simplified version of Property 4.1.1) implies that an invariant linear model between the target unit and donors persists across measurements with high probability; effectively, with probability at least $1 - r/(N-1)$, where r is the dimension of the latent unit and time factors. This is the key result that justifies RSC (and, more generally, learning a linear model). Therefore, under the standard SC setting, we establish that a linear synthetic control (almost) always exists and need not be assumed as an axiom as is traditionally done in the literature.

■ 6.2.3 Theoretical Performance Guarantees

Objective

Recall that our aim is to recover the underlying potential outcomes for our target unit 1 under control ($d = 1$), i.e., $M_{t1}^{(1)}$ for all $t > T_0$.

Notation. Since we are only interested in recovering the counterfactuals under the no-intervention state, we suppress dependencies on $d = 1$ for ease of notation, e.g., $\boldsymbol{Z}_{\text{pre}} = \boldsymbol{Z}_{\text{pre}}^{(1)}$ and $r_{\text{post}} = \text{rank}(\boldsymbol{M}_{\text{post}}^{(1)})$, where $\boldsymbol{M}_{\text{post}}^{(1)}$ is given by (4.1). Further, let $\beta^* = \beta^{(1,1)}$, where $\beta^{(1,1)}$ is given in (4.3).

Evaluation Metric

Hence, we evaluate the RSC algorithm based on its post-intervention squared prediction error. Specifically, we define the *post-intervention* (or *test*) error for unit $n = 1$ under the null-intervention $d = 1$ as

$$\mathcal{E}_{\text{post}}(\widehat{M}_1) = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} (\widehat{M}_{t1}^{(1)} - M_{t1}^{(1)})^2,$$

where $\widehat{M}_1 = [\widehat{M}_{t1}^{(1)} : t > T_0] = \text{RSC}(y_1, Z_{\text{pre}}, Z_{\text{post}}, r_{\text{pre}}, r_{\text{post}})$.

Remark. Although Property 4.1.5 assumes independent noise entries, our results are stated when ε_{tn} can be dependent across donors for a given t, i.e., only the target and donor noise must remain independent.

Post-intervention Prediction Error

Below, we present our main results, which bound RSC's post-intervention prediction error in both high probability and expectation under a special case (i.e., the standard matrix factor model) of the setting described in Chapter 4.

Theorem 6.2.1 (RSC Error in High-Probability). *Let Properties 4.1.1, 4.1.2, 4.1.3, 4.1.4, 4.1.5, 4.1.6 hold. Consider the unique β^* of minimum ℓ_2-norm that satisfies (4.3). For any $\delta > 0$ and some $C > 0$, if $\rho \geq \sqrt{\frac{C_1 C_2 r_{\text{pre}}}{N' \wedge T'}}$, then the following holds w.p. at least $1 - \delta$:*

$$\mathcal{E}_{\text{post}}(\widehat{M}_1) \leq \frac{r_{\text{pre}}}{r_{\text{post}}} \Big(\frac{C\sigma^2 r_{\text{pre}}}{T_0} + \frac{C_1 C_2 r_{\text{pre}} \sqrt{\log N'}}{\rho^4 (N' \wedge T')} \|\beta^*\|_1^2 + \frac{C_1^2 r_{\text{pre}} N'}{\rho^4 (N' \wedge T')^2} \|\beta^*\|_2^2 + \Delta \Big),$$

where $N' = N - 1$, $T' = T_0 \wedge (T - T_0)$,

$$\Delta = \frac{C_2}{\sqrt{T_0}} \|\beta^*\|_1, \quad C_1 = C(1 + \sigma^4)(1 + \gamma^2)(1 + K^2), \quad C_2 = C_1 K^2 (1 + \log^2(1/\delta)).$$

Proof. The result is immediate from Theorem 5.3.2. ■

Corollary 6.2.1 (RSC Error in Expectation). *Let the conditions of Theorem 6.2.1 hold. Then for any $\delta > 0$,*

$$\mathbb{E}[\mathcal{E}_{\text{post}}(\widehat{M}_1)] \leq \frac{r_{\text{pre}}}{r_{\text{post}}} \left(\frac{2\sigma^2 r_{\text{pre}}}{T_0} + \frac{C_3 C_4 r_{\text{pre}} \log^2(N'/\delta)}{\rho^4 (N' \wedge T')} \|\beta^*\|_1^2 + \frac{C_4^2 r_{\text{pre}} N'}{\rho^4 (N' \wedge T')^2} \|\beta^*\|_2^2 \right) + 4\delta,$$

where $C_3 = CK^2(1+\sigma^4)(1+\gamma^2)(1+K^2)$ and $C_4 = C(1+\sigma^2)(1+\gamma^2)(1+K^2)$.

Proof. The result is immediate from Corollary 5.3.1. ■

Intepretation. For simplicity, let $T_0 = \Theta(N) = \Theta(T)$. Ignoring log factors, Corollary 6.2.1 states that the post-intervention error decays linearly with T_0, in expectation. In words, Theorem 6.2.1 and Corollary 6.2.1 establish that RSC produces consistent counterfactual estimates of the potential outcomes in the absence of any intervention. To the best of our knowledge, Theorem 6.2.1 and Corollary 6.2.1 provide the first finite-sample analysis that captures the behavior of the post-intervention prediction error (with respect to the latent potential outcomes) of SC-like methods.

■ 6.2.4 Empirical Validation: Placebo Studies

We have provided the theoretical performance of RSC in Section 6.2.3, and shown it to be a consistent estimator of the unobservable counterfactuals. However, the question still remains: how does one determine the empirical performance of a counterfactual estimation method without access to ground-truth values? Although it is possible to use the pre-intervention data to cross-validate the performance of any estimation method, such a methodology ignores the actual period of interest, i.e., the post-intervention period, and is prone to over-fitted results that may not be indicative of the counterfactual performance. An alternate and more effective approach is to study the performance of an estimation method on units that do *not* experience the intervention, i.e., the donor units. Indeed, since RSC (like other SC variants) is designed to predict the counterfactuals in the absence of any intervention, performing RSC with the donor units (as opposed to the exposed unit) as the targets should ideally reproduce the observed trajectories; this is precisely the placebo studies proposed by Abadie and Gardeazabal (2003); Abadie et al. (2010). Thus, if the method is able to accurately estimate the observed post-intervention evolution of the donor units, it would be reasonable to assume that it would perform well in estimating the unobserved counterfactuals for the target unit of interest. This post-intervention period placebo study, or cross-validation (as its known within the machine learning/statistics literature), becomes our primary empirical metric of evaluation in all of our case studies.

6.3 Empirical Case Studies

Here, we present empirical results using the RSC method and several known datasets in the literature to explain the nuances of the results stated above.

6.3.1 Terrorism in Basque Country

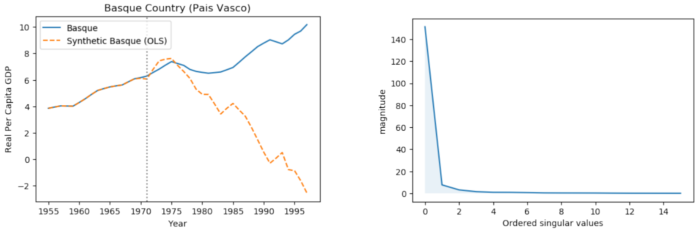
One canonical case study within the SC Literature investigates the impact of terrorism on the economy in Basque Country (see Abadie and Gardeazabal (2003)). Here, the target unit of interest is Basque Country, the donor pool consists of neighboring Spanish regions, and the intervention is represented by the first wave of terrorist activity in 1970. The aim in this study is to isolate the effect of terrorism on the GDP of Basque Country. That is, to evaluate the effect of terrorism, we aim to estimate the unobservable counterfactual GDP growth in the absence of terrorism for Basque Country using observations from the other Spanish regions, which are assumed to be unaffected by the political unrest.

Empirical Results and Key Takeaways

We will use two evaluation metrics: (i) Since we do not have access to Basque's counterfactual GDP post 1970 without terrorism, we will use the celebrated results of Abadie and Gardeazabal (2003) as our baseline; this is our chosen "ground-truth" because these counterfactual trajectories have been widely accepted by the econometrics community. (ii) We also perform placebo studies (i.e., cross-validation), as described in Section 6.2.4, by iteratively designating each neighboring Spanish region (donor for Basque) as the target.

Importance of Regularization. As stated above, we use the results of Abadie and Gardeazabal (2003) as the ground-truth counterfactuals. Recall that the classical SC method proposed by Abadie and Gardeazabal (2003) enforces the learnt model to have non-negative weights and sum to one. This offers two benefits: (i) qualitatively, the model offers an intuitive interpretation of the synthetic control unit (e.g., synthetic Basque is 85% Catalonia and 15% Madrid), and (ii) quantitatively, this form of regularization protects the model from overfitting to the data. RSC, on the other hand, removes the convexity constraint on the model and instead employs PCR to regularize (see Chapter 5.5.2 for details) and learn a linear synthetic control.

To highlight the importance of the PCA subroutine (and, more generally, regularization), we construct a synthetic Basque via vanilla OLS. As seen in Figure 6.1a, OLS clearly overfits to the pre-intervention training data and fails to extrapolate post-intervention. In



(a) Synthetic Basque via OLS.

(b) Spectrum of Basque's donor data.

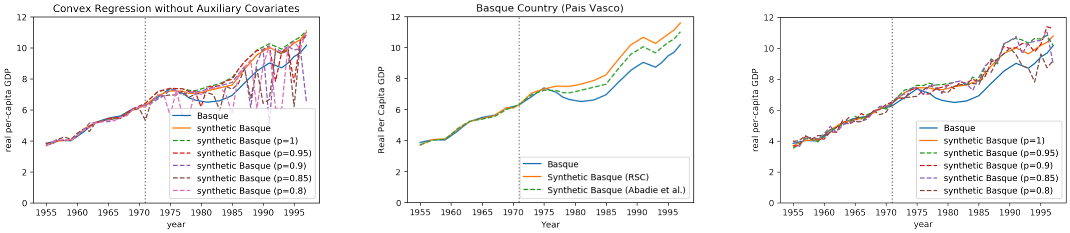
Figure 6.1: Plots highlight the importance of regularization and justification for PCR. Specifically, (a) illustrates how OLS overfits to the training data while (b) displays the low-dimensional structure of the donor data, which motivates the usage of PCR since it regresses on the reduced subspace spanned by the top principal components.

fact, the synthetic Basque GDP as predicted by OLS suggests that terrorism actually benefited the Basque economy in the long-term(!), which contradicts the conclusions drawn by the econometrics community.

The first step of PCR (i.e., PCA) is even more starkly empirically motivated by inspecting the singular value spectrum of the donor data, which is shown in Figure 6.1b. Clearly, the data exhibits low-dimensional structure with over 99% of the spectral energy captured in the top singular value in both settings, which fits the conditions under which our theoretical results imply low pre- and post-intervention prediction errors. Hence, it is reasonable to first extract the signal by filtering out the low principal components, which correspond to idiosyncratic noise, prior to learning a synthetic control; in other words, PCR is a natural and empirically justified regularization method to employ in this setting.

Robustness of RSC. To highlight the robustness of RSC, we begin by randomly obfuscating data, ranging from 5-20%, and plotting the resulting synthetic Basque GDPs predicted via convex regression on the outcome GDP data, i.e., the original Synthetic Control method *without* auxiliary covariates, in Figure 6.2a; here, the solid blue and orange lines represent the observed and synthetic Basque (predicted by Abadie and Gardeazabal (2003)), respectively, while the dashed lines represent the synthetic Basques (learned *without* auxiliary covariates) under varying levels of missing data. As seen from the figure, the original SC method is not robust to sparse observations, which may explain its dependency on auxiliary covariates to learn its model.

The resulting synthetic Basque as per RSC is shown in Figure 6.2b, which pleasingly closely matches that of Abadie and Gardeazabal (2003). Similarly, in Figure 6.2c, we



(a) Synthetic Basque via Abadie and Gardeazabal (2003) (without auxiliary covariates) under varying levels of missing data.

(b) Synthetic Basque via RSC and Abadie and Gardeazabal (2003).

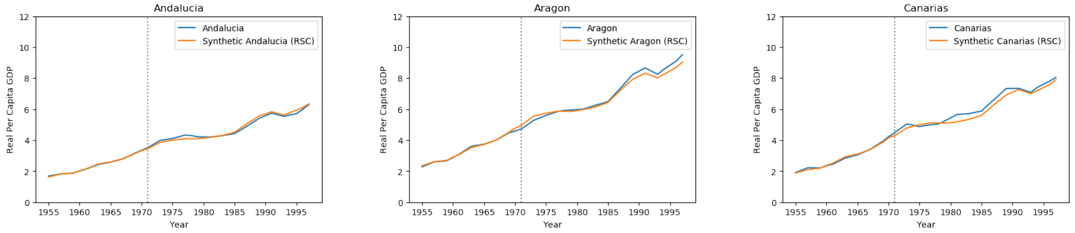
(c) Synthetic Basque via RSC under varying levels of missing data.

Figure 6.2: Counterfactual estimates of Basque Country's GDP in the absence of terrorism. While Figure 6.2b demonstrates that RSC (without covariate data) and Abadie and Gardeazabal (2003) (with covariate data) produce similar results when all observations are accessible, Figures 6.2a and 6.2c highlight RSC's robustness to sparsity compared to the classical SC method (when covariate data is withheld).

display various synthetic Basque GDPs after randomly obfuscating the donor observations. Across the varying levels of missing data from 5-20%, the synthetic Basque GDPs continue to resemble the baseline estimates of Abadie and Gardeazabal (2003) such that the same negative economic effects of terrorism can be drawn.

Importantly, we underscore that all of the results computed via RSC shown in Figures 6.2b and 6.2c only use the outcome data and *without* any access to the auxiliary covariate information that was required to achieve the results in Abadie and Gardeazabal (2003), i.e., only the per-capita GDP values are utilized in the PCR learning process. Hence, PCR exhibits desirable robustness properties with respect to missing and noisy data, and with less stringent data requirements to achieve similar counterfactual estimates.

Placebo Studies: Cross-Validation. In Table 6.1, we show the results of the hypothesis test and median R^2-score across all neighboring Spanish donor regions. The hypothesis test passes at a significance level of $\alpha = 0.05$, which suggests that we cannot reject the null hypothesis where the post-intervention donor subspace lies within the pre-intervention donor subspace; recall that this is the key condition that enables RSC to generalize to unseen data and produce reliable post-intervention counterfactual estimates. Pleasingly, the post-intervention median R^2-score of 0.84 also supports this claim, i.e., our cross-validation results indicate that RSC is able to accurately reproduce the observed economic trajectories for the donor regions – for reference, we display the predictions associated with three regions (namely, Andalucia, Aragon, and Canarias) in Figures 6.3a, 6.3b, and



(a) Synthetic Andalucia via RSC.

(b) Synthetic Aragon via RSC.

(c) Synthetic Canarias via RSC.

Figure 6.3: Validating RSC: donor Spanish regions unaffected by terrorist activity.

6.3c, respectively, which are representative of our general results. This further validates the counterfactual estimates of RSC for Basque Country.

Intervention	No terrorism
Hypo. Test ($\alpha = 0.05$)	Pass
R^2**-score**	0.84

Table 6.1: Hypothesis test and median R^2-score for RSC for Basque Country case study.

Key Takeaways. Importantly, the RSC model of the target region is always fit in the pre-intervention period. Still, the learnt model is able to accurately reproduce the post-intervention observations (as evidence of the hypothesis test results and cross-validation R^2-scores). This helps validate the RSC framework and the hypothesis test.

■ 6.3.2 California Proposition 99

Another popular case study investigates the impact of California's Proposition 99, an anti-tobacco legislation, on the per-capita cigarette consumption in California (cf. Abadie et al. (2010)). Here, the authors of Abadie et al. (2010) considered California as the target state, the collection of states in the U.S. that did not adopt some variant of a tobacco control program as the donor pool, and Proposition 99 (enacted in 1988) as the intervention. As with the Basque example, we will use the widely accepted counterfactual estimates of Abadie et al. (2010) as our baseline for California and also measure the efficacy of RSC via the placebo (cross-validation) studies.

Empirical Results and Key Takeaways

Robustness of RSC. To motivate the usage of PCR, we first plot the singular value spectrum of the California Prop. 99 dataset, seen in Figure 6.4. Notably, over 99% of the cumulative spectral energy is again captured by the top singular value, which fits the setting under

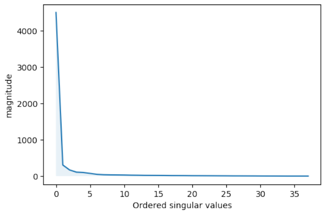
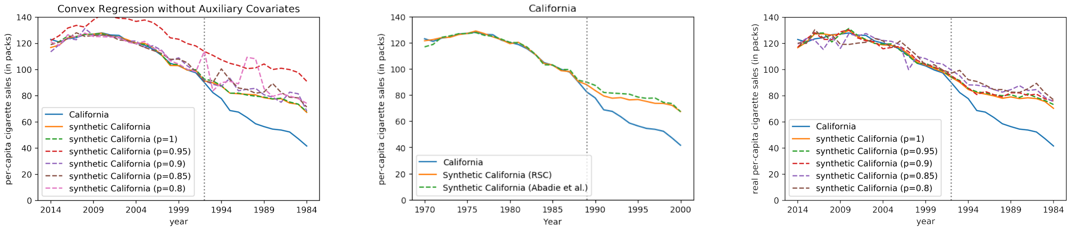


Figure 6.4: Spectrum of California's donor data, which exhibits highly low-dimensional structure.



(a) Synthetic California via Abadie et al. (2010) (without auxiliary covariates) under varying levels of missing data.

(b) Synthetic California via RSC and Abadie et al. (2010).

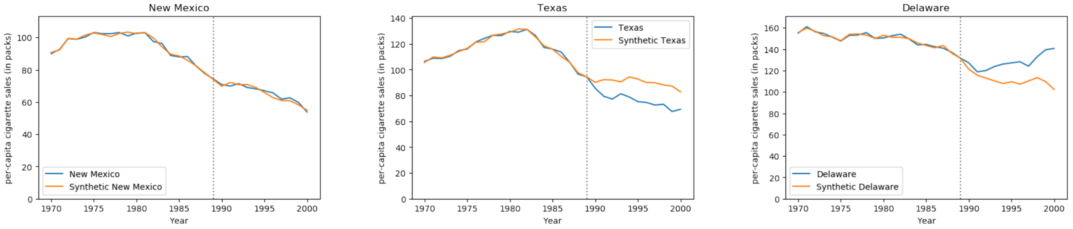
(c) Synthetic California via RSC under varying levels of missing data.

Figure 6.5: Counterfactual estimates of California's cigarette sales in the absence of Prop. 99. While Figure 6.5b demonstrates that RSC and Abadie et al. (2010) (with covariate data) produce similar results when all observations are accessible, Figures 6.5a and 6.5c highlight RSC's robustness to sparsity compared to the classical SC method (when covariate data is withheld).

which our theoretical results apply and motivates the application of PCR.

Further, we plot the resulting synthetic Californias learned via convex regression *without* auxiliary covariates and under varying levels of missing data (5-20%) in Figure 6.5a; similar to the Basque case study, this figure highlights the poor performance of the original SC method in the presence of missing data.

Empirically, we observe that the resulting synthetic California predicted via PCR, also displayed in Figure 6.5b, closely matches the baseline. Much like the previous Basque example, across the varying levels of missing data from 5-20%, the synthetic California per-capita cigarette consumption trajectories continue to mirror the baseline estimates of Abadie et al. (2010); even in the presence of missing data, the counterfactual estimates produced by RSC suggest that Prop. 99 successfully cut smoking in California. This is indeed expected from the theoretical analysis given the extremely low-dimensional structure of the data and the robustness of PCR.



(a) Synthetic New Mexico via RSC.

(b) Synthetic Texas via RSC.

(c) Synthetic Delaware via RSC.

Figure 6.6: Validating RSC: donor states without tobacco control programs (including raised state cigarette taxes).

Placebo Studies: Cross-Validation. In Table 6.2, we show the results of the hypothesis test and median R^2-score across all donor states (for the specific 38 states that were considered donors, please see Section 3.2 of Abadie et al. (2010)). Interestingly, the hypothesis test fails at a significance level of $\alpha = 0.05$, and so we can reject the null hypothesis. This suggests that RSC's synthetic control, which is learned during the pre-intervention period, should not generalize to the post-intervention regime. Correspondingly, the post-intervention median R^2-scores (across all donor states) suffers a low prediction accuracy of -0.58. For reference, we display the predictions associated with three states (namely, New Mexico, Texas, and Delaware) in Figures 6.6a, 6.6b, and 6.6c, respectively, which are representative of our general results. As we can see from these figures, the "counterfactual" estimates for Texas and Delaware do not match the observed trajectories, which is alarming.

Intervention	No Prop. 99
Hypo. Test ($\alpha = 0.05$)	Fail
R^2**-score**	-0.58

Table 6.2: Hypothesis test and prediction accuracy results for RSC in the context of California Proposition 99 study.

Key Takeaways. Although the impact of Prop. 99 on California is a canonical case study within the SC literature, our hypothesis test results suggest that the post-intervention data is "more complex" than the pre-intervention data; hence, a (linear) model learned during the pre-intervention regime should not generalize to the post-intervention regime. Coupled with the low cross-validation R^2-scores, our results possibly indicate that the counterfactual estimates for California (as shown in Figure 6.5b) may not be as reliable as hoped.

■ 6.4 Discussion

■ 6.4.1 Connection to Matrix Completion

We discuss the connection between the SC framework and that of matrix estimation. As discussed, the problem of estimating the unobservable counterfactuals for the target unit can be formulated as recovering a segment of a matrix whose rows correspond to time, columns correspond to units, and entries contain the potential outcomes under the null-intervention. While it may be of interest to de-noise the noisy realizations of our data (i.e., donor observations or pre-intervention target data), our primary concern is to recover the post-intervention counterfactuals, which are never accessible (including the noisy instantiations). Therefore, given the unique sparsity patterns of our data, it is not reasonable to simply impute the missing, counterfactual values via a direct application of standard matrix completion/estimation techniques (e.g., nuclear norm minimization or SVT on the entire dataset with both target and donors). Additionally, from a theoretical standpoint, performance guarantees often only hold with respect to the Frobenius and spectral norms across the entire matrix, and thus are unable to make any statements with respect to recovering the missing segment of interest. Instead, the utility of matrix estimation techniques lies in their ability to de-noise the donor observations, which can be viewed as covariates in the context of supervised learning, to extract the latent signal and assist in the model learning subroutine.

■ 6.4.2 Generalized Factor Models

Here, we consider a generalized factor model, or latent variable model (LVM), which is a natural extension of the linear factor model (a special case of Proposition 4.1.1) typically assumed within the SC literature. Throughout this section, for simplicity and ease of notation, let $\boldsymbol{M} = [M_{tn} : t \leq T, n \leq N]$ with $M_{tn} = M_{tn}^{(1)}$, i.e., $\boldsymbol{M}$ is the matrix of potential outcomes across all units and time under the null-intervention $d = 1$.

More formally, we say $\boldsymbol{M}$ is generated as per a LVM if for all (t, n),

$$M_{tn} = g(u_t, v_n), \tag{6.1}$$

where $u_t \in \mathbb{R}^{p_1}$ and $v_n \in \mathbb{R}^{p_2}$ are latent features that capture time and unit specific information, respectively, for some $p_1, p_2 \geq 1$; and the latent function $g : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}$ captures the model relationship; again, we note that if g is a linear function, then we recover the standard factor model and $\boldsymbol{M}$ is exactly low-rank. More generally, if g is "well-behaved" (e.g., Hölder continuous) and the latent spaces are compact, then Proposition 6.4.1 shows $\boldsymbol{M}$ is *approximately* low-rank.

Establishing (Approximate) Low-rankness

We first define the Hölder class of functions, which is widely adopted in the non-parametric regression literature (see Xu (2017); Tsybakov (2008)). Given a function $g : [0,1)^{p_1} \to \mathbb{R}$, and a multi-index $\kappa \in \mathbb{N}^{p_1}$, let the partial derivate of g at $x \in [0,1)^{p_1}$ (if it exists) be denoted as,

$$\nabla_\kappa g(x) = \frac{\partial^{|\kappa|} g(x)}{(\partial x)^\kappa} \tag{6.2}$$

Definition 6.4.1 (($\alpha, \mathcal{L}$)**-Hölder Class**). *Let $\alpha, \mathcal{L}$ be two positive numbers. The Hölder class $\mathcal{H}(\alpha, \mathcal{L})$ on $[0,1)^{p_1}$*[1] *is defined as the set of functions $g : [0,1)^{p_1} \to \mathbb{R}$ whose partial derivatives satisfy*

$$\sum_{\kappa : |\kappa| = \lfloor \alpha \rfloor} \frac{1}{\kappa!} \left| \nabla_\kappa g(x) - \nabla_\kappa g(x') \right| \leq \mathcal{L} \left\| x - x' \right\|_\infty^{\alpha - \lfloor \alpha \rfloor} \quad \text{for all } x, x' \in [0,1)^{p_1}. \tag{6.3}$$

Here, $\lfloor \alpha \rfloor$ denotes the largest integer strictly smaller than α.

Remark 6.4.1. *Note if $\alpha \in (0,1]$, then (6.3) is $(\alpha, \mathcal{L})$-Lipschitz, i.e.,*

$$\left| g(x) - g(x') \right| \leq \mathcal{L} \left\| x - x' \right\|_\infty^{\alpha - \lfloor \alpha \rfloor} \quad \text{for all } x, x' \in [0,1)^{p_1}.$$

However, for $\alpha > 1$, $(\alpha, \mathcal{L})$-Hölder no longer implies $(\alpha, \mathcal{L})$-Lipschitz.

Proposition 6.4.1 (Hölder-Smoothness Induces Approximate Low-rankness – adapted from Xu (2017)). *Let $\boldsymbol{M}$ satisfy (6.1) with $u_t, v_n \in [0,1)^{p_1}$ as latent parameters. Further, for all v_n, let $g(\cdot, v_n) \in \mathcal{H}(\alpha, \mathcal{L})$, as defined in (6.3). Then for any $\delta > 0$, there exists a low-rank matrix $\boldsymbol{A}$ of rank $r \leq C(\alpha, p_1)\delta^{-p_1}$ such that*

$$\left\| \boldsymbol{M} - \boldsymbol{A} \right\|_{\max} \leq \mathcal{L} \cdot \delta^\alpha. \tag{6.4}$$

Here, $C(\alpha, p_1)$ is a constant that depends only on α and p_1.

Interpretation. In words, Proposition 6.4.1 establishes that if the potential outcomes follow a LVM, then the corresponding potential outcomes matrix $\boldsymbol{M}$ is approximately low-rank.

[1] The domain is easily extended to any compact subset of $\mathbb{R}^{p_1}$.

Additionally, by setting $\delta = 1/(N \wedge T_0)$, it is guaranteed that the approximation error, $\left\|M - A\right\|_{\max}$, vanishes as more data is collected.

Remark 6.4.2. *We remark on the Hölder continuity of a typical linear factor model, i.e., $g(u_t, v_n) = \langle u_t, v_n \rangle$. It is easily seen that such a model satisfies Definition 6.4.1 for all $\alpha \in \mathbb{N}$ and $\mathcal{L} = C$ for some $C > 0$. Thus, one can think of Hölder continuous functions as generalizations of typical linear factor models to (sufficiently smooth) non-linear functions.*

Existence of (Approximate) SC

In what follows, we show that the approximate low-rank property of the underlying potential outcomes matrix implies the existence of an approximate linear synthetic control.

Proposition 6.4.2 (Existence of Approximate SC). *Assume the conditions of Proposition 6.4.1 hold. For intervention $d = 1$ and unit 1, suppose Property 4.1.4 holds for $A = [A_{tn}] = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell}$, where r and A are given by (6.4). Then, there exists a $\beta^* \in \mathbb{R}^{N-1}$ such that for all $t \in [T]$,*

$$M_{t1} = \sum_{n=2}^{N} \beta^*_{n-1} \cdot M_{tn} + \phi_t,$$

where $\phi_t \leq \left(C(\alpha, p_1)\mathcal{L}\left\|\beta^*\right\|_\infty\right) \cdot \delta^{(\alpha - p_1)}$.

Interpretation. Combined with Proposition 4.1.2, Proposition 6.4.2 establishes that (approximate) synthetic controls (almost) always exist under a LVM; additionally, the model mismatch error, ϕ, vanishes as more data is collected. Therefore, in a very general sense, pleasingly, a synthetic control almost always exists and need not be assumed as a hypothesis or axiom.

6.5 Proofs for Generalized Factor Model

6.5.1 Proof of Proposition 6.4.1

As previously stated, the following analysis is adapted from Xu (2017) and is stated here for completeness. Before we dive into the proofs, let us introduce some useful notation.

Definition 6.5.1 (Piecewise Polynomials). *Let $\mathcal{E}$ denote a partition of the cube $[0, 1)^d$ into a finite number ($|\mathcal{E}|$) of cubes Δ. Let $\ell \in \mathbb{N}$. Then $P_{\mathcal{E},\ell} : [0, 1)^d \to \mathbb{R}$ is a piecewise polynomial of degree ℓ if*

$$P_{\mathcal{E},\ell}(x) = \sum_{\Delta \in \mathcal{E}} P_{\Delta,\ell}(x) \cdot \mathbb{1}(x \in \Delta), \tag{6.5}$$

where $P_{\Delta,\ell}(x) : [0,1)^d \to \mathbb{R}$ *denotes a polynomial of degree at most* ℓ.

Proof. We will achieve our result by decomposing the proof into three parts. First, we will discretize the compact latent feature spaces. Then, we will show that g can be well approximated by piecewise polynomials. We conclude the proof by constructing $\boldsymbol{A}$ from these piecewise polynomials, which have been shown to be entry-wise close to $\boldsymbol{M}$, and establish its low-rank structure. For brevity, we will suppress the dependence of g on the latent feature v_n such that $g(u) := g(\cdot, v_n)$.

Step 1: Partitioning the latent spaces. For our purposes, it suffices to consider an equipartition of $[0,1)^{p_1}$. More precisely, for any $\tau \in \mathbb{N}$, we partition $[0,1)$ into $1/\tau$ half-open intervals of length $1/\tau$, i.e., $[0,1) = \cup_{i=1}^{\tau}[(i-1)/\tau, i/\tau)$. It follows that $[0,1)^{p_1}$ can be partitioned into τ^{p_1} cubes of the form $\otimes_{j=1}^{p_1}[(i_j-1)/\tau, i_j/\tau)$ with $i_j \in [\tau]$. Let $\mathcal{E}_n$ be such a partition of $[0,1)^{p_1}$ with $I_1, \ldots, I_{p_1}$ denoting all such cubes and $z_1, \ldots, z_{\tau^{p_1}} \in [0,1)^{p_1}$ denoting the centers of those cubes.

Step 2: Approximating g via piecewise polynomials. Let $\ell = \lfloor \alpha \rfloor$. For every cube I_i with $i \in [\tau^{p_1}]$, we define $P_{I_i,\ell}(u)$ as the degree-ℓ Taylor's series expansion of $g(u)$ centered at z_i:

$$P_{I_i,\ell}(u) = \sum_{\kappa:|\kappa|\leq\ell} \frac{1}{\kappa!}(u-z_i)^{\kappa}\nabla_{\kappa} g(z_i), \tag{6.6}$$

where $\kappa = (\kappa_1, \ldots, \kappa_{p_1})$ is a multi-index with $\kappa! = \prod_{i=1}^{p_1}\kappa_i!$, and $\nabla_{\kappa} g(z_i)$ is the partial derivative defined (6.2) evaluated at z_i. Further, we define a degree-ℓ piecewise polynomial as in (6.5):

$$P_{\mathcal{E}_n,\ell}(u) = \sum_{i=1}^{\tau^{p_1}} P_{I_i,\ell}(u) \cdot \mathbb{1}(u \in I_i), \tag{6.7}$$

where $P_{I_i,\ell}$ is defined as in (6.6).

We are now ready to show that g is well approximated by a piecewise polynomial. To that end, let $z_i' = \theta z_i + (1-\theta)u$ for every $i \in [\tau^{p_1}]$ and some $\theta \in (0,1)$. Since $g(u) \in \mathcal{H}(\alpha, \mathcal{L})$, it follows from Taylor's theorem (using the Lagrange remainder form) that

$$\begin{aligned}
&\sup_u |g(u) - P_{\mathcal{E}_n,\ell}(u)| \\
&= \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} |g(u) - P_{\mathcal{E}_n,\ell}(u)|
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} \left| \sum_{\kappa:|\kappa|<\ell} \frac{\nabla_\kappa g(z_i)}{\kappa!}(u-z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i')}{\kappa!}(u-z_i)^\kappa - P_{\mathcal{E}_n,\ell}(u) \right| \\
&= \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} \left| \sum_{\kappa:|\kappa|\le\ell} \frac{\nabla_\kappa g(z_i)}{\kappa!}(u-z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i') - \nabla_\kappa g(z_i)}{\kappa!}(u-z_i)^\kappa - P_{\mathcal{E}_n,\ell}(u) \right| \\
&= \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i') - \nabla_\kappa g(z_i)}{\kappa!}(u-z_i)^\kappa \right| \\
&\le \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} \left\| u - z_i \right\|_\infty^\ell \cdot \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i') - \nabla_\kappa g(z_i)}{\kappa!} \right| \\
&\le \mathcal{L} \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} \left\| u - z_i \right\|_\infty^\ell \cdot \left\| \theta z_i + (1-\theta)u - z_i \right\|_\infty^{\alpha-\ell} \\
&\le \mathcal{L} \sup_{i\in[\tau^{p_1}]} \sup_{u\in I_i} \left\| u - z_i \right\|_\infty^\alpha = \mathcal{L}\tau^{-\alpha}.
\end{aligned}
$$

Observe that we have used (6.3) to establish the second inequality above.

Step 3: Constructing A and establishing its low-rank structure. We now construct $A = [A_{tn}]$ as follows: for every (t, n), let

$$
A_{tn} = P_{\mathcal{E}_n,\ell}(u_t, v_n),
$$

where $P_{\mathcal{E}_n,\ell}$ is defined as in (6.7). Since $M_{tn} = g(u_t, v_n)$, it follows that

$$
\left\| M - A \right\|_{\max} \le \mathcal{L}\tau^{-\alpha},
$$

which was established in the previous section.

It remains to bound the rank of A. Since $P_{\mathcal{E}_n,\ell}$ is a piecewise polynomial of degree ℓ, it admits the following decomposition:

$$
A_{tn} = \sum_{i=1}^{\tau^{p_1}} \langle \Phi(u_t), \beta_{I_i,v_n} \rangle \cdot \mathbb{1}(u_t \in I_i),
$$

where

$$
\Phi(u_t) = (1, u_{t1}, \dots, u_{tp_1}, \dots, u_{t1}^\ell, \dots, u_{tp_1}^\ell)^T
$$

denotes the collection of all monomials of degree $|\kappa| \le \ell$; and β_{I_i,v_n} denotes the corresponding coefficient vector. Thus, for any fixed u_t, we have that

$$A = \sum_{i=1}^{\tau^{p_1}} \begin{bmatrix} \Phi^T(u_1) \cdot \mathbb{1}(u_1 \in I_i) \\ \vdots \\ \Phi^T(u_T) \cdot \mathbb{1}(u_T \in I_i) \end{bmatrix} \begin{bmatrix} \beta_{I_i, v_1} \dots \beta_{I_i, v_N} \end{bmatrix}.$$

Since there are $C(\alpha, p_1) := \sum_{i=0}^{\ell} \binom{i+p_1-1}{p_1-1}$ degree-ℓ monomials, $\phi(u_t)$ and β_{I_i, v_n} are of dimension at most $C(\alpha, p_1)$. As a result, the rank of A is bounded by $\tau^{p_1} \cdot C(\alpha, p_1)$. Setting $\tau = 1/\delta$ completes the proof. ■

■ 6.5.2 Proof of Proposition 6.4.2

Proof. Let $A = [A_{tn}] \in \mathbb{R}^{T \times N}$ be defined as in Proposition 6.4.1. By appealing to its SVD, A has the following representation: for all (t, n),

$$A_{tn} = \sum_{\ell=1}^{r} u_{t\ell} v_{n\ell}.$$

Therefore, it follows that there exists a β^* with $\|\beta^*\|_0 \leq r$ such that Property 4.1.4 holds; let us define $\mathcal{I} = \{n : \beta^*_n \neq 0\}$ as the support of β^*. This implies that

$$A_{t1} = \sum_{n \in \mathcal{I}} \beta^*_n A_{tn}.$$

Using the above with Proposition 6.4.1, we obtain

$$\begin{aligned} M_{t1} - \sum_{n \in \mathcal{I}} \beta^*_n M_{tn} &= (M_{t1} - \sum_{n \in \mathcal{I}} \beta^*_n A_{tn}) + (\sum_{n \in \mathcal{I}} \beta^*_n (A_{tn} - M_{tn})) \\ &= (M_{t1} - A_{t1}) + (\sum_{n \in \mathcal{I}} \beta^*_n (A_{tn} - M_{tn})) \\ &\leq \mathcal{L} \delta^{\alpha} (1 + r \|\beta^*\|_\infty). \end{aligned}$$

Noting $r \leq C(\alpha, p_1) \delta^{-p_1}$ completes the proof. ■

Chapter 7

Multi-dimensional RSC

7.1 Introduction

As discussed in Chapter 6.2.4, placebo (or cross-validation) studies can be used to evaluate the statistical performance of SC-like methods since ground-truth counterfactuals are never accessible. At the same time, through the lens of time series analysis, the same evaluation measures can be viewed as a forecasting methodology. That is, as long as the temporal or sequential dimension of the data is relative and not absolute, i.e., the donor pool has already undergone the future evolution in "time", SC-like methods can be used to forecast the future evolution for any unit of interest that is unexposed to treatments. However, even though RSC is proven to exhibit attractive theoretical and empirical properties, it (like other SC variants) may still suffer from poor estimation when the amount of training data (i.e., the length of the pre-intervention period) is too small.

Consider the problem of estimating demand in retail. Typically, the amount of data available for the outcome variable of interest is sparse; e.g., given the massive scale of users and products, the observed matrix of transactions has very few nonzero realizations. To avoid overfitting to the idiosyncrasies of the training data, it is commonplace to employ regularization when learning a model (e.g., convex regression a la classical SC or PCR); this reduces the variance of the estimator at the expense of higher bias. Still, algorithmic remedies such as regularization do not enable these methods to generalize if the training data is too small to capture the underlying signal or trend.

In such settings, a data remedy may be the only remaining option. While acquiring more data of the same variable type is frequently infeasible, other types of data (e.g., browse and search histories, responses to promotions to name a few) are often readily available. We refer to this as the problem of estimation (e.g., recovering counterfactuals or forecasting demand) with auxiliary metrics (i.e., data of different types, beyond the outcome variable of interest).

■ 7.1.1 Problem Statement

We are interested in outcomes (e.g., sales) associated with N units (e.g., retail stores) across T measurements (e.g., weeks). Continuing the interventional storyline of Chapter 6, we consider $D = 2$, where (without loss of generality) we denote the "no-intervention" state with $d = 1$ and let $d = 2$ represent an actual intervention (e.g., storewide sale). However, while we are again in the interest of estimating outcomes for a particular metric (e.g., true demand of umbrellas), we now have access to auxiliary metrics in the form of additional metrics (e.g., transactions for clothes), i.e., we let $P \geq 1$.

Observations

Throughout this chapter, we let unit $n = 1$ represent our target unit of interest, which is assumed to receive intervention $d = 2$ during the post-intervention period; thus, $\mathcal{I}^{(2)} = \{1\}$. Meanwhile, all other units form our universe of donors, yielding $\mathcal{I}^{(1)} = [N] \setminus \{1\}$ and $N^{(1)} = N - 1$. We encode our observations into a $T \times N \times 2 \times P$ tensor $\boldsymbol{Z} = [Z_{tn}^{(d,p)}]$, where

$$Z_{tn}^{(d,p)} = \begin{cases} Y_{tn}^{(1,p)} \cdot \pi_{tn}^{(1,p)}, & \text{for all } t \in [T], n > 1, d = 1, p \in [P] \\ Y_{t1}^{(1,p)} \cdot \pi_{t1}^{(1,p)}, & \text{for all } t \leq T_0, n = 1, d = 1, p \in [P] \\ Y_{t1}^{(2,p)} \cdot \pi_{t1}^{(2,p)}, & \text{for all } t > T_0, n = 1, d = 2, p \in [P] \\ \star, \quad \text{otherwise.} \end{cases}$$

Aim

Given $\boldsymbol{Z}$, our goal is to infer the potential outcomes in the *absence of any intervention* for the *target unit* under *metric* p^* during the post-intervention period, i.e., $M_{t1}^{(1,p^*)}$ for all $t > T_0$.

■ 7.2 Multi-dimensional Robust Synthetic Control

As the primary contribution of this chapter, we present multi-dimensional RSC (MRSC), a natural extension of RSC that incorporates auxiliary metrics. In what follows, we will show how MRSC offers a simple, theoretically justified approach in exploiting auxiliary data to overcome the limitations of sparsity and limited training (pre-intervention) data.

■ 7.2.1 Algorithm

As established in Chapter 6, RSC exhibits desirable robustness and generalization properties. Hence, we generalize RSC to utilize multiple metrics in a principled manner.

Notation. Recall that unit 1 is our target unit. In consistency with (4.1), we represent the pre- and post-intervention observation matrices associated with the donors for metric p, which remain in the no-intervention state ($d = 1$) across all T, as

$$\begin{aligned} \boldsymbol{Z}^{(1,p)}_{\text{pre}} &= [Z^{(1,p)}_{tn} : t \leq T_0, n > 1] \in \mathbb{R}^{T_0 \times (N-1)} \\ \boldsymbol{Z}^{(1,p)}_{\text{post}} &= [Z^{(1,p)}_{tn} : t > T_0, n > 1] \in \mathbb{R}^{(T-T_0) \times (N-1)}. \end{aligned}$$

Further, for every p, we denote $y^{(1,p)}_1 = [Y^{(1,p)}_{t1} : t \leq T_0]$ as the corresponding pre-intervention observations for our target unit, which is also observed under $d = 1$. We are now ready to state the MRSC algorithm, which holds for any metric p of interest. For simplicity, we consider estimating the counterfactuals for the target under p^*.

```
Algorithm 5: MRSC
Data: $\{(y_1^{(1,p)}, \boldsymbol{Z}_{\text{pre}}^{(1,p)}) : p \in [P]\}, \boldsymbol{Z}_{\text{post}}^{(1,p^*)}, k, k'$
Result: $\widehat{M}_1^{(1,p^*)} = [\widehat{M}_{t1}^{(1,p^*)} : t > T_0]$
    1. Concatenate:
        (a) donors: $\boldsymbol{Z}_{\text{pre}}^{(1)} \leftarrow [\boldsymbol{Z}_{\text{pre}}^{(1,p)} : p \in [P]] \in \mathbb{R}^{PT_0 \times (N-1)}$
        (b) target: $y_1^{(1)} \leftarrow [y_1^{(1,p)} : p \in [P]] \in \mathbb{R}^{PT_0}$
    2. Learn synthetic target model:
        (a) $\widehat{\beta} \leftarrow \text{PCR}(\boldsymbol{Z}_{\text{pre}}^{(1)}, y_1^{(1)}, k)$
    3. Predict counterfactual prediction outcomes:
        (a) $\widehat{\boldsymbol{M}}_{\text{post}}^{(1,p^*)} \leftarrow \text{HSVT}(\boldsymbol{Z}_{\text{post}}^{(1,p^*)}, k')$
        (b) $\widehat{M}_1^{(1,p^*)} \leftarrow \text{Truncate}(\widehat{\boldsymbol{M}}_{\text{post}}^{(1,p^*)} \widehat{\beta})$
```

Algorithmic Intuition

In words, MRSC simply performs RSC with the added pre-processing procedure of concatenating the pre-intervention donor data across all metrics. Effectively, this step augments the amount of training data.

Weighted Least Squares

MRSC, as stated in Algorithm 5, implicitly assumes that each data type is of equal importance, i.e., the model learning subroutine (PCR) assigns uniform weight to each measurement across all metrics. However, if it is known a priori that certain data types are more similar or important to the primary outcome variable of interest p^*, then Algorithm 5 can be modified to assign greater weights to those data types in the PCR step. More formally, if we denote our weighting matrix as

$$\boldsymbol{W} = \text{diag}\left(\underbrace{\frac{1}{w_1}, \dots, \frac{1}{w_1}}_{T_0}, \dots, \underbrace{\frac{1}{w_P}, \dots, \frac{1}{w_P}}_{T_0} \right) \in \mathbb{R}^{PT_0 \times PT_0},$$

where w_p represents the relative importance of metric p, then we can redefine

$$\widehat{\beta} \leftarrow \text{PCR}(\boldsymbol{W}^{1/2}\boldsymbol{Z}_{\text{pre}}^{(1)}, \boldsymbol{W}^{1/2}y_1^{(1)}, k).$$

In words, the model is now learned via weighted least squares in the reduced subspace spanned by the top principal components of the augmented pre-intervention donor data. We note that $w_1 = \cdots = w_P = 1$ recovers Algorithm 5. In general, the weights can be chosen in a data-driven manner via standard machine learning techniques such as cross-validation.

■ 7.2.2 Existence of SC Across Metrics

Consider an order-three tensor factor model (a simplified version of Property 4.1.1) where the potential outcomes correspond to the universe of units, time, and metrics under the null-intervention $d = 1$. Then Proposition 4.1.1 of Chapter 4 states that an invariant linear model between the target unit and donors holds across all measurements *and* metrics with high probability. This result suggests that the auxiliary metrics (i.e., metrics $p \neq p^*$) can effectively be viewed as additional measurements. Thus, from both a theoretical and algorithmic perspective, this justifies and motivates MRSC to concatenate the auxiliary data (thereby, performing "data augmentation") and learn a *single* linear model $\widehat{\beta}$.

■ 7.2.3 Theoretical Performance Guarantees

Objective

Recall that our aim is to recover the underlying potential outcomes for our target unit $n = 1$ under the null-intervention $d = 1$ associated with metric p^*, i.e., $M_{t1}^{(1,p^*)}$ for all $t > T_0$. As a benefit, we are also given access to auxiliary metrics $(y_1^{(1,p)}, \boldsymbol{Z}^{(1,p)})$ for $p \neq p^*$.

Notation. Throughout the rest of this chapter, let

$$\boldsymbol{M}_{\text{pre}}^{(1)} = [\boldsymbol{M}_{\text{pre}}^{(1,p)} : p \in [P]] \in \mathbb{R}^{PT_0 \times (N-1)}$$

denote the concatenation of potential outcomes under the null-intervention $d = 1$ for the pool of donor units across all metrics; let $r_{\text{pre}} = \text{rank}(\boldsymbol{M}_{\text{pre}}^{(1)})$. Since we are only interested in recovering the counterfactuals under the no-intervention state, we henceforth suppress dependencies on $d = 1$ for ease of notation, e.g., $\boldsymbol{Z}_{\text{pre}}^{(p)} = \boldsymbol{Z}_{\text{pre}}^{(1,p)}$ and $r_{\text{post}}^{(p)} = \text{rank}(\boldsymbol{M}_{\text{post}}^{(1,p)})$, where $\boldsymbol{M}_{\text{post}}^{(1,p)}$ is given by (4.1). Further, let $\beta^* = \beta^{(1,1)}$, where $\beta^{(1,1)}$ is given in (4.3).

Evaluation Metric

We evaluate MRSC based on its post-intervention squared prediction error. Specifically, we define the *post-intervention* error for unit $n = 1$ under the null-intervention $d = 1$ and metric p^* as

$$\mathcal{E}_{\text{post}}(\widehat{M}_1^{(p^*)}) = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} (\widehat{M}_{t1}^{(1,p^*)} - M_{t1}^{(1,p^*)})^2,$$

where $\widehat{M}_1^{(p^*)} = [\widehat{M}_{t1}^{(1,p^*)} : t > T_0] = \text{MRSC}(\{(y_1^{(p)}, \boldsymbol{Z}_{\text{pre}}^{(p)}) : p \in [P]\}, \boldsymbol{Z}_{\text{post}}^{(p^*)}, r_{\text{pre}}, r_{\text{post}}^{(p^*)});$

Post-intervention Prediction Error

We consider the tensor factor model under the null-intervention, a special case of the setting described in Chapter 4. However, rather than assuming a general sub-gaussian noise model (Property 4.1.5), we analyze the i.i.d. Gaussian contamination model instead. This is formalized by the property detailed below.

Property 7.2.1 (Gaussian noise). *Let $\varepsilon_{tn}^{(d,p)}$ be a sequence of independent mean zero Gaussian random variables with* $\text{Var}(\varepsilon_{tn}^{(d,p)}) = \sigma^2$.

We are now ready to present our main results, which bound MRSC's post-intervention prediction error in both high probability and expectation. We relegate the proofs to the end of this chapter.

Theorem 7.2.1 (MRSC Error in High-Probability)**.** *Let Properties 4.1.1, 4.1.2, 4.1.3, 4.1.4, 7.2.1, 4.1.6 hold. Consider the unique β^* of minimum ℓ_2-norm that satisfies* (4.3)*. For any $\delta > 0$ and some $C > 0$, if $\rho \geq \sqrt{\frac{C_1 C_2 r_{\text{pre}}}{N' \wedge T'}}$, then the following holds w.p. at least $1 - \delta$:*

$$\mathcal{E}_{\text{post}}(\widehat{M}_1^{(p^*)}) \leq \frac{r_{\text{pre}}}{r_{\text{post}}^{(p^*)}} \Big(\frac{C\sigma^2 r_{\text{pre}}}{T_0} + \frac{C_1 C_2 r_{\text{pre}} \sqrt{\log N'}}{\rho^4 (N' \wedge T')} \left\|\beta^*\right\|_1^2 + \frac{C_1^2 r_{\text{pre}} N'}{\rho^4 P T_0 (N' \wedge T')} \left\|\beta^*\right\|_2^2 + \Delta \Big),$$

where $N' = N - 1$, $T' = PT_0 \wedge (T - T_0)$,

$$\Delta = \frac{C_2}{\sqrt{PT_0}} \left\|\beta^*\right\|_1, \quad C_1 = C(1 + \sigma^4), \quad C_2 = C_1 \sigma^2 (1 + \log^2(1/\delta)).$$

Corollary 7.2.1 (MRSC Error in Expectation)**.** *Let the conditions of Theorem 7.2.1 hold. Then for any $\delta > 0$,*

$$\mathbb{E}[\mathcal{E}_{\text{post}}(\widehat{M}_1^{(p^*)})] \leq \frac{r_{\text{pre}}}{r_{\text{post}}^{(p^*)}} \Big(\frac{2\sigma^2 r_{\text{pre}}}{T_0} + \frac{C_3 C_4 r_{\text{pre}} \log^2(N'/\delta)}{\rho^4 (N' \wedge T')} \left\|\beta^*\right\|_1^2 + \frac{C_4^2 r_{\text{pre}} N'}{\rho^4 P T_0 (N' \wedge T')} \left\|\beta^*\right\|_2^2 \Big) + 4\delta,$$

where $C_3 = C(1 + \sigma^6)$ and $C_4 = C(1 + \sigma^4)$.

Interpretation. The impact of auxiliary metrics is made precise by the dependence on P. If $P = 1$, then we return to setting of Chapter 6 (without access to auxiliary metrics) and recover Theorem 6.2.1. However, for any $P > 1$, the generalization error (third term of Theorem 7.2.1) decreases linearly with P, the total number of metrics used in the model learning procedure.

To gain greater intuition, let $T_0 = \Theta(N) = \Theta(T)$. In such a setting, observe that the first two terms of Corollary 7.2.1, ignoring log factors, decay linearly with T_0. This suggests that the benefit of auxiliary metrics can only reduce the overall testing prediction error up to a certain point, irrespective of the amount of additional information. Hence, the benefit of utilizing auxiliary metrics is to help alleviate the problem of sparsity, as desired. More specifically, as opposed to requiring on the order of r_{pre} entries per sample in our training set, we may now only need to observe r_{pre}/P entries per sample. This establishes a trade-off in data acquisition; specifically, trading off sparsity of one type of data for data of a different type.

■ 7.3 Empirical Case Studies

■ 7.3.1 Forecasting in Retail

We consider the problem of forecasting weekly sales in retail. Here, we highlight a key utility of MRSC over RSC in the presence of sparse data. More specifically, our results demonstrate that when the pre-intervention period (training set) is short, then standard RSC methods fail to generalize well. On the other hand, by using auxiliary information from other metrics, MRSC effectively "augments" the training data, which allows it to overcome the difficulty of extrapolating from small sample sizes.

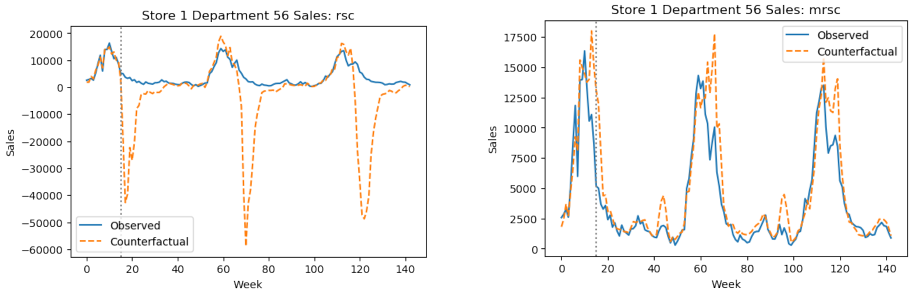
Experimental Setup

We consider the Walmart dataset, which contains $T = 143$ weekly sales information across $N = 45$ stores and $P = 81$ departments. We arbitrarily choose store one as the treatment unit, and introduce an "artificial" intervention at various points; this is done to study the effect of the pre-intervention period length on the predictive power for both MRSC and RSC methods. In particular, we consider the following pre-intervention points to be $15, 43$, and 108 weeks, representing small to large pre-intervention periods (roughly $10\%, 30\%$, and 75% of the entire time horizon T, respectively). Further, we consider three department subsets (representing three different metric subgroups): Departments $\{2, 5, 6, 7, 14, 23, 46, 55\}$, $\{17, 21, 22, 32, 55\}$, and $\{3, 16, 31, 56\}$.

Empirical Results

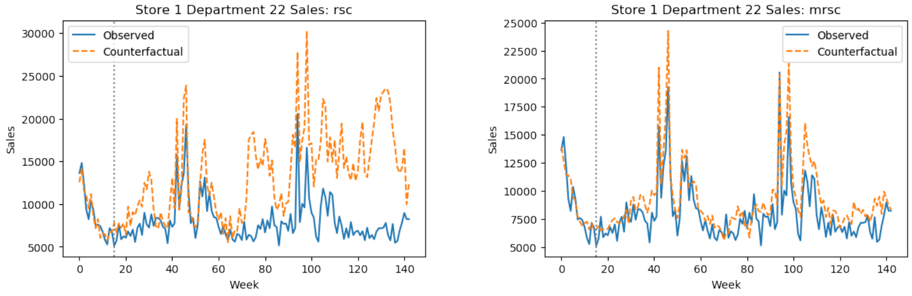
In Table 7.1, we show the effect of the pre-intervention length on the RSC and MRSC's ability to forecast. In particular, we compute the average pre-intervention (training) and post-intervention (testing) MSEs across each of the three departmental subgroups (as described above) for both methods and for varying pre-intervention lengths. Although the RSC method consistently achieves a smaller average pre-intervention error, the MRSC consistently outperforms the RSC method in the post-intervention regime, especially when the pre-intervention stage is short. This is in line with our theoretical findings of the post-intervention error behavior, as stated in Theorem 7.2.1 and Corollary 7.2.1; i.e., the benefit of incorporating multiple relevant metrics is exhibited by the MRSC algorithm's ability to generalize in the post-intervention regime despite high levels of sparsity.

We present Figures 7.1 and 7.2 to highlight two settings, departments 56 (left) and 22 (right), respectively, in which MRSC drastically outperforms RSC in extrapolating from a small training set ($T_0 = 15$ weeks). We highlight that the weekly sales axes between the subplots for each department, particularly department 56, are different; indeed, since the RSC algorithm was given such little training data, the RSC algorithm predicted negative sales values for department 56 and, hence, we have used different sales axes ranges to underscore the prediction quality gap between the two methods. As seen from these plots, the RSC method struggles to extrapolate beyond the training period since the pre-intervention period is short. In general, the RSC method compensates for lack of data by overfitting to the pre-intervention observations and, thus, misinterpreting noise for signal (as seen also by the smaller pre-intervention error in Table 7.1). Meanwhile, the MRSC overcomes this challenge by incorporating sales information from other departments. By effectively augmenting the pre-intervention period, MRSC becomes robust to sparse data. However, it is worth noting that both methods are able to extrapolate well in the presence of sufficient data.



(a) Dept. 56 (RSC) **(b)** Dept. 56 (MRSC)

Figure 7.1: MRSC and RSC forecasts for department 56 of store 1 using $T_0 = 15$ weeks.



(a) Dept. 22 (RSC) **(b)** Dept. 22 (MRSC)

Figure 7.2: MRSC and RSC forecasts for department 22 of store 1 using $T_0 = 15$ weeks.

	Train Error (10^6)		Test Error (10^6)	
T_0	RSC	RSC	RSC	MRSC
10%	**1.54**	3.89	21.0	**5.25**
30%	**2.21**	3.51	19.4	**4.62**
75%	**4.22**	5.33	3.32	**2.48**
10%	**0.67**	2.61	14.4	**2.48**
30%	**0.79**	1.21	2.13	**1.97**
75%	**1.18**	2.78	1.31	**0.77**
10%	**1.28**	6.10	84.6	**12.5**
30%	**2.60**	3.45	**3.72**	4.13
75%	**2.29**	2.65	4.92	**4.72**

Table 7.1: Average pre-intervention (train) and post-intervention (test) MSE for RSC and MRSC methods.

■ 7.4 Discussion

■ 7.4.1 Connection to Matrix & Tensor Completion

Much like the discussion in Chapter 6.4.1, we formalize our problem as recovering a segment of a matrix whose rows and columns correspond to time and units, respectively, and entries contain the potential outcomes in the absence of any intervention. Through this lens, we view side information as additional matrices of conforming dimension. Together with the primary matrix of interest (corresponding to what we often referred to as metric p^*), we can encode our data into an order-three tensor, where each frontal slice corresponds to a unique data type or metric. Since we hope to recover a specific segment of the frontal tensor slice, standard matrix and tensor estimation techniques (even those that incorporate side information a la Farias and Li (2019)) are again insufficient for our purposes. MRSC, on the other hand, is a natural consequence of this tensor perspective (particularly, under a tensor factor model), and is well-suited for our setting of interest.

■ 7.5 Proofs

Throughout, we adopt the notation established in Section 5.7 of Chapter 5 with the modification that the number of training samples is now kn for some $k \geq 1$.

■ 7.5.1 Preservation of Gaussians

Lemma 7.5.1. *Let $\boldsymbol{A}$ be an $m \times n$ matrix whose rows A_i are independent, mean-zero, isotropic Gaussian random vectors in $\mathbb{R}^n$, i.e., $A_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n \times n})$. Let $\boldsymbol{U} \in \mathbb{R}^{m \times k_1}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times k_2}$ be matrices with orthonormal columns. Then, $\boldsymbol{A}' = \boldsymbol{U}^T \boldsymbol{Q} \boldsymbol{V}$ is a $k_1 \times k_2$ matrix whose rows A'_i are independent, mean-zero, isotropic Gaussian random vectors in $\mathbb{R}^{k_2}$, i.e., $A'_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{k_2 \times k_2})$.*

Proof. To begin, consider the matrix $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{V} \in \mathbb{R}^{m \times k_2}$. Let $X_i = \langle A_i, \boldsymbol{V} \rangle$ denote the i-th row of $\boldsymbol{X}$. Then, it follows that $\mathbb{E}[X_i] = 0$ and

$$\text{cov}(X_i) = \boldsymbol{V}^T \text{cov}(A_i) \boldsymbol{V} = \sigma^2 \boldsymbol{I}_{k_2 \times k_2}.$$

Thus, the rows of $\boldsymbol{X}$ are independent, mean-zero, isotropic Gaussian random vectors in $\mathbb{R}^{k_2}$ distributed as $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_{k_2 \times k_2})$.

Now, consider the matrix $\boldsymbol{Y} = \boldsymbol{U}^T \boldsymbol{A} \in \mathbb{R}^{k_1 \times n}$. Let $A_j \in \mathbb{R}^m$ denote the j-th column of $\boldsymbol{A}$, and let $Y_j = \boldsymbol{U}^T A_j \in \mathbb{R}^{k_1}$ denote the j-th column of $\boldsymbol{Y}$. By assumption, it follows that the entries of A_j are independent Gaussian random variables with mean zero and variance σ^2. Hence, it follows that $\mathbb{E}[Y_j] = 0$ and

$$\text{cov}(Y_j) = \boldsymbol{U}^T \text{cov}(A_j) \boldsymbol{U} = \sigma^2 \boldsymbol{I}_{n \times n}.$$

Since uncorrelation implies independence for Gaussian distributions, the rows of $\boldsymbol{Y}$ are independent, mean-zero, isotropic Gaussian random vectors in $\mathbb{R}^n$ distributed as $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n \times n})$.

Observing that $\boldsymbol{A}' = \boldsymbol{Y}\boldsymbol{V} = \boldsymbol{U}^T \boldsymbol{X}$ completes the proof. ■

■ 7.5.2 Learning Subspaces

Below, we state an alternative version of Lemma 5.11.1 in Lemma 7.5.2, which provides a sharper bound under the i.i.d. Gaussian noise assumption (Property 7.2.1).

Notation. Let $\boldsymbol{U}_{M\perp} \in \mathbb{R}^{n \times (n-r)}$ and $\boldsymbol{V}_{M\perp} \in \mathbb{R}^{p \times (p-r)}$ denote the orthogonal complements to $\boldsymbol{U}_M$ and $\boldsymbol{V}_M$, respectively. Further, letting $\tilde{\boldsymbol{H}} = \boldsymbol{Y} - \rho \boldsymbol{M}$, we write

$$\tilde{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{U}_M & \boldsymbol{U}_{M\perp} \end{bmatrix} \cdot \begin{bmatrix} \tilde{\boldsymbol{H}}_{11} & \tilde{\boldsymbol{H}}_{12} \\ \tilde{\boldsymbol{H}}_{21} & \tilde{\boldsymbol{H}}_{22} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{V}_M^T \\ \boldsymbol{V}_{M\perp}^T \end{bmatrix}, \tag{7.1}$$

where

$$\begin{aligned}\tilde{\boldsymbol{H}}_{11} &= \boldsymbol{U}_M^T \tilde{\boldsymbol{H}} \boldsymbol{V}_M, \quad \tilde{\boldsymbol{H}}_{12} = \boldsymbol{U}_M^T \tilde{\boldsymbol{H}} \boldsymbol{V}_{M\perp}, \\ \tilde{\boldsymbol{H}}_{21} &= \boldsymbol{U}_{M\perp}^T \tilde{\boldsymbol{H}} \boldsymbol{V}_M, \quad \tilde{\boldsymbol{H}}_{22} = \boldsymbol{U}_{M\perp}^T \tilde{\boldsymbol{H}} \boldsymbol{V}_{M\perp}\end{aligned}$$

are matrices of dimensions $r \times r, r \times (p-r), (n-r) \times r$, and $(n-r) \times (p-r)$, respectively.

Lemma 7.5.2. *Suppose Property 4.1.1 holds, and* $\widehat{\boldsymbol{M}} = \text{HSVT}(\boldsymbol{Y}, r)$*. Then,*

$$\left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\| \le \frac{\left(\rho s_r + \left\|\tilde{\boldsymbol{H}}_{11}\right\|\right)\left\|\tilde{\boldsymbol{H}}_{12}\right\| + \left\|\tilde{\boldsymbol{H}}_{22}\right\| \cdot \left\|\tilde{\boldsymbol{H}}_{21}\right\|}{\left(\rho s_r - \left\|\tilde{\boldsymbol{H}}_{11}\right\|\right)^2 - \left\|\tilde{\boldsymbol{H}}_{22}\right\|^2 - \left(\left\|\tilde{\boldsymbol{H}}_{21}\right\|^2 \wedge \left\|\tilde{\boldsymbol{H}}_{12}\right\|^2\right)},$$

where $\tilde{\boldsymbol{H}}_{11}, \tilde{\boldsymbol{H}}_{12}, \tilde{\boldsymbol{H}}_{21}$*, and* $\tilde{\boldsymbol{H}}_{22}$ *are defined in* (7.1).

Proof. Since $\text{rank}(\boldsymbol{M}) = r$, it follows that

$$\boldsymbol{U}_M^T \boldsymbol{Y} \boldsymbol{V}_M = \boldsymbol{U}_M^T(\rho \boldsymbol{M} + \tilde{\boldsymbol{H}})\boldsymbol{V}_M = \rho \boldsymbol{S}_M + \boldsymbol{U}_M^T \tilde{\boldsymbol{H}} \boldsymbol{V}_M = \rho \boldsymbol{S}_M + \tilde{\boldsymbol{H}}_{11}.$$

Applying Weyl's inequality (Lemma 3.1.1), we obtain

$$\rho s_r - \left\|\tilde{\boldsymbol{H}}_{11}\right\| \le \alpha \le \rho s_r + \left\|\tilde{\boldsymbol{H}}_{11}\right\|,$$

where α is as defined in (3.5). Similarly, since $\boldsymbol{U}_{M\perp}$ is the orthogonal complement of $\boldsymbol{U}_M$ (equivalently, $\boldsymbol{V}_{M\perp}$ is orthogonal to $\boldsymbol{V}_M$), it holds that

$$\boldsymbol{U}_{M\perp}^T \boldsymbol{Y} \boldsymbol{V}_{M\perp} = \boldsymbol{U}_{M\perp}^T(\rho \boldsymbol{M} + \tilde{\boldsymbol{H}})\boldsymbol{V}_{M\perp} = \boldsymbol{U}_{M\perp}^T \tilde{\boldsymbol{H}} \boldsymbol{V}_{M\perp} = \tilde{\boldsymbol{H}}_{22},$$

which yields $\beta = \left\|\tilde{\boldsymbol{H}}_{22}\right\|$, where β is also defined as in (3.5).

By the construction of $\widehat{\boldsymbol{M}}$, $\mathcal{P}_{\widehat{V}_M}$ is an orthogonal projection onto the subspace spanned by the top r right singular vectors of $\boldsymbol{Y}$. Therefore, Theorem 3.1.2 gives

$$\left\|\sin\Theta(\widehat{\boldsymbol{V}}_M, \boldsymbol{V}_M)\right\| \le \frac{\left(\rho s_r + \left\|\tilde{\boldsymbol{H}}_{11}\right\|\right)\left\|\tilde{\boldsymbol{H}}_{12}\right\| + \left\|\tilde{\boldsymbol{H}}_{22}\right\| \cdot \left\|\tilde{\boldsymbol{H}}_{21}\right\|}{\left(\rho s_r - \left\|\tilde{\boldsymbol{H}}_{11}\right\|\right)^2 - \left\|\tilde{\boldsymbol{H}}_{22}\right\|^2 - \left(\left\|\tilde{\boldsymbol{H}}_{21}\right\|^2 \wedge \left\|\tilde{\boldsymbol{H}}_{12}\right\|^2\right)}.$$

This completes the proof. ■

■ 7.5.3 Proof of Theorem 7.2.1

Proof. We follow the proof of Theorem 5.3.2 with a slight modification. In particular, Lemma 7.5.1 states that the entries of $\tilde{H}_{12}$, given in (7.1), are independent, mean-zero Gaussian random variables with variance σ^2; thus, for any $\delta > 0$ and some $C > 0$, Theorem 3.2.2 states that with probability at least $1 - \delta$,

$$\left\|\tilde{H}_{12}\right\| \le \sqrt{C_1}\left(\sqrt{r} + \sqrt{p} + \sqrt{\log(1/\delta)}\right),$$

where $C_1 = C(1 + \sigma^4)$. Using the above inequality with Lemma 7.5.2, we modify (5.36) to obtain

$$\Lambda_M^2 \le \frac{C_1}{\rho^4}\left(\frac{r}{kn} + \frac{r\log(1/\delta)}{knp}\right).$$

Plugging the above into (5.37) yields

$$\{\text{term } 4 + \text{term } 5\} \le \frac{C_1^2}{\rho^4}\left(\frac{r}{kn}\left(1 + \frac{p}{m} + \frac{r}{r'kn}\right) + \frac{r\log^2(1/\delta)}{r'Knp}\right)\|\beta^*\|_2^2.$$

Therefore, it follows that (5.39) becomes

$$\Delta_{\text{gen}} \le \frac{C_1^2}{\rho^4}\frac{r}{r'}\left(\frac{rp}{kn(kn \wedge m \wedge p)} + \frac{r\log^2(1/\delta)}{kn(kn \wedge m \wedge p)}\right)\|\beta^*\|_2^2.$$

Collecting and simplifying the above results gives the following:

$$\mathcal{E}_{\text{test}}(\widehat{M}_{\text{test}}) \le \frac{r}{r'}\left(\frac{C\sigma^2 r}{n} + \frac{C_1C_2}{\rho^4}\frac{r\sqrt{\log p}}{kn \wedge m \wedge p}\|\beta^*\|_1^2 + \frac{C_1^2}{\rho^4}\frac{rp}{kn(kn \wedge m \wedge p)}\|\beta^*\|_2^2 + \Delta_1\right),$$

where $C_2 = C_1\sigma^2(1 + \log^2(1/\delta))$ and Δ_1 is given by (5.3). The proof is complete after relabeling constants.

■

Chapter 8

Synthetic Interventions

■ 8.1 Introduction

As the COVID-19 pandemic began to rapidly spread within the United States (U.S.), the U.S. government responded by implementing policies to enforce social distancing. Unfortunately, these policies only led to a less than 5% reduction in mobility goo, and many lives were tragically lost. This begs the questions: Would greater reductions in mobility, say 30% or 60%, have led to significantly better societal health outcomes? And moving forward, what trade-offs between health outcomes and economic impact can be achieved through different policies? Although it is infeasible to answer either question through actual experimentation, it is possible to leverage information from across the globe. Given that different regions and/or countries have implemented various policies, valuable observation data is readily available and can be used to answer these questions.

■ 8.1.1 Problem Statement

We are interested in outcomes (e.g., COVID-19 death counts) associated with N units (e.g., countries) across T measurements (e.g., days) and D possible interventions (e.g., different mobility restriction interventions). For simplicity, we consider the single metric case ($K = 1$) and suppress all dependencies on the metric. As before, we also denote the control or "no-intervention" state with $d = 1$ (e.g., no mobility restriction enacted).

Observations

We encode our observations into a $T \times N \times D$ tensor $\boldsymbol{Z} = [Z_{tn}^{(d)}]$, where

$$Z_{tn}^{(d)} = \begin{cases} Y_{tn}^{(1)} \cdot \pi_{tn}^{(1,p)}, & \text{for all } t \leq T_0, n, d = 1, p \in [P] \\ Y_{tn}^{(d)} \cdot \pi_{tn}^{(d,p)}, & \text{for all } t > T_0, n \in \mathcal{I}^{(d)}, d \in [D], p \in [P] \\ \star, & \text{otherwise;} \end{cases}$$

here, $\mathcal{I}^{(d)}$ is given by (1.1). In words, all units are under control during the pre-intervention phase ($t \leq T_0$), but is exposed to some intervention or remains under control during the post-intervention phase ($t > T_0$).

Aim

Given $\boldsymbol{Z}$, our goal is to infer the potential outcomes under *all* interventions for *every* unit during the post-intervention period, i.e., $M_{tn}^{(d)}$ for all n, d, and $t > T_0$.

8.1.2 Synthetic Control (SC), A Partial Solution

Recall that SC Abadie and Gardeazabal (2003); Abadie et al. (2010) provides a solution for a restricted setting: $D = 2$ with $\mathcal{I}^{(1)} = [N] \setminus \{1\}$, $\mathcal{I}^{(2)} = \{1\}$ (see Chapter 6 for details). That is, SC can only infer outcomes for unit 1 in the *absence* of any intervention, i.e., $M_{t1}^{(1)}$ for $t > T_0$. In our COVID example, this corresponds to the situation where the U.S. (i.e., unit 1 in this case) implemented a mobility restriction of less than 5% while all other countries did *nothing*. Using such observations, SC can only estimate the number of deaths in the U.S. if it had done *nothing* to combat COVID-19. Therefore, SC provides an incomplete answer to the COVID-19 question laid out above.

8.2 Synthetic Interventions (SI), A Complete Solution

In order to quantify the trade-offs of different *policies* before having to enact them, we need to estimate potential outcomes under *treatment*, as opposed to only *control*. As the primary contribution of this chapter and thesis, we introduce synthetic interventions (SI), which provides a solution to this important open problem. In short, SI estimates the potential outcome under control and *every* treatment of interest for *every* unit.

8.2.1 Algorithm

Methodologically, SI pleasingly turns out to be straightforward extension of SC, making it easy to implement. For a graphical depiction of the input and output of SI, please refer to Figure 1.4.

Notation. Suppose unit i is our target unit. In consistency with (4.1), we represent the pre- and post-intervention observation matrices associated with donors that receive d as

$$\boldsymbol{Z}_{\text{pre}}^{(d)} = [Z_{tn}^{(1)} : t \leq T_0, n \in \mathcal{I}^{(d)}] \in \mathbb{R}^{T_0 \times N^{(d)}}$$

$$\boldsymbol{Z}^{(d)}_{\text{post}} = [Z^{(d)}_{tn} : t > T_0, n \in \mathcal{I}^{(d)}] \in \mathbb{R}^{(T-T_0)\times N^{(d)}}.$$

We note that if unit i receives d, then we define $\mathcal{I}^{(d)} := \mathcal{I}^{(d)} \setminus \{i\}$, $N^{(d)} := \left|\mathcal{I}^{(d)} \setminus \{i\}\right|$. Finally, let $y^{(1)}_i = [Y^{(1)}_{ti} : t \leq T_0]$ denote the pre-intervention observations for our target unit i, which is observed under $d = 1$. We are now ready to state the SI algorithm in Algorithm 6, which holds for every unit i and intervention d of interest.

```
Algorithm 6: SI
  Data: $y^{(1)}_i, \boldsymbol{Z}^{(d)}_{\text{pre}}, \boldsymbol{Z}^{(d)}_{\text{post}}, k, k'$
  Result: $\widehat{M}^{(d)}_i = [\widehat{M}^{(d)}_{ti} : t > T_0]$
     1. Learn synthetic target model:
        (a) $\widehat{\beta}^{(d,i)} \leftarrow \text{PCR}(\boldsymbol{Z}^{(d)}_{\text{pre}}, y^{(1)}_i, k)$
     2. Predict counterfactual prediction outcomes:
        (a) $\widehat{\boldsymbol{M}}^{(d)}_{\text{post}} \leftarrow \text{HSVT}(\boldsymbol{Z}^{(d)}_{\text{post}}, k')$
        (b) $\widehat{M}^{(d)}_i \leftarrow \text{Truncate}(\widehat{\boldsymbol{M}}^{(d)}_{\text{post}}\widehat{\beta}^{(d,i)})$
```

Algorithmic Intuition

In words, much like (M)RSC (see Algorithms 4 and 5), SI first builds a synthetic version of the target unit i, represented by $\widehat{\beta}^{(d,i)}$, using the donors within $\mathcal{I}^{(d)}$ via PCR. Once the model is learned, SI rescales the observed outcome variables associated with the donor units within $\mathcal{I}^{(d)}$ during the post-intervention period according to $\widehat{\beta}^{(d,i)}$. Effectively, the re-scaling subroutine performs a synthetic intervention d for the target unit i and provides the corresponding counterfactual potential outcomes under such a setting. Importantly, we highlight that when $d = 1$, the model, $\widehat{\beta}^{(d,i)}$, effectively represents a synthetic control; however, for any $d \neq 1$ (corresponding to an actual intervention), this model now represents a synthetic treatment group.

Incorporating Auxiliary Metrics

SI can incorporate auxiliary data types by simply concatenating the additional measurements a la the pre-processing step of MRSC (given in Algorithm 5). The rest of the algorithm flows as described above.

■ 8.2.2 Existence of SI

Though SI is methodologically similar to SC in terms of learning a model to estimate counterfactual outcomes, it is conceptually significantly different. Specifically, as in SC, the model in SI is learnt using pre-intervention data under the no-intervention ($d = 1$) setting; however, to produce post-intervention counterfactual estimates, SI now applies the learnt model to *any* intervention d. A priori, it is not clear why the model can be *transferred* between interventions.

However, as stated by Proposition 4.1.1 of Chapter 4, our proposed tensor factor model (Property 4.1.1) implies that an invariant linear model between any target unit i and subgroup of donors $\mathcal{I}^{(d)}$ persists across both measurements and interventions whp – indeed, this is the key result that justifies SI. Algorithmically, it allows us to learn a linear model under any intervention framework during the pre-intervention period and then transfer the learned model to any other intervention framework during the post-intervention period.

■ 8.2.3 Theoretical Performance Guarantees

Objective

We state Theorem 8.2.1 and Corollary 8.2.1, which hold for *all* interventions d and units n. Thus, for simplicity and ease of notation, we restrict our attention to estimating the post-intervention counterfactual potential outcomes under a specific intervention d and for unit $n = 1$, i.e., we aim to recover $M_{t1}^{(d)}$ for all $t > T_0$.

Notation. Given the above, we suppress dependencies on d for ease of notation. Instead, to distinguish between the pre- and post-intervention data (corresponding to no-intervention and intervention d, respectively), we make explicit their dependencies through appropriate subscripts, e.g., $\boldsymbol{Z}_{\text{pre}} = \boldsymbol{Z}_{\text{pre}}^{(d)}$ and $r_{\text{post}} = \text{rank}(\boldsymbol{M}_{\text{post}}^{(d)})$, where $\boldsymbol{M}_{\text{post}}^{(d)}$ is given by (4.1). Further, let $\beta^* = \beta^{(d,1)}$, where $\beta^{(d,1)}$ is given in (4.3). To avoid confusion, we do not alter $N^{(d)}$.

Evaluation Metric

We evaluate SI based on its post-intervention squared prediction error. Specifically, we define the *post-intervention* (or *test*) error for unit $n = 1$ under intervention d as

$$\mathcal{E}_{\text{post}}(\widehat{M}_1) = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} (\widehat{M}_{t1}^{(d)} - M_{t1}^{(d)})^2,$$

where $\widehat{M}_1 = [\widehat{M}_{t1}^{(d)} : t > T_0] = \text{SI}(y_1, \boldsymbol{Z}_{\text{pre}}, \boldsymbol{Z}_{\text{post}}, r_{\text{pre}}, r_{\text{post}})$.

Remark. Although Property 4.1.5 assumes independent noise entries, our results are stated when ε_{tn} can be dependent across donors for a given t, i.e., only the target and donor noise must remain independent.

Post-intervention Prediction Error

We now state the post-intervention counterfactual prediction errors of SI under the setting described in Chapter 4.

Theorem 8.2.1 (SI Error in High-Probability). *Let Properties 4.1.1, 4.1.2, 4.1.3, 4.1.4, 4.1.5, 4.1.6 hold. Consider the unique β^* of minimum ℓ_2-norm that satisfies (4.3). For any $\delta > 0$ and some $C > 0$, if $\rho \geq \sqrt{\frac{C_1 C_2 r_{\text{pre}}}{N^{(d)} \wedge T'}}$, then the following holds w.p. at least $1 - \delta$:*

$$\mathcal{E}_{\text{post}}(\widehat{M}_1) \leq \frac{r_{\text{pre}}}{r_{\text{post}}} \left(\frac{C\sigma^2 r_{\text{pre}}}{T_0} + \frac{C_1 C_2 r_{\text{pre}} \sqrt{\log N^{(d)}}}{\rho^4 (N^{(d)} \wedge T')} \left\|\beta^*\right\|_1^2 + \frac{C_1^2 r_{\text{pre}} N^{(d)}}{\rho^4 (N^{(d)} \wedge T')^2} \left\|\beta^*\right\|_2^2 + \Delta \right),$$

where $T' = T_0 \wedge (T - T_0)$,

$$\Delta = \frac{C_2}{\sqrt{T_0}} \left\|\beta^*\right\|_1, \quad C_1 = C(1 + \sigma^4)(1 + \gamma^2)(1 + K^2), \quad C_2 = C_1 K^2 (1 + \log^2(1/\delta)).$$

Proof. The result is immediate from Theorem 5.3.2. ■

Corollary 8.2.1 (SI Error in Expectation). *Let the conditions of Theorem 8.2.1 hold. Then for any $\delta > 0$,*

$$\mathbb{E}[\mathcal{E}_{\text{post}}(\widehat{M}_1)] \leq \frac{r_{\text{pre}}}{r_{\text{post}}} \left(\frac{2\sigma^2 r_{\text{pre}}}{T_0} + \frac{C_3 C_4 r_{\text{pre}} \log^2(N^{(d)}/\delta)}{\rho^4 (N^{(d)} \wedge T')} \left\|\beta^*\right\|_1^2 + \frac{C_4^2 r_{\text{pre}} N^{(d)}}{\rho^4 (N^{(d)} \wedge T')^2} \left\|\beta^*\right\|_2^2 \right) + 4\delta,$$

where $C_3 = CK^2(1 + \sigma^4)(1 + \gamma^2)(1 + K^2)$ and $C_4 = C(1 + \sigma^2)(1 + \gamma^2)(1 + K^2)$.

Proof. The result is immediate from Corollary 5.3.1. ■

Intepretation. For simplicity, let $T_0 = \Theta(N^{(d)}) = \Theta(T)$. Ignoring log factors, Corollary 8.2.1 states that the post-intervention error decays linearly with T_0, in expectation. Again, we highlight that Theorem 8.2.1 and Corollary 8.2.1 *do not make any distributional assumptions*. While standard generalization error analyses anchor on i.i.d. data generating assumptions, we skirt such an assumption as potential outcomes from different interventions are likely to come from different distributions. Instead, we rely on a linear algebraic condition, which can be verified in a data-driven manner in practice, as described in Section 5.4 of Chapter 5.

Finally, it is not a coincidence that our theoretical performance guarantees for SI closely match that of RSC (see Theorem 6.2.1 and Corollary 6.2.1). Indeed, this phenomena reflects the fact that SI mirrors RSC methodologically, and the only significant departure of SI from RSC (and SC-like methods in general) is purely conceptual in nature.

■ 8.3 Empirical Case Studies

We extensively test the validity and widespread applicability of SI on real-world data. In particular, we consider four case studies: (i) analyzing the impact of mobility-restricting interventions in mitigating the COVID-19 pandemic with observational data; (ii) exploring the effect of different discount strategies to increase user engagement in an A/B testing framework for a large e-commerce company; (iii) studying how 20 different interventions affected immunization rates in Haryana, India as part of a large developmental economics study Banerjee et al. (2018) with RCT data; (iv) investigating the effect of drug therapies on cells in in-vitro studies with experimental data. Our results indicate that SI can not only be useful in guiding policy-makers as they weigh the trade-offs of different policy interventions, but also in performing personalized, data-efficient randomized control trials and drug discovery.

Quantifying Counterfactual Prediction Accuracy

To quantify the accuracy of the counterfactual predictions produced by SI, we need meaningful baselines to compare against. To that end, we define $R^2_{\text{rct}} = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{rct}}}$, where

$$\text{SS}_{\text{res}} = \sum_{t=T_0+1}^{T} (Y^{(d)}_{t1} - \widehat{M}^{(d)}_{t1})^2, \quad \text{SS}_{\text{rct}} = \sum_{t=T_0+1}^{T} (Y^{(d)}_{t1} - Y^{(d)}_{t,\text{rct}})^2, \quad Y^{(d)}_{t,\text{rct}} = \frac{1}{N^{(d)}} \sum_{n \in \mathcal{I}^{(d)}} Y^{(d)}_{tn}.$$

Interpretation. $Y^{(d)}_{t,\text{rct}}$ is the average outcome across all donors that experienced intervention d at time t. If the units were homogeneous (i.e., they all reacted identically to each intervention), then $Y^{(d)}_{t,\text{rct}}$ will be a good predictor of the counterfactual outcome for the target unit, i.e. $Y^{(d)}_{t,\text{rct}} \approx \widehat{M}^{(d)}_{t1}$, and R^2_{rct} will be correspondingly small. In other words, the R^2_{rct}-score captures the gain by "personalizing" the prediction to the target unit using the SI method over the natural baseline of taking the average outcome of all units who receive that particular intervention. Thus, $R^2_{\text{rct}} > 0$ indicates the *success of SI.*

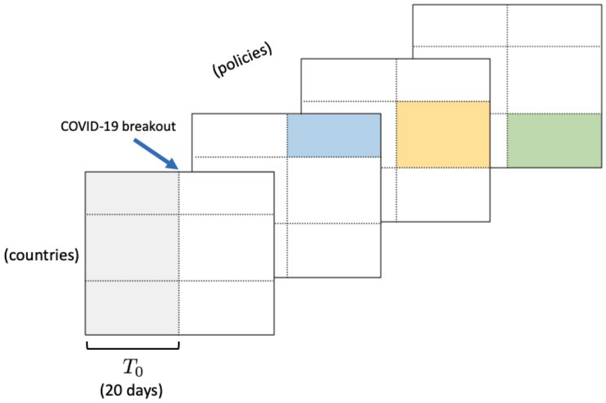


Figure 8.1: Observation pattern for COVID-19 case study.

■ 8.3.1 COVID-19: What-if Scenario Planning

Aim, Setup and Key Modeling Choices

We apply SI to study the impact of mobility restriction policies on COVID-19 related health outcomes at a national level. Below, we list our key modeling decisions.

Choosing Metric of Interest: Daily Death Counts. Due to its relative reliability and availability, we use daily COVID-19 related death counts as our outcome variable of interest. Another standard metric, number of daily infections, is much less reliable due to the inconsistencies in testing and reporting across regions.

Choosing Interventions of Interest: Daily Mobility Rates. Each country has 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. However, almost all such policies have been directed towards restricting how individuals move and interact. Thus, we adopt mobility as our notion of intervention, and investigate how a country's change in mobility level translates to the number of potential COVID-19 related deaths. To that end, we use Google's mobility reports goo to study the change in a country's mobility compared to their respective national baseline from January 2020.

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 N (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" (this will denote our intervention point, T_0) onwards, we consider a country's mobility score from Day -20 to Day -1. Given that the mobility score in goo is changing every day, we take the average mobility score of a country from Day -20 to Day -1 and then bucket it into the three distinct, mutually exclusive intervention groups defined as follows (see Figure 8.2 for a visual depiction of this clustering):

(a) *Low Mobility Restricting Intervention* – reduction in mobility is below 5% compared to national baseline from January 2020;

(b) *Moderate Mobility Restricting Intervention* – reduction in mobility is between 5% to 35% compared to national baseline from January 2020;

(c) *Severe Mobility Restricting Intervention* – reduction in mobility is greater than 35% compared to national baseline from January 2020.

Choosing Pre- and Post-Intervention Periods: Number of Deaths in Country. To apply SI, it is crucial to have well-defined pre- and post-intervention period; in particular, the effects of each country's enacted interventions should only be observed during the post-intervention period. Using Google's mobility reports, we verify that 20 days prior to cumulative 80 deaths in a country (and any time before), none of the selected countries enacted a mobility restricting intervention. Thus, we choose the day a country has cumulative 80 deaths as Day 0, and the pre- and post-intervention periods refer to the days before and after Day 0, respectively.

Observation Pattern. For a graphical depiction of the observation pattern, please refer to Figure 8.1.

Empirical Results and Key Takeaways

We apply SI using the setup above to produce counterfactual predictions of the daily death counts for 15 days following Day 0 under the three different mobility interventions of interest. This analysis is carried out for 27 countries selected as follows: we (i) only include countries whose mobility changes are tracked by Google mobility reports; (ii) remove countries that have enacted a mobility restricting intervention 20 days prior to Day 0; (iii) remove countries with not enough data in the pre-intervention period of interest. That is, countries that had less than 80 cumulative COVID-19 related deaths in the pre-intervention period. We then group the 27 countries into the three buckets defined above based on their average mobility score, as shown in Figure 8.2.

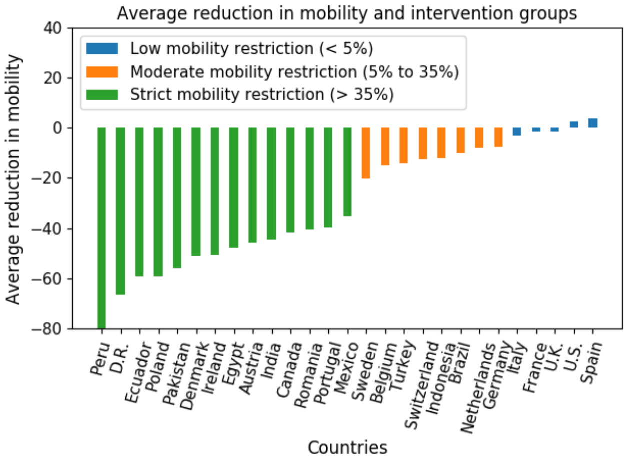
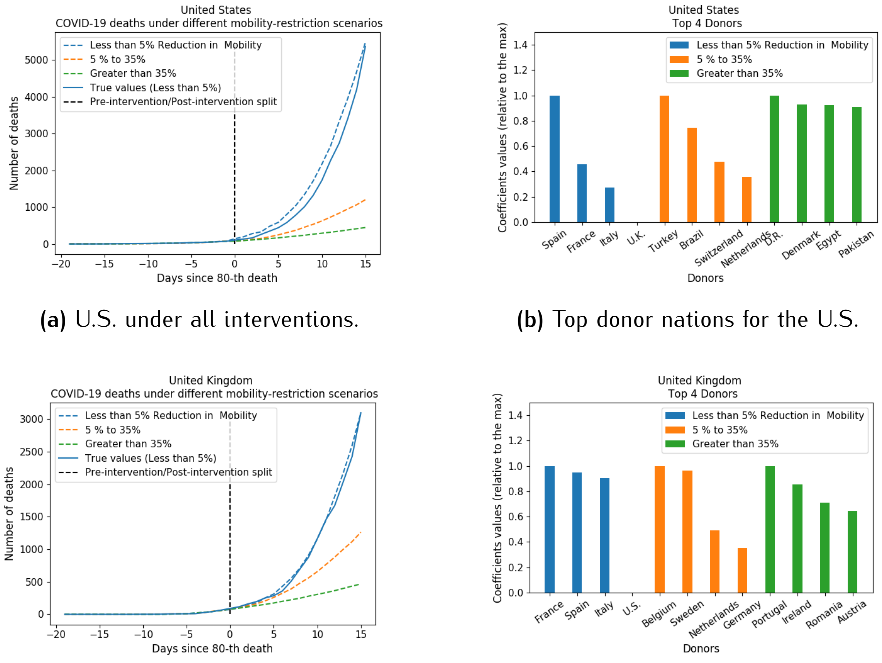


Figure 8.2: Average reduction in mobility and the assigned intervention group for the 27 countries.

Intervention	low	moderate	severe
Hypo. Test ($\alpha = 0.05$)	Pass	Pass	Pass
R^2_{rct}**-score**	0.74	0.14	0.12

Table 8.1: Hypothesis test and prediction accuracy results for SI in the context of COVID-19 under different levels of mobility restriction.



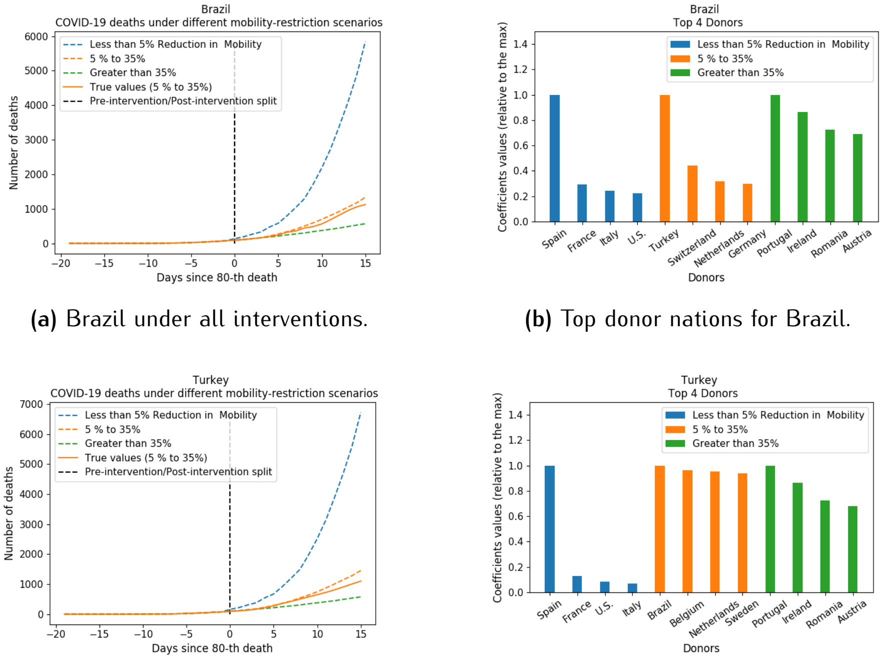
(a) U.S. under all interventions.

(b) Top donor nations for the U.S.

(c) U.K. under all interventions.

(d) Top donor nations for the U.K.

Figure 8.3: Validating SI: countries with low mobility restricting interventions.



(a) Brazil under all interventions.

(b) Top donor nations for Brazil.

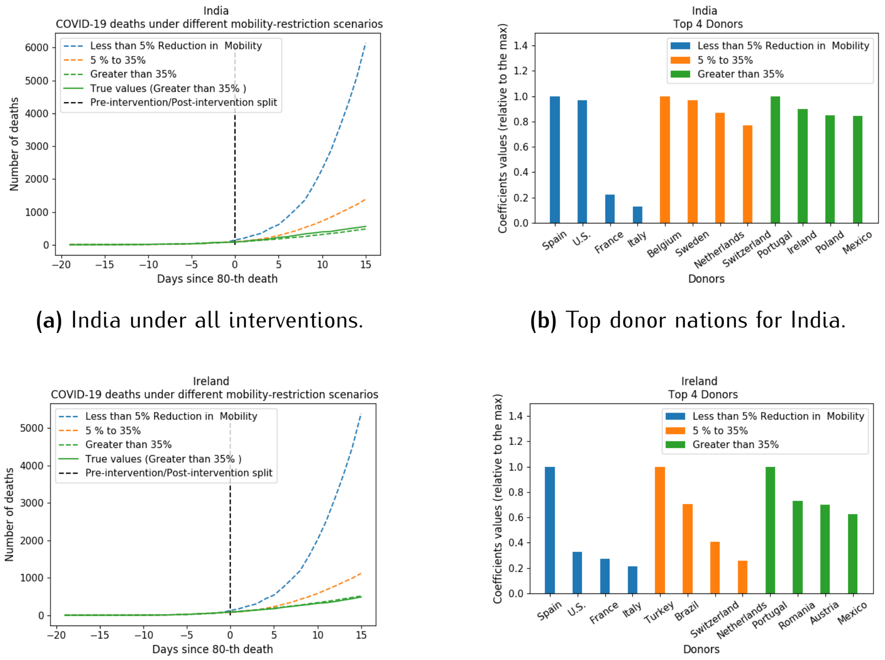
(c) Turkey under all interventions.

(d) Top donor nations for Turkey.

Figure 8.4: Validating SI: countries with moderate mobility restricting interventions.

Empirical Results. In Table 8.1, we show the results of the hypothesis test for the three mobility restricting interventions and the median R^2_{rct}-score for all 27 countries. The hypothesis test passes for all three interventions at a significance of $\alpha = 0.05$. A median R^2_{rct}-score of $[0.74, 0.14, 0.12]$ across the three interventions indicates there is indeed significant heterogeneity amongst the countries on how mobility interventions affect the national death trajectories. Thus, there is significant gains to be had by using SI over naively averaging the outcome across countries that experienced a particular level of mobility reduction.

For every mobility restriction level, we display the counterfactual predictions associated with two representative countries that enacted that intervention. We note similar results hold generally across all countries. For the low mobility restricting regime, we show results for the United States and the United Kingdom in Figures 8.3a and 8.3c, respectively. 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 predictions produced by SI closely matches the observed death rates in the post-intervention period. Similarly, for the moderate and severe mobility restricting regimes,



(a) India under all interventions.

(b) Top donor nations for India.

(c) Ireland under all interventions.

(d) Top donor nations for Ireland.

Figure 8.5: Validating SI: countries with severe mobility restricting interventions.

we display results for Turkey, Brazil, India, and Ireland in Figures 8.4c, 8.4a, 8.5a, and 8.5c, respectively. Again, the predictions produced from SI closely matches the observed death rates under all different interventions, i.e., mobility restrictions. For each of the countries listed above, we display their top four donor countries (under each intervention) that most closely resemble them. These are shown in Figures 8.3b, 8.3d, 8.4d, 8.4b, 8.5b, and 8.5d respectively.

Key Takeaways. Importantly, the SI model of the target country is fit in the pre-intervention period, when no intervention has yet occurred. Still, the learnt model transfers to an intervention 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 compared to the national baseline). After this point, gains by further restricting mobility seem to be diminishing. We hope this case study shows how SI can be used to guide important policy decisions.

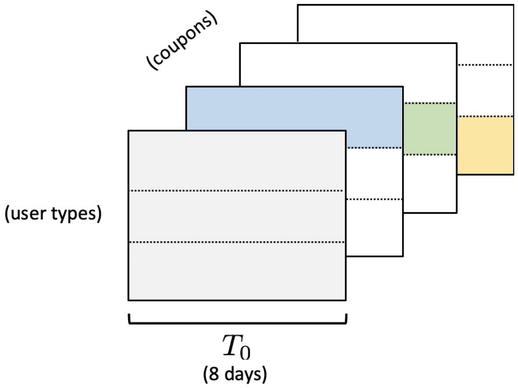


Figure 8.6: Observation pattern for A/B testing case study.

■ 8.3.2 Web A/B Testing: Towards Data Efficient RCTs

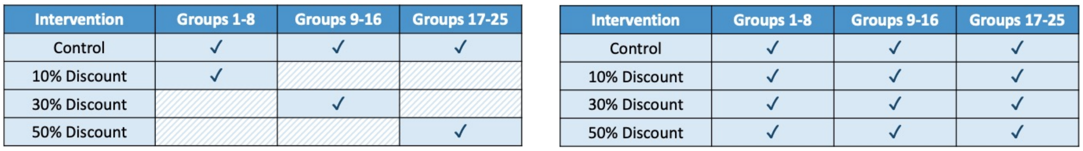
Aim, Setup and Key Modeling Choices

We consider an A/B testing dataset from a large e-commerce company[1] that issued different discount strategies (interventions) to engage its customer base: 10%, 30%, and 50% discounts over the regular subscription cost. Users were segmented into 25 groups ($\sim$ 10,000 users per group) based on the historical time and money spent on the platform. The aim of the e-commerce company was to find how these different levels of discounts affected user engagement for each of the 25 user groups. The A/B test was performed by randomly partitioning users in each of the 25 user groups into 4 sub-groups; these sub-groups corresponded to either one of the 3 discount strategies or a control group that received a 0% discount. User engagement in each of these 100 sub-groups (25 user groups multiplied by 4 discount strategies) was measured daily over 8 days.

Suitability of Case Study to Validate SI. This web A/B testing case study is particularly suited to validate SI as we get to observe the engagement levels of each customer group under each of the three discount strategies, i.e., we observe every "counterfactual" trajectory. This is in contrast to the COVID-19 case study where we only observe the death trajectory for a country for the particular intervention it enacted during the post-intervention period.

Choosing Pre- and Post-Intervention Periods. For each of the 25 user groups, we

[1] We anonymize the identity of the company due to privacy considerations.



Intervention	Groups 1-8	Groups 9-16	Groups 17-25
Control	✓	✓	✓
10% Discount	✓	▨	▨
30% Discount	▨	✓	▨
50% Discount	▨	▨	✓

(a) Experimental setup under SI.

Intervention	Groups 1-8	Groups 9-16	Groups 17-25
Control	✓	✓	✓
10% Discount	✓	✓	✓
30% Discount	✓	✓	✓
50% Discount	✓	✓	✓

(b) Experimental setup of e-commerce company.

Figure 8.7: Experimental setups for A/B testing case study.

denote the daily user engagement trajectories of the sub-groups associated with the control – those who do not receive a discount on their regular subscription – as the pre-intervention period. Correspondingly, for each of the 25 user groups, we denote the daily user trajectories associated with the 10%, 30%, and 50% discount coupons as the post-intervention period.

Choosing Donor Groups. We randomly partition the 25 user groups into three clusters, denoted as user groups 1-8, 9-16, and 17-25. For the 10% discount coupon strategy, we choose user groups 1-8 as our donor pool, i.e., we use their post-intervention data under a 10% discount to create the synthetic trajectories of user engagement for user groups 9-25 under a 10% discount. We do the same with the 30% and 50% discount coupon strategies, and user groups 9-16 and 17-25, respectively. See Figure 8.7a for a visual depiction of the set of experiments/observations the SI algorithm uses to make predictions.

Observation Pattern. For a graphical depiction of the observation pattern, please refer to Figure 8.6.

Empirical Results and Key Takeaways

We apply SI using the setup above to produce the "counterfactual" trajectories for each of the 25 user groups under the three discount strategies. We evaluate the accuracy under the 10% discount coupon strategy using only the estimated trajectories of user groups 9-25 (as we use user groups 1-8 as our donors). Similarly, we use the estimated trajectories of user groups 1-8 and 17-25 for the 30% discount coupon strategy, and user groups 1-16 for the 50% discount coupon strategy.

Empirical Results. In Table 8.2, we show the hypothesis test results for the three discount strategies and the median R^2_{rct}-score of the 25 user groups. The hypothesis test passes for all three interventions at a significance of $\alpha = 0.05$. Additionally, SI achieves a median R^2_{rct}-score of 0.98 across the three discount strategies. This indicates significant heterogeneity amongst the user groups in how they respond to discounts, and thus warrants having to run separate A/B tests for each of the 25 groups.

Intervention	10% discount	30% discount	50% discount
Hypo. Test ($\alpha = 0.05$)	Pass	Pass	Pass
R^2_{rct}**-score**	0.98	0.99	0.98

Table 8.2: Hypothesis test and prediction accuracy results for SI in the context of A/B testing.

Key Takeaways. Recall that there were a total of 100 distinct experiments run in the A/B testing framework as there were 25 user groups and 4 interventions (0%, 10%, 30%, and 50% discount coupons). However, the SI framework only required observations from 50 experiments. That is, two experiments for each of the 25 user user groups: one in the pre-intervention period (under 0% discount rate) and one in the post-intervention period (under exactly one of the three discount coupon strategies). See Figure 8.7b for a visual depiction of the experiments conducted by the e-commerce company in comparison to what is required by SI, as shown in Figure 8.7a.

More generally, if there are N user groups and D interventions, an *ideal* RCT performs $N \times D$ experiments to estimate the best "personalized" intervention for every user group. With SI, assuming the tensor factor model holds and $D \leq N$, one only needs to perform $2N$ experiments. Crucially, the number of required experiments does not scale with D, which becomes significant as the number of interventions, i.e, the level of personalization, grows. Also, if pre-intervention data has been or is being collected, then SI only requires N experiments. This can be significant when experimentation is costly (e.g., clinical trials).

■ 8.3.3 Development Economics: Towards "Personalized" RCTs

Aim, Setup and Key Modeling Choices

We use data from a large real-world development economics case study, which aimed to increase vaccination rates in seven districts in the state of Haryana, India. This study, carried out by the authors of Banerjee et al. (2018) in collaboration with the Haryana state government, is the first large scale evaluation of the effects of different types of interventions on childhood immunization rates. The Haryana immunization trials were conducted with 2523 villages, with data collected monthly over 13 months, and included a total of 74 different interventions. Each intervention can be encoded by a 3-dimensional discrete-valued vector where its entries represent different levels of (1) financial incentives, (2) social network influence, and (3) information campaigns to encourage vaccinations.

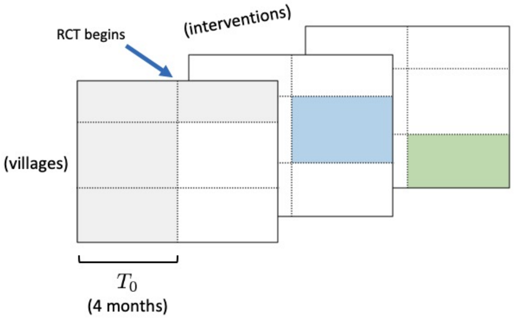


Figure 8.8: Observation pattern for development economics case study.

"Personalized" RCTs via SI. As is standard in RCTs, the authors in Banerjee et al. (2018) randomly partitioned the 2523 villages into 74 groups, corresponding to the 74 different interventions they aimed to study. They then measured the average increase in immunization rates for each of these 74 groups over the 13 month trial period. Subsequently, they made a single policy recommendation to the Haryana state government, corresponding to the intervention that yielded the highest average increase in immunization rates.

The aim of this case study is to estimate whether there would have been a greater uptake in immunization amongst the villages if, instead of a single policy recommendation for all villages, a tailored intervention recommendation was made for each village.

Data Pre-Processing. We restrict our attention to the 20 most frequent interventions, where the frequency of an intervention is measured by the number of villages that experienced said intervention, e.g., 175 villages experienced the most frequent intervention while 18 villages experienced the twentieth most frequent intervention. Let $\mathcal{D}$ denote the collection of these 20 interventions. There were $N = 1302$ villages that received one of the top 20 most frequent interventions. Based on conversations with the authors of Banerjee et al. (2018), it was appropriate to denote the first four months as the pre-intervention period, i.e., $T_0 = 4$ months.

Observation Pattern. For a graphical depiction of the observation pattern, please refer to Figure 8.8.

Empirical Results and Key Takeaways

We follow the same setup as in the COVID-19 case study. That is, we iterate over the 1302 villages such that each village is designated to be the target village for some iteration. In the pre-intervention period, we build a model of the target village under each of the twenty interventions using the appropriate donor village sub-groups. Then in the post-intervention phase, we estimate the counterfactual immunization rates of the target village under each intervention using data from the appropriate donor village sub-group and fitted linear model.

Empirical Results. In Tables 8.3 and 8.4, we show the results of the hypothesis test for the twenty interventions considered and the median R^2_{rct}-scores. The hypothesis test passes for all but four interventions at a significance of $\alpha = 0.05$. Indeed, the corresponding median R_{rct}-scores are among the lowest, with three of four being the minimum achieved scores. This highlights the use of the hypothesis test as a helpful robustness check for when to trust the counterfactual predictions produced by SI. For the remaining 17 interventions that do pass the hypothesis test, we generally see significantly higher R^2_{rct}-scores, indicating again that there is significant heterogeneity amongst villages.

Intervention Code	000	001	002	010	031	032	040	050	100	101
Hypo. Test ($\alpha = 0.05$)	Pass	Pass	Fail	Pass	Pass	Pass	Pass	Pass	Pass	Pass
R^2_{rct}**-score**	0.55	0.50	0.48	0.73	0.62	0.73	0.57	0.75	0.50	0.68

Table 8.3: Hypothesis test and prediction accuracy results for SI in the context of immunization case study for top 1-10 most frequent interventions.

Intervention Code	102	200	201	202	300	301	302	400	401	402
Hypo. Test ($\alpha = 0.05$)	Pass	Pass	Pass	Pass	Fail	Pass	Pass	Fail	Fail	Pass
R^2_{rct}**-score**	0.48	0.70	0.66	0.45	0.34	0.46	0.60	0.29	0.29	0.42

Table 8.4: Hypothesis test and prediction accuracy results for SI in the context of immunization case study for top 11-20 most frequent interventions.

Key Takeaways. The question we set out to answer was whether providing "personalized" intervention recommendations to each village would have led to significant increases in the immunization rates for that village over the single intervention recommendation made by the authors of Banerjee et al. (2018), as is standard practice in a RCT. Using the counterfactual estimates produced by SI, we define the average utility of intervention d for village n as

$$\widehat{U}(n, d) = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \widehat{M}^{(d)}_{tn}.$$

In words, for a particular intervention d, this is the average increase in immunization rates over the post-intervention phase. Further, let

$$\widehat{U}_{\text{rand}} = \frac{1}{N}\sum_{n=1}^{N}\widehat{U}(n, d_n), \quad \widehat{U}_{\text{rct}} = \max_{d\in\mathcal{D}}\frac{1}{N}\sum_{n=1}^{N}\widehat{U}(n, d), \quad \widehat{U}_{\text{tailored}} = \frac{1}{N}\sum_{n=1}^{N}\max_{d\in\mathcal{D}}\ \widehat{U}(n, d)$$

where $d_n \sim \text{Uniform}(\mathcal{D})$. In words, $\widehat{U}_{\text{rand}}$ represents the estimated average utility across all villages if a randomly sampled intervention had been administered. $\widehat{U}_{\text{rct}}$ represents the estimated best single intervention across all villages in hindsight (i.e., the RCT policy). Lastly, $\widehat{U}_{\text{tailored}}$ is the estimated average utility for each village under its optimal intervention.

Normalizing $\widehat{U}_{\text{rand}}$ as 1.0, we find $\widehat{U}_{\text{rct}}$ and $\widehat{U}_{\text{tailored}}$ are 1.3x and 2.8x higher, respectively, compared to $\widehat{U}_{\text{rand}}$. Thus, if the units of interest are heterogeneous, then using SI to produce tailored interventions can lead to large gains. We stress these are only estimated utilities as we, of course, never observe each village under all interventions. Lastly, we note the estimated single best policy that maximized $\widehat{U}_{\text{rct}}$ matched the policy recommendation made in Banerjee et al. (2018).

Recommendation Type	Average Utility
$\widehat{U}_{\text{rand}}$: Random Assignment	1.0
$\widehat{U}_{\text{rct}}$: Single-best RCT Policy	1.3
$\widehat{U}_{\text{tailored}}$: Personalized Recommendation	2.8

Table 8.5: Average utilities associated with three types of intervention interventions per village: random assignment, single-best RCT policy, and personalized intervention recommendation.

■ 8.3.4 In-Vitro Life Sciences: Drug Discovery

The standard drug delivery pipeline begins with exploratory, in-vitro studies on animal and/or human cell, which are used to determine the drug candidates that should be investigated in the clinical stages involving human subjects. Unfortunately, the traditional paradigm is known to suffer from inefficiency, high costs, and high failure rates, only to deliver "one-drug-fits-most" treatment options. This begs several important questions: (1) can we identify the most promising candidate therapies early on so we do not waste resources in the latter stages? (2) can we personalize our therapies based on the particular characteristics of individuals (i.e., achieve precision medicine)? In this study, we tackle the first question using in-vitro data.

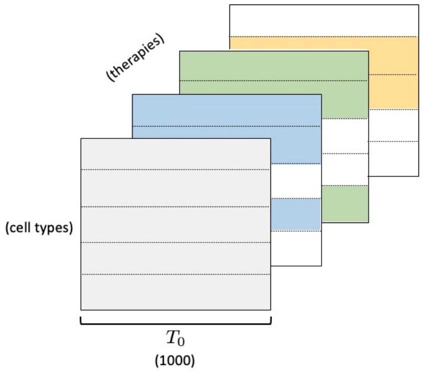


Figure 8.9: Observation pattern for in-vitro case study.

Aim, Setup and Key Modeling Choices

We continue our application of SI in the context of in-vitro studies. More specifically, we consider the task of predicting the effects of chemical therapies (drugs) on cell types, using the publicly available LINCS dataset. Due to sparsity concerns, we analyze a subset of the dataset that is comprised of the 20 most administered therapies and a control, DMS10, which yields 24 unique cell types. Each of the cells is observed under the control therapy and a subset of the 20 other therapies; in total, we observe the gene expressions for approximately 60% of all possible cell-therapy pairs. Therefore, our goal is to use our existing data (the 60% of observed gene expressions) to infer the gene expressions of the remaining untested cell-therapy pairs.

However, unlike the previous studies where each unit (cell) only experiences a single intervention (therapy) during the post-intervention period, the units in this study receive multiple interventions. In turn, we can learn relationships either across units or interventions. We detail the nuances below.

1. **Learning Relationships between Cells.** To gain a better intuition for the experimental setup, consider (for concreteness) the task of estimating the gene expression for cell 1 under therapy yellow. The same principles are then applied to all untested cell-therapy pairs.

   (a) *Choosing Donor Groups.* Given the discussion above, we define the donor cells for target cell 1 as the subset of cells that receive therapy yellow.

   (b) *Choosing Pre- and Post-intervention Data.* To begin, we consider the pre-intervention data as cell 1's observed gene expression under the control, DMS10.

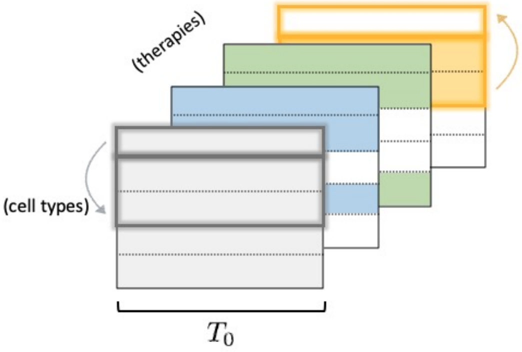


Figure 8.10: Experimental setup for SI on cells.

However, since cells are exposed to multiple therapies rather than just one, we identify the largest common subset of therapies (not including control) that were exposed to cell 1 and the donor cells (that receive therapy yellow). The gene expressions under the common subset of therapies can then be interpreted as "auxiliary metrics"; this effectively augments the pre-intervention data and, as suggested by our theoretical and empirical results in Chapter 7, behooves the SI estimator. Restating the above, we define the pre-intervention data as the collection of gene expressions under all commonly tested therapies and control. Correspondingly, the post-intervention donor data is represented by the observed gene expressions of our donor cells under therapy yellow. For a graphical depiction, please refer to Figure 8.10.

(c) *Algorithm.* To infer the gene expression of cell 1 under therapy yellow, we apply the standard SI algorithm (using side information) as before, i.e., we learn a linear model between cell 1 and the donor cells, and perform a synthetic intervention by rescaling the donor gene expressions under therapy yellow via the learnt (linear) coefficients. To distinguish this approach from the discussion to follow, we will refer to this method as SI *on cells*.

2. **Learning Relationships between Therapies.** Consider the above example of estimating cell 1's gene expressions under therapy yellow. However, rather than learning a linear model between cells, we will now exploit the symmetry of our tensor factor model and the structure of our data to learn a linear model between interventions.

   (a) *Choosing Donor Groups.* By symmetry, we define the donor therapies for therapy yellow as the subset of therapies (including control) that cell 1 experienced.

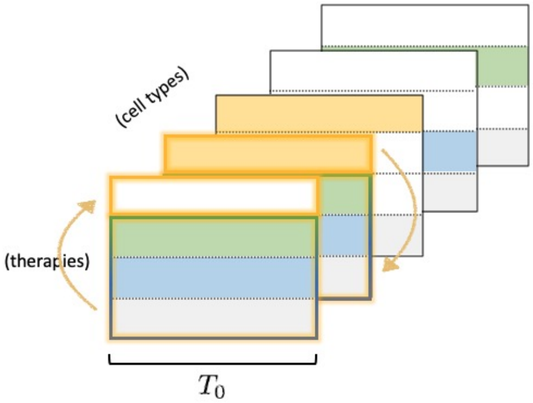


Figure 8.11: Experimental setup for SI on therapies.

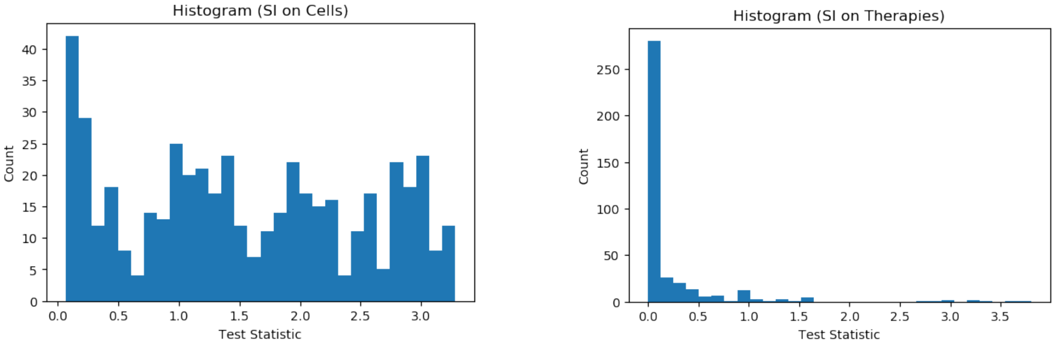
(b) *Choosing Pre- and Post-intervention Data.* In a similar spirit to the above setup, we first identify the largest common subset of cells that experienced the donor therapies and therapy yellow. Our pre-intervention donor data is then represented by the amalgam of gene expressions of these cells under the donor therapies, while our pre-intervention target data is represented by the gene expressions of these cells under therapy yellow. Therefore, our post-intervention donor data are the gene expressions of cell 1 under the donor therapies. For a graphical depiction, please refer to Figure 8.11.

(c) *Algorithm.* Under this new setup, we apply the SI algorithm along the therapies instead. That is, we learn a linear model between therapy yellow and the donor therapies using the data described above. Then, we use the learnt model to rescale the gene expressions of cell 1 under the donor therapies to effectively perform a synthetic intervention. As such, we refer to this method as SI *on therapies*.

Empirical Results and Key Takeaways

We apply SI on both cells and therapies to produce two different estimates of the gene expressions for each cell-therapy pair (not including control data). However, as before, we evaluate our method on the observed gene expressions. To do so, we iteratively hold out one tested cell-therapy pair and apply the two methods described above using the remaining observations to infer its resulting gene expression.

Empirical Results. We display a histogram of the test statistics of our hypothesis test for SI on cells and therapies in Figures 8.12a and 8.12b, respectively. Recalling Theorem 5.4.1



(a) Histogram of test statistics for SI on cells.

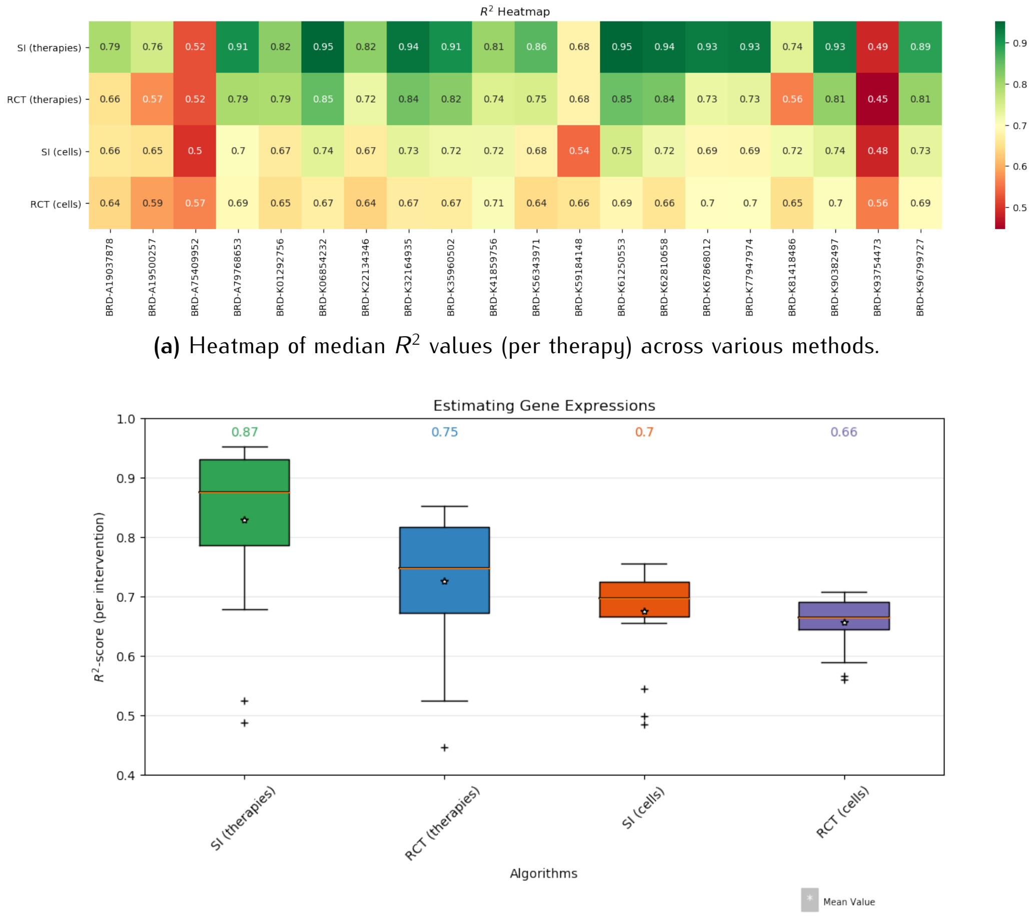
(b) Histogram of test statistics for SI on therapies.

Figure 8.12: Histogram of test statistics for both SI methods; namely, applied to cells and therapies. As suggested by Theorem 5.4.1, SI on therapies should outperform SI on cells, given the left skewness of the test statistics in (b).

and the discussion in Chapter 5.4, we remind the reader that, for all practical purposes, the closer the test statistic is to zero, the more likely SI will be able to generalize. As such, the histograms suggest that SI on therapies should outperform SI on cells since the post-intervention data, in the context of SI on cells, is likely to be more complex than that of the corresponding pre-intervention data.

We display a heatmap of median standard R^2-scores (across receiving cells) for each of the 20 therapies in Figure 8.13a, where greener entries correspond to higher scores and redder entries correspond to lower scores. In line with the previous case studies, we compare our two methods, SI on cells and on therapies, against their "RCT" estimator counterparts, i.e., the RCT estimator on cells simply takes the average gene expression levels across all donor cells as its estimate, and the RCT estimator on therapies is defined analogously. As evident from the figure, we identify that SI on therapies outperforms its competitors almost uniformly across all therapies. Pleasingly, this is in agreement with our hypothesis test results (namely, the histogram of test statistics), which continue to be a useful guide in testing for generalizability. For convenience, we further summarize Figure 8.13a via a boxplot in Figure 8.14b, which displays the median R^2-score, and corresponding lower and upper quantile bounds, for each method across all therapies.

Key Takeaways. Because experiments can be costly, therapies are commonly tested on a subset of cells. The therapies that are the most promising on average across the tested cells are then likely to move onto the clinical stages. Thus, our baseline "RCT" (average) estimators reflect a standard inference procedure utilized in modern practice. However,



(a) Heatmap of median R^2 values (per therapy) across various methods.

(b) Boxplot summary of (a).

Figure 8.13: In (a), we see that, almost uniformly across all therapies, SI on therapies is the highest performing approach. (b) summarizes the heatmap in (a) via a boxplot, which displays the median R^2–score across therapies for each method, with its corresponding upper and lower quantile bounds.

as shown in the figures above, SI (on therapies) can far exceed standard baselines with respect to inference. Therefore, in settings such as bench and clinical research, where any gain in R^2-values is of significance, SI can be largely beneficial for researchers. Specifically, SI has the ability to accurately infer the potential outcomes for untested cell-therapy pairs. In turn, we hope this may guide researchers (in a data-driven manner) on which promising pairs to invest in and perform experimentation. At the same time, our hypothesis tests can inform the researchers on which estimates to confidently trust in and which to disregard. In short, SI can effectively be viewed as a recommendation engine for researchers.

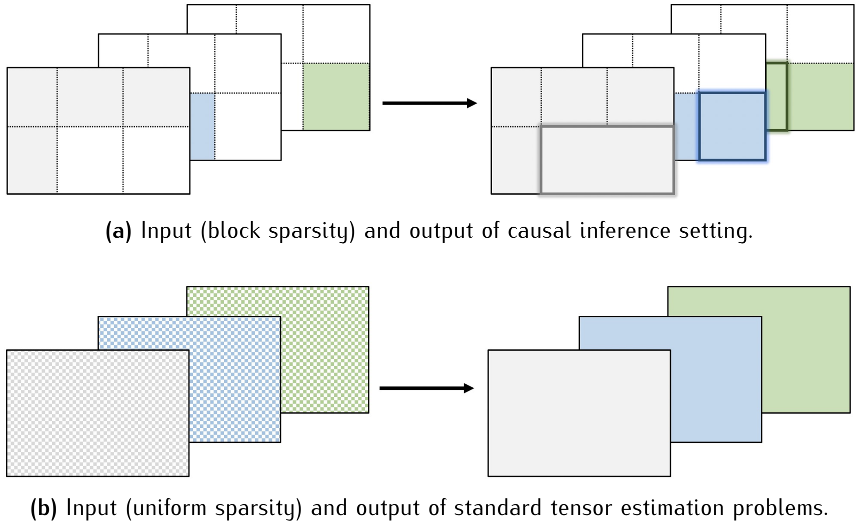
■ 8.4 Discussion

■ 8.4.1 Connection to Tensor Estimation

The penultimate goal of this chapter (and thesis) is to infer the counterfactual potential outcomes for all unit-intervention tuples during the post-intervention period. Therefore, unlike the standard SC settings (see Chapters 6.4.1 and 7.4.1) where the primary interest is in recovering a specific segment of a matrix, our problem is now formalized as recovering all of the "interesting" (post-intervention period) aspects of the order-three tensor whose dimensions correspond to measurements, units, and interventions[2]. Because each unit is only exposed to a single intervention or remains unaffected during the post-intervention period, the studies induce a block sparsity pattern. Further, we aim to provide guarantees across all post-intervention measurements for every unit-intervention tuple. As such, standard tensor estimation methods, which typically assume observations are revealed at random and commonly provided guarantees on average (with respect to Frobenius norm) across the entire tensor, may be ill-suited for our purposes.

On the other hand, SI is tactfully designed to handle the particular block sparsity and provides statistical guarantees for every unit-intervention pair, as desired. Since the units are assumed to operate under a common state for T_0 measurements and the measurement factors evolve from the pre- to post-intervention periods, SI explores and learns the correlations amongst the units during this pre-intervention phase. This is crucial as the latent unit factors are the only objects that are simultaneously diverse during the training period (allowing us to extract signal) yet remain fixed in the post-intervention period (enabling generalization).

[2] Recall that we analyze the single metric $P = 1$ for simplicity, but can easily extend to the multiple metric case.



(a) Input (block sparsity) and output of causal inference setting.

(b) Input (uniform sparsity) and output of standard tensor estimation problems.

Figure 8.14: Comparison of sparsity patterns and objectives of causal inference and standard tensor estimation problems.

■ 8.4.2 Connection to Transfer Learning & Transportability

Since the effects of interventions can vary, potential outcomes associated with the pre- and post-intervention periods may come from different domains. Additionally, we are only given access to the target unit's pre-intervention labels, as its post-intervention labels are precisely the unobservable counterfactuals we wish to estimate. Therefore, our problem of interest also places us within the transductive transfer learning setting and bears connections to transportability. That is, using the language of transfer learning, the *source* (pre-intervention) and *target* (post-intervention) domains may be different yet related, and only the source domain labels are available. Nevertheless, as we have proven in Theorem 8.2.1 and Corollary 8.2.1, SI can extrapolate from our observed outcomes, which may be observed under one interventional framework, to estimate the counterfactual potential outcomes under a distinct interventional framework. Importantly, we can apply our subspace inclusion hypothesis test (see Chapter 5.4) to validate when SI can reliably transfer models between frameworks in practice.

■ 8.4.3 Broader Impacts

What-if Scenario Planning

It is clear that the COVID-19 pandemic has led to an unprecedented disruption of modern society at a global scale. What is much less clear, however, is the effect that various interventions that have been put into place have had on health and economic outcomes. For example, perhaps a 30% and 60% clampdown in mobility have similar societal health outcomes, yet vastly different implications for the number of people who cannot go to work or file for unemployment. Having a clear understanding of the trade-offs between these interventions is crucial in charting a path forward on how to open up various sectors of society. A key challenge is that policy makers do not have the luxury of actually enacting a variety of interventions and seeing which has the optimal outcome. In fact, at a societal level, this is simply infeasible. Arguably, an even bigger challenge is that the COVID-19 pandemic, and the resulting policy choices ahead of us, are unprecedented in scale. Thus, it is difficult to reliably apply lessons from previous pandemics (e.g., SARS, H1N1). This is only further exacerbated when taking into the account the vastly different economic, cultural, and societal factors that make each town/city/state/country unique.

SI provides a data-driven and statistically principled way to perform *what-if* scenario planning, i.e., for policy makers to understand the trade-offs between different interventions before having to actually enact them. In essence, the SI method leverages information from different interventions that have already been enacted across the world and fits it to a policy maker's setting of interest – for example, to estimate the effect of mobility-restricting interventions on the U.S., we use daily death data from countries that enforced severe mobility restrictions to create a "synthetic low-mobility U.S." and predict the *counterfactual* trajectory of the U.S. if it had indeed applied a similar intervention. We highlight a few desirable attributes of this methodology.

- *Personalized Predictions* – SI takes into account the heterogeneity of the particular (geographical) purview of a policy-maker. For example, SI would provides different predictions for the effects of a 40% drop in mobility for the U.S. vs. India vs. Italy etc. based on the particulars of that country.

- *Simplicity & Interpretability* – SI relies on building "synthetic" versions of each geographical location under different interventions by simply using a weighted combination of geographical locations that did indeed enact such an intervention. Thus, SI requires relatively little hyper-parameter tuning. We hope that the simplicity, interpretability, and robustness (yet surprising accuracy) of SI will encourage policy makers to apply SI without fear of over-fitting to the idiosyncracies of their data.

- *Low Data Requirement* – SI produces accurate forecasts using only (i) a few "donor" regions, i.e., SI only requires a small number of regions, approximately 5-10, to have gone through the intervention of interest; (ii) measurements over a small number of time periods, approximately 10-30 days (this can be viewed as the amount of training data); (iii) data corresponding to the metric of interest from each geographical location (e.g., daily national death rates), i.e., it does not require additional covariate information about each location; however, if auxiliary information is available, then the SI method can naturally incorporate these data points in the model learning procedure as well.

Randomized Control Trials (RCTs)

Data-Efficient RCTs. Consider the setting with $N > 1$ types of customers coming to an e-commerce website, which has $D > 1$ types of promotions to offer. The goal is to find which of the D promotions is best suited for each of the N different customer types. Traditionally, this is be achieved by running $N \times D$ RCTs (i.e., A/B tests). As detailed in Section 8.3.2, using actual e-commerce A/B testing data, we show that SI can infer these $N \times D$ outcomes by running only $2N$ experiments[3] (assuming $D \leq N$); crucially, this does *not* depend on D.

Personalized RCTs. A core assumption in such RCTs is that a blanket policy works well for all units, i.e., all interventions have essentially the same effect on all units. However, this assumption is often violated, and the inherent diversity between different groups of people is increasingly taken into consideration. SI, on the other hand, provides personalized recommendations for each group, yet essentially only requires the same data as what is generated in a classical RCT. Indeed, as detailed in Section 8.3.3, we find that personalized recommendations can have significant gains over the optimal RCT policy (see Table 8.5 for details).

Beyond Traditional Paradigms. In general, it is worth noting that all of our results have direct implications for other important applications where RCT-like experiments are an integral part of the decision-making pipeline. In particular, there has been a large, recent wave towards precision (i.e., personalized) medicine. As discussed above, SI

[3] If one has or is already collecting pre-intervention data, then SI only requires N experiments. Further, we note that SI only requires outcomes across all N user types under a common intervention, i.e., the model does not necessarily have to be learned under a no-intervention setting.

provides a way to potentially perform personalized clinical trials without having to run an infeasible number of experiments on the various patient groups. Within clinical trials, patient recruitment and compliance is especially costly due to the monetary expenses and ethical considerations (e.g., placebo trials). Thus, the potential application of SI, especially in the context of personalized drug design or clinical trials, if successful, can have a large impact.

Chapter 9

Discussion, Conclusions, and Future Work

In this thesis, we reinterpret the classical potential outcomes framework through the lens of tensors (Chapter 4). As such, studies can be characterized by unique block sparsity patterns, and the problem of estimating counterfactuals is equivalent to tensor estimation. Under a low-rank assumption, we prove the existence of synthetic controls and interventions, and argue that PCR is a natural algorithm that arises from our setting of interest. In Chapter 5, we address a long-standing problem of showing PCR is surprisingly robust to a wide array of problems that plague large-scale modern datasets, including high-dimensionality and noisy, sparse, and mixed valued covariates. In particular, we provide meaningful non-asymptotic bounds for both the parameter estimation and test prediction errors for these settings, and furnish a data-driven hypothesis test to check when the key condition that enables generalizability holds. Having established the robustness of PCR, we present a robust variant of SC in Chapter 6 that uses PCR as a key subroutine to estimate counterfactual potential outcomes under control. As such, our finite-sample analysis of PCR immediately provides a theoretical foundation for the RSC estimator. For particularly sparse datasets, we present MRSC, an extension of RSC that incorporates auxiliary metrics in a principled manner, in Chapter 7. Finally, we present SI in Chapter 8, a computationally and statistically efficient method towards estimating counterfactual potential outcomes under both control *and treatment*. Consequently, SI enables effective decision making with notable applications towards what-if policy evaluation/scenario planning, drug discovery, and personalized, data efficient RCTs (A/B tests). Of course, there is still significant room for further improvement and extensions of this thesis, which we briefly discuss below.

■ 9.1 Algorithmic Fine Print

■ 9.1.1 Incorporating Covariates

In order for PCR (and thus RSC, MRSC, and SI) to recover the underlying model parameter (Theorem 5.3.1) and generalize to unseen data (Theorem 5.3.2 and Corollary 5.12.1), it is essential that PCA (the key subroutine of PCR) accurately estimates the latent subspace spanned by the top principal components of the underlying training covariates, $\boldsymbol{V}_{\text{train}} \in \mathbb{R}^{p \times r}$. At the same time, we may have access to auxiliary covariate information, which we denote as $\boldsymbol{A} \in \mathbb{R}^{p \times s}$. For concreteness, consider the setting SC and SI, where $\boldsymbol{V}_{\text{train}}$ describes the latent relationship between units and $\boldsymbol{A}$ contains s features per unit. Then, as suggested by Farias and Li (2019), the auxiliary information can be incorporated by expanding the estimated feature space $\widehat{\boldsymbol{V}}_{\text{train}}$ with $\boldsymbol{A}$, and then proceeding normally. In turn, this expansion should ideally better describe the relationship between units, which is crucial for generalization given that the key enabling condition required for PCR is $\text{span}(\boldsymbol{V}_{\text{test}}) \subseteq \text{span}(\boldsymbol{V}_{\text{train}})$ (see Chapters 5.3.5 and 5.4 for details).

■ 9.1.2 Finding a Low-Dimensional Representation

In essence, Chapter 5 shows that the PCA component of PCR is an effective pre-processing tool in finding a linear low-dimensional embedding of the covariates, which carries the added benefits of implicit de-noising, noise-model agnostic, and regularization. When the covariate data is "unstructured" (e.g., speech or video), however, it may be the case that more complex covariate pre-processing techniques a la variational auto-encoders or general adversarial networks are needed to discover useful nonlinear low-dimensional embeddings that can also achieve similar implicit benefits. Therefore, we hope that this work provides an architectural guideline and theoretical framework of first de-noising and then performing a prediction algorithm in the presence of more general datasets.

■ 9.1.3 Beyond PCA and HSVT

As per the discussion above, this thesis advocates for a simple algorithmic architecture that is comprised of two primary steps: (1) first, to de-noise the covariate (donor) data and (2) then learn a mapping from the de-noised covariates to the target for the purposes of prediction. While we argue that PCA and HSVT are powerful de-noising techniques with provable statistical guarantees, they are, by no means, the only options. Therefore, it is the analyst's discretion to decide which method (e.g., alternating least squares or nuclear norm minimization) is most appropriate for his or her setting of interest.

■ 9.2 Future Work

■ 9.2.1 Causal Forecasts under Novel Sequence of Interventions

As mentioned, SI (Chapter 8) represents the culmination of this thesis and has strong implications towards effective decision making across a wide net of applications. Nevertheless, there are inevitable limitations of SI. Namely, SI, like any causal inference procedure, produces counterfactuals for the *past*; at the same time, even though SI can estimate potential outcomes under both control and treatment, these estimates can only be constructed if some subset of units has actually undergone these interventions of interest. This enables SI to answer questions such as *what would have happened two weeks ago if the U.S. had imposed severe mobility restrictions*? From this perspective, SI can be viewed as a "causal imputation" algorithm. However, it may be of interest to produce "counterfactual *forecasts*" under a *novel* sequence of interventions. To elucidate this extension, consider the following contrived scenario: Suppose none of the countries ever transitioned from severe to low mobility restrictions. Then, an example question that is beyond the current scope of SI would be *what will happen to the U.S. a month from now if it had imposed high mobility restrictions two weeks ago and now relaxes to low mobility restrictions moving forward*? Because this a forecasting problem and none of the countries ever experienced this interventional shift, SI cannot address such a query. As a result, an interesting line of future work (which is currently underway) would be to build upon SI to overcome these limitations.

■ 9.2.2 Open Question

A primary contribution of this work is to view causal inference through a new lens. Specifically, we encode our data into a tensor and reformulate the problem of estimating counterfactuals into one of tensor completion. Although we have presented one algorithm (i.e., SI) that is able to recover the tensor under a tensor factor model, it remains an open problem as to which sparsity patterns and corresponding objectives are achievable. Answering this question may have profound consequences for new study frameworks and inference schemes.

Bibliography

Google LLC google covid-19 community mobility reports. https://www.google.com/covid19/mobility/. Accessed: 2020-04-20.

Synthetic control arms: the end of placebos? 2019. URL https://stories.abbvie.com/stories/synthetic-control-arm-end-placebos.htm?_ga=2.157423162.1271074098.1565222118-1727800855.1565222118.

A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 2003.

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American Statistical Association*, 2010.

A. Abadie, A. Diamond, and J. Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 2014.

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects (working paper). 2019.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015a.

Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*, 2015b.

Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *Advances in neural information processing systems*, 2016.

B. Adhikari and J. Alm. Evaluating the economic effects of flat tax reforms using synthetic control methods. *Southern Economic Association*, 2016.

Anish Agarwal, Muhammad J. Amjad, Devavrat Shah, and Dennis Shen. Model agnostic time series analysis via matrix estimation. *POMACS*, 2(3):40–79, 2018.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 2019.

Anish Agarwal, Abdullah Alomar, Romain Cosson, Devavrat Shah, and Dennis Shen. Synthetic interventions. *ArXiv*, abs/2006.07691, 2020a.

Anish Agarwal, Abdullah Alomar, Arnab Sarker, Devavrat Shah, Dennis Shen, and Cindy Yang. Two burning questions on covid-19: Did shutting down the economy help? can we (partially) reopen the economy without risking the second wave?, 2020b.

Anish Agarwal, Abdullah Alomar, and Devavrat Shah. On multivariate singular spectrum analysis, 2020c.

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

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.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham Kakade. A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory*, pages 867–881, 2013.

Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference in differences. *arXiv e-prints arXiv:1812.09970.*, 2018.

S. Athey and G. Imbens. The state of applied econometrics - causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2016.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, and Guido Imbens. Matrix completion methods for causal panel data models. 2017.

H. Aytug, M. Kutuk, A. Oduncu, and S. Togan. Twenty years of the eu-turkey customs union: A synthetic control method analysis. *Journal of Common Market Studies*, 2016.

Mohammad Taha Bahadori, Qi (Rose) Yu, and Yan Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3491–3499. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5429-fast-multivariate-spatio-temporal-analysis-via-low-rank-tensor-learning.pdf.

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.

BallotPedia. California proposition 63, background checks for ammunition purchases and large-capacity ammunition magazine ban (2016). *www.ballotpedia.org*, 2016. URL https://ballotpedia.org/California_Proposition_63,_Background_Checks_for_Ammunition_Purchases_and_Large-Capacity_Ammunition_Magazine_Ban_(2016).

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, John Floretta, Harini Kannan, Anna Schrimpf, and Maheshwor Shrestha. Improving immunization coverage through incentives, reminders, and social networks in india. 2018.

Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445, 2016.

AA. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming approaches to high-dimensional errors-in-variables models. *Journal of the Royal Statistical Society*, 79:939–956, 2017a.

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*, 2017b.

A. Billmeier and T. Nannicini. Assessing economic liberalization episodes: A synthetic control approach. *The Review of Economics and Statistics*, 2013.

Christopher M Bishop. Bayesian pca. In *Advances in neural information processing systems*, pages 382–388, 1999.

Christian Borgs, Jennifer T Chayes, Henry Cohn, and Shirshendu Ganguly. Consistent non-parametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015.

Christian Borgs, Jennifer Chayes, Christina E Lee, and Devavrat Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In *Advances in Neural Information Processing Systems*, pages 4718–4729, 2017.

T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.*, 46(1):60–89, 02 2018. doi: 10.1214/17-AOS1541. URL https://doi.org/10.1214/17-AOS1541.

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.

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. *The Annals of Statistics*, 43(1):177–214, 2015a.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43:177–214, 2015b.

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.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 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, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 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.

N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Paper No. 22791*, 2016.

Vivek Farias and Andrew Li. Learning preferences with side information. *Management Science*, 65, 05 2019. doi: 10.1287/mnsc.2018.3092.

Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low—rank tensor recovery via convex optimization. *Inverse Problems*, 27:025010, 01 2011. doi: 10.1088/0266-5611/27/2/025010.

Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is <inline-formula> <tex-math notation="tex">$4/\sqrt{3}$ </tex-math></inline-formula>. *IEEE Transactions on Information Theory*, 60(8):5040–5053, Aug 2014. ISSN 1557-9654. doi: 10.1109/tit.2014.2323359. URL http://dx.doi.org/10.1109/TIT.2014.2323359.

MA Hernán and JM Robins. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*, 2020.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 379–390. IEEE, 2017.

Cheng Hsiao, Shui-Ki Wan, and Yimeng Xie. Panel data approach vs synthetic control method. *Economics Letters*, 164:121–123, 2018.

Furong Huang, U. N. Niranjan, Mohammad Umar Hakeem, Animashree An, and kumar. Online tensor methods for learning latent variable models. *Journal of Machine Learning Research*, 16(86):2797–2835, 2015. URL http://jmlr.org/papers/v16/huang15a.html.

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.

Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data, 2014.

Ji Liu, P. Musialski, P. Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2114–2121, 2009.

Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society*, 31(3):300–303, 1982.

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.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009. doi: 10.1137/07070111X.

N. Kreif, R. Grieve, D. Hangartner, A. Turner, S. Nikolova, and M. Sutton. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, 2015.

Christina E. Lee, Yihua Li, Devavrat Shah, and Dogyoon Song. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems 29*, pages 2155–2163, 2016.

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.

M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing mri. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008.

Michael Lustig, David Donoho, and John Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 58:1182–95, 12 2007. doi: 10.1002/mrm.21391.

Patrick McGreevy. California voters approve gun control measure proposition 63. *Los Angeles Times*, Nov. 2016. URL http://www.latimes.com/nation/politics/trailguide/la-na-election-day-2016-proposition-63-gun-control-1478280771-htmlstory.html.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Master's Thesis*, 1923.

Donald B. Rubin Paul R. Rosenbaum. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

Philippe Rigollet and Jan-Christian Hütter. High dimensional statistics. 2017.

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. Matching to remove bias in observational studies. *Biometrics*, 29(1): 159–183, 1973.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974a.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974b.

J. Saunders, R. Lundberg, A. Braga, G. Ridgeway, and J. Miles. A synthetic control approach to evaluating place-based crime interventions. *Journal of Quantitative Criminology*, 2014.

Devavrat Shah and Dogyoon Song. Learning mixture model with missing values and its application to rankings. *arXiv preprint arXiv:1812.11917*, 2018.

Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation, 2020.

Karl Stratos. PhD thesis, 2016.

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.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

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.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Telfar Barnard L Baker MG Wilson N, Kvalsvig A. Case-fatality estimates for covid-19 calculated by using a lag time for fatality. *Emerg Infect Dis.*, 2020. URL https://doi.org/10.3201/eid2606.200320.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. *arXiv preprint arXiv:1709.03183*, 2017.

Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. *CoRR*, abs/1905.11971, 2019a. URL http://arxiv.org/abs/1905.11971.

Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Harnessing structures for value-based planning and reinforcement learning, 2019b.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, Apr 2015. ISSN 1464-3510. doi: 10.1093/biomet/asv008. URL http://dx.doi.org/10.1093/biomet/asv008.

Yuan Zhang and Regina Barzilay. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1213. URL https://www.aclweb.org/anthology/D15-1213.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.