Algorithmic Approaches to Nonparametric Causal Inference

by

Peter L. Cohen

B.A., Bowdoin College (2017)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Sloan School of Management
April 29, 2022

Certified by .....
Colin B. Fogarty
Assistant Professor
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science
Codirector, Operations Research Center

Algorithmic Approaches to Nonparametric Causal Inference

by

Peter L. Cohen

Submitted to the Sloan School of Management
on April 29, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

This thesis presents procedures for performing inferences of causal parameters across an array of contexts including observational studies, completely randomized designs, paired experiments, and covariate-adaptive designs. First, we discuss an application of convex optimization to conduct directional inference and sensitivity analyses in matched observational studies. We design an algorithm which maximizes the signal-to-noise ratio while accounting for unobserved confounding. We analyze the asymptotic distributional behavior of the algorithm's output to develop asymptotically valid hypothesis tests for causal effects. The resulting procedure achieves the maximal design sensitivity over a broad class of procedures. Second, we examine the role of feature information in drawing high-precision inferences of effects in completely randomized experiments. We construct a calibration technique based around linear regression which constructs imputation estimators with upper bounds on the asymptotic variance of the estimator. We show that this calibration procedure is applicable to any imputation estimator which may be semiparametric efficient and automatically certifies that the resulting nonlinear regression-adjusted estimator is at least as asymptotically precise as the difference in means; a feature that was previously not guaranteed for nonlinear regression-adjusted estimators under model misspecification. Third, we introduce Gaussian prepivoting: an algorithmic technique to construct test statistics for which randomization inference remains asymptotically valid even when symmetries underlying the randomization hypothesis are violated in the null. We demonstrate that randomization tests based upon prepivoted statistics are finite-sample exact under sharp nulls while they asymptotically control the probability of false rejection under weak nulls. This allows for the formation of confidence regions for treatment effects with simultaneous interpretations as exact confidence regions for homogeneous additive treatment effects and asymptotic confidence regions for heterogeneous additive effects; thereby unifying Fisherian and Neymanian inference for many experimental designs including rerandomized experiments. Fourth, we construct a nested

hierarchy of resampling algorithms which exploit probabilistic structure in superpopulation, fixed covariate, and finite population models to facilitate nonparametric inference for a wide variety of statistics in completely randomized designs. The resampling algorithms extend the classical bootstrap paradigm by leveraging modern results on regression-adjustment and optimal transport to achieve significant gains under fixed covariate and finite population models.

Thesis Supervisor: Colin B. Fogarty
Title: Assistant Professor

Acknowledgments

Those to whom I owe thanks are too great in number to name. The opportunity to pursue my Ph.D. at MIT stands as the culmination of years of work from dedicated educators who taught me to see the beauty of mathematics and the importance of its applications. I am enormously grateful to the efforts of those who guided me along the path that led to Cambridge.

First, I thank my advisor Colin Fogarty. His patient explanations and insightful comments have given me the opportunity to grow as a researcher. The breadth and depth of his knowledge across so many fields have been awe inspiring. Moreover, his humility, grace, and constantly kind attitude have served as a model for conduct. It is a combination of his talent as a scholar and his genuine personality that has made working with him a pleasure. Some of the most fun I had during my Ph.D. has been working through the ideas of a proof at a whiteboard with him; my only complaint is that COVID cut short these wonderful meetings.

Second, I thank the students of MIT's Operations Research Center, especially those of my cohort. Before coming to MIT, I knew that my Ph.D. would be an excellent academic experience, but it turned out to be an excellent social experience as well. From long days studying for qualifying exams to countless conversations over coffee, I am grateful to you all for your knowledge and – more importantly – your friendships.

Third, I thank my parents, Susan and Dan, my brother, David, and wonderful partner, Roya. In particular, I offer my sincere thanks to Roya for putting up with countless hours of listening to mathematics interspersed with bluegrass. My time here at MIT has been profoundly enriching and humbling; I could not have asked for a better support network, and I dedicate this thesis to them.

Contents

List of Figures

List of Tables

Chapter 0

Introduction

0.1 Overview

In brief, this thesis concerns statistically rigorous procedures for drawing causal conclusions from observed data. The standard caveat that *correlation does not imply causation* generally makes causal inference a challenging statistical question, however the practical relevance of determining cause-and-effect relationships cannot be understated. This thesis approaches causal inference across a range of contexts, but a consistent set of underlying principles runs throughout the chapters:

- *Randomization as a basis for inference*: Random assignment to treatment can be algorithmically leveraged to draw causal conclusions.

- *Algorithms based upon practitioner choice*: Practitioners bring crucial subject-matter knowledge to the table; the algorithms presented in this thesis rest on top of underlying choices made by users but retain statistical guarantees despite user choices in analyzing their data.

- *Practical interpretability matters*: Statistical conclusions are only as valuable as the actions they inform; practitioners must be able to understand the key elements of the inference procedures in order to effectively apply inferences towards real-world decision-making. Furthermore, causal claims ought to be "robust to misinterpretation".

- *Minimal assumptions are a necessity*: In some fields, assumptions of *i.i.d.* data, exchangeable observations, parametric generative models, etc. are commonplace and easy to justify. However, for many modern contexts – especially for practitioners in the social sciences and economics – such assumptions are burdensome if not prohibitive. Undergirding this work is a mathematical framework that avoids such assumptions while still showing that strong statistical results can be garnered nonetheless.

0.2 Thesis Organization

The thesis is organized into four main chapters:

1. *Multivariate One-sided Testing in Matched Observational Studies as an Adversarial Game*: In this chapter we study causal inference in observational studies; studies wherein the allocation of treatment is not under the control of the experimenter. We develop a sensitivity analysis procedure for detecting prespecified directions of treatment effect. The procedure is tightly related to a game-theoretic problem, and the Nash equilibrium of the game can be efficiently computed via convex optimization.

2. *No-harm Calibration for Generalized Oaxaca-Blinder Estimators*: In this chapter we study how a practitioner can leverage side-information to improve the precision of inferences for completely randomized experiments. In practice experimenters often record feature information for the participants in a study in addition to their primary outcome of interest; this chapter details how one can use this information to fruitfully sharpen one's inferential statements within a model-agnostic regression framework.

3. *Gaussian Prepivoting for Finite Population Causal Inference*: Dating back to the Fisher-Neyman controversy of the twentieth century, causal inference has typically split between two different frameworks. The distinction between the two camps is subtle, and is possibly confusing to practitioners. In this chapter we develop a single unified method for causal inference under both frameworks. Inferences drawn from this method are robust to practitioner misinterpretation: at worst a practitioner may make a statement which they believe to be finite-sample exact which turns out to only be asymptotically true.

4. *Hierarchical Resampling Procedures for Causal Inference* Resampling methods have played a crucial role in statistical inference since the pioneering bootstrap work of Efron

in the 1980s. In this chapter, we construct three resampling algorithms which provide bootstrap hypothesis tests and confidence intervals for causal parameters across three different probabilistic models commonly used in the causal inference community. The three resampling algorithms form a natural hierarchy wherein each algorithm is tailored to the probabilistic structure available in the model which motivated the algorithm but still provides valid inference in conditional submodels. Each resampling algorithm is easy to implement and confers the automaticity of the bootstrap to inference problems of practical relevance.

0.3 Chapter Details

Multivariate One-sided Testing in Matched Observational Studies as an Adversarial Game

Random allocation of treatment effect through a controlled process provides strong guarantees for subsequent statistical inferences. Consequently, it is in the interest of an experimenter to use an experimental design tailored to their applications (e.g., complete randomization, pair-matching, etc.). However, numerous situations necessitate that causal inferences be drawn from data for which the experimenter had no oversight of the treatment allocation process. For instance, in testing the harms of an addictive drug, it would be obviously unethical to randomly assign some individuals in a study to begin using the drug. Nonetheless it is crucial for public health practitioners to understand the impacts of such "treatments"; for such instances observational studies serve as a fruitful avenue for understanding treatment effects.

Causal claims in observational studies are inherently vulnerable to concerns of unmeasured confounding. For instance, imagine that there was a gene that simultaneously predisposes individuals to smoking tobacco and to developing lung cancer; such a factor may

impact the "assignment to treatment" (choice to smoke tobacco) and the outcome (development of lung cancer). Without controlling for this factor, analyses of smoking and lung cancer are merely associational and not causal. To control for any such factor the experimenters must assign the treatments themselves or know the probability of treatment for each individual. Since this is typically infeasible, [Ros02, Chapter 4] presents a *sensitivity analysis* for matched observational studies. Informally, sensitivity analyses ask: how strong of an unmeasured confounding feature would need to exist to invalidate the causal claims made with the observed data?

We present a multivariate one-sided sensitivity analysis for matched observational studies, appropriate when the researcher has specified that a given causal mechanism should manifest itself in effects on multiple outcome variables in a known direction. The test statistic can be thought of as the solution to an adversarial game, where the researcher determines the best linear combination of test statistics to combat nature's presentation of the worst-case pattern of hidden bias. The corresponding optimization problem is convex, and can be solved efficiently even for reasonably sized observational studies. Asymptotically the test statistic converges to a $\overline{\chi}^2$ (chi-bar-squared) distribution under the null, a common distribution in order restricted statistical inference. The test attains the largest possible design sensitivity over a class of so-called "coherent" test statistics, and facilitates one-sided sensitivity analyses for individual outcome variables while maintaining familywise error control through incorporation into closed testing procedures.

No-harm Calibration for Generalized Oaxaca-Blinder Estimators

It is standard practice to accumulate background information about subjects in a study in addition to recording the outcomes of interest. For instance, in healthcare experiments it is common to compile a large list of features for each individual at the onset of a study. Sometimes these features are directly used to inform the experimental design itself, e.g.,

matched designs. However, it is common to collect feature data and still perform a completely randomized experiment; in fact sometimes this is a necessity, e.g., if the collected data includes protected features which would be ethically prohibitive to base treatment decisions upon. Of course, in a completely randomized experiment there may be chance imbalances between the features of treated and control units; adjusting for these imbalances stands to potentially provide significant benefit in downstream analyses. *How does one go about adjusting for feature information?*

On a whole, the question above puts this work squarely in the field of *regression adjusted estimators*, a crowded field of research with a long history dating back to at least the 1970s with the work of [Oax73] and [Bli73]. In [Lin13] it was shown that linear regression of the outcomes upon the covariates, an indicator variable of treatment, and an interaction term facilitates a new *linear regression adjusted estimator* of treatment effect, $\hat{\tau}_{lin}$, which is guaranteed to have asymptotic efficiency no worse than the unadjusted difference in means between the treated and control groups, $\hat{\tau}_{unadj}$. Lin's result does not assume that the relationship between covariates and outcomes obeys a linear model. His result is model agnostic in the sense that his estimator $\hat{\tau}_{lin}$ is non-inferior to $\hat{\tau}_{unadj}$ uniformly with respect to the underlying relationship between covariates and outcomes. This surprising result cleverly leverages the orthogonalities of linear regression. Consequently, nonlinear regression adjustment procedures generally lack the non-inferiority of $\hat{\tau}_{lin}$. In many cases, the data suggests that a practitioner ought to adjust using some nonlinear model; e.g., $\{0, 1\}$-valued outcomes suggest logistic regression or $\mathbb{N}$-valued outcomes suggest Poisson regression. We present a novel technique which facilitates simple *interpretable* nonlinear regression adjusted estimators for treatment effect that are guaranteed to have asymptotic efficiency no worse than that of $\hat{\tau}_{unadj}$. In other words, we show that the results of [Lin13] are a special case of a general class of nonlinear adjustment algorithms; our work builds off of recent work on imputation-based estimators of [GB21].

Gaussian Prepivoting for Finite Population Causal Inference

When practitioners test for the presence of treatment effect, they must specify the null model they wish to test against. This amounts to defining what "no treatment effect" means; two different definitions exist in the literature: Neyman's weak null (equality in mean outcome between treatment and control) and Fisher's sharp null (potential outcomes are equal under treatment and control).

In finite population causal inference exact randomization tests can be constructed for such sharp null hypotheses, i.e., hypotheses which fully impute the missing potential outcomes. The mathematical convenience of randomization tests and their associated finite-sample guarantees are highly desirable, but oftentimes inference is instead desired for the weak null that the average of the treatment effects takes on a particular value while leaving the subject-specific treatment effects unspecified. Without proper care, tests valid for sharp null hypotheses may be anti-conservative should only the weak null hold, creating the risk of misinterpretation when randomization tests are deployed in practice.

We develop a general framework for unifying modes of inference for sharp and weak nulls, wherein a single procedure simultaneously delivers exact inference for sharp nulls and asymptotically valid inference for weak nulls. To do this, we employ randomization tests based upon prepivoted test statistics. Prepivoting, an idea introduced by Rudy Beran in the late 1980s, facilitates asymptotically pivotal statistics under the assumption that the sharp null holds [Ber87, Ber88]. For a large class of test statistics, we show that prepivoting rests on transforming a test statistic of interest by the push-forward of a sample-based Gaussian measure with a suitably estimated covariance parameter. The main result of this work is a proof that randomization tests using these prepivoted statistics provide asymptotically valid inference even under just Neyman's weak null. This means that a practitioner can use the randomization test of such a prepivoted statistic to test both Fisher's sharp null and Neyman's weak null: the only difference being that inferences under Fisher's sharp null

are finite-sample exact and those under Neyman's weak null are asymptotic. Consequently, statistical conclusions are robust to practitioner misinterpretation; at worst a result is stated to be finite-sample exact under Neyman's weak null when it is in-fact only asymptotic.

We demonstrate our method in a host of examples, including rerandomized designs and regression-adjusted estimators in completely randomized designs.

Hierarchical Resampling Procedures for Causal Inference

Numerous probabilistic models have been used for statistical inference within the potential outcomes framework of causal inference. In particular, the three most common models are the superpopulation model wherein units are *i.i.d.* draws from a fixed distribution, the fixed covariate model wherein the potential outcomes of a unit are drawn from a conditional distribution which depends upon the unit's features, and the finite population model wherein potential outcomes and features are deterministic and randomness is inherited exclusively from the treatment allocation process. Bootstrap procedures for the classical nonparametric Behrens–Fisher problem suggest that resampling may provide fruitful inferential algorithms in the superpopulation model; but research in resampling for inferences in the fixed covariate and finite population models remains at an early stage.

In this chapter we provide a novel reformulation of the classical *i.i.d.* bootstrap of Efron which recognizes that the bootstrap can itself be viewed in the light of a two-phase experimental design framework: first a full population of units is constructed via some resampling procedure and next an experiment is simulated upon this "bootstrapped" population. This reformulation suggests a potential direction forward for resampling algorithms at the fixed covariate and finite population levels: construct a full population by leveraging the probabilistic structure of the model and then simulate a completely randomized experiment as before. In the fixed covariate and finite population models the construction of the bootstrap population is tightly related to construction of imputation estimators and leverages results of

[Lin13] and [CF21]. The finite population resampling algorithm that we develop utilizes both imputation estimation, copula results of [Tch80], and optimal transport theory to construct a resampling procedure which achieves an asymptotically sharp bootstrap variance.

Bibliography

[Ber87] Rudolf Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468, 1987.

[Ber88] Rudolf Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

[Bli73] Alan S. Blinder. Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, 1973.

[CF21] Peter L. Cohen and Colin B. Fogarty. No-harm calibration for generalized oaxaca-blinder estimators, 2021.

[GB21] Kevin Guo and Guillaume Basse. The generalized Oaxaca-Blinder estimator. *Journal of the American Statistical Association*, (DOI 10.1080/01621459.2021.1941053):DOI 10.1080/01621459.2021.1941053, 2021.

[Lin13] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

[Oax73] Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973.

[Ros02] Paul R. Rosenbaum. *Observational studies*. Springer, New York, 2002.

[Tch80] Andre H. Tchen. Inequalities for distributions with given marginals. *The Annals of Probability*, 8(4):814–827, 1980.

Chapter 1

Multivariariate One-sided Testing in Matched Observational Studies as an Adversarial Game

(Joint work with Matt Olson)

Abstract

We present a multivariate one-sided sensitivity analysis for matched observational studies, appropriate when the researcher has specified that a given causal mechanism should manifest itself in effects on multiple outcome variables in a known direction. The test statistic can be thought of as the solution to an adversarial game, where the researcher determines the best linear combination of test statistics to combat nature's presentation of the worst-case pattern of hidden bias. The corresponding optimization problem is convex, and can be solved efficiently even for reasonably sized observational studies. Asymptotically the test statistic converges to a chi-bar-squared distribution under the null, a common distribution in order restricted statistical inference. The test attains the largest possible design sensitivity over a class of coherent test statistics, and facilitates one-sided sensitivity analyses for individual outcome variables while maintaining familywise error control through its incorporation into closed testing procedures.

1.1 On Multiplicity And Causality

Controlled randomization protects empirical evidence against a host of counterclaims. A significant finding may well be due to random chance alone, but cannot be dismissed on the grounds of biases unaccounted for by the study's design. Observational evidence provides no such assurance, and causal inference in observational studies involves ambiguity which randomization eschews: Is the association an effect, or is it bias from self-selection? Anticipating skepticism, a practitioner may take measures when planning an observational study to properly frame the debate, rendering certain criticism unwarranted should the practitioner's hypothesis be true. While ambiguity cannot be eliminated, quasi-experimental devices may be employed to help clarify the step from association to causation in observational studies; see [SCC02] and [Ros15] for an overview. One such device, known alternatively as pattern specificity, multiple operationalism, or coherence, advocates that observational studies be designed with the objective of confirming a complex pattern of predictions made by the causal theory in question. This is in keeping with Fisher's notion of elaborate theories, which advocates that the practitioner "envisage as many different consequences of [a causal

hypothesis's] truth as possible, and plan observational studies to discover whether each of these consequences is found to hold" [Coc65, Section 5, p. 252]. Complex predictions imperil the practitioner's hypothesis, as doubt is cast should any prediction fail in the observational study at hand. Should the evidence prove coherent with the theory's predictions, fortification is provided as attributing a complex pattern to hidden bias requires that hidden bias could reproduce the particular pattern of association.

One way in which a theory can be made elaborate is through predicting that an intervention will affect multiple outcome variables in a prespecified direction. While the practitioner hopes that each prediction holds, should certain predictions fail she would regardless like to quantify which components came to fruition as a means of refining understanding of the mechanism in question. With this comes the attending issues of multiple comparisons. Concerns over a loss in power from multiplicity control may lead practitioners to instead investigate the outcome they believe *a priori* will be most affected, reducing the extent to which Fisher's advice is followed.

The qualitative benefits of multiple outcomes in observational studies are thus at odds with the statistical corrections they require. This tension exists not only when assuming no hidden bias, but also in the sensitivity analysis where the researcher quantifies the magnitude of hidden bias required to overturn the study's conclusions. In what follows, we present a new method for sensitivity analysis in multivariate one-sided testing, appropriate when the researcher anticipates a particular direction of effects for multiple outcome variables. The test adaptively combines outcome-specific test statistics, has the optimal design sensitivity over a class of multivariate tests respecting coherence, and leads to substantial improvements in power when the researcher's prediction proves correct. The method greatly attenuates the impact of multiplicity control on power for testing individual outcome variables through its use in closed testing procedures [MEG76], facilitating the analysis of multiple outcomes for demonstrating coherence.

1.2 Hidden Bias In Matched Observational Studies

1.2.1 A Finely Stratified Experiment with Multiple Outcomes

There are I independent strata, the ith of which contains $n_i \geq 2$ individuals. Individual j in stratum i has a P-dimensional vector of observed covariates x_{ij}, along with an unobserved covariate u_{ij}, $0 \leq u_{ij} \leq 1$. The strata are formed such that $x_{ij} \approx x_{ij'}$ for any two individuals $j \neq j'$ in stratum i. We take Z_{ij} as the indicator of treatment for the jth individual in stratum i, such that $Z_{ij} = 1$ if assigned to treatment and $Z_{ij} = 0$ otherwise. Each strata contains one treated individual and $n_i - 1$ controls such that $\sum_{j=1}^{n_i} Z_{ij} = 1$ $(i = 1, ..., I)$. See [Fog18] for more on this particular class of stratified experiments, referred to as finely stratified experiments. Forthcoming developments readily extend to full-matched observational studies; see [Ros02, Ex. 4.12] for details.

Each individual has two vectors of potential outcomes of length K: the responses for each outcome variable under control $r_{Cij} = (r_{Cij1}, ..., r_{CijK})^T$, and the responses under treatment $r_{Tij} = (r_{Tij1}, ..., r_{TijK})^T$. The K-dimensional vector of treatment effects $\tau_{ij} = r_{Tij} - r_{Cij}$ is not observed; instead, we observe the vector $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$. Let $Z = (Z_{11}, ..., Z_{In_I})^T$ be the lexicographically ordered vector of treatment assignments of length N, and let the analogous hold for u along with r_{Ck}, r_{Tk} and R_k for $k = 1, ..., K$. The $N \times K$ matrix with lexicographically ordered rows containing R_{ij}^T is R. We assume that the response of individual j varies only with the treatment allocation to unit j and that the potential outcomes are well-defined; this is commonly referred to as the "stable unit treatment value assumption" [Rub86, Ros02].

Let $\mathcal{F} = \{r_{Cij}, r_{Tij}, x_{ij}, u_{ij} : i = 1, ..., I; j = 1, ..., n_i\}$ be a set containing the potential outcomes along with the measured and unmeasured covariates for each individual in the observational study. In what follows we consider inference conditional upon $\mathcal{F}$, such that a generative model for the potential outcomes is neither assumed nor required. Let $\Omega =$

$\{z : \sum_{j=1}^{n_i} z_{ij} = 1; i = 1, ..., I\}$ be the set of $\prod_{i=1}^{I} n_i$ treatment assignments adhering to the stratified design, and let $\mathcal{Z} = \{Z \in \Omega\}$ be the event that the observed treatment assignment satisfies this design. In a finely stratified experiment $\text{pr}(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = 1/n_i$ and $\text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ where $|A|$ is the cardinality of the set A.

1.2.2 A Model for Biased Treatment Assignment

Matched observational studies aim to mimic the finely stratified experiment described in Section 1.2.1. Matching algorithms assign individuals to matched sets on the basis of observed covariates such that $x_{ij} \approx x_{ij'}$ for individuals j and j' in the same matched set i; see [Han04] and [Zub12] among many for more on matching algorithms and the optimization problems underpinning them. A simple model for treatment assignment in an observational study states that before matching, individuals are assigned to treatment independently with unknown probabilities $\pi_{ij} = \text{pr}(Z_{ij} = 1 \mid \mathcal{F})$. While one may hope that $\pi_{ij} \approx \pi_{ij'}$ after matching, proceeding as such would be specious due to both the potential presence of unmeasured confounding and residual imbalances on the observed covariates in each matched set. The model of [Ros02, Chapter 4] stipulates that individuals in the same matched set may differ in their odds of assignment to treatment by at most a factor of Γ,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ij'})}{\pi_{ij'}(1 - \pi_{ij})} \leq \Gamma. \tag{1}$$

The parameter Γ controls the degree to which matching solely on observed covariates may have failed to align the assignment probabilities in each matched set. The value $\Gamma = 1$ returns a randomized finely stratified experiment, while $\Gamma > 1$ allows for a tilt in the randomization distribution to a degree controlled by Γ. For instance, $\Gamma = 2$ stipulates that individuals in the same matched set might truly differ in their odds of receiving the treatment by a factor of at most two. Returning attention to the matched structure by conditioning on $\mathcal{Z}$, this

model is equivalent to assuming

$$\mathbb{P}\left(Z = z \mid \mathcal{F},\, \mathcal{Z}\right) = \frac{\exp\left(\gamma z^T u\right)}{\sum_{b\in\Omega}\exp\left(\gamma b^T u\right)}, \tag{2}$$

where $\gamma = \log(\Gamma)$ and u lies in the N-dimensional unit cube, call it $\mathcal{U}$, embodying both differences in unobserved covariates and latent discrepancies in observed covariates after matching; see [Ros95] or [Ros02, Chapter 4] for a proof of this equivalence.

1.2.3 Sensitivity Analysis for a Particular Outcome

Assume without loss of generality that the outcomes have been recorded such that positive values for the treatment effects τ_{ijk} are predicted by the causal theory under study. For each outcome variable, we consider tests of the null hypothesis of non-positive treatment effects,

$$H_k : r_{Tijk} \leq r_{Cijk} \quad (i = 1, ..., I; j = 1, ..., n_i).$$

H_k is a composite null hypothesis. Elements of H_k include the null of a non-positive constant effect for all individuals, $r_{Tijk} = r_{Cijk} + \delta_k$ for any scalar $\delta_k \leq 0$; and certain models of tobit effects, such as $r_{Cijk} = \max\{r_{Tijk}, 0\}$. Fisher's sharp null of no effect is $\delta_k = 0$, thus representing the boundary of H_k. The composite null H_k is distinct from Neyman's weak null of no average treatment effect for the kth outcome variable. That said, both nulls allow for inference without prespecifying the particular pattern of effect heterogeneity. Neyman's null has been seen as a flexible way to test for existence of treatment effect while accommodating arbitrary effect heterogeneity. Unfortunately, testing Neyman's null on the kth outcome greatly constrains the test statistics available to the practitioner, requiring the use of a studentized difference-in-means or a regression-adjusted estimator [WD18]. These test statistics have poor theoretical properties when used in a sensitivity analysis. The null H_k is also more general than a sharp null, but can still be tested through randomization

inference using statistics such as m-tests with better theoretical properties in the potential presence of hidden bias [Ros07]. The null H_k is not limited to continuous outcome variables, and can also be employed with ordinal outcomes. In fact, our method may be used with potential outcomes of any partially ordered set. See Section 1.7 for further details.

We consider test statistics for each outcome variable which are effect increasing sum statistics. Sum statistics are statistics of the form $T_k(Z, R_k) = Z^T q_k$ where $q_k = q_k(R_k)$ is a pre-specified function of the observed responses R_k. A test statistic is effect increasing if $T_k(z, r_k^*) \geq T_k(z, r_k)$ whenever $(2z_{ij} - 1)(r_{ijk}^* - r_{ijk}) \geq 0$ for all i and j, where r_{ijk}^* denotes a different value of the potential outcome. In words, this means that if every treated unit did better with r_k^* than with r_k, and if every control did worse with r_k^* than with r_k, then the test statistic corresponding to the observed outcomes r_k^* would be larger than it would have been under r_k. Most familiar test statistics, including differences-in-means, rank tests, and m-tests are endowed with these properties; see [Ros02, Chapter 2.4.4] and [Ros16, Section 3.1] for additional examples.

If Fisher's sharp null is true then $R_k = r_{Ck}$, and hence $T_k(Z, R_k) = T_k(Z, r_{Ck})$. For a particular $\Gamma > 1$, the test statistic's null distribution under Fisher's sharp null is

$$\text{pr}\{T_k(Z, r_{Ck}) \geq v \mid \mathcal{F}, \mathcal{Z}\} = \sum_{z \in \Omega} 1\{T_k(Z, r_{Ck}) \geq v\} \frac{\exp\left(\gamma z^T u\right)}{\sum_{b \in \Omega} \exp\left(\gamma b^T u\right)}, \tag{3}$$

where $1(A)$ is an indicator that the condition A was met. At $\Gamma = 1$ (3) is simply the proportion of treatment assignments where the test statistic is greater than or equal to v, returning the usual randomization inference in a finely stratified experiment. For $\Gamma > 1$ (3) is unknown due to its dependence on the nuisance vector u. A sensitivity analysis proceeds for a particular Γ by maximizing (3) with $v = t_k$, the observed value of the test statistic for a particular Γ, resulting in the largest possible p-value for the desired inference subject to (1) holding at Γ. The practitioner then increases Γ until the test no longer rejects the null hypothesis. This changepoint value of Γ serves as a measure of how robust the study's finding

was to unmeasured confounding. See [GKR00] and [Ros18] for large-sample approaches for conducting a sensitivity analysis for Fisher's sharp null with a single outcome variable under (1). Since T_k is assumed effect increasing, the worst-case p-value for a sensitivity analysis for Fisher's sharp null attains the largest p-value over the composite null H_k. That is, a sensitivity analysis for Fisher's sharp null also provides a valid sensitivity analysis for H_k [CDM17, Prop. 1].

1.3 Sensitivity Analysis With Multiple Outcomes

1.3.1 A Directional Global Null Hypothesis

There are K hypotheses $H_1, ..., H_K$, one each for the null of non-positive treatment effects for the kth outcome variable. We concern ourselves with a level-α sensitivity analysis for the global null hypothesis that all K of these hypotheses are true,

$$H_0 : \bigwedge_{k=1}^{K} H_k. \tag{4}$$

Through closed testing [MEG76], a valid sensitivity analysis for (4) also facilitates tests of the outcome-specific hypotheses H_k while controlling the familywise error rate. See [FS16, Section 5] for more on closed testing procedures applied to sensitivity analyses.

1.3.2 Linear Combinations of Test Statistics and Their Distribution

In what follows it is useful to define $\varrho_{ij} = \mathbb{P}\left(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}\right)$. Under the global null (4) and recalling that our test statistics are of the form $T_k = Z^T q_k$ with q_k fixed under the global null, the expectation $\mu(\varrho)$ and covariance $\Sigma(\varrho)$ for the vector of test statistics $T = (T_1, \ldots, T_K)^T$

are

$$\mu(\varrho)_k = \sum_{i=1}^{I} \sum_{j=1}^{n_i} q_{ijk}\varrho_{ij}, \quad \Sigma(\varrho)_{k,\ell}$$
$$= \sum_{i=1}^{I} \left\{ \sum_{j=1}^{n_i} q_{ijk}q_{ij\ell}\varrho_{ij} - \left(\sum_{j=1}^{n_i} q_{ijk}\varrho_{ij} \right) \left(\sum_{j=1}^{n_i} q_{ij\ell}\varrho_{ij} \right) \right\}.$$

For a given vector of probabilities ϱ, under suitable conditions on the constants q_{ijk} the distribution of T is asymptotically multivariate normal through an application of the Cramér-Wold device. That is, for any fixed nonzero vector $\lambda = (\lambda_1, ..., \lambda_K)^T$ the standardized deviate $\lambda^T\{T - \mu(\varrho)\}/\left\{\lambda^T\Sigma(\varrho)\lambda\right\}^{1/2}$ is asymptotically standard normal.

The actual values of ϱ are unknown to the practitioner due to their dependence on hidden bias. Instead, the constraints imposed by the sensitivity model (1) on ϱ can be represented by a polyhedral set. For a particular Γ this set, call it $\mathcal{P}_\Gamma$, contains vectors ϱ such that (i) $\varrho_{ij} \geq 0$ $(i = 1, ..., I; j = 1, ..., n_i)$; (ii) $\sum_{i=1}^{n_i} \varrho_{ij} = 1$ $(i = 1, ..., I)$; and (iii) $s_i \leq \varrho_{ij} \leq \Gamma s_i$ for some s_i $(i = 1, ..., I; j = 1, ..., n_i)$. Conditions (i) and (ii) simply reflect that ϱ_{ij} are probabilities, while the s_i terms in (iii) arise from applying a Charnes-Cooper transformation [CC62] to (2).

1.3.3 Multivariate Sensitivity Analysis via a Two-Person Game

Let $t = (t_1, ..., t_K)^T$ be the observed vector of test statistics. In this subsection only, suppose interest lies not in a test of (4), but rather in the narrower intersection null that Fisher's sharp null holds for all K outcome variables. For fixed λ, a large-sample sensitivity analysis for Fisher's sharp null could be achieved by minimizing the standardized deviate $\lambda^T\{t - \mu(\varrho)\}/\{\lambda^T\Sigma(\varrho)\lambda\}^{1/2}$ over all ϱ such that $\varrho \in \mathcal{P}_\Gamma$, and assessing whether the minimal objective value exceeds the appropriate critical value from a standard normal.

With λ pre-specified, the sensitivity analysis imagines what would happen if the worst-

case, adversarial bias at a given level of Γ were present. If the practitioner fixes the linear combination λ ahead of time, she has no further recourse against such adversarial attacks. The practitioner may instead consider a two-person game of the form

$$a^*_{\Gamma,\Lambda} = \min_{\varrho \in \mathcal{P}_\Gamma} \sup_{\lambda \in \Lambda} \frac{\lambda^T \{t - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho) \lambda\}^{1/2}}, \tag{5}$$

where Λ is some subset of $\mathbb{R}^K$ without the zero vector. The adversary may be thought of as embodying future counterclaims regarding the study's conclusions. In keeping with the scientific method the investigator recognizes that her conclusions will be subjected to challenges by her peers, and through the sensitivity analysis assesses whether a particular counterclaim could possibly overturn the study's findings. The critic aligns the unobserved confounders to inflate the p-value for the performed inference, while the investigator may choose weights for each outcome within the constraints imposed by Λ in response to the configuration of unmeasured confounders selected by the critic. With regards to (5), as Γ grows during the process of a sensitivity analysis the outer minimization takes place over a sequence of growing feasible regions. In the sense of the two-player game, this corresponds to the adversary having more and more flexibility in assigning unfavorable treatment allocation distributions.

Most familiar large-sample sensitivity analyses for Fisher's sharp null hypothesis are instances of this game for particular choices of Λ. Setting $\Lambda = \{e_k\}$ where e_k is a vector with a 1 in the kth coordinate and zeroes elsewhere returns a univariate sensitivity analysis for the kth outcome with a greater-than alternative, while $-e_k$ would return the less-than alternative. When the test statistics T_K are rank tests, setting $\Lambda = \{1_K\}$ where 1_K is a vector containing K ones returns the coherent rank test of [Ros97]. When $\Lambda = \{e_1, ..., e_K\}$, the collection of standard basis vectors, (5) returns the method of [FS16] with greater-than alternatives, and $\Lambda = \{\pm e_1, ..., \pm e_K\}$ gives the same method with two-sided alternatives. The method of [Ros16] amounts to a choice of $\Lambda = \mathbb{R}^K \setminus \{0_K\}$, i.e. all possible linear

combinations except the vector 0_K containing K zeroes.

While appealing as a unifying framework for multivariate sensitivity analyses, the form (5) would be of little practical use if the corresponding optimization problem could not be readily solved. The problem (5) is not itself convex; however, consider replacing it with

$$b^*_{\Gamma,\Lambda} = \min_{\varrho \in \mathcal{P}_\Gamma} \sup_{\lambda \in \Lambda} \max \left[0, \frac{\lambda^T \{t - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho) \lambda\}^{1/2}} \right]^2, \tag{6}$$

and let $f(\lambda, \varrho) = \max[0, \lambda^T\{t - \mu(\varrho)\}/\{\lambda^T\Sigma(\varrho)\lambda\}^{1/2}]^2$. This replaces negative values for the standardized deviate with zero, and then takes the square of the result. It is a monotone non-decreasing transformation of the standardized deviate in general, and is strictly increasing whenever the standardized deviate is larger than zero. The following proposition, proved in the supplementary material, establishes convexity of (6).

Proposition 1.1. *The function $g(\varrho) = \sup_{\lambda \in \Lambda} f(\lambda, \varrho)$ is convex in ϱ for any set Λ without the zero vector.*

The proof requires showing that for any $\lambda \in \Lambda$, the function $f(\lambda, \varrho)$ is convex in ϱ. As the pointwise supremum over a potentially infinite set of convex functions is itself convex [BV04, Section 3.2.3], the result then follows. The convexity of $g(\varrho)$ allows for its minimization over the polyhedral set $\mathcal{P}_\Gamma$ such that the value $b^*_{\Gamma,\Lambda}$ in (6) can be computed in practice. For any Γ and Λ, a sensitivity analysis through (5) would proceed by comparing the value $a^*_{\Gamma,\Lambda}$ to a suitable critical value $c_{\alpha,\Lambda}$. Observe that $a^*_{\Gamma,\Lambda} = (b^*_{\Gamma,\Lambda})^{1/2}$ for $a^*_{\Gamma,\Lambda} \geq 0$. If $\alpha \leq 0.5$ then $c_{\alpha,\Lambda}$ is non-negative for any choice of Λ. Consequently $a^*_{\Gamma,\Lambda} \geq c_{\alpha,\Lambda}$, leading to a rejection of the null, if and only if $b^*_{\Gamma,\Lambda} \geq c^2_{\alpha,\Lambda}$ so long as $\alpha \leq 0.5$. Through this equivalence, a large-sample sensitivity analysis using (5) can proceed through the solution of the convex program (6).

1.3.4 The Practitioner's Price

The critical value $c_{\alpha,\Lambda}$ depends on the structure of Λ, through which it is seen that additional flexibility in the set Λ does not come without a cost. Intuition for the price to be paid may be formed at $\Gamma = 1$ in (5). When Λ is a singleton the asymptotic reference distribution is the standard normal. If Λ is instead a finite set with $|\Lambda| = L > 1$ simply comparing the optimal value of (4) to the $1 - \alpha$ quantile of a standard normal would not provide a level$-\alpha$ test due to multiplicity issues. One could proceed using a Bonferroni correction based on the L comparisons, which would inflate the critical value. When $\Lambda = \mathbb{R}^K \setminus \{0_K\}$, [Ros16] applies a result on quadratic forms of multivariate normals [e.g. Rao73, page 60, 1f.1(i)] to show that one must instead use the square root of a critical value from a χ^2_K distribution when conducting inference through (5). This result underpins Scheffé's method for multiplicity control while comparing all linear contrasts of a multivariate normal (Sch53). In the potential presence of hidden bias, the additional flexibility afforded by a richer set Λ often offsets the loss in power from controlling for multiple comparisons, particularly in large samples. We discuss this further in Section 1.5.1, but see also (FS16, Section 6) and (Ros16, Section 4).

1.4 The Null Distribution Over Coherent Combinations

1.4.1 Adaptive Linear Combinations over the Non-negative Orthant

By allowing the set Λ to be arbitrary, the developments Section 1.3.3 were presented with Fisher's sharp null in mind. A moment's reflection reveals that should inference instead concern the composite null (4) of non-positive effects for all outcome variables, the set Λ must be constrained to maintain the desired size of the procedure. If Λ allows for arbitrary linear combinations, evidence consistent with non-positive treatment effects for each outcome variable may nonetheless result in a rejection of the null hypothesis based on (5) beyond the

nominal rate by setting λ_k negative for each k. Directional control is lost without constraining the signs of the elements of Λ.

Following (Ros02, Section 9.4), we define a family of coherent test statistics by restricting the vector λ to lie in the non-negative orthant, $\Lambda_+ = \{\lambda : \lambda_k \geq 0 \ \ (k = 1..., K); \ \ \sum \lambda_k > 0\}$. The coherent test of (Ros97) with $\lambda = 1_K$ is a particular element of Λ_+. We instead consider a large-sample sensitivity analysis for (5) with $\Lambda = \Lambda_+$, hence optimizing over the entire space of coherent linear combinations. We describe a projected subgradient descent method for solving (6) with $\Lambda = \Lambda_+$ in the supplementary material. By a duality argument of (Sha03), an application of Weierstrass' Extreme Value Theorem implies that the value of the inner supremum in (5) and (6) is guaranteed to be defined for any ϱ; however, there may not be a feasible $\lambda \in \Lambda_+$ which achieves this supremum as Λ_+ is a blunt cone since it does not contain the origin. This result holds true if Λ_+ is replaced with any non-empty cone Λ. Subgradients are straightforward to compute, and projections onto $\mathcal{P}_\Gamma$ are facilitated by the constraints being separable across matched sets.

Let $\tilde{\varrho}$ be the true, though typically unknown, vector of assignment probabilities and consider the random variable

$$A_{\Lambda_+}(Z, R) = \sup_{\lambda \in \Lambda_+} \ \frac{\lambda^T \{T - \mu(\tilde{\varrho})\}}{\{\lambda^T \Sigma(\tilde{\varrho}) \lambda\}^{1/2}}. \tag{7}$$

Let R_Z denote the observed responses when the treatment assignment is Z. Let $G(v, R_Z)$ be the reference distribution based on the observed outcome R_Z assuming Fisher's sharp null,

$$G(v, R_Z) = \sum_{b \in \Omega} 1\{A_{\Lambda_+}(b, R_Z) \leq v\} \mathrm{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}), \tag{8}$$

and let $G^{-1}(1-\alpha, R_Z)$ be its $1-\alpha$ quantile. Observe that the reference distribution $G(v, R_Z)$ itself varies over elements of Ω through its dependence on R_Z if Fisher's sharp null is false.

Proposition 1.2, proved in the supplementary material, states that a valid test of the

composite null of non-positive effects H_0 can be achieved through the randomization distribution of A_{Λ_+} under the assumption of Fisher's sharp null. Through an analogous proof, the randomization distribution also provides an unbiased test against positive alternatives of the form $\tau_{ijk} \geq 0$ $(i = 1, .., I; j = 1, ..., n_i; k = 1, ..., K)$ with at least one strict inequality.

Proposition 1.2. *Suppose that the global null (4) of non-positive treatment effects is true and assume that the test statistics T_k $(k = 1, ..., K)$ are effect increasing. Then*

$$pr\{A_{\Lambda_+}(Z, R_Z) \geq G^{-1}(1 - \alpha, R_Z)\} \leq \alpha,$$

such that the reference distribution under Fisher's sharp null controls the Type I error rate for any element of the composite null H_0.

Both the observed value $A_{\Lambda_+} = a_{\Lambda_+}$ and the probabilities $\text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z})$ are unknown in the observational study at hand due to their dependence on the true conditional assignment probabilities $\tilde{\varrho}$. Through the solution to (5) we instead observe the value a^*_{Γ,Λ_+}, which bounds a_{Λ_+} from below so long as $\tilde{\varrho} \in \mathcal{P}_\Gamma$. That said, the true randomization distribution (8) typically remains unknown outside of a randomized experiment as it depends on $\tilde{\varrho}$. For many test statistics, such as those formed when Λ is a singleton, the asymptotic reference distribution does not depend on $\tilde{\varrho}$ after suitable standardization. In what follows, we consider the large-sample distribution of A_{Λ_+} under Fisher's sharp null.

1.4.2 The Chi-Bar-Squared Distribution

Comparing the optimal value of (5) with $\Lambda = \Lambda_+$ to the $1 - \alpha$ quantile of a standard normal would not provide a valid level$-\alpha$ sensitivity analysis, as it would not account for the optimization over coherent combinations. While one could proceed with the square root of the $1 - \alpha$ quantile of a χ^2_K distribution, doing so would be unduly conservative. The χ^2_K critical value allows for optimization over all linear combinations, while here we

have constrained ourselves to combinations lying in the non-negative orthant. Theorem 1 provides the appropriate reference distribution given this restriction.

Theorem 1. *Suppose that $I^{-1}\Sigma(\tilde{\varrho})$ has an positive definite limit M as $I \to \infty$ and the random vector $\Sigma(\tilde{\varrho})^{-1/2}\{T - \mu(\tilde{\varrho})\}$ converges in distribution to a K-dimensional vector of independent standard normals. Then, as $I \to \infty$ the random variable $A^2_{\Lambda_+}$ converges in distribution to a $\bar{\chi}^2(M^{-1}, \Lambda_+)$ random variable under Fisher's sharp null.*

The proof is deferred to the supplementary material. The supplementary material also contains a discussion of sufficient conditions such that $\Sigma(\tilde{\varrho})^{-1/2}\{T - \mu(\tilde{\varrho})\}$ converges in distribution to a multivariate normal, which amount to assumptions about the vectors of constants q_k $(k = 1, ..., K)$. For instance, one sufficient condition would be to stipulate that $I^{-1}\sum_{i=1}^{I}\sum_{j=1}^{n_i} q_{ijk}^4$ is uniformly bounded for all $I \in \mathbb{N}$ and all $k = 1, \ldots, K$.

The $\bar{\chi}^2$ ("chi-bar-squared") is a common family of distributions arising in order restricted statistical inference (SS02). To illustrate, let X be a mean zero K-variate normal random vector with positive definite covariance matrix V, and define the random variable

$$\bar{\chi}^2(V, \Lambda_+) = X^T V^{-1} X - \inf_{\theta \in \Lambda_+} (X - \theta)^T V^{-1} (X - \theta). \tag{9}$$

Letting θ denote the mean vector of a multivariate normal, (9) is equivalent to the likelihood ratio statistic for testing the null $H_0 : \theta_k = 0$ $(k = 1, ..., K)$ versus the alternative $H_a : \theta_k \geq 0$ $(k = 1, ..., K)$ with strict inequality in at least one component (Kud63). Observe that replacing Λ_+ with $\mathbb{R}^K$ in (9) would return $X^T V^{-1} X$, and with it the usual χ^2_K distribution. Computation of (9) requires solving a quadratic program, an easy task with modern solvers but one which historically limited the adoption of methods requiring the $\bar{\chi}^2$ distribution.

The cumulative distribution function of the $\bar{\chi}^2(V, \Lambda_+)$ is

$$\text{pr}\{\bar{\chi}^2(V, \Lambda_+) \leq c\} = \sum_{i=0}^{K} w_i(V, \Lambda_+) \mathbb{P}\left(\chi^2_i \leq c\right),$$

a mixture of χ^2_i distributions ($i = 0, ..., K$) with χ^2_0 representing a pointmass at zero. The ith weight $w_i(V, \Lambda_+)$ is equal to the probability that the vector $V^{-1/2}X$ has exactly i positive components. The weights depend upon the covariance V through the corresponding correlation matrix C: any two covariance matrices V' and V with the same correlation structure C yield the same weights for $\bar{\chi}^2$ [SS05, Proposition 3.6.1 (11)]. See (Kud63, RWD88); and (SS05) for more on the role of the $\bar{\chi}^2$ distribution in multivariate one-sided testing.

(Sha03) presents an extension of Scheffé's method for multiple comparisons to linear combinations subject to cone constraints such as lying in the non-negative orthant. Arguments therein show that strong duality holds in (7), such that the optimal value for (7), A_{Λ_+}, equals the optimal value of the dual. The optimal solution to the dual is

$$A_{\Lambda_+} = \left\{ h^T \Sigma(\tilde{\varrho}) h - \inf_{\lambda \in \Lambda_+} (h - \lambda)^T \Sigma(\tilde{\varrho})(h - \lambda) \right\}^{1/2}, \tag{10}$$

where $h = \Sigma^{-1}(\tilde{\varrho})\{T - \mu(\tilde{\varrho})\}$. Under mild conditions h is asymptotically multivariate normal with covariance equal to the limit of $I\Sigma^{-1}(\tilde{\varrho})$. Comparing (10) to (9) provides intuition for the $\bar{\chi}^2$ limiting distribution. Moving forwards, we refer to the procedure using $A_{\Lambda+}$ to facilitate inference as the $\bar{\chi}^2$-test. In the supplementary material, we present Type I error control simulations indicating that the $\bar{\chi}^2$ reference distribution provides a reasonable approximation to the true randomization distribution of $A_{\Lambda+}$ with moderate sample sizes.

1.4.3 The Critical Value and Its Dependence on the Unknown Assignment Probabilities

A large-sample sensitivity analysis can be conducted by comparing the optimal value of (5) over coherent linear combinations, a_{Γ,Λ_+} to the square root of the $1 - \alpha$ quantile of a $\bar{\chi}^2\{\Sigma^{-1}(\tilde{\varrho}), \Lambda_+\}$ distribution. Recalling that $\tilde{\varrho}$ is the true vector of assignment probabilities, we are faced with a difficulty encountered by neither a univariate sensitivity analysis for

a particular outcome nor the method of (Ros16): The asymptotic reference distribution depends on the assignment probabilities $\tilde{\varrho}$ through the covariance $\Sigma(\tilde{\varrho})$ even after proper normalization. While $\tilde{\varrho}$ is known in a randomized experiment, the purpose of a sensitivity analysis is to assess robustness of a study's findings as $\tilde{\varrho}$ is allow to vary within bounds imposed by Γ.

The dependence of the covariance on nuisance parameters is commonly encountered in applications of the $\bar{\chi}^2$ distribution (SS02, Section 2.2). One solution is to compute p-values through the bound $\text{pr}\{\bar{\chi}^2(V, \Lambda_+) \geq c\} \leq 0.5\{\mathbb{P}\left(\chi^2_{K-1} \geq c\right) + \mathbb{P}\left(\chi^2_K \geq c\right)\}$; see (Per69, Theorem 6.2) for a proof. This upper bound is attained in the limit as the correlation between all outcomes converges to one, and can itself be quite conservative in the presence of more moderate degrees of correlation typically observed in practical applications.

Motivated by the particular structure imposed by a sensitivity analysis, we instead use a two-stage procedure to better upper bound the worst-case critical value for each Γ. In a sensitivity analysis the range of the nuisance parameters $\tilde{\varrho}$ is controlled by Γ. At $\Gamma = 1$ $\tilde{\varrho}$ is entirely specified, such that in finely stratified experiments the appropriate $\bar{\chi}^2$ distribution is known. As Γ increases the bounds imposed by membership in $\mathcal{P}_\Gamma$ widen. For each pair of outcomes k and ℓ, we first find upper and lower bounds on the correlation between k and k' given $\tilde{\varrho} \in \mathcal{P}_\Gamma$, call them $C^{(\ell)}_{k,k',\Gamma}$ and $C^{(u)}_{k,k',\Gamma}$. We then maximize the $1-\alpha$ quantile of a $\bar{\chi}^2(C^{-1}, \Lambda_+)$ distribution over the correlation matrix C subject to $C^{(\ell)}_{k,k',\Gamma} \leq C_{k,k'} \leq C^{(u)}_{k,k',\Gamma}$ for all k, k' and C being a correlation matrix. See the supplementary material for implementation details along with a discussion of the case $K = 2$, where it is seen that the worst-case critical value is attained at the lower bound on the correlation. In practice, we find that this can provide meaningful improvements in the power of the procedure; see the supplementary material for an illustration.

1.5 Design Sensitivity And Power For The Chi-Bar Squared Test

1.5.1 Design Sensitivity

Suppose that the treatment in question actually has an effect in the direction of the alternative, and further that there is truly no hidden bias such that inference at $\Gamma = 1$ would be justified. As would be the case in practice, the researcher analyzing the observational study is unaware of these favorable conditions. Thus, she would like to reject the null hypothesis not only under the assumption of no unmeasured confounding, but also for values $\Gamma > 1$ to assess whether the rejection of the null is robust to certain degrees of hidden bias. The power of a level-α sensitivity analysis is the probability that the procedure correctly rejects the null hypothesis at some pre-specified value of $\Gamma \geq 1$. In what follows we will assume a stochastic generative model for the outcome variables, an assumption which greatly simplifies power calculations.

Under mild conditions, there is a value $\tilde{\Gamma}$ such that the power of a sensitivity analysis converges to one for all $\Gamma < \tilde{\Gamma}$, and converges to zero for all for all $\Gamma > \tilde{\Gamma}$; this value is called the design sensitivity of the test (Ros04). It quantifies the asymptotic ability of the test to discriminate treatment effect under the concern of bias in the treatment allocation process, and can vary substantially across choices of test statistics. For a fixed data generating model, a test with high design sensitivity is preferable to a test with low design sensitivity.

For fixed choices of the univariate test statistics $T_k = Z^T q_k$ $(k = 1, ..., K)$, we consider the design sensitivity of multivariate tests based upon (5) and their dependence on the set Λ. Theorem 2 shows that design sensitivity is a monotonic non-decreasing function with respect to the partial ordering over sets Λ given by inclusion.

Theorem 2. *Suppose* $\Lambda_1 \subseteq \Lambda_2$. *Under mild conditions, the design sensitivity of* (5) *using* $\Lambda = \Lambda_1$ *is less than or equal to the design sensitivity of* (5) *using* $\Lambda = \Lambda_2$.

The proof of Theorem 2 is deferred to the supplementary material. In light of Theorem 2, it may be tempting to take $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ in order to achieve the greatest design sensitivity. While this would result in a valid test of Fisher's sharp null, it does not provide a valid test of the null hypothesis of non-positive treatment effects: should the signs of λ_k be left unconstrained, evidence of a negative treatment effect may result in a large optimal value for (5). Restricting attention to the set of coherent linear combinations Λ_+, Theorem 2 gives rise to the following optimality property for the $\bar{\chi}^2$-test due to its optimizing over the entirety of Λ_+.

Corollary 1. *The $\bar{\chi}^2$-test achieves greatest design sensitivity among coherent tests based upon* (5) *with $\Lambda \subseteq \Lambda_+$*

The balance between using Λ_+ and using further constrained conic subsets of Λ_+ is informed by the practitioner's subject knowledge: constraints added to the λs do reduce design sensitivity relative to using Λ_+, but these constraints may be advisable when they reflect important subject-specific scientific knowledge. Changing the structure of the feasible region may necessitate changing the critical value of the test.

1.5.2 Finite-sample Power for Rejecting the Global Null

Corollary 1 illustrates that despite the larger critical value necessitated by the $\bar{\chi}^2$-test by optimizing over Λ_+, the $\bar{\chi}^2$-tests achieves the largest possible design sensitivity over the set of coherent multivariate tests. This reflects that in large samples bias trumps variance in the analysis of observational studies, such that the differences in critical values are rendered irrelevant in the limit. In moderate samples, the variance of the null distribution plays a larger role in the power of a sensitivity analysis, such that differences in critical values can make a more substantial difference for procedures with similar design sensitivities.

We present a simulation study comparing the power of a sensitivity analysis based upon the $\bar{\chi}^2$-test to two competitors: the method of (FS16); and the test using (5) with $\Lambda = \{1_K\}$,

which we refer to as the equal-weight test. Combining test statistics with equal weights is only sensible when the constituent test statistics T_k $(k = 1, .., K)$ reflect evidence against the null hypothesis on the same scale. This would be true of rank statistics as described in (Ros97), and would also be true of suitably scaled m-statistics of the type described in (Ros07); however, if one outcome is tested using a rank-sum statistic and another with an m-statistic for instance, the "equal-weight" test would give unreasonable weight to the rank-sum recorded outcome. The $\bar{\chi}^2$-test and the test of (FS16) do not require comparable scales for the test statistics as they are scale invariant.

The simulations are performed on $I = 300$ matched pairs with $K = 3$ outcomes. In each simulation, we generate I mean-zero unit-variance trivariate normal vectors of noise $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ equicorrelated with correlation ρ. We then create the vector of treated-minus-control paired differences in outcomes as $(Y_{i1}, Y_{i2}, Y_{i3})^T = (\tau_1, \tau_2, \tau_3)^T + (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ for different values of the treatment effects $(\tau_1, \tau_2, \tau_3)^T$. For each outcome variable, the employed test statistic is $T_k = \sum_{i=1}^{I} \text{sign}(Y_{ik}) \min(|Y_{ik}|/s_k, 2.5)$, where s_k is the median of $|Y_{ik}|$ $(i = 1, .., I)$. This amounts to a choice of a m-statistic with Huber's ψ-function, as described in (Ros07).

Table 1.1 presents the values of the treatment effects and the correlation employed in the simulation study. For each combination of parameters, it further provides the design sensitivity for the $\bar{\chi}^2$-test and the equal-weight test. While there is no known formula for the design sensitivity of the procedure of (FS16), it is lower-bounded by the the maximal design sensitivities of the three univariate tests; this value is also presented in the table. The table reflects Corollary 1: for each combination of parameters, the design sensitivity for the $\bar{\chi}^2$-test is greater than or equal to that of the equal-weight test and the maximal univariate test. Further, there is no consistent ordering between the equal-weight test and the max of the univariate tests, as the corresponding sets Λ for neither test is a subset of the other.

Figure 1-1 presents the estimated power curves of the three tests as a function of $\Gamma > 1$ in these simulation settings at $I = 300$, with 2000 simulations for each combination of param-

	$\bar{\chi}^2$-Test		Equal-Weight Test		Max Univariate	
	$\rho = 0$	$\rho = 0.2$	$\rho = 0$	$\rho = 0.2$	$\rho = 0$	$\rho = 0.2$
$\tau = (0.25, 0.25, 0.25)^T$	2.9	2.4	2.9	2.4	1.9	1.9
$\tau = (0.10, 0.10, 0.50)^T$	3.6	3.4	2.6	2.2	3.4	3.4
$\tau = (0.02, 0.20, 0.50)^T$	3.8	3.5	2.8	2.4	3.4	3.4

Table 1.1: Design sensitivities for $\bar{\chi}^2$-test, the equal-weight test, and the largest of the three univariate tests under both independence and moderate positive correlation between outcomes.

eters. The correlation between paired differences varies across the columns from $\rho = 0$ (left) to $\rho = 0.2$ (right), while the treatment effects vary down the rows. The first row corresponds to $\tau_1 = \tau_2 = \tau_3$ and I, and here it is seen that the equal-weight test outperforms both the $\bar{\chi}^2$-test and (FS16). When the treatment effects are equal the linear combination $\lambda = 1_K$ attains the largest design sensitivity, and by restricting Λ to only this linear combination the lower critical value employed by the equal-weighted test improves power over that attained by the $\bar{\chi}^2$-test. When one of the three outcomes is strongly affected by the treatment while the other two are minimally impacted, as in the second row of the figure, the method of (FS16) and the $\bar{\chi}^2$-test perform similarly, while the equal-weight test lags behind. The test statistics returned by the $\bar{\chi}^2$-test are larger, but this is offset relative to the method of (FS16) by the larger critical value necessitated. When the treatment effects are staggered between the three outcomes as in the third row, the $\bar{\chi}^2$-test outperforms both (FS16) and the equal-weight test, particularly in the case of independence between the outcome variables. Optimizing over Λ_+ increases the value of the test statistic over both competitors, such that the flexibility is well worth the price of a larger critical value.

The simulations indicate that while the $\bar{\chi}^2$-test must have optimal power in the limit as asserted by Corollary 1, it need not have the best finite-sample performance. In some cases the equal-weight test can outperform it, while in others it is outperformed by the method of (FS16). Importantly the $\bar{\chi}^2$-test was never the worst of the three methods considered, and the simulations show that *a priori* restricting the set of combinations Λ under consideration

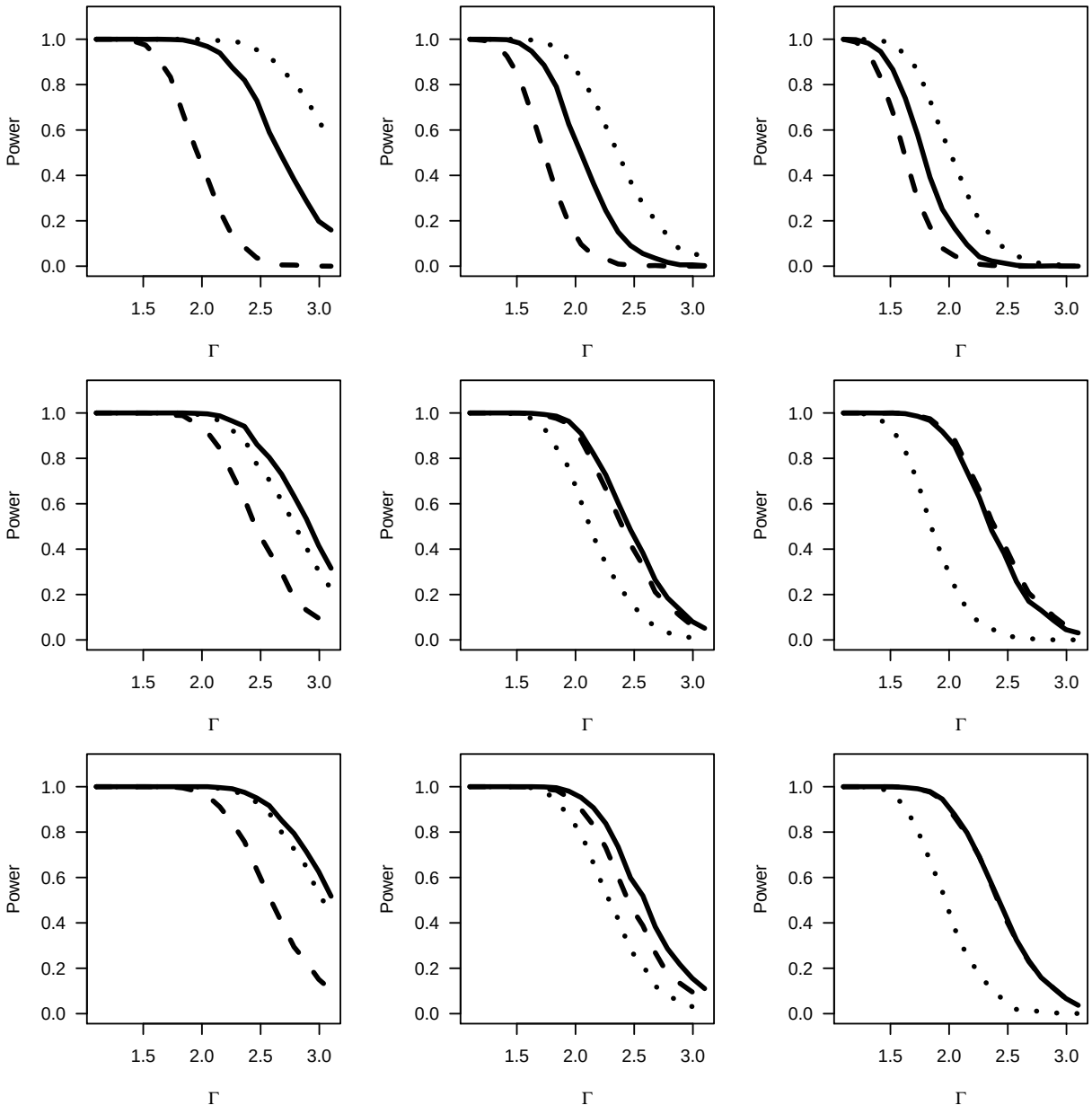

Figure 1-1: Power comparisons between the method of (FS16) (dashed), the $\bar{\chi}^2$-test of this paper (solid), and the equal-weight test (dotted) as Γ increases with $I = 300$. The first row has $\tau_1 = \tau_2 = \tau_3 = 0.25$; the second row has $\tau_1 = \tau_2 = 0.1$ and $\tau_3 = 0.5$; and the third row has $\tau_1 = 0.05$, $\tau_2 = 0.2$, and $\tau_3 = 0.5$. The left column has $\rho = -0.2$, the center has $\rho = 0$, and the right has $\rho = 0.2$.

can substantially reduce power should the choice of Λ be poor. For instance, the equally-weighted test performs poorly in the second and third rows of Figure 1-1, while the method of (FS16) is markedly worse than the other methods in the first row. The $\bar{\chi}^2$-test does pay a price in terms of an increased critical value, but this price acts as insurance against an unwise choice of Λ. Theorem 2 offers asymptotic assurance that the $\bar{\chi}^2$-test performs optimally in terms of design sensitivity; furthermore, the results of Table 1.1 and Figure 1-1 demonstrate that the $\bar{\chi}^2$-test performs well across a broad range of treatment effect regimes without sacrificing asymptotic optimality.

In the supplementary material, we present additional simulations with $I = 1000$ matched pairs which begin to show convergence of behavior of the tests under comparison to their design sensitivities. We further illustrate the potential for improvements in power for testing outcome-specific null hypotheses through incorporating the $\bar{\chi}^2$-test into a closed testing framework, as described in (FS16, Section 6).

1.6 Illustrations Of Multivariate One-Sided Sensitivity Analysis

1.6.1 The Role of Coherence in Two Observational Studies

We now consider the role of multiple outcomes in two observational studies. Both examples are drawn from The National Health and Nutrition Examination Survey (NHANES) and study physiological impacts of cigarette smoking. One study investigates the impact of smoking on two measures of periodontal disease, while the other looks at whether smoking increases urinary metabolite levels of four carcinogens. In both examples, the alternative hypothesis is that smoking should have a positive treatment effect on each of the outcome variables measured. Rosenbaum remarks that "If incoherence presents a substantial obstacle to a claim that the treatment caused its ostensible effects, then the absence of incoherence -

that is, coherence - should entail some strengthening of that claim" (Ros10, p. 119). Should the evidence suggest ostensible effects of smoking incompatible with positive effects for each outcome variable, smoking's place in the causal pathway would be cast into doubt. Should the outcomes all be affected in the predicted direction, this would provide further evidence for smoking's role in the causal mechanism.

In both observational studies and for each outcome variable, we use an m-test based upon Huber's ψ-function to conduct inference with the default choices for parameters in the `senmv` function in the `sensitivitymv` package in `R`.

1.6.2 Smoking and Periodontal Disease

It has been suggested that up to 42% of cases of periodontal disease can be attributed to smoking (TA00); however, as the evidence is observational in nature this association may well be explained away by other intrinsic differences between smokers and non-smokers. Using the 2011-2012 NHANES survey, (Ros16) paired $I = 441$ smoking individuals to non-smokers who were similar on the basis of education, income, race, age and gender. Two outcome variables pertaining to dental health were recorded, one each for upper and lower teeth. In this context, coherence would amount to demonstrating that smoking negatively impacted dental health in both the upper and lower teeth. Such a coherent hypothesis strengthens the causal claim that cigarette smoking is detrimental to periodontal health. Should smoking only appear to impact upper teeth but not lower teeth, for instance, such incoherence would cast into doubt whether smoking is truly to blame.

At $\alpha = 0.05$, the overall null hypothesis of non-positive treatment effects was rejected up until $\Gamma = 2.36$ when using the $\bar{\chi}^2$-test, while the equal-weight test was able to reject until $\Gamma = 2.54$. By selecting $\Lambda = \Lambda_+$ Theorem 1 gives that the appropriate asymptotic null distribution is the $\bar{\chi}^2$ distribution, while for the equal-weight test the asymptotic null distribution is the standard normal. The $1 - \alpha$ quantile of the standard normal lies below

the square root of the $1-\alpha$ quantile of any $\bar{\chi}^2$ distribution, such that the equal-weight test is able to employ a smaller critical value. With this restriction comes the risk that equally weighting the outcome variables may be suboptimal. In this particular observational study, it comes as little surprise that with periodontal disease in upper and lower teeth the risk was worth the while: there is little reason to suspect that magnitude of effects on upper and lower teeth should differ. Sensitivity analysis using the method of (FS16) achieves significance up to $\Gamma = 2.32$, a slightly lower value than the $\bar{\chi}^2$-test. The method of (FS16) takes Λ as the set of standard unit basis vectors for $\mathbb{R}^2$ in this case, and does not combine the two related measures of periodontal disease. Despite the method of (FS16) also having a smaller critical value than the $\bar{\chi}^2$-test, in this example this was offset by the additional flexibility afforded by the $\bar{\chi}^2$-test in optimizing over Λ_+. The sensitivity analysis using the $\bar{\chi}^2$-test took 13 seconds to complete on a personal laptop with a 2.60GHz processor with 16GB of RAM for this data set.

1.6.3 Smoking and Polycyclic Aromatic Hydrocarbons

Polycyclic Aromatic Hydrocarbons (PAHs) are a class of organic compounds formed during incomplete combustion which have been labeled potentially carcinogenic to humans (BGH$^+$02). We examine urinary concentrations of four different PAH metabolites in 432 smokers and 1206 non-smokers recorded in NHANES 2007-2008. The four metabolites are 1-hydroxyphenanthrene (1-Phen), 3-hydroxyphenanthrene (3-Phen), 1-hydroxypyrene (1-Pyr), and 9-hydroxyfluorene (9-Fluo). Full matching (Han04) was employed to adjust for a host of measured covariates thought to impact one's decision to smoke and one's exposure to PAHs; see the supplementary material for additional details. We then proceed with inference assessing whether cigarette use increases urinary concentrations of these four PAH metabolites. As tobacco smoke contains all of these PAHs, an incoherent result that none or only some urinary concentrations of PAH metabolites are higher in smokers than in non-smokers be

discovered would cast into question whether the association between smoking cigarettes and urinary PAH concentrations was actually causal. At $\alpha = 0.05$, a sensitivity analysis using the $\bar{\chi}^2$-test yielded significance up to $\Gamma = 6.28$, whereas for the equal-weight test with $\Lambda = \{1_K\}$ the sensitivity analysis was only able to reject up to $\Gamma = 5.38$. Despite the smaller critical value, in this case restricting oneself to equal weighting led to a markedly lower changepoint value of Γ than did the $\bar{\chi}^2$-test. The method of (FS16) rejected until $\Gamma = 6.18$.

At $\Gamma = 6.28$, our procedure for upper bounding the worst-case critical value for the $\bar{\chi}^2$-test as described in Section 1.4.3 returns a bound of 2.20 for the test based upon $a^*_{6.18,\Lambda_+}$ in (5). To illustrate the improvements from this approach, the square root of the 0.95 quantile of a χ^2_4 is 3.08, while employing the conservative bound from (Per69, Theorem 6.2) yields a critical value of 2.96. The $\bar{\chi}^2$ sensitivity analysis ran in about 20 minutes on a personal laptop with a 2.60GHz processor with 16GB of RAM. The length of runtime is dependent upon several factors including the number of strata, the size of the strata, the number of outcome variables, and the number of values of Γ tested in the sensitivity analysis.

1.6.4 Improvements in Tests of Individual Null Hypotheses

Rejecting the global null hypothesis confirms to the experimenter that at least one of the outcome variables is impacted by treatment in the direction of the alternative. However in order to appraise a coherent pattern of treatment impact an experimenter will need to examine the local null hypotheses of treatment impact upon each of the outcomes individually. Correcting for multiple comparisons can be facilitated through many techniques; here we juxtapose embedding the $\bar{\chi}^2$-test into a closed testing framework against performing K individual sensitivity analyses, one for each outcome variable, while employing a Bonferroni correction.

Table 1.2 details the changepoint Γ values for each individual outcome of the periodontal data and the PAH data while controlling the familywise error rate at $\alpha = 0.05$. The $\bar{\chi}^2$-test

	Periodontal Disease		Polycyclic Aromatic Hydrocarbons			
	Lower Teeth	Upper Teeth	1-Phen	3-Phen	1-Pyr	9-Fluo
Closed Testing	2.26	1.82	2.13	5.28	5.25	5.78
Bonferroni	2.17	1.76	1.99	4.88	4.84	5.31
Uncorrected	2.26	1.82	2.13	5.28	5.25	5.78

Table 1.2: Comparison of the closed test changepoint Γ versus Bonferroni corrected sensitivity analysis changepoint Γ for the data examples at $\alpha = 0.05$. The last row is the benchmark given by conducting individual tests at α without correction for multiplicity.

embedded into a closed testing framework outperformed the Bonferroni corrected tests for each outcome. Table 1.2 also includes the changepoint Γ values returned by the univariate sensitivity analyses without a Bonferroni correction, i.e. with each outcome tested at $\alpha = 0.05$. The table reveals that through embedding the $\bar{\chi}^2$-test in a closed testing procedure, in both studies we are able to report the same robustness to unmeasured confounding that would have been attained had we not controlled for multiple comparisons in the first place. Due to the improvements in power along the closed testing path furnished by the $\bar{\chi}^2$-test, there is no cost for evaluating coherence of all outcome variables relative to the best univariate outcome analysis.

1.7 Discussion

While we have tailored our presentation to continuous outcome variables, our test is equally applicable with binary outcomes and ordinal outcomes. In fact, potential outcomes of any partially ordered set are amenable to this composite null, and the remaining proofs of this paper hold true so long as the test statistics considered are effect increasing. See (Ros02, Section 2.8.5) for more on effect increasing statistics for partially ordered outcomes. The composite null H_k for the kth outcome variable requires an ordered structure to the potential outcomes, and since Proposition 1.2 relies only upon effect-increasingness of the test statistic $T_k(\cdot, \cdot)$, the result remains valid as long as one has a suitable partial ordering for the values

of the potential outcomes.

The $\bar{\chi}^2$-test we develop is not immediately applicable to testing Neyman's weak null. Interestingly, even assuming strong ignorability as would be the case in a randomized experiment, it is possible for the Type I error rate to exceed α under the weak null. The procedure we present uses a critical value from the asymptotic form of a randomization distribution assuming the sharp null as the sharp null attains the supremum p-value over H_0 in (4). If instead only Neyman's weak null is true for all K outcomes but H_0 is not it is possible that unspecified effect heterogeneity would cause the reference distribution used by our procedure to not stochastically dominate the randomization distribution, leading to an invalid procedure. Unlike the univariate case and the multivariate case with two-sided alternatives, a simple studentization does *not* fix the problem even asymptotically, as the studentized reference distribution depends upon the correlation between the outcome variables. This parallels known results for multivariate permutation tests conducted in the absence of a group invariance assumption (CR16). An ongoing area of the authors' research is examining the extent to which bootstrap prepivoting may be applicable to create a test that is both exact under H_0 and asymptotically valid for Neyman's weak null at $\Gamma = 1$, but as of yet no extension to cases of potential unmeasured confounding has been developed. The extension of sensitivity analyses to such contexts remains an interesting and important open question.

Our use of the $\bar{\chi}^2$-test in conjunction with closed testing provides a sensitivity analysis for testing patterns of directed effect among a moderate number of outcomes, as is common in many public health, econometric, and policy applications. Unfortunately, the combinatorial blow-up inherent to closed testing prohibits large-scale multiplicity control of the sort required for applications to data sets of the scale encountered in genome-wide association studies. Even in regimes for which closed testing is computationally infeasible, the interpretation of sensitivity analyses as two-player games lends meaningful intuition and will hopefully stimulate further algorithmic development.

Another quasiexperimental device related to the pattern-specificity approach taken in

this paper is the multiple evidence factor approach of (Ros17), though some important mathematical differences exist between the two approaches: most notably, the approach taken in this paper does not rely upon the group-theoretic structure of the symmetry group of the possible treatment allocations.

Supplementary Material

Below we include additional information which contains theoretical results, proofs, simulation studies, further algorithmic details, additional insight into the $\bar{\chi}^2$ distribution, further information on the observational study on smoking and polycyclic aromatic hydrocarbons, and an `R` script for implementing the method proposed in this work.

1.8 Proof Of Main Results

1.8.1 Proposition 1.1

Proposition 1.1. *The function $g(\varrho) = \sup_{\lambda \in \Lambda} f(\lambda, \varrho)$ is convex in ϱ for any set Λ without the zero vector.*

In order to show that (6) is convex in ϱ we first prove a lemma.

Lemma 1.1. *For a fixed $\lambda \in \mathbb{R}^K$ the function $d(\varrho) = \lambda^T \Sigma(\varrho) \lambda$ is a concave function of ϱ.*

Proof of Lemma 1.1. Define Q_i to be the K-by-n_i matrix where the (k, j)th entry is q_{ijk}. Then the Hessian matrix of $d(\varrho)$ with respect to the variables in the ith strata is

$$\nabla^2_{\varrho_{ij};\,(j=1,\ldots,n_i)} d(\varrho) = \frac{-1}{2} Q_i^T \lambda \lambda^T Q_i.$$

This is negative semi-definite. By independence between strata, the full Hessian $\nabla^2_{\varrho} f(\varrho)$ is the direct sum of the Hessians associated to each stratum. Thus, the full Hessian matrix is a block diagonal matrix wherein each block is negative semi-definite. Since the eigenvalues of a block diagonal matrix are the collection of eigenvalues of its constituent blocks, we have that the full Hessian must be negative semi-definite as well. As a consequence, $d(\varrho)$ is a concave function of ϱ. □

Proof of Proposition 1.1. The identity function $x \mapsto x$ is convex as a function of x. Since the point-wise maximum of convex functions is convex $\max\{0, x\}$ is convex as a function of x. The quadratic function $a \mapsto a^2$ is convex and increasing on the non-negative real line so by (BV04, 3.10) the function $\psi(x) = [\max\{0, x\}]^2$ is a convex function of x.

The perspective of a function $\psi(x)$ is defined to be $\phi(x, v) = v\psi(x/v)$ for $v > 0$; by (BV04, 3.2.6) the perspective of a convex function is convex as well. Computing the perspective of ψ follows as

$$\begin{aligned}
\phi(x, v) &= v\psi(x/v) \\
&= v\left\{\max(0, x/v)\right\}^2 \\
&= v\left\{\frac{\max(0, x)}{v}\right\}^2 \\
&= \frac{\max(0, x)^2}{v}.
\end{aligned}$$

Thus, $\phi(x, v) = \max(0, x)^2/v$ is convex. Now, consider any fixed $\lambda \geq 0$ and t of dimension K. $\mu(\varrho)$ is a linear function of ϱ. Since affine transformations of linear functions are convex, $\lambda^T\{t - \mu(\varrho)\}$ is convex. Furthermore, $\lambda^T\Sigma(\varrho)\lambda$ is concave in ϱ by Lemma 1.1. By (BV04, 3.15), since $\phi(x, v)$ is non-decreasing in x and non-increasing in v the function

$$f(\lambda, \varrho) = \phi\left(\lambda^T\{t - \mu(\varrho)\}, \lambda^T\Sigma(\varrho)\lambda\right) = \frac{\max\left[0, \lambda^T\left\{t - \mu(\varrho)\right\}\right]^2}{\lambda^T\Sigma(\varrho)\lambda}$$

is convex in ϱ. As $g(\varrho)$ is the point-wise supremum over all $\lambda \in \Lambda$ of $f(\lambda, \varrho)$, by (BV04, 3.7) $g(\varrho)$ is convex in ϱ as desired. The requirement that Λ excludes the zero vector ensures that for any positive definite $\Sigma(\varrho)$ the denominator is always defined and thus $g(\varrho)$ is defined. $\square$

1.8.2 Proposition 1.2

Here and elsewhere in the supplement, Λ_+ is once again defined to be the non-negative orthant in $\mathbb{R}^K$ excluding the zero vector, that is $\Lambda_+ = \{\lambda : \lambda_k \geq 0 \;\; (k = 1..., K); \;\; \sum \lambda_k > 0\}$.

Proposition 1.2. *Suppose that the global null (4) of non-positive treatment effects is true and assume that the test statistics T_k $(k = 1, ..., K)$ are effect increasing. Then*

$$pr\{A_{\Lambda_+}(Z, R_Z) \geq G^{-1}(1 - \alpha, R_Z)\} \leq \alpha,$$

such that the reference distribution under Fisher's sharp null controls the Type I error rate for any element of the composite null H_0.

Proof.

$$\begin{aligned}
&\text{pr}\{A_{\Lambda_+}(Z, R_Z) > G^{-1}(1 - \alpha, R_Z)\} \\
&= \sum_{z \in \Omega} 1\{A_{\Lambda_+}(z, R_z) > G^{-1}(1 - \alpha, R_z)\} \text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \\
&= \sum_{b \in \Omega} \left[\sum_{z \in \Omega} 1\{A_{\Lambda_+}(z, R_z) > G^{-1}(1 - \alpha, R_z)\} \text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \right] \text{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}) \\
&\leq \sum_{b \in \Omega} \left[\sum_{z \in \Omega} 1\{A_{\Lambda_+}(b, R_z) > G^{-1}(1 - \alpha, R_z)\} \text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \right] \text{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}) \\
&= \sum_{z \in \Omega} \left[\sum_{b \in \Omega} 1\{A_{\Lambda_+}(b, R_z) > G^{-1}(1 - \alpha, R_z)\} \text{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}) \right] \text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \\
&\leq \alpha \sum_{z \in \Omega} \text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = \alpha.
\end{aligned}$$

The third line simply multiplies by one in the form of $\sum_{b \in \Omega} \text{pr}(Z = b \mid \mathcal{F}, \mathcal{Z})$. The fourth line uses that the test statistics are effect increasing. After rearranging the order of summation in the fifth line, the sixth follows by definition as it simply uses that for any particular z, $G^{-1}(1 - \alpha, R_z)$ is the $1 - \alpha$ quantile corresponding to $G(v, R_z) = \sum_{b \in \Omega} 1\{A_{\Lambda_+}(b, R_z) \leq$

$v\}\text{pr}(Z = b \mid \mathcal{F}, \mathcal{Z})$.

□

1.8.3 Theorem 1

For ease of notation we suppress conditioning on $\mathcal{F}$ and $\mathcal{Z}$ when writing expectations and covariances in this section. We again define $T_k = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Z_{ij} q_{ijk}$, and let $\tilde{\varrho}$ represent the true vector of conditional assignment probabilities. For precision quantities such as $\tilde{\varrho}$ should be subscripted by I to denote their dependence on the sample size; this is omitted for improved readibility.

Theorem A.3. *Suppose that $I^{-1}\Sigma(\tilde{\varrho})$ has a positive definite limit M as $I \to \infty$ and the random vector $\Sigma(\tilde{\varrho})^{-1/2}\{T - \mu(\tilde{\varrho})\}$ converges in distribution to a K-dimensional vector of independent standard normals. Then, as $I \to \infty$ the random variable $A^2_{\Lambda_+}$ converges in distribution to a $\bar{\chi}^2(M^{-1}, \Lambda_+)$ random variable under Fisher's sharp null.*

Before proving Theorem 1, we establish conditions under which the random vector $\Sigma(\tilde{\varrho})^{-1/2}\{T - \mu(\tilde{\varrho})\}$ has a multivariate normal limiting distribution.

Lemma 1.2. *Suppose that there exists a $\delta > 0$ for which*

$$\sum_{i=1}^{I} \mathbb{E}\left[\mid \sum_{j=1}^{n_i} q_{ijk} Z_{ij} - \sum_{j=1}^{n_i} q_{ijk} \varrho^*_{ij} \mid^{2+\delta} \right] = O(I) \tag{11}$$

for all k and all I, and that $I^{-1}\Sigma(\tilde{\varrho})$ has an positive definite limit M as $I \to \infty$. Then as $I \to \infty$ the random vector $\Sigma(\tilde{\varrho})^{-1/2}\{T - \mu(\tilde{\varrho})\}$ converges in distribution to a K-variate vector of independent standard normals.

Proof of Lemma 1.2. Define $X_i = (X_{i1}, \ldots, X_{iK})^T$ where $X_{ik} = \sum_{j=1}^{n_i} q_{ijk} Z_{ij}$. Denote $\mu_i(\tilde{\varrho}) = \mathbb{E}[X_i]$ and $\Sigma_i(\tilde{\varrho}) = E(X_i X_i^T) - E(X_i)E(X_i)^T$, such that $\sum_{i=1}^{I} \mu_i(\tilde{\varrho}) = \mu(\tilde{\varrho}) = E(T)$ and $\sum_{i=1}^{I} \Sigma_i(\tilde{\varrho}) = \Sigma(\tilde{\varrho}) = \text{cov}(T)$.

By the Cramér-Wold device it suffices to consider the distribution of the univariate random variable $I^{-1/2}\sum_{i=1}^{I}\lambda^T\{X_i-\mu_i(\tilde{\varrho})\}$ for a fixed, non-zero, $\lambda\in\mathbb{R}^K$. By independence between strata, the random variables $\lambda^T\{X_i-\mu_i(\tilde{\varrho})\}$ are independent but not necessarily identically distributed. The variance of $I^{-1/2}\sum_{i=1}^{I}\lambda^T\{X_i-\mu_i(\tilde{\varrho})\}$ is $I^{-1}\sum_{i=1}^{I}\lambda^T\Sigma_i(\tilde{\varrho})\lambda$. By hypothesis $I^{-1}\Sigma(\tilde{\varrho})$ has an positive definite limit M as $I\to\infty$ so

$$\lim_{I\to\infty}\frac{1}{\left(I^{-1}\sum_{i=1}^{I}\lambda^T\Sigma_i(\tilde{\varrho})\lambda\right)^{\frac{2+\delta}{2}}}=\frac{1}{(\lambda^TM\lambda)^{\frac{2+\delta}{2}}}>0. \tag{12}$$

Furthermore, (11) and the c_r-inequality imply that

$$\lim_{I\to\infty}I^{-\frac{2+\delta}{2}}\sum_{i=1}^{I}\mathbb{E}\left[\mid\sum_{j=1}^{i}q_{ijk}Z_{ij}-\sum_{j=1}^{i}q_{ijk}\varrho_{ij}\mid^{2+\delta}\right]=0. \tag{13}$$

Combining (12) and (13) gives that

$$\lim_{I\to\infty}\frac{1}{\left(\sum_{i=1}^{I}\lambda^T\Sigma_i(\tilde{\varrho})\lambda\right)^{\frac{2+\delta}{2}}}\sum_{i=1}^{I}\mathbb{E}\left[\mid\sum_{j=1}^{i}q_{ijk}Z_{ij}-\sum_{j=1}^{i}q_{ijk}\varrho_{ij}\mid^{2+\delta}\right]=0.$$

The Lyapunov central limit theorem then implies that

$$\frac{\sum_{i=1}^{I}\lambda^T\{X_i-\mu_i(\tilde{\varrho})\}}{\left\{\sum_{i=1}^{I}\lambda^T\Sigma_i(\tilde{\varrho})\lambda\right\}^{1/2}}$$

converges in distribution to the standard univariate normal. Hence, the Cramér-Wold device establishes that $\Sigma(\tilde{\varrho})^{-1/2}\{T-\mu(\tilde{\varrho})\}$ converges in distribution to a K-variate vector of independent standard normals. □

The sufficient criterion given above, that $I^{-1}\sum_{i=1}^{I}\sum_{j=1}^{n_i}q_{ijk}^4$ is uniformly bounded for all I and all $k=1,\ldots,K$, satisfies the conditions of Lemma 1.2 with $\delta=2$ since Z_{ij} is binary

and $0 \leq \tilde{\varrho}_{ij} \leq 1$ for all i and j.

For many statistics, such as an m-statistic using Huber's ψ function, q_{ijk} are bounded for all i, j and k. In these cases, asymptotic normality would hold if the stratum sizes n_i were bounded, for instance. When the underlying q_{ijk} varies as a function of I as with various rank tests, the proof given above is insufficient. In such cases, a triangular array version of the central limit theorem must be applied and the sufficient conditions adapted accordingly to guarantee asymptotic normality as $I \to \infty$.

Proof of Theorem 1. Consider the random variable

$$D^2_{\Lambda_+} = h^T \Sigma(\tilde{\varrho}) h - \inf_{\lambda \in \Lambda_+} (h - \lambda)^T \Sigma(\tilde{\varrho})(h - \lambda), \tag{14}$$

where $h = \Sigma(\tilde{\varrho})^{-1}\{T - \mu(\tilde{\varrho})\}$. Assume no degeneracy between the test statistics, such that the covariance matrix $\Sigma(\tilde{\varrho})$ is positive definite for all I. For $\Sigma(\tilde{\varrho})$ positive definite, the program

$$\inf_{\lambda \in \Lambda_+} (h - \lambda)^T \Sigma(\tilde{\varrho})(h - \lambda) \tag{15}$$

is convex. Since the feasible region of (15) is Λ_+, the relative interior of the feasible region is non-empty (BV04, Section 2.1.3) and Slater's condition holds (BV04, Section 5.2.3). Consequently, there is no duality gap and the Karush-Kuhn-Tucker conditions are both necessary and sufficient for optimality (BV04, Section 5.5.3). As the objective function of (15) is a quadratic form, it is a smooth function of the arguments h, λ, and $\Sigma(\tilde{\varrho})$. Thus, the Karush-Kuhn-Tucker conditions stipulate that an optimal λ is the root of continuous functions of h and $\Sigma(\tilde{\varrho})$. Since the solutions to the Karush-Kuhn-Tucker conditions are continuous functions of h and $\Sigma(\tilde{\varrho})$, the optima of (15) are continuous functions of h and $\Sigma(\tilde{\varrho})$.

(Sha03) uses that strong duality holds for (14) to give rise to the identity

$$D^2_{\Lambda_+} = \sup_{\lambda \in \Lambda_+} \frac{\left[\lambda^T \{T - \mu(\tilde{\varrho})\}\right]^2}{\lambda^T \Sigma(\tilde{\varrho}) \lambda}. \tag{16}$$

From this, it is seen by the definition of $A^2_{\Lambda_+}$ in (7) that $D^2_{\Lambda_+} = A^2_{\Lambda_+}$.

(Sha03) shows that if Y has a multivariate normal distribution with mean vector θ and known non-singular covariance matrix V then

$$\sup_{\lambda \in \Lambda_+} \frac{\left\{\lambda^T (Y - \theta)\right\}^2}{\lambda^T V \lambda} \sim \bar{\chi}^2(V^{-1}, \Lambda_+). \tag{17}$$

Since $I^{-1}\Sigma(\tilde{\varrho}) \to M$ as $I \to \infty$, it follows that $I^{1/2}\Sigma(\tilde{\varrho})^{-1/2} \to M^{-1/2}$. By Lemma 1.2 the random vector $\Sigma(\tilde{\varrho})^{-1/2} \{T - \mu(\tilde{\varrho})\}$ converges in distribution to a K-variate vector of independent standard normals. By Slutsky's Lemma $I^{1/2}h$ converges in distribution to the mean-zero multivariate normal distribution with covariance M^{-1}. By continuity of the function taking h to the optima of (15) along with (16), the mapping

$$\Sigma(\tilde{\varrho})^{-1/2} \{T - \mu(\tilde{\varrho})\} \mapsto \sup_{\lambda \in \Lambda_+} \frac{\left\{\lambda^T (T - \mu(\tilde{\varrho}))\right\}^2}{\lambda^T \Sigma(\tilde{\varrho}) \lambda}$$

is continuous. Exploiting Slutsky's Lemma, the Continuous Mapping Theorem, and (17) yields that $A^2_{\Lambda_+}$ converges in distribution to a $\bar{\chi}^2(M^{-1}, \Lambda_+)$ random variable as desired. □

1.8.4 Theorem 2

Theorem A.4. *Suppose $\Lambda_1 \subseteq \Lambda_2$. Under mild conditions, the design sensitivity of* (5) *using $\Lambda = \Lambda_1$ is less than or equal to the design sensitivity of* (5) *using $\Lambda = \Lambda_2$.*

Proof. Define $\tilde{\Gamma}_\Lambda$ as the design sensitivity of the test using $a^*_{\Gamma,\Lambda}$ as a test statistic. To avoid triviality, suppose that the design sensitivities $\tilde{\Gamma}_{\Lambda_1}$ and $\tilde{\Gamma}_{\Lambda_2}$ both exist; see (Ros04) and

(Ros13) for mild conditions for existence of the design sensitivity. Let A_{Γ,Λ_i} be the random variable giving rise to the observation a^*_{Γ,Λ_i} in (6) for $i = 1, 2$, that is

$$A^*_{\Gamma,\Lambda_i} = \min_{\varrho\in\mathcal{P}_\Gamma} \sup_{\lambda\in\Lambda_i} \frac{\lambda^T\{T-\mu(\varrho)\}}{\{\lambda^T\Sigma(\varrho)\lambda\}^{1/2}}.$$

Since $\Lambda_1 \subseteq \Lambda_2$, for any ϱ we have

$$\sup_{\lambda\in\Lambda_1} \frac{\lambda^T\{T-\mu(\varrho)\}}{\{\lambda^T\Sigma(\varrho)\lambda\}^{1/2}} \leq \sup_{\lambda\in\Lambda_2} \frac{\lambda^T\{T-\mu(\varrho)\}}{\{\lambda^T\Sigma(\varrho)\lambda\}^{1/2}},$$

such that $A^*_{\Gamma,\Lambda_1} \leq A^*_{\Gamma,\Lambda_2}$. Consider any $\Gamma < \tilde{\Gamma}_{\Lambda_1}$. By the definition of design sensitivity, for a sensitivity analysis conducted at Γ we have that $\text{pr}(A_{\Gamma,\Lambda_1} \geq k \mid \mathcal{Z})$ tends to one as $I \to \infty$ for any scalar k. Since $A^*_{\Gamma,\Lambda_1} \leq A^*_{\Gamma,\Lambda_2}$, for any $\Gamma < \tilde{\Gamma}_{\Lambda_1}$ the power of the test based upon A^*_{Γ,Λ_2}, $\text{pr}(A_{\Gamma,\Lambda_2} \geq k \mid \mathcal{Z})$, also tends to one as $I \to \infty$ for any k. Thus, $\tilde{\Gamma}_{\Lambda_2} \geq \tilde{\Gamma}_{\Lambda_1}$ as desired.

□

1.9 Additional Simulations

1.9.1 The General Setup of the Simulation Studies

In this section we present additional simulation studies to further illustrate the results presented above. All of the simulation studies are conducted with some number I pairs, and some number K outcome variables, equicorrelated with correlation controlled by a parameter ρ. For each outcome variable, the employed test statistic is $T_k = \sum_{i=1}^{I} \text{sign}(Y_{ik}) \min(|Y_{ik}|/s_k, 2.5)$, where s_k is the median of $|Y_{ik}|$ $(i = 1, .., I)$. This amounts to a choice of a m-statistic with Huber's ψ-function, as described in (Ros07).

1.9.2 Rejecting the Global Null with $I = 1000$ Pairs

In Section 1.5.2 the $\bar{\chi}^2$-test was compared to the equal-weight test and the test of (FS16) with $I = 300$ matched pairs. To highlight the large-sample properties of the test, we include Figure 1-2. As I increases, the power curves converge pointwise to step functions, evaluating to 1 if Γ is below the design sensitivity and zero otherwise (Ros13). This indicates that the gap between the equal-weight test and the $\bar{\chi}^2$-test observed in the first row of Figure 1-1 and Figure 1-2 will shrink as I increases, and will disappear in the limit. This trend can be appraised visually by comparing the disparity observed in the first row of Figure 1-1 in Section 1.6 where $I = 300$ to the first row of Figure 1-2 where $I = 1000$. As a consequence Theorem 2 and of the pointwise convergence of the power curve to the indicator function of the event Γ less than the design sensitivity, the power curve of the $\bar{\chi}^2$-test will converge to that of the most powerful test at any fixed Γ among all coherent tests.

1.9.3 Rejecting Individual Nulls Through Closed Testing

An experimenter may want to test not only the global null hypothesis H_0 of (4) but also the K individual null hypotheses $H_1, \ldots, H_K$. To achieve this at level α, she may use a closed-testing framework (MEG76). Then, in order to test H_i at level α, she performs α-level tests all hypotheses of the form $H_i \wedge \left(\bigwedge_{k \in S_i}\right)$ with S_i the set of all possible subsets of the numbers $1, \ldots, K$ excluding i; she then rejects H_i if all of these tests rejected. Another standard method to test both the global null and each individual null would be to conduct a Bonferroni-corrected test of the global null and then use the results of the corrected individual tests to reject each H_k. Figure 1-3 examines the performance of these two methods against the test of only H_1 when $\tau_1 = 0.5$, $\tau_2 = 0.2$, $\tau_3 = 0.05$ and equicorrelation between the paired differences at at $\rho = 0.2$. The comparison to the test of only H_1 is an unfair comparison in that testing only H_1 at level-α does not control the family-wise error rate at α when examining all $k = 1, \ldots, K$. However, the test of H_1 alone at level-α achieves the highest

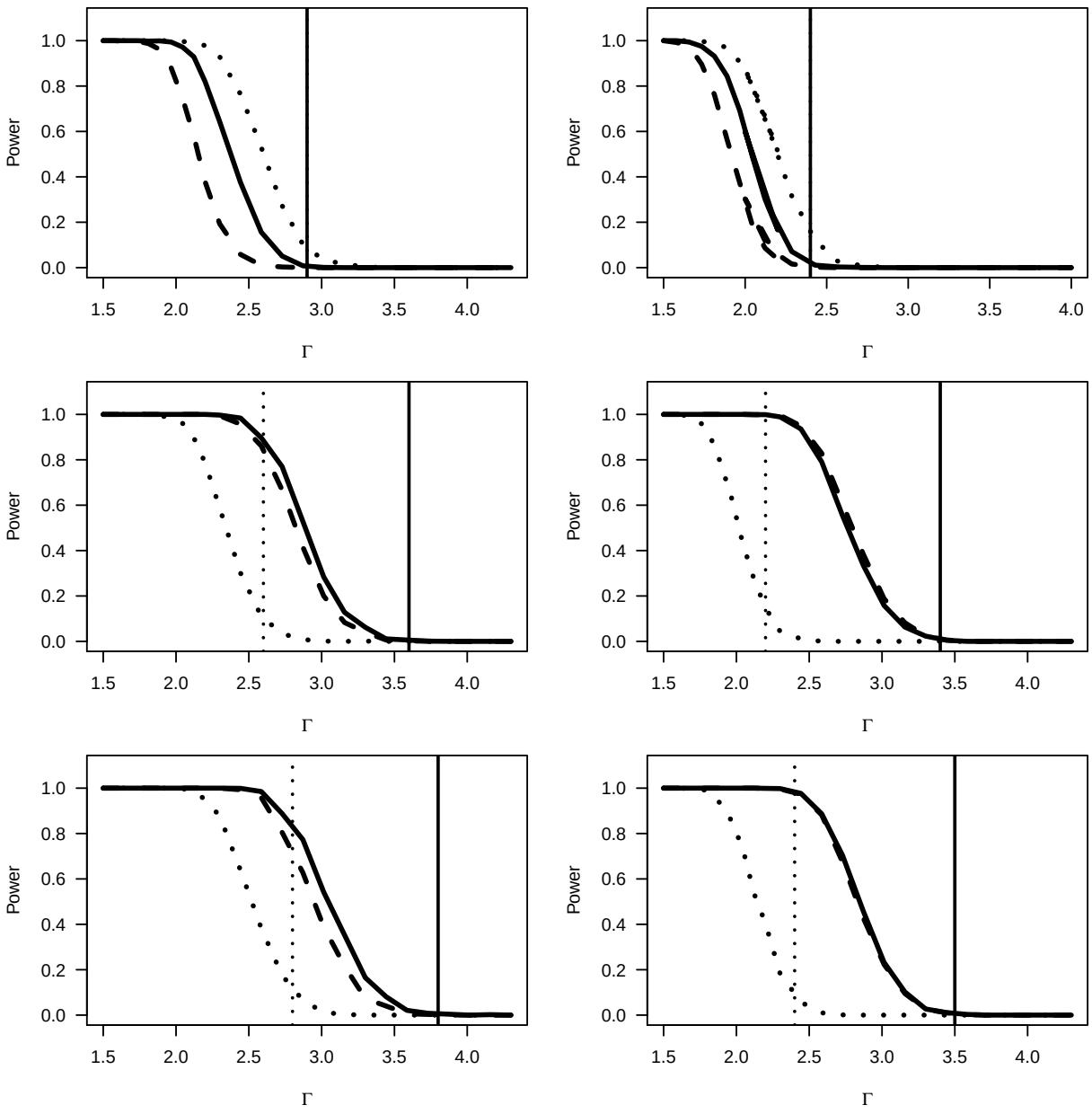

Figure 1-2: Power comparisons between the method of (FS16) (dashed), the method of this paper (solid), and the equal-weight test (dotted) as Γ increases with $I = 1000$. The first row has $\tau_1 = \tau_2 = \tau_3 = 0.25$. The second row has $\tau_1 = \tau_2 = 0.1$ and $\tau_3 = 0.5$. The third row has $\tau_1 = 0.05$, $\tau_2 = 0.2$, and $\tau_3 = 0.5$. Figures on the left have $\rho = 0$ while on the right $\rho = 0.2$. For each fixed set of parameters, power simulations were performed on 1000 simulated data sets. The design sensitivity of the equal-weight test is the dotted vertical line and the design sensitivity of the $\bar{\chi}^2$-test is the solid vertical line. In the first row, these two design sensitivities are the same and are shown by the single solid vertical line.

τ_2	ρ	(FS16)	$\Lambda = \Lambda_+$	$\Lambda = \mathbb{R}^K \setminus \{0_K\}$
-0.5	0	0	0	0.694
-0.25	0	0	0	0.454
0	0	0.026	0.018	0.46
-0.5	0.5	0	0	0.546
-0.25	0.5	0	0	0.424
0	0.5	0.018	0.016	0.496

Table 1.3: Type I error rates for the method of (FS16), the $\bar{\chi}^2$-test of this paper, and the test taking $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ using $\alpha = 0.05$. All tests performed with $I = 20$ matched pairs, $\tau_1 = -0.5$, and $\Gamma = 1$. For each set of parameters the power was estimated based upon 500 simulations.

power possible for any testing procedure that tests H_k as it does not employ any corrections to control the family-wise error rate. Thus, comparison to the test of H_1 alone at level-α serves as a comparison to an idealized benchmark, the absolute limit of statistical power that one may achieve when testing H_1 using a particular test statistic.

At all values of I examined, the closed test outperforms the Bonferroni-corrected test. Furthermore, as I increases, the closed test approaches the same power as the individual test without correction. Thus, for sufficiently large studies, empirical results suggest that a closed testing framework allows the experimenter to test both the global null and the individual null at level-α with minimal loss of power from multiple comparisons relative to testing only the individual null.

1.9.4 Type I Error Control in Small Samples Using the Asymptotic Reference Distribution

In this simulation, we assess the Type I error rate with $I = 20$ matched pairs at $\Gamma = 1$. In each simulation, the global null of non-positive treatment effects is true. Table 1.3 details simulated Type I error rates for the method of (FS16), the $\bar{\chi}^2$-test of this paper, and the unconstrained test taking $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ for $K = 2$ outcome variables with a range of

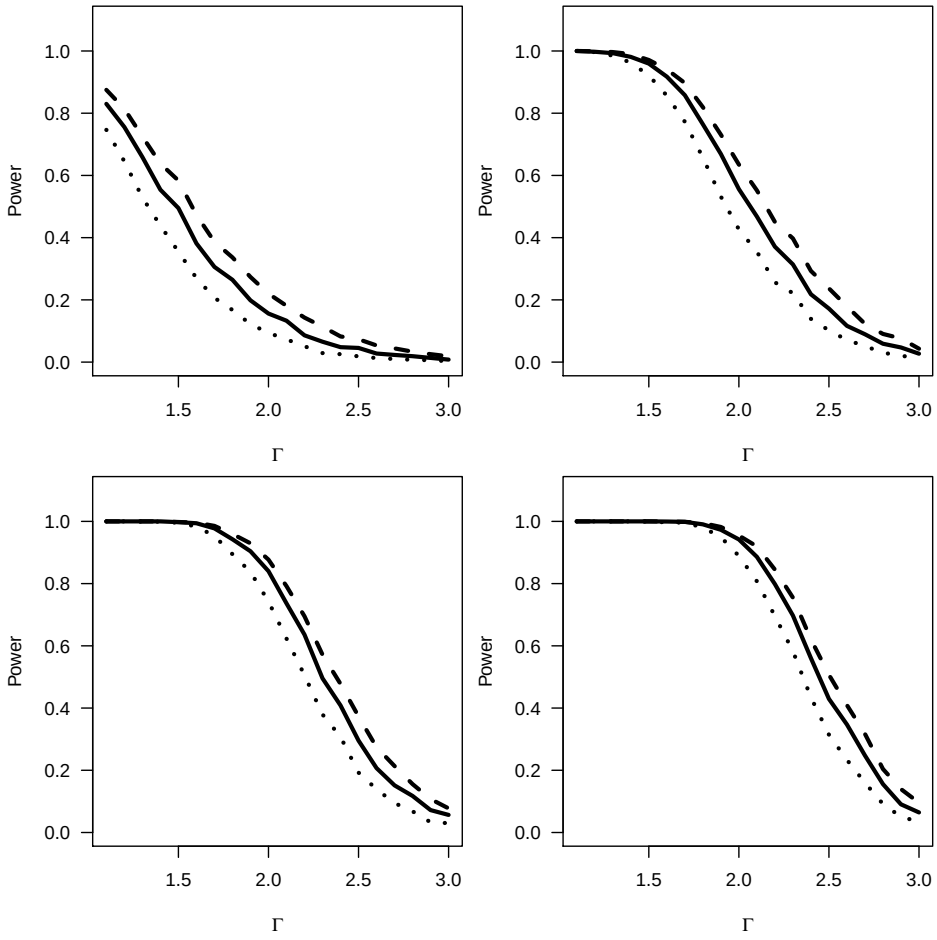

Figure 1-3: Power comparisons between embedding the $\bar{\chi}^2$-test into a closed testing framework (solid), performing a Bonferroni-corrected test (dotted), and performing an uncorrected test (dashed) as I increases. All simulations performed with $\tau_1 = 0.5$, $\tau_2 = 0.2$, $\tau_3 = 0.05$ with equicorrelation at $\rho = 0.2$ testing H_1. All data is with normal noise and tested with Huber's ψ-function as the underlying statistic. Additional parameters listed clockwise from the top-left: $I = 50$, $I = 150$, $I = 250$, and $I = 350$. For each sample size, power simulations were performed on 2000 simulated data sets.

different parameter values.

Both the method of (FS16) and the $\bar{\chi}^2$-test control the Type I error rate at α even when in the finite sample regime while using the asymptotic reference distribution. As alluded to in Section 1.4.1 the test taking $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ fails to control the Type I error rate at α since allowing λ to have an unconstrained sign in each coordinate removes the ability to discriminate positive treatment effects from negative treatment effects. This further motivates the restriction to the set of coherent combinations Λ_+.

1.9.5 Non-Normal and Larger K Simulations

We include several additional simulations, using a larger number of outcomes and experimenting with heavy-tailed noise in the data generating distribution. In order to demonstrate the method's properties on studies with more outcomes, we conducted tests with $K = 4$. Choosing $K = 4$ still allows for interpretable regimes of treatment effect relative magnitudes while expanding from the trivariate case. We conducted tests under normality as in Section 1.5 as well as under non-normal conditions.

To generate the normal 4-variate data we followed the same procedure as outlined in Section 1.5, but using four τ's and four ε's. To conduct the non-normal tests, we elected to experiment with heavy-tailed noise. This was implemented via substituting t_5-distributed noise $(\varepsilon_1, \ldots, \varepsilon_4)$ in place of normal noise in the procedure of Section 5.2. This process mirrored the t-distributed simulation construction of (Ros16). We conducted tests for $\tau = (.1, .1, .1, .5), (.1, .1, .5, .5)$, and $(.1, .25, .25, .5)$. These treatment effect relative magnitudes were selected to highlight the strength of the $\bar{\chi}^2$-test when:

- One treatment effect is much larger than the others but none are of negligible magnitude (this is the case of $\tau = (.1, .1, .1, .5)$).
- There are several highly impacted outcomes, but there remain several outcomes for which treatment effect is small (this is the case of $\tau = (.1, .1, .5, .5)$). This regime does

not exist in the trivariate case.

- The treatment effects are spread across multiple magnitude scales; selecting all to be equally weighted is apt to perform poorly, but selecting only the largest is unlikely to perform as well as optimizing for weighting in accordance with their magnitudes (this is the case of $\tau = (.1, .25, .25, .5)$).

For all of the test considered, $I = 300$, $n_i = 2$ for all i, and $\rho = 0.2$. Figure 1-4 presents the results of these simulations.

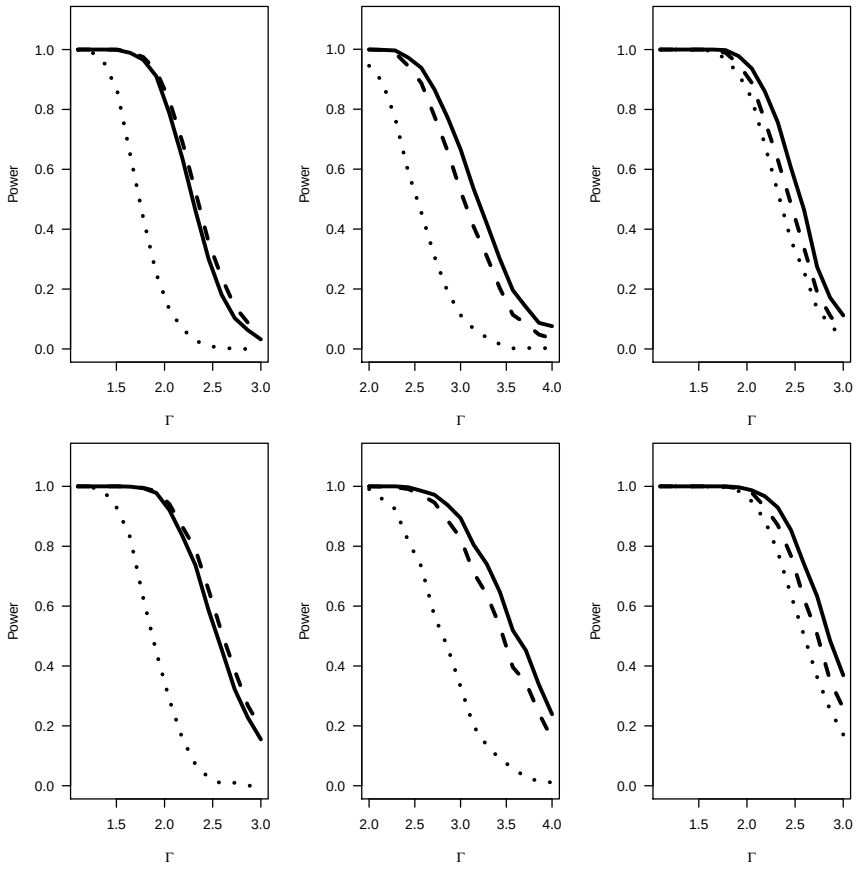

Figure 1-4: Power comparisons between the method of (FS16) (dashed), the $\bar{\chi}^2$-test of this paper (solid), and the equal-weight test (dotted) as Γ increases with $I = 300$. The first column has $\tau = (.1, .1, .1, .5)$; the second column has $\tau = (.1, .1, .5, .5)$; and the third column has $\tau = (.1, .25, .25, .5)$. The top row is generated under the Gaussian data-generating process and the bottom row is generated under the t_5 data-generating process.

Despite the heavy-tails, the relative performance of the $\bar{\chi}^2$-method stays the same as under the Gaussian data-generating process. Moreover, the performance of the $\bar{\chi}^2$-test relative to the equal-weight test and the basis-vector test accords well with the intuition developed in Section 1.5.

- In the first column, the equal-weight test is apt to under-perform due to the strong disparity in treatment effects across the four outcomes. Since there is one "stand-out" effect the lower critical value of the basis-vector test accounts for the slight increase performance edge over the $\bar{\chi}^2$-test.

- In the second column, the equal-weight test fares poorly for the same reasons as before. Since there are several outcomes that are strongly impacted, the $\bar{\chi}^2$-test outperforms the basis-vector test which weights only one outcome.

- In the third column, the spread of treatment effect magnitudes across different regimes again accounts for the strong performance of the $\bar{\chi}^2$-statistic over the other two, as it flexibly weights each outcome in accordance with the degree of treatment effect.

1.10 Algorithmic Details For Conducting The Sensitivity Analysis

The optimization problem in (6) is solved via a projected subgradient descent algorithm. (Sho85) contains a detailed introduction to subgradient methods. The algorithm begins with some initial feasible $\varrho_{(0)}$, solves for an optimal λ under the fixed $\varrho_{(0)}$, computes a subgradient of the objective at the optimal λ, and uses the subgradient to project onto the feasible region thereby locating a $\varrho_{(1)}$. The procedure iterates until convergence criteria are satisfied.

Formally, given a feasible $\varrho_{(n)}$ we compute

$$\lambda^*_{\varrho_{(n)}} = \sup_{\lambda \in \Lambda_+} \frac{\left[\max\{0, \lambda^T(T - \mu(\varrho_{(n)}))\}\right]^2}{\lambda^T \Sigma(\varrho_{(n)})\lambda}. \tag{18}$$

To compute (18), results from (Sha03) are leveraged to allow efficient computation of

$$\sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\varrho_{(n)})\}}{\{\lambda^T \Sigma(\varrho_{(n)}\lambda\}^{1/2}},$$

by solving a single quadratic program. In the event that

$$\sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\varrho_{(n)})\}}{\{\lambda^T \Sigma(\varrho_{(n)})\lambda\}^{1/2}} > 0$$

$\lambda^*_{\varrho_{(n)}}$ is set to the optimizing choice of λ. However, when

$$\sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\varrho_{(n)})\}}{\{\lambda^T \Sigma(\varrho_{(n)})\lambda\}^{1/2}} \leq 0$$

there exists a feasible ϱ such that the test fails to reject the sharp null, and thus no further iterations of the subgradient method are needed.

By (HUL13), if $f(x) = \sup_{j \in J} f_j(x)$ where each $f_j(x)$ is a convex function, $f(x) = f_{j^*}(x)$, and $g \in \partial f_{j^*}(x)$, then $g \in \partial f(x)$. In less technical terms, to compute a subgradient of a function which is the point-wise supremum of a many convex functions, one first finds a function $f_{j^*}(\cdot)$ which achieves the maximum value at the desired point x and then one computes a subgradient of this function. As such, at the optimal value λ^* one computes that the subgradient of the objective function with respect to the variables $\varrho_i = (\varrho_{i1}, \ldots, \varrho_{in_i})$ is

$$g = \frac{h_1(\varrho)\partial_{\varrho_i} h_2(\varrho) - h_2(\varrho)\partial_{\varrho_i} h_1(\varrho)}{h_2(\varrho)^2},$$

where

$$
\begin{aligned}
h_1(\varrho) &= \left(\lambda^{*T}\left(T-\mu(\varrho)\right)\right)^2, \\
h_2(\varrho) &= \lambda^{*T}\Sigma(\varrho)\lambda^*, \\
\partial_{\varrho_i} h_1(\varrho) &= -2(Q_i^T\lambda^*)\lambda^{*T}(T-\mu(\varrho)), \\
\partial_{\varrho_i} h_2(\varrho) &= (Q_i^T\lambda^*)\circ(Q_i^T\lambda^*) - 2(Q_i^T\lambda^*)(Q_i^T\lambda^*)^T\varrho_i,
\end{aligned}
$$

where Q_i is the K-by-n_i matrix where the (k,j)th entry is q_{ijk} and $\circ$ denotes the coordinate-wise product operation.

Armed with the solution to the inner maximization and the form of the subgradient g, we can now detail the projected subgradient descent method.

1. Initialize a feasible $\rho_{(0)}$, pick $t_0 > 0$ and $n = 1$

2. Repeat until convergence:

   (a) Find $\lambda_{\rho_{(n-1)}}$ by solving (18),

   (b) Compute the subgradient g from (1.10) using $\lambda_{\rho_{(n-1)}}$,

   (c) Define $\varrho_{(n)}$ to be the projection of $\rho_{(n-1)} - t_{n-1}g$ onto the feasible region,

   (d) Update the parameters: $t_n = t_0/\sqrt{n}$ and $n = n+1$.

Since the objective function is convex and the feasible set is also convex, any local optimum is a global optimum as well. In practical execution on both synthetic and real data sets convergence has been observed after few iterations.

1.11 The Chi-Bar-Squared Distribution

1.11.1 Finding a Better Critical Value

While the subgradient method solves (6) and Theorem 1 gives that the asymptotic distribution of $A^2_{\Lambda_+}$ is $\bar{\chi}^2$, the weights of the limiting distribution are still unknown. Comparing the value of (6) against the $1-\alpha$ quantile arising from the bound

$$\text{pr}\{\bar{\chi}^2(V,\Lambda_+) \geq c\} \leq 0.5\{\mathbb{P}\left(\chi^2_{K-1} \geq c\right) + \mathbb{P}\left(\chi^2_K \geq c\right)\} \tag{19}$$

would control the Type I error. While improving over a critical value based on a χ^2_K distribution, the bounds through (19) are still unduly conservative. We now describe an algorithm which exploits the particular structure of the sensitivity analysis problem to dramatically improve the critical value.

By directly computing upper and lower bounds on the correlation between T_k and T_ℓ for each $k, \ell = 1, \ldots, K$ one can compute coordinate-wise upper and lower bounds on on the overall correlation matrix $\text{diag}\{\Sigma(\varrho)\}^{-1/2}\Sigma(\varrho)\text{diag}\{\Sigma(\varrho)\}^{-1/2}$, where $\text{diag}\{\Sigma(\varrho)\}$ contains the diagonal elements of $\Sigma(\varrho)$ on its diagonals but has zeroes on its off-diagonals. Since the weights of the $\bar{\chi}^2$ distribution depend on $\Sigma(\varrho)$ only through its correlation matrix (SS05), one can directly optimize over bounds on the marginal correlations to find the most conservative $1-\alpha$ critical value associated to a correlation matrix within the bounds. This optimization can be performed via either numerical approximation of gradients or by directly computing gradients of the p-value function with respect to the correlations. Such gradients are accessible due to Plackett's identity (Pla54) and can be calculated with assistance of functions in the `mvtnorm` package within `R` for evaluating orthant probabilities and the density of the multivariate normal. Optimizing over the space of correlation matrices yields significant improvement over the critical value drawn from previous bound. Figure 1-5 highlights the differences between using the $1-\alpha$ quantile from a χ^2 distribution, the $1-\alpha$ quantile from

the naive bound based upon (19), and using the optimal $1-\alpha$ quantile with $K = 3$ outcome variables. By using the most conservative $1-\alpha$ quantile within the upper and lower bounds on the correlation matrix the Type I error rate is asymptotically controlled at α.

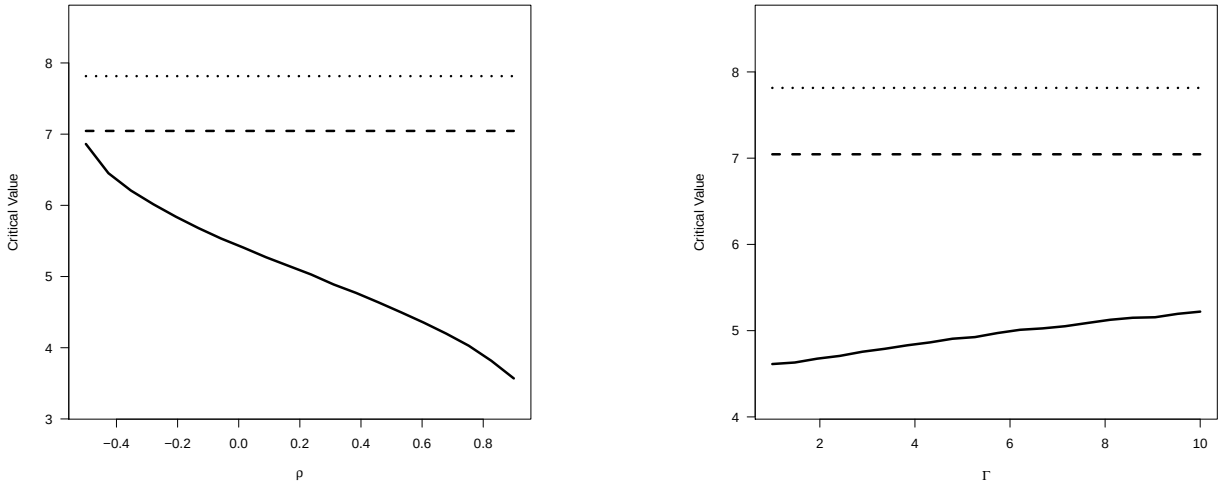

Figure 1-5: $1-\alpha$ quantiles for $\alpha = 0.05$ generated for the trivariate scenario $I = 300$, $\tau_1 = \tau_2 = \tau_3 = 0$. On the left, Γ is fixed at 1 while ρ varies over $[-0.5, 0.9]$. On the right, ρ is fixed at 0.5 while Γ ranges from 1 to 10. In both figures the χ^2_3 $1-\alpha$ quantile is the dotted line, that of the naive bound derived from (19) is the dashed line, and the $1-\alpha$ quantile coming from optimizing over feasible correlation matrices is the solid line.

There is a true, but generally unknown, underlying correlation structure between the test statistics that depends upon the true vector of conditional probabilities $\tilde{\varrho}$. Thus, the true $1-\alpha$ quantile from the $\bar{\chi}^2$ distribution with weights based on the true correlation would not change with the value of Γ employed in the sensitivity analysis. As the true unmeasured confounders are unknown, we instead find a conservative critical value based upon the feasible values for ϱ at a given Γ. As Γ grows so too does the feasible region for the probabilities $\mathcal{P}_\Gamma$; consequently the conservative critical value increases with Γ as well. This explains the trend in the right-hand panel of Figure 1-5.

1.11.2 The Worst-case Correlation with Bivariate Outcomes

In the case for $K = 2$, an elementary proof establishes a closed form of the optimizing correlation matrix subject to box constraints.

Theorem A.5. *Suppose that $K = 2$ and $[\ell, u] \subseteq (-1, 1)$. Over all matrices M in the set*

$$S = \left\{ \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \, : \, \rho \in [\ell, u] \right\}$$

the matrix $\begin{bmatrix} 1 & \ell \\ \ell & 1 \end{bmatrix}$ *achieves the most conservative (largest possible) $1-\alpha$ quantile of $\bar{\chi}^2(M^{-1}, \Lambda_+)$.*

Proof. Say that $X \sim \bar{\chi}^2(M^{-1}, \Lambda_+)$ for $M^{-1} \in S$. From (SS02) the probability

$$\begin{aligned} \mathbb{P}\left(X \leq c\right) &= w_0(2, M^{-1})\mathbb{P}\left(\chi_0^2 \leq c\right) + \frac{1}{2}\mathbb{P}\left(\chi_1^2 \leq c\right) + w_2(2, M^{-1})\mathbb{P}\left(\chi_2^2 \leq c\right) \\ &= w_2(2, M)\mathbb{P}\left(\chi_0^2 \leq c\right) + \frac{1}{2}\mathbb{P}\left(\chi_1^2 \leq c\right) + w_0(2, M)\mathbb{P}\left(\chi_2^2 \leq c\right) \end{aligned}$$

where each χ_i^2 is an independent random variable with χ_i^2 distribution. The value $w_{2-i}(2, M)$ is the probability that the projection, under the norm induced by the quadratic form $x^T M x$, of a standard bivariate normal random vector onto the non-negative orthant has exactly $2-i$ positive components. By an argument presented in (SS02), this interpretation of $w_{2-i}(2, M)$ is equivalent to defining $w_{2-i}(2, M)$ as the probability that a standard bivariate normal random variable Z falls into $R_i = \left\{x \in \mathbb{R}^2 \mid \sum_{k=1}^2 1(b_k > 0) = i\right\}$ where $b = M^{1/2}z$. Since $w_2(2, M) + w_0(2, M) = 1$ and $\mathbb{P}\left(\chi_0^2 \leq c\right) \geq \mathbb{P}\left(\chi_2^2 \leq c\right)$ for all scalars c it suffices to maximize $w_2(2, M)$.

Taking

$$M = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

gives that maximizing $w_2(2, M)$ is equivalent to maximizing the area R_2 in Figure 1-6 The

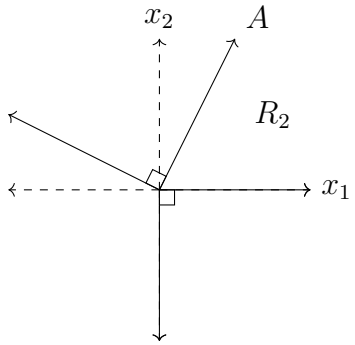

Figure 1-6: Pictorial representation of the region R_2. The upper right boundary of R_2 is the line given by A.

slope of the line A is $\rho - \rho^{-1}$ when $\rho \neq 0$ and A is vertical when $\rho = 0$. Maximizing R_2 corresponds to taking ρ as small as possible within $[\ell, u]$. Thus the matrix $\rho = \ell$ achieves the most conservative $1 - \alpha$ critical value of $\bar{\chi}^2(M^{-1}, \Lambda_+)$.

□

1.11.3 A Bivariate Illustration of the Chi-bar-squared Distribution

Consider a mean-zero bivariate normal with covariance V and consider the distribution of $\bar{\chi}^2(V, \Lambda_+)$. By the law of total probability,

$$\mathbb{P}\left(\bar{\chi}^2(V, \Lambda_+) \leq c\right) = \sum_{i=0}^{2} \text{pr}\{\bar{\chi}^2(V, \Lambda_+) \leq c \mid X \in R_i\}\text{pr}(X \in R_i),$$

where R_0, R_1, and R_2 are disjoint coverings of $\mathbb{R}^2$. Let $b = V^{-1/2}x$, and set

$$R_i = \left\{ x \in \mathbb{R}^2 \mid \sum_{k=1}^{2} 1(b_k > 0) = i \right\};$$

this is shown in Figure 1-7.

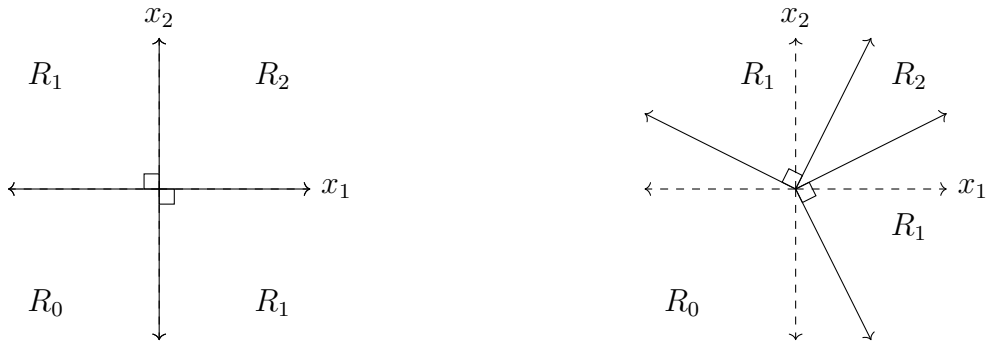

Figure 1-7: The regions corresponding to different distributional forms of the likelihood ratio statistic. In the left image $V = I_{2\times 2}$; the right image illustrates the general case, in this case the correlation is -0.8.

Within each R_i, $\bar{\chi}^2(V, \Lambda_+) \sim \chi^2_i$, where χ^2_0 is a point mass at zero. The weights of the $\bar{\chi}^2_K(V, \Lambda_+)$ are determined by the probability of falling into each partition, and are seen to depend on the covariance V.

Expansive literature exists on the $\bar{\chi}^2$ distribution. The paper (Kud63) introduces the topic in the context of order constrained one-sided tests; Chapter 3 of (SS05) contains detailed examples and derivations as well as a collection of many contemporary results; and (Sha85) discusses the weights $w_i(k, V, C)$ extensively.

1.12 Matching Details For Smoking And Polycyclic Aromatic Hydrocarbons

Individuals were classified as cigarette smokers or as non-cigarette-smokers in accordance with the criteria used in (FS16). This divided the population of 1638 total individuals into 432 cigarette smokers and 1206 non-cigarette-smokers. The population of non-smokers did include those who may have smoked in the past but had stopped smoking by the time of the survey, as well as individuals who had never smoked cigarettes. The individuals were placed into matched groups using a full matching procedure (Ros10, Section 8.5); thus each group

contained a single treated unit and multiple control units or a single control unit and multiple treated units. Pre-treatment covariates were selected based upon recent medical research. To form the fully-matched sets, propensity score caliper with a rank-based Mahalanobis distance for within-caliper distance was used. The caliper was set at 0.08 and logistic regression was performed to estimate propensity scores (Ros10, Section 8). See (FS16, Appendix A) for further implementation details.

Bibliography

[BGH+02] Carl-Elis Boström, Per Gerde, Annika Hanberg, Bengt Jernström, Christer Johansson, Titus Kyrklund, Agneta Rannug, Margareta Törnqvist, Katarina Victorin, and Roger Westerholm. Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environmental Health Perspectives*, 110:451–488, 2002.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[CC62] A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Res. Logist. Quart.*, 9:181–186, 1962.

[CDM17] Devin Caughey, Allan Dafoe, and Luke Miratrix. Beyond the Sharp Null: Randomization Inference, Bounded Null Hypotheses, and Confidence Intervals for Maximum Effects. *arXiv e-prints*, page arXiv:1709.07339, Sep 2017.

[Coc65] William G Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

[CR16] EunYi Chung and Joseph P. Romano. Multivariate and multiple permutation tests. *J. Econometrics*, 193(1):76–91, 2016.

[Fog18] Colin B Fogarty. On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:1035–1056, 2018.

[FS16] Colin B. Fogarty and Dylan S. Small. Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.*, 111(516):1820–1830, 2016.

[GKR00] Joseph L Gastwirth, Abba M Krieger, and Paul R Rosenbaum. Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):545–555, 2000.

[Han04] Ben B Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

[HUL13] Jean-Baptiste Hiriart-Urruty and Claude Lemarechal. *Convex analysis and minimization algorithms I: Fundamentals.*, volume 305. Springer Science and Business Media, 2013.

[Kud63] Akio Kudô. A multivariate analogue of the one-sided test. *Biometrika*, 50:403–418, 1963.

[MEG76] Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[Per69] Michael D Perlman. One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, 40(2):549–567, 1969.

[Pla54] R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41:351–360, 1954.

[Rao73] C. Radhakrishna Rao. *Linear statistical inference and its applications*. New York: John Wiley & Sons, second edition, 1973.

[Ros95] Paul R Rosenbaum. Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association*, 90(432):1424–1431, 1995.

[Ros97] Paul R Rosenbaum. Signed rank statistics for coherent predictions. *Biometrics*, 53(2):556–566, 1997.

[Ros02] Paul R Rosenbaum. *Observational studies*. Springer, New York, 2002.

[Ros04] Paul R Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

[Ros07] Paul R Rosenbaum. Sensitivity analysis for M-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464, 2007.

[Ros10] Paul R. Rosenbaum. *Design of observational studies*. Springer Series in Statistics. Springer, New York, 2010.

[Ros13] Paul R. Rosenbaum. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1):118–127, 2013.

[Ros15] Paul R Rosenbaum. How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Application*, 2:21–48, 2015.

[Ros16] Paul R Rosenbaum. Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. *The Annals of Applied Statistics*, 10(3):1447–1471, 2016.

[Ros17] Paul R. Rosenbaum. The general structure of evidence factors in observational studies. *Statist. Sci.*, 32(4):514–530, 2017.

[Ros18] Paul R. Rosenbaum. Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. *Annals of Applied Statistics*, to appear, 2018.

[Rub86] Donald B. Rubin. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[RWD88] Tim Robertson, FT Wright, and RL Dykstra. *Order restricted statistical inference*. John Wiley & Sons, New York, 1988.

[SCC02] WR Shadish, Thomas D Cook, and DT Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth, Belmont, 2002.

[Sch53] Henry Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110, 1953.

[Sha85] Alexander Shapiro. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1):133–144, 1985.

[Sha03] Alexander Shapiro. Scheffe's method for constructing simultaneous confidence intervals subject to cone constraints. *Statist. Probab. Lett.*, 64(4):403–406, 2003.

[Sho85] N. Z. Shor. *Minimization methods for nondifferentiable functions*, volume 3 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński.

[SS02] Pranab K. Sen and Mervyn J. Silvapulle. An appraisal of some aspects of statistical inference under inequality constraints. *J. Statist. Plann. Inference*, 107(1-2):3–43, 2002.

[SS05] Mervyn J. Silvapulle and Pranab K. Sen. *Constrained statistical inference*. Hoboken: John Wiley & Sons, 2005.

[TA00] Scott L. Tomar and Samira Asma. Smoking-attributable periodontitis in the United States: Findings from NHANES III. *Journal of Periodontology*, 71(5):743–751, 2000.

[WD18] Jason Wu and Peng Ding. Randomization Tests for Weak Null Hypotheses. *arXiv e-prints*, page arXiv:1809.07419, Sep 2018.

[Zub12] José R Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.

Chapter 2

No-harm Calibration for Generalized Oaxaca-Blinder Estimators

Abstract

In randomized experiments, adjusting for observed features when estimating treatment effects has been proposed as a way to improve asymptotic efficiency. However, only linear regression has been proven to form an estimate of the average treatment effect that is asymptotically no less efficient than the treated-minus-control difference in means regardless of the true data generating process. Randomized treatment assignment provides this "do-no-harm" property, with neither truth of a linear model nor a generative model for the outcomes being required. We present a general calibration method which confers the same no-harm property onto estimators leveraging a broad class of nonlinear models. This recovers the usual regression-adjusted estimator when ordinary least squares is used, and further provides non-inferior treatment effect estimators using methods such as logistic and Poisson regression. The resulting estimators are non-inferior to both the difference in means estimator and to treatment effect estimators that have not undergone calibration. We show that our estimator is asymptotically equivalent to an inverse probability weighted estimator using a logit link with predicted potential outcomes as covariates. In a simulation study, we demonstrate that common nonlinear estimators without our calibration procedure may perform markedly worse than both the calibrated estimator and the unadjusted difference in means.

2.1 Introduction

In completely randomized experiments, (Lin13) demonstrated that linear regression employing treatment-by-covariate interactions can be used to estimate the sample average treatment effect while adjusting for baseline features. The orthogonalities arising in the geometry of linear regression in concert with the act of randomization yield a "do-no-harm" property for the resulting estimator: assuming neither the existence of a true linear model nor a generative model for the outcome variables, the regression-adjusted estimator's asymptotic variance is never larger than that of the usual difference in means estimator. While the resulting estimator is asymptotically no less efficient than the difference in means regardless of the data generating process, linear regression seems ill-suited for modeling phenomena such as binary or count data. In such contexts, leveraging a nonlinear model such as logistic or Poisson regression may seem more natural.

Lin's (2013) regression-adjusted estimator can be viewed as an imputation estimator:

the practitioner uses linear regressions of outcomes on covariates to impute counterfactual outcomes, and then takes the difference in means between the imputed populations as their estimate of the treatment effect. Building upon the work of (Oax73) and (Bli73) among others, (GB21) present a general theory for leveraging "simple" nonlinear models to impute missing potential outcomes, providing conditions for consistency and asymptotic normality for the resulting treatment effect estimators, coined *generalized Oaxaca-Blinder estimators*. That said, (GB21) were not able to establish a non-inferiority property for these nonlinear estimators relative to the difference in means estimator, raising the concern that imputation with more general prediction functions could degrade inference relative to no adjustment whatsoever.

A related class of treatment effect estimators are *model standardization estimators*, which take the difference in the averages of the predicted values for the potential outcomes under treatment and control as an effect estimate. These are equivalent to Oaxaca-Blinder estimators when the average of the fitted values for those receiving treatment and control equal the average of the observed outcomes for those individuals; this holds automatically for generalized linear models. When using nonlinear adjustment, standardized estimators have been shown to have non-inferior asymptotic efficiency relative to the difference in means estimator under a superpopulation model when assuming correct specification of the conditional mean function; see for instance (RvdL10) or (NW21, Theorem 7.1). Unfortunately, this class of estimators does not generally provide non-inferior treatment effect estimates under misspecification. Negi and Wooldridge remark that "we do not have theoretical results to show when the nonlinear [regression adjustment] methods unambiguously improve asymptotic efficiency in case of misspecification" (NW21, p. 526). (LD20) generalized linear model standardization results to high-dimensional data; they retain the asymptotic non-inferiority of (Lin13) but their proofs rest upon the linearity of the prediction functions. This leaves a major gap between linear and nonlinear regression adjustment in randomized experiments.

We provide a calibration procedure that confers non-inferiority after nonlinear regression

adjustment under both the finite-population and superpopulation framework for causal inference. To the best of our knowledge this is the first procedure for conferring non-inferiority to generalized Oaxaca-Blinder estimators and model standardization estimators while remaining agnostic to the truth of an underlying nonlinear model. Our procedure simply feeds the predicted values for the potential outcomes under both treatment and control from a nonlinear model as covariates from which to form the linear regression-adjusted estimator of (Lin13), and provides the same non-inferiority guarantees under conditions on the prediction functions outlined in this work. We show through simulation that without this calibration step generalized Oaxaca-Blinder estimators can perform markedly worse than both the calibrated estimator and the unadjusted difference in means. This leads us to strongly recommend the use of our procedure in providing the natural extension of adjustment in randomized experiments from linear to nonlinear models. Not only are calibrated estimators non-inferior to both the difference in means estimator and the uncalibrated estimator, but also without calibration generalized Oaxaca-Blinder estimators can perform worse than the difference in means even when using simple, commonly deployed nonlinear models. We further discuss how calibration may be used in concert with adjustment strategies leveraging flexible nonlinear methods without corrupting desirable properties such as semiparametric efficiency.

2.2 Notation And Review

2.2.1 Notation for Completely Randomized Designs

We begin under the finite-population approach to causal inference where randomized treatment allocation alone justifies our results without assuming a generative model for outcomes or features; as will be discussed, our findings also hold under common superpopulation formulations for inference on both the population average treatment effect and the conditional

average treatment effect. An experimental population is comprised of N units. For the ith unit there are two scalar potential outcomes: what would have been observed under control, $y_i(0)$; and what would have been observed under treatment, $y_i(1)$. Randomness only enters the experiment through the allocation of treatment over the population of N individuals, n_0 of whom receive the control and n_1 of whom receive the treatment. Let Z_i denote the treatment indicator of the ith unit: $Z_i = 1$ if the ith unit receives treatment otherwise $Z_i = 0$. In a completely randomized experiment, the vector $Z = (Z_1, \ldots, Z_N)^{\mathrm{T}}$ is distributed uniformly over $\Omega_{CRE} = \left\{ z \in \{0,1\}^N \ : \ \sum_{i=1}^N z_i = n_1 \right\}$. Asymptotics are taken with respect to a sequence of finite experimental populations. For each N the characteristics of the finite population may change, as may the ratio of treated to control units; this dependence is generally suppressed in the notation that follows. We assume that $n_1/N \to p$ satisfying $0 < p < 1$ as $N \to \infty$.

An experimenter draws a treatment allocation Z uniformly from Ω_{CRE}. Under the stable-unit treatment value assumption (Rub80), she then observes $y_1(Z_1), \ldots, y_N(Z_N)$, the potential outcomes corresponding to the observed treatment assignment. The sample average treatment effect for the N units is $\overline{\tau}_{\text{SATE}} = N^{-1} \sum_{i=1}^N \{y_i(1) - y_i(0)\}$, and is unknown because $y_i(0)$ and $y_i(1)$ cannot be jointly observed for any unit i. The conventional estimator for $\overline{\tau}_{\text{SATE}}$ is $\hat{\tau}_{unadj} = n_1^{-1} \sum_{i=1}^N Z_i y_i(Z_i) - n_0^{-1} \sum_{i=1}^N (1 - Z_i) y_i(Z_i)$, the treated-minus-control difference in means (Ney23). Under mild regularity conditions on the sequence of finite populations, $N^{1/2} \left(\hat{\tau}_{unadj} - \overline{\tau}_{\text{SATE}} \right)$ obeys a central limit theorem (LD17).

2.2.2 The Generalized Oaxaca-Blinder Estimator

Oftentimes baseline covariates $x_i \in \mathbb{R}^k$ are collected for each unit in the study. While $\hat{\tau}_{unadj}$ is unbiased without adjustment its variance may be inflated due to post-randomization imbalances on covariates predictive of the outcome. Imputation estimators use covariate information to impute the counterfactual $y_i(1 - Z_i)$ with the objective of reducing estimator

variance. Suppose one trains predictors for the outcomes under control and treatment, $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$ respectively. Following (GB21), define imputed outcomes $\hat{y}_i(z)$ for $z = 0, 1$ and the resulting generalized Oaxaca-Blinder estimator as

$$\hat{\tau}_{gOB} = N^{-1} \sum_{i=1}^{N} \{\hat{y}_i(1) - \hat{y}_i(0)\}; \quad \hat{y}_i(z) = \begin{cases} y_i(Z_i) & \text{if } Z_i = z \\ \hat{\mu}_z(x_i) & \text{if } Z_i \neq z \end{cases}. \tag{1}$$

Both the difference-in-means estimator and Lin's (2013) regression-adjusted estimator arise from particular choices for $\hat{\mu}_1$ and $\hat{\mu}_0$. (GB21) present a set of sufficient conditions which facilitate analysis of the limiting distribution for $N^{1/2}(\hat{\tau}_{gOB} - \bar{\tau})$. We generalize their conditions slightly. Two important assumptions, "stability" and "vanishing error processes", allow for asymptotic reformulations of the estimators in terms of certain residuals. A third, "prediction unbiasedness," ensures robustness of Oaxaca-Blinder estimators to model misspecification.

Assumption 1 (Stability)**.** *For $z = 0, 1$, there exists a deterministic sequence of functions $\{\dot{\mu}_z^{(N)}\}_{N \in \mathbb{N}}$ such that $\left|\left|\hat{\mu}_z - \dot{\mu}_z^{(N)}\right|\right|_N := \left\{N^{-1} \sum_{i=1}^{N} ||\dot{\mu}_z^{(N)}(x_i) - \hat{\mu}_z(x_i)||^2\right\}^{1/2} = o_p(1)$. For notational simplicity, we generally drop the superscripted index and write $\dot{\mu}_z$.*

Assumption 2 (Vanishing Error Process)**.** *For a function $f : \mathbb{R}^k \to \mathbb{R}$ define*[1]

$$\mathcal{G}_{N,z}(f) = N^{-1/2} \sum_{i=1}^{N} \left(\frac{\mathbb{1}_{\{Z_i = z\}} f(x_i)}{n_z/N} - f(x_i) \right).$$

Assume that, for $z \in \{0, 1\}$, the error stochastic process $|\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)|$ vanishes in probability; formally

$$|\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)| = o_P(1).$$

Assumption 3 (Prediction Unbiasedness)**.** $\sum_{i:Z_i=z} \hat{\mu}_z(x_i) = \sum_{i:Z_i=z} y_i(Z_i)$ *for* $z = 0, 1$.

[1]We use the notation $\mathcal{G}_{N,z}$ to follow that of (GB21), but the stochastic process $\{\mathcal{G}_{N,z}(\cdot)\}$ indexed by functions f has been implicitly examined elsewhere in the literature, e.g., (Rot20).

If $\hat{\mu}_z$ is the solution to some empirical risk minimization procedure, then a natural candidate for $\dot{\mu}_z$ is the population-level risk minimizer and Assumption 1 reflects the standard goal that the empirical risk minimizers approximate the population risk minimizers as the sample size grows. The functions $\dot{\mu}_0$ and $\dot{\mu}_1$ need not reflect any true relationship between outcomes and covariates. Assumption 2 is quite general; in the supplementary material we provide concrete sufficient conditions based upon an entropy bound of (vdVW11) or cross-fitting. Assumption 3 holds for many choices of nonlinear models, including generalized linear models. Under Assumption 3 the estimator may be written as a model standardization estimator, with $\hat{\tau}_{gOB} = N^{-1} \sum_{i=1}^{N} \{\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\}$.

Assumptions 1, 2, and 3 are sufficient to establish consistency and asymptotically linear representations for generalized Oaxaca-Blinder estimators (GB21, Theorems 2-3). Further regularity conditions are required to imply asymptotic normality (GB21, Corollary 1). Based upon the assumptions of (Lin13) and (Fre08b) we assume the following about the potential outcomes and the functions $\dot{\mu}_z$:

Assumption 4 (Limiting Means and Variances)**.** *The mean vector and covariance matrix of* $(y_i(0), y_i(1), \dot{\mu}_0(x_i), \dot{\mu}_0(x_i))^{\mathrm{T}}$ *have limiting values. For instance, for* $z = 0, 1$ *there exists a limiting value* $\bar{y}(z)_\infty$ *such that* $\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} y_i(z) = \bar{y}(z)_\infty$.

Assumption 5 (Bounded Fourth Moments)**.** *There exists some* $C < \infty$ *for which, for all* $z = 0, 1$ *and all* N, $N^{-1} \sum_{i=1}^{N} \{y_i(z)\}^4 < C$ *and* $N^{-1} \sum_{i=1}^{N} \{\dot{\mu}_z(x_i)\}^4 < C$.

2.3 Linear Calibration

Assumptions 1 - 5 are not sufficient for $\hat{\tau}_{gOB}$ to be non-inferior to the unadjusted difference in means estimator; see Section 2.6 for an illustration with Poisson regression. We now describe a simple transformation of $\hat{\mu}_z$ that provides a "do-no-harm" property after nonlinear adjustment. For each unit i, create the pseudo-feature vector $\tilde{x}_i = (\hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$ contain-

ing the predicted potential outcomes under control and treatment. Then, for $z = 0, 1$, define $\hat{\mu}_{OLS,z}(x_i)$ as the prediction equation from a least squares regression of $y_i(Z_i)$ on $\tilde{x}_i$ along with an intercept for those units i such that $Z_i = z$, yielding for $z = 0, 1$

$$\begin{aligned} \hat{\mu}_{OLS,z}(x_i) &= \hat{\alpha}_z + \hat{\beta}_{z,0}\hat{\mu}_0(x_i) + \hat{\beta}_{z,1}\hat{\mu}_1(x_i); \\ (\hat{\alpha}_z, \hat{\beta}_{z,0}, \hat{\beta}_{z,1})^{\mathrm{T}} &\in \underset{(\alpha_z,\beta_{z,0},\beta_{z,1})^{\mathrm{T}}}{\arg\min} \sum_{i:Z_i=z} \{y_i(z) - \alpha_z - \beta_{z,0}\hat{\mu}_0(x_i) - \beta_{z,1}\hat{\mu}_1(x_i)\}^2. \end{aligned} \tag{2}$$

Finally, form the treatment effect estimator (1) using $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$. Equivalently, simply calculate Lin's (2013) regression-adjusted estimator with the features $\tilde{x}_i = (\hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$.

We call the resulting estimator the linearly-calibrated Oaxaca-Blinder estimator, denoted by $\hat{\tau}_{cal}$; see the supplementary material for pseudocode. The approach is similar to that of (GB21, Equation 8), but importantly differs in that under their approach $\hat{\mu}_{1-z}(x_i)$ is not included as a predictor variable in the regressions for individuals with $Z_i = z$. By including both prediction functions in $\tilde{x}_i$, $\hat{\tau}_{cal}$ attains non-inferiority.

Theorem 1. *Suppose that Assumptions 1, 2, 4, and 5 hold. Then,* $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}})$ *converges in distribution to a mean-zero Gaussian random variable. Furthermore the asymptotic variance of* $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}})$ *is no larger than that of* $N^{1/2}(\hat{\tau}_{unadj} - \overline{\tau}_{\text{SATE}})$.

Theorem 1 does not require knowledge of the true relationship between outcomes and covariates, and the models $\hat{\mu}_0$ and $\hat{\mu}_1$ can be arbitrarily misspecified. Assumption 3 is not required for $\hat{\mu}_0$ and $\hat{\mu}_1$ because prediction unbiasedness always holds after applying our procedure due to the inclusion of the intercept terms. The non-inferiority statement in Theorem 1 is proven in a manner similar to Corollary 1.1 of (Lin13); however, care must be taken to account for randomness in the incorporation of the derived covariates $\tilde{x}_i = (\hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$. The proof further demonstrates that the sufficient conditions for asymptotic Gaussianity of $N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}})$ provided by Assumptions 1 - 5 are also sufficient for the asymptotic Gaussianity of $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}})$. That is, under these assumptions non-

inferiority can be achieved "for free" through our calibration step. A similar argument yields the following comparisons between $\hat{\tau}_{cal}$ and both the non-calibrated estimator $\hat{\tau}_{gOB}$ under Assumption 3 and the singly-calibrated estimator suggested in (GB21, Equation 8), denoted $\hat{\tau}_{GBcal}$.

Theorem 2. *Under the assumptions of Theorem 1 and for given estimators $\hat{\mu}_0$ and $\hat{\mu}_1$, the linearly-calibrated estimator $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ has an asymptotic variance that is no larger than that of $N^{1/2}\left(\hat{\tau}_{GBcal} - \overline{\tau}_{\text{SATE}}\right)$. Further enforcing Assumption 3, $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ has an asymptotic variance that is no larger than that of $N^{1/2}\left(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}}\right)$.*

2.4 Further Insight Into Linear Calibration

As the calibration step may appear unusual, it is first worthwhile to consider what occurs when $\hat{\mu}_1$ and $\hat{\mu}_0$ are fit by separate ordinary least squares regressions with intercepts in the treated and control groups. In this case, $\hat{\tau}_{cal}$, $\hat{\tau}_{gOB}$, and $\hat{\tau}_{GBcal}$ are identical to Lin's (2013) estimator as the resulting prediction equations remain linear in the covariates themselves; see the supplementary materials for a formal proof. Including $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$ as predictors when calculating $\hat{\tau}_{cal}$ also yields an insightful asymptotic equivalence with an inverse probability weighted (IPW) estimator. Suppose that despite having run a randomized experiment with known assignment probabilities, one fits a logistic regression model for the probability that $Z_i = 1$ with $\tilde{x}_i = \left(\hat{\mu}_0\left(x_i\right), \hat{\mu}_1\left(x_i\right)\right)^{\mathrm{T}}$ as covariates along with an intercept. Call the resulting predicted probabilities $\hat{\pi}(\tilde{x}_i)$. Consider the IPW estimator $\hat{\tau}_{ipw} = N^{-1}\sum_{i=1}^{N} Z_i y_i(Z_i)/\hat{\pi}(\tilde{x}_i) - N^{-1}\sum_{i=1}^{N}(1 - Z_i)y_i(Z_i)/\{1 - \hat{\pi}(\tilde{x}_i)\}$.

Theorem 3. *Under the assumptions of Theorem 1 $N^{1/2}(\hat{\tau}_{cal} - \hat{\tau}_{ipw}) = o_p(1)$. That is, the two estimators are asymptotically equivalent.*

In light of the proof of Theorem 1, the result follows immediately from Corollary 1 of (SLL14) and is omitted; see also (HIR03) for related results on IPW estimators. In random-

ized experiments, inverse probability weighted estimators adjust for chance imbalances on variables contained within the propensity score model by reweighting individuals using their predicted probabilities of treatment. Imbalances on covariates are problematic only in so far as the covariates are predictive of the potential outcomes. By including both $\hat{\mu}_0$ and $\hat{\mu}_1$ within the propensity score model, $\hat{\tau}_{ipw}$ adjusts for chance imbalances on the predicted values for the potential outcomes under treatment and control. Both (RLSR12) and (CR15) use predicted potential outcomes into propensity score models to establish non-inferior treatment effect estimators under a superpopulation. By the asymptotic equivalence provided by Theorem 3, $\hat{\tau}_{cal}$ can also be viewed in this light. Importantly, this equivalence does not generally hold for the entire class of Oaxaca-Blinder estimators $\hat{\tau}_{gOB}$. The estimator $\hat{\tau}_{GBcal}$ suggested in (GB21, Equation 8) is equivalent to a peculiar IPW estimator where the treated and control outcomes are weighted with estimated probabilities stemming from *different* logit models, with the treated (resp. control) outcomes weighted by estimated probabilities where only fitted values under treatment (resp. control) are used as covariates.

2.5 Calibration And Non-Inferiority Under Superpopulation Models

Our results have viewed the potential outcomes and covariates as fixed, with the only randomness coming from random assignment. As the recruitment process for inclusion in randomized experiments often amounts to a convenience sample, we view this as a natural framework for inference. Alternative frameworks view $(y_i(1), y_i(0), x_i)$ as independent and identically distributed draws from some distribution P. Still others view x_i as fixed but $(y_i(1), y_i(0))$ as independently distributed from some conditional distribution $P_{y(1),y(0)|x}$. The corresponding estimands are the population average treatment effect (PATE) and the conditional average treatment effect (CATE), respectively (Imb04). Theorems 1 - 3 apply with $\bar{\tau}$ replaced by

$\overline{\tau}_{\text{PATE}}$ or $\overline{\tau}_{\text{CATE}}$ and with our regularity conditions reformulated depending upon the super-population framework. See the supplementary materials for details.

When the triples $(y_i(1), y_i(0), x_i)$ are viewed as independent and identically distributed draws from a distribution P, alternative estimators leveraging nonlinear adjustment exist which either guarantee non-inferiority under model misspecification, or leverage cross-fitting to attain semiparametric efficiency bounds. Methods guarding against model misspecification in parametric models include (Tan10), (RLSR12) and (CR15). These are not imputation estimators, require explicitly fitting a separate propensity score model, and in the case of (RLSR12) require solving a non-convex optimization problem. Our calibration step returns an imputation estimator, and it can be implemented using off-the-shelf statistical software, simply requiring an initial (potentially nonlinear) regression adjustment within each treatment group followed by a linear regression using the fitted values as covariates. Calibration may be deployed in concert with augmented inverse probability weighted (AIPW) and targeted maximum likelihood estimators (TMLE) within randomized experiments. For instance, one may replace $\hat{\mu}_z$ by $\hat{\mu}_{OLS,z}$ of (2) to produce an AIPW estimator which is guaranteed to be non-inferior to both the uncalibrated AIPW estimator and the unadjusted difference in means. Should the original AIPW estimator achieve the semiparametric efficiency bound, so too will the estimator after calibration. Should the uncalibrated AIPW estimator be based upon a misspecified model, calibration can provide an improvement over both the uncalibrated AIPW estimator and the unadjusted estimator. The relationship between calibration and semiparametric efficiency is explored in detail in the supplementary materials. Furthermore, in the supplementary materials we discuss the use of sample splitting to acheive finite sample unbiasedness of calibrated estimators under superpopulation models, thereby extending the results of (WGB18) and (Rot20). We also include simulations to emphasize these points.

2.6 Illustrating The Improvements From Linear Calibration

To illustrate both the benefits of linear calibration with the prediction equations for both potential outcomes and the potential peril of proceeding without our calibration step, we present a simulation study using Poisson regression. The sth of S data sets contains N individuals upon whom an experimenter performs a completely randomized experiment with $n_1 = \lceil pN \rceil$ treated units. In our simulations $p = 0.8$. Each unit has a scalar covariate x_i, generated as independent and identically distributed draws from a Uniform random variable on $[-5, 5]$. We then generate the potential outcomes under treatment and control for each individual independently as $y_i(1) \sim Poisson\{\exp(x_i)\}$ and $y_i(0) \sim Poisson\{72 - 0.45\exp(x_i)\}$, where $Poisson(\lambda)$ is a Poisson distribution with rate λ. The Poisson regression model is thus correctly specified for the potential outcomes under treatment, but incorrectly specified for those under control.

For each data set, we draw B treatment assignment allocations. An experimenter observes $y_i(Z_i)$ and continuous covariates x_i for each unit. Using the observed responses after each randomized treatment allocation, we estimate the prediction functions $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$ via separate Poisson regressions of $y_i(Z_i)$ on x_i in the subgroups where $Z_i = 0$ and $Z_i = 1$ respectively. With the outcomes and response functions in tow, we form the difference-in-means estimator $\hat{\tau}_{unadj}$, generalized Oaxaca-Blinder estimator $\hat{\tau}_{gOB}$, the singly-calibrated estimator of (GB21, Equation 8) $\hat{\tau}_{GBcal}$, and our linearly-calibrated estimator $\hat{\tau}_{cal}$.

Table 2.1 compares the averages (over $s = 1, \ldots, S$) of the ratios of the variances for the adjusted estimators to the unadjusted estimator when setting $S = 1000$, $B = 1000$, and varying N. Even at $N = 10,000$, both $\hat{\tau}_{gOB}$ and $\hat{\tau}_{GBcal}$ have markedly larger variances than the unadjusted difference in means estimator. Contrast this with our proposed estimator $\hat{\tau}_{cal}$, which in this simulation study provides a substantial reduction in variance relative to

	$\hat{\text{var}}(\hat{\tau}_{gOB})/\hat{\text{var}}(\hat{\tau}_{unadj})$	$\hat{\text{var}}(\hat{\tau}_{GBcal})/\hat{\text{var}}(\hat{\tau}_{unadj})$	$\hat{\text{var}}(\hat{\tau}_{cal})/\hat{\text{var}}(\hat{\tau}_{unadj})$
$N = 200$	1.732	1.717	0.703
$N = 500$	1.692	1.685	0.665
$N = 1000$	1.675	1.670	0.659
$N = 10000$	1.660	1.657	0.654

Table 2.1: Ratios of Monte Carlo variances for $\hat{\tau}_{cal}$, $\hat{\tau}_{GBcal}$, and $\hat{\tau}_{gOB}$ to the difference in means estimator $\hat{\tau}_{unadj}$ for various experiment sizes N. Each variance is based upon $B = 1000$ simulated treatment allocations for a given set of potential outcomes and covariates. Results are averaged over $S = 1000$ simulated data sets.

the difference in means, $\hat{\tau}_{gOB}$ and $\hat{\tau}_{GBcal}$. This highlights the importance of including the prediction functions for both potential outcomes in the calibration step (2) after nonlinear adjustment. In the supplementary material we include a simulation using logistic regression which shows the same qualitative phenomena. We also include analysis of real-world data using Poisson regression adjustment investigating the effectiveness of a chemotherapeutic agent.

2.7 Discussion

Linear calibration maps $\hat{\tau}_{gOB}$ to $\hat{\tau}_{cal}$ in such a way that asymptotic non-inferiority is guaranteed. The linearly-calibrated estimator can provide hypothesis tests and construct confidence intervals using the standard errors proposed in (GB21, Section 3.3) and a Gaussian approximation; detailed discussion of variance estimators for $\hat{\tau}_{cal}$ is provided in the supplementary materials. Extending (ZD21), one can further provide inference that is exact under the sharp null of no effect for any individual while remaining asymptotically valid for the sample average treatment effect. Calibration using only $\hat{\mu}_0$ and $\hat{\mu}_1$ fits into a more general class of algorithms wherein calibration is performed on the vectors $\ddot{x}_i = (f(x_i), \hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$ instead of just $\tilde{x}_i$. Taking $f : \mathbb{R}^k \to \mathbb{R}^\ell$ adds an additional ℓ features to $\tilde{x}_i$. Setting $f(x_i) = x_i$ yields an estimator which is asymptotically no less efficient than $\hat{\tau}_{unadj}$, $\hat{\tau}_{gOB}$, $\hat{\tau}_{cal}$, and Lin's (2013)

estimator simultaneously and is akin to the estimator of (CR15); see the supplementary materials for details.

Supplementary Material

Below we include additional information which contains proofs, a discussion of variance estimation, pseudocode, superpopulation results, and a data example. Code written in `R` is available at:
`https://github.com/PeterLCohen/OaxacaBlinderCalibration` to implement the method and to reproduce the simulations.

2.8 Extensions And Further Results

2.8.1 Feature Engineering

While $\hat{\tau}_{cal}$ is no less asymptotically efficient than $\hat{\tau}_{gOB}$ and $\hat{\tau}_{unadj}$, there are no guarantees that $\hat{\tau}_{cal}$ offers an improvement over Lin's (2013) regression-adjusted estimator, $\hat{\tau}_{lin}$, in terms of asymptotic variance. Intuitively, if the prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are extremely poor predictors of the potential outcomes, then linear regression on the raw features may substantially outperform even what calibration is able to correct. Must an experimenter decide *a priori* whether to use $\hat{\tau}_{cal}$ or $\hat{\tau}_{lin}$ when she seeks an estimator of $\overline{\tau}_{\text{SATE}}$ that is certain to be no less efficient than $\hat{\tau}_{unadj}$? In fact, there is an estimator which is non-inferior to both $\hat{\tau}_{cal}$ and $\hat{\tau}_{lin}$.

In the calibration algorithm pseudo-feature vectors were defined as

$$\tilde{x}_i = \left(\hat{\mu}_0(x_i), \hat{\mu}_1(x_i)\right)^{\mathrm{T}}.$$

Instead, take $\ddot{x}_i = \left(f(x_i), \hat{\mu}_0(x_i), \hat{\mu}_1(x_i)\right)^{\mathrm{T}}$ where $f : \mathbb{R}^k \to \mathbb{R}^\ell$ for some fixed ℓ. The function f adds an additional ℓ features to $\tilde{x}_i$; perhaps incorporating tailored feature engineering guided by domain knowledge. Define $\hat{\tau}_{cal2}$ to be the calibrated Oaxaca-Blinder estimator

based upon $\ddot{x}_i$ instead of $\tilde{x}_i$; i.e., incorporate the additional features $f(x_i)$ in the second-stage linear regression of calibration.

Theorem A.4. *Assume the regularity conditions of Theorem 1. So long as the random vectors* $(f(x_i), \hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$ *are sufficiently regular a central limit theorem applies to*

$$N^{1/2}\left(\hat{\tau}_{cal2} - \overline{\tau}_{\text{SATE}}\right)$$

and it is non-inferior to both $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ *and* $N^{1/2}\left(\hat{\tau}_{lin} - \overline{\tau}_{\text{SATE}}\right)$ *using the engineered features* $f(x_i)$.

Corollary A.1. *Take* $f(x_i) = x_i$*; the resulting estimator* $\hat{\tau}_{cal2}$ *is asymptotically no less efficient than* $\hat{\tau}_{cal}$ *and the standard regression adjusted estimator of treatment effect,* $\hat{\tau}_{lin}$.

We include proof of Theorem A.4 in Section 2.12 below, along with a more precise statement of regularity conditions.

2.8.2 Idempotence

Let $\mathcal{O}$ denote the set of Oaxaca-Blinder estimators for $\hat{\mu}_0$ and $\hat{\mu}_1$ satisfying Assumptions 1 and 7. Calibration can be thought of as a mapping $\varphi : \mathcal{O} \to \mathcal{O}$ wherein $\hat{\tau}_{gOB} \overset{\varphi}{\mapsto} \hat{\tau}_{cal}$; repeated iterations of φ induce dynamics on $\mathcal{O}$. A natural question is: do repeated applications of φ provide additional improvements in terms of asymptotic efficiency?

Theorem A.5. φ *is an idempotent map on* $\mathcal{O}$*, i.e.,* $\varphi \circ \varphi = \varphi$.

Theorem A.5 demonstrates a desirable feature: since φ is idempotent, all of the improvement possible through calibration is achieved in one application of φ.

Proof. This proof uses the same line of reasoning as that of Proposition A.2, which we include below in Section 2.12. See Remark 2 for the details of this argument. □

2.9 Regularity Conditions

In this section we lay out regularity conditions on the potential outcomes and covariates that are sufficient for analyzing the asymptotic distributions of the estimators encountered in the preceding sections. Finite population inference asymptotics are taken with respect to a sequence of probability spaces which vary with the size of the finite population, N. For each N, there are deterministic potential outcomes and covariates for each of the N individuals; randomness enters the model only through the treatment allocation, Z. Our results center around completely randomized experiments; i.e., $Z \sim \text{Unif}(\Omega_{CRE})$. A basic requirement is that the completely randomized experiments are not asymptotically degenerate.

Assumption A.6 (Non-degeneracy)**.** *The proportion of treated units, n_1/N, limits to $p \in (0,1)$ as $N \to \infty$.*

Our main set of assumptions concern the prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$. These assumptions play into the asymptotic analysis of generalized Oaxaca-Blinder estimators. The following two regularity conditions match those of (GB21).

Assumption 1 (Stability)**.** *There exists a deterministic sequence of functions $\{\dot{\mu}_1^{(N)}\}_{N\in\mathbb{N}}$ such that*

$$\left(\frac{1}{N}\sum_{i=1}^{N} ||\dot{\mu}_1^{(N)}(x_i) - \hat{\mu}_1(x_i)||^2\right)^{1/2} = o_P(1).$$

The left-hand-side of the formula above satisfies the properties of a norm on functions; this norm is denoted $||\cdot||_N$ and so an equivalent statement is that $\left|\left|\hat{\mu}_1 - \dot{\mu}_1^{(N)}\right|\right|_N = o_P(1)$.

We assume that an analogous sequence, $\{\dot{\mu}_0^{(N)}\}_{N\in\mathbb{N}}$, exists for $\hat{\mu}_0$. For notational simplicity we drop the superscripted index and write $\dot{\mu}_0$ and $\dot{\mu}_1$, but the dependence upon N remains an important background detail.

An instructive example of this assumption in practice is when $\hat{\mu}_0$ and $\hat{\mu}_1$ are derived via linear regression as in (Lin13). The deterministic sequence $\{\dot{\mu}_1^{(N)}\}_{N\in\mathbb{N}}$ can be taken as the population-level ordinary least squares linear predictor of treated outcome given covariates.

Assumption 2 (Vanishing Error Process)**.** *For a function $f : \mathbb{R}^k \to \mathbb{R}$ define*[2]

$$\mathcal{G}_{N,z}(f) = N^{-1/2} \sum_{i=1}^{N} \left(\frac{\mathbb{1}_{\{Z_i=z\}} f(x_i)}{n_z/N} - f(x_i) \right).$$

Assume that, for $z \in \{0,1\}$, the error stochastic process $|\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)|$ vanishes in probability; formally $|\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)| = o_P(1)$.

We include the prediction unbiasedness assumption of (GB21) in order to rigorously discuss asymptotic results for $\hat{\tau}_{gOB}$.

Assumption 3 (Prediction Unbiasedness)**.** *For $z = 0, 1$,*

$$\sum_{i:Z_i=z} \hat{\mu}_z(x_i) = \sum_{i:Z_i=z} y_i(Z_i).$$

The next assumptions constrain the sets of potential outcomes and imputed responses so that central limit behaviour holds for the estimators considered above. Asymptotic theory for finite population inference builds upon combinatorial analogues for classical probability theory results; see (Mad48), (ER59), (H́60), and (Hoe51) among many others. (LD17) developed results for finite population central limit theorems in causal inference; the regularity conditions of their work have become standard in the literature and form the basis for our regularity conditions. The conditions below naturally generalize those of (Lin13) and (Fre08a, Fre08b) to the nonlinear imputation context of (GB21).

[2]We use the notation $\mathcal{G}_{N,z}$ to follow that of (GB21), but the stochastic process $\{\mathcal{G}_{N,z}(\cdot)\}$ indexed by functions f has been implicitly examined elsewhere in the literature, e.g., (Rot20).

Assumption 4 (Limiting Means and Variances)**.** *The mean vector and covariance matrix of* $(y_i(0), y_i(1), \dot{\mu}_0(x_i), \dot{\mu}_0(x_i))^{\mathrm{T}}$ *have limiting values. For instance, for* $z = 0, 1$ *there exists a limiting value* $\bar{y}(z)_\infty$ *such that* $\lim_{N\to\infty} N^{-1} \sum_{i=1}^N y_i(z) = \bar{y}(z)_\infty$.

Assumption 5 (Bounded Fourth Moments)**.** *There exists some* $C < \infty$ *for which, for all* $z = 0, 1$ *and all* N, $N^{-1} \sum_{i=1}^N \{y_i(z)\}^4 < C$ *and* $N^{-1} \sum_{i=1}^N \{\dot{\mu}_z(x_i)\}^4 < C$.

For comparing $\hat{\tau}_{cal}$ to $\hat{\tau}_{lin}$ we will, at times, need to constrain the covariates x_i directly so that Lin's regressions have appropriate asymptotic properties. For such applications we make the assumption:

Assumption 6. *The mean vector and covariance matrix of* $(y_i(0), y_i(1), x_i^{\mathrm{T}})^{\mathrm{T}}$ *have limiting values. Furthermore, the bounded fourth moment assumption applies componentwise to the raw covariates, i.e.,* $N^{-1} \sum_{i=1}^N \{x_{ij}\}^4 < C$. *where* x_{ij} *denotes the* j*th coordinate of* x_i.

We use the "bar" notation to denote the mean, i.e., $\bar{y}(z) = N^{-1} \sum_{i=1}^N y_i(z)$. Let $\Sigma_{y(0)}$ denote the finite population covariance of the control outcomes, $\Sigma_{y(1)}$ denote its treated analogue, and $\Sigma_{y(z)x}$ the finite population covariance matrix of the joint vectors $(y_i(z), x_i^{\mathrm{T}})^{\mathrm{T}}$. In light of this notation, Assumptions 4 and 6 imply that $\lim_{N\to\infty} \bar{y}(z) = \bar{y}_\infty(z)$, $\lim_{N\to\infty} \Sigma_{y(z)} = \Sigma_{y(z),\infty}$, $\lim_{N\to\infty} \Sigma_{y(z)x} = \Sigma_{y(z)x,\infty}$ for $z \in \{0, 1\}$, etc. The limiting covariance matrices are assumed to be positive definite.

In the proofs we maintain the implicit assumption that the prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are non-collinear and asymptotically almost surely non-constant when evaluated over the set $\{x_i\}_{i=1}^N$. We make the same assumption for $\dot{\mu}_0$ and $\dot{\mu}_1$. Extensions to the rank deficient case are straightforward and are discussed in Section 2.14. Rank deficiency can occur in practice: for instance, if both $\hat{\mu}_0$ and $\hat{\mu}_1$ are linear regressions and x_i is scalar, then $\dot{\mu}_0$ and $\dot{\mu}_1$ are perfectly collinear. Importantly, handling the rank-deficient case does not require additional assumptions, but would complicate the notation used in the proofs.

For comparison with the assumptions of (GB21) and (Rot20) we include the following entropy-based assumption.

Assumption 7 (Typically Simple Realizations)**.** *There exists a sequence of sets of functions $\mathscr{A}_{N,0}$, which may vary with N, such that the random function $\hat{\mu}_0$ falls into this class asymptotically almost surely. Formally, $\mathbb{P}\left(\hat{\mu}_0 \in \mathscr{A}_{N,0}\right) \to 1$. Furthermore, the sets of functions are "small" in the sense that*

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{A}_{N,0}, ||\cdot||_N, s\right)}\, ds < \infty$$

where $\mathscr{N}\left(\mathscr{A}_{N,0}, ||\cdot||_N, s\right)$ is the s-covering number of $\mathscr{A}_{N,0}$ under the metric induced by $||\cdot||_N$. An analogous statement holds for $\hat{\mu}_1$ with a sequence of sets $\mathscr{A}_{N,1}$.

The inequality in Assumption 7 is common in central limit theorems for stochastic processes: it integrates the square root of the maximal entropy without bracketing of $\mathscr{A}_{N,z}$; see (LT91, Chapter 11), (vdVW96, Chapter 2), and (vdVW11) for background on the theory and uses of covering numbers and entropy bounds. The particular upper bound on the region of integration in Assumption 7 is unimportant, as shown in Lemma A.10. As a result, we say Assumption 7 holds so long as there is any $D > 0$ for which

$$\int_0^D \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{A}_{N,z}, ||\cdot||_N, s\right)}\, ds < \infty$$

for both $z \in \{0, 1\}$.

2.10 Helpful Technical Results

Lemma A.1. *Let $\{a_i\}_{i=1}^N$ and $\{b_i\}_{i=1}^N$ be two sets of fixed scalars. Let $Z \sim$ Unif(Ω_{CRE}). The variance of*

$$\frac{1}{n_1}\sum_{i=1}^N Z_i a_i - \frac{1}{n_0}\sum_{i=1}^N (1 - Z_i) b_i \tag{3}$$

is

$$\frac{n_0 n_1}{N} \cdot \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{a_i}{n_1} + \frac{b_i}{n_0} - \overline{\frac{a}{n_1} + \frac{b}{n_0}} \right)^2$$

where $\overline{v} = N^{-1} \sum_{j=1}^{N} v_i$ *for* $v \in \mathbb{R}^N$.

Proof. We start with a simple algebraic manipulation:

$$\frac{1}{n_1} \sum_{i=1}^{N} Z_i a_i - \frac{1}{n_0} \sum_{i=1}^{N} (1 - Z_i) b_i = \sum_{i=1}^{N} Z_i \left(\frac{a_i}{n_1} + \frac{b_i}{n_0} \right) - \sum_{i=1}^{N} \frac{b_i}{n_0}.$$

Since the second term on the right is constant with respect to Z it plays no part in the variance of (3) so

$$\mathbb{V}\left(\frac{1}{n_1} \sum_{i=1}^{N} Z_i a_i - \frac{1}{n_0} \sum_{i=1}^{N} (1 - Z_i) b_i \right) = \mathbb{V}\left(\sum_{i=1}^{N} Z_i \left(\frac{a_i}{n_1} + \frac{b_i}{n_0} \right) \right).$$

The term on the right is the variance of the population total for drawing a sample of size n_1 without replacement from the set

$$\left\{ \left(\frac{a_1}{n_1} + \frac{b_1}{n_0} \right), \ldots, \left(\frac{a_N}{n_1} + \frac{b_N}{n_0} \right) \right\}.$$

This variance is well understood in the survey-sampling community; see, for instance, (MHL16). Specifically,

$$\mathbb{V}\left(\sum_{i=1}^{N} Z_i \left(\frac{a_i}{n_1} + \frac{b_i}{n_0} \right) \right) = \frac{n_0 n_1}{N} \cdot \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{a_i}{n_1} + \frac{b_i}{n_0} - \overline{\frac{a}{n_1} + \frac{b}{n_0}} \right)^2$$

□

Lemma A.2. *Consider any two scalars,* $w_0, v_0 \in \mathbb{R}$ *and any two vectors* $W, V \in \mathbb{R}^\ell$. *For a*

set of features $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^\ell$ *define the "residuals"*

$$r_i(0) = y_i(0) - \left(w_0 + W^T\chi_i\right)$$
$$r_i(1) = y_i(1) - \left(v_0 + V^T\chi_i\right).$$

i) The variance of

$$\frac{1}{n_1}\sum_{i=1}^N Z_i r_i(1) - \frac{1}{n_0}\sum_{i=1}^N (1-Z_i) r_i(0) \tag{4}$$

is minimized at

$$(w_0^*, W^*) \in \arg\min_{w_0, W} \left[\sum_{i=1}^N \left\{y_i(0) - \left(w_0 + W^T\chi_i\right)\right\}^2\right], \tag{5}$$

$$(v_0^*, V^*) \in \arg\min_{v_0, V} \left[\sum_{i=1}^N \left\{y_i(1) - \left(v_0 + V^T\chi_i\right)\right\}^2\right]. \tag{6}$$

ii) Let $\mathcal{X} = [\chi_1, \ldots, \chi_N]^T$*. The variance of* (4) *achieves a strict global minimum at* w_0^**,* v_0^**,* W^**, and* V^* *when the matrix* $\mathcal{X}^T\mathcal{X}$ *is nonsingular.*

Proof. By Lemma A.1, the variance of (4) is proportional to

$$\sum_{i=1}^N \left(\frac{y_i(1) - \left(v_0 + V^T\chi_i\right)}{n_1} + \frac{y_i(0) - \left(w_0 + W^T\chi_i\right)}{n_0} - \right.$$
$$\left. \frac{y(1) - (v_0 + V^T\chi)}{n_1} + \frac{y(0) - (w_0 + W^T\chi)}{n_0} \right)^2.$$

Rearranging terms gives that the variance of (4) is proportional to

$$\sum_{i=1}^{N}\left(\left(\frac{y_i(1)}{n_1}+\frac{y_i(0)}{n_0}\right)-\left(\left(\frac{v_0}{n_1}+\frac{w_0}{n_0}\right)+\left(\frac{V}{n_1}+\frac{W}{n_0}\right)^T\chi_i\right)-\overline{\left(\frac{y(1)}{n_1}+\frac{y(0)}{n_0}\right)-\left(\left(\frac{v_0}{n_1}+\frac{w_0}{n_0}\right)+\left(\frac{V}{n_1}+\frac{W}{n_0}\right)^T\chi\right)}\right)^2. \quad (7)$$

Since w_0^* and W^* are the intercept and slope, respectively, of the ordinary least squares regression of $y_i(0)$ on χ_i it follows that $n_0^{-1}w_0^*$ and $n_0^{-1}W^*$ are the intercept and slope, respectively, of the ordinary least squares regression of $n_0^{-1}y_i(0)$ on χ_i. Likewise, $n_1^{-1}v_0^*$ and $n_1^{-1}V^*$ are the intercept and slope, respectively, of the ordinary least squares regression of $n_1^{-1}y_i(1)$ on χ_i.

Consider now the regression of $n_1^{-1}y_i(1)+n_0^{-1}y_i(0)$ on χ_i. Writing the design matrix of this regression as $\mathcal{X}$, the ordinary least squares slope is

$$\begin{aligned}(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\left(\frac{y(1)}{n_1}+\frac{y(0)}{n_0}\right)&=(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\frac{y(1)}{n_1}+(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\frac{y(0)}{n_0}\\&=\frac{V^*}{n_1}+\frac{W^*}{n_0}.\end{aligned}$$

The second equality holds for the following reason: the first term on the right is the slope coefficient in the regression of $n_1^{-1}y_i(1)$ on χ_i and so it must match $n_1^{-1}V^*$, similar reasoning applies to the second term.[3] Likewise, it follows that the intercept of the regression of $n_1^{-1}y_i(1)+n_0^{-1}y_i(0)$ on χ_i is given by $n_1^{-1}v_0^*+n_0^{-1}w_0^*$. In total, the argument above implies that the ordinary least squares regression predictor of $n_1^{-1}y_i(1)+n_0^{-1}y_i(0)$ based upon χ_i is

$$\left(\left(\frac{v_0^*}{n_1}+\frac{w_0^*}{n_0}\right)+\left(\frac{V^*}{n_1}+\frac{W^*}{n_0}\right)^T\chi_i\right).$$

[3]In the rank-deficient case that $\mathcal{X}^T\mathcal{X}$ is not invertible, these slope terms are not uniquely defined; however, picking a canonical pseudoinverse, e.g., the Moore-Penrose pseudoinverse, obviates this concern.

Equation (7) is simply the variance of the residuals from attempting to predict $n_1^{-1}y_i(1) + n_0^{-1}y_i(0)$ via

$$\left(\left(\frac{v_0}{n_1} + \frac{w_0}{n_0}\right) + \left(\frac{V}{n_1} + \frac{W}{n_0}\right)^T \chi_i\right).$$

Since ordinary least squares linear regression minimizes the variance of the residuals, it follows that (7) is minimized at w_0^*, v_0^*, W^*, and V^*. Consequently, the variance of (4) is minimized at w_0^*, v_0^*, W^*, and V^*, as required to show Part i.

The result of Part ii follows from the uniqueness of the global minima in the full-rank regression problems (5) and (6). □

Lemma A.3. *Consider the quantities*

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\arg\min} \left[\sum_{i\,:\,Z_i=1} \left\{ y_i(Z_i) - \left(\beta_0 + \beta_1^T \begin{bmatrix}\hat{\mu}_0(x_i) \\ \hat{\mu}_1(x_i)\end{bmatrix}\right)\right\}^2\right],$$

$$\dot{\beta} = (\dot{\beta}_0, \dot{\beta}_1) = \underset{\beta_0,\beta_1}{\arg\min} \left[\sum_{i=1}^{N} \left\{ y_i(1) - \left(\beta_0 + \beta_1^T \begin{bmatrix}\dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i)\end{bmatrix}\right)\right\}^2\right].$$

Then $||\hat{\beta} - \dot{\beta}||_2 = o_P(1)$ *under Assumptions 1 - 5.*[4]

Proof. Define the design matrices

$$\hat{U}_1 = \begin{bmatrix} 1 & \hat{\mu}_0(x_{i_1}) & \hat{\mu}_1(x_{i_1}) \\ \vdots & \vdots & \vdots \\ 1 & \hat{\mu}_0(x_{i_{n_1}}) & \hat{\mu}_1(x_{i_{n_1}}) \end{bmatrix} \text{ and } \dot{U}_1 = \begin{bmatrix} 1 & \dot{\mu}_0(x_1) & \dot{\mu}_1(x_1) \\ \vdots & \vdots & \vdots \\ 1 & \dot{\mu}_0(x_N) & \dot{\mu}_1(x_N) \end{bmatrix}$$

[4]This lemma generalizes Lemma 7 of (GB21) and also incorporates both the imputed outcomes.

where $i_1, \ldots, i_{n_1}$ are the indices of the treated units. Standard ordinary least squares regression theory gives closed-form solutions for $\hat{\beta}$ and $\dot{\beta}$ in terms of $\hat{U}_1$ and $\dot{U}_1$, respectively, via

$$\hat{\beta} = \left(\hat{U}_1^{\mathrm{T}}\hat{U}_1\right)^{-1}\hat{U}_1^{\mathrm{T}}\begin{bmatrix} y_{i_1}(1) \\ \vdots \\ y_{i_{n_1}}(1) \end{bmatrix} \text{ and } \dot{\beta} = \left(\dot{U}_1^{\mathrm{T}}\dot{U}_1\right)^{-1}\dot{U}_1^{\mathrm{T}}\begin{bmatrix} y_1(1) \\ \vdots \\ y_N(1) \end{bmatrix}.$$

Conveniently rewriting these regression coefficients by multiplying by one gives

$$\hat{\beta} = \left(\frac{1}{n_1}\hat{U}_1^{\mathrm{T}}\hat{U}_1\right)^{-1}\frac{1}{n_1}\hat{U}_1^{\mathrm{T}}\begin{bmatrix} y_{i_1}(1) \\ \vdots \\ y_{i_{n_1}}(1) \end{bmatrix} \text{ and } \dot{\beta} = \left(\frac{1}{N}\dot{U}_1^{\mathrm{T}}\dot{U}_1\right)^{-1}\frac{1}{N}\dot{U}_1^{\mathrm{T}}\begin{bmatrix} y_1(1) \\ \vdots \\ y_N(1) \end{bmatrix}.$$

We first show that

$$\left\| \frac{1}{n_1}\hat{U}_1^{\mathrm{T}}\begin{bmatrix} y_{i_1}(1) \\ \vdots \\ y_{i_{n_1}}(1) \end{bmatrix} - \frac{1}{N}\dot{U}_1^{\mathrm{T}}\begin{bmatrix} y_1(1) \\ \vdots \\ y_N(1) \end{bmatrix} \right\|_2 = o_P(1). \tag{8}$$

These are vectors in $\mathbb{R}^3$. To show (8) we show that the difference in each coordinate is vanishing. The first coordinate is

$$\left| \frac{1}{n_1}\sum_{Z_i=1} y_i(1) - \frac{1}{N}\sum_{i=1}^{N} y_i(1) \right|$$

which converges in probability to zero by the weak law of large numbers for finite populations; see for instance (Lin13, Lemma A.1). The second term is

$$\left| \frac{1}{n_1}\sum_{Z_i=1} \hat{\mu}_0(x_i)y_i(1) - \frac{1}{N}\sum_{i=1}^{N} \dot{\mu}_0(x_i)y_i(1) \right|. \tag{9}$$

This term vanishes in probability by an application of the triangle inequality, the Cauchy-Schwarz inequality, and a finite population law of large numbers. The details proceed analogously to those in the proof of Lemma 7 (on page 35) of the arXiv version of (GB21)[5] except we use $\hat{\mu}_0(\cdot)$ in place of $\exp(\hat{\theta}_0^{\mathrm{T}}\cdot)$; this presents no problem because we have assumed the stability of $\hat{\mu}_0$. The third term is

$$\left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i) y_i(1) - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_1(x_i) y_i(1) \right|$$

which vanishes for exactly the same reason as the second term. In total, we have shown that (8) holds.

Next, we turn attention to showing that

$$\left|\left| \frac{1}{n_1} \hat{U}_1^{\mathrm{T}} \hat{U}_1 - \frac{1}{N} \dot{U}_1^{\mathrm{T}} \dot{U}_1 \right|\right|_F = o_P(1). \tag{10}$$

The matrix in (10) is a 3-by-3 symmetric matrix; the term $\frac{1}{n_1} \hat{U}_1^{\mathrm{T}} \hat{U}_1$ takes the form

$$\begin{bmatrix} \frac{1}{n_1} \sum_{Z_i=1} 1 & \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_0(x_i) & \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i) \\ * & \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_0(x_i)^2 & \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_0(x_i) \hat{\mu}_1(x_i) \\ * & * & \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i)^2 \end{bmatrix}$$

where stars indicate symmetry. The matrix $\frac{1}{N} \dot{U}_1^{\mathrm{T}} \dot{U}_1$ has an analogous construction where $\dot{\mu}$ replaces $\hat{\mu}$, sums $\sum_{i=1}^{N}$ replace $\sum_{Z_i=1}$, and N replaces n_1.

In the matrix of (10), the top-left corner element is $\left| \frac{1}{n_1} \sum_{Z_i=1} 1 - \frac{1}{N} \sum_{i=1}^{N} 1 \right|$ which is

[5]The arXiv version of (GB21) can be found at `https://arxiv.org/pdf/2004.11615.pdf`.

deterministically zero. The remaining two diagonal elements are

$$\left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_0(x_i)^2 - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_0(x_i)^2 \right|,$$
$$\left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i)^2 - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_1(x_i)^2 \right|.$$

As before, the convergence of each of these terms in probability to zero is guaranteed by following the argument used to prove that the denominator terms of Lemma 7 from (GB21) vanish. The only difference is that $\hat{\mu}_z(\cdot)$ stands in place of $\exp(\hat{\theta}_z^{\mathrm{T}} \cdot)$, but this presents no obstacle as we have assumed that the prediction functions $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$ are stable.

By the symmetry of the matrix in (10), it suffices to show that

$$\left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_z(x_i) - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_z(x_i) \right| = o_P(1) \quad \text{for } z \in \{0, 1\}, \tag{11}$$
$$\left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i)\hat{\mu}_0(x_i) - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_1(x_i)\dot{\mu}_0(x_i) \right| = o_P(1). \tag{12}$$

The same argument used to show that (9) vanishes in probability applies to (11); replacing each $y_i(1)$ in that argument gives the desired result. To show, (12) first add zero to the left-hand-side

$$\left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i)\hat{\mu}_0(x_i) - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_1(x_i)\dot{\mu}_0(x_i) \right| = \left| \frac{1}{n_1} \sum_{Z_i=1} \hat{\mu}_1(x_i)\hat{\mu}_0(x_i) - \right.$$
$$\left. \frac{1}{n_1} \sum_{Z_i=1} \dot{\mu}_1(x_i)\dot{\mu}_0(x_i) + \frac{1}{n_1} \sum_{Z_i=1} \dot{\mu}_1(x_i)\dot{\mu}_0(x_i) - \frac{1}{N} \sum_{i=1}^{N} \dot{\mu}_1(x_i)\dot{\mu}_0(x_i) \right|.$$

Applying the triangle inequality for the right-hand-side then bounds above by

$$\underbrace{\left|\frac{1}{n_1}\sum_{Z_i=1}\left(\hat{\mu}_1(x_i)\hat{\mu}_0(x_i)-\dot{\mu}_1(x_i)\dot{\mu}_0(x_i)\right)\right|}_{\text{Term 1}}+\underbrace{\left|\frac{1}{n_1}\sum_{Z_i=1}\dot{\mu}_1(x_i)\dot{\mu}_0(x_i)-\frac{1}{N}\sum_{i=1}^{N}\dot{\mu}_1(x_i)\dot{\mu}_0(x_i)\right|}_{\text{Term 2}}.$$

Just as in (GB21, Lemma 7) the Cauchy-Schwarz inequality and the stability of $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$ imply that Term 1 vanishes in probability, and the finite population law of large numbers implies that Term 2 vanishes. In total, we have shown (10).

The matrix inverse map $M \mapsto M^{-1}$ is continuous over the set of positive definite matrices with the metric defined by the Frobenious norm. Assuming that the continuous mapping theorem applies (vdV98, Chapter 18) (i.e., assume that the matrices $\frac{1}{n_1}\hat{U}_1^{\mathrm{T}}\hat{U}_1$ and $\frac{1}{N}\dot{U}_1^{\mathrm{T}}\dot{U}_1$ have their minimum eigenvalue bounded away from zero) then (10) implies that

$$\left|\left|\left(\frac{1}{n_1}\hat{U}_1^{\mathrm{T}}\hat{U}_1\right)^{-1}-\left(\frac{1}{N}\dot{U}_1^{\mathrm{T}}\dot{U}_1\right)^{-1}\right|\right|_F = o_P(1). \tag{13}$$

Finally, combining (8) with (13) yields that $||\hat{\beta}-\dot{\beta}||_2 = o_P(1)$ as desired. □

Lemma A.3 applies to the ordinary least squares linear regression coefficients computed within the treated group. The same logic applies to the ordinary least squares linear regression coefficients computed within the control group.

Remark 1. Suppose that the design matrices used in the proof of Lemma A.3 were replaced with

$$\hat{U}_1 = \begin{bmatrix} 1 & \hat{\mu}_0(x_{i_1}) & \hat{\mu}_1(x_{i_1}) & f(x_{i_1})^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \hat{\mu}_0(x_{i_{n_1}}) & \hat{\mu}_1(x_{i_{n_1}}) & f(x_{i_{n_1}})^{\mathrm{T}} \end{bmatrix} \text{ and } \dot{U}_1 = \begin{bmatrix} 1 & \dot{\mu}_0(x_1) & \dot{\mu}_1(x_1) & f(x_1)^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \dot{\mu}_0(x_N) & \dot{\mu}_1(x_N) & f(x_N)^{\mathrm{T}} \end{bmatrix}.$$

As long as the vectors $(\dot{\mu}_0(x_1), \dot{\mu}_1(x_1), f(x_1)^{\mathrm{T}})^{\mathrm{T}}$ satisfy the regularity conditions applied origi-

nally (Assumption 4 and Assumption 5) they are amenable to the law of large numbers proofs used in the componentwise analyses for the proof of Lemma A.3. In particular, when $f(\cdot)$ is the identity function, this requirement reduces to Assumption 6 jointly with Assumptions 4 and 5.

Consequently, the proof of Lemma A.3 is equally useful for showing the consistency of the ordinary least squares linear regression coefficients used in $\hat{\tau}_{cal2}$.

Lemma A.4. *Assumptions 4 and 5 imply that the quantity*

$$\left|\left|\begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N$$

is bounded uniformly in N.

Proof. By Assumption 5 there exists some finite constant which upper bounds

$$\frac{\sum_{i=1}^{N}\left(\dot{\mu}_z(x_i) - N^{-1}\sum_{i=1}^{N}\dot{\mu}_z(x_i)\right)^4}{N}$$

for both $z \in \{0,1\}$ and all N. Bounded fourth central moments imply bounded second central moments, so there exists some constant which upper bounds

$$\frac{\sum_{i=1}^{N}\left(\dot{\mu}_z(x_i) - N^{-1}\sum_{i=1}^{N}\dot{\mu}_z(x_i)\right)^2}{N}$$

for both $z \in \{0,1\}$ and all N. Decomposing the variance into the difference between the

second moment and the squared first moment yields that the quantities

$$\frac{1}{N}\sum_{i=1}^{N}\dot{\mu}_z(x_i)^2 - \left\{\frac{1}{N}\sum_{i=1}^{N}\dot{\mu}_z(x_i)\right\}^2 \tag{14}$$

are uniformly bounded above by a constant.

Assumption 4 implies that the quantities $N^{-1}\sum_{i=1}^{N}\dot{\mu}_z(x_i)$ limit to fixed values for both $z \in \{0,1\}$; combining this with (14) yields that $\frac{1}{N}\sum_{i=1}^{N}\dot{\mu}_z(x_i)^2$ is uniformly bounded above by some constant. This implies that $\frac{1}{N}\sum_{i=1}^{N}\left(\dot{\mu}_0(x_i)^2 + \dot{\mu}_1(x_i)^2\right)$ is uniformly bounded. Lastly,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(\dot{\mu}_0(x_i)^2 + \dot{\mu}_1(x_i)^2\right) &= \frac{1}{N}\sum_{i=1}^{N}\left|\left|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2^2 \\ &= \left|\left|\begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N^2\end{aligned}$$

which concludes the proof. □

Lemma A.5. *Let $\{\hat{\mu}_z^{(N)}\}_{N\in\mathbb{N}}$ be a sequence of prediction unbiased functions and assume that there exists some sequence $\{\dot{\mu}_z^{(N)}\}_{N\in\mathbb{N}}$ for which $||\hat{\mu}_z^{(N)} - \dot{\mu}_z^{(N)}||_N$ converges in probability to 0 with respect to randomness in the superpopulation model (i.e., randomness in Z, the potential outcomes and the covariates). Further assume that $N^{-1}\sum_{i=1}^{N}\left(\dot{\mu}_z^{(N)}(x_i) - y_i(z)\right)^2 = o_P(N)$. Under the preceding conditions, without loss of generality we may assume that $\mathbb{E}\left[\dot{\mu}_z^{(N)}(x_i) - y_i(z)\right] = 0$.*

Proof. Let $\{\dot{\mu}_z^{(N)}\}_{N\in\mathbb{N}}$ for which $||\hat{\mu}_z^{(N)} - \dot{\mu}_z^{(N)}||_N$ converges in probability to 0 and

$$N^{-1}\sum_{i=1}^{N}\left(\dot{\mu}_z^{(N)}(x_i) - y_i(z)\right)^2 = o_P(n).$$

By the strong law of large numbers,

$$\left| N^{-1} \sum_{i=1}^{N} \left(\dot{\mu}_z^{(N)}(x_i)) - y_i(z) \right) - \mathbb{E}\left[\dot{\mu}_z^{(N)}(x_i) - y_i(z) \right] \right|$$

converges almost surely to zero. By the argument of Lemma 3 in the appendix of (GB21), $\left| N^{-1} \sum_{i=1}^{N} \left(\dot{\mu}_z^{(N)}(x_i)) - y_i(z) \right) \right| = o_P(1)$. Furthermore, $\mathbb{E}\left[\dot{\mu}_z^{(N)}(x_i) - y_i(z) \right]$ is deterministic; so we must have that $\left| \mathbb{E}\left[\dot{\mu}_z^{(N)}(x_i) - y_i(z) \right] \right| = o(1)$. Consequently, we may center $\dot{\mu}_z^{(N)}$ by subtracting off $\mathbb{E}\left[\dot{\mu}_z^{(N)}(x_i) - y_i(z) \right]$; defining

$$\dot{\nu}_z^{(N)} = \dot{\mu}_z^{(N)} - \mathbb{E}\left[\dot{\mu}_z^{(N)}(x_i) - y_i(z) \right]$$

we retain that $||\hat{\mu}_z^{(N)} - \dot{\nu}_z^{(N)}||_N$ converges in probability to 0 and

$$N^{-1} \sum_{i=1}^{N} \left(\dot{\nu}_z^{(N)}(x_i) - y_i(z) \right)^2 = o_P(n).$$

□

Lemma A.6. *Consider a sequence of random elements* $\{(\mathcal{A}_N, \mathcal{B}_N)\}_{N \in \mathbb{N}}$ *and a function* f *for which* $f(\mathcal{A}_N, \mathcal{B}_N) \in \mathbb{R}$ *for all* $N \in \mathbb{N}$. *If* $f(a_N, \mathcal{B}_N)$ *converges in distribution to the random variable* χ *for all sequences* $\{a_N\}_{N \in \mathbb{N}}$ *with* a_N *in some measurable set* A_N *such that* $\mathbb{P}\left(\mathcal{A}_N \in A_N\right) = 1$, *then* $f(\mathcal{A}_N, \mathcal{B}_N)$ *converges in distribution to* χ.

Proof. For notation, let $\mathcal{L}(X)$ denote the law of a random variable X and let $\mathcal{L}(X \mid Y)$ denote the conditional law of X given Y. The *bounded Lipschitz metric* is a metric on the space of probability measures;

$$d(F, G) := \sup_{f \in \mathcal{F}_{BL}} \left| \int f \, dF - \int f \, dG \right|,$$

where $\mathcal{F}_{BL}$ is the class of 1-Lipschitz functions mapping into $[-1, 1]$.

The bounded Lipschitz metric metrizes weak convergence of probability measures (vdVW96, Theorem 1.12.4). Consequently, the fact that $f(a_N, \mathcal{B}_N)$ converges in distribution to the random variable χ for $\mathcal{A}_N$-almost all $\{a_N\}_{N\in\mathbb{N}}$

$$d\left(\mathcal{L}\left(f(a_N, \mathcal{B}_N)\right), \mathcal{L}(\chi)\right) \to 0$$

for $\mathcal{A}_N$-almost all $\{a_N\}_{N\in\mathbb{N}}$. Equivalently, the random variable $d\left(\mathcal{L}\left(f(\mathcal{A}_N, \mathcal{B}_N) \mid \mathcal{A}_N\right), \mathcal{L}(\chi)\right)$ converges almost surely to 0 with respect to randomness in $\mathcal{A}_N$. As almost sure convergence implies convergence in probability, $d\left(\mathcal{L}\left(f(\mathcal{A}_N, \mathcal{B}_N) \mid \mathcal{A}_N\right), \mathcal{L}(\chi)\right)$ converges in probability to 0 with respect to randomness in $\mathcal{A}_N$. Theorem 4.1 of (DDCZ13) then implies that $d\left(\mathcal{L}\left(f(\mathcal{A}_N, \mathcal{B}_N)\right), \mathcal{L}(\chi)\right)$ converges to 0. Finally, again using that the bounded Lipschitz metric metrizes weak convergence we conclude that unconditionally $f(\mathcal{A}_N, \mathcal{B}_N)$ converges in distribution to χ. □

2.11 Error Processes, Asymptotic Linearity, And Calibration

For a function $f : \mathbb{R}^k \to \mathbb{R}$ (GB21) define

$$\mathcal{G}_{N,z}(f) = N^{-1/2} \sum_{i=1}^{N} \left(\frac{\mathbb{1}_{\{Z_i = z\}} f(x_i)}{n_z/N} - f(x_i) \right).$$

The collection $\{\mathcal{G}_{N,z}\}_{f\in F}$ forms a stochastic process indexed by functions f ranging over some family of functions F. Showing that $|\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)|$ decays quickly in probability is crucial for the central limit theorems undergirding asymptotic inference for imputation-based estimators. This general principle is true across finite population, fixed covariate, and superpopulation models. The first consequence of Assumption 2 is that, under mild regularity conditions, the error process for the calibrated estimator $\hat{\mu}_{OLS,z}$ vanishes at the same rate.

Without loss of generality – for the sake of readability – we focus only on the prediction equations in the treated group ($z = 1$) and so we follow the notation of Lemma A.3.

Define

$$\dot{\mu}_{OLS,1}(\cdot) = \dot{\beta}_0 + \dot{\beta}_1^{\mathrm{T}} \begin{bmatrix} \dot{\mu}_0(\cdot) \\ \dot{\mu}_1(\cdot) \end{bmatrix}.$$

Proposition A.1. *For $z \in \{0, 1\}$ suppose that the random variables $\hat{\beta}_0$ and $\hat{\beta}_1$ converge in probability to fixed values $\dot{\beta}_0$ and $\dot{\beta}_1$, respectively. Under Assumption 2 it immediately holds that*

$$|\mathcal{G}_{N,1}(\dot{\mu}_{OLS,1}) - \mathcal{G}_{N,1}(\hat{\mu}_{OLS,1})| = o_P(1).$$

Of course, an analogous result holds for the control quantities as well.

Proof. Recall the definition

$$\hat{\mu}_{OLS,1}(\cdot) = \hat{\beta}_0 + \hat{\beta}_1^{\mathrm{T}} \begin{bmatrix} \hat{\mu}_0(\cdot) \\ \hat{\mu}_1(\cdot) \end{bmatrix}.$$

Notice that $\mathcal{G}_{N,1}(\cdot)$ is linear in its argument in the sense that

$$\mathcal{G}_{N,1}(f + g) = \mathcal{G}_{N,1}(f) + \mathcal{G}_{N,1}(g),$$
$$\mathcal{G}_{N,1}(cf) = c\mathcal{G}_{N,1}(f) \text{ for } c \in \mathbb{R},$$

so the quantity $|\mathcal{G}_{N,1}(\dot{\mu}_{OLS,1}) - \mathcal{G}_{N,1}(\hat{\mu}_{OLS,1})|$ decomposes naturally across the linear combination of terms in $\dot{\mu}_{OLS,1}$ and $\hat{\mu}_{OLS,1}$. The assumption of convergence in probability of $\hat{\beta}_0$ and $\hat{\beta}_1$ to $\dot{\beta}_0$ and $\dot{\beta}_1$, respectively, combined with Slutsky's theorem and Assumption 2 concludes the proof. □

In its most abstract sense, the typical analysis of an imputation-based estimator proceeds by writing the estimator in terms of the difference in means of "residuals"

$\{\dot{\epsilon}_i(1) = y_i(1) - \dot{\mu}_1(x_i)\}_{i=1}^N$ and $\{\dot{\epsilon}_i(0) = y_i(0) - \dot{\mu}_0(x_i)\}_{i=1}^N$, possibly some correction factor, and an error term related to $\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)$. For example, in the finite population context of (GB21)

$$\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}} = \underbrace{\frac{1}{n_1} \sum_{i\,:\,Z_i=1} \dot{\epsilon}_i(1) - \frac{1}{n_0} \sum_{i\,:\,Z_i=1} \dot{\epsilon}_i(0)}_{\text{Difference in Residual Means}} + \underbrace{N^{-1/2} \left(\mathcal{G}_{N,1}(\dot{\mu}_1) - \mathcal{G}_{N,1}(\hat{\mu}_1)\right) - N^{-1/2} \left(\mathcal{G}_{N,0}(\dot{\mu}_0) - \mathcal{G}_{N,0}(\hat{\mu}_0)\right)}_{\text{Error Term}}.$$

Likewise, in the context of (Rot20)

$$\hat{\tau}_{rothe} - \overline{\tau}_{\text{PATE}} = \underbrace{\frac{1}{n_1} \sum_{i\,:\,Z_i=1} \dot{\epsilon}_i(1) - \frac{1}{n_0} \sum_{i\,:\,Z_i=1} \dot{\epsilon}_i(0)}_{\text{Difference in Residual Means}} + \underbrace{\frac{1}{N} \sum_{i=1}^{N} \left(\dot{\mu}_1(x_i) - \dot{\mu}_0(x_i)\right)}_{\text{Prediction Unbiasedness Correction Term}} - \overline{\tau}_{\text{PATE}} + \underbrace{N^{-1/2} \left(\mathcal{G}_{N,1}(\dot{\mu}_1) - \mathcal{G}_{N,1}(\hat{\mu}_1)\right) - N^{-1/2} \left(\mathcal{G}_{N,0}(\dot{\mu}_0) - \mathcal{G}_{N,0}(\hat{\mu}_0)\right)}_{\text{Error Term}}.$$

We provide a formal definition of $\hat{\tau}_{rothe}$ and a detailed analysis of calibrating $\hat{\tau}_{rothe}$ in Section 2.19.

Under Assumption 2 these error terms are $o_P(N^{-1/2})$ and so asymptotic analyses of such estimators can safely ignore the error terms even after scaling by $N^{1/2}$; consequently central limit theorems hold for estimators such as $N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}})$ and $N^{1/2}(\hat{\tau}_{rothe} - \overline{\tau}_{\text{PATE}})$ under mild conditions.

2.12 Main Proofs

Above we remarked upon the special case of $\hat{\tau}_{cal}$ when the prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are the linear regression prediction functions of outcome based upon covariates, fitted separately in the treated and control groups. We now formalize this and provide proof. The identity $\hat{\tau}_{gOB} = \hat{\tau}_{lin}$ in this case is well known; see for instance (DL18) or (GB21) and the identity $\hat{\tau}_{GBcal} = \hat{\tau}_{lin}$ follows from similar logic to that presented below.

Proposition A.2. *Take the original prediction functions to be the linear predictors $\hat{\mu}_1$ and $\hat{\mu}_0$ derived from separate regressions in the treated and control groups, respectively; the calibrated estimator $\hat{\tau}_{cal}$ yields exactly $\hat{\tau}_{lin}$.*

Proof. For $z \in \{0, 1\}$ the original linear prediction function $\hat{\mu}_z$ is defined as

$$\hat{\mu}_z(x) = \beta_z^{\mathrm{T}} x + \gamma_z$$

where β_z and γ_z are the solutions to the L_2-norm empirical risk minimization problem

$$(\beta_z, \gamma_z) = \underset{\substack{\beta_z \in \mathbb{R}^k \\ \gamma_z \in \mathbb{R}}}{\arg\min} \left[\sum_{i \,:\, Z_i = z} \left(y_i(z) - \beta_z^{\mathrm{T}} x_i - \gamma_z \right)^2 \right]. \tag{15}$$

Let $\tilde{x}_i = \left(\hat{\mu}_0(x_i), \hat{\mu}_1(x_i)\right)^{\mathrm{T}}$. The calibrated predictor $\hat{\mu}_{OLS,z}$ is defined similarly as $\hat{\mu}_{OLS,z}(\tilde{x}) = \beta_{OLS,z}^{\mathrm{T}} \tilde{x} + \gamma_{OLS,z}$ for

$$(\beta_{OLS,z}, \gamma_{OLS,z}) = \underset{\substack{B_z \in \mathbb{R}^2 \\ \eta_z \in \mathbb{R}}}{\arg\min} \left[\sum_{i \,:\, Z_i = z} \left(y_i(z) - B_z^{\mathrm{T}} \tilde{x}_i - \eta_z \right)^2 \right]. \tag{16}$$

Writing $B_z \in \mathbb{R}^2$ as the vector $(a_z, b_z)^{\mathrm{T}}$ we expand the objective function of (16) via

$$\begin{aligned}\sum_{i\,:\,Z_i=z} \left(y_i(z) - B_z^{\mathrm{T}}\tilde{x}_i - \eta_z\right)^2 &= \sum_{i\,:\,Z_i=z} \Big\{ y_i(z) - \left(a_z\hat{\mu}_0(x_i) + b_z\hat{\mu}_1(x_i) + \eta_z\right) \Big\}^2 \\ &= \sum_{i\,:\,Z_i=z} \left\{ y_i(z) - \left(a_z\beta_0 + b_z\beta_1\right)^{\mathrm{T}} x_i - \left(a_z\gamma_0 + b_z\gamma_1 + \eta_z\right) \right\}^2 .\end{aligned}$$

Since $a_z\gamma_0 + b_z\gamma_1 + \eta_z$ ranges over all of $\mathbb{R}$ just as η_z does, by a change of variables we can write $\hat{\mu}_{OLS,z}(\tilde{x})$ directly as a function of the original covariates. This reduces to $ß_{OLS,z}^{\mathrm{T}} x + \gamma_{OLS,z}$ derived from the solution of

$$(ß_{OLS,z}, \gamma_{OLS,z}) = \underset{\substack{ß_z \in \mathrm{Span}(\beta_0,\beta_1) \\ \eta_z \in \mathbb{R}}}{\arg\min} \left[\sum_{i\,:\,Z_i=z} \left\{ y_i(z) - ß_z^{\mathrm{T}} x_i - \eta_z \right\}^2 \right]. \tag{17}$$

where $\mathrm{Span}(\beta_0, \beta_1)$ denotes the linear subspace of $\mathbb{R}^k$ generated by β_0 and β_1.

Trivially $\beta_0, \beta_1 \in \mathrm{Span}(\beta_0, \beta_1)$ and, by construction, these offer the solutions to the unconstrained optimization problems of (15) for $z = 0$ and $z = 1$, respectively; thus β_0 and β_1 offer feasible optimal solutions to the constrained problem (17) as well for $z = 0$ and $z = 1$, respectively. Consequently, $ß_{OLS,z} = \beta_z$ and $\gamma_{OLS,z} = \gamma_z$ for $z \in \{0, 1\}$. Thus, $\hat{\mu}_{OLS,0}(\tilde{x}_i) = \hat{\mu}_0(x_i)$ and likewise $\hat{\mu}_{OLS,1}(\tilde{x}_i) = \hat{\mu}_1(x_i)$. From this it follows that $\hat{\tau}_{cal} = \hat{\tau}_{lin}$. □

Remark 2. The proof of Proposition A.2 amounts to demonstrating that when the original prediction algorithms $\hat{\mu}_0$ and $\hat{\mu}_1$ are affine functions of the features x_i the optimality of the ordinary least squares regression coefficients implies that $\hat{\tau}_{cal}$ must match $\hat{\tau}_{lin}$.

Suppose that one calibrates $\hat{\mu}_0$ and $\hat{\mu}_1$ to form the estimator $\hat{\tau}_{cal}$ and the predictors $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$. In light of Theorem A.5, imagine that one calibrates again: in other words, one forms new predictors $\hat{\mu}_{OLS2,0}$ and $\hat{\mu}_{OLS2,1}$ by regressing upon $(\hat{\mu}_{OLS,0}(x_i), \hat{\mu}_{OLS,1}(x_i))^{\mathrm{T}}$ in the treated and control groups separately. Mirroring the logic of the proof above, we can write $\hat{\mu}_{OLS2,0}$ and $\hat{\mu}_{OLS2,1}$ directly as affine functions of the original prediction functions

$(\hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$ and the optimality of the original linear regressions used to form $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ implies that

$$\hat{\mu}_{OLS,0}(x_i) = \hat{\mu}_{OLS2,0}(x_i); \quad \hat{\mu}_{OLS,1}(x_i) = \hat{\mu}_{OLS2,1}(x_i).$$

This implies the idempotence of calibration.

For the purpose of easy computation, we present the following proposition. Section 2.13 below provides further implementation details.

Proposition A.3. *The calibrated estimator can be written exclusively in terms of the predicted values:* $\hat{\tau}_{cal} = N^{-1} \sum_{i=1}^{N} \{\hat{\mu}_{OLS,1}(x_i) - \hat{\mu}_{OLS,0}(x_i)\}$.

Proof. This follows from Proposition A.4; which we present below. □

Proposition A.4. *With probability one, the prediction functions* $\hat{\mu}_{OLS,0}$ *and* $\hat{\mu}_{OLS,1}$ *satisfy*

$$\sum_{i=1}^{N} \mathbb{1}_{\{Z_i=z\}} \hat{\mu}_{OLS,z}(x_i) = \sum_{i=1}^{N} \mathbb{1}_{\{Z_i=z\}} y_i(z). \tag{18}$$

In the terminology of (GB21, Definition 1) $\hat{\mu}_{OLS,0}$ *and* $\hat{\mu}_{OLS,1}$ *are prediction unbiased.*

Proof. This is a direct consequence of the first order optimality conditions for the intercept term in the regressions defining $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$. □

Remark 3. Since the proof of Proposition A.4 relied on the first-order optimality condition for the intercept term in the regressions defining $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ the logic of the proof applies equally well when $\tilde{x}_i$ is replaced with $\ddot{x}_i$, thereby easily incorporating feature engineering.

In what follows we focus attention on the resulting prediction equations in the treated group ($z = 1$) after linear calibration, and will at times introduce quantities with the dependence on z suppressed for readability. The proofs for the control group are analogous. Recall from before the quantities

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \left[\sum_{i \,:\, Z_i = 1} \left\{ y_i(Z_i) - \left(\beta_0 + \beta_1^T \begin{bmatrix} \hat{\mu}_0(x_i) \\ \hat{\mu}_1(x_i) \end{bmatrix} \right) \right\}^2 \right], \tag{19}$$

$$\dot{\beta} = (\dot{\beta}_0, \dot{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \left[\sum_{i=1}^{N} \left\{ y_i(1) - \left(\beta_0 + \beta_1^T \begin{bmatrix} \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} \right) \right\}^2 \right]. \tag{20}$$

The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are the ordinary least squares linear regression intercept and slope coefficients, respectively, for the sample regression of treated outcomes on the imputed values $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$. Consequently $\hat{\beta}$ can be computed based upon the observed data. In contrast, $\dot{\beta}_0$ and $\dot{\beta}_1$ are the population-level ordinary least squares regression intercept and slope coefficients, respectively, for the regression of all treated potential outcomes upon the "imputed" values $\dot{\mu}_0(x_i)$ and $\dot{\mu}_1(x_i)$. Importantly, $\dot{\beta}$ is not generally computable from the observed data for two reasons:

1. The minimization problem defining $\dot{\beta}$ requires knowledge of the treated potential outcomes for each individual, not just those who received treatment.

2. In practice, the functions $\dot{\mu}_0(x_i)$ and $\dot{\mu}_1(x_i)$ are frequently unknown. For instance, they can often be taken to be solutions to population-level risk minimization procedures; see (GB21) for some examples of this.

Define $\dot{\mu}_{OLS,1}(\cdot)$ to be $\dot{\beta}_0 + \dot{\beta}_1^{\mathrm{T}} \begin{bmatrix} \dot{\mu}_0(\cdot) \\ \dot{\mu}_1(\cdot) \end{bmatrix}$.

Proposition A.5. *If the original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are stable in the sense of Assumption 1, then under Assumptions 4 and 5 the calibrated prediction functions $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ are also stable.*

Proof. Our proof focuses on $\hat{\mu}_{OLS,1}$ and uses the notation of (19) and (20); the proof for $\hat{\mu}_{OLS,0}$ follows the same logic, but requires exchanging control quantities with treated quantities in the obvious places.

To show stability of $\hat{\mu}_{OLS,1}$ we will show that

$$\frac{1}{N}\sum_{i=1}^{N}\left|\hat{\mu}_{OLS,1}(x_i) - \dot{\mu}_{OLS,1}(x_i)\right|^2 = o_P(1).$$

This amounts to examining

$$\frac{1}{N}\sum_{i=1}^{N}\left|\left(\hat{\beta}_0 + \hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix}\right) - \left(\dot{\beta}_0 + \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right)\right|^2. \tag{21}$$

Clearly this quantity is non-negative, so our objective is to provide an $o_P(1)$ upper bound. By the triangle inequality (21) is upper bounded by

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\left|\hat{\beta}_0 - \dot{\beta}_0\right| + \left|\hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right\}^2.$$

By Lemma A.3 the term $\left|\hat{\beta}_0 - \dot{\beta}_0\right| = o_P(1)$ and so (21) is upper bounded by

$$\frac{1}{N}\sum_{i=1}^{N}\left\{o_P(1) + \left|\hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right\}^2. \tag{22}$$

In light of (22) we turn our focus to the term

$$\begin{aligned}
\left|\hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right| &= \left|\hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix} + \right. \\
&\qquad \left. \hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix} - \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right| \\
&\leq \left|\hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right| + \\
&\qquad \left|\hat{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix} - \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right| \\
&\leq \left|\left|\hat{\beta}_1\right|\right|_2 \left|\left|\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 + \left|\left|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 \left|\left|\hat{\beta}_1^{\mathrm{T}} - \dot{\beta}_1^{\mathrm{T}}\right|\right|_2 \\
&\leq \left|\left|\hat{\beta}_1\right|\right|_2 \left|\left|\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 + \left|\left|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 o_P(1) \\
&\leq O_P(1)\left|\left|\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 + \left|\left|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 o_P(1). \qquad (23)
\end{aligned}$$

The second line follows from the triangle inequality, the third line from the Cauchy-Schwarz inequality, and the fourth line from Lemma A.3. The last line follows from $\left|\left|\hat{\beta}_1\right|\right|_2 = O_P(1)$. To see why this is the case, notice that by standard theory for ordinary least squares linear regression, under Assumptions 4 and 5, $\dot{\beta}_1$ converges to a fixed vector in $\mathbb{R}^2$; so the consistency of $\hat{\beta}_1$ implies that $\left|\left|\hat{\beta}_1\right|\right|_2 = O_P(1)$.

Combining (22) with (23) gives that (21) is upper bounded by

$$\frac{1}{N}\sum_{i=1}^{N}\left\{o_P(1) + O_P(1)\left|\left|\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 + \left|\left|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right|\right|_2 o_P(1)\right\}^2. \qquad (24)$$

Applying the inequality $2a^2 + 2b^2 \geq (a+b)^2$ twice yields that (24) is bounded above by

$$\frac{4}{N}\sum_{i=1}^{N}\left[\left\{O_P(1)\left\|\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right\|_2\right\}^2 + \left\{\left\|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right\|_2 o_P(1)\right\}^2 + o_P(1)\right] =$$

$$\underbrace{\frac{O_P(1)}{N}\sum_{i=1}^{N}\left\|\begin{bmatrix}\hat{\mu}_0(x_i)\\ \hat{\mu}_1(x_i)\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right\|_2^2}_{\text{Term 1}} + \underbrace{\frac{o_P(1)}{N}\sum_{i=1}^{N}\left\|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right\|_2^2}_{\text{Term 2}} + o_P(1).$$

Term 1 is vanishing in probability by the stability of $\hat{\mu}_0$ and $\hat{\mu}_1$. By Assumptions 4 and 5,

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right\|_2^2$$

is limits to a constant, so Term 2 vanishes in probability as well. In total, we have established that 21 is $o_P(1)$ which concludes the proof. □

Remark 4. If Assumptions 4 and 5 apply to the feature engineered pseudo-covariates $(f(x_i)^{\mathrm{T}}, \dot{\mu}_0(x_i), \dot{\mu}_1(x_i))^{\mathrm{T}}$ then the same logic applies to the calibration regressions performed in the formation of $\hat{\tau}_{cal2}$. Consequently, the calibrated prediction functions used for the feature-engineered estimator $\hat{\tau}_{cal2}$ are prediction unbiased (in the sense of (GB21, Definition 1)) and stable.

We restate the theorems from above with the precise regularity conditions required.

Theorem 1. *Suppose that the completely randomized sampling is non-degenerate (Assumption A.6). If the original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are stable with vanishing error processes (Assumptions 1 and 2) and satisfy Assumptions 4 and 5, then $\hat{\tau}_{cal}$ is consistent, obeys a central limit theorem, and is asymptotically no less efficient than $\hat{\tau}_{unadj}$.*

Proof. Proposition A.4 establishes that $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ are prediction unbiased (in the sense of (GB21, Definition 1)). Proposition A.5 implies the stability of the prediction functions $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$. Proposition A.1 implies that $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ satisfy the vanishing error process assumption. Thus, the argument of Theorem 2 of (GB21) guarantees the consistency of $\hat{\tau}_{cal}$ as long as

$$MSE_N(z) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i(z) - \dot{\mu}_{OLS,z}(x_i)\right)^2 = o(N) \quad \text{for } z \in \{0,1\} \tag{25}$$

a requirement which is met by Assumptions 4 and 5.

The consistency of $\hat{\tau}_{cal}$ does not rely upon whether or not any of the regression models $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{\mu}_{OLS,0}$, and $\hat{\mu}_{OLS,1}$ are "well-specified" with respect to the data-generating process that gave rise to the potential outcomes and covariates. In other words, consistency is derived without any special regard for knowing how the data came to be.

Furthermore, since Proposition A.1 implies that $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ have vanishing error processes the argument of (GB21, Theorem 3) implies that $\hat{\tau}_{cal}$ has an asymptotically linear reformulation:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i(1) - y_i(1)\right) = \frac{1}{n_1}\sum_{i\,:\,Z_i=1}\underbrace{\left(y_i(1) - \dot{\mu}_{OLS,1}(x_i)\right)}_{\dot{\epsilon}_i(1)} + o_P(N^{-1/2}), \tag{26}$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i(0) - y_i(0)\right) = \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\underbrace{\left(y_i(0) - \dot{\mu}_{OLS,0}(x_i)\right)}_{\dot{\epsilon}_i(0)} + o_P(N^{-1/2}). \tag{27}$$

Equations (26) and (27) show that the calibrated estimator $\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}$ is equivalent, up

to an error of $o_P(N^{-1/2})$, to the difference in means estimator for a population where the potential outcomes are given by the residuals $\dot{\epsilon}_i(1)$ and $\dot{\epsilon}_i(0)$ instead of the original potential outcomes. By standard central limit theorems for the difference in means, e.g., (LD17), it follows that under the conditions of Corollary 1 in (GB21)

$$N^{1/2}\left(\frac{\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}}{\sigma_N}\right)\xrightarrow{d}\mathcal{N}\left(0,1\right),$$

where

$$\sigma_N^2=\left(\frac{1}{n_1}\right)MSE_N(1)+\left(\frac{1}{n_0}\right)MSE_N(0)-\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\dot{\epsilon}_i(1)-\dot{\epsilon}_i(0)\right)^2. \tag{28}$$

Assumptions 4 and 5 are sufficient conditions for Corollary 1 of (GB21).

What remains to be shown is that $\hat{\tau}_{cal}$ is asymptotically no less efficient than $\hat{\tau}_{unadj}$. In other words, we must show that

$$\lim_{N\to\infty}N\sigma_N^2\leq\frac{\Sigma_{y(1),\infty}}{p}+\frac{\Sigma_{y(0),\infty}}{1-p}-\Sigma_{\tau,\infty} \tag{29}$$

where $\Sigma_{\tau,\infty}$ is the limiting variance of $\{\tau_i=y_i(1)-y_i(0)\}_{i=1}^N$. The limit of $N\sigma_N^2$ is guaranteed to exist due to Assumption 4.

Consider forming Lin's regression adjusted estimator by regressing the observed outcomes $y_i(Z_i)$ on the deterministic features $(\dot{\mu}_0(x_i),\dot{\mu}_1(x_i))^{\text{T}}$ separately in the treated and control groups, imputing counterfactuals based upon the corresponding regressions, and then taking the difference in means across the imputed populations. Call this estimator $\hat{T}_{lin}$. $\hat{T}_{lin}$ differs from $\hat{\tau}_{cal}$ in that the features used for regression are not those estimated from the sample, i.e., $(\hat{\mu}_0(x_i),\hat{\mu}_1(x_i))^{\text{T}}$. Nonetheless, Theorem 1 of (Lin13) gives that the asymptotic variance of $N^{1/2}\left(\hat{T}_{lin}-\overline{\tau}_{\text{SATE}}\right)$ is $\lim_{N\to\infty}N\sigma_N^2$. Corollary 1 of (Lin13) then implies the non-inferiority of $\hat{T}_{lin}$ relative to the unadjusted difference in means, which is equivalent to the inequality

of 29. This indirectly shows the non-inferiority of $\hat{\tau}_{cal}$ relative to the unadjusted difference in means $\hat{\tau}_{unadj}$.

□

Theorem 2. *Under the regularity conditions of Theorem 1 and for a given set of prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$, the calibrated estimator $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}})$ has an asymptotic variance that is no larger than that of both $N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}})$ and $N^{1/2}(\hat{\tau}_{GBcal} - \overline{\tau}_{\text{SATE}})$.*

Proof. By Theorem 3 of (GB21), $N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}})$ differs by $o_P(1)$ from the difference in means statistic

$$\frac{1}{n_1} \sum_{i\,:\,Z_i=1} \dot{\epsilon}_i(1) - \frac{1}{n_0} \sum_{i\,:\,Z_i=0} \dot{\epsilon}_i(0) \tag{30}$$

where $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(x_i)$ for $z \in \{0, 1\}$.

In contrast, by (26) and (27), $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}})$ differs by $o_P(1)$ from the difference in means statistic

$$\frac{1}{n_1} \sum_{i\,:\,Z_i=1} \dot{\epsilon}_i(1) - \frac{1}{n_0} \sum_{i\,:\,Z_i=0} \dot{\epsilon}_i(0). \tag{31}$$

Unpacking the notation of (30)

$$\begin{aligned} N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}}) &= \frac{1}{n_1} \sum_{i\,:\,Z_i=1} (y_i(1) - \dot{\mu}_1(x_i)) - \frac{1}{n_0} \sum_{i\,:\,Z_i=0} (y_i(0) - \dot{\mu}_0(x_i)) + o_P(1) \\ &= \frac{1}{n_1} \sum_{i\,:\,Z_i=1} \left(y_i(1) - \mathbf{e}_2^{\mathrm{T}} \begin{bmatrix} \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} \right) - \\ &\qquad \frac{1}{n_0} \sum_{i\,:\,Z_i=0} \left(y_i(0) - \mathbf{e}_1^{\mathrm{T}} \begin{bmatrix} \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} \right) + o_P(1) \end{aligned} \tag{32}$$

where $\mathbf{e}_i$ is the i^{th} standard basis vector. The same logic applied to (31) yields that

$$\begin{aligned} N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right) &= \frac{1}{n_1}\sum_{i\,:\,Z_i=1}\left(y_i(1) - \dot{\mu}_{OLS,1}(x_i)\right) - \\ &\qquad \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\left(y_i(0) - \dot{\mu}_{OLS,0}(x_i)\right) + o_P(1) \\ &= \frac{1}{n_1}\sum_{i\,:\,Z_i=1}\left\{y_i(1) - \left(\dot{\beta}_0 + \dot{\beta}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right)\right\} - \\ &\qquad \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\left\{y_i(0) - \left(\dot{B}_0 + \dot{B}_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right)\right\} + o_P(1) \end{aligned} \tag{33}$$

where $\dot{\beta}_0$ is the intercept of $\dot{\mu}_{OLS,1}(\cdot)$, $\dot{\beta}_1$ is the slope of $\dot{\mu}_{OLS,1}(\cdot)$, and $(\dot{B}_0, \dot{B}_1)$ are defined similarly for $\dot{\mu}_{OLS,0}(\cdot)$.

Notice that both (32) and (33) are, up to an $o_P(1)$ difference, of the form

$$\begin{aligned} &\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\left\{y_i(1) - \left(\beta_0 + \beta_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right)\right\} - \\ &\qquad\qquad \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\left\{y_i(0) - \left(B_0 + B_1^{\mathrm{T}}\begin{bmatrix}\dot{\mu}_0(x_i)\\ \dot{\mu}_1(x_i)\end{bmatrix}\right)\right\}, \end{aligned} \tag{34}$$

with the only difference being that in (32) we take $(\beta_1, B_1) = (\mathbf{e}_2, \mathbf{e}_1)$ and $(\beta_0, B_0) = (0, 0)$ whereas in (33) we take $(\beta_1, B_1) = \left(\dot{\beta}_1, \dot{B}_1\right)$ and $(\beta_0, B_0) = (\dot{\beta}_0, \dot{B}_0)$. Since, the $o_P(1)$ term contributes nothing to the asymptotic variance of either quantity of interest, we neglect it in the remainder of our analysis.

By the argument presented in Section 4.1 of (Lin13) and Lemma A.2 the variance of (34) is minimized when $(\beta_0, \beta_1, B_0, B_1)$ is taken to be the population ordinary least squares linear regression intercepts and slopes. This is exactly the case for (33), whereas (32) presents a feasible, but not necessarily optimal, solution to the ordinary least squares prob-

lem. Consequently, the asymptotic variance of $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ does not exceed that of $N^{1/2}\left(\hat{\tau}_{gOB} - \overline{\tau}_{\text{SATE}}\right)$ and the inequality is strict when the population least squares problem has a unique optimal solution which does not degenerate to $|\hat{\tau}_{cal} - \hat{\tau}_{gOB}| = o_p(N^{-1/2})$.

The same style of proof also shows that the asymptotic variance of $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ does not exceed that of $N^{1/2}\left(\hat{\tau}_{GBcal} - \overline{\tau}_{\text{SATE}}\right)$. □

Here we restate Theorem A.4 with precise regularity conditions and provide its proof.

Theorem A.1. *Assume the regularity conditions of Theorem 1. So long as the random vectors* $\left(f(x_i)^{\mathrm{T}}, \hat{\mu}_0(x_i), \hat{\mu}_1(x_i)\right)^{\mathrm{T}}$ *satisfy Assumptions 4 and 5, a central limit theorem applies to* $N^{1/2}\left(\hat{\tau}_{cal2} - \overline{\tau}_{\text{SATE}}\right)$ *and it is non-inferior to both* $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ *and* $N^{1/2}\left(\hat{\tau}_{lin} - \overline{\tau}_{\text{SATE}}\right)$ *using the engineered features* $f(x_i)$.

Proof. The mechanics of proving consistency and central limit behavior for $\hat{\tau}_{cal2}$ mirror those of the proof for Theorem 1. Stability follows from Proposition A.5 using the adaptation in Remarks 4 and 1. Vanishing of the error processes for the prediction functions of $\hat{\tau}_{cal2}$ follows from modifying Proposition A.1 in the natural way to account for the additional features $f(x_i)$.

The asymptotic variance of $\hat{\tau}_{cal2}$ is the same as the asymptotic variance of Lin's regression adjusted estimator which regresses upon $\left(f(x_i)^{\mathrm{T}}, \dot{\mu}_0(x_i), \dot{\mu}_1(x_i)\right)$; call this estimator $\hat{T}_{f,\dot{\mu}}$. Likewise, the asymptotic variance of $\hat{\tau}_{cal}$ is the same as the same as the asymptotic variance of the regression adjusted estimator which regresses upon $\left(\dot{\mu}_0(x_i), \dot{\mu}_1(x_i)\right)$; in the proof of Theorem 1 we called this estimator $\hat{T}_{lin}$. By Lemma A.2 the inclusion of the additional features $f(x_i)$ guarantees that $\hat{T}_{f,\dot{\mu}}$ is non-inferior to $\hat{T}_{lin}$. Similarly, the inclusion of the additional features $\left(\dot{\mu}_0(x_i), \dot{\mu}_1(x_i)\right)$ guarantees that $\hat{T}_{f,\dot{\mu}}$ is non-inferior to $\hat{\tau}_{lin}$ using the engineered

features $f(x_i)$. Both of these statements rest upon the observation that the asymptotic variance of the regression adjusted estimator is never increased by including additional features; this follows easily from Lemma A.2 by observing that regressions without the added features are equivalent to regressions with the extra features but with coefficients restricted to zero for such features. □

2.13 Implementation

In Algorithm 1 we include pseudocode for the construction of $\hat{\tau}_{cal}$ in order to facilitate easy implementation.

Algorithm 1: Computation of the calibrated Oaxaca-Blinder estimator.

Input: An observed treatment allocation Z, with observed responses $y_1(Z_1), \ldots, y_N(Z_N)$ and covariates $x_1, \ldots, x_N$.

Result: The calibrated estimator $\hat{\tau}_{cal}$.

Step 1: Initial predictions

Train the initial prediction algorithms $\hat{\mu}_0$ and $\hat{\mu}_1$;

for $i = 1, \ldots, N$ **do**
  Impute the outcomes $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$.

end

Step 2: Calibrate prediction functions

for $z \in \{0, 1\}$ **do**
  Let $\hat{\mu}_{OLS,z}$ be the linear regression of $y_i(Z_i)$ on $(\hat{\mu}_0(x_i), \hat{\mu}_1(x_i))^{\mathrm{T}}$ in the group $\{i \,:\, Z_i = z\}$.

end

return $\hat{\tau}_{cal} = N^{-1} \sum_{i=1}^{N} \{\hat{\mu}_{OLS,1}(x_i) - \hat{\mu}_{OLS,0}(x_i)\}$

Additionally, code written in `R` is available at:

`https://github.com/PeterLCohen/OaxacaBlinderCalibration` to demonstrate computation of $\hat{\tau}_{cal}$ and $\hat{\tau}_{cal2}$ and to provide examples of their use on real data sets.

2.14 Rank-Deficiency

Throughout the proofs above we worked under the assumption that regressions, both those defining $\hat{\beta}$ and $\dot{\beta}$, are not rank-deficient. From the perspective of (19) and (20) this assumption guaranteed unique solutions to the empirical risk minimization problems undergirding ordinary least squares regression. If instead these regressions had linear dependencies between their features, these minimization problems would have uncountably many solutions. This would present a mathematical impediment to the style of proofs used in the preceding sections, but would not invalidate the general results.

In the case that the regression problems (19) and (20) have uncountably many solutions, there is no uniquely identified choice of $\hat{\beta}^{(N)}$ or $\dot{\beta}^{(N)}$ to make the consistency statement $||\hat{\beta}^{(N)} - \dot{\beta}^{(N)}||_2 = o_P(1)$ make sense. However, the predictions $\hat{\mu}_{OLS,z}(x_i)$ and $\dot{\mu}_{OLS,z}(x_i)$ are still well-defined for all x_i. Because of this, the residuals $\dot{\epsilon}_i(z)$ remain well defined; and so intuition derived from the asymptotically linear expansions (26) and (27) leads one to expect no change to our main results.

In order to rectify the proofs for stability, vanishing error processes, and non-inferiority to accord with the uncountable solution-space to the ordinary least squares regression problems (19) and (20) one can pick a canonical representative from the class of optimal solutions. For this, the linear regression coefficients given by the Moore-Penrose pseudoinverse (Pen55) suffice; see (GVL13, Chapter 5.5) for the general relationship between rank-deficient least squares and matrix pseudoinverses. Specifically, let the Moore-Penrose pseudoinverse of a matrix M be M^+. Then take the canonical regression coefficients to be

$$\hat{\beta}^+ = \left(\hat{U}_1^{\mathrm{T}}\hat{U}_1\right)^+ \hat{U}_1^{\mathrm{T}} \begin{bmatrix} y_{i_1}(1) \\ \vdots \\ y_{i_{n_1}}(1) \end{bmatrix} \text{ and } \dot{\beta}^+ = \left(\dot{U}_1^{\mathrm{T}}\dot{U}_1\right)^+ \dot{U}_1^{\mathrm{T}} \begin{bmatrix} y_1(1) \\ \vdots \\ y_N(1) \end{bmatrix}.$$

The proofs for stability, vanishing error processes, and non-inferiority follow through as

before but with $\hat{\beta}^+$ and $\dot{\beta}^+$ replacing $\hat{\beta}$ and $\dot{\beta}$, respectively. A minor technical consideration is required: the continuous mapping theorem must be applicable as it was in Lemma A.3. Unlike the matrix inverse map $M \mapsto M^{-1}$, the Moore-Penrose pseudoinverse map $M \mapsto M^+$ is not a continuous map over the positive semidefinite cone under the metric induced by the Frobenius norm (Ste69). The discontinuity of $M \mapsto M^+$ can be resolved by stipulating that the there exists some $\aleph \in \mathbb{N}$ such that (almost surely) rank $\left(\hat{U}_1^{\mathrm{T}}\hat{U}_1\right)$ and rank $\left(\dot{U}_1^{\mathrm{T}}\dot{U}_1\right)$ are equal to a common constant for all populations of size $N \geq \aleph$ (Rak97); then the continuous mapping theorem for general metric spaces of (vdV98, Theorem 18.11) applies as needed to carry through the rest of the proofs in the same style as before.

2.15 Variance Estimation And Inference Under The Finite Population Model

The asymptotic Gaussianity of $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}})$ facilitates inference for the sample average treatment effect using $\hat{\tau}_{cal}$. To proceed, we provide an asymptotically conservative variance estimator for $\hat{\tau}_{cal}$. Define

$$\hat{\Sigma}_z := \frac{1}{n_z - 1} \sum_{i\,:\,Z_i = z} \left(y_i(z) - \hat{\mu}_z(x_i)\right)^2,$$
$$\hat{V} = \frac{\hat{\Sigma}_1}{n_1} + \frac{\hat{\Sigma}_0}{n_0}.$$

By Theorem 4 of (GB21), $\hat{V}$ provides an asymptotically conservative estimate of the variance of the generalized Oaxaca-Blinder estimator using imputation functions $\hat{\mu}_0$ and $\hat{\mu}_1$; their result holds for the finite population model and accounts only for variability arising from the stochasticity of the treatment allocation process. Consequently, the variance of the calibrated estimator, which is a particular form of generalized Oaxaca-Blinder estimator, is estimated

by

$$\hat{\Sigma}_{z,cal} := \frac{1}{n_z - 1} \sum_{i\,:\,Z_i = z} \left(y_i(z) - \hat{\mu}_{OLS,z}(x_i)\right)^2,$$
$$\hat{V}_{cal} = \frac{\hat{\Sigma}_{1,cal}}{n_1} + \frac{\hat{\Sigma}_{0,cal}}{n_0}. \tag{35}$$

In a finite population $\hat{V}_{cal}$ is potentially asymptotically conservative relative to the true variance of $\hat{\tau}_{cal}$, and so asymptotically valid – though potentially conservative – confidence intervals may be formed based upon the usual normal approximation. In Section 2.16 we discuss variance estimators which account for stochasticity in the potential outcomes themselves.

2.16 Linear Calibration In Alternative Models

2.16.1 Superpopulation and Fixed-Covariate Models

The asymptotic non-inferiority of linearly calibrated generalized Oaxaca-Blinder estimators is not tied to the finite population model. Here we detail two alternative common models for which linear calibration guarantees asymptotic non-inferiority of generalized Oaxaca-Blinder estimators: the superpopulation model and the fixed-covariate model. In the superpopulation model the vector of potential outcomes and covariates $(y_i(0), y_i(1), x_i)$ is an independent and identically distributed draw from some fixed distribution for each $i \in \{1, \ldots, N\}$. In the fixed-covariate model the ith unit's covariates x_i are deterministic while the potential outcomes $(y_i(0), y_i(1))$ are independent draws from some conditional distribution given x_i. In the superpopulation framework, the targeted estimand is the population average treatment effect (PATE), $\overline{\tau}_{\text{PATE}} = E[y_i(1) - y_i(0)]$. Under the fixed covariate model, the estimand of interest is the conditional average treatment effect (CATE), $\overline{\tau}_{\text{CATE}} = N^{-1} \sum_{i=1}^{N} E[y_i(1) - y_i(0) \mid x_i]$. In both of these models linear calibration functions exactly the same as in the finite popu-

lation model: estimate the nonlinear prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ on the observed data and subsequently form Lin's linearly-adjusted estimator using the imputed values $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$ in lieu of the covariates x_i. The natural analogs of Theorem 1 hold so long as the required assumptions are modified to reflect the chosen probabilistic framework.

2.16.2 Linear Calibration for the Population Average Treatment Effect

Under the superpopulation model, Assumption 1 (Stability) remains unchanged. Likewise Assumption 2 stays the same, though the stochastic process now inherits randomness from the potential outcomes and covariates in addition to the randomness in treatment allocation. Assumptions 4 and 5 require simple modifications to adapt to randomness in the potential outcomes and covariates.

Assumption 8 (Superpopulation Limiting Means and Variances)**.** *The mean vector and covariance matrix of* $(y_i(0), y_i(1), \dot{\mu}_0(x_i), \dot{\mu}_0(x_i))^{\mathrm{T}}$ *have limiting values. For instance, for* $z = 0, 1$ *there exists a limiting value* $\bar{y}(z)_\infty$ *such that* $\lim_{N\to\infty} \mathbb{E}\left[y_i(z)\right] = \bar{y}(z)_\infty$.

Assumption 9 (Superpopulation Bounded Fourth Moments)**.** *There exists some* $C < \infty$ *for which, for all* $z = 0, 1$ *and all* N, $\mathbb{E}\left[y_i(z)^4\right] < C$ *and* $\mathbb{E}\left[\dot{\mu}_z(x_i)^4\right] < C$.

In the superpopulation case, the definition of $\dot{\beta}$ given in Lemma A.3 requires a minor change to

$$\dot{\beta} = (\dot{\beta}_0, \dot{\beta}_1) = \underset{\beta_0,\beta_1}{\arg\min}\, \mathbb{E}\left[\left\{ y_i(1) - \left(\beta_0 + \beta_1^T \begin{bmatrix} \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} \right) \right\}^2 \right].$$

Then $||\hat{\beta} - \dot{\beta}||_2 = o_P(1)$ by the standard consistency of sample ordinary least squares regression coefficients under Assumptions 8 and 9; this consistency holds even without assuming the truth of a linear model (BBB$^+$19, Proposition 7).

Lemma A.7 (Asymptotically Linear Expansions around the SATE)**.** *Under Assumptions A.6, 1, and 2 the random variable* $N^{-1}\sum_{i=1}^{N}\left(\hat{\mu}_z(x_i) - y_i(z)\right)$ *is asymptotically linear in the sense that*

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_z(x_i) - y_i(z)\right) = \frac{1}{n_z}\sum_{i\,:\,Z_i=z}\dot{\epsilon}_i(z) + o_p(N^{-1/2})$$

where $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(x_i)$.

Proof. By the exact same reasoning as in (GB21, Proof of Theorem 3), rewrite $N^{-1}\sum_{i=1}^{N}\left(\hat{\mu}_z(x_i) - y_i(z)\right)$ as

$$\frac{1}{n_z}\sum_{i\,:\,Z_i=z}\dot{\epsilon}_i(z) + \frac{1}{N}\Bigg(\underbrace{\sum_{i=1}^{N}\left(\frac{Z_i\dot{\mu}_z(x_i)}{(n_z/N)} - \dot{\mu}_z(x_i)\right)}_{N^{1/2}\mathcal{G}_{N,z}(\dot{\mu}_z)} - \underbrace{\sum_{i=1}^{N}\left(\frac{Z_i\hat{\mu}_z(x_i)}{(n_z/N)} - \hat{\mu}_z(x_i)\right)}_{N^{1/2}\mathcal{G}_{N,z}(\hat{\mu}_z)}\Bigg).$$

Consequently, the desired result holds so long as $|\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z)| = o_P(1)$; this holds by Assumption 2. □

Theorem A.2. *Under the superpopulation model, subject to Assumptions A.6, 1, 2, 8 and 9,* $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}}\right)$ *obeys a central limit theorem.*

Proof. We start with an algebraic decomposition:

$$\begin{aligned}
N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}}\right) &= N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,1}(x_i) - \hat{\mu}_{OLS,0}(x_i)\right) - \overline{\tau}_{\text{PATE}}\right) \\
&= N^{1/2}\Bigg(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,1}(x_i) - \hat{\mu}_{OLS,0}(x_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(y_i(1) - y_i(0)\right) + \\
&\qquad \frac{1}{N}\sum_{i=1}^{N}\left(y_i(1) - y_i(0)\right) - \overline{\tau}_{\text{PATE}}\Bigg) \\
&= N^{1/2}\Bigg(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,1}(x_i) - y_i(1)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,0}(x_i) - y_i(0)\right) + \\
&\qquad \frac{1}{N}\sum_{i=1}^{N}\left(y_i(1) - y_i(0)\right) - \overline{\tau}_{\text{PATE}}\Bigg)
\end{aligned}$$

By Lemma A.7 this final term equals

$$N^{1/2}\Bigg(\frac{1}{N}\sum_{i=1}^{N} Z_i N n_1^{-1}\dot{\epsilon}_i(1) - \frac{1}{N}\sum_{i=1}^{N}(1 - Z_i) N n_0^{-1}\dot{\epsilon}_i(0) + \frac{1}{N}\sum_{i=1}^{N}\left(y_i(1) - y_i(0)\right) - \overline{\tau}_{\text{PATE}}\Bigg) + o_P(1)$$

where $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_{OLS,z}(x_i)$. By the reasoning of Lemma A.5 we stipulate that $\mathbb{E}\left[\dot{\epsilon}_i(z)\right] = 0$ for $z \in \{0, 1\}$. Rearranging the formula above yields

$$N^{1/2}\Bigg(\frac{1}{N}\sum_{i=1}^{N} Z_i N\left(n_1^{-1}\dot{\epsilon}_i(1) + n_0^{-1}\dot{\epsilon}_i(0)\right) - \frac{1}{N}\sum_{i=1}^{N} N n_0^{-1}\dot{\epsilon}_i(0) + \frac{1}{N}\sum_{i=1}^{N}\left(y_i(1) - y_i(0)\right) - \overline{\tau}_{\text{PATE}}\Bigg) + o_P(1)$$

Consider the random variable

$$\varphi(Z_i) := Z_i N \left(n_1^{-1}\dot{\epsilon}_i(1) + n_0^{-1}\dot{\epsilon}_i(0)\right) - N n_0^{-1}\dot{\epsilon}_i(0) + (y_i(1) - y_i(0))\,;$$

to prove the desired result it suffices to show a central limit theorem for $\varphi(Z_i)$. To do this we will start by conditioning upon the event $Z = \tilde{z}$ for some $\tilde{z} \in \Omega_{CRE}$; we will show that a central limit theorem applies to $\varphi(\tilde{z}_i)$ almost surely with respect to the conditioning event $Z = \tilde{z} \in \Omega_{CRE}$; then we will leverage this conditional central limit theorem to deduce that an unconditional central limit theorem applies to $\varphi(Z_i)$.

Condition on the event $Z = \tilde{z}$ for some $\tilde{z} \in \Omega_{CRE}$ and recall that Z_i is independent of $(y_i(0), y_i(1), x_i)$ so this conditioning changes nothing of the distribution of the $\dot{\epsilon}_i(z)$ and the $y_i(z)$ for all $i \in \{1, \ldots, N\}$ and $z \in \{0, 1\}$. We examine the conditional expectation $\mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]$.

$$\begin{aligned}
\mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right] &= \mathbb{E}\left[Z_i N \left(n_1^{-1}\dot{\epsilon}_i(1) + n_0^{-1}\dot{\epsilon}_i(0)\right) \mid Z = \tilde{z}\right] - \\
&\qquad \mathbb{E}\left[N n_0^{-1}\dot{\epsilon}_i(0) \mid Z = \tilde{z}\right] + \mathbb{E}\left[(y_i(1) - y_i(0)) \mid Z = \tilde{z}\right] \\
&= \tilde{z}_i \underbrace{\mathbb{E}\left[N \left(n_1^{-1}\dot{\epsilon}_i(1) + n_0^{-1}\dot{\epsilon}_i(0)\right)\right]}_{0} - \underbrace{\mathbb{E}\left[N n_0^{-1}\dot{\epsilon}_i(0)\right]}_{0} + \\
&\qquad \underbrace{\mathbb{E}\left[(y_i(1) - y_i(0))\right]}_{\overline{\tau}_{\text{PATE}}} \\
&= \overline{\tau}_{\text{PATE}}.
\end{aligned}$$

The random variables $\varphi(Z_1), \ldots, \varphi(Z_n)$ are a conditionally independent family of random variables given the event $Z = \tilde{z}$; however, they are not identically distributed. For all i such that $\tilde{z}_i = 0$ the random variable $\varphi(\tilde{z}_i)$ is equal in distribution to $-N n_0^{-1}\dot{\epsilon}_i(0) + (y_i(1) - y_i(0))$. For the i such that $\tilde{z}_i = 1$ the random variable $\varphi(\tilde{z}_i)$ is equal in distribution

to $N\left(n_1^{-1}\dot{\epsilon}_i(1) + n_0^{-1}\dot{\epsilon}_i(0)\right) - Nn_0^{-1}\dot{\epsilon}_i(0) + (y_i(1) - y_i(0))$. Motivated by this, we study

$$\begin{aligned} N^{-1/2}\left(\sum_{i=1}^{N}\varphi(\tilde{z}_i) - \overline{\tau}_{\text{PATE}}\right) &= N^{-1/2}\left(\sum_{i=1}^{N}\varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right) \\ &= \left(N^{-1/2}n_0^{1/2}\right) n_0^{-1/2}\left(\sum_{i\,:\,\tilde{z}_i=0}\varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right) + \\ &\qquad \left(N^{-1/2}n_1^{1/2}\right) n_1^{-1/2}\left(\sum_{i\,:\,\tilde{z}_i=1}\varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right). \end{aligned}$$

The first term on the right-hand-side is independent from the second term on the right-hand-side; furthermore:

- The first term is a $n_0^{-1/2}$-scaled sum of n_0 independent terms each equal in distribution to $-Nn_0^{-1}\dot{\epsilon}_i(0) + (y_i(1) - y_i(0))$;
- The second term is a $n_1^{-1/2}$-scaled sum of n_1 independent terms each equal in distribution to $N\left(n_1^{-1}\dot{\epsilon}_i(1) + n_0^{-1}\dot{\epsilon}_i(0)\right) - Nn_0^{-1}\dot{\epsilon}_i(0) + (y_i(1) - y_i(0))$.

Recall that $n_1/N \to p$ and $n_0/N \to (1-p)$; then apply the Lindeberg–Lévy central limit theorem (Dur19, Theorem 3.4.1) to the two terms

$$\begin{aligned} &n_0^{-1/2}\left(\sum_{i\,:\,\tilde{z}_i=0}\varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right), \\ &n_1^{-1/2}\left(\sum_{i\,:\,\tilde{z}_i=1}\varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right) \end{aligned}$$

separately. Denote the limiting variance of the first term by s_0 and the limiting variance of the second term by s_1 and notice that these quantities do not depend on the particular choice of $\tilde{z}$. Then, regardless of $\tilde{z}$'s value we have that $N^{-1/2}\left(\sum_{i=1}^{N}\varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right)$ converges weakly to the random variable $(1-p)^{-1/2}A + p^{-1/2}B$ where $A \sim \mathcal{N}(0, s_0)$ independent of

$B \sim \mathcal{N}(0, s_1)$. By the independence of A and B we have that, for all $\tilde{z} \in \Omega_{CRE}$,

$$N^{-1/2}\left(\sum_{i=1}^{N} \varphi(\tilde{z}_i) - \mathbb{E}\left[\varphi(Z_i) \mid Z = \tilde{z}\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{s_0}{1-p} + \frac{s_1}{p}\right).$$

In other words, $N^{-1/2}\left(\sum_{i=1}^{N} \varphi(Z_i) - \overline{\tau}_{\text{PATE}}\right)$ conditional upon Z converges weakly to $\mathcal{N}\left(0, (1-p)^{-1}s_0 + p^{-1}s_1\right)$ almost surely with respect to randomness in Z. Consequently, by Lemma A.6, $N^{-1/2}\left(\sum_{i=1}^{N} \varphi(Z_i) - \overline{\tau}_{\text{PATE}}\right)$ converges weakly to $\mathcal{N}\left(0, (1-p)^{-1}s_0 + p^{-1}s_1\right)$ unconditionally on Z. Unwinding the definition of $\varphi(Z_i)$ establishes that $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}}\right)$ obeys a central limit theorem.

□

Theorem A.3. *Assume the conditions of Theorem A.2 hold. Further suppose that the original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are prediction unbiased. Then $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})$, $N^{1/2}(\hat{\tau}_{GBcal} - \overline{\tau}_{\text{PATE}})$, $N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{PATE}})$, and $N^{1/2}(\hat{\tau}_{unadj} - \overline{\tau}_{\text{PATE}})$ all obey central limit theorems and $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})$ has asymptotic variance no greater than any of the other three.*

Proof. The central limit theorem for $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})$ is proved in Theorem A.2. The proof for the central limit theorems of $N^{1/2}(\hat{\tau}_{GBcal} - \overline{\tau}_{\text{PATE}})$ follows a similar arc to that of Theorem A.2; the proof of the central limit theorem for $N^{1/2}(\hat{\tau}_{gOB} - \overline{\tau}_{\text{PATE}})$ also follows similar reasoning, but relies upon the assumption that the original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are prediction unbiased. The assumption that the original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are prediction unbiased is not needed for $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})$ and $N^{1/2}(\hat{\tau}_{GBcal} - \overline{\tau}_{\text{PATE}})$ since these two calibration procedures automatically confer prediction unbiasedness due to the first-order optimality conditions of linear regression. Finally, the central limit theorem for $N^{1/2}(\hat{\tau}_{unadj} - \overline{\tau}_{\text{PATE}})$ is a classical consequence of the Lyapunov central limit theorem.

By the law of total variance, for any measurable event $\mathcal{F}$, the variance of $N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})$

decomposes as

$$\mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})\right)=\mathbb{E}\left[\mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})\mid\mathcal{F}\right)\right]+\mathbb{V}\left(\mathbb{E}\left[N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})\mid\mathcal{F}\right]\right).$$

Taking $\mathcal{F}$ to be the event $\{(y_i(0),y_i(1),x_i)=(\mathbf{y}_i(0),\mathbf{y}_i(1),\mathbf{x}_i)\,i=1,\ldots,N\}$ yields the decomposition of the variance of $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})$ into the expected variance of $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})$ given a fixed finite population and the variance of the expectation of $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})$ given that fixed finite population. Since $\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}$ conditioned on $\mathcal{F}$ has mean on the order of $o_P(N^{-1/2})$ it follows that $\lim_{N\to\infty}\mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})\right)$ equals

$$\underbrace{\lim_{N\to\infty}\mathbb{E}\left[\mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})\mid\mathcal{F}\right)\right]}_{\text{Term 1}}+\underbrace{\lim_{N\to\infty}\mathbb{V}\left(N^{1/2}(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{PATE}})\right)}_{\text{Term 2}}. \tag{36}$$

Term 1 is the limiting expected finite population variance of $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})$ while Term 2 is the limiting variance of $N^{1/2}(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{PATE}})$. In other words, Term 2 is the asymptotic N-scaled variance of the SATE around the PATE. The same variance decomposition idea applies to $N^{1/2}(\hat{\tau}_{GBcal}-\overline{\tau}_{\text{PATE}})$, $N^{1/2}(\hat{\tau}_{gOB}-\overline{\tau}_{\text{PATE}})$, and $N^{1/2}(\hat{\tau}_{unadj}-\overline{\tau}_{\text{PATE}})$; for each different estimator the form of Term 1 adapts to the particular estimator at hand but Term 2 is exactly the same. Thus, the differences in asymptotic variance under the superpopulation model are controlled only by the differences in finite population variance. Consequently, the finite population analysis already performed implies that $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}})$ has asymptotic variance no greater than that of $N^{1/2}(\hat{\tau}_{GBcal}-\overline{\tau}_{\text{PATE}})$, $N^{1/2}(\hat{\tau}_{gOB}-\overline{\tau}_{\text{PATE}})$, and $N^{1/2}(\hat{\tau}_{unadj}-\overline{\tau}_{\text{PATE}})$. □

The decomposition of variance in (36) points to an important deficiency of $\hat{V}_{cal}$ of (35): while the variance estimator $\hat{V}_{cal}$ of (35) is guaranteed to be asymptotically conservative in the finite population model, it need not be valid if additional randomness is incorporated into the data generating process. Indeed, if there is randomness in the potential outcomes

themselves, $N\hat{V}_{cal}$ may converge in probability to a constant which is strictly smaller than the limiting variance of $N^{1/2}(\hat{\tau}_{cal} - N^{-1}\sum_{i=1}^{N}\mathbb{E}\left[y_i(1) - y_i(0)\right])$. Fortunately, (36) also suggests a simple rectification of this anti-conservativeness. By asymptotic linearity and the usual finite population decomposition of variance (DFM19), we have that

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}}) \mid \mathcal{F}\right)\right] &= \frac{\Sigma_{\dot{\epsilon}(1),\infty}}{p} + \frac{\Sigma_{\dot{\epsilon}(0),\infty}}{1-p} - \Sigma_{\delta,\infty}, \\
\Sigma_{\dot{\epsilon}(z),\infty} &= \lim_{N\to\infty}\frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left[\left(\dot{\epsilon}_i(z) - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\dot{\epsilon}_j(z)\right]\right)^2\right] \quad \text{for } z \in \{0,1\}, \\
\Sigma_{\delta,\infty} &= \lim_{N\to\infty}\frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left[\left((\dot{\epsilon}_i(1) - \dot{\epsilon}_i(0)) - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\dot{\epsilon}_j(1) - \dot{\epsilon}_j(0)\right]\right)^2\right].
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{V}\left(N^{1/2}(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{PATE}})\right) &= \Sigma_{\tau,\infty}, \\
\Sigma_{\tau,\infty} &= \lim_{N\to\infty}\frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left[\left((y_i(1) - y_i(0)) - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[y_j(1) - y_j(0)\right]\right)^2\right].
\end{aligned}$$

Combining the two previous observations with (36) yields that

$$\lim_{N\to\infty}\mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})\right) = \frac{\Sigma_{\dot{\epsilon}(1),\infty}}{p} + \frac{\Sigma_{\dot{\epsilon}(0),\infty}}{1-p} - \Sigma_{\delta,\infty} + \Sigma_{\tau,\infty}.$$

By the classical partitioning of variance in linear regression the total sum of squares is the sum of the residual sum of squares and the explained sum of squares; in our context this

means that

$$\begin{aligned}\Sigma_{\tau,\infty} &= \Sigma_{\delta,\infty} + \Sigma_{fitted,\infty},\\ \Sigma_{fitted,\infty} &= \lim_{N\to\infty} \frac{1}{N-1}\sum_{i=1}^{N} \mathbb{E}\Bigg[\Bigg((\dot{\mu}_{OLS,1}(x_i) - \dot{\mu}_{OLS,0}(x_i)) - \\ &\qquad \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}\left[\dot{\mu}_{OLS,1}(x_j) - \dot{\mu}_{OLS,0}(x_j)\right] \Bigg)^2\Bigg].\end{aligned}$$

In total, we have that

$$\lim_{N\to\infty} \mathbb{V}\left(N^{1/2}(\hat{\tau}_{cal} - \overline{\tau}_{\text{PATE}})\right) = \frac{\Sigma_{\dot{\epsilon}(1),\infty}}{p} + \frac{\Sigma_{\dot{\epsilon}(0),\infty}}{1-p} + \Sigma_{fitted,\infty}.$$

Consequently, to adapt the variance estimator of (35) to superpopulation inference, one must incorporate the variance of the predicted treatment effects; this forms the superpopulation variance estimator

$$\begin{aligned}\hat{V}_{cal,sup} = &\frac{\hat{\Sigma}_{1,cal}}{n_1} + \frac{\hat{\Sigma}_{0,cal}}{n_0} + \\ &\frac{1}{N-1}\sum_{i=1}^{N}\left((\hat{\mu}_{OLS,1}(x_i) - \hat{\mu}_{OLS,0}(x_i)) - \frac{1}{N}\sum_{j=1}^{N}(\hat{\mu}_{OLS,1}(x_j) - \hat{\mu}_{OLS,0}(x_j))\right)^2. \quad (37)\end{aligned}$$

Notice that the decomposition of variance into $p^{-1}\Sigma_{\dot{\epsilon}(1),\infty} + (1-p)^{-1}\Sigma_{\dot{\epsilon}(0),\infty} + \Sigma_{fitted,\infty}$ relies upon the orthogonality of residuals $y_i(z) - \dot{\mu}_z(x_i)$ and predicted values $\dot{\mu}_z(x_i)$ due to the first order optimality condition of ordinary least squares linear regression, so such a decomposition is not generally applicable to uncalibrated estimators; this provides another attractive property for calibrated estimators.[6]

[6]For a generic superpopulation variance estimator in the context of uncalibrated imputation estimators see (Rot20, Section 3.3).

2.16.3 Linear Calibration for the Conditional Average Treatment Effect

We start by presenting several highly general regularity conditions; further on we present an example of a generative model which satisfies the required regularity conditions. Our regularity conditions in the fixed-covariate model are presented with respect to conditioning upon the potential outcomes in addition to the already implicitly determined covariates. For each population of size N with deterministic covariates $\{x_i\}_{i=1}^N$, consider conditioning upon some realization of the potential outcomes

$$\{(y_i(0), y_i(1)) = (\mathbf{y}_i(0), \mathbf{y}_i(1))\, i = 1, \ldots, N\}\,. \tag{38}$$

Assumption 10 (Fixed-Covariate Limiting Means and Variances)**.** *For all conditioning events of the form* (38) *except for a set of measure zero under the fixed covariate model we require that for* $z = 0, 1$ *there exists a limiting value* $\bar{y}(z)_\infty$ *such that* $\lim_{N\to\infty} N^{-1} \sum_{i=1}^N y_i(z) = \bar{y}(z)_\infty$*. Likewise, for almost all conditioning events of the form* (38) *there exists a common limiting positive semidefinite matrix* Σ *such that*

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^N \left(\begin{bmatrix} y_i(0) \\ y_i(1) \\ \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} - N^{-1} \sum_{j=1}^N \begin{bmatrix} y_j(0) \\ y_j(1) \\ \dot{\mu}_0(x_j) \\ \dot{\mu}_1(x_j) \end{bmatrix} \right) \otimes \left(\begin{bmatrix} y_i(0) \\ y_i(1) \\ \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} - N^{-1} \sum_{j=1}^N \begin{bmatrix} y_j(0) \\ y_j(1) \\ \dot{\mu}_0(x_j) \\ \dot{\mu}_1(x_j) \end{bmatrix} \right) = \Sigma. \tag{39}$$

Assumption 11 (Fixed-Covariate Bounded Fourth Moments)**.** *For all conditioning events*

of the form (38) *except for a set of measure zero under the fixed covariate model we require that for* $z = 0, 1$ *there exists some* $C < \infty$ *such that for all* $N \in \mathbb{N}$, $N^{-1} \sum_{i=1}^{N} y_i(z)^4 < C$ *and* $N^{-1} \sum_{i=1}^{N} \dot{\mu}_z(x_i)^4 < C$.

In the fixed-covariate case, the definition of $\dot{\beta}$ given in Lemma A.3 requires a minor change to

$$\dot{\beta} = (\dot{\beta}_0, \dot{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^{N} \mathbb{E}\left[\left\{ y_i(1) - \left(\beta_0 + \beta_1^T \begin{bmatrix} \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} \right) \right\}^2 \,\middle|\, x_i \right].$$

Assumptions 10 and 11 facilitate consistency of ordinary least squares coefficients in the fixed covariate model.

Lemma A.8 (Lindeberg Condition)**.** *Under Assumptions 10 and 11, the potential outcomes* $(y_i(0), y_i(1))$ *given* x_i *jointly satisfy the conditions of Lindeberg's central limit theorem.*

Proof. Define $s_N^2(z) = \sum_{i=1}^{N} \mathbb{V}\left(y_i(z) \mid x_i\right)$ where $\mathbb{V}\left(y_i(z) \mid x_i\right)$ denotes the variance of $y_i(z)$ given the covariates x_i. We will show that Lyapounov's condition (LR05, Equation 11.12) holds at $\delta = 2$ for the potential outcomes; formally, for $z \in \{0, 1\}$ and $\delta = 2$

$$\lim_{N \to \infty} \frac{1}{s_N^{2+\delta}(z)} \sum_{i=1}^{N} \mathbb{E}\left[\left| y_i(z) \right|^{2+\delta} \mid x_i \right] = 0. \tag{40}$$

Rewrite (40) as

$$\lim_{N \to \infty} \underbrace{\frac{N}{s_N^{2+\delta}(z)}}_{\text{Term 1}} \cdot \underbrace{\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\left| y_i(z) \right|^{2+\delta} \mid x_i \right]}_{\text{Term 2}}.$$

Term 2 is bounded above by C for all N by Assumption 11 and Kolmogorov's strong law of large numbers for non-identically distributed sequences (Fel68, Section 10.7), so it suffices to show that Term 1 vanishes as $N \to \infty$. By Assumption 10 and Kolmogorov's strong

law, $N^{-1}\sum_{i=1}^{N}\mathbb{V}\left(y_i \mid x_i\right)$ limits to a positive constant which we denote $\Sigma_{y(z)}$.[7] Consequently, $Ns_N^{-2} \to \Sigma_{y(z)}^{-1} > 0$ and $s_N^2 = \Theta(N)$. From this, it is immediate that $Ns_N^{-(2+\delta)} \to 0$ as $N \to \infty$ for $\delta = 2$. In total, this establishes (40). Since (40) is sufficient for the Lindeberg condition (LR05, Page 427), the result follows. □

Remark 5. Since the residuals $(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))$ defined by $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(x_i)$ are deterministic translations of the potential outcomes $(y_i(0), y_i(1))$ in the fixed-covariate model, Lemma A.8 immediately implies that the residuals $(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))$ jointly satisfy the conditions of Lindeberg's central limit theorem.

The stability assumption is also taken in accordance with the conditioning events of (38).

Assumption 12 (Stability)**.** *For all conditioning events of the form* (38) *except for a set of measure zero under the fixed covariate model there exists a deterministic sequence of functions* $\{\dot{\mu}_1^{(N)}\}_{N\in\mathbb{N}}$ *such that*

$$\left(\frac{1}{N}\sum_{i=1}^{N}||\dot{\mu}_1^{(N)}(x_i) - \hat{\mu}_1(x_i)||^2\right)^{1/2} = o_P(1). \tag{41}$$

We assume that an analogous sequence, $\{\dot{\mu}_0^{(N)}\}_{N\in\mathbb{N}}$*, exists for* $\hat{\mu}_0$*.*

In Assumption 12 the randomness on both sides of (41) is only with respect to Z.

Remark 6. The regularity conditions Assumptions 10 and 11 do not prescribe a particular generative model; they are working-level mathematical ingredients in our subsequent proofs. In order to complete the picture, we detail a conventional generative model which satisfies Assumptions 10 and 11.

For each finite population the N units have covariates $\{x_i\}_{i=1}^N$ and the potential outcomes of unit i are independent of all units j for $j \neq i$. The pair of potential outcomes $(y_i(0), y_i(1))$

[7] The use of Kolmogorov's strong law can be replaced by the bounded convergence theorem for both the arguments pertaining to Term 1 and Term 2; Assumption 11 establishes the required bounds.

are distributed according the the conditional distribution P_{x_i}. Consequently, for any two individuals who exactly match on their covariates, their potential outcome pairs are independent and identically distributed. Formally, $x_i = x_j$ implies that $(y_i(0), y_i(1))$ is equal in distribution to $(y_j(0), y_j(1))$. Let expectations under the distribution P_{x_i} be denoted as E_{x_i}; similarly, denote variances by var_{x_i}.

Assumption 10 codifies the need for a strong law of large numbers for the means and variances of the realized populations under the fixed covariate model. Assumption 10 is highly general, but can be implied by moment conditions on the potential outcomes themselves. In particular suppose that $\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \mathrm{E}_{x_i}[y_i(z)]$ and $\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \mathrm{var}_{x_i}(y_i(z))$ exist and are finite. This is sufficient to guarantee the first condition of Assumption 10; the proof proceeds by application of Kolmogorov's strong law of large numbers for non-identically distributed sequences. Similar reasoning establishes that if there exists a limiting positive semidefinite matrix Σ such that

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} \mathrm{E}_{x_i} \left[\left(\begin{bmatrix} y_i(0) \\ y_i(1) \\ \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} - N^{-1} \sum_{j=1}^{N} \mathrm{E}_{x_j} \left[\begin{bmatrix} y_j(0) \\ y_j(1) \\ \dot{\mu}_0(x_j) \\ \dot{\mu}_1(x_j) \end{bmatrix} \right] \right) \right.$$
$$\left. \otimes \left(\begin{bmatrix} y_i(0) \\ y_i(1) \\ \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} - N^{-1} \sum_{j=1}^{N} \mathrm{E}_{x_j} \left[\begin{bmatrix} y_j(0) \\ y_j(1) \\ \dot{\mu}_0(x_j) \\ \dot{\mu}_1(x_j) \end{bmatrix} \right] \right) \right] = \Sigma$$

and sufficient control of higher order moments (e.g., coordinate-wise fourth moments) is assumed then the second condition of Assumption 10 again follows by Kolmogorov's law of

large numbers applied now to the entries of the matrix in (39). The moment condition

$$N^{-1}\sum_{i=1}^{N}\mathrm{E}_{x_i}\left[y_i(z)^4\right]\to c_z \text{ for } z\in\{0,1\}$$

for a constant c_z and the requirement that Kolmogorov's condition (Fel68, Eqn. 7.2) holds for the random variables $y_i(z)$ establishes Assumption 11 by Kolmogorov's law of large numbers and simultaneously serves in establishing the second condition of Assumption 10. Since the covariates are non-stochastic we make the usual assumption that $N^{-1}\sum_{i=1}^{N}\dot{\mu}_z(x_i)^4<C$ for $z\in\{0,1\}$; this exactly mirrors our earlier finite population analysis.

Lemma A.9 (Asymptotically Linear Expansions around the SATE)**.** *Under Assumptions A.6, 2, and 12 the random variable* $N^{-1}\sum_{i=1}^{N}\left(\hat{\mu}_z(x_i)-y_i(z)\right)$ *is asymptotically linear in the sense that, for* $\dot{\epsilon}_i(z)=y_i(z)-\dot{\mu}_z(x_i)$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_z(x_i)-y_i(z)\right)=\frac{1}{n_z}\sum_{i\,:\,Z_i=z}\dot{\epsilon}_i(z)+o_p(N^{-1/2}).$$

The proof of Lemma A.9 mostly mirrors that of Lemma A.7; the main working ingredient is the vanishing error process argument facilitated by Assumption 2.[8]

For the analysis of central limit theorems under the fixed-covariate model it is mathematically convenient to adopt the probabilistic joint-model-design framework of (RBK05). Consider a probability space $(\Phi,\mathscr{F},P)$ from which we form a population of N individuals with potential outcome $y_i(z)=\mathcal{Y}_z(\omega_i)$ for $z\in\{0,1\}$ and covariates $x_i=\mathcal{X}(\omega_i)$ for $\mathcal{Y}_z$ and $\mathcal{X}$ measurable functions of $\omega_i\in\Phi$. Let $\mathcal{F}_{cov}=\{\omega\in\Phi\ :\ \mathcal{X}(\omega_i)=\mathbf{x}_i \text{ for } i=1,\ldots,N\}$; $\mathcal{F}_{cov}$ is the event that the covariates of the N individuals are given by the deterministic values $\{\mathbf{x}_i\}_{i=1}^{N}$. Let $P_{\mathcal{F}_{cov}}$ be the conditional probability measure derived from P conditioned on

[8] In the fixed covariate model, the error process $\mathcal{G}_{N,z}(\dot{\mu}_z)-\mathcal{G}_{N,z}(\hat{\mu}_z)$ inherits randomness only from stochasticity in the outcomes and treatment allocation while the covariates x_i are viewed as deterministic vectors.

the event $\mathcal{F}_{cov}$. In the case that $\mathcal{F}_{cov}$ is an event of P-measure zero, we tacitly assume that there exists a well-defined regular conditional probability measure and take $P_{\mathcal{F}_{cov}}$ to be this conditional; see (CT97, Section 7.2) for more details on this technical issue. Inferences under the fixed-covariate model take $(\Omega, \mathscr{F}, P_{\mathcal{F}_{cov}})$ to generate the outcomes $y_i(z) = \mathcal{Y}_z(\omega_i)$ for $z \in \{0, 1\}$ and implicitly constrain the covariates $x_i = \mathcal{X}(\omega_i) = \mathbf{x}_i$ for $i = 1, \ldots, N$.

Theorem A.4. *Under the fixed-covariate model, subject to Assumptions A.6, 2, 10, 11, and 12,* $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right)$ *obeys a central limit theorem.*

Proof. We start out with the simple observation that

$$N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right) = N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right) + N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right).$$

Leveraging the asymptotic linearity of Lemma A.9 the first term on the right-hand-side can be replaced with the difference in means of the residuals $\dot{\epsilon}_i(1)$ and $\dot{\epsilon}_i(0)$ plus an $o_P(1)$ error term:

$$N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right) = N^{1/2}\left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right) + o_P(1) + N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right).$$

The $o_P(1)$ error term has no impact on the asymptotic distributional behaviour of $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right)$. Thus, to show a central limit theorem for $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right)$ it suffices to show that

1. Conditionally upon the potential outcomes, the term

$$N^{1/2}\left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right)$$

converges weakly in probability to a fixed Gaussian distribution and term

$$N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$$

obeys a central limit theorem.

2. The terms

$$N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right) + o_P(1) = N^{1/2}\left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right)$$

and $N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$ are asymptotically independent in the sense that their limiting joint distribution is the product of the two limiting marginal distributions.

We tackle 1 first. By Lemma A.8 and the Lindeberg central limit theorem (LR05, Theorem 11.2.5) it follows that $N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$ converges in distribution to a Gaussian distribution; denote this limiting distribution as $\mathcal{N}\left(0, s_m\right)$.

Next, we show that $N^{1/2}\left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right)$ converges weakly in probability to a fixed Gaussian distribution.

Under Assumption 12 and the assumption that $N^{-1}\sum_{i=1}^{N}\left(\dot{\mu}_z(x_i) - y_i(z)\right)^2 = o(N)$ as a numeric sequence for almost all conditioning events of the potential outcomes, by Lemma 3 in the appendix of (GB21) we can, without loss of generality, stipulate that $N^{-1}\sum_{i=1}^{N}\dot{\epsilon}_i(z) = 0$ for $z \in \{0, 1\}$ almost surely with respect to the conditioning (38). Under Assumptions 10 and 11 the finite population analysis provided in Theorem 1 shows that

$$N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i N n_1^{-1}\dot{\epsilon}_i(1) - \frac{1}{N}\sum_{i=1}^{N}(1 - Z_i) N n_0^{-1}\dot{\epsilon}_i(0)\right)$$

converges weakly to a centered Gaussian distribution with variance given by the limit of σ_N^2 defined in (28). This limit exists by Assumption 10 and is common to all condition-

ing events of the form (38) up to a set of measure zero; we denote it by s_d. Consequently, $N^{1/2}\left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1)-\frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right)$ converges weakly in probability to a random variable with distribution $\mathcal{N}(0,s_d)$.

Finally, we turn to 2. By Theorem 5.1 (iii) of (RBK05) it follows that the random vector $\left(N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}),N^{1/2}(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{CATE}})\right)$ converges in distribution to $(\mathcal{C},\mathcal{D})\sim\mathcal{N}(0,s_d)\otimes\mathcal{N}(0,s_m)$.[9]

By the continuous mapping theorem (vdV98, Theorem 18.11),

$$N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}})+N^{1/2}(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{CATE}})$$

converges in distribution to $\mathcal{C}+\mathcal{D}$. Since the sum of independent Gaussian random variables is itself Gaussian we have that $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}})+N^{1/2}(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{CATE}})$ converges in distribution to $\mathcal{N}(0,s_d+s_m)$. In turn, this implies that $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{CATE}})$ converges in distribution to $\mathcal{N}(0,s_d+s_m)$. □

Theorem A.5. *Assume the conditions of Theorem A.4 hold. Further suppose that the original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are prediction unbiased. Then $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{CATE}})$, $N^{1/2}(\hat{\tau}_{GBcal}-\overline{\tau}_{\text{CATE}})$, $N^{1/2}(\hat{\tau}_{gOB}-\overline{\tau}_{\text{CATE}})$, and $N^{1/2}(\hat{\tau}_{unadj}-\overline{\tau}_{\text{CATE}})$ all obey central limit theorems and $N^{1/2}(\hat{\tau}_{cal}-\overline{\tau}_{\text{CATE}})$ has asymptotic variance no greater than any of the other three.*

Proof. The proof thematically mirrors that of Theorem A.3. The first three central limit theorems are justified by Theorem A.4 and analogous variants for $N^{1/2}(\hat{\tau}_{GBcal}-\overline{\tau}_{\text{CATE}})$ and $N^{1/2}(\hat{\tau}_{gOB}-\overline{\tau}_{\text{CATE}})$. The fourth central limit theorem is justified by Lemma A.8 and the Lindeberg central limit theorem (LR05, Theorem 11.2.5). □

Variance estimation and the construction of confidence intervals proceeds via the variance estimator of (35) under analogous reasoning.

[9]The original work of (RBK05) focuses on survey-sampling; however, nothing of their result Theorem 5.1 (iii) relies upon the survey-sampling framework of having only a single potential outcome, so we apply their result to the causal inference context of multiple potential outcomes.

2.17 Further Simulations

2.17.1 An Example with Logistic Regression

In Section 2.6, we provided a simulation study to demonstrate the practical benefits of our calibration procedure while simultaneously showing the risks of uncalibrated estimators. We used an example based upon Poisson regression; to further highlight the concerns of using uncalibrated estimates we now provide a simulation using logistic regression.

The sth of S data sets contains N individuals upon whom an experimenter performs a completely randomized experiment with $n_1 = \lceil pN \rceil$ treated units. In our simulations $p = 0.8$. Each unit has a scalar covariate x_i, generated as independent and identically distributed draws from a Uniform random variable on $[-8, 8]$. We then generate the potential outcomes under treatment and control for each individual independently as $y_i(1) \sim Bern\{f(x_i)\}$ and $y_i(0) \sim Bern\{-0.4 * (f(x_i) - 1)\}$, where $Bern(c)$ is a Bernoulli distribution with probability of success c. We take

$$f(x) = \frac{\exp(-3 + 2x)}{1 + \exp(-3 + 2x)}.$$

Consequently, the logistic regression model is correctly specified for the potential outcomes under treatment, but incorrectly specified for those under control.

For each data set, these values are left fixed while the remaining randomness in the simulation arises only from treatment allocation. An experimenter observes only the count data $y_i(Z_i)$ and continuous covariates x_i for each unit. Using the observed responses after each randomization of treatment allocation, we estimate the prediction functions $\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$ via separate logistic regressions of $y_i(Z_i)$ on x_i in the subgroups where $Z_i = 0$ and $Z_i = 1$, respectively. We form the difference-in-means estimator $\hat{\tau}_{unadj}$, generalized Oaxaca-Blinder estimator $\hat{\tau}_{gOB}$, the singly-calibrated estimator of (GB21, Equation 8) $\hat{\tau}_{GBcal}$, and our linearly-calibrated estimator $\hat{\tau}_{cal}$.

Table 2.2 compares the averages (over $s = 1, ..., S$) of the ratios of the variances for

	$\hat{\text{var}}(\hat{\tau}_{gOB})/\hat{\text{var}}(\hat{\tau}_{unadj})$	$\hat{\text{var}}(\hat{\tau}_{GBcal})/\hat{\text{var}}(\hat{\tau}_{unadj})$	$\hat{\text{var}}(\hat{\tau}_{cal})/\hat{\text{var}}(\hat{\tau}_{unadj})$
$N = 200$	1.077	1.076	1.031
$N = 500$	1.056	1.054	0.993
$N = 1000$	1.050	1.047	0.981
$N = 10000$	1.043	1.041	0.970

Table 2.2: Ratios of Monte Carlo variances for $\hat{\tau}_{cal}$, $\hat{\tau}_{GBcal}$, and $\hat{\tau}_{gOB}$ to the difference in means estimator $\hat{\tau}_{unadj}$ for various experiment sizes N. Each variance is based upon $B = 1000$ simulated treatment allocations for a given set of potential outcomes and covariates. Results are averaged over $S = 1000$ simulated data sets.

the adjusted estimators to the unadjusted estimator when setting $S = 1000$, $B = 1000$, and varying N. Qualitatively, the results of Table 2.2 mirror those of the Poisson regression simulation in Section 2.6: uncalibrated generalized Oaxaca-Blinder estimators can fare worse than the simple difference in means and the singly-calibrated estimator $\hat{\tau}_{GBcal}$ fails to correct this issue; however, our calibration procedure asymptotically improves upon $\hat{\tau}_{unadj}$ while leveraging the desired nonlinear model.

2.17.2 Poisson Regression Calibration in Alternative Models

To highlight the application of calibration in superpopulation and fixed-covariate models, we recreate the Poisson regression example from Section 2.6 at all three levels of inference. For simulations in the superpopulation, new covariates and potential outcomes are redrawn in each of the SB simulated data sets. For simulation in the fixed-covariate model new covariates are constructed in the sth simulation, but are held fixed – while potential outcomes are redrawn conditional upon these covariates – for each randomization of treatment allocation $1, \ldots, B$ in the sth simulation. Each simulation is conducted with $N = 10000$. Variances are reported for appropriately centered versions $\hat{\tau}_{cal}$, $\hat{\tau}_{GBcal}$, and $\hat{\tau}_{gOB}$. As an example, for the sth simulation in the fixed-covariate case we compute the ratio of the variance of $\hat{\tau}_{cal} - \overline{\tau}^{(s)}_{\text{CATE}}$ and the variance of $\hat{\tau}_{unadj} - \overline{\tau}^{(s)}_{\text{CATE}}$ where $\overline{\tau}^{(s)}_{\text{CATE}}$ denotes the conditional average treatment effect in the sth simulated population. Table 2.3 summarizes our results. Even in the fixed-covariate

	$\hat{\text{var}}(\hat{\tau}_{gOB})/\hat{\text{var}}(\hat{\tau}_{unadj})$	$\hat{\text{var}}(\hat{\tau}_{GBcal})/\hat{\text{var}}(\hat{\tau}_{unadj})$	$\hat{\text{var}}(\hat{\tau}_{cal})/\hat{\text{var}}(\hat{\tau}_{unadj})$
SATE	1.660	1.657	0.654
CATE	1.549	1.547	0.711
PATE	1.114	1.114	0.941

Table 2.3: Ratios of Monte Carlo variances for $\hat{\tau}_{cal}$, $\hat{\tau}_{GBcal}$, and $\hat{\tau}_{gOB}$ to the difference in means estimator $\hat{\tau}_{unadj}$ under different generative models. Each variance is based upon $B = 1000$ simulated treatment allocations; results are averaged over $S = 1000$ simulated data sets.

and superpopulation models, uncalibrated Oaxaca-Blinder estimators may suffer from inflated asymptotic variances relative to the unadjusted difference in means and the single calibration of $\hat{\tau}_{GBcal}$ fails to correct the issue. Fortunately, our linear calibration procedure succeeds in all three models, as evidenced by the third column of Table 2.3. The impact of calibration is most noticeable in the finite population framework but is nonetheless profound in all three models.

In the third column of Table 2.3 $\hat{\text{var}}(\hat{\tau}_{cal})/\hat{\text{var}}(\hat{\tau}_{unadj})$ increases as one goes from inference for the SATE to the CATE and finally to the PATE. This trend is a fundamental reflection of asymptotic variances in the three models under consideration. By the law of total variance, for any measurable event $\mathcal{F}$, the variance of $\hat{\tau}_{unadj}$ decomposes as

$$\mathbb{V}\left(N^{1/2}\hat{\tau}_{unadj}\right) = \mathbb{E}\left[\mathbb{V}\left(N^{1/2}\hat{\tau}_{unadj} \mid \mathcal{F}\right)\right] + N\mathbb{V}\left(\mathbb{E}\left[\hat{\tau}_{unadj} \mid \mathcal{F}\right]\right).$$

Taking $\mathcal{F}$ to be the event $\{(y_i(0), y_i(1), x_i) = (\mathbf{y}_i(0), \mathbf{y}_i(1), \mathbf{x}_i)\, i = 1, \ldots, N\}$ yields the decomposition of the variance of $\hat{\tau}_{unadj}$ into the expected variance of $\hat{\tau}_{unadj}$ given a fixed finite population and the variance of the expectation of $\hat{\tau}_{unadj}$ given that fixed finite population. Since the difference in means is unbiased for the sample average treatment effect given $\mathcal{F}$ it follows that the second term is just the variance of the sample average treatment effect, $\mathbb{V}\left(\overline{\tau}_{\text{SATE}}\right)$. This decomposition is valid regardless of the underlying model of the data: superpopulation, fixed-covariate, or finite population. For the sake of explanation, we discuss the difference between the SATE row and the PATE row, but the same reasoning applies

to the differences between the SATE and CATE rows and the CATE and PATE rows of Table 2.3.

In the finite population model there is no variance of the sample average treatment effect whatsoever since it depends only on fixed values; in other words, for the top row of $N\mathbb{V}\left(\mathbb{E}\left[\hat{\tau}_{unadj} \mid \mathcal{F}\right]\right) = 0$. In contrast with the finite population case, under a superpopulation model the sample average treatment effect may vary; in all but the most degenerate cases $N\mathbb{V}\left(\overline{\tau}_{\text{SATE}}\right) > 0$. Furthermore, under the finite population conditions of (Lin13) and standard fourth-moment regularity conditions in the superpopulation model $\mathbb{E}\left[\mathbb{V}\left(N^{1/2}\hat{\tau}_{unadj} \mid \mathcal{F}\right)\right]$ has the same limit in the finite population model and the superpopulation model.

The same variance decomposition can be applied to the estimators $\hat{\tau}_{gOB}$, $\hat{\tau}_{GBcal}$, and $\hat{\tau}_{cal}$. Without loss of generality, we discuss the case of $\hat{\tau}_{cal}$. Since $\hat{\tau}_{cal}$ is asymptotically unbiased in all three models, the term $N\mathbb{V}\left(\mathbb{E}\left[\hat{\tau}_{cal} \mid \mathcal{F}\right]\right)$ limits to the N-scaled variance of the sample average treatment effect in all three models. As before, under mild regularity conditions $\mathbb{E}\left[\mathbb{V}\left(N^{1/2}\hat{\tau}_{cal} \mid \mathcal{F}\right)\right]$ has the same limit in the finite population model and the superpopulation model. Thus, the difference in asymptotic variances between $\hat{\tau}_{cal}$ and $\hat{\tau}_{unadj}$ is driven by the limiting difference between $\mathbb{E}\left[\mathbb{V}\left(N^{1/2}\hat{\tau}_{cal} \mid \mathcal{F}\right)\right]$ and $\mathbb{E}\left[\mathbb{V}\left(N^{1/2}\hat{\tau}_{unadj} \mid \mathcal{F}\right)\right]$. By the reasoning above, this limiting difference is the same in both the finite population model and in a superpopulation model. Taking the results for large N as reflective of their asymptotic behavior the difference between the top-right and bottom-right elements of Table 2.3 can be explained as: the difference between the numerator and denominator remains the same while the magnitude of both the numerator and the denominator are larger in the PATE row than in the SATE row. Analogous reasoning applies to the SATE versus CATE rows and the CATE versus PATE rows and across the other columns as well.

2.18 A Case Study On Tumor Recurrence Of Bladder Cancers

The Veterans Administration Cooperative Urological Research Group (VACURG) conducted a completely randomized clinical trial to examine the effectiveness of treatment against recurrence of bladder cancers. Each patient enrolled in the study had superficial bladder tumors are the start of the study; the tumors were removed transurethrally before the patients were assigned to one of three treatment conditions: placebo pills, pyridoxine (vitamin B_6) pills, or periodic treatment with thiotepa (a chemotherapeutic agent). The patients returned for follow-up visits and the existence of recurrent tumors observed in these follow-ups was tabulated; although – at times – more than one tumor was observed during a follow-up appointment the number of such tumors is not the primary object of study, only their presence or not is recorded as a binary outcome in each follow-up. After a recurrent tumor was observed, it was removed and the treatment regimen assigned to that individual was continued. For further details see (AH85, Chapter 45). Our analysis focuses upon the placebo group, with 47 individuals, and the thiotepa treatment group, with 38 individuals.

The primary outcome of the study is the count of the number of recurrences, so Poisson regression is a natural adjustment model. Covariate information collected at the start of the experiment includes the initial number of tumors and the diameter of the largest of these. The number of months over which the patient attended follow-up appointments was recorded as well as the survival status of the patient at the conclusion of the study. We control for the log-number of follow-up months, the number of initial tumors, and the diameter of the largest initial tumor. We compare the unadjusted difference in means, the uncalibrated generalized Oaxaca-Blinder estimator of (GB21), the singly-calibrated estimator of (GB21), and our calibrated estimator. Table 2.4 displays the point estimate of treatment effect and the corresponding estimated variance. The variances displayed along the right-hand column

	Point Estimate	Variance
$\hat{\tau}_{unadj}$	-0.667	0.190
$\hat{\tau}_{gOB}$	-0.775	0.123
$\hat{\tau}_{GBcal}$	-0.784	0.122
$\hat{\tau}_{cal}$	-0.778	0.120

Table 2.4: Point estimates and estimated variances of $\hat{\tau}_{unadj}$, $\hat{\tau}_{gOB}$, and $\hat{\tau}_{GBcal}$, and $\hat{\tau}_{cal}$ on the VACURG bladder tumor recurrence data set.

of Table 2.4 demonstrate the substantial benefit of controlling for features. Only $\hat{\tau}_{cal}$ is guaranteed to be non-inferior to $\hat{\tau}_{unadj}$; the performance of $\hat{\tau}_{gOB}$ and $\hat{\tau}_{GBcal}$ is not generally guaranteed. Moreover, $\hat{\tau}_{cal}$ never has asymptotic variance which exceeds that of $\hat{\tau}_{gOB}$ and $\hat{\tau}_{GBcal}$; this is observed even in this sample.

2.19 Calibration And Semiparametric Efficiency

In superpopulation models, a great deal of regression adjustment literature has focused upon semiparametric efficiency of estimators. Below we include a brief survey of some of this literature and demonstrate the relationship between semiparametric efficient estimators and the calibration procedure.

(Hah98) takes a superpopulation approach to inference; his formulation aligns with the superpoplation framework of (RBK05). Units of the population are N independent and identically distributed tuples $(y_i(0), y_i(1), x_i)$ and the object of inference is the population average treatment effect $\mathbb{E}[y_1(1) - y_1(0)]$. We present his main semiparametric efficiency bound adapted to the context of completely randomized experiments below; it can be envisioned in the same light as the Cramér-Rao bound as it provides a lower bound on the asymptotic variance of any regular estimator of $\mathbb{E}[y_1(1) - y_1(0)]$.

Theorem A.6 ((Hah98), Theorem 1)**.** *In a completely randomized experiment, the asymp-*

totic variance of any regular estimator sequence for $\mathbb{E}\left[y_1(1) - y_1(0)\right]$ *is bounded below by*

$$\mathbb{E}\left[\frac{\mathbb{V}\left(y_i(1) \mid x_i\right)}{p} + \frac{\mathbb{V}\left(y_i(0) \mid x_i\right)}{1-p} + \left(\mathbb{E}\left[y_1(1) - y_1(0) \mid x_i\right] - \mathbb{E}\left[y_1(1) - y_1(0)\right]\right)^2\right]. \tag{42}$$

Any sequence of regular estimators for $\mathbb{E}\left[y_1(1) - y_1(0)\right]$ which achieves an asymptotic variance of (42) in the limit is said to be (asymptotically) *semiparametric efficient*. Write the conditional expectation of the outcomes given the covariates as

$$\mu_z^{true}(x) = \mathbb{E}\left[y_i(z) \mid x_i = x\right] \quad \text{for } z \in \{0, 1\}.$$

Citing a result by (Hah98), (Rot20) remarks that any semiparametric efficient regular estimator of $\mathbb{E}\left[y_1(1) - y_1(0)\right]$ is necessarily of the form $N^{-1}\sum_{i=1}^{N}\psi_i(\mu_0^{true}, \mu_1^{true}) + o_p(N^{-1/2})$ where

$$\psi_i(\mu_0^{true}, \mu_1^{true}) := \left(\mu_1^{true}(x_i) - \mu_0^{true}(x_i)\right) + \frac{Z_i\left(y_i(Z_i) - \mu_1^{true}(x_i)\right)}{(n_1/N)} - \frac{(1-Z_i)\left(y_i(Z_i) - \mu_0^{true}(x_i)\right)}{(n_0/N)}. \tag{43}$$

We show that any estimator of the form $N^{-1}\sum_{i=1}^{N}\psi_i(\hat{\mu}_0, \hat{\mu}_1) + o_p(N^{-1/2})$ obtains non-inferiority after calibration; even though $N^{-1}\sum_{i=1}^{N}\psi_i(\hat{\mu}_0, \hat{\mu}_1) + o_p(N^{-1/2})$ originally could have been asymptotically inferior to the unadjusted difference in means. Consequently, any estimator which has hope of semiparametric efficiency can be made non-inferior to the difference in means via our calibration procedure. If it was the case that the original estimator sequence happened to be semiparametric efficient, then the asymptotic variance of the calibrated estimator will coincide with the semiparametric efficiency bound; however for regular estimators which do not achieve the semiparametric efficiency bound calibration automatically yields asymptotic non-inferiority to the difference in means. We formalize this statement below.

(Rot20) considers estimators of the form $\hat{\tau}_{rothe} = N^{-1} \sum_{i=1}^{N} \psi_i(\hat{\mu}_0, \hat{\mu}_1)$ where $\hat{\mu}_0$ and $\hat{\mu}_1$ attempt to estimate the true conditional expectation functions μ_0^{true} and μ_1^{true}, respectively. Consider calibrating $\hat{\tau}_{rothe}$ by defining $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ according to the usual linear calibration formula and forming

$$\hat{\tau}_{rothe,cal} = N^{-1} \sum_{i=1}^{N} \psi_i(\hat{\mu}_{OLS,0}, \hat{\mu}_{OLS,1}).$$

Theorem A.7. *Subject to Assumptions 1, 2, 8, and 9:*

- $\hat{\tau}_{rothe,cal}$ *is asymptotically linear in the sense that*

$$N^{1/2} \left(\hat{\tau}_{rothe,cal} - \overline{\tau}_{\text{PATE}} \right) = N^{-1/2} \sum_{i=1}^{N} \left(\psi_i(\dot{\mu}_{OLS,0}, \dot{\mu}_{OLS,1}) - \overline{\tau}_{\text{PATE}} \right) + o_P(1). \tag{44}$$

- *The asymptotic variance of* $N^{1/2} \left(\hat{\tau}_{rothe,cal} - \overline{\tau}_{\text{PATE}} \right)$ *lies in the closed interval* $[\Sigma_l, \Sigma_u]$ *where* Σ_l *is the semiparametric efficiency bound of* (42) *and* Σ_u *is the asymptotic variance of the* $N^{1/2} \left(\hat{\tau}_{unadj} - \overline{\tau}_{\text{PATE}} \right)$*. Furthermore, the asymptotic variance of*

$$N^{1/2} \left(\hat{\tau}_{rothe,cal} - \overline{\tau}_{\text{PATE}} \right)$$

is no greater than the asymptotic variance of $N^{1/2} \left(\hat{\tau}_{rothe} - \overline{\tau}_{\text{PATE}} \right)$*, so in the case that* $\hat{\tau}_{rothe}$ *is more asymptotically precise than* $\hat{\tau}_{unadj}$ *we can replace* Σ_u *with the limiting variance of* $N^{1/2} \left(\hat{\tau}_{rothe} - \overline{\tau}_{\text{PATE}} \right)$ *and thereby shrink the interval even further.*

Proof. The stability assumption (Assumption 1) coupled with the moment assumptions Assumptions 8, and 9 establish that the calibrated imputation functions $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ are stable themselves; the argument mirrors the finite population case as before. Consequently, $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ satisfy Assumption 1 of (Rot20) (which is basically a variant of our stability assumption). Furthermore, the vanishing error process assumption (Assumption 2)

coupled with Assumptions 1, 8, and 9 establishes

$$N^{1/2}\left(\hat{\tau}_{rothe,cal} - \overline{\tau}_{\text{PATE}}\right) = N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_{OLS,0}, \dot{\mu}_{OLS,1}) - \overline{\tau}_{\text{PATE}}\right) + o_P(1).$$

Now we turn attention to the second claim of the theorem. The lower bound is an automatic consequence of (Rot20, Corollary 1) and (Hah98, Theorem 1); thus we turn to the upper bound. By the earlier asymptotic linearity result (see (44) for the calibrated case and (Rot20, Theorem 1) for the uncalibrated case), the asymptotic variance of $N^{1/2}\left(\hat{\tau}_{rothe} - \overline{\tau}_{\text{PATE}}\right)$ equals that of $N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right)$; consequently, it suffices to examine the asymptotic variance of $N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right)$. We leverage the law of total variance via a conditioning argument; as in Section 2.17 let $\mathcal{F}$ to be the event $\{(y_i(0), y_i(1), x_i) = (\mathbf{y}_i(0), \mathbf{y}_i(1), \mathbf{x}_i)\, i = 1, \ldots, N\}$.

$$\begin{aligned}\mathbb{V}\left(N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right)\right) = \\ \mathbb{E}\left[\mathbb{V}\left(N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right) \,\middle|\, \mathcal{F}\right)\right] + \\ \mathbb{V}\left(\mathbb{E}\left[N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right) \,\middle|\, \mathcal{F}\right]\right)\end{aligned}$$

Given $\mathcal{F}$ the only randomness in $N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right)$ comes from the random allocation of treatment assignment, Z, and since

$$\psi_i(\dot{\mu}_0, \dot{\mu}_1) = \dot{\mu}_1(x_i) - \dot{\mu}_0(x_i) + \frac{Z_i\left(y_i(Z_i) - \dot{\mu}_1(x_i)\right)}{(n_1/N)} - \frac{(1 - Z_i)\left(y_i(Z_i) - \dot{\mu}_0(x_i)\right)}{(n_0/N)}, \tag{45}$$

it follows that

$$\mathbb{E}\left[N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0, \dot{\mu}_1) - \overline{\tau}_{\text{PATE}}\right) \,\middle|\, \mathcal{F}\right] = N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i(1) - y_i(0)) - \overline{\tau}_{\text{PATE}}\right).$$

Consequently, the term $\mathbb{V}\left(\mathbb{E}\left[N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0,\dot{\mu}_1)-\overline{\tau}_{\text{PATE}}\right)\,\middle|\,\mathcal{F}\right]\right)$ has no dependence upon $\dot{\mu}_0$ and $\dot{\mu}_1$; it is just determined by the variance of the sample average treatment effect. Formally,

$$\mathbb{V}\left(\mathbb{E}\left[N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0,\dot{\mu}_1)-\overline{\tau}_{\text{PATE}}\right)\,\middle|\,\mathcal{F}\right]\right)=\mathbb{V}\left(N^{1/2}\left(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{PATE}}\right)\right).$$

Furthermore, by inspection of (45), the conditional variance of $\psi_i(\dot{\mu}_0,\dot{\mu}_1)$ given $\mathcal{F}$ is only dependent upon variability in $\frac{Z_i(y_i(Z_i)-\dot{\mu}_1(x_i))}{(n_1/N)}-\frac{(1-Z_i)(y_i(Z_i)-\dot{\mu}_0(x_i))}{(n_0/N)}$ inherited from randomness in the Z_i, so

$$\mathbb{V}\left(N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0,\dot{\mu}_1)-\overline{\tau}_{\text{PATE}}\right)\,\middle|\,\mathcal{F}\right)=$$
$$\mathbb{V}\left(N^{-1/2}\sum_{i=1}^{N}\left(\frac{Z_i\left(y_i(Z_i)-\dot{\mu}_1(x_i)\right)}{(n_1/N)}-\frac{(1-Z_i)\left(y_i(Z_i)-\dot{\mu}_0(x_i)\right)}{(n_0/N)}-\overline{\tau}_{\text{PATE}}\right)\,\middle|\,\mathcal{F}\right).$$

In total, we have shown that

$$\mathbb{V}\left(N^{-1/2}\sum_{i=1}^{N}\left(\psi_i(\dot{\mu}_0,\dot{\mu}_1)-\overline{\tau}_{\text{PATE}}\right)\right)=$$
$$\mathbb{E}\left[\mathbb{V}\left(N^{-1/2}\sum_{i=1}^{N}\left(\frac{Z_i\left(y_i(Z_i)-\dot{\mu}_1(x_i)\right)}{(n_1/N)}-\frac{(1-Z_i)\left(y_i(Z_i)-\dot{\mu}_0(x_i)\right)}{(n_0/N)}-\overline{\tau}_{\text{PATE}}\right)\,\middle|\,\mathcal{F}\right)\right]+$$
$$\mathbb{V}\left(N^{1/2}\left(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{PATE}}\right)\right). \quad (46)$$

Comparing (46) to the variance decomposition of Theorem A.3 and taking the limit as $N\to\infty$ yields that the asymptotic variance of $N^{1/2}\left(\hat{\tau}_{rothe}-\overline{\tau}_{\text{PATE}}\right)$ matches that of the $N^{1/2}\left(\hat{\tau}_{gOB}-\overline{\tau}_{\text{PATE}}\right)$ where the generalized Oaxaca-Blinder estimator is computed using the same imputation functions $\hat{\mu}_0$ and $\hat{\mu}_1$ that $\hat{\tau}_{rothe}$ uses. Consequently, the non-inferiority results for $N^{1/2}\left(\hat{\tau}_{cal}-\overline{\tau}_{\text{PATE}}\right)$ proven in Theorem A.3 translate to $N^{1/2}\left(\hat{\tau}_{rothe,cal}-\overline{\tau}_{\text{PATE}}\right)$ as

well. This establishes the upper bound on the asymptotic variance of $\hat{\tau}_{rothe,cal}$ and thereby completes the proof. □

Remark 7. The second claim of Theorem A.7 substantially improves upon the result of (Rot20, Corollary 1, Part ii) which provides only the lower bound. The upper bound of Theorem A.7 guarantees two things:

1. A practitioner is certain that their calibrated estimator is asymptotically no less efficient than the difference in means (or the uncalibrated estimator) regardless of model misspecification.

2. If the original estimator of $\hat{\tau}_{rothe}$ was indeed a semiparamteric efficient estimator then so too is the new calibrated estimator $\hat{\tau}_{rothe,cal}$.

Informally stated, the result of (Hah98) establishes that any estimator which has hope of semiparamteric efficiency must be of the form $\hat{\tau}_{rothe}$ for some choice of $\hat{\mu}_0$ and $\hat{\mu}_1$ and Theorem A.7 goes on to establish that any such estimator can be imbued with non-inferiority via the calibration procedure.

2.20 Cross-Fitting And Calibration

Cross-fitting is an algorithmic procedure based upon randomly splitting the sample into multiple portions, often called “folds", computing prediction functions on some portion of these folds, and then applying the prediction functions to data from the other folds. Since the prediction function is trained on different data than it is applied to, it is independent of the data it is applied to under standard superpopulation models. This independence provides numerous benefits from a theoretical angle and has established cross-fitting as a common and powerful tool in statistical literature. In particular, (CCD$^+$18) demonstrated that an appropriate use of cross-fitting could achieve strong statistical inference guarantees

while eschewing classical Donsker-style entropy conditions. Here we demonstrate that cross-fitting is compatible with calibration to yield non-inferiority results for a wide array of prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ which need not satisfy the "typically simple realizations" entropy condition. Our discussion centers around superpopulation inference in the style of Section 2.16.2.

We begin with some new notation. Superscripts of $(-i)$ indicate that the associated random function is independent of the ith data point; for example $\hat{\mu}_1^{(-i)}(\cdot)$ is a prediction function of treated outcomes which is independent of the ith observation $(y_i(Z_i), x_i)$. In practice, $\hat{\mu}_1^{(-i)}(\cdot)$ is usually computed by fitting the random prediction function $\hat{\mu}_1$ on the data set of $N-1$ individuals which excludes the ith individual. In line with the estimator $\hat{\tau}_{rothe}$ and the work of (WDTT16), define the "leave-one-out" estimator

$$\hat{\tau}_{loo} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_1^{(-i)}(x_i) - \hat{\mu}_0^{(-i)}(x_i)\right) + \frac{1}{n_1}\sum_{i\,:\,Z_i=1}\left(y_i(Z_i) - \hat{\mu}_1^{(-i)}(x_i)\right) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\left(y_i(Z_i) - \hat{\mu}_0^{(-i)}(x_i)\right).$$

To define the calibrated version of $\hat{\tau}_{loo}$, some care is needed to account for the sample splitting in both the original prediction functions $(\hat{\mu}_0, \hat{\mu}_1)$ and in the calibrated prediction functions. The leave-one-out calibrated prediction function is defined as

$$\hat{\mu}_{OLS,z}^{(-i)}(x_i) = \hat{\alpha}_z^{(-i)} + \hat{\beta}_{z,0}^{(-i)}\hat{\mu}_0^{(-i)}(x_i) + \hat{\beta}_{z,1}^{(-i)}\hat{\mu}_1^{(-i)}(x_i); \tag{47}$$

$$(\hat{\alpha}_z^{(-i)}, \hat{\beta}_{z,0}^{(-i)}, \hat{\beta}_{z,1}^{(-i)})^{\mathrm{T}} \in \underset{(\alpha_z,\beta_{z,0},\beta_{z,1})^{\mathrm{T}}}{\arg\min} \sum_{\substack{j:Z_j=z \\ j\neq i}} \{y_j(z) - \alpha_z - \beta_{z,0}\hat{\mu}_0^{(-i)}(x_j) - \beta_{z,1}\hat{\mu}_1^{(-i)}(x_j)\}^2.$$

The calibrated prediction function is then

$$\hat{\tau}_{loo,cal} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,1}^{(-i)}(x_i) - \hat{\mu}_{OLS,0}^{(-i)}(x_i)\right) + \frac{1}{n_1}\sum_{i\,:\,Z_i=1}\left(y_i(Z_i) - \hat{\mu}_{OLS,1}^{(-i)}(x_i)\right) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\left(y_i(Z_i) - \hat{\mu}_{OLS,0}^{(-i)}(x_i)\right).$$

In the context of leave-one-out estimation, we slightly modify the notation of the error process in Assumption 2. Specifically, write

$$\mathcal{G}_{N,z}(\hat{\mu}_z) = N^{-1/2}\sum_{i=1}^{N}\left(\frac{\mathbb{1}_{\{Z_i=z\}}\hat{\mu}_z^{(-i)}(x_i)}{n_z/N} - \hat{\mu}_z^{(-i)}(x_i)\right).$$

This modification of $\mathcal{G}_{N,z}(\hat{\mu}_z)$ is done to take into account the fact that different prediction functions may be used for different individuals in the population. In the case of leave-one-out estimation, there are $2N$ different prediction functions $\left(\hat{\mu}_0^{(-1)}, \ldots, \hat{\mu}_0^{(-N)}, \hat{\mu}_1^{(-1)}, \ldots, \hat{\mu}_1^{(-N)}\right)$. This change does not modify any of the structure of our previous proofs. In fact, we could have originally defined the process $\mathcal{G}_{N,z}$ to account for different prediction functions at each i, but this would have introduced needless notational burden for the previous proofs wherein the functions $\hat{\mu}_0$ and $\hat{\mu}_1$ do not depend upon i.

Theorem A.8. *Consider the superpopulation model model of Section 2.16.2. Suppose that Assumptions A.6, 2, 8, and 9 hold; then $\hat{\tau}_{loo,cal}$ obeys a central limit theorem and the asymptotic variance of $N^{1/2}\left(\hat{\tau}_{loo,cal} - \overline{\tau}_{\text{PATE}}\right)$ lies in the closed interval $[\Sigma_l, \Sigma_u]$ where Σ_l is the semiparametric efficiency bound of (42) and Σ_u is the asymptotic variance of the $N^{1/2}\left(\hat{\tau}_{unadj} - \overline{\tau}_{\text{PATE}}\right)$. Furthermore, the asymptotic variance of $N^{1/2}\left(\hat{\tau}_{loo,cal} - \overline{\tau}_{\text{PATE}}\right)$ is no greater than that of $N^{1/2}\left(\hat{\tau}_{loo} - \overline{\tau}_{\text{PATE}}\right)$; so in the case that $\hat{\tau}_{loo}$ is more asymptotically precise than $\hat{\tau}_{unadj}$ we can replace Σ_u with the limiting variance of $N^{1/2}\left(\hat{\tau}_{loo} - \overline{\tau}_{\text{PATE}}\right)$ and thereby shrink the interval even further.*

Proof. As in our previous proofs, we begin by showing that the estimators $\hat{\tau}_{loo}$ and $\hat{\tau}_{loo,cal}$ are asymptotically linear under the assumed regularity conditions. Quite similarly to (WDTT16)

$$\hat{\tau}_{loo} = \frac{1}{N}\left(\dot{\mu}_1(x_i) - \dot{\mu}_0(x_i)\right) + \frac{1}{n_1}\sum_{i\,:\,Z_i=1}\left(y_i(Z_i) - \dot{\mu}_1(x_i)\right) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\left(y_i(Z_i) - \dot{\mu}_0(x_i)\right) + R$$

where R is defined as

$$R = \sum_{i=1}^{N} \frac{(-1)^{Z_i}}{n_{Z_i}}\left(\frac{n_0}{N}\left(\hat{\mu}_1^{(-i)}(x_i) - \dot{\mu}_1(x_i)\right) + \frac{n_1}{N}\left(\hat{\mu}_0^{(-i)}(x_i) - \dot{\mu}_0(x_i)\right)\right).$$

The quantity $\hat{\tau}_{loo} - R$ exactly agrees with $N^{-1}\sum_{i=1}^{N}\psi_i(\dot{\mu}_0, \dot{\mu}_1)$, and so analysis of $\hat{\tau}_{loo}$ reduces to analysis of $N^{-1}\sum_{i=1}^{N}\psi_i(\dot{\mu}_0, \dot{\mu}_1)$ so long as R vanishes asymptotically at a sufficiently fast rate. Algebraic manipulation shows that

$$R = \frac{1}{N^{1/2}}\left(\mathcal{G}_{N,1}(\dot{\mu}_1) - \mathcal{G}_{N,1}(\hat{\mu}_1)\right) - \frac{1}{N^{1/2}}\left(\mathcal{G}_{N,0}(\dot{\mu}_0) - \mathcal{G}_{N,0}(\hat{\mu}_0)\right).$$

Consequently, by the vanishing error process assumption, $|R| = o_P(N^{-1/2})$. Similarly, by Lemma A.1 since $|R| = o_P(N^{-1/2})$ it follows that $|R_{cal}| = o_P(N^{-1/2})$ as well where

$$R_{cal} = \sum_{i=1}^{N} \frac{(-1)^{Z_i}}{n_{Z_i}}\left(\frac{n_0}{N}\left(\hat{\mu}_{OLS,1}^{(-i)}(x_i) - \dot{\mu}_{OLS,1}(x_i)\right) + \frac{n_1}{N}\left(\hat{\mu}_{OLS,0}^{(-i)}(x_i) - \dot{\mu}_{OLS,0}(x_i)\right)\right).$$

Consequently,

$$\hat{\tau}_{loo} = N^{-1}\sum_{i=1}^{N}\psi_i(\dot{\mu}_0, \dot{\mu}_1) + o_P(N^{-1/2}),$$
$$\hat{\tau}_{loo,cal} = N^{-1}\sum_{i=1}^{N}\psi_i(\dot{\mu}_{OLS,0}, \dot{\mu}_{OLS,1}) + o_P(N^{-1/2}),$$

and so the conclusions of Theorem A.7 hold automatically for $\hat{\tau}_{loo}$ and $\hat{\tau}_{loo,cal}$ as well.

□

Below we present sufficient conditions for the theorem. We begin with two definitions related to those of (WDTT16).

Definition 1. An estimator $\hat{\mu}_z$ is "jackknife compatible" if

$$\mathbb{E}\left[\sum_{i\,:\,Z_i=z}\left(\hat{\mu}_z^{(-i)}(x_{new})-\hat{\mu}_z(x_{new})\right)^2\right]=o(n_z)$$

for a new data point x_{new} drawn independently of the observed data.

Definition 2. An estimator $\hat{\mu}_z$ is "risk consistent" if

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_z^{(-i)}(x_i)-\dot{\mu}_z(x_i)\right)^2=o_P(1).$$

Proposition A.6. *Assume that $\hat{\mu}_0$ and $\hat{\mu}_1$ are jackknife compatible and risk consistent, then Assumption 2 holds.*

Proof. As before, define

$$R=\sum_{i=1}^{N}\frac{(-1)^{Z_i}}{n_{Z_i}}\left(\frac{n_0}{N}\left(\hat{\mu}_1^{(-i)}(x_i)-\dot{\mu}_1(x_i)\right)+\frac{n_1}{N}\left(\hat{\mu}_0^{(-i)}(x_i)-\dot{\mu}_0(x_i)\right)\right).$$

The argument of (WDTT16, Proof of Theorem 5) shows that $\mathbb{E}\left[R^2\right]=o(N^{-1})$.[10] In fact, a more detailed examination of their argument shows that both $\mathbb{E}\left[R_0^2\right]=o(N^{-1})$ and $\mathbb{E}\left[R_1^2\right]=$

[10]In fact, a small modification is needed to adapt the argument of (WDTT16) to our context; simply replace their $\mu^{(z)}$ with $\dot{\mu}_z$ for $z\in\{0,1\}$.

$o(N^{-1})$ for

$$R_0 = \sum_{i=1}^{N} \frac{(-1)^{Z_i}}{n_{Z_i}} \left(\frac{n_0}{N} \left(\hat{\mu}_1^{(-i)}(x_i) - \dot{\mu}_1(x_i) \right) \right).$$
$$R_1 = \sum_{i=1}^{N} \frac{(-1)^{Z_i}}{n_{Z_i}} \left(\frac{n_1}{N} \left(\hat{\mu}_0^{(-i)}(x_i) - \dot{\mu}_0(x_i) \right) \right).$$

Take $z \in \{0, 1\}$. For any $\varepsilon > 0$ Chebyshev's inequality implies that

$$\mathbb{P}\left(N^{1/2} |R_z| > \varepsilon \right) \leq \mathbb{E}\left[R_z^2 \right] N \varepsilon^{-2}.$$

Since $\mathbb{E}\left[R_z^2\right] = o(N^{-1})$ the right-hand-side vanishes so $R_z = o_P(N^{-1/2})$. Algebraic rearrangement yields that $R_z = \frac{1}{N^{1/2}} \left(\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z) \right)$; so $\mathcal{G}_{N,z}(\dot{\mu}_z) - \mathcal{G}_{N,z}(\hat{\mu}_z) = o_P(1)$ which concludes the proof.

□

Remark 8. Our definition of risk consistency differs sharply from that of (WDTT16) in that we are concerned only with the squared distance between the leave-one-out prediction $\hat{\mu}_z^{(-i)}(x_i)$ and some fixed function $\dot{\mu}_z(x_i)$ where $\dot{\mu}_z$ need not be the true conditional mean of $y_i(z)$ given the covariates. Consequently, even under arbitrary model misspecification $\hat{\tau}_{loo,cal}$ achieves asymptotic non-inferiority to both $\hat{\tau}_{loo}$ and $\hat{\tau}_{unadj}$. Moreover, the leave-one-out cross-fitting procedure allows one to avoid entropy conditions (cf. (GB21, Rot20)). This allows practitioners to use complex machine learning algorithms – subject to Definitions 1 and 2 – to form the initial estimators $\hat{\mu}_0$ and $\hat{\mu}_1$. See (WDTT16) for examples of such estimators; one such example is the subsampled random forest estimator of (Hil11).

Remark 9. A further advantage of leave-one-out estimation is in its control of finite sample bias; we discuss this issue in a superpopulation setting and focus on estimators of the form $\hat{\tau}_{rothe} = N^{-1} \sum_{i=1}^{N} \psi_i(\hat{\mu}_0, \hat{\mu}_1)$. In general, imputation-based estimators can introduce biases in the sense that $\mathbb{E}\left[\hat{\tau}_{rothe} - \overline{\tau}_{\text{PATE}}\right] \neq 0$. The central limit theorem for estimators of the form

$N^{-1}\sum_{i=1}^{N}\psi_i(\hat{\mu}_0,\hat{\mu}_1)$ implies that this bias is asymptotically vanishing. However, (WGB18) highlights that practical cases exist for which finite sample biases are unacceptable. Under mild conditions, leave-one-out estimation of the form $\hat{\tau}_{loo}$ – which is equivalent to the LOOP estimator of (WGB18) – achieves finite sample exact unbiasedness; see (WGB18) or (Rot20, Corollary 3) for proof. Furthermore, our analyses of $\hat{\tau}_{loo,cal}$ and $\hat{\tau}_{rothe,cal}$ demonstrate that calibration can be applied to simultaneously obtain non-inferiority guarantees for such estimators without any assumption of correct model specification. Together this implies that under the conditions of (Rot20, Corollary 3) or those of (WGB18) the leave-one-out calibrated estimator $\hat{\tau}_{loo,cal}$ has no finite sample bias and is asymptotically non-inferior to the difference in means.

We include simulations which mirror the superpopulation Poisson regression simulations of Table 2.3; we evaluate the performance of two sample-splitting estimators $\hat{\tau}_{loo}$ and $\hat{\tau}_{loo,cal}$. The results are summarized in Table 2.5. The first columns highlight that – in this particular simulation setting – $\hat{\tau}_{gOB}$ and $\hat{\tau}_{loo}$ are less efficient than the unadjusted difference in means $\hat{\tau}_{unadj}$. However, calibration of the leave-one-out estimator results in a new estimator $\hat{\tau}_{loo,cal}$ which is asymptotically no less efficient than the difference in means. The final column highlights that $\hat{\tau}_{loo,cal}$ and $\hat{\tau}_{cal}$ are asymptotically equivalent under the conditions of Theorem A.8. Both $\hat{\tau}_{loo}$ and $\hat{\tau}_{loo,cal}$ are guaranteed to have zero bias (WGB18, Rot20), so the third column of Table 2.5 demonstrates that via calibration we can construct an estimator with both desirable robustness properties: unbiasedness and asymptotic non-inferiority.

	$\frac{\hat{\text{MSE}}(\hat{\tau}_{gOB})}{\hat{\text{MSE}}(\hat{\tau}_{unadj})}$	$\frac{\hat{\text{MSE}}(\hat{\tau}_{loo})}{\hat{\text{MSE}}(\hat{\tau}_{unadj})}$	$\frac{\hat{\text{MSE}}(\hat{\tau}_{loo,cal})}{\hat{\text{MSE}}(\hat{\tau}_{unadj})}$	$\frac{\hat{\text{MSE}}(\hat{\tau}_{cal})}{\hat{\text{MSE}}(\hat{\tau}_{unadj})}$
$N = 200$	1.131	1.139	0.950	0.949
$N = 500$	1.119	1.121	0.948	0.948

Table 2.5: Ratios of Monte Carlo mean-square-errors for $\hat{\tau}_{cal}$, $\hat{\tau}_{gOB}$, $\hat{\tau}_{loo}$, and $\hat{\tau}_{loo,cal}$ to the difference in means estimator $\hat{\tau}_{unadj}$ for various experiment sizes N. Each mean-square-error is based upon $B = 100$ simulated experiments. Results are averaged over $S = 100$ simulations.

To illustrate the benefits of calibration even when flexible machine learning methods are used for imputation, below we present a simulation using random forests as the original imputation functions $\hat{\mu}_0$ and $\hat{\mu}_1$. In this simulation, we used the `Boston` data set from (VR02). In each simulation of we uniformly sampled N observations with replacement from the `Boston` data set; crime was used as the outcome of interest and the remaining information was taken as features. Independently, we drew a treatment allocation $Z \sim \mathit{Unif}(\Omega)$ with $n_1 = \lfloor pN \rfloor$ for $p = 0.6$. By construction, Fisher's sharp null holds. We use a sample splitting estimation procedure wherein prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are trained on half of the data and then applied to the other half to form the treatment effect estimator; the roles of the training and testing sets are interchanged and the results are averaged together to form the final estimator. This follows the 2-fold procedure of (WDTT16). The original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are random forests. We compare the results against the analogous calibrated estimator which linearly calibrates the prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ within the training sets; Table 2.6 summarizes the results. In particular, although both estimators achieve the semiparametric efficiency bound in the limit the calibrated estimator displays smaller variances across the simulations of Table 2.6. Table 2.7 repeats the simulations of Table 2.6 but incorporates treatment effect heterogeneity by adding independent exponential noise with rate 1 to the treated outcomes and subtracting independent exponential noise with rate 1 from the control outcomes. The ties between the variances of the calibrated and uncalibrated estimators presented in Table 2.7 demonstrates the benign effect of calibration when the original uncalibrated estimator was performing well to begin with.

	$\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF})$	$\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF,cal})$
$N = 500$	137.27	134.98
$N = 1000$	118.19	116.13
$N = 2000$	83.02	80.50

Table 2.6: **(Sharp Null Simulations)** Monte Carlo variances for the sample-split random forest imputation estimator, $\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF})$, and its calibrated analogue, $\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF,cal})$, for various experiment sizes N. Each variance is based upon 1000 simulated experiments.

	$\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF})$	$\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF,cal})$
$N = 500$	139.73	139.08
$N = 1000$	119.91	119.47
$N = 2000$	90.11	90.16

Table 2.7: **(Weak Null Simulations)** Monte Carlo variances for the sample-split random forest imputation estimator, $\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF})$, and its calibrated analogue, $\hat{\text{var}}(\sqrt{N}\hat{\tau}_{SSRF,cal})$, for various experiment sizes N. Each variance is based upon 1000 simulated experiments.

2.21 An Alternative Framework Via Entropy Conditions

Assumption 2 is sufficient for our results, but may be challengingly abstract. For the sake of clarity in our explication, we include some auxiliary results based upon Assumption 7 which demonstrate how to work directly with entropy conditions in the proofs of this paper. In the process, we establish several results which demonstrate that entropy conditions and Vapnik–Chervonenkis dimension conditions are sufficient for Assumption 2. We start with a few technical lemmas.

2.21.1 Some Technical Lemmas on Entropy Conditions

Lemma A.10. *For a sequence of function classes* $\{\mathscr{F}_N\}_{N\in\mathbb{N}}$ *suppose that*

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{F}_N, ||\cdot||_N, s)}\, ds < \infty.$$

Then it follows that for any $D > 0$

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, \frac{s}{D}\right)}\, ds < \infty.$$

Proof. By the change of variables $s = D^{-1}t$

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, \frac{s}{D}\right)}\, ds = \frac{1}{D}\int_0^D \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt.$$

We break the proof into two pieces: $D \geq 1$ versus $D \in (0,1)$. For now we focus on the first case, so assume that $D \geq 1$. The result is trivial when $D = 1$, so take $D > 1$ and let $\lfloor D \rfloor$ denote the greatest integer below D. It follows that

$$\begin{aligned}\int_0^D \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt = \int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt + \\ \int_1^{\lfloor D \rfloor} \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt + \\ \int_{\lfloor D \rfloor}^D \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt.\end{aligned}$$

The first term is guaranteed to be finite by assumption. To bound the second term, notice that $\mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)$ is a non-increasing function of t and so

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt \geq \int_\ell^{\ell+1} \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt. \tag{48}$$

for any $\ell \in \mathbb{N}$. Thus, the second term is no greater than

$$(\lfloor D \rfloor - 1)\int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt$$

which is certain to be finite. Lastly, because $\mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right) \geq 1$ it follows that the integrand of the third term is non-negative and so

$$\int_{\lfloor D \rfloor}^D \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt \leq \int_{\lfloor D \rfloor}^{\lfloor D \rfloor + 1} \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt;$$

the right-hand-side of this inequality is finite by (48). Consequently, when $D \geq 1$ the desired

result holds.

Now suppose that $D \in (0,1)$. As noted before, the integrand

$$\sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}$$

is non-negative and so

$$\int_0^D \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt \leq \int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, t\right)}\, dt;$$

the right-hand-side of this inequality is finite by assumption. Thus, when $D \in (0,1)$ the desired result holds, thereby concluding the proof. □

Lemma A.11. *Suppose that the prediction functions $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$ have typically simple realizations (Assumption 7). Then, so too does the joint prediction function*

$$x \mapsto \begin{bmatrix} \hat{\mu}_0(x) \\ \hat{\mu}_1(x) \end{bmatrix}.$$

Proof. Say that $\hat{\mu}_0$ satisfies Assumption 7 for the sequence of function classes $\{\mathscr{A}_{N,0}\}_{N\in\mathbb{N}}$ and $\hat{\mu}_1$ satisfies Assumption 7 for the sequence of function classes $\{\mathscr{A}_{N,1}\}_{N\in\mathbb{N}}$.

Consider the class of functions $\{\mathscr{C}_N\}_{N\in\mathbb{N}}$ for $\mathscr{C}_N = \mathscr{A}_{N,0} \times \mathscr{A}_{N,1}$.

By assumption

$$\mathbb{P}\left(\hat{\mu}_0 \in \mathscr{A}_{N,0}\right) \to 1 \quad \text{and} \quad \mathbb{P}\left(\hat{\mu}_1 \in \mathscr{A}_{N,1}\right) \to 1.$$

Because $\mathbb{P}\left(\begin{bmatrix} \hat{\mu}_0(\cdot) \\ \hat{\mu}_1(\cdot) \end{bmatrix} \in \mathscr{C}_N\right) \geq 1 - \mathbb{P}\left(\hat{\mu}_0 \notin \mathscr{A}_{N,0}\right) - \mathbb{P}\left(\hat{\mu}_1 \notin \mathscr{A}_{N,1}\right)$ it follows that

$$\mathbb{P}\left(\begin{bmatrix} \hat{\mu}_0(\cdot) \\ \hat{\mu}_1(\cdot) \end{bmatrix} \in \mathscr{C}_N\right) \to 1.$$

Now, we show that

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{C}_N, ||\cdot||_N, s)}\, ds < \infty.$$

Consider $\begin{bmatrix}\mu_0(\cdot)\\ \mu_1(\cdot)\end{bmatrix}, \begin{bmatrix}\nu_0(\cdot)\\ \nu_1(\cdot)\end{bmatrix} \in \mathscr{C}_n$; by definition

$$\begin{aligned}
\left|\left|\begin{bmatrix}\mu_0(\cdot)\\ \mu_1(\cdot)\end{bmatrix} - \begin{bmatrix}\nu_0(\cdot)\\ \nu_1(\cdot)\end{bmatrix}\right|\right|_N &= \left(\frac{1}{N}\sum_{i=1}^N \left|\left|\begin{bmatrix}\mu_0(x_i)\\ \mu_1(x_i)\end{bmatrix} - \begin{bmatrix}\nu_0(x_i)\\ \nu_1(x_i)\end{bmatrix}\right|\right|_2^2\right)^{1/2} \\
&= \left[\frac{1}{N}\sum_{i=1}^N \left\{(\mu_0(x_i)-\nu_0(x_i))^2 + (\mu_1(x_i)-\nu_1(x_i))^2\right\}\right]^{1/2} \\
&= \left[\frac{1}{N}\sum_{i=1}^N \{\mu_0(x_i)-\nu_0(x_i)\}^2 + \frac{1}{N}\sum_{i=1}^N \{\mu_1(x_i)-\nu_1(x_i)\}^2\right]^{1/2} \\
&\leq \left[\frac{1}{N}\sum_{i=1}^N \{\mu_0(x_i)-\nu_0(x_i)\}^2\right]^{1/2} + \\
&\qquad \left[\frac{1}{N}\sum_{i=1}^N \{\mu_1(x_i)-\nu_1(x_i)\}^2\right]^{1/2} \qquad (49) \\
&= ||\mu_0(\cdot)-\nu_0(\cdot)||_N + ||\mu_1(\cdot)-\nu_1(\cdot)||_N
\end{aligned}$$

where (49) follows by the subadditivity of the square root.

Consequently, $\mathscr{N}(\mathscr{C}_N, ||\cdot||_N, s) \leq \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2)\mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2)$ and so the

monotonicity of the logarithm gives that

$$
\begin{aligned}
&\int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{C}_N, ||\cdot||_N, s)}\, ds \\
&\leq \int_0^1 \sup_N \sqrt{\log\left(\mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2)\mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2)\right)}\, ds \\
&= \int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2) + \log \mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2)}\, ds \\
&\leq \int_0^1 \sup_N \Big(\sqrt{\log \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2)} + \\
&\qquad\qquad \sqrt{\log \mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2)}\Big)\, ds \\
&\leq \int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2)} + \\
&\qquad\qquad \int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2)}\, ds.
\end{aligned}
$$

The last line is guaranteed to be finite exactly because $\hat{\mu}_0$ satisfies Assumption 7 for the sequence of function classes $\{\mathscr{A}_{N,0}\}_{N\in\mathbb{N}}$ and $\hat{\mu}_1$ satisfies Assumption 7 for the sequence of function classes $\{\mathscr{A}_{N,1}\}_{N\in\mathbb{N}}$. □

Remark 10 (Multiplicative Bounds for Covering Numbers)**.** In the proof of Lemma A.11 we remarked that

$$
\left|\left|\begin{bmatrix}\mu_0(\cdot)\\ \mu_1(\cdot)\end{bmatrix} - \begin{bmatrix}\nu_0(\cdot)\\ \nu_1(\cdot)\end{bmatrix}\right|\right|_N \leq ||\mu_0(\cdot) - \nu_0(\cdot)||_N + ||\mu_1(\cdot) - \nu_1(\cdot)||_N \tag{50}
$$

implies that

$$
\mathscr{N}(\mathscr{C}_N, ||\cdot||_N, s) \leq \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2)\mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2). \tag{51}
$$

This reasoning plays an important role in our proofs and is of independent interest since covering numbers play an significant role in numerous areas of probability.

To avoid triviality, assume that both $\mathscr{A}_{N,0}$ and $\mathscr{A}_{N,1}$ are non-empty. Suppose that the points $a_1, \ldots, a_\ell$ provide a minimal cardinality $(s/2)$-cover of $\mathscr{A}_{N,0}$ and the points $b_1, \ldots, b_{\ell'}$ provide a minimal cardinality $(s/2)$-cover of $\mathscr{A}_{N,1}$. Consider the set of points

$$C = \{a_1, \ldots, a_\ell\} \times \{b_1, \ldots, b_{\ell'}\};$$

this set has cardinality $\ell\ell'$. Fix a point $c = (a, b) \in \mathscr{C}_N$. Since $a_1, \ldots, a_\ell$ and $b_1, \ldots, b_{\ell'}$ are $(s/2)$-covers there exists at least one point (a_i, b_j) for which

$$||a - a_i||_N \leq \frac{s}{2}$$
$$||b - b_j||_N \leq \frac{s}{2}.$$

By (50) $||c - (a_i, b_j)^{\mathrm{T}}||_N$ is bounded above by

$$||a - a_i||_N + ||b - b_j||_N$$

which is bounded above by s due to our choice of a_i and b_j. This implies that C is a valid s-cover of $\mathscr{C}_N$ which has cardinality $\ell\ell'$. Since C is a feasible s-cover it follows that the minimal cardinality s-cover of $\mathscr{C}_N$ must have cardinality no greater than $\ell\ell'$. However, since $\{a_1, \ldots, a_\ell\}$ is a minimal cardinality $(s/2)$-cover of $\mathscr{A}_{N,0}$ it follows that $\ell = \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s/2)$; likewise $\ell' = \mathscr{N}(\mathscr{A}_{N,1}, ||\cdot||_N, s/2)$. Consequently (51) holds.

Iterating the logic above implies the following theorem.

Theorem A.9. *Suppose that the set $T \subseteq T_1 \times \cdots \times T_\ell$; let $\pi_i : T \to T_i$ be the i^{th} coordinate projection. Suppose that (T, d) is a metric space such that*

$$d(a, b) \leq \sum_{i=1}^{\ell} d_i\left(\pi_i(a), \pi_i(b)\right)$$

where $d_i(\cdot,\cdot)$ denotes some metric on T_i. Then

$$\mathscr{N}(T,d,s) \leq \prod_{i=1}^{\ell} \mathscr{N}\left(T_i, d_i, \frac{s}{\ell}\right). \tag{52}$$

Proposition A.7. *If the prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$ are stable, have typically simple realizations, and satisfy Assumptions 4 and 5, then the prediction functions $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ also have typically simple realizations.*

Proof. Say that $\hat{\mu}_0$ satisfies Assumption 7 for the sequence of function classes $\{\mathscr{A}_{N,0}\}_{N\in\mathbb{N}}$ and $\hat{\mu}_1$ satisfies Assumption 7 for the sequence of function classes $\{\mathscr{A}_{N,1}\}_{N\in\mathbb{N}}$.

Let $\mathscr{C}_N = \mathscr{A}_{N,0} \times \mathscr{A}_{N,1}$. Define the sequence of function classes $\{\mathscr{F}_N\}_{N\in\mathbb{N}}$ via

$$\mathscr{F}_N = \left\{ \mu_{\beta_0,\beta_1}(x) = \beta_0 + \beta_1^{\mathrm{T}} \begin{bmatrix} \mu_0(x) \\ \mu_1(x) \end{bmatrix} \;\middle|\; |\beta_0 - \dot{\beta}_0^{(N)}| \leq 1;\ ||\beta_1 - \dot{\beta}_1^{(N)}||_2 \leq 1; \right.$$
$$\left. \left|\left|\left|\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix} - \begin{bmatrix} \dot{\mu}_0 \\ \dot{\mu}_1 \end{bmatrix}\right|\right|\right|_N \leq 1;\ \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix} \in \mathscr{C}_N \right\} \tag{53}$$

First, we show that $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ are asymptotically almost surely elements of $\mathscr{F}_N$. Since the proofs are basically the same for both $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$, we present the proof only for $\hat{\mu}_{OLS,1}$ and use the notation of (19) and (20). By the consistency of the ordinary least squares linear regression coefficients, Lemma A.3, it follows that

$$\lim_{N\to\infty} \mathbb{P}\left(|\hat{\beta}_0 - \dot{\beta}_0^{(N)}| > 1 \ \ \& \ \ ||\hat{\beta}_1 - \dot{\beta}_1^{(N)}||_2 > 1 \right) = 0. \tag{54}$$

By the joint stability of $\hat{\mu}_0$ and $\hat{\mu}_1$

$$\lim_{N\to\infty} \mathbb{P}\left(\left|\left|\left| \begin{bmatrix} \hat{\mu}_0 \\ \hat{\mu}_1 \end{bmatrix} - \begin{bmatrix} \dot{\mu}_0 \\ \dot{\mu}_1 \end{bmatrix} \right|\right|\right|_N > 1 \right) = 0. \tag{55}$$

Since $\hat{\mu}_0$ and $\hat{\mu}_1$ have are typically simple with respect to $\{\mathscr{A}_{N,0}\}_{N\in\mathbb{N}}$ and $\{\mathscr{A}_{N,1}\}_{N\in\mathbb{N}}$, respectively, it follows from Lemma A.11 that

$$\lim_{N\to\infty} \mathbb{P}\left(\begin{bmatrix}\hat{\mu}_0(\cdot)\\ \hat{\mu}_1(\cdot)\end{bmatrix} \in \mathscr{C}_N\right) = 1. \tag{56}$$

By Boole's inequality

$$\mathbb{P}\left(\hat{\mu}_{OLS,1} \in \mathscr{F}_N\right) \geq 1 - \mathbb{P}\left(|\hat{\beta}_0 - \dot{\beta}_0^{(N)}| > 1 \;\;\&\;\; ||\hat{\beta}_1 - \dot{\beta}_1^{(N)}|| > 1\right) - \mathbb{P}\left(\left|\left|\begin{bmatrix}\hat{\mu}_0\\ \hat{\mu}_1\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N > 1\right) - \mathbb{P}\left(\begin{bmatrix}\hat{\mu}_0(\cdot)\\ \hat{\mu}_1(\cdot)\end{bmatrix} \notin \mathscr{C}_N\right)$$

for each $N \in \mathbb{N}$, so it follows from (54), (55), and (56) that $\lim_{N\to\infty} \mathbb{P}\left(\hat{\mu}_{OLS,1} \in \mathscr{F}_N\right) = 1$. A mirrored proof yields that $\lim_{N\to\infty} \mathbb{P}\left(\hat{\mu}_{OLS,0} \in \mathscr{F}_N\right) = 1$.

All that remains to be shown is that

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}\left(\mathscr{F}_N, ||\cdot||_N, s\right)}\, ds < \infty.$$

To start, we examine two functions $f, g \in \mathscr{F}_N$ defined by

$$f(x) = \beta_{0f} + \beta_{1f}^{\mathrm{T}} \begin{bmatrix}\mu_{0f}(x)\\ \mu_{1f}(x)\end{bmatrix}$$

$$g(x) = \beta_{0g} + \beta_{1g}^{\mathrm{T}} \begin{bmatrix}\mu_{0g}(x)\\ \mu_{1g}(x)\end{bmatrix}.$$

The norm of their difference is

$$
\begin{aligned}
||f-g||_N &= \left|\left|(\beta_{0f}-\beta_{0g}) + \left(\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \beta_{1g}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right)\right|\right|_N \\
&\leq ||\beta_{0f}-\beta_{0g}||_N + \left|\left|\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \beta_{1g}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \qquad (57)\\
&= |\beta_{0f}-\beta_{0g}| + \left|\left|\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \beta_{1g}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \qquad (58)
\end{aligned}
$$

where (57) is due to the triangle inequality and (58) follows from the definition of $||\cdot||_N$ for constant functions. Furthermore

$$
\begin{aligned}
&\left|\left|\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \beta_{1g}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \\
&= \left|\left|\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix} + \beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix} - \beta_{1g}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \\
&\leq \left|\left|\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N + \left|\left|\beta_{1f}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix} - \beta_{1g}^{\mathrm{T}}\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \qquad (59)\\
&\leq ||\beta_{1f}||_N \left|\left|\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N + \left|\left|\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N ||\beta_{1f}-\beta_{1g}||_N \qquad (60)\\
&= ||\beta_{1f}||_2 \left|\left|\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N + \left|\left|\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N ||\beta_{1f}-\beta_{1g}||_2 \qquad (61)
\end{aligned}
$$

where (59) is due to the triangle inequality, (60) is due to the Cauchy-Schwarz inequality,

and (61) is due to the definition of $||\cdot||_N$ for constant functions. Because $f, g \in \mathscr{F}_N$

$$\left|\beta_{1f} - \dot{\beta}_1^{(N)}\right| \leq 1 \quad \text{and} \quad \left|\left|\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix} - \begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N \leq 1$$

so

$$||\beta_{1f}||_2 \leq 1 + \left|\left|\dot{\beta}_1^{(N)}\right|\right|_2 \quad \text{and} \quad \left|\left|\begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \leq 1 + \left|\left|\begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N .$$

Thus (61) can be bounded above by

$$\left(1 + \left|\left|\dot{\beta}_1^{(N)}\right|\right|_2\right) \left|\left|\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N + \left(1 + \left|\left|\begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N\right) ||\beta_{1f} - \beta_{1g}||_2 .$$

By Assumptions 4 and 5 and standard ordinary least squares regression results $\left|\left|\dot{\beta}_1^{(N)}\right|\right|_2$ is bounded uniformly in N. By Assumptions 4 and 5 and Lemma A.4 the quantity $\left|\left|\begin{bmatrix}\dot{\mu}_0\\ \dot{\mu}_1\end{bmatrix}\right|\right|_N$ is also bounded uniformly in N, it follows that (61) is bounded above by

$$\kappa \left(\left|\left|\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N + ||\beta_{1f} - \beta_{1g}||_2\right) \tag{62}$$

for some κ which does not depend upon N. Combining (58) with (62) yields that

$$||f - g||_N \leq D\,|\beta_{0f} - \beta_{0g}| + D\left|\left|\beta_{1f}^{\mathrm{T}} - \beta_{1g}^{\mathrm{T}}\right|\right|_2 + D\left|\left|\begin{bmatrix}\mu_{0f}\\ \mu_{1f}\end{bmatrix} - \begin{bmatrix}\mu_{0g}\\ \mu_{1g}\end{bmatrix}\right|\right|_N \tag{63}$$

for $D = \max\{1, \kappa\}$ which does not depend upon N.

Using (63) and Theorem A.9 we can bound the s-covering number of $\mathscr{F}_N$ as

$$\mathscr{N}(\mathscr{F}_N, ||\cdot||_N, s) \leq \mathscr{N}\left(\mathcal{B}(\dot{\beta}_0^{(N)}), |\cdot|, \frac{s}{3D}\right) \times \mathscr{N}\left(\mathcal{B}(\dot{\beta}_1^{(N)}), ||\cdot||_2, \frac{s}{3D}\right) \times \mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, \frac{s}{3D}\right) \quad (64)$$

where $\mathcal{B}(\dot{\beta}_0^{(N)})$ is the unit ball around $\dot{\beta}_0^{(N)}$ and $\mathcal{B}(\dot{\beta}_1^{(N)})$ is the unit ball around $\dot{\beta}_1^{(N)}$. Since the $s/(3D)$-covering number of the unit ball in $\mathbb{R}^m$ under the ℓ^2-norm is bounded above by $(1+6D/s)^m$ (GB21, Example 2) it follows from (64) that for all $s \in (0,1)$

$$\mathscr{N}(\mathscr{F}_N, ||\cdot||_N, s) \leq (1+6D/s) \times (1+6D/s)^2 \times \mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, s/(3D)\right). \quad (65)$$

By the monotonicity of the logarithm (65) implies that

$$\begin{aligned} &\int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{F}_N, ||\cdot||_N, s)}\, ds \quad (66)\\ &\leq \int_0^1 \sup_N \sqrt{\log\left((1+6D/s)^3 \mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, s/(3D)\right)\right)}\, ds \\ &= \int_0^1 \sup_N \sqrt{3\log(1+6D/s) + \log\left(\mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, s/(3D)\right)\right)}\, ds \end{aligned}$$

By the subadditivity of the square-root, the last line can be bounded above by

$$\begin{aligned} &\int_0^1 \sup_N \left(\sqrt{3\log(1+6D/s)} + \sqrt{\log\left(\mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, s/(3D)\right)\right)}\right) ds \\ &= \underbrace{\int_0^1 \sqrt{3\log(1+6D/s)}\, ds}_{a} + \underbrace{\int_0^1 \sup_N \sqrt{\log\left(\mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, s/(3D)\right)\right)}\, ds}_{b}. \quad (67) \end{aligned}$$

The term a is finite for all $D > 0$. The term b is finite by Lemmas A.10 and A.11. Thus, $\int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{F}_N, ||\cdot||_N, s)}\, ds < \infty$ and so $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ have typically simple realizations.

□

Remark 11. For a deterministic sequence of functions $\left\{f^{(N)}\right\}_{N\in\mathbb{N}}$ which map from $\mathbb{R}^k$ to $\mathbb{R}^\ell$ define the function class

$$\mathscr{F}_N = \left\{ \mu_{\beta_0,\beta_1}(x) = \beta_0 + \beta_1^{\mathrm{T}} \begin{bmatrix} \mu_0(x) \\ \mu_1(x) \end{bmatrix} + \beta_2^{\mathrm{T}} f^{(N)}(x) \;\middle|\; |\beta_0 - \dot{\beta}_0^{(N)}| \leq 1; \right.$$
$$\left. ||\beta_1 - \dot{\beta}_1^{(N)}||_2 \leq 1; ||\beta_2 - \dot{\beta}_2^{(N)}||_2 \leq 1; \left|\left| \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix} - \begin{bmatrix} \dot{\mu}_0 \\ \dot{\mu}_1 \end{bmatrix} \right|\right|_N \leq 1; \; \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix} \in \mathscr{C}_N \right\} \quad (68)$$

where $\dot{\beta}^{(N)}$ is derived from the population-level ordinary least squares linear regression including the engineered features $f^{(N)}(x_i)$; i.e., for the case of calibration in the treated population replace (20) with

$$\begin{aligned} \dot{\beta}^{(N)} &= (\dot{\beta}_0^{(N)}, \dot{\beta}_1^{(N)}, \dot{\beta}_2^{(N)}) \\ &= \underset{\beta_0,\beta_1,\beta_2}{\arg\min} \left[\sum_{i=1}^{N} \left\{ y_i(1) - \left(\beta_0 + \beta_1^T \begin{bmatrix} \dot{\mu}_0(x_i) \\ \dot{\mu}_1(x_i) \end{bmatrix} + \beta_2^{\mathrm{T}} f^{(N)}(x_i) \right) \right\}^2 \right]. \end{aligned}$$

Suppose that the engineered feature vectors $f(x_i)$ satisfy Assumptions 4 and 5, and the required linear regressions are not ill-defined. Under Assumptions 4 and 5 following the same line of reasoning used in the proof of Proposition A.7 yields that

$$\int_0^1 \sup_N \sqrt{\log \mathscr{N}(\mathscr{F}_N, ||\cdot||_N, s)}\, ds \leq \int_0^1 \sqrt{(3+\ell)\log(1+8D/s)}\, ds + \int_0^1 \sup_N \sqrt{\log\left(\mathscr{N}\left(\mathscr{C}_N, ||\cdot||_N, s/(4D)\right)\right)}\, ds < \infty.$$

Consequently, the prediction functions undergirding $\hat{\tau}_{cal2}$ have typically simple realizations.

2.21.2 Entropy Analysis in Finite Population Models

The work of (GB21) proceeds under a finite population model and the entropy condition Assumption 7, so we direct the interested reader to their explication. We highlight a two points that are relevant in the context of calibration; namely

- Lemma A.11 establishes that the mapping from covariates to "pseudo-covariates"

$$(\hat{\mu}_0(x_i), \hat{\mu}_1(x_i))$$

  automatically inherits typically simple realizations from the two original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$;

- Proposition A.7 leverages Lemma A.11 to conclude that the prediction functions $\hat{\mu}_{OLS,0}$ and $\hat{\mu}_{OLS,1}$ inherit typically simple realizations from the two original prediction functions $\hat{\mu}_0$ and $\hat{\mu}_1$.

2.21.3 Entropy Analysis in Superpopulation Models

In the superpopulation model of Section 2.16.2, the entropy condition Assumption 7 requires mild modification to account for randomness in potential outcomes and covariates.

Assumption 13 (Superpopulation Typically Simple Realizations)**.** *There exists a sequence of sets of functions $\mathscr{A}_{N,0}$, which may vary with N, such that the random function $\hat{\mu}_0$ falls into this class asymptotically almost surely. Formally, $\mathbb{P}\left(\hat{\mu}_0 \in \mathscr{A}_{N,0}\right) \to 1$. Furthermore, the sets of functions are "small" in the sense that*

$$\int_0^1 \mathbb{E}\left[\sup_N \sqrt{\log \mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s)}\right] ds < \infty$$

where $\mathscr{N}(\mathscr{A}_{N,0}, ||\cdot||_N, s)$ is the s-covering number of $\mathscr{A}_{N,0}$ under the random metric induced

by $||\cdot||_N$*, and the expectation is taken with respect to randomness in* $\{x_i\}_{i=1}^N$*. An analogous statement holds for* $\hat{\mu}_1$ *with a sequence of sets* $\mathscr{A}_{N,1}$*.*

The following lemma reproduces Lemma A.7 but directly uses the entropy condition Assumption 13 instead of assuming *a priori* that the error process vanishes as Assumption 2 does. A byproduct of this this result is that Assumption 13 is a sufficient condition for Assumption 2 in a superpopulation model.

Lemma A.12 (Superpopulation Linear Expansions Via Entropy Bounds)**.** *Under Assumptions A.6, 1, and 13 the random variable* $N^{-1}\sum_{i=1}^N (\hat{\mu}_z(x_i) - y_i(z))$ *is asymptotically linear in the sense that*

$$\frac{1}{N}\sum_{i=1}^N (\hat{\mu}_z(x_i) - y_i(z)) = \frac{1}{n_z}\sum_{i\,:\,Z_i=z} \dot{\epsilon}_i(z) + o_p(N^{-1/2})$$

where $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(x_i)$*.*

Proof. By the exact same reasoning as in (GB21, Proof of Theorem 3), rewrite

$$N^{-1}\sum_{i=1}^N (\hat{\mu}_z(x_i) - y_i(z))$$

as

$$\frac{1}{n_1}\sum_{i\,:\,Z_i=z} \dot{\epsilon}_i(z) + \frac{1}{N}\Bigg(\underbrace{\sum_{i=1}^N \left(\frac{Z_i\dot{\mu}_z(x_i)}{(n_z/N)} - \dot{\mu}_z(x_i)\right)}_{N^{1/2}\mathbb{G}(\dot{\mu}_z)} - \underbrace{\sum_{i=1}^N \left(\frac{Z_i\hat{\mu}_z(x_i)}{(n_z/N)} - \hat{\mu}_z(x_i)\right)}_{N^{1/2}\mathbb{G}(\hat{\mu}_z)}\Bigg).$$

Consequently, the desired result holds so long as we can show that $|\mathbb{G}(\dot{\mu}_z) - \mathbb{G}(\hat{\mu}_z)| = o_p(1)$. Unlike the proof in (GB21), the processes $\mathbb{G}(\dot{\mu}_z)$ and $\mathbb{G}(\hat{\mu}_z)$ inherit randomness from more than just Z; randomness enters through Z, $\{y_i(z)\}_{i=1}^N$, and $\{x_i\}_{i=1}^N$. For this, we use a Massart-like inequality (FHMM86, pg. 73-109) similar to that of Lemma S1 in (BLZ$^+$16). In the original formulation of (BLZ$^+$16) randomness is only due to Z. Consequently, for

some fixed universal constant $\kappa > 0$, by the logic of (GB21, Proof of Theorem 3)

$$\mathbb{P}\left(\sup_{\substack{\mu \in \mathscr{A}_{N,z} \\ ||\mu - \dot{\mu}_z||_N \leq r}} |\mathbb{G}(\dot{\mu}_z) - \mathbb{G}(\hat{\mu}_z)| > \varepsilon \,\middle|\, \{x_i, y_i(0), y_i(1)\}_{i=1}^N\right) \leq \kappa \frac{r}{\varepsilon} + \kappa \int_0^r \sup_N \left(\log \mathscr{N}(\mathscr{A}_{N,z}, ||\cdot||_N, s)\right)^{1/2} ds. \quad (69)$$

The norm $||\cdot||_N$ on the right-hand-side of (69) is random since it depends upon the realization of $\{x_i\}_{i=1}^N$. By the law of iterated expectation,

$$\mathbb{P}\left(\sup_{\substack{\mu \in \mathscr{A}_{N,z} \\ ||\mu - \dot{\mu}_z||_N \leq r}} |\mathbb{G}(\dot{\mu}_z) - \mathbb{G}(\hat{\mu}_z)| > \varepsilon\right) \leq \kappa \frac{r}{\varepsilon} + \kappa \mathbb{E}\left[\int_0^r \sup_N \left(\log \mathscr{N}(\mathscr{A}_{N,z}, ||\cdot||_N, s)\right)^{1/2} ds\right], \quad (70)$$

where the expectation on the right is with respect to randomness in $\{x_i\}_{i=1}^N$.[11] The Fubini-Tonelli theorem (Dur19, Theorem 1.7.2) justifies exchanging the expectation with the integral, yielding the upper bound of

$$\kappa \frac{r}{\varepsilon} + \kappa \int_0^r \mathbb{E}\left[\sup_N \left(\log \mathscr{N}(\mathscr{A}_{N,z}, ||\cdot||_N, s)\right)^{1/2} ds\right].$$

The remaining logic of (GB21, Proof of Theorem 3) combined with the superpopulation entropy condition of Assumption 13 yields that $|\mathbb{G}(\dot{\mu}_z) - \mathbb{G}(\hat{\mu}_z)| = o_p(1)$; asymptotic linearity follows immediately. □

Remark 12. The condition of Assumption 13 is implied by the uniform entropy bound of

[11] By the law of iterated expectation, the expectation on the right-hand-side of (70) is with respect to randomness in $\{x_i, y_i(0), y_i(1)\}_{i=1}^N$; however, the right-hand-side of the inequality has no dependence upon $\{y_i(0), y_i(1)\}_{i=1}^N$ and so the expectation can be taken to be only over $\{x_i\}_{i=1}^N$ without loss of generality.

Equation (2.5.1) in (vdVW96):

$$\int_0^\infty \sup_N \sup_Q \left(\log \mathscr{N}(\mathscr{A}_{N,z}, L_2(Q), s \, ||\mathscr{A}_{N,z}||_{Q,2}) \right)^{1/2} ds < \infty$$

where the supremum over Q ranges over all finitely discrete probability measures on $\mathbb{R}^k$. The log-covering number $\log \mathscr{N}(\mathscr{A}_{N,z}, L_2(Q), s \, ||\mathscr{A}_{N,z}||_{Q,2})$ vanishes when s exceeds one, since $\mathscr{A}_{N,z}$, which must be of finite diameter, can be covered by a single ball with diameter greater than $||\mathscr{A}_{N,z}||_{Q,2}$. Thus, the integral in question can be rewritten as

$$\int_0^1 \sup_N \sup_Q \left(\log \mathscr{N}(\mathscr{A}_{N,z}, L_2(Q), s \, ||\mathscr{A}_{N,z}||_{Q,2}) \right)^{1/2} ds. \tag{71}$$

Uniform entropy conditions are tightly related to Vapnik-Chervonenkis dimension (vdVW96, Theorem 2.6.7) for $s \in (0, 1)$

$$\mathscr{N}(\mathscr{A}_{N,z}, L_2(Q), s \, ||\mathscr{A}_{N,z}||_{Q,2}) \leq \tilde{C} \mathcal{VC}(\mathscr{A}_{N,z}) (16e)^{\mathcal{VC}(\mathscr{A}_{N,z})} \left(\frac{1}{s} \right)^{2(\mathcal{VC}(\mathscr{A}_{N,z})-1)},$$

where $\tilde{C}$ is a constant and $\mathcal{VC}(\mathscr{A}_{N,z})$ denotes the Vapnik-Chervonenkis dimension of the subgraphs of functions in the function class $\mathscr{A}_{N,z}$. Consequently

$$\log \mathscr{N}(\mathscr{A}_{N,z}, L_2(Q), s \, ||\mathscr{A}_{N,z}||_{Q,2}) \leq \log \tilde{C} + \log \mathcal{VC}(\mathscr{A}_{N,z}) + \mathcal{VC}(\mathscr{A}_{N,z}) \log(16e) + \log \left(\left(\frac{1}{s} \right)^{2(\mathcal{VC}(\mathscr{A}_{N,z})-1)} \right).$$

The finiteness of (71) is guaranteed as long as

$$\int_0^1 \sup_N \left(\log \tilde{C} + \log \mathcal{VC}(\mathscr{A}_{N,z}) + \right.$$
$$\left. \mathcal{VC}(\mathscr{A}_{N,z}) \log(16e) + \log\left(\left(\frac{1}{s} \right)^{2(\mathcal{VC}(\mathscr{A}_{N,z})-1)} \right) \right)^{1/2} ds < \infty.$$

Assume that $\mathcal{VC}(\mathscr{A}_{N,z})$ is bounded above by some constant independent of N. By subadditivity of the square root, a sufficient condition for the required finiteness of the integral above is to require that

$$\int_0^1 \sup_N \left(-2(\mathcal{VC}(\mathscr{A}_{N,z}) - 1) \log s\right)^{1/2} ds = \sup_N \left(\mathcal{VC}(\mathscr{A}_{N,z}) - 1\right)^{1/2} \left(\frac{\pi}{2}\right) ds < \infty.$$

This holds so long as $\sup_N \mathcal{VC}(\mathscr{A}_{N,z}) \geq 1$, which is true by definition.

Assumption 14. *For $z \in \{0, 1\}$, the Vapnik-Chervonenkis dimension of the sub-graphs of the function class $\mathscr{A}_{N,z}$ is bounded above by some finite constant which does not depend upon N.*

By the discussion above, Assumption 14 is a sufficient condition for Assumption 13 and consequently Vapnik-Chervonenkis conditions are sufficient to force the vanishing of the error process from Assumption 2.

2.21.4 Entropy Analysis in Fixed Covariate Models

Lemma A.13 (Fixed Covariate Linear Expansions via Entropy Bounds)**.** *Under Assumptions A.6, 7, and 12 the random variable $N^{-1} \sum_{i=1}^N (\hat{\mu}_z(x_i) - y_i(z))$ is conditionally almost surely asymptotically linear in the sense that, for $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(x_i)$*

$$\frac{1}{N} \sum_{i=1}^N (\hat{\mu}_z(x_i) - y_i(z)) = \frac{1}{n_z} \sum_{i \,:\, Z_i = z} \dot{\epsilon}_i(z) + o_p(N^{-1/2})$$

holds for almost all conditioning events of the form (38).

To prove Lemma A.13 condition on (38) and then apply the proof of Theorem 3 from (GB21) to the conditional random functions $\hat{\mu}_z$ given $\{(y_i(0), y_i(1)) = (\mathbf{y}_i(0), \mathbf{y}_i(1)) \, i = 1, \ldots, N\}$.

Theorem A.10. *Under the fixed-covariate model, subject to Assumptions A.6, 7, 10, 11, and 12,* $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right)$ *obeys a central limit theorem.*

Proof. We start out with the simple observation that

$$N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right) = N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right) + N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right).$$

Thus, to show a central limit theorem for $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right)$ it suffices to show that

1. Conditionally upon the potential outcomes, the term $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ converges weakly in probability to a fixed Gaussian distribution and term $N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$ obeys a central limit theorem.

2. The terms $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ and $N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$ are asymptotically independent in the sense that their limiting joint distribution is the product of the two limiting marginal distributions.

We tackle 1 first. By Lemma A.8 and the Lindeberg central limit theorem (LR05, Theorem 11.2.5) it follows that $N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$ converges in distribution to a Gaussian distribution; denote this limiting distribution as $\mathcal{N}(0, s_m)$.

Next, we show that $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)$ converges weakly in probability to a fixed Gaussian

distribution. We start with an algebraic rearrangement of $N^{1/2}\left(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}\right)$:

$$\begin{aligned} N^{1/2}\left(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}\right) &= N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,1}(x_i)-\hat{\mu}_{OLS,0}(x_i)\right)-\frac{1}{N}\sum_{i=1}^{N}\left(y_i(1)-y_i(0)\right)\right) \\ &= N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,1}(x_i)-y_i(1)\right)-\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{OLS,0}(x_i)-y_i(0)\right)\right). \end{aligned}$$

For each population of size N, condition upon some realization of the potential outcomes $\{(y_i(0),y_i(1))=(\mathbf{y}_i(0),\mathbf{y}_i(1))\, i=1,\ldots,N\}$. By Lemma A.9 for almost all such conditioning events, we have that

$$N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}Z_i N n_1^{-1}\dot{\epsilon}_i(1)-\frac{1}{N}\sum_{i=1}^{N}(1-Z_i)Nn_0^{-1}\dot{\epsilon}_i(0)\right)+o_P(1) \tag{72}$$

where $\dot{\epsilon}_i(z)=y_i(z)-\dot{\mu}_{OLS,z}(x_i)$ and the randomness in the $o_P(1)$ term is only with respect to randomness in Z since we condition upon the covariates implicitly in the fixed-covariate model.

Under Assumption 12 and the assumption that $N^{-1}\sum_{i=1}^{N}\left(\dot{\mu}_z(x_i)-y_i(z)\right)^2=o(N)$ as a numeric sequence for almost all conditioning events of the potential outcomes, by Lemma 3 in the appendix of (GB21) we can, without loss of generality, stipulate that $N^{-1}\sum_{i=1}^{N}\dot{\epsilon}_i(z)=0$ for $z\in\{0,1\}$ almost surely with respect to the conditioning (38).

Under Assumptions 10 and 11 the finite population analysis provided in Theorem 1 shows that $N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}Z_iNn_1^{-1}\dot{\epsilon}_i(1)-\frac{1}{N}\sum_{i=1}^{N}(1-Z_i)Nn_0^{-1}\dot{\epsilon}_i(0)\right)$ converges weakly to a centered Gaussian distribution with variance given by the limit of σ_N^2 defined in (28). This limit exists by Assumption 10 and is common to all conditioning events of the form (38) up to a set of measure zero; we denote it by s_d. Consequently, $N^{1/2}\left(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}\right)$ converges weakly in probability to a $\mathcal{N}\left(0,s_d\right)$.

Finally, we turn to 2. By Theorem 5.1 (iii) of (RBK05) it follows that the random vector $\left(N^{1/2}\left(\hat{\tau}_{cal}-\overline{\tau}_{\text{SATE}}\right),N^{1/2}\left(\overline{\tau}_{\text{SATE}}-\overline{\tau}_{\text{CATE}}\right)\right)$ converges in distribution to $(\mathcal{C},\mathcal{D})\sim\mathcal{N}\left(0,s_d\right)\otimes$

$\mathcal{N}(0, s_m)$.[12]

By the continuous mapping theorem (vdV98, Theorem 18.11),

$$N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right) + N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$$

converges in distribution to $\mathcal{C}+\mathcal{D}$. Since the sum of independent Gaussian random variables is itself Gaussian we have that $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{SATE}}\right)+N^{1/2}\left(\overline{\tau}_{\text{SATE}} - \overline{\tau}_{\text{CATE}}\right)$ converges in distribution to $\mathcal{N}(0, s_d + s_m)$. In turn, this implies that $N^{1/2}\left(\hat{\tau}_{cal} - \overline{\tau}_{\text{CATE}}\right)$ converges in distribution to $\mathcal{N}(0, s_d + s_m)$. □

Bibliography

[AH85] D. F. Andrews and A. M. Herzberg. *Data*. Springer New York, 1985.

[BBB+19] Andreas Buja, Richard Berk, Lawrence Brown, Edward George, Emil Pitkin, Mikhail Traskin, Linda Zhao, and Kai Zhang. Models as approximations i: Consequences illustrated with linear regression, 2019.

[Bli73] Alan S. Blinder. Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, 1973.

[BLZ+16] Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon, and Bin Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[CR15] Elizabeth Colantuoni and Michael Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine*, 34(18):2602–2617, 2015.

[12]The original work of (RBK05) focuses on survey-sampling; however, nothing of their result Theorem 5.1 (iii) relies upon the survey-sampling framework of having only a single potential outcome, so we apply their result to the causal inference context of multiple potential outcomes.

[CT97] Yuan Shih Chow and Henry Teicher. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. Independence, interchangeability, martingales.

[DDCZ13] Lutz Dümbgen and Perla Del Conte-Zerial. On low-dimensional projections of high-dimensional distributions. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 91–104. Inst. Math. Statist., Beachwood, OH, 2013.

[DFM19] Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

[DL18] Peng Ding and Fan Li. Causal inference: A missing data perspective. *Statist. Sci.*, 33(2):214–237, 05 2018.

[Dur19] Rick Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. Fifth edition of [MR1068527].

[ER59] Paul Erdos and Alfréd Rényi. On the central limit theorem for samples from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 4:49–61, 1959.

[Fel68] William Feller. *An introduction to probability theory and its applications. Vol. I.* John Wiley & Sons, Inc., New York-London-Sydney, third edition, 1968.

[FHMM86] X. Fernique, B. Heinkel, M. B. Marcus, and P.-A. Meyer, editors. *Geometrical and statistical aspects of probability in Banach spaces*, volume 1193 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986.

[Fre08a] David A. Freedman. On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.*, 2(1):176–196, 2008.

[Fre08b] David A. Freedman. On regression adjustments to experimental data. *Adv. in Appl. Math.*, 40(2):180–193, 2008.

[GB21] Kevin Guo and Guillaume Basse. The generalized Oaxaca-Blinder estimator. *Journal of the American Statistical Association*, page DOI: 10.1080/01621459.2021.1941053, 2021.

[GVL13] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[H́60] Jaroslav Hájek. Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:361–374, 1960.

[Hah98] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

[Hil11] Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[HIR03] Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[Hoe51] Wassily Hoeffding. A combinatorial central limit theorem. *Ann. Math. Statist.*, 22(4):558–566, 12 1951.

[Imb04] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[LD17] Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.*, 112(520):1759–1769, 2017.

[LD20] Lihua Lei and Peng Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828, 12 2020.

[Lin13] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

[LR05] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, Berlin, 1991.

[Mad48] William G. Madow. On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Statist.*, 19(4):535–545, 12 1948.

[MHL16] Zeinab Mashreghi, David Haziza, and Christian Léger. A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10(none):1 – 52, 2016.

[Ney23] Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (in Polish). *Roczniki Nauk Roiniczych*, X:1–51, 1923. Reprinted in Statistical Science, 1990, 5:463-480.

[NW21] Akanksha Negi and Jeffrey M. Wooldridge. Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5):504–534, 2021.

[Oax73] Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973.

[Pen55] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[Rak97] Vladimir Rakocevic. On continuity of the Moore-Penrose and Drazin inverses. volume 49, pages 163–172. 1997. 4th Symposium on Mathematical Analysis and Its Applications (Arandelovac, 1997).

[RBK05] Susana Rubin-Bleuer and Ioana Schiopu Kratina. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789 – 2810, 2005.

[RLSR12] Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M. Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

[Rot20] Christoph Rothe. Flexible covariate adjustments in randomized experiments, 2020. Available at `http://www.christophrothe.net/papers/fca_apr2020.pdf`.

[Rub80] Donald B Rubin. Comment on "Randomization analysis of experimental data: The Fisher randomization test". *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[RvdL10] Michael Rosenblum and Mark J. van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1), jan 2010.

[SLL14] Changyu Shen, Xiaochun Li, and Lingling Li. Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in medicine*, 33(4):555–568, 2014.

[Ste69] G. W. Stewart. On the continuity of the generalized inverse. *SIAM J. Appl. Math.*, 17:33–45, 1969.

[Tan10] Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

[vdV98] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[vdVW11] Aad van der Vaart and Jon A. Wellner. A local maximal inequality under uniform entropy. *Electron. J. Stat.*, 5:192–203, 2011.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[WDTT16] Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

[WGB18] Edward Wu and Johann A. Gagnon-Bartsch. The LOOP estimator: Adjusting for covariates in randomized experiments. *Evaluation Review*, 42:458 – 488, 2018.

[ZD21] Anqi Zhao and Peng Ding. Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2):278–294, 2021.

Chapter 3

Gaussian Prepivoting for Finite Population Causal Inference

Abstract

In finite population causal inference exact randomization tests can be constructed for sharp null hypotheses, hypotheses which impute the missing potential outcomes. Oftentimes inference is instead desired for the weak null that the sample average of the treatment effects takes on a particular value while leaving the subject-specific treatment effects unspecified. Tests valid for sharp null hypotheses can be anti-conservative should only the weak null hold. We develop a general framework for unifying modes of inference for sharp and weak nulls, wherein a single procedure simultaneously delivers exact inference for sharp nulls and asymptotically valid inference for weak nulls. We employ randomization tests based upon prepivoted test statistics, wherein a test statistic is first transformed by a suitably constructed cumulative distribution function and its randomization distribution assuming the sharp null is then enumerated. For a large class of test statistics, we show that prepivoting may be accomplished by employing the push-forward of a sample-based Gaussian measure based upon a suitable covariance estimator. The approach enumerates the randomization distribution (assuming the sharp null) of a P-value for a large-sample test known to be valid under the weak null, and uses the resulting randomization distribution for inference. The versatility of the method is demonstrated through many examples, including rerandomized designs and regression-adjusted estimators in completely randomized designs.

3.1 Introduction

In finite population causal inference two distinct hypotheses of "no treatment effect" are commonly tested: Fisher's sharp null and Neyman's weak null. Fisher's sharp null of no effect refers to the null that the responses under treatment and under control are the same for *all* individuals in the study (Ros02). The sharp null imputes the missing values of the potential outcomes for all individuals, in so doing facilitating the use of randomization tests to provide exact inference with randomization alone acting as the basis for inference (Fis35). Neyman's weak null instead specifies only that the average of the treatment effects for the individuals in the experiment equals zero, while allowing for heterogeneity in the unit-specific effects. The missing potential outcomes are no longer imputed under the weak null, such that the randomization distribution under the weak null remains unknown. Consequently, inference for the weak null has historically proceeded using asymptotically conservative analytical

approximations to the limiting distribution of the treated-minus-control difference in means.

While the exactness attained under the sharp null is appealing, randomization tests have been criticized for the seemingly restricted nature of the conclusions to which the researcher is entitled should the null be rejected (CDM17). While the researcher may suggest that the treatment effect is not zero for all individuals, generally one is not entitled to a statement of whether the treatment effect is positive or negative on average for the individuals in the study. To address this, a recent literature has emerged on how randomization tests may be modified to maintain asymptotic validity under the weak null hypothesis. The resulting methods provide a *single* testing procedure that is asymptotically valid for the weak null hypothesis, while maintaining exactness should the sharp null also be true (Din17, LRR17, DD18, WD18, Fog20).

The existing literature attains this unified mode of inference largely on a case-by-case basis: for a given experimental design, a specially catered test statistic is constructed so that the corresponding randomization test under the sharp null maintains asymptotic validity under the weak null. In this work, we provide both general conditions under which the unification may be achieved and a general methodology for attaining it. The central idea is to leverage prepivoting, an idea introduced in (Ber87, Ber88). For most commonly employed experimental designs and test statistics, the reference distribution generated by the prepivoted statistic under the assumption of the sharp null asymptotically stochastically dominates the true, but unknowable, randomization distribution under the weak null, yielding asymptotically conservative inference for the weak null while maintaining exactness under the sharp null hypothesis. As we demonstrate, prepivoting succeeds in many scenarios where other common resolutions such as studentization prove inadequate.

At a high level, prepivoting takes a test statistic T_0 and composes it with a cumulative distribution function $\hat{F}$ constructed from the observed data, forming the new test statistic $T_1 = \hat{F}(T_0)$. If $\hat{F}$ were a consistent estimate of T_0's limit distribution, $\hat{F}(T_0)$ would, through an asymptotic application of the probability integral transform, tend to a standard uniform.

Under the weak null hypothesis, the true distribution function for common test statistics T_0 cannot generally be consistently estimated. Fortunately, as developed in Section 3.5 a distribution function for a random variable that asymptotically stochastically dominates T_0 may be constructed. For most common test statistics for the weak null hypothesis, under conditions outlined in Section 3.5 this dominating distribution function amounts to a suitable pushforward of a multivariate Gaussian measure constructed using a conservative covariance estimator. Using this estimated distribution function, $T_1 = \hat{F}(T_0)$ is instead stochastically dominated by a standard uniform in the limit. Observe that through this construction, the prepivoted test $T_1 = \hat{F}(T_0)$ is precisely one minus the large sample p-value for the test statistic T_0 leveraging the central limit theorem. Rather than using this p-value to reach a conclusion by comparing its value to the desired α, we instead use the reference distribution of this large-sample p-value enumerated over all possible randomizations assuming the sharp null holds. This reference distribution generally converges pointwise to the standard uniform distribution function for commonly used covariance estimators. As a result, inference is guaranteed to be asymptotically conservative under the weak null while maintaining exactness under the sharp null. The general takeaway is that rather than looking at the randomization distribution of a test statistic itself under the sharp null, one should instead enumerate the randomization distribution of one minus an asymptotically valid p-value to restore validity of Fisher randomization tests when only the weak null holds.

In Section 3.2 we introduce notation for finite population causal inference and detail some standard assumptions. Section 3.3 defines the reference distribution assuming the truth of Fisher's sharp null and juxtaposes it with its true though unknowable randomization distribution under Neyman's weak null of no effect on average. After an overview of useful asymptotic results on completely randomized designs in Section 3.4, Section 3.5 introduces Gaussian prepivoting in the context of suitably constructed functions of treated-minus control difference in means. Section 3.6 provides examples of and insight into prepivoting using Gaussian measure. Section 3.7 extends these results to other asymptotically linear estima-

tors including regression-adjusted estimators, while Section 3.8 provides simulation studies highlighting the benefits of Gaussian prepivoting.

3.2 Notation And Review

3.2.1 Notation for Finite Population Causal Inference

While the developments in this work apply quite generally across common experimental designs and with two or more levels of the treatment, in this work we focus on completely randomized experiments and rerandomized experiments with two treatments; see the supplementary materials for extensions to paired designs and to completely randomized designs with multi-valued treatments. Consider a collection of N individuals, where n_1 receive treatment and $n_0 = N - n_1$ receive the control. For the ith individual, the random variable Z_i is the treatment indicator, taking the value 1 if the ith individual receives treatment and 0 otherwise. We assume that the stable unit treatment value assumption holds, such that there is no interference and that there are no hidden levels of the treatment (Rub80). The ith individual has two deterministic potential outcomes: $\mathbf{y}_i(1)$, the d-dimensional outcome under treatment, and $\mathbf{y}_i(0)$ the d-dimensional outcome under control. Furthermore, the ith unit has deterministic covariates $\mathbf{x}_i \in \mathbb{R}^k$. The jth coordinate of $\mathbf{y}_i(z)$ is $y_{ij}(z)$, and the analogous statement holds for x_{ij}. The random vector $\mathbf{Z}$ represents $(Z_1, \ldots, Z_N)^{\mathrm{T}}$; likewise $\mathbf{y}(\mathbf{1}) = (\mathbf{y}_1(1), \ldots, \mathbf{y}_N(1))^{\mathrm{T}}$ and $\mathbf{y}(\mathbf{0}) = (\mathbf{y}_1(0), \ldots, \mathbf{y}_N(0))^{\mathrm{T}}$. Under the finite population model the potential outcomes are viewed as fixed across randomizations, and the only randomness enters through $\mathbf{Z}$, the treatment allocation. For a discussion of the finite population inference framework, we suggest (DLM17). The observed outcome-vector for individual i is $\mathbf{y}_i(Z_i)$ and the collection of these is denoted $\mathbf{y}(\mathbf{Z})$. Causal inference with multiple outcomes is becoming increasingly common in modern applications ranging from drug repurposing studies to A/B tests assessing the impact of competing web page designs on various user

engagement metrics. See (TPSGN09) and (TPM11) for concrete examples of causal inference with multiple endpoints in biomedical sciences, and see (DFM19) for a reference on the underlying mathematics of multivariate potential outcome models.

The vector of treatment effects for the ith individual is $\boldsymbol{\tau}_i = \mathbf{y}_i(1) - \mathbf{y}_i(0)$. The average treatment effect for the individuals in the experiment is $\bar{\boldsymbol{\tau}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\tau}_i$. As the two potential outcomes are not jointly observable, $\boldsymbol{\tau}_i$ is unknown for all individuals. Neyman's weak null of no treatment effect on average is $H_N : \bar{\boldsymbol{\tau}} = \mathbf{0}$, while Fisher's sharp null further stipulates $H_F : \boldsymbol{\tau}_i = \mathbf{0}$ $(i = 1, \ldots, N)$ such that the treatment made no difference among any of the d outcomes measured. We implicitly define the alternative hypothesis as that which complements the null, so for H_N the alternative is $H_A : \overline{\boldsymbol{\tau}} \neq \mathbf{0}$ and for H_F the alternative is $H_A : \exists\, i$ s.t. $\boldsymbol{\tau}_i \neq \mathbf{0}$. Consequently, our tests are non-directional; this differs from the one-sided bounded alternatives tested by (CDM17). Furthermore, the one-sided bounded alternatives of (CDM17) bound each individual's treatment effect, whereas we are interested in unifying inference for both individual effects and aggregate effects.

For any matrix $\mathbf{r} \in \mathbb{R}^{N \times d}$ and any binary vector $\mathbf{W}$ with $\sum_{i=1}^{N} W_i = n_1$, we define the function

$$\hat{\tau}(\mathbf{r}, \mathbf{W}) = \frac{1}{n_1} \sum_{i=1}^{N} W_i \mathbf{r}_i - \frac{1}{n_0} \sum_{i=1}^{N} (1 - W_i) \mathbf{r}_i.$$

Using this notation, the observed treated-minus-control difference in means for the outcome variables is $\hat{\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and is often denoted by $\hat{\boldsymbol{\tau}}$ as shorthand. In general, "hats" are used to denote functions of observed quantities whose limiting properties will eventually be studied herein. Define $\bar{\mathbf{y}}(0) = N^{-1} \sum_{i=1}^{N} \mathbf{y}_i(0)$ and $\bar{\mathbf{y}}(1) = N^{-1} \sum_{i=1}^{N} \mathbf{y}_i(1)$ to be the average potential outcomes for the N individuals in the study population. Likewise, we define the covariance

matrices

$$\Sigma_{y(z)} = (N-1)^{-1} \sum_{i=1}^{N} (\mathbf{y}_i(z) - \bar{\mathbf{y}}(z))(\mathbf{y}_i(z) - \bar{\mathbf{y}}(z))^{\mathrm{T}}, \ z \in \{0,1\};$$
$$\Sigma_{\tau} = (N-1)^{-1} \sum_{i=1}^{N} (\boldsymbol{\tau}_i - \bar{\boldsymbol{\tau}})(\boldsymbol{\tau}_i - \bar{\boldsymbol{\tau}})^{\mathrm{T}}.$$

To emphasize the distinction between functions of observed outcomes and functions of covariates, we define the function $\hat{\delta}(\mathbf{x}, \mathbf{W})$ with binary $\mathbf{W}$ such that $\sum_{i=1}^{N} W_i = n_1$ as

$$\hat{\delta}(\mathbf{x}, \mathbf{W}) = \frac{1}{n_1} \sum_{i=1}^{N} W_i \mathbf{x}_i - \frac{1}{n_0} \sum_{i=1}^{N} (1 - W_i) \mathbf{x}_i.$$

The function $\hat{\delta}(\mathbf{x}, \mathbf{W})$ is a special case of $\hat{\tau}(\mathbf{r}, \mathbf{W})$. The observed difference in means for covariates is $\hat{\delta}(\mathbf{x}, \mathbf{Z})$, abbreviated as $\hat{\boldsymbol{\delta}}$. The finite population mean of the covariates is $\bar{\mathbf{x}} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_i$. The finite population covariance matrix for the covariates is Σ_x, defined by simply replacing $\mathbf{y}_i(z)$ with $\mathbf{x}_i$ and $\bar{\mathbf{y}}(z)$ with $\bar{\mathbf{x}}$ in the definition of $\Sigma_{y(z)}$. The finite population covariance between potential outcomes and covariates is $\Sigma_{y(z)x}$ for $z = 0, 1$, and the covariance between treatment effects and covariates is $\Sigma_{\tau x}$. Asymptotic arguments that follow will imagine a single sequence of finite populations of increasing size, with $N \to \infty$. As a result, quantities such as Σ_{τ} themselves vary as $N \to \infty$ and should be denoted by $\Sigma_{\tau,N}$ to reflect this. Generally the dependence is suppressed to reduce notational clutter; however, we do employ the notation $\Sigma_{\tau,\infty}$ to denote the limiting value of $\Sigma_{\tau,N}$ as $N \to \infty$, and likewise for other finite population quantities. For more on the finite population model for causal inference, see (IR15) and (DLM17) among many.

3.2.2 Rerandomized Designs and Balance Criterion

The set of all possible treatment assignments $\mathbf{Z}$ is denoted by Ω, and is determined by the experimental design. In completely randomized experiments, covariates are not used to inform the chosen treatment assignment and $\Omega_{CRE} = \{\mathbf{z} : \sum_{i=1}^{N} z_i = n_1\}$. To mitigate the risk of significant covariate imbalance, (MR12) suggest instead building covariate balance into the treatment allocation process through rerandomization. The study is conducted by collecting covariate data for the study participants, determining a measure of imbalance and a threshold for deciding what imbalances are acceptable, and selecting a treatment allocation uniformly over the set of allocations satisfying the balance criterion (LDR18). Stringent balance criterion reduce the cardinality of Ω by eliminating undesirable assignments, with the hopes of improving precision as a consequence. Naturally, randomization inference must take into account that the allowable realizations of $\mathbf{Z}$ depend upon the condition that covariate balance is met.

A *balance criterion* is a Boolean-valued function $\phi(\cdot)$, where $\phi(\sqrt{N}\hat{\boldsymbol{\delta}}) = 1$ is taken to mean that the treatment allocation $\mathbf{Z}$ which results in the particular realization of $\hat{\boldsymbol{\delta}}$ under consideration satisfies appropriate covariate balance. We impose the following restriction on ϕ:

Condition 1. *$\phi : \mathbb{R}^k \mapsto \{0, 1\}$ is an indicator function such that the set $M = \{\mathbf{b} : \phi(\mathbf{b}) = 1\}$ is closed, convex, mirror-symmetric about the origin (i.e. $\mathbf{b} \in M \Leftrightarrow -\mathbf{b} \in M$) with non-empty interior.*

3.2.3 Regularity Conditions

We make the following assumptions about the structure of the finite populations and experimental designs as N goes to infinity. These assumptions are for the most part standard in the literature; see, for instance, (WD18).

Assumption 1. *The proportion n_1/N limits to $p \in (0,1)$ as $N \to \infty$.*

Assumption 2. *All finite population means and covariances having limiting values for both the potential outcomes and the covariates. For instance,* $\lim_{N\to\infty} \bar{\mathbf{y}}(z) = \bar{\mathbf{y}}_\infty(z)$ *for* $z \in \{0,1\}$ *and* $\lim_{N\to\infty} \Sigma_{y(1)} = \Sigma_{y(1),\infty}$.

Assumption 3. *There exists some $C < \infty$ for which, for all $z \in \{0,1\}$, all $j = 1,..,d$ and all N,*

$$\frac{\sum_{i=1}^{N} \left(y_{ij}(z) - \overline{y}_j(z)\right)^4}{N} < C,$$

where $\overline{y}_j(z)$ denotes the jth coordinate of $\overline{\mathbf{y}}(z)$. Further, the above holds for the covariates with x_{ij} replacing $y_{ij}(z)$ above for $j = 1,..,k$.

Assumption 3 is used to obtain finite population-inference strong laws of large numbers for mean and variance estimators. Such an assumption is made at times for mathematical convenience to simplify the analysis of certain random distributions and may hold under weaker assumptions. Assumption 3 is commonplace in the literature on finite population causal inference; see, for instance, (WD18, Lin13, Fre08a, Fre08b).

3.2.4 A Technical Note on the Convergence of Random Measures

A random sequence of probability measures $\hat{\mu}_N$ on S converges weakly in probability to a deterministic probability measure μ if $\int_S f\, d\hat{\mu}_n \xrightarrow{p} \int_S f\, d\mu$ for all continuous bounded functions $f : S \to \mathbb{R}$ (DDCZ13, Section 2). Aspects of the Portmanteau Theorem (vdVW96, Theorem 1.3.4) extend to weak convergence in probability of random measures (DDCZ13, Cra02). Most importantly for our purposes is that if $\{\hat{F}_N\}$ are random cumulative distribution functions and F is a fixed cumulative distribution function, then their associated measures converge weakly in probability if and only if $\hat{F}_N(t) \xrightarrow{p} F(t)$ for all t which are continuity points of F; we take this as the definition of weak convergence in probability for random cumulative distribution functions.

3.3 Randomization Distributions And Tests

3.3.1 Randomization Distributions

Consider a scalar test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$, a function of the observed responses and the treatment assignment received. The *randomization distribution* for the test statistic T is

$$\mathscr{R}_T(t) = \frac{1}{|\Omega|} \sum_{\mathbf{w} \in \Omega} \mathbb{1}_{\{T(\mathbf{y}(\mathbf{w}), \mathbf{w}) \leq t\}}. \tag{1}$$

$\mathscr{R}_T$ is the true cumulative distribution function of $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ with respect to the randomness in the treatment allocation $\mathbf{Z}$, distributed uniformly over Ω. If we had access to $\mathscr{R}_T$ under the null hypothesis in question, we could make direct use of it to provide inference that is exact in finite samples, proceeding without dependence on asymptotics. Under Fisher's sharp null hypothesis, $\mathscr{R}_T$ *is* specified by the observed outcomes as $\mathbf{y}(\mathbf{Z}) = \mathbf{y}(\mathbf{w})$ for any $\mathbf{w} \in \Omega$. Unfortunately, the distribution is generally unknown under the weak null, as the weak null merely constrains the missing potential outcomes without determining them.

3.3.2 Randomization Tests Assuming the Sharp Null

In practice an experimenter draws a single realization of $\mathbf{Z}$, in so doing only revealing the values of the potential outcomes corresponding to the observed assignment. Suppose that regardless of whether or not Fisher's sharp null hypothesis actually holds, the researcher considers use of the randomization distribution to which she or he would be entitled if the sharp null were true. This reference distribution takes the form

$$\hat{\mathscr{P}}_T(t) = \frac{1}{|\Omega|} \sum_{\mathbf{w} \in \Omega} \mathbb{1}_{\{T(\mathbf{y}(\mathbf{Z}), \mathbf{w}) \leq t\}}. \tag{2}$$

While $\mathscr{R}_T = \hat{\mathscr{P}}_T$ under the sharp null, under the weak null $\hat{\mathscr{P}}_T$ is a random distribution function as it varies with $\mathbf{Z}$. Inference using $\hat{\mathscr{P}}_T$ proceeds as though $\mathbf{y}(\mathbf{Z})$ would have been the observed response for any $\mathbf{w} \in \Omega$. As the true response $\mathbf{y}(\mathbf{w})$ under assignment $\mathbf{w}$ need not align with $\mathbf{y}(\mathbf{Z})$, $\hat{\mathscr{P}}_T$ does not actually reflect the true randomization distribution under the weak null. This gives rise to potentially anti-conservative inference should $\hat{\mathscr{P}}_T$ be used to test the weak null hypothesis.

For $\alpha \in (0, 1)$ define the Fisher randomization test of nominal level α by

$$\varphi_T(\alpha) = \mathbb{1}_{\left\{T(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) \geq \hat{\mathscr{P}}_T^{-1}(1-\alpha)\right\}}. \tag{3}$$

Under the sharp null, $\mathbb{E}\{\varphi_T(\alpha)\} \leq \alpha$ for any sample size as $\hat{\mathscr{P}}_T = \mathscr{R}_T$. Throughout this paper, we examine the extent to which certain choices of test statistics entitle us to asymptotic Type I error control at α when $\varphi_T(\alpha)$ is used to conduct inference but only the weak null holds. For a given test statistic T, we will often proceed by juxtaposing its true limiting behavior under the randomization distribution $\mathscr{R}_T$ with the limiting behavior of $\hat{\mathscr{P}}_T$, the randomization distribution if we (incorrectly) assumed that the sharp null held.

3.3.3 Towards a Unified Mode of Inference

Suppose that for a test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ based upon the observed outcomes $\mathbf{y}(\mathbf{Z})$ and the treatment allocation $\mathbf{Z}$,

(a) $\hat{\mathscr{P}}_T$ converges weakly in probability to a fixed distribution $\mathscr{P}_{T,\infty}$ as $N \to \infty$; and

(b) $\mathscr{R}_T$ converges pointwise to a fixed distribution $\mathscr{R}_{T,\infty}$ at all continuity points of $\mathscr{R}_{T,\infty}$. Formally, $\mathscr{R}_T(t) \to \mathscr{R}_{T,\infty}(t) \quad \forall\, t \in \text{cont}(\mathscr{R}_{T,\infty})$, where $\text{cont}(\mathscr{R}_{T,\infty})$ is the set of continuity points of $\mathscr{R}_{T,\infty}$.

The test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is called *asymptotically sharp-dominant* if, under H_N, $\mathscr{P}_{T,\infty}(t) \leq \mathscr{R}_{T,\infty}(t)$ for any scalar t. This implies that the $(1 - \alpha)$ quantile of $\mathscr{P}_{T,\infty}$ is at or above the

$(1-\alpha)$ quantile of $\mathscr{R}_{T,\infty}$. If $T(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is asymptotically sharp-dominant, then inference based upon the reference distribution $\hat{\mathscr{P}}_T$ will be asymptotically conservative even if only H_N holds (WD18, Proposition 4), satisfying $\limsup \mathbb{E}\{\varphi_T(\alpha)\} \leq \alpha$ as $N \to \infty$ while maintaining exactness should the sharp null be true.

Many common test statistics are not asymptotically sharp-dominant over all elements of the weak null. For instance, with univariate potential outcomes and under a completely randomized design with imbalanced treated and control groups, the absolute difference in means $T(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) = \sqrt{N}|\hat{\tau}|$ is not generally asymptotically sharp-dominant as there exist sequences of potential outcomes satisfying the weak null such that $\liminf \mathbb{E}\{\varphi_T(\alpha)\} > \alpha$; see (Din17), (WD18, Cor. 3), or (LRR17) for details. For this test statistic, simply studentizing by the usual standard error estimator ensures sharp dominance. However, studentization fails to generalize to other more complicated test statistics and complex experimental designs. Significant efforts have recovered appropriate studentization techniques for some test statistics, but each test statistic requires its own separate analysis (WD18). For some experimental designs, studentizing the difference in means is not sufficient to regain asymptotically valid inference even in the univariate case; we explore this topic in Section 3.5.2 and Section 3.8.1 in the context of rerandomization. In Section 3.5, we present a general method called Gaussian prepivoting which both recovers studentization when it alone would be sufficient, but also yields asymptotic sharp-dominance in circumstances where studentization would be insufficient. Before describing the method, we recall a few important results on the difference in means in completely randomized designs which underpin the success of Gaussian prepivoting.

3.4 Useful Results For The Difference-In-Means In Completely Randomized Designs

3.4.1 Asymptotic Normality and Conservative Covariance Estimation for the Randomization Distribution

Consider the distribution of $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})^{\mathrm{T}}$ in a completely randomized design. Under Assumptions 1, 2, and 3, a finite population central limit theorem applies (LD17), and $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})^{\mathrm{T}}$ converges in distribution to a mean-zero multivariate Gaussian with covariance matrix V of the form

$$
\begin{aligned}
V &= \begin{pmatrix} V_{\tau\tau} & V_{\tau\delta} \\ V_{\delta\tau} & V_{\delta\delta} \end{pmatrix}; \\
V_{\tau\tau} &= p^{-1}\Sigma_{y(1),\infty} + (1-p)^{-1}\Sigma_{y(0),\infty} - \Sigma_{\tau,\infty}; \\
V_{\delta\delta} &= \{p(1-p)\}^{-1}\Sigma_{x,\infty}; \\
V_{\tau\delta} &= p^{-1}\Sigma_{y(1)x,\infty} + (1-p)^{-1}\Sigma_{y(0)x,\infty} = V_{\delta\tau}^{\mathrm{T}}.
\end{aligned}
$$

While $V_{\delta\delta}$ and $V_{\tau\delta}$ can be consistently estimated, $V_{\tau\tau}$ cannot be in the presence of effect heterogeneity due to its dependence on Σ_τ, the covariance of the unobserved treatment effects. Consequently, one cannot consistently estimate the probability that $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}})$ falls within a given region $\mathcal{B}$. While consistent variance estimates are not available, there are several covariance estimators $\hat{V}_{\tau\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ for $V_{\tau\tau}$ satisfying $\hat{V}_{\tau\tau} - V_{\tau\tau} \xrightarrow{p} \Delta$ for some $\Delta \succeq 0$ under Assumptions 1 - 3 in completely randomized designs. These estimators typically have the property that $\Sigma_{\tau\tau} = 0$ implies consistency, rather than asymptotic conservativeness; see (DFM19) for more details. So while the matrix V cannot generally be consistently estimated, one can construct an estimate converging in probability to a matrix $\bar{\bar{V}}$ of the form

$$\bar{\bar{V}} = \begin{pmatrix} V_{\tau\tau} + \Delta & V_{\tau\delta} \\ V_{\delta\tau} & V_{\delta\delta} \end{pmatrix}$$

with $\Delta \succeq 0$.

As an illustration, consider the conventional covariance estimator for the difference in means in a two-sample problem, $\hat{V}_{\tau\tau} = N\left(\hat{\Sigma}_{y(1)}/n_1 + \hat{\Sigma}_{y(0)}/n_0\right)$ with

$$\hat{\Sigma}_{y(1)} = \frac{1}{n_1 - 1}\sum_{i=1}^{N} Z_i \left(y_i(1) - n_1^{-1}\sum_{i=1}^{N} Z_i y_i(1)\right)\left(y_i(1) - n_1^{-1}\sum_{i=1}^{N} Z_i y_i(1)\right)^{\mathrm{T}}$$

and the analogous for $\hat{\Sigma}_{y(0)}$. Under both completely randomized experiments and rerandomized experiments with balance criterion satisfying Condition 1, this estimator satisfies $\hat{V}_{\tau\tau} - V_{\tau\tau} \xrightarrow{p} \Sigma_{\tau,\infty} \succeq 0$ under Assumptions 1 - 3.

3.4.2 Limiting Behavior of the Reference Distribution

Suppose we have a completely randomized design, and consider the random variable

$$\begin{aligned}
&\{\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W})\}^{\mathrm{T}} \\
&= \sqrt{N}\Bigg\{\frac{1}{n_1}\sum_{i=1}^{N} W_i \tilde{\mathbf{y}}_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1 - W_i)\tilde{\mathbf{y}}_i(Z_i), \\
&\qquad\qquad \frac{1}{n_1}\sum_{i=1}^{N} W_i \mathbf{x}_i - \frac{1}{n_0}\sum_{i=1}^{N}(1 - W_i)\mathbf{x}_i\Bigg\}^{\mathrm{T}}
\end{aligned}$$

where $\mathbf{Z}$ and $\mathbf{W}$ are independent, identically distributed, and drawn uniformly from Ω and $\tilde{\mathbf{y}}_i(Z_i) = \mathbf{y}_i(Z_i) - Z_i\bar{\boldsymbol{\tau}}$, such that $\tilde{\mathbf{y}}(\mathbf{Z}) = \mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\boldsymbol{\tau}}^{\mathrm{T}}$.

Proposition 1. *Subject to Assumptions 1, 2, and 3, under a completely randomized design*

the distribution of $\{\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W})\}^{\mathrm{T}} \mid \mathbf{Z}$ *converges weakly in probability to a multivariate Gaussian measure, with mean zero and covariance* $\tilde{V}$ *of the form*

$$\begin{aligned}
\tilde{V} &= \begin{pmatrix} \tilde{V}_{\tau\tau} & \tilde{V}_{\tau\delta} \\ \tilde{V}_{\delta\tau} & \tilde{V}_{\delta\delta} \end{pmatrix}; \\
\tilde{V}_{\tau\tau} &= (1-p)^{-1}\Sigma_{y(1),\infty} + p^{-1}\Sigma_{y(0),\infty}; \\
\tilde{V}_{\delta\delta} &= \{p(1-p)\}^{-1}\Sigma_{x,\infty}; \\
\tilde{V}_{\tau\delta} &= (1-p)^{-1}\Sigma_{y(1)x,\infty} + p^{-1}\Sigma_{y(0)x,\infty} = \tilde{V}_{\delta\tau}^{\mathrm{T}}.
\end{aligned}$$

The proof of this statement is contained within the proof of Theorem 1 in (WD18) and is omitted. Under the sharp null, $\tilde{V} = V$ as $\mathbf{y}_i(1) = \mathbf{y}_i(0)$ for all i. Under the weak null however, while $\tilde{V}_{\delta\delta} = V_{\delta\delta}$ generally $\tilde{V}_{\tau\tau} \neq V_{\tau\tau}$ and $\tilde{V}_{\delta\tau} \neq V_{\delta\tau}$. The divergence between V and $\tilde{V}$ can render randomization tests for the weak null hypothesis anti-conservative; examples are given in Section 3.5.2. We now describe how prepivoting may be used to guarantee asymptotic correctness when inference for the weak null hypothesis is conducted using a reference distribution generated under the sharp null.

3.5 Gaussian Prepivoting

3.5.1 Prepivoting with an Estimated Pushforward Measure

Consider functions $f_\eta : \mathbb{R}^d \to \mathbb{R}$ subject to the following requirement:

Condition 2. *For any* $\eta \in \Xi$, $f_\eta(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}_+$ *is continuous, quasi-convex, and nonnegative with* $f_\eta(\mathbf{t}) = f_\eta(-\mathbf{t})$ *for all* $\mathbf{t} \in \mathbb{R}^d$. *Furthermore,* $f_\eta(\mathbf{t})$ *is jointly continuous in* η *and* $\mathbf{t}$.

We begin with statistics for H_N of the form

$$T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}}), \tag{4}$$

where $\hat{\xi} = \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$ satisfies the following condition for some set Ξ:

Condition 3. *With* $\mathbf{W}, \mathbf{Z}$ *independent and each uniformly distributed over* Ω*,*

$$\hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} \xi; \;\; \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \xrightarrow{p} \tilde{\xi},$$

for some $\xi, \tilde{\xi} \in \Xi$*.*

As will be shown in Section 3.5.2, several commonly encountered statistics for Neyman's null are of this form. A detailed discussion of Condition 2 is included in the supplementary materials. Suppose further that one employs a covariance estimator $\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$ with the following property:

Condition 4. *With* $\mathbf{W}, \mathbf{Z}$ *independent, both uniformly distributed over* Ω*, and for some* $\Delta \succeq 0$*,* $\Delta \in \mathbb{R}^{d \times d}$*,*

$$\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) - V \xrightarrow{p} \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix}; \;\; \hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) - \tilde{V} \xrightarrow{p} 0_{(d+k),(d+k)}.$$

As a concrete example satisfying Conditions 2-4, suppose that $f_\eta(\mathbf{t}) = \mathbf{t}^{\mathrm{T}} \eta^{-1} \mathbf{t}^{\mathrm{T}}$ and $\hat{\xi}(\mathbf{y}(\mathbf{z}), \mathbf{w}) = \hat{V}_{Neyman}(\mathbf{y}(\mathbf{z}), \mathbf{w})$ with $\hat{V}_{Neyman}$ denoting the usual Neyman variance estimator of (Ney90); numerous other examples are included in Section 3.5.2. Observe that when assuming the weak null for the purpose of testing, $\tilde{\mathbf{y}}(\mathbf{Z}) = \mathbf{y}(\mathbf{Z})$ and $\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}$. Gaussian prepivoting transforms the test statistic $T(\mathbf{y}(\mathbf{z}), \mathbf{w}) = f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{z}), \mathbf{w}))$ into a new statistic of the form

$$G(\mathbf{y}(\mathbf{z}), \mathbf{w}) = \frac{\gamma^{(d+k)}_{\mathbf{0}, \hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{w})} \left\{ (\mathbf{a}, \mathbf{b})^{\mathrm{T}} : f_{\hat{\xi}}(\mathbf{a}) \leq T(\mathbf{y}(\mathbf{z}), \mathbf{w}) \;\wedge\; \phi(\mathbf{b}) = 1 \right\}}{\gamma^{(k)}_{\mathbf{0}, \hat{V}_{\delta\delta}} \{ \mathbf{b} : \phi(\mathbf{b}) = 1 \}} \tag{5}$$

where $\gamma^{(p)}_{\mu, \Sigma}(\mathcal{B})$ is the p-dimensional Gaussian measure of a set $\mathcal{B}$ with mean parameter $\boldsymbol{\mu}$ and

covariance Σ, i.e.

$$\gamma_{\mu,\Sigma}^{(p)}(\mathcal{B}) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \int_{\mathbf{x}\in\mathcal{B}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\, dx.$$

For $(\mathbf{A},\mathbf{B})^{\mathrm{T}}$ jointly multivariate normal with mean zero and covariance $\hat{V}$, $\mathbf{A}\in\mathbb{R}^d$, $\mathbf{B}\in\mathbb{R}^k$, $G(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ represents the $f_{\hat{\xi}}$-pushforward measure of $\mathbf{A} \mid \phi(\mathbf{B}) = 1$ evaluated on the set $(-\infty, T(\mathbf{y}(\mathbf{Z}),\mathbf{Z})]$. That is, $G(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ treats $f_{\hat{\xi}}$ and $\hat{V}$ as fixed and computes the conditional probability that $f_{\hat{\xi}}(\mathbf{A})$ falls at or below the observed value for $T(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) = f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}})$ given that $\varphi(\mathbf{B}) = 1$. From the perspective of hypothesis testing, $G(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is 1 minus the large-sample p-value for $T(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ leveraging the finite population central limit theorem and the estimated covariance $\hat{V}$.

We now describe how to use the prepivoted statistic $G(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ to provide a single procedure that is both exact for H_F and asymptotically conservative for H_N. In order to provide a precise implementation of this, we give detailed pseudocode in Algorithm 1 and provide example code through the supplementary materials. First, we compute the prepivoted test statistic $G(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ given the observed data; this proceeds according to Equation 5 and is Step 1 of Algorithm 1. Next, we construct the reference distribution $\hat{\mathscr{P}}_G(\cdot)$, the construction of which requires imputing counterfactual outcomes as if Fisher's sharp null held; this is Step 2 of Algorithm 1. Finally, the p-value for testing H_N is computed and we reject the null when this lies below or at the nominal level α.

Algorithm 1: Inference for the weak null through Gaussian prepivoting

Input: An observed treatment allocation $\mathbf{z}$, with observed responses $\mathbf{y}(\mathbf{z})$, test statistic $T(\mathbf{y}(\mathbf{z}), \mathbf{z}) = f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}}_{obs})$ and covariance estimator $\hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{z})$

Result: The p-value for the Gaussian prepivoted test statistic

Step 1: The observed prepivoted statistic

Compute $f_{\hat{\xi}(\mathbf{y}(\mathbf{z}),\mathbf{z})}(\cdot)$; $\hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{z})$. Compute

$$g_{\mathbf{z}} = \frac{\gamma^{(d+k)}_{\mathbf{0},\hat{V}(\mathbf{y}(\mathbf{z}),\mathbf{z})}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : f_{\hat{\xi}(\mathbf{y}(\mathbf{z}),\mathbf{z})}(\mathbf{a}) \leq T(\mathbf{y}(\mathbf{z}), \mathbf{z}) \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma^{(k)}_{\mathbf{0},\hat{V}_{\delta\delta}(\mathbf{y}(\mathbf{z}),\mathbf{z})}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}$$

Step 2: The reference distribution $\hat{\mathscr{P}}_G$

for $\mathbf{w} \in \Omega$ **do**

Compute $f_{\hat{\xi}(\mathbf{y}(\mathbf{z}),\mathbf{w})}(\cdot)$; $\hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{w})$.

Compute

$$g_{\mathbf{w}} = \frac{\gamma^{(d+k)}_{\mathbf{0},\hat{V}(\mathbf{y}(\mathbf{z}),\mathbf{w})}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : f_{\hat{\xi}(\mathbf{y}(\mathbf{z}),\mathbf{w})}(\mathbf{a}) \leq T(\mathbf{y}(\mathbf{z}), \mathbf{w}) \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma^{(k)}_{\mathbf{0},\hat{V}_{\delta\delta}(\mathbf{y}(\mathbf{z}),\mathbf{w})}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}$$

end

return

$$p_{val} = \frac{1}{|\Omega|}\sum_{\mathbf{w}\in\Omega} \mathbb{1}_{\{g_{\mathbf{w}} \geq g_{\mathbf{z}}\}};$$

$$\varphi_G(\alpha) = \mathbb{1}_{\{p_{val} \leq \alpha\}}.$$

Observe that $1 - g_{\mathbf{z}}$ defined within Algorithm 1 is the usual large-sample p-value based upon a Gaussian approximation and using the covariance estimator $\hat{V}$. The large-sample test compares $1 - g_{\mathbf{z}}$ to α, the desired Type I error rate, and rejects if $1 - g_{\mathbf{z}} \leq \alpha \Leftrightarrow g_{\mathbf{z}} \geq$

$1-\alpha$. The Gaussian prepivoted randomization test instead rejects if $g_{\mathbf{z}} \geq \hat{\mathscr{P}}_G^{-1}(1-\alpha)$. The following Theorem, in concert with Lemma 11.2.1 of (LR05), show under our assumptions $\hat{\mathscr{P}}_G^{-1}(1-\alpha) \xrightarrow{p} 1-\alpha$, such that the prepivoted randomization test is asymptotically equivalent to large sample test under the weak null. By using $\hat{\mathscr{P}}_G^{-1}(1-\alpha)$ instead of $1-\alpha$, exactness under the sharp null is preserved.

Theorem 1. *Suppose we have either a completely randomized design or a rerandomized design with balance criterion ϕ satisfying Condition 1. Suppose $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is of the form (4) for some f_η and $\hat{\xi}$ satisfying Conditions 2 and 3. Suppose further that we employ a covariance estimator $\hat{V}$ satisfying Condition 4 when forming the prepivoted test statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. Then, under $H_N : \bar{\boldsymbol{\tau}} = 0$ and under Assumptions 1-3, $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in distribution to a random variable $\tilde{U}$ taking values in $[0,1]$ satisfying*

$$\mathbb{P}(\tilde{U} \leq t) \geq t,$$

for all $t \in [0,1]$. Furthermore, the distribution $\hat{\mathscr{P}}_G(t)$ satisfies $\hat{\mathscr{P}}_G(t) \xrightarrow{p} t$ for all $t \in [0,1]$.

Corollary 1. *Under the conditions of Theorem 1; the prepivoted statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is asymptotically sharp dominant regardless of whether the base statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ was. Consequently, p-values derived under $\hat{\mathscr{P}}_G$ via Algorithm 1 are guaranteed to be exact under H_F and asymptotically conservative under just H_N.*

Theorem 1 states that under the weak null, $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in distribution to a random variable which is stochastically dominated by the standard uniform. Meanwhile, the reference distribution for $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ constructed assuming (incorrectly) that the sharp null holds converges pointwise to the distribution function of a standard uniform. As a result, the randomization distribution for $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is asymptotically sharp-dominant: the reference distribution generated in this manner yields asymptotically conservative inference for the weak null hypothesis, while maintaining exactness should the sharp null also hold. By

exploiting the duality between hypothesis testing and confidence sets Theorem 1 provides the basis for generating exact and asymptotically conservative confidence sets for treatment effect; this is explored in the supplementary materials.

Remark 1. Consider the function

$$\hat{F}(t) = \frac{\gamma_{\mathbf{0},\hat{V}}^{(d+k)}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : f_{\hat{\xi}}(\mathbf{a}) \leq t \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma_{\mathbf{0},\hat{V}_{\delta\delta}}^{(k)}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}$$

the estimated distribution function for $f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}}) \mid \phi(\sqrt{N}\hat{\boldsymbol{\delta}}) = 1$ based upon a finite population central limit theorem. In special cases, the function $\hat{F}(t)$ may have a known closed form. This is true of the test statistics which are sharp-dominated by a χ^2_d distribution considered in (WD18), for example. Should this not be the case, one can approximate $\hat{F}(\cdot)$ by way of Monte-Carlo approximation, replacing the measures $\gamma_{\mathbf{0},\hat{V}}$ and $\gamma_{\mathbf{0},\hat{V}_{\delta\delta}}$ with estimates based upon a B draws from a multivariate normal with mean $\mathbf{0}$ and covariance $\hat{V}$ when enumerating the reference distribution. Importantly, such Monte-Carlo approximation does *not* corrupt finite-sample exactness under Fisher's sharp null.

3.5.2 Examples of Gaussian prepivoting

Through a series of examples, we now provide illustrations of the transformations achieved by (5). As will be demonstrated, the form recovers several randomization tests previously known to be valid for weak null hypotheses in the literature while providing a basis for new randomization tests for weak nulls using other test statistics. These examples serve four objectives: (i) unify previous *ad hoc* solutions under the framework of Gaussian prepivoting; (ii) provide an alternative approach to already valid procedures; (iii) highlight that prepivoting can succeed even where studentization fails; and (iv) extend randomization inference for H_F and H_N to new experimental designs.

Example 1 (Absolute difference in means)**.** Let $\sqrt{N}\hat{\tau}$ be univariate, consider a completely randomized design with no rerandomization, and let $T_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \sqrt{N}|\hat{\tau}|$, with $f_\eta(t) = |t|$ and $\hat{\xi} = 1$. The randomization distribution for $T_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is not asymptotically sharp-dominant, such that employing the reference distribution assuming that the sharp null holds may lead to anti-conservative inference. The conventional fix is to studentize $\sqrt{N}|\hat{\tau}|$ using a variance estimator estimator satisfying Condition 4, forming instead $T_{Stu}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \sqrt{N}|\hat{\tau}|/\sqrt{\hat{V}_{\tau\tau}}$ (LRR17).

As $\phi(\cdot) = 1$ deterministically in a completely randomized design, Gaussian prepivoting via (5) yields the test statistic

$$G_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \gamma^{(1)}_{0,\hat{V}_{\tau\tau}}\{a : |a| \leq \sqrt{N}|\hat{\tau}|\} = 1 - 2\Phi\left(-\frac{\sqrt{N}|\hat{\tau}|}{\sqrt{\hat{V}_{\tau\tau}}}\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function. For any $\mathbf{Z}$, the pairs $\{G_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{w}), T_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{w})\}$ have rank correlation equal to 1 when computed for all $\mathbf{w} \in \Omega$. As a result, the reference distribution using the studentized difference in means assuming the sharp null will furnish identical p-values to those attained using Gaussian prepivoting. That is, in the univariate case Gaussian prepivoting is equivalent to studentization for completely randomized designs. This highlights objectives (i) and (ii).

Example 2 (Multivariate studentization)**.** Let $\sqrt{N}\hat{\boldsymbol{\tau}}$ now be multivariate and suppose we have a completely randomized design. (WD18) suggest the test statistic

$$T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right)^{\mathrm{T}} \hat{V}_{\tau\tau}^{-1} \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right), \tag{6}$$

with $\hat{V}_{\tau\tau} = \frac{N}{n_1}\hat{\Sigma}_{y(1)} + \frac{N}{n_0}\hat{\Sigma}_{y(0)}$. For this test statistic, $f_\eta(\mathbf{t}) = \mathbf{t}^{\mathrm{T}}\eta^{-1}\mathbf{t}$ and $\hat{\xi} = \hat{V}_{\tau\tau}$. (WD18) show that under our assumptions, under the weak null this test statistic converges in distribution to $\sum_{i=1}^{d} w_i\zeta_i^2$ where $w_i \in [0, 1]$ are weights and $\zeta_1, \ldots, \zeta_d \overset{iid}{\sim} \mathcal{N}(0, 1)$ while the reference distribution of $T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ attained assuming that the sharp null holds converges

weakly in probability to the χ^2_d-distribution. As a result, $T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is asymptotically sharp-dominant, and its reference distribution assuming the sharp null may be used for inference for the weak null hypothesis. Here, Gaussian prepivoting produces

$$G_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \gamma^{(d)}_{\mathbf{0}, \hat{V}_{\tau\tau}}\{\mathbf{a} : \mathbf{a}^{\mathrm{T}} \hat{V}^{-1}_{\tau\tau} \mathbf{a} \leq T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\} = F_d\{T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\},$$

where $F_d(\cdot)$ is the distribution function of a χ^2_d random variable. For any $\mathbf{Z}$, the pairs $\{G_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{w}), T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{w})\}$ have rank correlation equal to 1 when computed for all $\mathbf{w} \in \Omega$, such that Gaussian prepivoting yields equivalent inference to that attained using the distribution of $T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ under the sharp null. This demonstrates objective (ii).

Suppose instead that, erroneously, a practitioner proceeded with the more typical form of Hotelling's T-squared statistic employing a pooled covariance estimator,

$$T_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right)^{\mathrm{T}} \left(\hat{V}_{Pool}\right)^{-1} \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right);$$
$$\hat{V}_{Pool} = \left(\frac{N}{n_0} + \frac{N}{n_1}\right) \left(\frac{(n_1 - 1)\hat{\Sigma}_{y(1)} + (n_0 - 1)\hat{\Sigma}_{y(0)}}{n_1 + n_0 - 2}\right).$$

For this test statistic, $f_\eta(\mathbf{t}) = \mathbf{t}^{\mathrm{T}} \eta^{-1} \mathbf{t}$ as before, but $\hat{\xi} = \hat{V}_{Pool}$. In this case, $T_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is not asymptotically sharp-dominant, such that the reference distribution using this statistic and assuming the sharp null may yield invalid inference. Gaussian prepivoting returns the test statistic

$$G_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \gamma^{(d)}_{\mathbf{0}, \hat{V}_{\tau\tau}}\{\mathbf{a} : \mathbf{a}^{\mathrm{T}} \hat{V}^{-1}_{Pool} \mathbf{a} \leq T_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\}.$$

Importantly, $G_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ continues to use the Gaussian measure computed with the covariance matrix $\hat{V}_{\tau\tau}$ in forming the suitable transformation, despite the fact that the pooled covariance matrix is used in forming $T_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. For fixed $\mathbf{Z}$, $G_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{w})$ generally will not have perfect rank correlation with $T_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{w})$ when computed over $\mathbf{w} \in \Omega$,

such that the two randomization tests assuming the sharp null no longer furnish identical p-values. This divergence is necessary: while $T_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is not asymptotically sharp-dominant, Theorem 1 asserts that
$G_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is, such that the reference distribution for $G_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ assuming the sharp null yields asymptotically conservative inference for the weak null. Gaussian prepivoting can thus restore asymptotic validity to a test statistic employing improper studentization, illustrating objective (iii).

Example 3 (Max absolute t-statistic)**.** Consider again multivariate $\sqrt{N}\hat{\boldsymbol{\tau}}$ and a completely randomized design, and consider the test statistic

$$T_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \max_{1 \leq j \leq d} \frac{\sqrt{N}|\hat{\tau}_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}},$$

where $\hat{V}_{\tau\tau,jj}$ is the jj element of $\hat{V}_{\tau\tau}$. For this statistic, $f_{\boldsymbol{\eta}}(\mathbf{t}) = \max_{1 \leq j \leq d} |t_j|/\eta_j$, and $\hat{\boldsymbol{\xi}} = (\hat{V}_{\tau\tau,11}^{1/2}, ..., \hat{V}_{\tau\tau,dd}^{1/2})^{\mathrm{T}}$. For $d \geq 2$, $T_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is not asymptotically sharp-dominant under the weak null: the reference distribution generated under the sharp null depends upon the correlation matrix corresponding to $\tilde{V}$, while the true randomization distribution is governed by the correlations encoded within V. The Gaussian prepivoted correction takes the form

$$G_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \gamma^{(d)}_{\mathbf{0}, \hat{V}_{\tau\tau}} \left\{ \mathbf{a} : \max_{1 \leq j \leq d} \frac{|a_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}} \leq \max_{1 \leq j \leq d} \frac{\sqrt{N}|\hat{\tau}_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}} \right\},$$

which composes $T_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ with the distribution function for $\max |A_j|/\sqrt{\hat{V}_{\tau\tau,jj}}$, $j = 1, .., d$, when A is multivariate Gaussian with mean zero and covariance $\hat{V}_{\tau\tau}$. Gaussian prepivoting rectifies the insufficiency of the studentization in $T_{|max|}$, thereby providing an example of objective (iii).

Example 4 (Rerandomization)**.** Let $\sqrt{N}\hat{\tau}$ be univariate and suppose we now consider a

rerandomized design with balance criterion ϕ satisfying Condition 1. Consider the absolute difference in means, $f_{\hat{\xi}}(\sqrt{N}\hat{\tau}) = \sqrt{N}|\hat{\tau}|$, such that $\hat{\xi} = 1$. Gaussian prepivoting yields the test statistic

$$G_{Re}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \frac{\gamma^{(1+k)}_{\mathbf{0},\hat{V}} \left\{ (\mathbf{a}, \mathbf{b})^{\mathrm{T}} : |a| \leq \sqrt{N}|\hat{\tau}| \ \wedge \ \phi(\mathbf{b}) = 1 \right\}}{\gamma^{(k)}_{\mathbf{0},\hat{V}_{\delta\delta}} \{\mathbf{b} : \phi(\mathbf{b}) = 1\}}$$

For completely randomized designs with $\phi(\cdot) = 1$ deterministically, Gaussian prepivoting is equivalent to studentizing as described in Example 1. In general rerandomized designs however, observe that the transformation depends upon the particular form of the balance criterion ϕ, and that the reference distribution will depend upon the relationship between the potential outcomes and the covariates used in the balance criterion. As a result, it will generally not be the case that the reference distribution of $G_{Re}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ under the sharp null yields equivalent inference to that attained using $\sqrt{N}|\hat{\tau}|/\sqrt{\hat{V}_{\tau\tau}}$. This suggests that in rerandomized designs, studentization alone is insufficient for attaining an asymptotically sharp-dominant test statistic. In Section 3.8.1, we show this through an example in the case of Mahalanobis rerandomization. Lemmas A15 and A16 of (LDR18) show that under our conditions, probability limits for estimators $\hat{V}$ derived under complete randomization are generally preserved under rerandomized designs. Once again, Theorem 1 ensures that $G_{Re}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ will be asymptotically sharp-dominant, such that the randomization distribution assuming the sharp null may be employed for inference for the weak null. The development of a finite sample exact method for testing H_F which is asymptotically valid for testing H_N in rerandomized designs is novel, but its construction is extremely simple within the framework of Gaussian prepivoting; this highlights Gaussian prepivoting's portability to designs outside of just completely randomized experiments. In the supplementary materials we provide two more examples of this portability: one for matched-pair designs and one for experiments with any finite number of treatment arms. This highlights objective (iv).

For the interested reader, in the supplementary materials we include this same collection of examples written directly in the form of Gaussian integrals, and we include verification of Conditions 2-4.

3.6 Gaussian Comparison, Stochastic Dominance, And The Probability Integral Transform

3.6.1 Gaussian Comparison and Anderson's Theorem

We now highlight the essential technical ingredients underpinning the success of Gaussian prepivoting. Consider two mean-zero multivariate Gaussian vectors $(\mathbf{A}_1, \mathbf{B}_1)^{\mathrm{T}}$ and $(\mathbf{A}_2, \mathbf{B}_2)^{\mathrm{T}}$, with covariances

$$M_1 = \begin{pmatrix} \Lambda_{aa}^{(1)} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}; \quad M_2 = \begin{pmatrix} \Lambda_{aa}^{(2)} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb}, \end{pmatrix},$$

satisfying $\Lambda_{aa}^{(2)} - \Lambda_{aa}^{(1)} \succeq 0$ and $\Lambda_{bb} \succ 0$; the inequalities are stated with respect to the Loewner partial order on positive semidefinite matrices. Let the dimensions of $\mathbf{A}_j$ and $\mathbf{B}_j$ be d and k respectively for $j = 1, 2$. Compare the tail probabilities for

$$f(\mathbf{A}_1) \mid \phi(\mathbf{B}_1) = 1 \text{ and } f(\mathbf{A}_2) \mid \phi(\mathbf{B}_2) = 1,$$

where ϕ and f satisfy Conditions 1 and Condition 2 respectively. The following result is a straightforward corollary of Anderson's (1955) theorem for multivariate Gaussians (And55); see also Theorem 4.2.5 of (Ton90).

Lemma 1. *Under the stated conditions, for any scalar v,*

$$\mathbb{P}\left\{f(\mathbf{A}_1) \geq v \mid \phi(\mathbf{B}_1) = 1\right\} \leq \mathbb{P}\left\{f(\mathbf{A}_2) \geq v \mid \phi(\mathbf{B}_2) = 1\right\}.$$

The result follows immediately from Anderson's theorem after noting that the set $\mathcal{B}_v = \{(\mathbf{a}, \mathbf{b})^{\mathrm{T}} : f(\mathbf{a}) \leq v \ \wedge \ \phi(\mathbf{b}) = 1\}$ is convex and mirror-symmetric for any v. This can be seen through our assumption that $f(\cdot)$ is quasi-convex and mirror-symmetric, such that its sublevel sets are convex and mirror symmetric. We further have that $\mathbb{P}(\phi(\mathbf{B}_1) = 1) = \mathbb{P}(\phi(\mathbf{B}_2) = 1) > 0$ given the structure of the covariance matrices M_1 and M_2 and Condition 1, completing the proof.

3.6.2 Stochastic Dominance and the Probability Integral Transform

For two real valued random variables S and T, S *(first order) stochastically dominates* T if $F_S(a) \leq F_T(a)$ for all $a \in \mathbb{R}$, where F_S and F_T are the distribution functions of S and T respectively.

Suppose now that S and T are continuous and that S stochastically dominates T. By the probability integral transform, the distribution of $F_T(T)$ would be standard uniform. The following proposition considers transforming the random variable T not by its own distribution function, but rather by the distribution function of S, its stochastically dominating random variable.

Lemma 2. *Suppose that S, T are continuous random variables and that S stochastically dominates T. Then, $F_S(T)$ is stochastically dominated by a standard uniform random variable.*

Proof. For any $t \in [0, 1]$, $\mathbb{P}\{F_S(T) \leq t\} = \mathbb{P}\{T \leq F_S^{-1}(t)\} \geq \mathbb{P}\{S \leq F_S^{-1}(t)\} = t$. □

In the setup of Section 3.6.1, under Conditions 1 and 2 we have by Proposition 1 that $f(\mathbf{A}_2) \mid \phi(\mathbf{B}_2) = 1$ stochastically dominates $f(\mathbf{A}_1) \mid \phi(\mathbf{B}_1) = 1$. Consequently, composing

$f(\mathbf{A}_1) \mid \phi(\mathbf{B}_1) = 1$ with the distribution function of $f(\mathbf{A}_2) \mid \phi(\mathbf{B}_2) = 1$ would yield a random variable that is stochastically dominated by a standard uniform.

3.6.3 A Proof Sketch for Theorem 1

While a formal proof of Theorem 1 is deferred to the supplementary materials, here we provide an informal sketch in light of Lemmas 1 and 2. Under Assumptions 1 - 3 and Condition 1, $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}})$ converges in distribution to $\mathbf{A}_1 \mid \phi(\mathbf{B_1}) = 1$, where $(\mathbf{A}_1, \mathbf{B}_1)^{\mathrm{T}}$ are jointly multivariate normal with covariance V. Recall that $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}})$ for some f_η satisfying Condition 2 for all $\eta \in \Xi$, some $\hat{\xi}$ satisfying Condition 3, and with a balance criterion ϕ satisfying Condition 1. By Condition 3 and the assumption of the weak null, we have that $\hat{\xi}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in probability to ξ. Therefore, under the weak null, by Lemma 1 the limiting distribution of $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ would be stochastically dominated by that of $f_\xi(\mathbf{A}_2) \mid \phi(\mathbf{B}_2) = 1$ for any $(\mathbf{A}_2, \mathbf{B}_2)^{\mathrm{T}}$ multivariate Gaussian with covariance matrix

$$\bar{\bar{V}} = V + \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix}$$

with $\Delta \succeq 0$. The transformation

$$\bar{\bar{G}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \frac{\gamma^{(d+k)}_{\mathbf{0},\bar{\bar{V}}} \left\{ (\mathbf{a}, \mathbf{b})^{\mathrm{T}} : f_{\hat{\xi}}(a) \leq f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}}) \ \wedge \ \phi(\mathbf{b}) = 1 \right\}}{\gamma^{(k)}_{\mathbf{0},\bar{\bar{V}}_{\delta\delta}} \{ \mathbf{b} : \phi(\mathbf{b}) = 1 \}}$$

transforms $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ by the distribution function of a random variable which stochastically dominates its limiting distribution. By Lemma 2 and the continuous mapping theorem, asymptotically $\bar{\bar{G}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is stochastically dominated by a standard uniform. By Condition 4, the covariance estimator $\hat{V}$ used in forming $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ has a probability limit of the required form for stochastic dominance. Therefore, another application of the continuous mapping theorem yields that $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - \bar{\bar{G}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = o_p(1)$, such that by Slutsky's

Theorem $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is itself stochastically dominated by a standard uniform.

Meanwhile, Proposition 1 and Condition 1 yield that under the weak null the distribution of $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{W}) \mid \mathbf{Z}$ converges weakly in probability to the distribution of $\tilde{\mathbf{A}} \mid \phi(\tilde{\mathbf{B}}) = 1$, where $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})^{\mathrm{T}}$ are jointly multivariate Gaussian with mean zero and covariance $\tilde{V}$. The distribution of $f_{\hat{\xi}(\mathbf{y}(\mathbf{Z}),\mathbf{W})}\{\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{W})\} \mid \mathbf{Z}$ is precisely $\hat{\mathscr{P}}_T$, the reference distribution assuming the sharp null holds for the test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}})$. By Condition 4, $\hat{V}(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ converges in probability to $\tilde{V}$ itself. Further, by Condition 3 $\hat{\xi}(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ converges in probability to $\tilde{\xi}$. Applying the continuous mapping theorem and Slutsky's Theorem for randomization distributions (CR16, Lemmas A5-A6), one sees that Gaussian prepivoting furnishes a transformation that amounts to, asymptotically, an application of the probability integral transform. As a result, $\hat{\mathscr{P}}_G(t)$ converges in probability to t, the distribution function of the standard uniform, for all $t \in [0, 1]$.

3.7 Extensions To Asymptotically Linear Estimators

Theorem 1 may be extended to estimators other than the difference in means. Consider an estimator $\breve{\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ such that

$$\sqrt{N}\{\breve{\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - \bar{\boldsymbol{\tau}}\} = \sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} Z_i \mathbf{r}_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1 - Z_i)\mathbf{r}_i(Z_i)\right) + o_p(1)$$

for some constants $\{\mathbf{r}_i(0), \mathbf{r}_i(1)\}_{i=1}^{N}$ which may change with N and that satisfy

$$(1/N)\sum_{i=1}^{N}(\mathbf{r}_i(1) - \mathbf{r}_i(0)) = 0$$

along with Assumptions 2 and 3. Suppose further that $\breve{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$, $\mathbf{W}$ independent from $\mathbf{Z}$ and drawn uniformly from Ω, satisfies

$$\sqrt{N}\breve{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) = \sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} W_i \tilde{\mathbf{r}}_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1-W_i)\tilde{\mathbf{r}}_i(Z_i)\right) + o_p(1)$$

for potentially distinct constants $\{\tilde{\mathbf{r}}_i(0), \tilde{\mathbf{r}}_i(1)\}_{i=1}^N$ which may change with N that satisfy $(1/N)\sum_{i=1}^N(\tilde{\mathbf{r}}_i(1) - \tilde{\mathbf{r}}_i(0)) = 0$ along with Assumptions 2 and 3. Observe that the difference in means estimator satisfies these conditions with $\mathbf{r}_i(z) = \tilde{\mathbf{r}}_i(z) = \mathbf{y}_i(z) - z\bar{\boldsymbol{\tau}}$ for $z \in \{0,1\}$. Let $\boldsymbol{\tau}_{ri} = \mathbf{r}_i(1) - \mathbf{r}_i(0)$. Let $\Sigma_{r(z)}, \Sigma_{\tau_r}, \Sigma_{r(z)x}, \Sigma_{\tau_r x}$ be the analogues of $\Sigma_{y(z)}$, Σ_τ, $\Sigma_{y(z)x}$ and $\Sigma_{\tau x}$ for $z \in \{0,1\}$, and let the same hold with r replaced by $\tilde{r}$. Define $V^{(r)}$ and $\tilde{V}^{(\tilde{r})}$ as the analogues of V and $\tilde{V}$, computed now based upon $\mathbf{r}(z)$ and $\tilde{\mathbf{r}}(z)$ instead of $\mathbf{y}(z)$ and $\tilde{\mathbf{y}}(z)$ for $z \in \{0,1\}$.

Consider a test statistic for the weak null of the form $\breve{T}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = f_{\hat{\xi}}(\sqrt{N}\breve{\tau})$ for some f_η satisfying Condition 2 and $\hat{\xi}$ satisfying Condition 3, and suppose that there exists a covariance estimator $\breve{V}$ satisfying Condition 4 with V and $\tilde{V}$ replaced by $V^{(r)}$ and $\tilde{V}^{(r)}$. The Gaussian prepivoted test statistic is

$$\breve{G}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \frac{\gamma^{(d+k)}_{\mathbf{0},\breve{V}}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : f_{\hat{\xi}}(\mathbf{a}) \leq \breve{T}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma^{(k)}_{\mathbf{0},\breve{V}_{\delta\delta}}\{\mathbf{b} : \phi(\mathbf{b}) = 1\}}$$

Theorem 2. *Suppose that Neyman's null, $H_N : \bar{\boldsymbol{\tau}} = 0$, holds. Then, under the described restrictions on $\breve{T}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and $\breve{V}$ and under Assumption 1 and with Assumptions 2 and 3 applied to $\mathbf{r}_i(z)$, $z = \{0,1\}$, $\breve{G}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in distribution to a random variable $\breve{U}$ taking values in $[0,1]$ satisfying*

$$\mathbb{P}(\breve{U} \leq t) \geq t,$$

for all $t \in [0,1]$. Furthermore, the distribution $\hat{\mathscr{P}}_{\breve{G}}(t)$ satisfies $\hat{\mathscr{P}}_{\breve{G}}(t) \xrightarrow{p} t$ for all $t \in [0,1]$.

In the supplementary materials, we illustrate that the regression-adjusted average treatment effect estimator and its corresponding estimated variance presented in (Lin13) can be viewed in this form. As a result, Theorem 2 provides justification for the use of the prepivoted randomization distribution of a regression-adjusted estimator.

3.8 Simulation Studies

3.8.1 Studentization and Prepivoting in Rerandomized Designs

In the bth of B iterations, we draw, for $i = 1, ..., N$, covariates iid as

$$\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}\left(0, \begin{pmatrix} 1.0 & 0.8 & 0.2 \\ 0.8 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix}\right).$$

Given these covariates, we draw $r_i(0)$ and $r_i(1)$ as

$$r_i(0) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \epsilon_i(0); \;\; r_i(1) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \epsilon_i(1),$$

where $\beta_0 = -(6.4, -4.0, -2.4)$, $\beta_1 = c(0.2, 0.4, 0.6)^{\mathrm{T}}$, $\epsilon_i(0) \overset{iid}{\sim} -\mathcal{E}(1) + 1$, $\epsilon_i(1) \overset{iid}{\sim} -\mathcal{E}(1/10) + 10$, $\epsilon_i(0)$ independent of $\epsilon_i(1)$, and $\mathcal{E}(\lambda)$ representing an exponential distribution with rate λ.

We form the potential outcomes under treatment and control in two distinct ways, one in which the sharp null holds and one in which only the weak null holds:

Sharp Null: $y_i(1) = y_i(0) = r_i(1)$

Weak Null: $y_i(1) = r_i(1)$; $y_i(0) = r_i(0) + \bar{r}(1) - \bar{r}(0)$

Of the N individuals, $n_1 = 0.2N$ receive the treatment and $n_0 = 0.8N$ receive the control. We

use a Mahalanobis-based rerandomized design, with criterion $\phi(\sqrt{N}\hat{\boldsymbol{\delta}}) = \mathbb{1}_{\left\{(\sqrt{N}\hat{\boldsymbol{\delta}})^{\mathrm{T}} V_{\delta\delta}^{-1} (\sqrt{N}\hat{\boldsymbol{\delta}}) \leq 1\right\}}$. This balance criterion reduces the cardinality of Ω by roughly 80% relative to a completely randomized design. For each b, we draw a single $\mathbf{Z} \in \Omega$, and proceed with inference using the reference distribution of the following test statistics under the incorrect assumption that the sharp null holds:

1. Absolute difference in means, unstudentized

2. Absolute difference in means, studentized

3. Gaussian prepivoting the absolute difference in means, studentized

The true reference distributions assuming the sharp null are replaced by Monte-Carlo estimates with 1000 draws from Ω for each b, and the desired Type I error rate is $\alpha = 0.05$. We also perform inference using the large-sample reference distribution for the absolute studentized difference in means in a rerandomized design; see (LDR18) for more details. As a covariance estimator $\hat{V}$, we use the conventional unpooled covariance estimator for $(\sqrt{N}\hat{\tau}, \sqrt{N}\hat{\boldsymbol{\delta}})^{\mathrm{T}}$ in a two-sample design. For the generative models reflecting the sharp and weak nulls, we proceed with both $N = 50$ and $N = 1000$ to compare performance in small and large sample regimes. For each N, we conduct $B = 5000$ simulations.

Table 3.1 contains the results of the simulation study. Under the sharp null with $N = 50$, we see the benefits of using a randomization test: the randomization tests based upon the unstudentized, studentized, and prepivoted absolute difference in means all resulted in a Type I error rate of 0.05 (up to noise from the Monte-Carlo simulation) as desired. Contrast this with the large-sample test, which had an estimated Type I error rate of 0.110 under the sharp null hypothesis. Figure 3-1 explains the deficiency of the large-sample test by comparing the true distribution for the large-sample p-values to the standard uniform distribution. As is seen, at $N = 50$ small p-values are more likely to occur than what the standard uniform would predict at any point $t \in [0, 1]$, resulting in inflated Type I error rates. By $N = 1000$, the

Table 3.1: Inference after rerandomization. The rows describe the simulation settings, which vary between the sharp and weak nulls holding and between small and large sample sizes. The first three columns represent the performance of randomization tests assuming the sharp null hypothesis and using the unstudentized absolute difference in means, absolute studentized difference in means, and Gaussian prepivoted absolute difference in means respectively to perform inference. The last column is a large-sample test which is asymptotically valid for the weak null, based upon (LDR18). The desired Type I error rate in all settings is $\alpha = 0.05$.

	Randomization Test			Large-Sample
	No Stu.	Stu.	Pre.	
Sharp, $N = 50$	0.053	0.050	0.051	0.110
Sharp, $N = 1000$	0.052	0.048	0.048	0.054
Weak, $N = 50$	0.073	0.114	0.037	0.058
Weak, $N = 1000$	0.070	0.083	0.018	0.019

asymptotic approximation performs much better, as the true distribution of p-values lies on top of the standard uniform. Gaussian prepivoting uses 1 minus these large-sample p-values as the test statistic whose randomization distribution is enumerated, such that the solid line in Figure 3-1 reflects 1 minus the randomization distribution of the Gaussian prepivoted test statistic. As Gaussian prepivoting uses a randomization test under the sharp null, the solid line also reflects the reference distribution employed for performing inference. That these coincide is a consequence of the sharp null holding, such that the randomization tests are exact tests for any sample size.

Under the weak null, we see in Table 3.1 that even at $N = 1000$, the unstudentized and studentized randomization tests erroneously assuming the sharp null have inflated Type I error rates. This pattern will persist even asymptotically, as in this simulation setup these test statistics are not asymptotically sharp-dominant. This may come as a surprise, as in completely randomized designs studentizing *does* furnish asymptotic sharp dominance. As evidenced here, the impact of covariates on the limiting distribution in rerandomized experiments invalidates studentization as a mechanism for attaining asymptotic sharp dominance. Figure 3-2 illustrates this in the case of the studentized test statistic. We see in the top panel

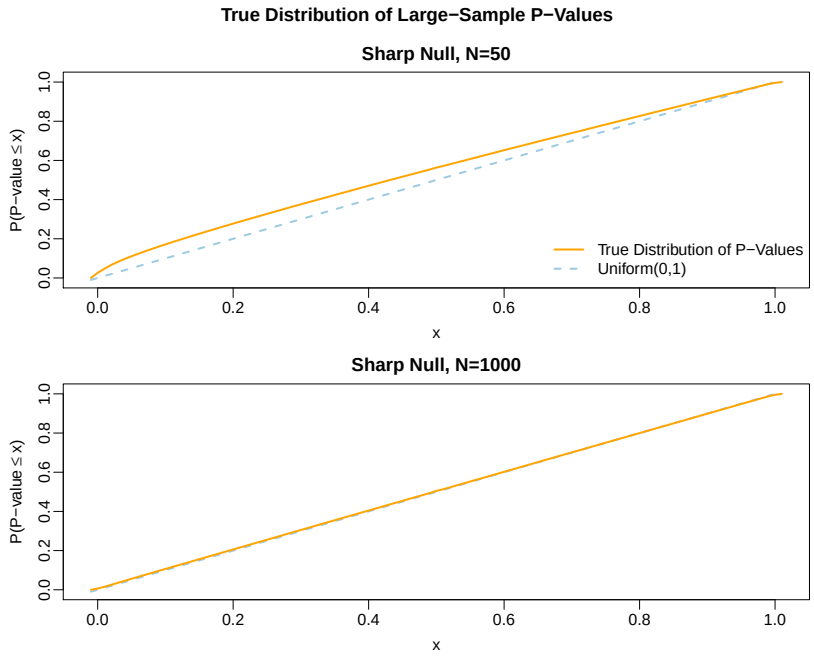

Figure 3-1: Randomization distribution of the large-sample p-values under the sharp null (solid) compared to a standard uniform distribution (dashed) at $N = 50$ (top) and $N = 1000$ (bottom). At $N = 50$, it is more likely to observe a small P-value than what the uniform distribution would suggest, yielding the inflated Type I error rate.

that the true distribution function for the studentized test statistic lies below that of the reference distribution assuming the sharp null, such that the right-tail probabilities are larger for the true randomization distribution than they are for the reference distribution. This yields anti-conservative inference. We see in the bottom panel of Figure 3-2 that through use of Gaussian prepivoting, asymptotic conservativeness has been restored: the true randomization distribution of the prepivoted test statistic is stochastically dominated by the reference distribution assuming the sharp null, as predicted by Theorem 1. We further see that the cumulative distribution assuming the sharp null is converging to the distribution function of the standard uniform (a straight line between 0 and 1), again reflecting Theorem 1. Table 3.1 further shows that the Gaussian prepivoted test and the large-sample test have very similar rejection rates at $N = 1000$, reflecting the asymptotic equivalence of the two

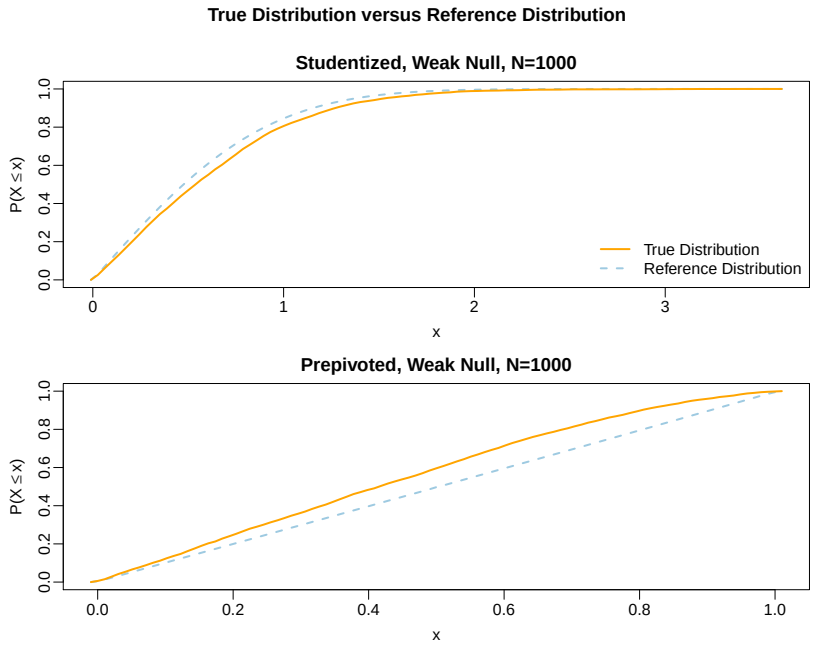

Figure 3-2: True randomization distribution under the weak null (solid) versus the reference distribution assuming the sharp null (dashed) for the studentized (top) and Gaussian prepivoted (bottom) test statistics with a rerandomized design. To yield valid randomization tests under the weak null, the solid line needs to lie above the dotted line, such that the solid line attributes less mass in the right tail than the dotted line does

methods under the weak null.

3.8.2 A Comparison of Multivariate Tests

In each iteration $b = 1, ..., B$, we draw $\{\mathbf{r}_i(1)\}_{i=1}^N$ and $\{\mathbf{r}_i(0)\}_{i=1}^N$ independent from one another and *iid* from mean zero equicorrelated multivariate normals of dimension $k = 25$ with marginal variances one. The correlation coefficients governing $\mathbf{r}_i(1)$ and $\mathbf{r}_i(0)$ are 0 and 0.95 respectively. We will have two simulation settings, one each for the sharp and weak null:

Sharp Null: $\mathbf{y}_i(1) = \mathbf{y}_i(0) = \mathbf{r}_i(1)$.

Weak Null: $\mathbf{y}_i(1) = \mathbf{r}_i(1)$; $\mathbf{y}_i(0) = \mathbf{r}_i(0) + \bar{\mathbf{r}}(1) - \bar{\mathbf{r}}(0)$.

In both settings, $n_1 = 0.2N$ individuals receive the treatment and $n_0 = 0.8N$ receive the control. We consider a completely randomized design, and proceed with inference using the reference distribution of the following test statistics under the (erroneous) assumption that the sharp null holds:

1. Hotelling's T-squared, unpooled covariance

2. Hotelling's T-squared, pooled covariance

3. Max absolute t-statistic, unpooled standard error

For each candidate test, we proceed with the randomization distribution both of the untransformed test statistic and the Gaussian prepivoted test statistic. These tests are conducted using Monte-carlo simulation to generate the reference distributions, with 1000 draws from Ω for each iteration b. In addition to the two types of randomization tests, we also compute a large-sample p-value for each test which is asymptotically valid under the weak null hypothesis. As a covariance estimator $\hat{V}$, we use the conventional unpooled covariance estimator for $\sqrt{N}\hat{\boldsymbol{\tau}}$. For each test, we seek to maintain the Type I error rate at or below $\alpha = 0.05$. For the generative models reflecting the sharp and weak nulls we proceed with both $N = 300$ and $N = 5000$ to compare performance as N increases. For each N, we conduct $B = 5000$ simulations.

Table 3.2 gives the estimated Type I error rates for the candidate tests. We first note the poor performance of the large-sample tests under both the sharp and weak null with $N = 300$. For instance, the large-sample p-values constructed using the unpooled, Hotelling procedure are attained using a χ^2_{25} distribution and have estimated Type I error rates of 0.321 under the sharp null for $N = 300$, and of 0.270 under the weak null for $N = 300$ despite the desired control at $\alpha = 0.05$. By $N = 5000$, the large-sample tests all have estimated Type I

Table 3.2: Inference in completely randomized designs with multiple outcomes. The rows describe the simulation settings, which vary between the sharp and weak nulls holding and between small and large sample sizes. There are three sets of columns, one corresponding to each of the three test statistics under consideration. For each set of columns, the column labeled "FRT" represents the Fisher randomization test using that test statistic. The column labeled "Pre." instead reflects the Fisher randomization test after applying Gaussian prepivoting to the original test statistic. The last column, labeled "LS," is a large-sample test which is asymptotically valid for the weak null. The desired Type I error rate in all settings is $\alpha = 0.05$.

	Hotelling, Unpooled			Hotelling, Pooled			Max t-stat		
	FRT	Pre.	LS	FRT	Pre.	LS	FRT	Pre.	LS
H_F, $N = 300$	0.050	0.050	0.321	0.052	0.047	0.086	0.051	0.050	0.068
H_F, $N = 5000$	0.044	0.044	0.053	0.047	0.042	0.045	0.046	0.045	0.048
H_N, $N = 300$	0.117	0.117	0.270	0.975	0.166	0.157	0.020	0.006	0.008
H_N, $N = 5000$	0.003	0.003	0.003	0.951	0.005	0.005	0.021	0.005	0.005

error rates approaching the nominal level under the sharp null, and below the nominal level under the weak null.

Naturally, all randomization tests attain (up to Monte-Carlo error) the desired Type I error rate under the sharp null at both $N = 300$ and $N = 5000$, highlighting the appeal of the randomization tests. Under the weak null, we see that the randomization test based upon the Hotelling T-statistic with a pooled covariance fails to control the Type I error rate even at $N = 5000$, reflecting that the test statistic is not asymptotically sharp-dominant. While the randomization test based on the max t-statistic controls the Type I error rate in these simulations, this is not guaranteed in general: in the supplementary materials we conduct this simulation at $\alpha = 0.25$, where anti-conservativeness of the max t-statistic arises. For both of these test statistics, applying Gaussian prepivoting restores guaranteed asymptotic conservativeness and results in test statistics whose performance closely aligns with the large-sample tests, a reflection of Theorem 1. For the test based upon Hotelling's T statistic with an unpooled covariance estimator, observe that the Type I error rates for the randomization tests with and without Gaussian prepivoting are identical in all four scenarios

tested. As discussed in Example 2 of Section 3.5.2, this is because Gaussian prepivoting is unnecessary for this particular test statistic: Hotelling's T statistic with an unpooled covariance estimator is already asymptotically sharp-dominant as proven in (WD18). Applying Gaussian prepivoting recovers an equivalent randomization test, furnishing identical p-values for any observed outcomes $\mathbf{y}(\mathbf{Z})$ for completely randomized designs.

In the supplementary materials we provide a theoretical analysis of the statistical power of Gaussian prepivoting and include simulations to demonstrate the power in practice. We also provide analysis of real-world data from the Student Achievement and Retention experiment of (ALO09).

3.9 Discussion

3.9.1 An Open Question: Multivariate One-sided Testing in Finite Population Causal Inference

The restrictions on the function f_η outlined in Condition 2 require a quasi-convex, continuous function that is mirror-symmetric about the origin. This restriction results in convex, mirror-symmetric sublevel sets for f_η and facilitates the application of Anderson's theorem, such that dominance in the Loewner order on covariance matrices translates to the stochastic dominance under the weak null. While the restrictions on f_η are sensible with two-sided alternatives, they preclude testing directional alternatives because of the mirror symmetry condition. For instance, suppose one wanted to test the null hypothesis $\bar{\tau}_i \leq 0$ for all $i = 1, .., d$ versus the alternative that for at least one i $(i = 1, .., d)$, $\bar{\tau}_i > 0$. In the univariate case, choosing $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \hat{\tau}/\hat{V}_{\tau\tau}^{1/2}$ does not provide a valid one-sided test for all α. That said, it *does* provide a valid test for $\alpha \leq 0.5$, such that for any reasonable value for α to be deployed in practice a one-sided test is possible.

Suppose we have multivariate potential outcomes and consider the test statistic

$$T_{max}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \max_{1 \leq i \leq d} \ \hat{\tau}_i / \hat{V}_{\tau\tau,ii}^{1/2},$$

with $\hat{V}_{\tau\tau}$ satisfying Condition 4. Consider the Gaussian prepivoted test statistic $G_{max}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. The following is, to the best of our knowledge, an open question: is it the case that, for any $\alpha \leq 0.5$, G_{max} is asymptotically sharp-dominant, in that $\limsup \ \mathbb{E}\{\varphi_{G_{max}}(\alpha)\} \leq \alpha$? Under the assumptions imposed in this work, the answer would be true should the following conjecture on Gaussian comparisons hold:

Conjecture 1. *Let* $\mathbf{X} = (X_1, ..., X_d)$*, and* $\mathbf{Y} = (Y_1, ..., Y_d)$ *be d-dimensional multivariate Gaussian vectors, with a common mean* $\boldsymbol{\mu} = (\mu_1, ..., \mu_d)$ *but distinct covariances* Σ^X *and* Σ^Y*, with ij entries* σ_{ij}^X *and* σ_{ij}^Y*, respectively. Let* $\gamma_{ij}^X = \mathbb{E}\{(X_i - X_j)^2\}$ *and* $\gamma_{ij}^Y = \mathbb{E}\{(Y_i - Y_j)^2\}$*. Define med* $(\max_i Y_i)$ *as the median of* $\max_{1 \leq i \leq d} Y_i$*, i.e. the value a such that* $\mathbb{P}\left(\max_{1 \leq i \leq d} Y_i \leq a\right) = 0.5$*. Suppose that* $\sigma_{ii}^Y \geq \sigma_{ii}^X$ *for all i and that* $\gamma_{ij}^Y \geq \gamma_{ij}^X$ *for all i, j. Consider any point* $c \geq$ *med* $(\max_i Y_i)$*. Then,*

$$\mathbb{P}\left(\max_{1 \leq i \leq d} X_i \geq c\right) \leq (?) \ \mathbb{P}\left(\max_{1 \leq i \leq d} Y_i \geq c\right).$$

The conjecture is true in the univariate case. Under the assumptions of this conjecture, the Sudakov-Fernique inequality (AT09, Theorem 2.2.5) asserts that

$$\mathbb{E}\{\max_{1 \leq i \leq d} X_i\} \leq \mathbb{E}\{\max_{1 \leq i \leq d} Y_i\}.$$

Should we further assume $\sigma_{ii}^X = \sigma_{ii}^Y$, the result holds for all points c through Slepian's lemma (Sle62, Ton90, Theorem 5.1.7). Unfortunately, a refined result about tail probabilities above the median does not appear to be available in the literature under the conditions outlined in the conjecture. A potential path forward may be a modification of the soft-max proof of

the Sudakov-Fernique inequality found in (Cha05).

3.9.2 Summary

In this work, we present a general framework for designing randomization tests that are both exact for Fisher's sharp null and are asymptotically conservative for Neyman's weak null in completely randomized experiments and rerandomized designs. Loosely stated, the approach may be summarized as follows: if one has access to a large-sample test that is asymptotically conservative under Neyman's weak null, then a Fisher randomization test using the p-value produced by that large-sample test will maintain asymptotic correctness under the weak null while additionally restoring exactness should the sharp null be true. As the Fisher randomization distribution of these p-values converges weakly in probability to a uniform, the resulting randomization test assuming the sharp will have the same large-sample performance under the weak null as large-sample test itself, and will further have the same asymptotic power under local alternatives as the large-sample test. We show that Gaussian prepivoting exactly recovers several randomization tests known to be valid under the weak null, while providing a general approach to restore asymptotic correctness to randomization tests for a large class of test statistics. Importantly, our framework immediately provides valid randomization tests of the weak null hypothesis in rerandomized designs, absent from the literature until now.

Supplementary Material

Below we include additional information which contains theoretical results, proofs, simulation studies, further algorithmic details, discussion of statistical power and confidence intervals, and a real-world data example. We also provide reference to an `R` script for implementing the method proposed in this work.

3.10 Useful Lemmas

Lemma A.3. *For any Borel measurable set $B \subseteq \mathbb{R}^\ell$, the centered Gaussian measure of B is a continuous function in terms of the covariance parameter. In other words, $\gamma^\ell_{\mathbf{0},\Sigma}(B)$ is a continuous function of Σ over the positive definite cone of $\ell \times \ell$ real matrices with metric induced by the Frobenius norm.*

Proof. Denote the space of positive definite real $\ell \times \ell$ matrices by S^ℓ_{++}; this is a metric space under the metric induced by the Frobenius norm. Consider a sequence of matrices $\Sigma_N \in S^\ell_{++}$ for which $\Sigma_N \to \Sigma$. By definition for any Borel measurable set $B \subseteq \mathbb{R}^\ell$

$$\gamma^\ell_{\mathbf{0},\Sigma_N}(B) = \int_B \frac{1}{\sqrt{2\pi}^\ell} \frac{1}{\sqrt{\det(\Sigma_N)}} \exp\left(\frac{-\mathbf{x}^T \Sigma_N^{-1} \mathbf{x}}{2}\right) d\mathbf{x}.$$

The function $f(M) = \det(M)^{-1/2}$ is continuous over the positive definite cone of $\ell \times \ell$ matrices. Thus, since $\Sigma_N \to \Sigma$ it follows that

$$\frac{1}{\sqrt{2\pi}^\ell} \frac{1}{\sqrt{\det(\Sigma_N)}} \to \frac{1}{\sqrt{2\pi}^\ell} \frac{1}{\sqrt{\det(\Sigma)}}. \tag{7}$$

All that remains to be examined is the limiting behavior of

$$\int_B \exp\left(\frac{-\mathbf{x}^{\mathrm{T}}\Sigma_N^{-1}\mathbf{x}}{2}\right) d\mathbf{x}.$$

For $(M, \mathbf{x}) \in S_{++}^{\ell} \times \mathbb{R}^{\ell}$ the function $g(M, \mathbf{x}) = \exp(-\mathbf{x}^{\mathrm{T}}M^{-1}\mathbf{x}/2)$ is a jointly continuous of both $\mathbf{x}$ and M. Consequently, for all $\mathbf{x} \in \mathbb{R}^{\ell}$

$$\exp\left(\frac{-\mathbf{x}^{\mathrm{T}}\Sigma_N^{-1}\mathbf{x}}{2}\right) \to \exp\left(\frac{-\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x}}{2}\right).$$

Since all convergent sequences are bounded there exits a positive semidefinite matrix Σ_* that is greater than or equal to (in the Loewner partial order) all Σ_N. Thus, $\Sigma_N^{-1} \succeq \Sigma_*^{-1}$ for all $N \in \mathbb{N}$. Consequently $g(\Sigma_N, \mathbf{x})$ is dominated by $g(\Sigma_*, \mathbf{x})$ for all N and all $\mathbf{x} \in \mathbb{R}^{\ell}$. Thus, Lebesgue's dominated convergence theorem implies that

$$\int_B \exp\left(\frac{-\mathbf{x}^{\mathrm{T}}\Sigma_N^{-1}\mathbf{x}}{2}\right) d\mathbf{x} \to \int_B \exp\left(\frac{-\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x}}{2}\right) d\mathbf{x}. \tag{8}$$

Combining (7) and (8) implies that for all sequences $\Sigma_N \in S_{++}^{\ell}$ such that $\Sigma_N \to \Sigma$

$$\gamma_{\mathbf{0},\Sigma_N}^{\ell}(B) \to \gamma_{\mathbf{0},\Sigma}^{\ell}(B). \tag{9}$$

(9) establishes that $\gamma_{\mathbf{0},\Sigma}^{\ell}(B)$ is a sequentially continuous function of the parameter Σ for all $\Sigma \in S_{++}^{\ell}$. Sequential continuity in a metric space is equivalent to continuity (GM07, Theorem 5.31); so $\gamma_{\mathbf{0},\Sigma}^{\ell}(B)$ is a continuous function of the parameter Σ for all $\Sigma \in S_{++}^{\ell}$.

□

Lemma A.4. *Let a function $f_\eta(\cdot)$ satisfy Condition 6 and let $\phi(\cdot)$ satisfy Condition 5. Let the matrix $V \in S_{++}^{(d+k)}$ be defined blockwise as*

$$V = \begin{pmatrix} V_{\tau\tau} & V_{\tau\delta} \\ V_{\delta\tau} & V_{\delta\delta} \end{pmatrix}.$$

Consider

$$h(V, \eta, x) = \frac{\gamma_{\mathbf{0},V}^{(d+k)} \left\{(\mathbf{a}, \mathbf{b})^{\mathrm{T}} : f_\eta(\mathbf{a}) \leq x \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma_{\mathbf{0},V_{\delta\delta}}^{(k)} \left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}.$$

The function $h(V, \eta, x)$ is a continuous function of V, η, and x jointly.

Proof. Because V is positive definite, $V_{\delta\delta}$ must be as well. Thus, the centered Gaussian measure $\gamma_{\mathbf{0},V_{\delta\delta}}^{(k)}(\cdot)$ is non-singular. Furthermore, because $\phi(\cdot)$ satisfies Condition 5, the set $\{\mathbf{b} : \phi(\mathbf{b}) = 1\}$ is Borel measurable with positive Lebesgue measure. Thus, $\gamma_{\mathbf{0},V_{\delta\delta}}^{(k)} \{\mathbf{b} : \phi(\mathbf{b}) = 1\}$ is positive. Moreover, Lemma A.3 establishes that $\gamma_{\mathbf{0},V_{\delta\delta}}^{(k)} \{\mathbf{b} : \phi(\mathbf{b}) = 1\}$ is a continuous function of $V_{\delta\delta}$, and thus of V.

Consider the function $\kappa : (\eta, x) \mapsto \{(\mathbf{a}, \mathbf{b})^{\mathrm{T}} : f_\eta(\mathbf{a}) \leq x \ \wedge \ \phi(\mathbf{b}) = 1\}$. The range of κ is the set of Borel measurable sets in $\mathbb{R}^{(d+k)}$. This space can be imbued with the metric[1]

$$d(B, B') = \mu(B \triangledown B')$$

where $B \triangledown B'$ is the symmetric difference of B and B' and $\mu(\cdot)$ is Lebesgue measure on $\mathbb{R}^{(d+k)}$; this is sometimes called the *Fréchet–Nikodým–Aronszajn distance* (CK17, Section 4). Consider sequences of η_N which converge to η and x_N which converge to x. Let B_N denote $\kappa(\eta_N, x_N)$; the set-theoretic limit of B_N converges to B under $d(B, B')$. This relies upon the continuity of $f_\eta(\mathbf{a})$ in η. Thus, κ is sequentially continuous in η and x jointly. Sequential continuity in a metric space is equivalent to continuity (GM07, Theorem 5.31);

[1] Actually, $d(B, B')$ is a pseudo-metric unless one considers two sets equal if their symmetric difference is of measure zero. We take this convention since – by absolute continuity – sets of Lebesgue measure zero are of Gaussian measure zero as well.

so κ is jointly continuous in η and x.

The numerator of $h(V, \eta, x)$ is the composition of $\gamma_{\mathbf{0},V}^{(d+k)}(B)$ with $\kappa(\eta, x)$; the former is continuous in V by Lemma A.3 and in B by the absolute continuity of Gaussian measure, and the later is jointly continuous in η and x. Thus, the numerator of $h(V, \eta, x)$ is jointly continuous in V, η, and x. Since the denominator of $h(V, \eta, x)$ is a continuous function of V that is always positive, the function $h(V, \eta, x)$ itself is a jointly continuous function of V, η, and x. □

3.11 Proof Of Main Results

3.11.1 A Reminder: Assumptions and Conditions

We rely on some regularity conditions from above which we restate below for convenience.

Assumption 1. *The proportion n_1/N limits to $p \in (0, 1)$ as $N \to \infty$.*

Assumption 2. *All finite population means and covariances have limiting values for both the potential outcomes and the covariates. For instance,* $\lim_{N\to\infty} \bar{\mathbf{y}}(z) = \bar{\mathbf{y}}_\infty(z)$ *for* $z \in \{0, 1\}$ *and* $\lim_{N\to\infty} \Sigma_{y(1)} = \Sigma_{y(1),\infty}$.

Assumption 3 of Section 3.2.3 is known to be stronger than necessary for certain results. Here we split Assumption 3 into two parts. We do this to show exactly which results can rely upon a weaker assumption and which results seem to rely upon the stronger assumption.

Assumption 3(a). *The worst-case squared distance from the average potential outcome is $o(N)$; i.e.,*

$$\lim_{N\to\infty} \max_{\substack{z\in\{0,1\} \\ j\in\{1,\ldots,d\}}} \max_{i\in\{1,\ldots,N\}} \frac{(y_{ij}(z) - \bar{y}_j(z))^2}{N} = 0.$$

Further, the above holds for the covariates with x_{ij} replacing $y_{ij}(z)$ above for $j = 1, .., k$.

Assumption 3(b). *There exists some $C < \infty$ for which, for all $z \in \{0,1\}$, all $j = 1,..,d$ and all N,*

$$\frac{\sum_{i=1}^{N} \left(y_{ij}(z) - \overline{y}_j(z)\right)^4}{N} < C$$

Further, the above holds for the covariates with x_{ij} replacing $y_{ij}(z)$ above for $j = 1,..,k$.

Assumption 3(b) implies Assumption 3(a) (WD18, Proposition 1). Assumption 3(b) is made at times for mathematical convenience to simplify the analysis of certain random distributions; though it remains an open question whether such results hold under weaker assumptions.

Recall the $\tilde{\mathbf{y}}(Z_i)$ is defined as

$$\tilde{\mathbf{y}}_i(Z_i) = \mathbf{y}_i(Z_i) - Z_i\bar{\boldsymbol{\tau}},$$

such that $\tilde{\mathbf{y}}(\mathbf{Z}) = \mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\boldsymbol{\tau}}^{\mathrm{T}}$. Further recall the following conditions from Section 3.2.3.

Condition 5. *$\phi : \mathbb{R}^k \mapsto \{0,1\}$ is an indicator function such that the set $M = \{\mathbf{b} : \phi(\mathbf{b}) = 1\}$ is closed, convex, and mirror-symmetric about the origin (i.e., $\mathbf{b} \in M \Leftrightarrow -\mathbf{b} \in M$) with non-empty interior.*

Condition 6. *For any $\eta \in \Xi$, $f_\eta(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}_+$ is continuous, quasi-convex, and nonnegative with $f_\eta(\mathbf{t}) = f_\eta(-\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^d$. Furthermore, $f_\eta(\mathbf{t})$ is jointly continuous in η and $\mathbf{t}$.*

Condition 7. *With $\mathbf{W}, \mathbf{Z}$ independent and each uniformly distributed over Ω,*

$$\hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} \xi; \;\; \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \xrightarrow{p} \tilde{\xi},$$

for some $\xi, \tilde{\xi} \in \Xi$.

Condition 8. *With* $\mathbf{W}, \mathbf{Z}$ *independent, both uniformly distributed over* Ω*, and for some* $\Delta \succeq 0$*,* $\Delta \in \mathbb{R}^{d\times d}$*,*

$$\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) - V \xrightarrow{p} \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix}; \quad \hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) - \tilde{V} \xrightarrow{p} 0_{(d+k),(d+k)}.$$

Oftentimes in the proofs it will implicitly be assumed that the weak null holds. For that reason, $\hat{\xi}$ and $\hat{V}$ may be written with $\mathbf{y}(\mathbf{Z})$ as inputs rather than $\tilde{\mathbf{y}}(\mathbf{Z})$. Let Ω_{CRE} denote the set of allowable treatment allocation vectors $\mathbf{z}$ for a completely randomized experiment. Formally

$$\Omega_{CRE} = \left\{ \mathbf{z} \in \{0,1\}^N \;\middle|\; \sum_{i=1}^{N} z_i = n_1 \right\}.$$

3.11.2 A Remark on Limiting Distributions for Rerandomized Designs

A completely randomized experiment can be considered a rerandomized experiment for which $\phi(\cdot)$ is identically one. This trivial balance criterion satisfies Condition 5.[2] When $\phi(\cdot)$ is not vacuous, the interesting case for rerandomized designs, limiting distributions in completely randomized designs continue to provide corresponding limiting distributions after rerandomization under Condition 5.

By the finite population central limit theorem of (LD17), $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})^{\mathrm{T}}$ is asymptotically distributed according to a mean-zero multivariate Gaussian distribution with covariance

[2]When no covariate information is collected, this statement is then vacuous, but in such a context the comparison to a rerandomized experiment is also missing.

matrix V, where

$$\begin{aligned}
V &= \begin{pmatrix} V_{\tau\tau} & V_{\tau\delta} \\ V_{\delta\tau} & V_{\delta\delta} \end{pmatrix}; \\
V_{\tau\tau} &= p^{-1}\Sigma_{y(1),\infty} + (1-p)^{-1}\Sigma_{y(0),\infty} - \Sigma_{\tau,\infty}; \\
V_{\delta\delta} &= \{p(1-p)\}^{-1}\Sigma_{x,\infty}; \\
V_{\tau\delta} &= p^{-1}\Sigma_{y(1)x,\infty} + (1-p)^{-1}\Sigma_{y(0)x,\infty} = V_{\delta\tau}^{\mathrm{T}}.
\end{aligned}$$

Conditioning according to appropriate balance holding requires that $V_{\delta\delta} \succ 0$. In this case, the conditional probability of $\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) \in B$ subject to $\phi(\sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{Z})) = 1$ limits to

$$\frac{\gamma_{\mathbf{0},V}^{(d+k)}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : \mathbf{a} \in B \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma_{\mathbf{0},V_{\delta\delta}}^{(k)}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}} \tag{10}$$

for any Borel measurable set B.

Likewise, by Proposition1 and Lemma 4.1 of (DDCZ13), the conditional probability of $\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \in B$ subject to $\phi(\sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W})) = 1$ limits to

$$\frac{\gamma_{\mathbf{0},\tilde{V}}^{(d+k)}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : \mathbf{a} \in B \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma_{\mathbf{0},\tilde{V}_{\delta\delta}}^{(k)}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}. \tag{11}$$

The finite population central limit theorem of (LD17) and Proposition 1 are statements about joint convergence in distribution for the scaled differences in means for the observed outcomes and for the covariates. Passing to convergence in distribution conditional upon $\phi(\sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{Z})) = 1$ or $\phi(\sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W})) = 1$ described in (10) and (11) rests upon the continuity-set argument used in the proof of Proposition A1 in (LDR18). Condition 5 guarantees that such arguments remain valid: in particular the set M defined within Condition 5 is of positive Lebesgue measure. This allows results for completely randomized designs to provide asymptotics when Ω_{CRE} is replaced with Ω from a general rerandomized design.

3.11.3 Proof of Theorem 1

Theorem 1. *Suppose we have either a completely randomized design or a rerandomized design with balance criterion ϕ satisfying Condition 5. Suppose $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is of the form $f_{\hat{\xi}}(\sqrt{N}\hat{\boldsymbol{\tau}})$ for some f_ξ and $\hat{\xi}$ satisfying Conditions 6 and 7. Suppose further that we employ a covariance estimator $\hat{V}$ satisfying Condition 8 when forming the prepivoted test statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. Then, under Neyman's null $H_N : \bar{\boldsymbol{\tau}} = 0$ and under Assumptions 1 - 3(a), $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in distribution to a random variable $\tilde{U}$ taking values in $[0, 1]$ satisfying*

$$\mathbb{P}(\tilde{U} \leq t) \geq t,$$

for all $t \in [0, 1]$. Furthermore, strengthening Assumption 3(a) to Assumption 3(b), the distribution $\hat{\mathscr{P}}_G(t)$ satisfies

$$\hat{\mathscr{P}}_G(t) \xrightarrow{p} t$$

for all $t \in [0, 1]$.

Proof of Theorem 1. A completely randomized experiment can be viewed as a rerandomized experiment for which $\phi(\mathbf{b}) = 1$ for all $\mathbf{b} \in \mathbb{R}^k$; this ϕ satisfies Condition 5. As such, the proof below proceeds with general ϕ satisfying Condition 5 – making no distinction between rerandomized designs and completely randomized design.

First, we focus on the randomization distribution of the prepivoted test statistic; in other words, we examine the limiting distribution of $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ under H_N. By the finite population central limit theorem of (LD17) in a completely randomized design or a rerandomized design with ϕ satisfying Condition 5 the $\sqrt{N}$-scaled difference in means, $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})^T$, converges

in distribution to $\mathcal{N}(0, V)$ with

$$\begin{aligned}
V &= \begin{pmatrix} V_{\tau\tau} & V_{\tau\delta} \\ V_{\delta\tau} & V_{\delta\delta} \end{pmatrix}; \\
V_{\tau\tau} &= p^{-1}\Sigma_{y(1),\infty} + (1-p)^{-1}\Sigma_{y(0),\infty} - \Sigma_{\tau,\infty}; \\
V_{\delta\delta} &= \{p(1-p)\}^{-1}\Sigma_{x,\infty}; \\
V_{\tau\delta} &= p^{-1}\Sigma_{y(1)x,\infty} + (1-p)^{-1}\Sigma_{y(0)x,\infty} = V_{\delta\tau}^{\mathrm{T}}.
\end{aligned}$$

Furthermore, by Condition 5 and Corollary A1 of (LDR18), we have that for $\mathbf{Z}$ instead uniform over Ω (accounting for the rerandomized design), $\sqrt{N}(\hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}}) \xrightarrow{d} \mathbf{C}$, where $\mathbf{C}$ follows the distribution of $\mathbf{A} \mid \phi(\mathbf{B}) = 1$ for $\mathbf{A} \in \mathbb{R}^d$, $\mathbf{B} \in \mathbb{R}^k$, and $(\mathbf{A}, \mathbf{B})^{\mathrm{T}}$ multivariate Gaussian with covariance V and mean zero.

By Condition 7 $\hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} \xi$ and by Condition 8

$$\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} V + \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix} =: \bar{\bar{V}}.$$

Leveraging Lemma A.4 and the continuous mapping theorem, under H_N

$$h\left(\hat{V}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}), \hat{\xi}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}), \sqrt{N}\hat{\tau}\right) \xrightarrow{d} h\left(V + \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix}, \xi, \mathbf{C}\right)$$

where $\mathbf{C}$ distributed as before. Unwinding the notation of $h(\cdot,\cdot,\cdot)$ gives that $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in distribution to

$$\frac{\gamma^{(d+k)}_{\mathbf{0},\bar{\bar{V}}}\left\{(\mathbf{a},\mathbf{b})^{\mathrm{T}} : f_\xi(\mathbf{a}) \le f_\xi(\mathbf{C}) \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma^{(k)}_{\mathbf{0},\bar{\bar{V}}_{\delta\delta}}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}. \tag{12}$$

If we had known to plug in V for $\hat{V}$, (12) would exactly amount to applying the f_ξ-

pushforward of the Gaussian measure $\gamma_{\mathbf{0},V}^{(d+k)}$ conditional on $\phi(\boldsymbol{b}) = 1$, which would result in a uniform random variable since this is just the asymptotic probability integral transform for $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ given that $\phi(\sqrt{N}\hat{\boldsymbol{\delta}}) = 1$. However, we do not know V and instead estimate it conservatively using a $\hat{V}$ that satisfies Condition 8; this results in the discrepancy between the covariance of $\mathbf{C}$ versus the covariance used in the Gaussian measure $\gamma_{\mathbf{0},\bar{\bar{V}}}^{(d+k)}$ in (12). Consequently, (12) amounts to f_ξ-pushforward of the Gaussian measure $\gamma_{\mathbf{0},\bar{\bar{V}}}^{(d+k)}$ in the numerator (the denominator stays the same in both cases since the bottom right block of both $\bar{\bar{V}}$ and V is $V_{\delta\delta}$). Since $\bar{\bar{V}} \succeq V$, it follows by Lemma 1 (and Anderson's theorem more generally) that the numerator of (12) is no larger than the numerator of (12) with $\bar{\bar{V}}$ replaced by V. Then, since applying the f_ξ-pushforward of the Gaussian measure $\gamma_{\mathbf{0},V}^{(d+k)}$ conditional on $\phi(\boldsymbol{b}) = 1$ results in a uniform random variable, it follows that (12) is stochastically dominated by a uniform random variable from Lemma 2 in the text. In other words, $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges in distribution to a random variable $\tilde{U}$ taking values in $[0, 1]$ satisfying $\mathbb{P}(\tilde{U} \leq t) \geq t$ for all $t \in [0, 1]$.

Now we turn our attention to the limiting value of $\hat{\mathscr{P}}_G(t)$ for any t. Relying upon the result of Proposition 1 – which requires Assumptions 1, 2, and 3(b) – in a completely randomized design the distribution of $\{\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W})\}^{\mathrm{T}} \mid \mathbf{Z}$ converges weakly in probability to a multivariate Gaussian measure, with mean zero and covariance

$$\tilde{V} = \begin{pmatrix} \tilde{V}_{\tau\tau} & \tilde{V}_{\tau\delta} \\ \tilde{V}_{\delta\tau} & \tilde{V}_{\delta\delta} \end{pmatrix}.$$

By (DDCZ13, Lemma 4.1), this is equivalent to

$$\begin{bmatrix} \{\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W})\}^{\mathrm{T}} \\ \{\sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}'), \sqrt{N}\hat{\delta}(\mathbf{x}, \mathbf{W}')\}^{\mathrm{T}} \end{bmatrix} \xrightarrow{d} \{(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}), (\tilde{\mathbf{A}}', \tilde{\mathbf{B}}')\}^{\mathrm{T}} \tag{13}$$

where $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{W}'$ are independent and uniformly distributed over Ω_{CRE} and $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})^{\mathrm{T}}$

and $(\tilde{\mathbf{A}}', \tilde{\mathbf{B}}')^{\mathrm{T}}$ are independent and identically distributed multivariate Gaussians with mean zero and covariance $\tilde{V}$. By the conditions on ϕ outlined in Condition 5, we further have that for $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{W}'$ independently drawn from Ω (now accounting for the restrictions imposed by rerandomization),

$$\begin{bmatrix} \sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \\ \sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \end{bmatrix} \xrightarrow{d} (\mathbf{D}, \mathbf{D}'), \tag{14}$$

where $(\mathbf{D}, \mathbf{D}')$ are independent and identically distributed from the conditional distribution of $\tilde{\mathbf{A}} \mid \phi(\tilde{\mathbf{B}}) = 1$.

By Conditions 7 and 8

$$\begin{bmatrix} \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \\ \hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \\ \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}') \\ \hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}') \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \tilde{\xi} \\ \tilde{V} \\ \tilde{\xi} \\ \tilde{V} \end{bmatrix}. \tag{15}$$

Moreover, (14) and (15) hold jointly. Thus, the continuous mapping theorem implies that

$$\begin{bmatrix} h\left(\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})\right) \\ h\left(\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}'), \hat{\xi}(\mathbf{y}(\mathbf{Z}), \mathbf{W}'), \sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}')\right) \end{bmatrix} \\ \downarrow d \\ \begin{bmatrix} h\left(\tilde{V}, \tilde{\xi}, \mathbf{D}\right) \\ h\left(\tilde{V}, \tilde{\xi}, \mathbf{D}'\right) \end{bmatrix} \tag{16}$$

where $\mathbf{D}$ and $\mathbf{D}'$ are distributed as before.

Recall that under the weak null, $\tilde{\mathbf{y}}(\mathbf{Z}) = \mathbf{y}(\mathbf{Z})$ and

$$h\left(\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \hat{\xi}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}), \sqrt{N}\hat{\tau}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})\right)$$

is precisely $G(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ as previously defined. Observe that $h(\tilde{V}, \tilde{\xi}, \mathbf{D})$ takes the form

$$h\left(\tilde{V}, \tilde{\xi}, \mathbf{D}\right) = \frac{\gamma_{\mathbf{0},\tilde{V}}^{(d+k)}\left\{(\mathbf{a}, \mathbf{b})^{\mathrm{T}} : f_{\tilde{\xi}}(\mathbf{a}) \leq f_{\tilde{\xi}}(\mathbf{D}) \ \wedge \ \phi(\mathbf{b}) = 1\right\}}{\gamma_{\mathbf{0},\tilde{V}_{\delta\delta}}^{(k)}\left\{\mathbf{b} : \phi(\mathbf{b}) = 1\right\}}. \tag{17}$$

The logic applied to (12) applies similarly to (17) except for the fact that the mismatch in the covariance of $\mathbf{C}$ and $\gamma_{\mathbf{0},\bar{\tilde{V}}}^{(d+k)}$ of (12) no longer exists in (17) since $\mathbf{D}$ is derived from $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})^{\mathrm{T}} \sim \mathcal{N}\left(0, \tilde{V}\right)$ and the Gaussian measure $\gamma_{\mathbf{0},\tilde{V}}^{(d+k)}$ is applied. As remarked earlier, since the internal covariance matches the external covariance $h\left(\tilde{V}, \tilde{\xi}, \mathbf{D}\right)$ is uniformly distributed over $[0, 1]$. Applying Lemma 4.1 of (DDCZ13) to (16) thus implies that $\hat{\mathscr{P}}_G$ converges weakly in probability to Unif$[0, 1]$. In other words, $\hat{\mathscr{P}}_G(t) \xrightarrow{p} t$ for all $t \in [0, 1]$. □

3.11.4 Theorem 2

Theorem 2 reduces to the proof of Theorem 1 by recognizing the r_i and $\tilde{r}_i$ as potential outcomes satisfying the required assumptions. The asymptotically vanishing factor $o_P(1)$ in the definitions of $\sqrt{N}\{\breve{\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - \bar{\boldsymbol{\tau}}\}$ and $\sqrt{N}\breve{\tau}(\mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\boldsymbol{\tau}}^{\mathrm{T}}, \mathbf{W})$ plays no role in the analysis of their limiting distributions, thereby allowing for application of the same proofs used to show Proposition 1 and Theorem 1.

3.12 Gaussian Prepivoting After Regression Adjustment

3.12.1 Regression Adjustment in Completely Randomized Experiments

In completely randomized experiments with covariate information, a common practice is to use regression-based estimators for treatment effects to improve efficiency. Assume that k is fixed and smaller than N, and let the potential outcomes be univariate. Define $\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ to be the estimated coefficient on Z_i in an ordinary least squares regression of $y_i(Z_i)$ on

Z_i, $(\mathbf{x}_i - \bar{\mathbf{x}})$, and $Z_i(\mathbf{x}_i - \bar{\mathbf{x}})$. (Lin13) shows that under suitable regularity conditions, $\hat{\tau}_{reg}$ is $\sqrt{N}$-consistent for $\bar{\tau}$ and has an asymptotic variance that is no larger than that of $\hat{\tau}$. Importantly, this result holds true *without* assuming that the linear model inspiring $\hat{\tau}_{reg}$ is actually true.

Let

$$Q_1 = \lim_{N\to\infty} \left(\sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} (y_i(1) - \overline{y}(1)) \right)$$

be the limit of the OLS slopes for potential outcome under treatment regressed upon covariates, and define Q_0 analogously for the potential outcomes under control. The population level treatment residuals based upon the limiting slopes are then defined as

$$\begin{aligned} \varepsilon_i(1) &= (y_i(1) - \overline{y}(1)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} Q_1; \\ \varepsilon_i(0) &= (y_i(0) - \overline{y}(0)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} Q_0. \end{aligned}$$

Let $\tilde{Q} = pQ_1 + (1-p)Q_0$ and further define

$$\begin{aligned} \tilde{\varepsilon}_i(1) &= (y_i(1) - \overline{y}(1)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} \tilde{Q}; \\ \tilde{\varepsilon}_i(0) &= (y_i(0) - \overline{y}(0)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} \tilde{Q}. \end{aligned} \tag{18}$$

Proposition A.2. *Suppose Assumption 1 holds, and suppose further that Assumptions 2 and 3(b) hold for the potential outcomes and covariates. Then,*

$$\begin{aligned} \sqrt{N}\left\{\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - \bar{\boldsymbol{\tau}}\right\} &= \sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} Z_i \varepsilon_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1-Z_i)\varepsilon_i(Z_i) \right) + o_p(1) \\ \sqrt{N}\left\{\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\boldsymbol{\tau}}^{\mathrm{T}}, \mathbf{W})\right\} &= \sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} W_i \tilde{\varepsilon}_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1-W_i)\tilde{\varepsilon}_i(Z_i) \right) + o_p(1) \end{aligned}$$

Let $\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ be the ith sample residual from a regression of $\tilde{\mathbf{y}}(\mathbf{Z})$ on W_i, $(\mathbf{x}_i - \bar{\mathbf{x}})$,

and $W_i(\mathbf{x}_i - \bar{\mathbf{x}})$. Using the sample residuals $\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ form the variance estimators

$$\hat{\sigma}_0^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) = \frac{1}{n_0 - 1} \sum_{i=1}^{N} (1 - W_i) \left\{ \hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) - \frac{1}{n_0} \sum_{j=1}^{N} (1 - W_j) \hat{\varepsilon}_j(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \right\}^2$$
$$\hat{\sigma}_1^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) = \frac{1}{n_1 - 1} \sum_{i=1}^{N} W_i \left\{ \hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) - \frac{1}{n_1} \sum_{j=1}^{N} W_j \hat{\varepsilon}_j(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \right\}^2$$

For the $\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$'s form $\hat{\sigma}_0^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$ and $\hat{\sigma}_1^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$ analogously but replace $\mathbf{W}$ with $\mathbf{Z}$.

Consider the variance estimators

$$\hat{V}_{reg}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) = \frac{N}{n_1} \hat{\sigma}_1^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z}) + \frac{N}{n_0} \hat{\sigma}_0^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$$
$$\hat{V}_{reg}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) = \frac{N}{n_1} \hat{\sigma}_1^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) + \frac{N}{n_0} \hat{\sigma}_0^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}).$$

Observe that $\hat{\sigma}_j^2(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \hat{\sigma}_j^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$ for $j = 0, 1$ regardless of whether or not the weak null holds, but that $\hat{\sigma}_j^2(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \neq \hat{\sigma}_j^2(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ unless the weak null holds.

Proposition A.3. *$\hat{V}_{reg}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ satisfies Condition 8 with $V_{\tau\tau}$ replaced by $V_{\tau\tau}^{(\varepsilon)}$ and $\tilde{V}_{\tau\tau}$ replaced by $\tilde{V}_{\tau\tau}^{(\tilde{\varepsilon})}$. The particular form of Δ, the degree to which $\hat{V}_{reg}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{Z})$ is asymptotically conservative, is*

$$\Delta = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} (\tau_i - \bar{\tau} - (\mathbf{x}_i - \mathbf{x})^{\mathrm{T}} (Q_1 - Q_0))^2.$$

By Theorem 2, one may apply Gaussian prepivoting to $\sqrt{N}\hat{\tau}_{reg}$ using $\hat{V}_{reg}$ and any function $f_{\hat{\xi}}$ satisfying Condition 6 and 7; for instance, take $f_{\hat{\xi}}(\sqrt{N}\hat{\tau}_{reg}) = \sqrt{N}|\hat{\tau}_{reg}|$. Note that other asymptotically equivalent forms for $\hat{V}_{reg}$ to the one given here exist. For example, Section 5 of (Lin13) suggests using the sandwich variance estimator corresponding to $\hat{\tau}_{reg}$.

3.12.2 Proof of Proposition 2

We begin with the following Lemma:

Lemma A.5. *If Assumptions 2 and 3(a) hold for the potential outcomes and covariates, then Assumptions 2 and 3(a) hold for the collection of $\varepsilon_i(z)$. Likewise, if Assumptions 2 and 3(b) hold for the potential outcomes and covariates, then Assumptions 2 and 3(b) hold for the collection of $\tilde{\varepsilon}_i(z)$.*

Proof. For each N, expanding by the definition of $\varepsilon_i(1)$ yields

$$\bar{\varepsilon}(1) = N^{-1} \sum_{i=1}^{N} \left((y_i(1) - \overline{y}(1)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} Q_1 \right) = 0;$$
$$\Sigma_{\varepsilon(1)} = (N-1)^{-1} \sum_{i=1}^{N} \left(y_i(1) - \overline{y}(1) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} Q_1 \right)^2.$$

By inspection, Assumption 2 holds for the collection of $\varepsilon_i(1)$ so long as the potential outcomes and covariates satisfy Assumption 2. Similar proofs establish Assumption 2 for $\varepsilon_i(0)$, $\tilde{\varepsilon}_i(0)$, and $\tilde{\varepsilon}_i(1)$.

Suppose that Assumption 3(a) holds for the potential outcomes and covariates. Then

$$\lim_{N\to\infty} \max_{z\in\{0,1\}} \max_{i\in\{1,\ldots,N\}} \frac{\left(y_i(z) - \bar{y}(z)\right)^2}{N} = 0 \tag{19}$$

and

$$\lim_{N\to\infty} \max_{j\in\{1,\ldots,k\}} \max_{i\in\{1,\ldots,N\}} \frac{\left(x_{ij} - \bar{x}_j\right)^2}{N} = 0.$$

As a consequence of the second statement,

$$\lim_{N\to\infty} \max_{i\in\{1,\ldots,N\}} \frac{\sum_{j=1}^{d} \left(x_{ij} - \bar{x}_j\right)^2}{N} = 0$$

and so, by the Cauchy-Schwarz inequality,

$$\lim_{N\to\infty}\max_{i\in\{1,\ldots,N\}}\frac{\left(\mathbf{x}_i^{\mathrm{T}}Q_1-\bar{\mathbf{x}}^{\mathrm{T}}Q_1\right)^2}{N} \quad\leq\quad \lim_{N\to\infty}\max_{i\in\{1,\ldots,N\}}\frac{||Q_1||_2^2\sum_{j=1}^{d}\left(x_{ij}-\bar{x}_j\right)^2}{N} \quad=\quad 0. \quad (20)$$

Because $\left((y_i(1)-\overline{y}(1))-(\mathbf{x}_i-\overline{\mathbf{x}})^{\mathrm{T}}Q_1\right)^2 \leq 2(y_i(1)-\overline{y}(1))^2+2\left((\mathbf{x}_i-\overline{\mathbf{x}})^{\mathrm{T}}Q_1\right)^2$ it follows from (19) and (20) that Assumption 3(a) holds for the collection of $\varepsilon_i(z)$.

Now suppose that Assumption 3(b) holds for the potential outcomes and covariates: there exists some $C<\infty$ for which, for all $z\in\{0,1\}$ and all N,

$$\frac{\sum_{i=1}^{N}\left(y_{ij}(z)-\overline{y}_j(z)\right)^4}{N}<C \quad \forall j=1,..,d$$

and

$$\frac{\sum_{i=1}^{N}\left(x_{ij}-\overline{x}_j\right)^4}{N}<C \quad \forall j=1,..,k.$$

Modifying the argument from above to accommodate $\tilde{Q}$ instead of Q_1 and applying Hölder's inequality gives the desired result. Specifically, Hölder's inequality implies that

$$\left(\mathbf{x}_i^{\mathrm{T}}\tilde{Q}-\bar{\mathbf{x}}^{\mathrm{T}}\tilde{Q}\right)^4\leq C_Q||\mathbf{x}_i-\overline{\mathbf{x}}||_4^4.$$

where C_Q is a constant that does not change with N and depends only upon $\tilde{Q}$. Combining this inequality with Assumption 3(b) on the potential outcomes then gives that Assumption 3(b) holds for the collection of $\tilde{\varepsilon}_i(z)$. □

We split the proof of Proposition A.2 into two: Proposition A.2(a) and Proposition A.2(b).

Proposition A.2(a).

$$\sqrt{N}\left\{\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})-\bar{\tau}\right\}=\sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N}Z_i\varepsilon_i(Z_i)-\frac{1}{n_0}\sum_{i=1}^{N}(1-Z_i)\varepsilon_i(Z_i)\right)+o_p(1)$$

Proof. By Lemma A.3 of (Lin13),

$$\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) - \bar{\tau} = \frac{1}{n_1}\sum_{i=1}^{N} Z_i\hat{\varepsilon}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) - \frac{1}{n_0}\sum_{i=1}^{N}(1-Z_i)\hat{\varepsilon}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$$

where the sample residuals $\hat{\varepsilon}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ are derived from the regression of $y_i(Z_i)$ on Z_i, $(\mathbf{x}_i - \bar{\mathbf{x}})$, and $Z_i(\mathbf{x}_i - \bar{\mathbf{x}})$. Let $\hat{Q}_1(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ be the sample slope coefficient in the OLS regression of $y_i(Z_i)$ on $\mathbf{x}_i$ in the group of individuals for which $Z_i = 1$; similarly, let $\hat{Q}_0(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ be the sample slope coefficient in the population OLS regression of $y_i(Z_i)$ on $\mathbf{x}_i$ in the group of individuals for which $Z_i = 0$ (Lin13).

Define

$$\hat{\varepsilon}_i(1) = (y_i(1) - \overline{y}(1)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\hat{Q}_1(\mathbf{y}(\mathbf{Z}),\mathbf{Z});$$
$$\hat{\varepsilon}_i(0) = (y_i(0) - \overline{y}(0)) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\hat{Q}_0(\mathbf{y}(\mathbf{Z}),\mathbf{Z});$$

these are random and depend upon $\mathbf{Z}$. The sample residual $\hat{\varepsilon}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is $\hat{\varepsilon}_i(Z_i)$.

By standard OLS theory the slope coefficient matrix $\hat{Q}_1(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is defined by

$$\hat{Q}_1(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) = \left(\frac{1}{n_1-1}\sum_{i=1}^{N} Z_i\left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N} Z_j\mathbf{x}_j\right)\left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N} Z_j\mathbf{x}_j\right)^{T}\right)^{-1} \times \left(\frac{1}{n_1-1}\sum_{i=1}^{N} Z_i\left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N} Z_j\mathbf{x}_j\right)\left(y_i(Z_i) - n_1^{-1}\sum_{j=1}^{N} Z_j y(Z_j)\right)\right)$$

$\hat{Q}_0(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is defined analogously.

By weak laws of large numbers for covariance matrices in finite populations, $\hat{Q}_0(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ and $\hat{Q}_1(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ converge in probability to Q_0 and Q_1, respectively (Lin13, Lemma A.5).

Thus,

$$\hat{\varepsilon}_i(1) - \varepsilon_i(1) = o_P(1)$$
$$\hat{\varepsilon}_i(0) - \varepsilon_i(0) = o_P(1)$$

From this, it follows that

$$\sqrt{N}\left\{\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - \bar{\tau}\right\} = \sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} Z_i\varepsilon_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1 - Z_i)\varepsilon_i(Z_i)\right) + o_p(1)$$

This proof closely parallels the logic used in the proof for Theorem 1 of (Lin13). □

Before proving the remaining component of Proposition A.2 we provide a convenient lemma.

Consider a function $g : \Omega \times \Omega \to \mathbb{R}$. Let $\mathbf{Z}$ and $\mathbf{W}$ independently distributed uniformly over Ω. Define two properties:

Property A. The random variable $g(\mathbf{Z}, \mathbf{W}) \mid \mathbf{Z} = \mathbf{z}$ converges in probability to c for all conditioning sets $\{\mathbf{z}\}_{N\in\mathbb{N}}$ except for a set of measure zero.

Property B. The random variable $g(\mathbf{Z}, \mathbf{W})$ converges in probability to c with respect to randomness in both $\mathbf{Z}$ and $\mathbf{W}$.

Lemma A.6. *Consider a function $g : \Omega \times \Omega \to \mathbb{R}$. For $\mathbf{Z}$ and $\mathbf{W}$ independently distributed uniformly over Ω Property A implies Property B.*

Proof. Assume that Property A holds. Fix $\varepsilon > 0$; then

$$\mathbb{P}_{\mathbf{W}|\mathbf{Z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon \mid \mathbf{Z}\right) \xrightarrow{a.s.} 0. \tag{21}$$

Consider $\mathbb{P}_{\mathbf{W},\mathbf{Z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon\right)$; by the law of total probability

$$\begin{aligned}\mathbb{P}_{\mathbf{Z},\mathbf{W}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon\right) &= \sum_{\mathbf{z}\in\Omega} \mathbb{P}_{\mathbf{W}|\mathbf{Z}=\mathbf{z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon \mid \mathbf{Z}=\mathbf{z}\right)\mathbb{P}_{\mathbf{Z}}\left(\mathbf{Z}=\mathbf{z}\right) \\ &= \mathbb{E}_{\mathbf{Z}}\left[\mathbb{P}_{\mathbf{W}|\mathbf{Z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon \mid \mathbf{Z}\right)\right]\end{aligned}$$

Since $\mathbb{P}_{\mathbf{W}|\mathbf{Z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon \mid \mathbf{Z}\right) \xrightarrow{a.s.} 0$ and $\mathbb{P}_{\mathbf{W}|\mathbf{Z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon \mid \mathbf{Z}\right) \in [0,1]$ the bounded convergence theorem implies that

$$\lim_{N\to\infty} \mathbb{E}_{\mathbf{Z}}\left[\mathbb{P}_{\mathbf{W}|\mathbf{Z}}\left(|g(\mathbf{Z},\mathbf{W}) - c| \geq \varepsilon \mid \mathbf{Z}\right)\right] = \mathbb{E}_{\mathbf{Z}}\left[0\right] = 0$$

Thus, $g(\mathbf{Z},\mathbf{W})$ converges in probability to c with respect to randomness in both $\mathbf{Z}$ and $\mathbf{W}$. □

Proposition A.2(b).

$$\sqrt{N}\left\{\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\tau}, \mathbf{W})\right\} = \sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} W_i\tilde{\varepsilon}_i(Z_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1-W_i)\tilde{\varepsilon}_i(Z_i)\right) + o_p(1)$$

Proof. By definition $\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\tau}, \mathbf{W})$ is the estimated coefficient on W_i in an ordinary least squares regression of $y_i(Z_i) - Z_i\bar{\tau}$ on W_i, $(\mathbf{x}_i - \bar{\mathbf{x}})$, and $W_i(\mathbf{x}_i - \bar{\mathbf{x}})$.

By the same logic that gave rise to Lemma A.3 of (Lin13),

$$\hat{\tau}_{reg}(\mathbf{y}(\mathbf{Z}) - \mathbf{Z}\bar{\tau}, \mathbf{W}) = \frac{1}{n_1}\sum_{i=1}^{N} W_i\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) - \frac{1}{n_0}\sum_{i=1}^{N}(1-W_i)\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$$

where the sample residuals $\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ are derived from the regression of $y_i(Z_i) - Z_i\bar{\tau}$ on

W_i, $(\mathbf{x}_i - \bar{\mathbf{x}})$, and $W_i(\mathbf{x}_i - \bar{\mathbf{x}})$. Let $\hat{Q}_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ be the sample slope coefficient in the OLS regression of $y_i(Z_i) - Z_i\bar{\tau}$ on $\mathbf{x}_i$ in the group of individuals for which $W_i = 1$; similarly, let $\hat{Q}_0(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ be the sample slope coefficient in the OLS regression of $y_i(Z_i) - Z_i\bar{\tau}$ on $\mathbf{x}_i$ in the group of individuals for which $W_i = 0$. For convenience of notation, denote $N^{-1}\sum_{i=1}^{N}\tilde{y}_i(Z_i)$ by $\overline{\tilde{y}}(\mathbf{Z})$.

Consequently

$$\hat{\varepsilon}_i(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) = \begin{cases} \tilde{y}_i(Z_i) - \overline{\tilde{y}}(\mathbf{Z}) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\hat{Q}_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}); & \text{if } W_i = 1 \\ \tilde{y}_i(Z_i) - \overline{\tilde{y}}(\mathbf{Z}) - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\hat{Q}_0(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}); & \text{if } W_i = 0. \end{cases}$$

these are random and depend upon both $\mathbf{Z}$ and $\mathbf{W}$.

By standard OLS theory the slope coefficient matrix $\hat{Q}_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ is

$$\hat{Q}_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) = \left(\frac{1}{n_1 - 1}\sum_{i=1}^{N} W_i \left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N} W_j\mathbf{x}_j\right)\left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N} W_j\mathbf{x}_j\right)^{\mathrm{T}}\right)^{-1} \times$$
$$\left(\frac{1}{n_1 - 1}\sum_{i=1}^{N} W_i \left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N} W_j\mathbf{x}_j\right)\left((y_i(Z_i) - Z_i\bar{\tau}) - n_1^{-1}\sum_{j=1}^{N} W_j\left(y_j(Z_j) - Z_j\bar{\tau}\right)\right)\right)$$

In Lemma A.5 of (Lin13), it is shown that the first term of $\hat{Q}_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ converges in probability to $\Sigma_{x,\infty}^{-1}$. Now we turn our analysis to the second term of $\hat{Q}_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$; denote this term by $M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$.

The centering of the potential outcomes under treatment that occurred when translating $y_i(z)$ to $\tilde{y}_i(z)$ does not impact Assumptions 1, 2, 3(a), and 3(b). Thus, the finite population

strong law for second moments (WD18, Lemma A.3, Part ii) applies to the sample covariances

$$\hat{\Sigma}_{\tilde{y}(1)x} = \frac{1}{n_1 - 1} \sum_{i=1}^{N} Z_i \left(\mathbf{x}_i - n_1^{-1} \sum_{j=1}^{N} Z_j \mathbf{x}_j \right) \left(\tilde{y}_i(1) - n_1^{-1} \sum_{j=1}^{N} Z_j \tilde{y}_j(1) \right)^{\mathrm{T}}$$

$$\hat{\Sigma}_{\tilde{y}(0)x} = \frac{1}{n_0 - 1} \sum_{i=1}^{N} (1 - Z_i) \left(\mathbf{x}_i - n_0^{-1} \sum_{j=1}^{N} (1 - Z_j) \mathbf{x}_j \right) \left(\tilde{y}_i(0) - n_0^{-1} \sum_{j=1}^{N} (1 - Z_j) \tilde{y}_j(0) \right)^{\mathrm{T}}.$$

Since the centering of the potential outcomes under treatment that occurred when translating $y_i(z)$ to $\tilde{y}_i(z)$ does not impact the above covariance structure, it follows from Lemma A.3 of (WD18) that $\hat{\Sigma}_{\tilde{y}(1)x} \xrightarrow{a.s.} \Sigma_{y(1)x,\infty}$ and $\hat{\Sigma}_{\tilde{y}(0)x} \xrightarrow{a.s.} \Sigma_{y(0)x,\infty}$ (This statement relies upon Assumptions 1, 2, and 3(b).). Condition on a sequence of treatment allocations $\{\mathbf{Z}\}_{N \in \mathbb{N}}$ for the growing sequence of experiments such that $\hat{\Sigma}_{\tilde{y}(1)x} \mid \mathbf{Z} \to \Sigma_{y(1)x,\infty}$ and $\hat{\Sigma}_{\tilde{y}(0)x} \mid \mathbf{Z} \to \Sigma_{y(0)x,\infty}$; this requirement is met for all $\mathbf{Z}$ except for a set of measure zero.

Fix the treatment allocations $\{\mathbf{Z}\}_{N \in \mathbb{N}}$; after this conditioning we are left with fully determined "imputed potential outcomes":

- $\{\tilde{y}_i(Z_i)\}_{i=1}^{N}$ for the "imputed treatment potential outcomes"
- $\{\tilde{y}_i(Z_i)\}_{i=1}^{N}$ for the "imputed control potential outcomes"

The imputed population can be envisioned as the population that an experiment would imagine to exist if she observed outcomes $\tilde{y}(\mathbf{Z})$ and believed that Fisher's sharp null held. Consider $\mathbf{W}$ as a treatment allocation for this imputed population. Under this interpretation $M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ is the sample covariance between covariates and the imputed outcomes observed under "treatment" $W_i = 1$. Instead of working with $M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$, we first focus attention to the underlying quantity that $M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ seeks to estimate: the covariance between covariates and the imputed potential outcomes $\{\tilde{y}_i(Z_i)\}_{i=1}^{N}$; we proceed with analysis

based upon a fixed sequence of treatment allocations $\mathbf{Z}$. This quantity is

$$\begin{aligned}\Sigma_{imputed,\,\tilde{y}(1)x} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - N^{-1}\sum_{j=1}^{N}\mathbf{x}_j\right)\left(\tilde{y}_i(Z_i) - N^{-1}\sum_{j=1}^{N}\tilde{y}_j(Z_j)\right)^{\mathrm{T}} \\ &= \frac{1}{N-1}\sum_{i|Z_i=1}\left(\mathbf{x}_i - N^{-1}\sum_{j=1}^{N}\mathbf{x}_j\right)\left(\tilde{y}_i(1) - N^{-1}\sum_{j=1}^{N}\tilde{y}_j(Z_j)\right)^{\mathrm{T}} \\ &\quad + \frac{1}{N-1}\sum_{i|Z_i=0}\left(\mathbf{x}_i - N^{-1}\sum_{j=1}^{N}\mathbf{x}_j\right)\left(\tilde{y}_i(0) - N^{-1}\sum_{j=1}^{N}\tilde{y}_j(Z_j)\right)^{\mathrm{T}}.\end{aligned}$$

By the strong laws for the sample means, this shares the same limit as

$$\begin{aligned}&\frac{1}{N-1}\sum_{i|Z_i=1}\left(\mathbf{x}_i - n_1^{-1}\sum_{j=1}^{N}Z_j\mathbf{x}_j\right)\left(\tilde{y}_i(1) - n_1^{-1}\sum_{j=1}^{N}Z_j\tilde{y}_j(1)\right)^{\mathrm{T}} \\ &\qquad + \frac{1}{N-1}\sum_{i|Z_i=0}\left(\mathbf{x}_i - n_0^{-1}\sum_{j=1}^{N}(1-Z_j)\mathbf{x}_j\right)\left(\tilde{y}_i(0) - n_0^{-1}\sum_{j=1}^{N}(1-Z_j)\tilde{y}_j(0)\right)^{\mathrm{T}}.\end{aligned}$$

In turn, these two terms can be rewritten as

$$\frac{n_1-1}{N-1}\hat{\Sigma}_{\tilde{y}(1)x} + \frac{n_0-1}{N-1}\hat{\Sigma}_{\tilde{y}(0)x}$$

which limits to $p\Sigma_{y(1)x,\infty} + (1-p)\Sigma_{y(0)x,\infty}$ for all $\mathbf{Z}$ except for a set of measure zero. Since the centering of the potential outcomes under treatment that occurred when translating $y_i(z)$ to $\tilde{y}_i(z)$ does not impact Assumptions 1, 2, 3(a), and 3(b) it follows from Lemma 1 of (Lin13) that

$$M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \mid \mathbf{Z} \xrightarrow{p} p\Sigma_{y(1)x,\infty} + (1-p)\Sigma_{y(0)x,\infty}$$

almost surely in $\mathbf{Z}$; combining this with Lemma A.6 implies that

$$M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \xrightarrow{p} p\Sigma_{y(1)x,\infty} + (1-p)\Sigma_{y(0)x,\infty}.$$

Thus

$$\hat{Q}_1(\mathbf{y}(\mathbf{Z}), \mathbf{W}) \xrightarrow{p} \Sigma_{x,\infty}^{-1} \left(p\Sigma_{y(1)x,\infty} + (1-p)\Sigma_{y(0)x,\infty} \right) = \tilde{Q}.$$

The remainder of the proof proceeds in direct analogy with the proof used for Proposition A.2(a). □

Remark 2. The utility of Lemma A.6 in the proof of Proposition A.2(b) arose from our choice to analyze $M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ through conditioning upon treatment allocation $\mathbf{Z}$. With this conditioning argument, Assumption 3(b) is leveraged to attain strong laws with respect to randomness in $\mathbf{Z}$; these guarantee that arguments based upon conditioning on $\mathbf{Z} = \mathbf{z}$ hold for all but a set of measure zero. An alternative approach to arrive at the statement $M_1(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \xrightarrow{p} p\Sigma_{y(1)x,\infty} + (1-p)\Sigma_{y(0)x,\infty}$ may be to work unconditionally: appealing to a suitable weak law while allowing for randomness in both $\mathbf{Z}$ and $\mathbf{W}$. With an approach of this nature, Assumption 3(b) may be stronger than necessary.

3.12.3 Proof of Proposition 3

First we show that

$$\hat{V}_{reg}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - V_{\tau\tau}^{(\varepsilon)} \xrightarrow{p} \Delta, \tag{22}$$

with Δ defined in the statement of Proposition 3. From the proof of Proposition A.2(a)

$$\begin{aligned}\hat{V}_{reg}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = {} & \frac{N}{n_1}\frac{1}{n_1 - 1}\sum_{i=1}^{N} Z_i \left(\varepsilon_i(1) - \frac{1}{n_1}\sum_{j=1}^{N} Z_j \varepsilon_j(1) \right)^2 \\ & + \frac{N}{n_0}\frac{1}{n_0 - 1}\sum_{i=1}^{N} (1 - Z_i) \left(\varepsilon_i(0) - \frac{1}{n_0}\sum_{j=1}^{N} (1 - Z_j)\varepsilon_j(0) \right)^2 \\ & + o_P(1).\end{aligned}$$

Since $N/n_1 \to p^{-1}$ and $N/n_0 \to (1-p)^{-1}$ this has the same limit as $N \to \infty$ as

$$\frac{1}{p}\frac{1}{n_1-1}\sum_{i=1}^{N} Z_i\left(\varepsilon_i(1)-\frac{1}{n_1}\sum_{j=1}^{N} Z_j\varepsilon_j(1)\right)^2 + \frac{1}{1-p}\frac{1}{n_0-1}\sum_{i=1}^{N}(1-Z_i)\left(\varepsilon_i(0)-\frac{1}{n_0}\sum_{j=1}^{N}(1-Z_j)\varepsilon_j(0)\right)^2.$$

Thus, (22) holds by the weak law of large numbers for second moments (Lin13, Lemma A.1) and second part of Theorem 2 from (Lin13).

Next we show that

$$\hat{V}_{reg}(\mathbf{y}(\mathbf{Z}),\mathbf{W}) - \tilde{V}_{\tau\tau}^{(\tilde{\varepsilon})} \xrightarrow{p} 0. \tag{23}$$

By the proof of Proposition A.2(b)

$$\hat{V}_{reg}(\mathbf{y}(\mathbf{Z}),\mathbf{W}) = \frac{N}{n_1}\frac{1}{n_1-1}\sum_{i=1}^{N} W_i\left(\tilde{\varepsilon}_i(Z_i)-\frac{1}{n_1}\sum_{j=1}^{N} W_j\tilde{\varepsilon}_j(Z_j)\right)^2 + \frac{N}{n_0}\frac{1}{n_0-1}\sum_{i=1}^{N}(1-W_i)\left(\tilde{\varepsilon}_i(Z_i)-\frac{1}{n_0}\sum_{j=1}^{N}(1-W_j)\tilde{\varepsilon}_j(Z_j)\right)^2 + o_P(1). \tag{24}$$

By conditioning upon $\mathbf{Z}$, an argument similar to that used to analyze $M_1(\tilde{\mathbf{y}}(\mathbf{Z}),\mathbf{W})$ in the proof of Proposition A.2(b) can then be applied to compute the probability limits of the first two terms in (24) almost surely with respect to the conditioning variable $\mathbf{Z}$. Then leveraging Lemma A.6 yields that the probability limit is the same when considering randomness in both $\mathbf{Z}$ and $\mathbf{W}$. Finally, using $N/n_1 \to p^{-1}$ and $N/n_0 \to (1-p)^{-1}$ yields that

$$\hat{V}_{reg}(\mathbf{y}(\mathbf{Z}),\mathbf{W}) \xrightarrow{p} \frac{1}{p}\Sigma_{\tilde{\varepsilon}(1),\infty} + \frac{1}{1-p}\Sigma_{\tilde{\varepsilon}(1),\infty} = \tilde{V}_{\tau\tau}^{(\tilde{\varepsilon})}.$$

3.13 An Example For Paired Designs

Above we focus upon rerandomized experimental designs. Since a completely randomized experiment is simply a rerandomized experiment with trivial balance criterion, the results from above automatically apply to completely randomized experiments as well. However, Gaussian prepivoting is not limited to just these contexts. Here we illustrate the utility of Gaussian prepivoting for a matched-pair experimental design. Before describing the exact details of Gaussian prepivoting for paired designs, we prove a generalization of Theorem 1 and Theorem 2.

Suppose that $\mathbf{A}(\cdot, \cdot)$ is a function such that under the weak null H_N and for some positive definite matrices V and $\tilde{V}$,

$$\begin{pmatrix} \mathbf{A}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \\ \sqrt{N}\hat{\boldsymbol{\delta}}(\mathbf{x}, \mathbf{Z}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} V_{AA} & V_{A\delta} \\ V_{\delta A} & V_{\delta\delta} \end{pmatrix}\right); \tag{25}$$

$$\begin{pmatrix} \mathbf{A}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) \\ \sqrt{N}\hat{\boldsymbol{\delta}}(\mathbf{x}, \mathbf{W}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \tilde{V}_{AA} & \tilde{V}_{A\delta} \\ \tilde{V}_{\delta A} & \tilde{V}_{\delta\delta} \end{pmatrix}\right). \tag{26}$$

Remark 3. With the introduction of the covariance matrices in (25) and (26) we must modify Condition 8 slightly. It now becomes: with $\mathbf{W}, \mathbf{Z}$ independent, both uniformly distributed over Ω, and for some $\Delta \succeq 0$, $\Delta \in \mathbb{R}^{d\times d}$,

$$\hat{V}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) - \begin{pmatrix} V_{AA} & V_{A\delta} \\ V_{\delta A} & V_{\delta\delta} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \Delta & 0_{d,k} \\ 0_{k,d} & 0_{k,k} \end{pmatrix};$$

$$\hat{V}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W}) - \begin{pmatrix} \tilde{V}_{AA} & \tilde{V}_{A\delta} \\ \tilde{V}_{\delta A} & \tilde{V}_{\delta\delta} \end{pmatrix} \xrightarrow{p} 0_{(d+k),(d+k)}.$$

Theorem A.3. *Suppose that $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = f_{\hat{\xi}}(\mathbf{A}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}))$ for some f_η and $\hat{\xi}$ satisfying Conditions 6 and 7 . If we employ a covariance estimator $\hat{V}$ satisfying the revised Condition*

8 when forming the prepivoted test statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ *then, under* $H_N : \bar{\boldsymbol{\tau}} = 0$ *and the assumption that* (25) *holds,* $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ *converges in distribution to a random variable* $\tilde{U}$ *taking values in* $[0, 1]$ *satisfying*

$$\mathbb{P}(\tilde{U} \leq t) \geq t,$$

for all $t \in [0, 1]$. *Furthermore, if* (26) *holds, then the distribution* $\hat{\mathscr{P}}_G(t)$ *satisfies*

$$\hat{\mathscr{P}}_G(t) \xrightarrow{p} t$$

for all $t \in [0, 1]$.

Proof. The proof of this theorem proceeds exactly as that of Theorem 1, but with $\sqrt{N}\hat{\boldsymbol{\tau}}(\cdot, \cdot)$ replaced by $\mathbf{A}(\cdot, \cdot)$. □

The structure assumed in Theorem A.3 is commonly encountered in finite population causal inference across a host of experimental designs. Armed with Theorem A.3 we turn to the problem of Gaussian prepivoting in paired designs. Consider a population with I matched pairs of individuals, so that the total population size is $N = 2I$. Attributes of the j^{th} unit in the i^{th} pair are subscripted with ij; e.g. potential outcomes are $\mathbf{y}_{ij}(0)$ and $\mathbf{y}_{ij}(1)$. For simplicity take $d = 1$, though these results are not bound to the univariate case. For a paired design $\Omega = \Omega_{pair}$ with

$$\Omega_{pair} := \left\{ \mathbf{z} \in \{0, 1\}^N \;\middle|\; \sum_j \mathbf{z}_{ij} = 1 \; \forall \, i = 1, \ldots, I \right\}.$$

In words, allowable treatment allocations assign one unit of each pair to treatment and the remaining unit of the pair to control. The average observed treated-minus-control difference

in outcomes is

$$\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) := \frac{1}{I}\sum_{i=1}^{I}\Bigg(\underbrace{(2Z_{i1}-1)\Big(y_{i1}(Z_{i1})-y_{i2}(Z_{i2})\Big)}_{\mathscr{T}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z})}\Bigg),$$

with $\mathscr{T}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ representing the treated-minus-control difference in pair i. Subject to Conditions 1 and 2 of (Fog18) – which are the paired-design analogues of our Assumptions 2 and 3(b) – the random variable $\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ obeys a finite population central limit theorem. This finite population central limit theorem for $\sqrt{I}\left(\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})-\overline{\boldsymbol{\tau}}\right)$ can be derived from Theorem 1 of (Fog18) by dropping the regression-assisting terms.

We estimate the variance of $\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ via its classical Neyman-style estimator

$$\hat{V}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) := \frac{1}{I-1}\sum_{i=1}^{I}\Big(\mathscr{T}_i(\mathbf{y}(\mathbf{Z}),\mathbf{Z})-\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})\Big)^2.$$

(Ima08) shows that $\hat{V}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is conservative for the variance of $\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$. Under standard regularity conditions (Fog20, Appendix Lemma 8) there exists a constant $\nu^2>0$ such that $\sqrt{N}\hat{\boldsymbol{\tau}}_{pair}(\tilde{\mathbf{y}}(\mathbf{Z}),\mathbf{W})$ converges in distribution to $\mathcal{N}(0,\nu^2)$ and by (Fog20, Appendix Lemma 11)

$$\hat{V}_{pair}(\tilde{\mathbf{y}}(\mathbf{Z}),\mathbf{W}) := \frac{1}{I-1}\sum_{i=1}^{I}\Big(\mathscr{T}_i(\tilde{\mathbf{y}}(\mathbf{Z}),\mathbf{W})-\hat{\boldsymbol{\tau}}_{pair}(\tilde{\mathbf{y}}(\mathbf{Z}),\mathbf{W})\Big)^2 \xrightarrow{p} \nu^2.$$

Select a function $f_\eta(\cdot)$ which satisfies Conditions 6 and 7, then form the prepivoted test statistic

$$G_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) = \gamma^{(1)}_{0,\hat{V}_{pair}}\left\{\mathbf{a} : f_{\hat{\xi}}(\mathbf{a}) \le f_{\hat{\xi}}(\hat{\boldsymbol{\tau}}_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}))\right\}.$$

Theorem A.3 applies to $G_{pair}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ and so prepivoting naturally extends to paired experimental designs. In fact, for paired designs there are numerous candidates for the variance

estimator $\hat{V}$ that extend beyond the Neyman-style estimator $\hat{V}_{pair}$; for examples, see the regression-assisted variance estimators of (Fog18) or the pairs of pairs estimator discussed in (AI08) and (FLKK19).

3.14 Experiments With Many Treatments

Theorems 1 and 2 are not limited to experiments with only two treatment arms (e.g. treatment versus control). In this section, we show that these results extend naturally to experiments with an arbitrary finite number of treatment arms. For simplicity we present notation and results for completely randomized designs, but extensions are available to rerandomized designs as in the two-armed case. Special cases of balance criteria for general multi-armed rerandomized designs are discussed in (MR12, Section 5.2); for rerandomization in factorial experiments (LDR20) and (BDR16) provide extensive literature.

Consider an experiment with A arms. The treatment indicator for each individual, Z_i, now takes values in $\{0, \ldots, A-1\}$. The potential outcomes under the various treatment options are $\mathbf{y}_i(0), \ldots, \mathbf{y}_i(A-1)$. For convenience denote $\{0, \ldots, A-1\}$ by $[A-1]$. For fixed values $n_0, \ldots, n_{A-1} \in \mathbb{N}$ which sum to N let $\Omega_{CRE,A}$ be the set of treatment allocation vectors

$$\Omega_{CRE,A} := \left\{ \mathbf{z} \in [A-1]^N \;\middle|\; \sum_{i=1}^{N} \mathbb{1}_{\{z_i = a\}} = n_a, \; \forall\, a \in [A-1] \right\}.$$

Modify Assumption 1 to be that $N^{-1} n_a \to p_a$ with $p_a \in (0,1)$ for all a. In words, no treatment arm is asymptotically degenerate. The remaining assumptions are modified to hold for $z \in [A-1]$ instead of just $z \in \{0,1\}$. In the multi-arm setting we redefine Fisher's sharp null to be

$$H_F^{(A)} : \mathbf{y}_i(0) = \mathbf{y}_i(1) = \cdots = \mathbf{y}_i(A-1) \; \forall\, i = 1, \ldots, N.$$

The corresponding generalization of Neyman's weak null is

$$H_N^{(A)} : N^{-1} \sum_{i=1}^{N} \mathbf{y}_i(0) = N^{-1} \sum_{i=1}^{N} \mathbf{y}_i(1) = \cdots = N^{-1} \sum_{i=1}^{N} \mathbf{y}_i(A-1).$$

See (DD18) for discussion of this generalization of the sharp and weak nulls. Further generalizations of these nulls can be found in (WD18).

Denote the vector of the average observed outcome in treatment group a by

$$\hat{\bar{\mathbf{y}}}(a) = n_a^{-1} \sum_{i \,:\, Z_i = a} \mathbf{y}_i(a)$$

and the $d \times A$ matrix of all such averages by

$$\hat{\bar{\mathbf{Y}}} = \left[\hat{\bar{\mathbf{y}}}(0) \quad \cdots \hat{\bar{\mathbf{y}}}(A-1) \right].$$

Consider a matrix $C_{\mathbf{y}}$ of dimensions $A \times d'$ for some d'. We stipulate that this matrices is comprised of column-wise contrasts; i.e., each column contains some non-zero element but sums to zero. In place of $\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ we now turn to the weighted treatment effect estimator

$$\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \text{vec}\left(\hat{\bar{\mathbf{Y}}} C_{\mathbf{y}} \right)$$

where the vec$(\cdot)$ operator reshapes d-by-d' matrices to (dd')-length vectors by vertically concatenating the columns. In the classical two-armed experiment $C_{\mathbf{y}} = [-1, 1]^{\mathrm{T}}$ returns the standard difference in means as $\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$.

We extend the notation $\hat{\Sigma}_{y(a)}$ to denote the sample variance estimator in the population which received treatment arm a. Mimicking the argument of (WD18, Section 2.3) with the natural extension to multivariate outcomes gives that an asymptotically conservative

covariance estimator for $\sqrt{N}\text{vec}\left(\hat{\hat{\mathbf{Y}}}\right)$ is

$$\hat{D}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) = \bigoplus_{a\in[A-1]} \left(\frac{N}{n_a}\hat{\Sigma}_{y(a)}\right),$$

where $\oplus$ denotes the direct sum of matrices, resulting in a block-diagonal matrix of dimension $Ad \times Ad$ with $(N/n_{a-1})\hat{\Sigma}_{y(a-1)}$ in the a^{th} of A blocks. (HS79) present numerous algebraic properties of the Kronecker product and the vectorization operator; exploiting their equation (6) yields

$$\text{vec}\left(\hat{\hat{\mathbf{Y}}}C_{\mathbf{y}}\right) = (C_{\mathbf{y}}^{\text{T}} \otimes I)\text{vec}(\hat{\hat{\mathbf{Y}}})$$

Consequently, to produce a Neyman-style conservative covariance estimator for $\sqrt{N}\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ we form

$$\begin{aligned}\hat{V}_C(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) &:= (C_{\mathbf{y}}^{\text{T}} \otimes I)\left(\bigoplus_{a\in[A-1]} \left(\frac{N}{n_a}\hat{\Sigma}_{y(a)}\right)\right)(C_{\mathbf{y}}^{\text{T}} \otimes I)^{\text{T}} \\ &= (C_{\mathbf{y}}^{\text{T}} \otimes I)\hat{D}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})(C_{\mathbf{y}}^{\text{T}} \otimes I)^{\text{T}}\end{aligned}$$

Since $\hat{D}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is an asymptotically conservative covariance estimator for $\sqrt{N}\text{vec}\left(\hat{\hat{\mathbf{Y}}}\right)$, the covariance estimator $\hat{V}_C(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$ is conservative in the flavor of Condition 8 but with the natural modifications taken to account for our focus on $\sqrt{N}\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}),\mathbf{Z})$.

An analysis of of $\hat{D}(\mathbf{y}(\mathbf{Z}),\mathbf{W})$ under $H_N^{(A)}$ in the univariate outcomes case is included in Appendix A2 of (WD18, Appendix page 6). The extension to multivariate outcomes follows the same reasoning, replacing scalar variance estimators with matrix-valued covariance estimators. Their analysis takes a perspective conditional upon $\mathbf{Z}$, but our Lemma A.6 ports their results into our unconditional framework. Furthermore, (WD18, Appendix A2) provides a detailed analysis of the asymptotic behavior of $\sqrt{N}(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}),\mathbf{W})$ conditional upon

$\mathbf{Z}$ and $H_N^{(A)}$. Using Hoeffding's Lemma – see for instance (DDCZ13, Lemma 4.1) – in an approach mirroring that used in our proof of Theorem 1 the analysis of (WD18) provides an unconditional understanding of $\sqrt{N}(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ under the weak null. Combining their results gives that

$$\hat{V}_C(\mathbf{y}(\mathbf{Z}), \mathbf{W}) - \mathbb{V}\left(\sqrt{N}(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{W})\right) \xrightarrow{p} \mathbf{0}_{(dd')\times(dd')} \tag{27}$$

where $\mathbb{V}\left(\sqrt{N}(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{W})\right)$ denotes the variance of $\sqrt{N}(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{W})$.

These results lay the basis for applying Gaussian prepivoting to test statistics $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ of the form $f_{\hat{\xi}}\left(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\right)$ with $f_{\hat{\xi}}$ satisfying Conditions 6 and 7. The prepivoted test statistic takes its usual form

$$G_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \gamma_{\mathbf{0}, \hat{V}_C}^{(dd')}\left\{\mathbf{a} : f_{\hat{\xi}}(\mathbf{a}) \leq f_{\hat{\xi}}\left(\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\right)\right\}.$$

The central limit behavior of $\sqrt{N}\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and asymptotic conservativeness of the Neyman-style variance estimator $\hat{V}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ apply to Theorem A.3 to show that the true distribution $\mathscr{R}_{G_C}$ is asymptotically stochastically dominated by the standard uniform distribution. The central limit behavior of $\sqrt{N}\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ in conjunction with (27) implies that the reference distribution $\mathscr{P}_{G_C}$ limits weakly in probability to the standard uniform distribution; thereby furnishing inferences which are exact for $H_F^{(A)}$ and asymptotically conservative for $H_N^{(A)}$.

The only major difference between the results for $A > 2$ and $A = 2$ is the use of more general finite population central limit theorems for $\hat{\boldsymbol{\tau}}_C(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ in place of those for $\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$; such a central limit theorem is given by (LD17, Theorem 5). Through particular choices for $C_{\mathbf{y}}$, f_η, and $\hat{\xi}$, we can recover the test statistic proposed in (DD18) for testing the weak null in multi-armed trials while additionally providing an alternative fix to the usual F-statistic to restore asymptotic sharp-dominance. Similarly, the same reasoning can be applied to asymptotically linear estimators (cf. Section 3.7) and the proof of Theorem 2 does not

change substantially.

3.15 Exact And Asymptotically Valid Confidence Sets

Our results readily extend to constructing confidence intervals which are both asymptotically conservative for the sample average treatment effect and exact if a constant treatment effect model holds. To this end, we first describe how to test the hypotheses $H_{F,\mathbf{c}} : \boldsymbol{\tau}_i = \mathbf{c} \, \forall \, i = 1, \ldots, N$ and $H_{N,\mathbf{c}} : \overline{\boldsymbol{\tau}} = \mathbf{c}$ for any fixed $\mathbf{c} \in \mathbb{R}^d$. Define $\tilde{\mathbf{y}}^{\mathbf{c}}(\mathbf{Z})$ to be $\tilde{\mathbf{y}}_i^{\mathbf{c}}(Z_i) = \mathbf{y}_i(Z_i) - Z_i\mathbf{c}$. Then by replacing $\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and $\hat{\boldsymbol{\tau}}(\tilde{\mathbf{y}}(\mathbf{Z}), \mathbf{W})$ with $\hat{\boldsymbol{\tau}}(\tilde{\mathbf{y}}^{\mathbf{c}}(\mathbf{Z}), \mathbf{Z})$ and $\hat{\boldsymbol{\tau}}(\tilde{\mathbf{y}}^{\mathbf{c}}(\mathbf{Z}), \mathbf{W})$, respectively, in Equation 4 the test for H_F and H_N developed in Section 3.5 yields a single procedure that is exact for for $H_{F,\mathbf{c}}$ and asymptotically conservative for $H_{N,\mathbf{c}}$.

Now, first suppose that one is willing to assume a constant effect model and desires a confidence set for the value $\mathbf{c}$ such that $\boldsymbol{\tau}_i = \mathbf{c}$ for all $i = 1, \ldots, N$. Second, suppose that one wants a confidence set for the average treatment effect without assuming a constant effect; i.e., a confidence set for the value $\mathbf{c}$ for which $\overline{\boldsymbol{\tau}} = \mathbf{c}$. Fix a confidence level $1-\alpha \in (0, 1)$ and let $C(T, \mathbf{y}(\mathbf{Z}), \mathbf{Z})$ be the acceptance region for the test of $H_{F,\mathbf{c}}$ conducted at level α based upon the test statistic $T(\cdot, \cdot)$ evaluated over $\mathbf{c} \in \mathbb{R}^d$.

Theorem 1 implies that randomization inference using the prepivoted test statistic $G(\cdot, \cdot)$ yields exact tests for $H_{F,\mathbf{c}}$ and asymptotically valid tests for $H_{N,\mathbf{c}}$. Leveraging the duality between hypothesis testing and confidence sets (CB90, Theorem 9.2.2) implies the following corollary of Theorem 1.

Corollary A.2. *Assume that the regularity conditions of Theorem 1 hold and consider the set $C(G, \mathbf{y}(\mathbf{Z}), \mathbf{Z})$ formed by inverting the randomization test of $H_{F,\mathbf{c}}$ conducted using the prepivoted test statistic $G(\tilde{\mathbf{y}}^{\mathbf{c}}(\mathbf{Z}), \cdot)$. $C(G, \mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is both an asymptotically conservative confidence set for the sample average treatment effect and an exact confidence set under the additional assumption of constant treatment effects.*

Corollary A.2 implies that inverting hypothesis tests based on Gaussian prepivoting yields a single confidence set with both $1-\alpha$ coverage for under a constant effect model at all finite N and also at least $1-\alpha$ asymptotic coverage for the sample average treatment effect. The extension to generating confidence sets based upon asymptotically linear estimators of the form $\breve{T}$ follows similarly but rests upon Theorem 2 instead of Theorem 1. Consequently, regression adjustment can be incorporated into the confidence set generating procedure.

3.16 Additional Simulations

3.16.1 The Generative Model

Theorem 1 and its generalizations concern the finite sample and asymptotic Type I error rates of testing H_F and H_N. Here we provide additional simulations to highlight the potential for anti-conservative inference in the absence of prepivoting and to investigate the statistical power of the Fisher Randomization Test based upon prepivoted test statistics.

Our simulations proceed similarly to those of Section 3.8. For completeness, we detail the simulation set-up here. In each iteration $b = 1, ..., B$, we draw $\{\mathbf{r}_i(1)\}_{i=1}^N$ and $\{\mathbf{r}_i(0)\}_{i=1}^N$ independent from one another and *iid* from mean zero equicorrelated multivariate normals of dimension $k = 25$ with marginal variances one. The correlation coefficients governing $\mathbf{r}_i(1)$ and $\mathbf{r}_i(0)$ are 0 and 0.95 respectively. For both our type I error and power simulations we will have two simulation settings, one with constant treatment effects and one with heterogeneous treatment effects:

Constant Effects: $\mathbf{y}_i(1) - \boldsymbol{\tau} = \mathbf{y}_i(0) = \mathbf{r}_i(1)$.

Heterogeneous Effects: $\mathbf{y}_i(1) = \mathbf{r}_i(1)$; $\mathbf{y}_i(0) + \overline{\boldsymbol{\tau}} = \mathbf{r}_i(0) + \bar{\mathbf{r}}(1) - \bar{\mathbf{r}}(0)$.

The experimental design is that of a completely randomized experiment in which $n_1 = 0.2N$. We test for treatment effect using randomization inference based upon the following three

statistics:

1. Hotelling's T-squared, unpooled covariance,

2. Hotelling's T-squared, pooled covariance,

3. Max absolute t-statistic, unpooled standard error.

3.16.2 Type I Error Rates

When $\boldsymbol{\tau} = \mathbf{0}$ in the constant effects simulation setting in Section 3.16.1, Fisher's sharp null holds; likewise taking $\overline{\boldsymbol{\tau}} = \mathbf{0}$ under heterogeneous effects enforces Neyman's weak null. In these contexts, we reexamine the Type I error rate simulations of Section 3.8, but at $\alpha = 0.25$ instead of 0.05. While this value of α is larger than typical in scientific practice, the larger value of α allows frequent rejections of the null hypothesis despite the inherent conservativeness of inference under the finite population model. We stress that the conservativeness of these tests rests upon the finite population inference framework and not upon the mechanics of Gaussian prepivoting: the non-identifiability of $\Sigma_{\tau,\infty}$ forces conservativeness of *any* procedure which asymptotically guarantees Type I error rate control under H_N. We conduct simulations with $N = 300$ and $N = 5000$; for each N, we conduct $B = 5000$ simulations. These tests are conducted using Monte-carlo simulation to generate the reference distributions, with 1000 draws from Ω_{CRE} for each iteration b. Table 3.3 presents Type I error rates under simulation with $\boldsymbol{\tau} = \overline{\boldsymbol{\tau}} = \mathbf{0}$.

As observed both in Table 3.3 above and in Table 3.2, the Fisher Randomization Test using the pooled Hotelling T-statistic demonstrates significant anti-conservativeness under H_N. Moreover, we see from Table 3.3 that at $\alpha = 0.25$ the Fisher Randomization Test using the max t-statistic is also anti-conservative under H_N, a problem that persists even when $N = 5000$. As a demonstration of Theorem 1, the Fisher Randomization Tests using the prepivoted versions of T_{pool} and $T_{|max|}$ control the Type I error rate under Neyman's null for

Table 3.3: Type I error rates in completely randomized designs with multiple outcomes. The rows describe the simulation settings, which vary between the sharp and weak nulls holding and between small and large sample sizes. There are three sets of columns, one corresponding to each of the three test statistics under consideration. For each set of columns, the column labeled "FRT" represents the Fisher Randomization Test using that test statistic. The column labeled "Pre." instead reflects the Fisher Randomization Test after applying Gaussian prepivoting to the original test statistic. The last column, labeled "LS," is a large-sample test which is asymptotically valid for the weak null. The desired Type I error rate in all settings is $\alpha = 0.25$. For all columns $\boldsymbol{\tau} = \overline{\boldsymbol{\tau}} = \mathbf{0}$.

	Hotelling, Unpooled			Hotelling, Pooled			Max t-stat		
	FRT	Pre.	LS	FRT	Pre.	LS	FRT	Pre.	LS
H_F, $N = 300$	0.244	0.244	0.630	0.251	0.249	0.365	0.254	0.252	0.300
H_F, $N = 5000$	0.247	0.247	0.270	0.248	0.243	0.257	0.251	0.247	0.255
H_N, $N = 300$	0.320	0.320	0.538	0.996	0.361	0.433	0.321	0.071	0.082
H_N, $N = 5000$	0.049	0.049	0.056	0.990	0.064	0.067	0.308	0.060	0.064

large N.

3.16.3 Power after Prepivoting

Below we provide a theoretical discussion of the power of the Fisher Randomization Test based on $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and use simulations to highlight key aspects of its statistical power in practice.

When prepivoting is not necessary because the test statistic being deployed is already sharp dominant and pivotal, its use does not affect the power of the test. Suppose that the Fisher Randomization Test using the pivotal test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ provides exact inferences under H_F and asymptotically valid inferences under H_N. Examples of such test statistics include the studentized absolute difference in means for $d = 1$ and its multivariate analogue $T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$; see (WD18) for further examples. Let $\mathcal{N}_{f_{\hat{\xi}}, \hat{V}}$ denote the $f_{\hat{\xi}}$-pushforward of the Gaussian measure $\gamma^{(d+k)}_{\mathbf{0}, \hat{V}}$. Since $f_{\hat{\xi}}$ takes values in $\mathbb{R}$ the pushforward measure $\mathcal{N}_{f_{\hat{\xi}}, \hat{V}}$ is a distribution on the real line, and so – in a slight abuse of notation – we write its corresponding cumulative distribution function evaluated at $t \in \mathbb{R}$ as $\mathcal{N}_{f_{\hat{\xi}}, \hat{V}}(t)$. For

a completely randomized experiment $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \mathcal{N}_{f_{\hat{\xi}}, \hat{V}}(T(\mathbf{y}(\mathbf{Z}), \mathbf{Z}))$. If $\mathcal{N}_{f_{\hat{\xi}}, \hat{V}}$ is pivotal in the sense that its distribution does not depend upon unknown parameters requiring estimation, then $G(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ is a fixed continuous non-decreasing transformation of $T(\mathbf{y}(\mathbf{Z}), \mathbf{W})$. Consequently, for any fixed $\mathbf{Z}$ the pair $(G(\mathbf{y}(\mathbf{Z}), \mathbf{w}), T(\mathbf{y}(\mathbf{Z}), \mathbf{w}))$ has rank correlation 1 when enumerated over $\mathbf{w} \in \Omega$ and so p-values derived under $\hat{\mathscr{P}}_T$ exactly match those under $\hat{\mathscr{P}}_G$. In this case, prepivoting has no impact upon the power of the test: a test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ with high power under the alternative will yield $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ with high power as well. An example of such a case is T_{χ^2}.

However, as demonstrated in Section 3.5, there are cases for which $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ cannot be used for randomization inference under H_N because it is not asymptotically sharp dominant. Examples of this include $T_{pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and $T_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. Even in these cases, the asymptotic power of the Fisher Randomization Test using $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ can be computed. Regardless of pivotality, the test statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ itself is the complement of a p-value for an asymptotically valid test of H_N. In other words, $1 - G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ can be used directly as a p-value for testing H_N with asymptotic control of the Type I error rate (simply reject H_N if $1 - G(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \leq \alpha$.) The power of this large-sample test must be computed on a case-by-case basis since it is reliant on the structure of the underlying test statistic $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. Because the reference distribution employed by Gaussian prepivoting converges to a standard uniform even under the alternative, the asymptotic power of the randomization test using $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ converges to the power of this large-sample test of H_N, with the added benefit that the randomization test is exact for finite N under H_F. Randomization inference after Gaussian prepivoting leverages an asymptotically valid test for H_N and furnishes exact inference for H_F with no sacrifice in asymptotic power against H_N. In other words, for test statistics $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ satisfying Conditions 6 and 7 exactness under H_F can be achieved for free, with limiting power remaining equal to that of the large-sample test upon which $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is based.

Table 3.4 presents power simulations under the set-up detailed above for $\boldsymbol{\tau} = \overline{\boldsymbol{\tau}} = 0.05\mathbf{e}$,

Table 3.4: Power simulations in completely randomized designs with multiple outcomes. The rows describe the simulation settings, which vary between constant (labelled C) and heterogeneous (labelled H) effects and between small and large sample sizes. There are three sets of columns, one corresponding to each of the three test statistics under consideration. For each set of columns, the column labeled "FRT" represents the Fisher Randomization Test using that test statistic. The column labeled "Pre." instead reflects the Fisher Randomization Test after applying Gaussian prepivoting to the original test statistic. The last column, labeled "LS," is a large-sample test which is asymptotically valid for the weak null. The desired Type I error rate in all settings is $\alpha = 0.25$. For all columns $\boldsymbol{\tau} = \overline{\boldsymbol{\tau}} = 0.05\mathbf{e}$ where $\mathbf{e}$ is the vector of all ones.

	Hotelling, Unpooled			Hotelling, Pooled			Max t-stat		
	FRT	Pre.	LS	FRT	Pre.	LS	FRT	Pre.	LS
C, $N = 300$	0.378	0.378	0.748	0.389	0.384	0.521	0.339	0.335	0.393
C, $N = 5000$	1	1	1	1	1	1	0.969	0.968	0.971
H, $N = 300$	0.360	0.360	0.574	0.995	0.391	0.452	0.458	0.130	0.149
H, $N = 5000$	0.421	0.421	0.448	0.995	0.080	0.086	0.993	0.861	0.868

where $\mathbf{e}$ denotes the vector of all ones. Under constant effects, the power of all of the tests is high and the Type I error rate is controlled for H_F because we are using Fisher Randomization Tests. Although the power of the Fisher Randomization Tests using T_{pool} and $T_{|max|}$ is very high for heterogeneous effects, as observed in Table 3.3 the randomization tests of T_{pool} and $T_{|max|}$ do not control the Type I error rate for testing H_N even asymptotically. However, for the tests which do asymptotically control the Type I error rate under H_N the randomization test of $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ has power observed to be close to that of the large-sample test. Furthermore, the gap in power between the two diminishes as N increases. As stated above, this is because the critical value deployed by the randomization test of $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is converging to 0.75 as N increases, while the large-sample test rejects for $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \geq 0.75$. This further highlights the asymptotic equivalence between the two approaches.

The results for the prepivoted test based upon the usual (pooled) Hotelling test yield two interesting observations. First, the pooled test has markedly worse power than the unpooled or max-t statistics, as the use of the pooled covariance matrix in forming the test statistics

amounts to a choice of a suboptimal norm for constructing the test. Second, the power actually decreases for both the prepivoted randomization test and the large-sample test when going from $N = 300$ to $N = 5000$. This can be attributed to the large-sample approximation being quite poor for the pooled test at $N = 300$ in the setting under consideration. As shown in the simulations in Section 3.8 and in Table 3.3, the Type I error rate exceeds the nominal level when the weak null is true at $N = 300$, but falls below it at $N = 5000$. As N increases the large-sample approximation becomes better, hence restoring the conservativeness of the test under the null. This behavior also drives the apparent reduction in power in the above table. The power still tends to 1 for the pooled Hotelling test as $N \to \infty$ in the generative model yielding this simulation study.

3.17 Gaussian Integral Formulation

Above we used the notation $\gamma^{(\ell)}_{\boldsymbol{\mu},\Sigma}(\mathscr{B})$ to denote the measure of a Borel-measurable set $\mathscr{B} \subseteq \mathbb{R}^\ell$ under Gaussian measure with mean $\boldsymbol{\mu}$ and covariance Σ. Here we provide equivalent formulations of the example Gaussian prepivoted test statistics examined in Section 3.5, but instead of using $\gamma^{(\ell)}_{\boldsymbol{\mu},\Sigma}$ we directly write the corresponding Gaussian integrals.

Example 5 (Absolute difference in means)**.** Let $\sqrt{N}\hat{\tau}$ be univariate, consider a completely randomized design with no rerandomization, and let $T_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \sqrt{N}|\hat{\tau}|$, such that $f_\eta(t) = |t|$ and $\hat{\xi} = 1$. Gaussian prepivoting yields the test statistic

$$\begin{aligned} G_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) &= \gamma^{(1)}_{0,\hat{V}_{\tau\tau}}\{a : |a| \leq \sqrt{N}|\hat{\tau}|\} \\ &= \frac{1}{\sqrt{2\pi\hat{V}_{\tau\tau}}} \int_{-\sqrt{N}|\hat{\tau}|}^{\sqrt{N}|\hat{\tau}|} \exp\left(\frac{-a^2}{2\hat{V}_{\tau\tau}}\right) da \\ &= 1 - 2\Phi\left(-\frac{\sqrt{N}|\hat{\tau}|}{\sqrt{\hat{V}_{\tau\tau}}}\right), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Example 6 (Multivariate studentization)**.** Let $\sqrt{N}\hat{\boldsymbol{\tau}}$ now be multivariate and suppose we have a completely randomized design; consider the test statistic

$$T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right)^{\mathrm{T}} \hat{V}_{\tau\tau}^{-1} \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right); \tag{28}$$
$$\hat{V}_{\tau\tau} = \frac{N}{n_1}\hat{\Sigma}_{y(1)} + \frac{N}{n_0}\hat{\Sigma}_{y(0)}.$$

For this test statistic, $f_\eta(\mathbf{t}) = \mathbf{t}^{\mathrm{T}}\eta^{-1}\mathbf{t}$ and $\hat{\xi} = \hat{V}_{\tau\tau}$. Gaussian prepivoting produces

$$\begin{aligned} G_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) &= \gamma^{(d)}_{\mathbf{0},\hat{V}_{\tau\tau}}\{\mathbf{a} : \mathbf{a}^{\mathrm{T}}\hat{V}_{\tau\tau}^{-1}\mathbf{a} \le T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\} \\ &= \frac{1}{\sqrt{(2\pi)^d \det(\hat{V}_{\tau\tau})}} \int_{\mathbb{R}^d} \mathbb{1}_{\left\{\mathbf{a}^{\mathrm{T}}\hat{V}_{\tau\tau}^{-1}\mathbf{a} \le T_{\chi^2}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})\right\}} \exp\left(\frac{-\mathbf{a}^{\mathrm{T}}\hat{V}_{\tau\tau}^{-1}\mathbf{a}}{2}\right) d\mathbf{a} \qquad (29) \\ &= F_d\{T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\}, \end{aligned}$$

where $F_d(\cdot)$ is the distribution function of a χ^2_d random variable.

Example 7 (Max absolute t-statistic)**.** Consider again multivariate $\sqrt{N}\hat{\boldsymbol{\tau}}$ in a completely randomized design and the test statistic

$$T_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \max_{1\le j\le d} \frac{\sqrt{N}|\hat{\tau}_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}},$$

where $\hat{V}_{\tau\tau,jj}$ is the jj element of $\hat{V}_{\tau\tau}$. For this statistic, $f_{\boldsymbol{\eta}}(\mathbf{t}) = \max_{1\le j\le d}|t_j|/\eta_j$, and $\hat{\boldsymbol{\xi}} =$

$(\hat{V}_{\tau\tau,11}^{1/2}, ..., \hat{V}_{\tau\tau,dd}^{1/2})^{\mathrm{T}}$. After Gaussian prepivoting we are left with

$$\begin{aligned} G_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) &= \gamma_{\mathbf{0},\hat{V}_{\tau\tau}}^{(d)} \left\{ \mathbf{a} : \max_{1\leq j\leq d} \frac{|a_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}} \leq \max_{1\leq j\leq d} \frac{\sqrt{N}|\hat{\tau}_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}} \right\} \\ &= \frac{1}{\sqrt{(2\pi)^d \det(\hat{V}_{\tau\tau})}} \times \\ &\quad \int_{\mathbb{R}^d} \mathbb{1}_{\left\{ \max_{1\leq j\leq d} \frac{|a_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}} \leq T_{|max|}(\mathbf{y}(\mathbf{Z}),\mathbf{Z}) \right\}} \exp\left(\frac{-\mathbf{a}^{\mathrm{T}} \hat{V}_{\tau\tau}^{-1} \mathbf{a}}{2} \right) d\mathbf{a}. \end{aligned} \tag{30}$$

Importantly, (29) and (30) differ only in the support of the Gaussian integral. The same Gaussian measure is used; the difference is that the support of (29) is an ellipsoid while the support of (30) is a hyperrectangle.

Example 8 (Rerandomization)**.** Let $\sqrt{N}\hat{\tau}$ be univariate and suppose we now consider a rerandomized design with balance criterion ϕ satisfying Condition 1. Consider the absolute difference in means, $f_{\hat{\xi}}(\sqrt{N}\hat{\tau}) = \sqrt{N}|\hat{\tau}|$, such that $\hat{\xi} = 1$. Gaussian prepivoting yields the test statistic

$$\begin{aligned} G_{Re}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) &= \frac{\gamma_{\mathbf{0},\hat{V}}^{(1+k)} \left\{ (\mathbf{a}, \mathbf{b})^{\mathrm{T}} : |a| \leq \sqrt{N}|\hat{\tau}| \ \wedge \ \phi(\mathbf{b}) = 1 \right\}}{\gamma_{\mathbf{0},\hat{V}_{\delta\delta}}^{(k)} \left\{ \mathbf{b} : \phi(\mathbf{b}) = 1 \right\}} \\ &= \frac{\frac{1}{\sqrt{(2\pi)^{(k+1)} \det(\hat{V})}} \int_{\mathbb{R}^k} \left(\phi(\mathbf{b}) \int_{-\sqrt{N}|\hat{\tau}|}^{\sqrt{N}|\hat{\tau}|} \exp\left(\frac{-[a\,\mathbf{b}^{\mathrm{T}}] \hat{V}^{-1} [a\,\mathbf{b}^{\mathrm{T}}]^{\mathrm{T}}}{2} \right) da \right) d\mathbf{b}}{\frac{1}{\sqrt{(2\pi)^k \det(\hat{V}_{\delta\delta})}} \int_{\mathbb{R}^k} \phi(\mathbf{b}) \exp\left(\frac{-\mathbf{b}^{\mathrm{T}} \hat{V}_{\delta\delta}^{-1} \mathbf{b}}{2} \right) d\mathbf{b}}. \end{aligned} \tag{31}$$

Since $\phi(\cdot)$ is a boolean-valued function it directly constrains the support of the Gaussian integrals in (31).

3.18 Discussing Condition 2

Recall Condition 2:

Condition 2. *For any $\eta \in \Xi$, $f_\eta(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}_+$ is continuous, quasi-convex, and nonnegative with $f_\eta(\mathbf{t}) = f_\eta(-\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^d$. Furthermore, $f_\eta(\mathbf{t})$ is jointly continuous in η and $\mathbf{t}$.*

Each condition on f_η plays an important role in the underlying mechanics of Gaussian prepivoting, and each deserves some degree of attention. First, the joint continuity of $f_\eta(\mathbf{t})$ in η and $\mathbf{t}$ plays a critical role in the asymptotic behavior of the test statistic T. When computing the asymptotic distributional behavior of $T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ we leverage the central limit theorem governing $\sqrt{N}\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and the continuous mapping theorem to obtain the distributional limit of

$$f_{\hat{\xi}(\mathbf{y}(\mathbf{Z}),\mathbf{Z})}\left(\sqrt{N}\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})\right).$$

Without the joint continuity of $f_\eta(\mathbf{t})$, such a generic asymptotic result would not be feasible. The same reasoning shows the utility of Condition 2's joint continuity requirement when analyzing

$$f_{\hat{\xi}(\mathbf{y}(\mathbf{Z}),\mathbf{W})}\left(\sqrt{N}\hat{\boldsymbol{\tau}}(\mathbf{y}(\mathbf{Z}), \mathbf{W})\right).$$

In this sense, the joint continuity assumption is of technical importance for deriving asymptotic distributional behavior. The quasi-convexity and mirror symmetry assumptions are of a more fundamental nature to our results; they are inextricably linked to Anderson's 1955 theorem for multivariate Gaussians (And55) and so they play a crucial role in guaranteeing the asymptotic sharp dominance of $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. A quasi-convex function is a function with convex sublevel sets; for those unfamiliar with quasi-convex functions, we suggest the excellent review of (ADSZ10, Chapter 3). A simple example function from $\mathbb{R}^d \to \mathbb{R}$ that is both quasi-convex and mirror symmetric about the origin is the Euclidean norm $\mathbf{t} \mapsto ||\mathbf{t}||$. Generalizing slightly more, if $f_\eta(\mathbf{t})$ is any seminorm on $\mathbb{R}^d$ which is jointly continuous in η and $\mathbf{t}$, then $f_\eta(\cdot)$ satisfies Condition 2. In fact, this is nearly a complete characterization;

we will show that the criteria of Condition 2 stipulate that $f_\eta(\cdot)$ is tightly related to a semi-norm; though $f_\eta(\cdot)$ need not be a seminorm itself. Our discussion centers around the case of a completely randomized experiment, but this restriction is only for the sake of explication; similar reasoning applies in the rerandomized case as well.

Consider a convex set $\mathcal{U} \subset \mathbb{R}^d$; suppose that $\mathcal{U}$ is *balanced* in the sense that $c\mathcal{U} \subseteq \mathcal{U}$ for all scalars $u \in [-1, 1]$. Such a set $\mathcal{U}$ is necessarily mirror-symmetric about the origin, and so Anderson's theorem states that:

> If $\mathcal{X} \sim \mathcal{N}(\mathbf{0}, S_\mathcal{X})$ and $\mathcal{Y} \sim \mathcal{N}(\mathbf{0}, S_\mathcal{Y})$ are non-degenerate with $S_\mathcal{Y} - S_\mathcal{X} \succeq 0$, then $\mathbb{P}(\mathcal{X} \in \mathcal{U}) \geq \mathbb{P}(\mathcal{Y} \in \mathcal{U})$.

Moreover, such a set $\mathcal{U}$ defines a seminorm on $\mathbb{R}^d$ via its Minkowski functional $\rho_\mathcal{U}(\mathbf{t}) = \inf_{k>0}\{\mathbf{t} \in k\mathcal{U}\}$. In light of this, Anderson's theorem can be rewritten as:

> If $\mathcal{X} \sim \mathcal{N}(\mathbf{0}, S_\mathcal{X})$ and $\mathcal{Y} \sim \mathcal{N}(\mathbf{0}, S_\mathcal{Y})$ are non-degenerate with $S_\mathcal{Y} - S_\mathcal{X} \succeq 0$, then $\mathbb{P}(\rho_\mathcal{U}(\mathcal{X}) \leq 1) \geq \mathbb{P}(\rho_\mathcal{U}(\mathcal{Y}) \leq 1)$.

Denote the preimage of a set S under f_η by $f_\eta^{-1}(S)$. By quasi-convexity and symmetry of f_η, the set $f_{\hat{\xi}}^{-1}([-\infty, T])$ is convex and symmetric about the origin for any $T \in \mathbb{R}$. Specifically, taking $\mathcal{U} = f_{\hat{\xi}}^{-1}([-\infty, T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})])$ yields a random seminorm $\rho_\mathcal{U}$. Finally, taking $S_\mathcal{Y} = \bar{\bar{V}}$ and $S_\mathcal{X} = V$ gives exactly that the randomization distribution of the Gaussian prepivoted test statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ is asymptotically dominated by the uniform distribution.

In other words, Anderson's theorem can be rephrased to say that when $S_\mathcal{Y} - S_\mathcal{X} \succeq 0$ the random variable $\mathcal{X}$ is more concentrated in any semi-norm than $\mathcal{Y}$; Condition 2 is designed exactly so that the random set $\mathcal{U} = f_{\hat{\xi}}^{-1}([-\infty, T(\mathbf{y}(\mathbf{Z}), \mathbf{Z})])$ generates a seminorm via its Minkowski functional.

3.19 Details Of Examples

In Section 3.5.2, we provide several examples of test statistics which are amenable to Gaussian prepivoting. Here we provide details to verify the conditions of Theorem 1 for these examples.

Define the standard Neyman covariance estimator

$$\hat{V}(\mathbf{y}(\mathbf{z}), \mathbf{w}) = \begin{bmatrix} \hat{V}_{\tau\tau}(\mathbf{y}(\mathbf{z}), \mathbf{w}) & \hat{V}_{\tau\delta}(\mathbf{y}(\mathbf{z}), \mathbf{w}) \\ \hat{V}_{\tau\delta}(\mathbf{y}(\mathbf{z}), \mathbf{w})^{\mathrm{T}} & \hat{V}_{\delta\delta}(\mathbf{y}(\mathbf{z}), \mathbf{w}) \end{bmatrix}$$

where

$$\begin{aligned}
\hat{V}_{\tau\tau}(\mathbf{y}(\mathbf{z}), \mathbf{w}) &= N \left(\frac{\hat{\Sigma}_{y(1)}(\mathbf{y}(\mathbf{z}), \mathbf{w})}{n_1} + \frac{\hat{\Sigma}_{y(0)}(\mathbf{y}(\mathbf{z}), \mathbf{w})}{n_0} \right), \\
\hat{\Sigma}_{y(1)}(\mathbf{y}(\mathbf{z}), \mathbf{w}) &= \frac{1}{n_1 - 1} \sum_{i \,:\, w_i = 1} \left(\mathbf{y}_i(z_i) - \frac{1}{n_1} \sum_{j \,:\, w_j = 1} \mathbf{y}_j(z_j) \right) \left(\mathbf{y}_i(z_i) - \frac{1}{n_1} \sum_{j \,:\, w_j = 1} \mathbf{y}_j(z_j) \right)^{\mathrm{T}}, \\
\hat{\Sigma}_{y(0)}(\mathbf{y}(\mathbf{z}), \mathbf{w}) &= \frac{1}{n_0 - 1} \sum_{i \,:\, w_i = 0} \left(\mathbf{y}_i(z_i) - \frac{1}{n_0} \sum_{j \,:\, w_j = 0} \mathbf{y}_j(z_j) \right) \left(\mathbf{y}_i(z_i) - \frac{1}{n_0} \sum_{j \,:\, w_j = 0} \mathbf{y}_j(z_j) \right)^{\mathrm{T}}.
\end{aligned}$$

and the other blocks are defined analogously.

Lemma A.7. *The Neyman covariance estimator satisfies Condition 4*

Proof. Limiting conservativeness of $\hat{V}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ rests upon the conservativeness of $\hat{V}_{\tau\tau}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$, a well known fact dating back to Neyman himself in the scalar case (Ney90). The vector version of this result is noted in (DFM19, Section 2.2) and relies upon the consistency of the sample covariance estimators $\hat{\Sigma}_{y(0)}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and $\hat{\Sigma}_{y(1)}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$. The consistency of $\hat{\Sigma}_{y(0)}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ and $\hat{\Sigma}_{y(1)}(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ (and their related quantities for the other blocks) is a consequence of Assumptions 1-3 and (Lin13, Appendix Lemma 1).

Verifying the second part of Condition 4 requires examining the limiting behavior of $\hat{V}(\mathbf{y}(\mathbf{Z}), \mathbf{W})$. Such an analysis can be found in (WD18, Appendix A2); while their work

focuses on the scalar-outcome many-treatment case, the techniques convert straightforwardly to the vector-outcome treated-versus-control case. □

Lemma A.7 establishes that the covariance estimators used for prepivoting in Section 3.5 are indeed in accordance with Condition 4. Next, we examine each example provided in Section 3.5.2 to establish that Conditions 2 and 3 are met.

Example 9 (Absolute Difference in Means)**.** In a completely randomized experiment, we consider $T_{DiM}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \sqrt{N}|\hat{\tau}|$, with $f_\eta(t) = |t|$ and $\hat{\xi} = 1$. Condition 3 is trivially satisfied since $\hat{\xi}$ is not stochastic. Condition 2 follows from the continuity, convexity, non-negativity, and symmetry of the absolute value function.

Example 10 (Multivariate studentization)**.** Let $\sqrt{N}\hat{\boldsymbol{\tau}}$ now be multivariate and suppose we have a completely randomized design. We examine the statistic

$$T_{\chi^2}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right)^{\mathrm{T}} \hat{V}_{\tau\tau}^{-1} \left(\sqrt{N}\hat{\boldsymbol{\tau}}\right), \tag{32}$$

with $\hat{V}_{\tau\tau} = \frac{N}{n_1}\hat{\Sigma}_{y(1)} + \frac{N}{n_0}\hat{\Sigma}_{y(0)}$. For this test statistic, $f_\eta(\mathbf{t}) = \mathbf{t}^{\mathrm{T}}\eta^{-1}\mathbf{t}$ and $\hat{\xi} = \hat{V}_{\tau\tau}$. Since $\hat{\xi}$ matches the top-left block of the Neyman covariance estimator Lemma A.7 shows that Condition 3 is met. Condition 2 holds because the quadratic form $f_\eta(\mathbf{t}) = \mathbf{t}^{\mathrm{T}}\eta^{-1}\mathbf{t}$ is certainly mirror symmetric, jointly continuous and convex by standard results for quadratic forms, and is non-negative since η is positive definite, and so its inverse must be as well.

The analysis for T_{pool} follows similar logic, but with the added observation that

$$\hat{V}_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} \frac{\Sigma_{y(0),\infty}}{p} + \frac{\Sigma_{y(1),\infty}}{1-p}$$
$$\hat{V}_{Pool}(\mathbf{y}(\mathbf{Z}), \mathbf{W}) \xrightarrow{p} \frac{\Sigma_{y(0),\infty}}{p} + \frac{\Sigma_{y(1),\infty}}{1-p}.$$

This follows because, under our assumptions,

$$\hat{\Sigma}_{y(1)}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} \Sigma_{y(1),\infty},$$
$$\hat{\Sigma}_{y(0)}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) \xrightarrow{p} \Sigma_{y(0),\infty}$$

while both $\hat{\Sigma}_{y(1)}(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ and $\hat{\Sigma}_{y(0)}(\mathbf{y}(\mathbf{Z}), \mathbf{W})$ converge in probability to $p\Sigma_{y(1),\infty} + (1 - p)\Sigma_{y(0),\infty}$.

Example 11 (Max absolute t-statistic)**.** Consider again multivariate $\sqrt{N}\hat{\boldsymbol{\tau}}$ and a completely randomized design. The max-absolute t-statistic is

$$T_{|max|}(\mathbf{y}(\mathbf{Z}), \mathbf{Z}) = \max_{1 \leq j \leq d} \frac{\sqrt{N}|\hat{\tau}_j|}{\sqrt{\hat{V}_{\tau\tau,jj}}},$$

where $\hat{V}_{\tau\tau,jj}$ is the jj element of $\hat{V}_{\tau\tau}$. For this statistic, $f_{\boldsymbol{\eta}}(\mathbf{t}) = \max_{1 \leq j \leq d} |t_j|/\eta_j$, and $\hat{\boldsymbol{\xi}} = (\hat{V}^{1/2}_{\tau\tau,11}, ..., \hat{V}^{1/2}_{\tau\tau,dd})^{\mathrm{T}}$. Since $\hat{\boldsymbol{\xi}}$ is the square-root of the diagonal elements of $\hat{V}_{\tau\tau}$, Lemma A.7 again establishes Condition 3. Certainly each coordinate projection $|t_j|/\eta_j$ is jointly continuous in $\eta_j > 0$ and t_j. Taking the maximum over these functions preserves continuity. The maximum of linear functions is convex so $f_{\boldsymbol{\eta}}(\mathbf{t})$ is quasi-convex. Non-negativity and mirror symmetry are trivial algebraic properties inherited from the coordinate-wise absolute value function. Thus, Condition 2 holds.

Example 12 (Rerandomization)**.** Consider a rerandomized design with balance criterion ϕ satisfying Condition 1 and let $\sqrt{N}\hat{\tau}$ be univariate. Consider the absolute difference in means, $f_{\hat{\xi}}(\sqrt{N}\hat{\tau}) = \sqrt{N}|\hat{\tau}|$, such that $\hat{\xi} = 1$. As before, since $\hat{\xi}$ is non-stochastic Condition 3 is immediate. The continuity, quasi-convexity, non-negativity, and symmetry of f_η are immediate consequences of the properties of the absolute value function. Thus, Condition 2 holds.

3.20 A Case Study With Educational Data

We demonstrate inference using Gaussian prepivoting in a completely randomized experiment. (ALO09) implemented a moderate-scale completely randomized experiment to test the effectiveness of several strategies intended to boost academic performance. Their experiment, the so-called *Student Achievement and Retention* (STAR) project, enrolled incoming first-year undergraduate students – except those with high-school grade point average (GPA) in the top 25% – in one of three treatment arms: a student support program, a financial incentive program, or both. Allocation to the programs was performed completely at random. Numerous demographic features of program participants were collected; we focus specifically on the participants' reported genders and high-school GPAs. The primary outcomes of the study were first-year GPA and second-year GPA. Further details on the nature of the interventions and the specific demographic features collected can be found in (ALO09). The data collected in the STAR project is publicly available in the online supplement to (ALO09).

(ALO09) found no evidence to suggest that the program was effective at improving educational outcomes among participants who identified as men. Lin (Lin13) used regression-adjusted estimators to examine inference for the marginal effect of offering financial incentives given that support services were offered; his analysis focuses on only the male participants of the study. Lin performs several simulations under the assumption that Fisher's sharp null holds, but he remarks that

> Chung and Romano (2011a, 2011b) discuss and extend a literature on permutation tests that do remain valid asymptotically when the null hypothesis is weakened. One such test is based on the permutation distribution of a heteroskedasticity-robust t-statistic. Exploration of this approach under the Neyman model (with and without covariate adjustment) would be valuable.

Gaussian prepivoting allows us to meet and exceed this objective: a permutation-testing framework can be applied with asymptotic validity under H_N for a wide class of statistics

– including multivariate statistics for which studentization alone is insufficient to restore asymptotic conservativeness. Theorem 1 guarantees finite sample exactness under H_F and asymptotic conservativeness under H_N of the prepivoted test statistic $G(\mathbf{y}(\mathbf{Z}), \mathbf{Z})$ subject to mild conditions; moreover, Section 3.12 of the supplement extends this result to the context of regression-adjusted estimators.

We re-analyze the data studied by (ALO09) through the lens of Gaussian prepivoting. Instead of restricting to the univariate outcome of first-year GPA, we examine the effect of treatment on both first-year and second-year GPA. We implement prepivoting using the test statistics of Section 3.5:

- the Euclidean 2-norm of the difference in means, denoted $T_{||\cdot||_2}$,
- the multivariate studentized statistics T_{χ^2} and T_{Pool},
- the maximum absolute t-statistic $T_{|max|}$.

Furthermore, we implement prepivoting in the cases above using the regression adjusted estimator of the difference in means – regressing on high-school GPA – instead of the naïve difference in means. In total $N = 141$ male-identifying participants have complete covariate and outcome data (high-school GPA, first and second year GPAs, respectively) and were offered *at least* support services. Of these individuals, $n_1 = 55$ were offered both support services and financial incentives while $n_0 = 86$ received only the offer for support services.

Table 3.5 contains p-values of the Fisher Randomization Test before and after prepivoting. The p-values obtained after prepivoting provide exact inference for H_F; moreover, the asymptotic results of Theorems 1 and 2 suggest that these p-values are likely to provide conservative inference for H_N. We stress that without Gaussian prepivoting only for T_{χ^2} would the Fisher Randomization Test be appropriate for Neymanian inference. The other three base statistics of Table 3.5 can exhibit asymptotically anti-conservative inference with the Fisher Randomization Test under H_N. For both $T_{||\cdot||_2}$ and T_{Pool} – with and

Table 3.5: p-values of the Fisher Randomization Test with and without using Gaussian prepivoting. The left two numerical columns use the Fisher Randomization Test directly on the base statistic without prepivoting. The right two numerical columns apply Gaussian prepivoting to the base statistic before using the Fisher Randomization Test. In "With Adj." columns linear regression adjustment using high-school GPA was applied to estimate the difference in means; "Without Adj." columns perform no regression adjustment.

	No Prepivoting		Prepivoting	
Base Statistic	Without Adj.	With Adj.	Without Adj.	With Adj.
$T_{\|\|\cdot\|\|_2}$	0.140	0.095	0.154	0.104
T_{χ^2}	0.159	0.126	0.159	0.126
T_{Pool}	0.141	0.107	0.153	0.121
$T_{\|max\|}$	0.181	0.129	0.174	0.122

without regression adjustment – the p-value obtained after prepivoting is no less than than the p-value derived without first prepivoting. These increased p-values suggest that the non-prepivoted procedures for $T_{||\cdot||_2}$ and T_{Pool} may have been anti-conservative. In fact, with $\alpha = 0.1$ an experimenter erroneously using the Fisher Randomization Test based upon regression adjustment with $T_{||\cdot||_2}$ to test H_N would have rejected the null of no average effect. Once prepivoting is applied the practitioner is asymptotically entitled to test H_N with the Fisher Randomization Test and we observe that the procedure no longer rejects H_N, thereby rectifying the potentially anti-conservative nature of the preceding result.

The code to implement our analysis is provided online to facilitate reproducibility.

3.21 Software

Code written in `R` that builds the figures above is available online at `https://github.com/PeterLCohen/PrepivotingCode`. Furthermore, at the same location, we provide concrete examples – also written in `R` code – illustrating how one might choose to implement Gaussian prepivoting from scratch. For simplicity, we present prepivoting the absolute $\sqrt{N}$-scaled difference in means for a univariate completely randomized design. In other words, we

exactly demonstrate the implementation of Algorithm 1 for Gaussian prepivoting used in Example 1 of Section 3.5. At the same location, we provide `R` code to reproduce the results of the data analysis in Section 3.20.

Bibliography

[ADSZ10] Mordecai Avriel, Walter E. Diewert, Siegfried Schaible, and Israel Zang. *Generalized Concavity*. Society for Industrial and Applied Mathematics, 2010.

[AI08] Alberto Abadie and Guido W. Imbens. Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, (91/92):175–187, 2008.

[ALO09] Joshua Angrist, Daniel Lang, and Philip Oreopoulos. Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–63, January 2009.

[And55] T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.*, 6:170–176, 1955.

[AT09] Robert J Adler and Jonathan E Taylor. *Random Fields and Geometry*. Springer Science & Business Media, 2009.

[BDR16] Zach Branson, Tirthankar Dasgupta, and Donald B. Rubin. Improving covariate balance in 2 k factorial designs via rerandomization with an application to a new york city department of education high school study. *Ann. Appl. Stat.*, 10(4):1958–1976, 12 2016.

[Ber87] Rudolf Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468, 1987.

[Ber88] Rudolf Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

[CB90] George Casella and Roger L. Berger. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[CDM17] Devin Caughey, Allan Dafoe, and Luke Miratrix. Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*, 2017.

[Cha05] Sourav Chatterjee. An error bound in the Sudakov-Fernique inequality. *arXiv preprint math/0510424*, 2005.

[CK17] Aura Conci and Carlos Kubrusly. Distances between sets—a survey. *Adv. Math. Sci. Appl.*, 26(1):1–18, 2017.

[CR16] EunYi Chung and Joseph P. Romano. Multivariate and multiple permutation tests. *J. Econometrics*, 193(1):76–91, 2016.

[Cra02] Hans Crauel. *Random probability measures on Polish spaces*, volume 11 of *Stochastics Monographs*. Taylor & Francis, London, 2002.

[DD18] Peng Ding and Tirthankar Dasgupta. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, 105(1):45–56, 2018.

[DDCZ13] Lutz Dümbgen and Perla Del Conte-Zerial. On low-dimensional projections of high-dimensional distributions. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 91–104. Inst. Math. Statist., Beachwood, OH, 2013.

[DFM19] Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

[Din17] Peng Ding. A paradox from randomization-based causal inference. *Statistical Science*, 32(3):331–345, 2017.

[DLM17] Peng Ding, Xinran Li, and Luke W. Miratrix. Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2), apr 2017.

[Fis35] Ronald Aylmer Fisher. *The Design of Experiments*. Oliver & Boyd, Edinburgh, 1935.

[FLKK19] Colin B Fogarty, Kwonsang Lee, Rachel R Kelz, and Luke J Keele. Biased encouragements and heterogeneous effects in an instrumental variable study of emergency general surgical outcomes. *arXiv preprint arXiv:1909.09533*, 2019.

[Fog18] Colin B Fogarty. Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4):994–1000, 2018.

[Fog20] Colin B Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, 115(531):1518–1530, 2020.

[Fre08a] David A. Freedman. On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.*, 2(1):176–196, 2008.

[Fre08b] David A. Freedman. On regression adjustments to experimental data. *Adv. in Appl. Math.*, 40(2):180–193, 2008.

[GM07] Mariano Giaquinta and Giuseppe Modica. *Mathematical analysis*. Birkhäuser Boston, Inc., Boston, MA, 2007. Linear and metric structures and continuity.

[HS79] Harold V. Henderson and S. R. Searle. Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 7(1):65–81, 1979.

[Ima08] Kosuke Imai. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Stat. Med.*, 27(24):4857–4873, 2008.

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[LD17] Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.*, 112(520):1759–1769, 2017.

[LDR18] Xinran Li, Peng Ding, and Donald B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

[LDR20] Xinran Li, Peng Ding, and Donald B. Rubin. Rerandomization in 2 k factorial experiments. *Ann. Statist.*, 48(1):43–63, 02 2020.

[Lin13] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

[LR05] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[LRR17] Wen Wei Loh, Thomas S Richardson, and James M Robins. An apparent paradox explained. *Statistical Science*, 32(3):356–361, 2017.

[MR12] Kari Lock Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *Ann. Statist.*, 40(2):1263–1282, 04 2012.

[Ney90] Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.*, 5(4):465–472, 1990. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed.

[Ros02] Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer-Verlag, New York, Second edition, 2002.

[Rub80] Donald B Rubin. Comment on "Randomization analysis of experimental data: The Fisher randomization test". *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[Sle62] D. Slepian. The one-sided barrier problem for Gaussian noise. *The Bell System Technical Journal*, 41(2):463–501, March 1962.

[Ton90] Y. L. Tong. *The Multivariate Normal Distribution*. Springer Series in Statistics. Springer-Verlag, New York, 1990.

[TPM11] Armando Teixeira-Pinto and Laura Mauri. Statistical analysis of noncommensurate multiple outcomes. *Circulation: Cardiovascular Quality and Outcomes*, 4(6):650–656, 2011.

[TPSGN09] A. Teixeira-Pinto, J. Siddique, R. Gibbons, and S. L. Normand. Statistical Approaches to Modeling Multiple Outcomes In Psychiatric Studies. *Psychiatr Ann*, 39(7):729–735, Jul 2009.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[WD18] Jason Wu and Peng Ding. Randomization tests for weak null hypotheses. *arXiv e-prints*, page arXiv:1809.07419, Sep 2018.

Chapter 4

Hierarchical Resampling Procedures for Causal Inference

Abstract

Causal inference has traditionally divided across several schools of thought: superpopulation inference, fixed covariate inference, and finite population inference. The work of (RBK05) rigorously specifies a probabilistic structure for which these three models can be viewed as a nested hierarchy defined by increasingly large conditioning events. Inference in the superpopulation setting using resampling procedures has been well understood, but resampling methods in the fixed covariate and finite population contexts remain nascent areas of research. We construct a family of resampling procedures which provide asymptotically valid inference for a wide variety of test statistics while respecting the hierarchical structure of the three models. A natural byproduct of our analysis is a novel derivation of Neyman's classical variance estimator for the difference in means in completely randomized experiments and two new variance estimators for the difference in means that exploit conditioning event structure to outperform Neyman's classical estimator in fixed covariate models and finite population models. We illustrate the generality of our techniques through a host of examples; and provide simulation studies to illustrate practical performance.

4.1 Introduction

Beginning with the work of Efron in the 1980s, bootstrap resampling has played a central role in modern statistical inference. For excellent introductions to the field, we direct the reader towards (Efr82, ST95). Initial applications of the bootstrap were primarily directed towards inference for independent and identically distributed data, though quickly the methodology branched out to accommodate data which lacks the regularity of *i.i.d.* data; e.g., (Liu88, LS95, MHL16). In this paper we focus upon resampling procedures in the context of nonparametric inference for causal parameters. Our objective is to provide computationally efficient procedures for asymptotic hypothesis testing and confidence interval formation which can be applied with minimal assumptions on the structure of the data generating process; such results are immediately applicable to a broad variety of applied fields ranging from biostatistics to econometrics, public policy, and many more.

Several different models exist for causal inference; these models vary in terms of which quantities in the model are random. Superpopulation inference takes each individual to have

potential outcomes and covariates which are random; fixed covariate inference takes only the potential outcomes to be random; and finite population inference takes both the potential outcomes and covariates to be fixed - with the only randomness entering the model through the treatment allocation mechanism. Frequently these models are viewed as disparate in the sense that inferential procedures are designed to perform well in one of the models while disregarding the performance - or even validity - of the procedure in the other models. However, these models form a natural hierarchy: superpopulation inference takes all quantities to be random, finite population inference takes only treatment allocation to be random, and fixed covariate inference bridges the gap between the two. The work of (RBK05) codifies a probabilistic framework wherein finite population inference is viewed as a conditional sub-model of fixed covariate inference and likewise fixed covariate inference is a conditional sub-model of superpopulation inference. This establishes a natural hierarchy based upon conditioning upon progressively larger events. Ideally one would tailor their resampling algorithm to the available structure of one's preferred inferential framework; as a rough heuristic it is typically held that mimicking the data generating process results in sensible resampling algorithms (HW91). Perhaps this is most easily seen in the superpopulation setting where observed data points are independent and identically distributed; there the natural bootstrap algorithm resamples by drawing independent and identically distributed draws from the observed empirical distributions. Adapting resampling algorithms to the fixed covariate and finite population framework is more subtle; we provide a hierarchy of bootstrap resampling procedures – each one building upon the last – which accord with the structure of the three probabilistic models. Our resampling algorithms provide asymptotically valid hypothesis tests against weak null hypotheses without ascribing to parametric or other rigid modeling assumptions.

4.2 Notation

We consider a population of N total number of units; n_1 receive treatment and $n_0 = N - n_1$ receive control. The i^{th} unit has control potential outcome $y_i(0) \in \mathbb{R}^d$, treated potential outcome $y_i(1) \in \mathbb{R}^d$, and features $x_i \in \mathbb{R}^k$. We write expectations as $\mathbb{E}[\cdot]$ and variances are written as $\mathbb{V}(\cdot)$. The space of all probability measures on a set S is $\mathcal{M}(S)$ and the δ-point-mass at $x \in S$ is written δ_x. For $\mathcal{X}$ a random variable taking in values in S the probability measure defining $\mathcal{X}$ is $\mathcal{L}(\mathcal{X})$ which takes values in $\mathcal{M}(S)$. The Gaussian measure with mean μ and covariance matrix Σ is written $\gamma_{\mu,\Sigma}$. For a vector A the subvector $A_{[I]}$ is the vector of A's coordinates indexed by $i \in I$. The set $I_0 = \{1, \ldots, d, 2d+1, \ldots, 2d+k\}$ indexes the coordinates of the control outcomes and the covariates jointly. Likewise, the set $I_1 = \{d+1, \ldots, 2d, 2d+1, \ldots, 2d+k\}$ indexes the coordinates of the treated outcomes and the covariates jointly. For a vector (y, x) the projection operator onto the coordinates of y is written Π_y so that $\Pi_y(y, x) = y$. The indicator function of an event E is written $\mathbb{1}_{\{E\}}$. If a sequence of random variables $\mathcal{X}^{(N)}$ converges in probability to a random variable $\mathcal{X}$, then we write $plim_{N\to\infty} \mathcal{X}^{(N)} = \mathcal{X}$. Given a multiset S of cardinality $|S| < \infty$ define $\mathcal{P}(N, S)$ to be the collection of all multisets S' of cardinality N with elements drawn from S. Importantly, $\mathcal{P}(N, S)$ is defined even when $N > |S|$. The collection $\mathcal{P}(N, S)$ is the collection of possible samples (with replacement) of size N drawn from S.

One of our key results is a relationship between conservative covariance estimation and first-order stochastic dominance. To rigorously define conservativeness for $d \times d$ covariance matrices when $d \geq 2$ we turn to the *Loewner partial order*. Given two positive semidefinite matrices M_1 and M_2: M_1 is no smaller than M_2 in the Loewner partial order if and only if $M_1 - M_2$ is positive semidefinite; we write $M_1 \succeq M_2$.

Our bootstrap consistency results utilize the *bounded-Lipschitz* metric. The bounded-

Lipschitz metric between two probability measures μ and ν be defined as

$$\rho_{BL}(\mu,\nu) := \sup_{\substack{f\in Lip_1 \\ ||f||_\infty \leq 1}} \left| \int f(x)d\mu(x) - \int f(x)d\nu(x) \right|.$$

Weak converges of measures in $\mathcal{M}(\mathbb{R})$ is metrized by the bounded-Lipschitz metric (vdVW96, Theorem 1.12.4).

4.3 Probabilistic Framework

4.3.1 A Nested Sequence of Models

We adapt the work of (RBK05) to create a nested sequence of probability spaces upon which our analyses will proceed. Consider a probability space $(\Phi, \mathcal{F}, P)$ from which we form a population of N individuals with potential outcome $y_i(z) = \mathcal{Y}_{z,i}(\omega)$ for $z \in \{0,1\}$ and covariates $x_i = \mathcal{X}_i(\omega)$ for $\mathcal{Y}_{z,i} : \Phi \to \mathbb{R}^d$ and $\mathcal{X}_i : \Phi \to \mathbb{R}^k$ measurable functions of $\omega \in \Phi$.[1] The generative model of the collection $\{(y_i(0), y_i(1), x_i)\}_{i=1}^N$ forms our superpopulation model. We take the $(y_i(0), y_i(1), x_i)$ to be independent and identically distributed.

To form the fixed covariate model, let

$$\mathscr{C} = \{\omega \in \Phi \ : \ \mathcal{X}(\omega_i) = \mathbf{x}_i \text{ for } i = 1, \ldots, N\}.$$

The set $\mathscr{C}$ is the event that the covariates of the N individuals are given by the deterministic values $\{\mathbf{x}_i\}_{i=1}^N$. Let $P_{\mathscr{C}}$ be the conditional probability measure derived from P conditioned on the event $\mathscr{C}$. In the case that $\mathscr{C}$ is an event of P-measure zero, we tacitly assume that there exists a well-defined regular conditional probability measure and take $P_{\mathscr{C}}$ to be this conditional; see (CT97, Section 7.2) for more details on this technical issue. Inferences

[1] We take k and d as fixed quantities which do not grow with N.

under the fixed-covariate model take $(\Omega, \mathcal{F}, P_{\mathscr{C}})$ to generate the outcomes $y_i(z) = \mathcal{Y}_z(\omega_i)$ for $z \in \{0,1\}$ and implicitly constrain the covariates $x_i = \mathcal{X}(\omega_i) = \mathbf{x}_i$ for $i = 1, \ldots, N$. We write the conditional distribution of $(y_i(0), y_i(1))$ given $\mathbf{x}_i$ as $P_{\mathbf{x}_i}$.

Finally, the finite population framework takes a fully conditional viewpoint by conditioning upon

$$\mathscr{F} = \{\omega \in \Phi \ : \ \mathcal{X}(\omega_i) = \mathbf{x}_i, \mathcal{Y}_z(\omega_i) = \mathbf{y}_i(z) \text{ for } i = 1, \ldots, N \ ; \ z \in \{0,1\}\}\,.$$

In this case, the entire population of N individuals is fully deterministic.

4.3.2 Population Conditional Measures and Average Treatment Effects

In the superpopulation model, the probability measure $\mathcal{L}\left((y_1(0), y_1(1), x_1)\right)$ encodes all of the required information to understand the generative model of

$$\{(y_i(0), y_i(1), x_i)\}_{i=1}^N\,.$$

On the other extreme, in the finite population model the N points in $\mathbb{R}^{d+d+k}$ given by $\{(\mathbf{y}_i(0), \mathbf{y}_i(1), \mathbf{x}_i)\}_{i=1}^N$ contain all of the information of the model. Interpolating between these two extremes, in the fixed covariate model, the N points $\{\mathbf{x}_i\}_{i=1}^N$ and the N conditional probability measures $\{P_{\mathbf{x}_i}\}_{i=1}^N$ fully dictate the structure of the model. Below, we define a single object parameterized by a P-measurable conditioning event $\mathscr{S} \in \mathcal{F}$ that exactly encapsulates these notions.

Definition 3. For a measurable event $\mathscr{S}$ the *population $\mathscr{S}$-conditional measure* is the uniform mixture over the conditional measures on $\mathbb{R}^{d+d+k}$ induced by $(y_i(0), y_i(1), x_i)$ given the event $\mathscr{S}$; when $\mathscr{S}$ is of P-measure zero we assume that there exists a well-defined regular

conditional probability measure and take this conditional (CT97, Section 7.2). Denote this measure as $\mathcal{P}_{\mathscr{S}}$; formally,

$$\mathcal{P}_{\mathscr{S}} = \sum_{i=1}^{N} N^{-1} \mathcal{L}\left(y_i(0), y_i(1), x_i\right).$$

Let $\Delta : \mathcal{M}(\mathbb{R}^{d+d+k}) \to \mathbb{R}^d$ be the linear operator which maps a measure to the expected difference between the second d coordinates and the first d coordinates; formally,

$$\Delta(M) = \mathbb{E}\left[A_{[I_1]} - A_{[I_0]}\right] \quad \textit{for } A \sim M.$$

Definition 4. The *population $\mathscr{S}$-average treatment effect* is the expected difference between the second and first d coordinates under the population $\mathscr{S}$-conditional measure; we denote this $\overline{\tau}_{\mathscr{S}}$. Formally

$$\overline{\tau}_{\mathscr{S}} = \Delta\left(\mathcal{P}_{\mathscr{S}}\right) = \frac{1}{N} \sum_{i=1} \mathbb{E}\left[y_i(1) - y_i(0) \mid \mathscr{S}\right].$$

Definitions 3 and 4 provide a unified perspective upon common features of the three usual levels of inference:

- **(Superpopulation Model)** When $\mathscr{S} = \emptyset$ all of the $(y_i(0), y_i(1), x_i)$ are *i.i.d.* so the population $\emptyset$-conditional measure is the probability measure defining $(y_1(0), y_1(1), x_1)$. The target of inference is the *population average treatment effect*,

$$\overline{\tau}_{\text{PATE}} = \mathbb{E}\left[y_i(1) - y_i(0)\right] = \overline{\tau}_{\emptyset}.$$

- **(Fixed Covariate Model)** When $\mathscr{S} = \mathscr{C}$ the population $\mathscr{C}$-conditional measure is the uniform mixture over the distributions of $(y_i(0), y_i(1), \mathbf{x}_i)$ where $(y_i(0), y_i(1)) \sim P_{\mathbf{x}_i}$;

thus, the population $\mathscr{C}$-conditional measure is

$$\mathcal{P}_{\mathscr{C}} = \frac{1}{N}\sum_{i=1}^{N}(P_{\mathbf{x}_i}, \delta_{\mathbf{x}_i}),$$

where $(P_{\mathbf{x}_i}, \delta_{\mathbf{x}_i})$ is the measure on $\mathbb{R}^{d+d+k}$ where the first $2d$ coordinates are distributed according to $P_{\mathbf{x}_i}$ and the final k coordinates are deterministically $\mathbf{x}_i$. The target of inference is the *conditional average treatment effect*

$$\overline{\tau}_{\text{CATE}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[y_i(1) - y_i(0) \mid x_i\right] = \overline{\tau}_{\mathscr{C}}.$$

- **(Finite Population Model)** When $\mathscr{S} = \mathscr{F}$ the points $\{(y_i(0), y_i(1), x_i)\}_{i=1}^{N}$ are deterministic and the population $\mathscr{F}$-conditional measure is just the uniform measure over these points; thus, the population $\mathscr{F}$-conditional measure is

$$\mathcal{P}_{\mathscr{F}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{(\mathbf{y}_i(0),\mathbf{y}_i(1),\mathbf{x}_i)}.$$

  The target of inference is the *sample average treatment effect*

$$\overline{\tau}_{\text{SATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i(1) - \mathbf{y}_i(0)\right) = \overline{\tau}_{\mathscr{F}}.$$

Importantly, the central objects of our interest can be represented as the image of a linear operator applied to the population $\mathscr{S}$-conditional measure. Consequently, construction of empirical approximations to the population $\mathscr{S}$-conditional measure will play crucially into our analyses.

In each of the three models of Section 4.3.1, we define Neyman's weak null as follows.

Definition 5. For a P-measurable event $\mathscr{S}$, Neyman's $\mathscr{S}$-weak null is the null hypothesis

$$H_{N,\mathscr{S}} : \overline{\tau}_{\mathscr{S}} = 0.$$

Consequently, Neyman's $\emptyset$-weak null is the null $\overline{\tau}_{\text{PATE}} = 0$, Neyman's $\mathscr{C}$-weak null is the null $\overline{\tau}_{\text{CATE}} = 0$, and Neyman's $\mathscr{F}$-weak null is the null $\overline{\tau}_{\text{SATE}} = 0$.

Most common test statistics deployed in practice for inferences of $H_{N,\mathscr{S}}$ are of the form

$$T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z})) = f_{\hat{\xi}}\left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)\right) = f_{\hat{\xi}}\left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \Delta\left(\mathcal{P}_{\mathscr{S}}\right)\right)\right), \tag{1}$$

where $\hat{\tau}(\mathbf{y}(\mathbf{Z})) = n_1^{-1}\sum_{i\,:\,Z_i=1} y_i(Z_i) - n_0^{-1}\sum_{i\,:\,Z_i=0} y_i(Z_i)$ is the observed difference in means and f_η satisfies the following condition:

Condition 3. *For any $\eta \in \Xi$, $f_\eta(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}_+$ is continuous, quasi-convex, and nonnegative with $f_\eta(t) = f_\eta(-t)$ for all $t \in \mathbb{R}^d$. Furthermore, $f_\eta(t)$ is jointly continuous in η and t.*

We allow for f to depend upon a data-dependent parameter $\hat{\xi}(\mathbf{y}(\mathbf{Z}))$ which takes values in some metric space Ξ. Later on we will introduce a mild regularity condition for $\hat{\xi}(\cdot)$ in order to enforce its compatibility with the resampling algorithms we develop. We include several example test statistics that are common for applied use which satisfy Condition 3 and are compatible with the methods and further conditions which we introduce below.

Example 13 (Absolute Difference in Means)**.** Suppose the potential outcomes are univariate, then the absolute difference in means $\sqrt{N}\,|\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}|$ satisfies Condition 3 with $f_\eta(t) = |t|$ and $\hat{\xi} = 1$.

Example 14 (Multivariate Studentized Statistic)**.** Take $d \geq 1$ and consider the multivariate

studentized statistic

$$T^{(\chi^2)}(\mathbf{y}(\mathbf{Z})) = \left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)\right)^T \hat{V}_{\tau\tau}^{-1} \left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)\right); \quad (2)$$
$$\hat{V}_{\tau\tau} = \frac{N}{n_1}\hat{\Sigma}_{y(1)} + \frac{N}{n_0}\hat{\Sigma}_{y(0)}.$$

with $\hat{\Sigma}_{y(Z)}$ the sample covariance matrix of the $y(z)$ potential outcomes. For this test statistic, $f_\eta(t) = t^T\eta^{-1}t$ and $\hat{\xi} = \hat{V}_{\tau\tau}$. This test statistic benefits from numerous desirable properties and is proposed for use in applied multivariate causal inference contexts (WD21).

Example 15 (Pooled Studentized Statistic)**.** Again with $d \geq 1$ a practitioner may instead consider proceeding with the more typical Hotelling T-squared statistic based upon a pooled covariance estimator

$$T^{(Pool)}(\mathbf{y}(\mathbf{Z})) = \left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)\right)^T \left(\hat{V}_{Pool}\right)^{-1} \left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)\right);$$
$$\hat{V}_{Pool} = \left(\frac{N}{n_0} + \frac{N}{n_1}\right)\left(\frac{(n_1-1)\hat{\Sigma}_{y(1)} + (n_0-1)\hat{\Sigma}_{y(0)}}{n_1+n_0-2}\right).$$

For this test statistic, $f_\eta(t) = t^T\eta^{-1}t$ as before, but $\hat{\xi} = \hat{V}_{Pool}$. While this test statistic lacks the desirable properties of $T^{(\chi^2)}$ detailed by (WD21) our methods below will still provide asymptotically valid inferences for practitioners wishing to use $T^{(Pool)}$.

Example 16 (Maximum Absolute t-Statistic)**.** As an alternative to $T^{(\chi^2)}$ and $T^{(Pool)}$ one may instead seek to extract the largest signal-to-noise ratio across each of the outcomes and use this as evidence of treatment effect; this amounts to

$$T^{(|max|)}(\mathbf{y}(\mathbf{Z})) = \max_{1\leq j\leq d} \frac{\sqrt{N}\left|\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right|_j}{\sqrt{\hat{V}_{\tau\tau,jj}}},$$

where $\hat{V}_{\tau\tau,jj}$ is the jj^{th} element of $\hat{V}_{\tau\tau}$. For this statistic, $f_\eta(t) = \max_{1\leq j\leq d}|t_j|/\eta_j$, and

$\hat{\xi} = (\hat{V}^{1/2}_{\tau\tau,11}, \ldots, \hat{V}^{1/2}_{\tau\tau,dd})^T$. In the univariate case, this coincides with $T^{(\chi^2)}$ but its behavior diverges sharply when $d > 1$.

4.3.3 The Experimental Design

The algorithms of this paper are developed in the context of asymptotically non-degenerate completely randomized experiments; some of the results are amenable to different experimental designs with only minor modification. The set of allowable treatment allocation vectors is $\Omega \subseteq \{0,1\}^N$ and we take our treatment allocation mechanism to be $Z \sim \mathit{Unif}(\Omega)$. We consider completely randomized experiments such that

$$\Omega = \left\{ z \in \{0,1\}^N \ : \ \sum_{i=1}^{N} z_i = n_1 \right\}.$$

These the completely randomized design of $Z \sim \mathit{Unif}(\Omega)$ is thus the experimental design which selects uniformly at random n_1 units to receive treatment and leaves the remaining $n_0 = N - n_1$ units to receive control. Generalizations to any finite number of treatment arms is not challenging; details of the modifications required are included in (CF22, Appendix Section E). We require that the treatment allocation is asymptotically non-degenerate; i.e., there exists $p \in (0,1)$ such that $\lim_{N\to\infty} N^{-1} n_1 = p$.

4.4 Conservative Resampling Algorithms And Error Rate Control

In fixed covariate and finite population models first-order correct inference against $H_{N,\mathscr{S}}$ for test statistics of the form (1) is impossible without further assumptions. This is fundamentally due to the unidentifiability of the variability in $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ driven by variance in the treatment effects $y_i(1) - y_i(0)$, an observation which dates back to Jerzy Neyman (Ney90).

Instead, one is forced to seek potentially conservative inference: while desiring to control the asymptotic Type I error rate at exactly some prescribed $\alpha \in (0,1)$ it is generally only possible to guarantee that the Type I error rate may be less than or equal to α in the limit. Below we construct a general understanding resampling algorithms which may not be first-order correct but nonetheless retain asymptotic control of the probability of false rejection.

Let $\mathcal{A}_T(Z)$ denote a algorithm which takes observed units $\{(y_i(Z_i), x_i\}_{i=1}^N$ as inputs and generates a random variable conditional upon $\{(y_i(Z_i), x_i\}_{i=1}^N$. We consider algorithms which create some imputed population based upon the observations $\{(y_i(Z_i), x_i\}_{i=1}^N$ and evaluates some test statistic $T_{\mathscr{S}}(\cdot)$ upon this new population; as such we denote the output of $\mathcal{A}_T(Z)$ as $T^*(y^*(Z))$. In the remainder of our analyses we examine the conditional distribution of $T^*(y^*(Z))$ given $\{(y_i(Z_i), x_i\}_{i=1}^N$ and Z; through this lens one can view $\mathcal{A}_T(Z)$ as constructing a distributional estimator $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ which captures the salient features of the inference problem at hand. In classical *i.i.d.* models, resampling algorithms are frequently designed such that $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ converges weakly either almost surely or in probability to the true distribution of the test statistic at hand; frequently such analyses leverage Glivenko-Cantelli-style results and an argument of (LR05, Theorem 15.4.1). Below we show that such techniques are applicable for superpopulation level inference problems, but are fundamentally impeded by the unidentifiable variance of treatment effects in fixed covariate and finite population models. Instead we develop a framework based upon first-order stochastic dominance (SS07).

Definition 6. We say that a resampling procedure $\mathcal{A}_T(Z)$ which generates a distributional estimator $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ is *(strongly) consistently conservative* if

$$\rho_{BL}\left(\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}(\mathcal{X})\right) \xrightarrow{a.s.} 0$$

where $\mathcal{L}(\mathcal{X})$ (first-order) stochastically dominates the weak limit of $\mathcal{L}\left(T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))\right)$ under the null $H_{N,\mathscr{S}}$.

A strongly consistently conservative resampling algorithm automatically provides hypothesis tests with guaranteed control of the probability of false rejection. Let Q^* denote the quantile function of $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ so that $Q^*(1-\alpha)$ is the $(1-\alpha)^{\text{th}}$ quantile of $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$. Define the test $\varphi_\alpha(\mathbf{y}(\mathbf{Z}))$ for $\alpha \in (0,1)$ as

$$\varphi_\alpha(\mathbf{y}(\mathbf{Z})) = \mathbb{1}_{\{T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z})) \geq Q^*(1-\alpha)\}}.$$

The test $\varphi_\alpha(\mathbf{y}(\mathbf{Z}))$ rejects the null when $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ exceeds the $(1-\alpha)^{\text{th}}$ quantile of

$$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right).$$

Informally stated, the probability of false rejection is controlled by strongly consistently conservative resampling algorithms.

Theorem 1. *If the data is generated according to $\mathcal{P}_{\mathscr{S}}$, $\mathcal{A}_T(Z)$ is strongly consistently conservative, and both $\mathcal{L}(\mathcal{X})$ and the weak limit of $\mathcal{L}\left(T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))\right)$ under the null $H_{N,\mathscr{S}}$ possess continuous and strictly increasing cumulative distribution functions then*

$$\lim_{N\to\infty} \mathbb{E}\left[\varphi_\alpha(\mathbf{y}(\mathbf{Z})) \mid H_{N,\mathscr{S}}\right] \leq \alpha \quad \forall \alpha \in (0,1).$$

The main contribution of this paper is twofold: the construction of a generic formulation of consistently conservative resampling procedures for roots of the form (1) and a collection of three example resampling procedures which are tailored to the superpopulation, fixed covariate, and finite population frameworks, respectively.

4.5 Constructing Conservative Resampling Algorithms Via Variance Overestimation

Although Definition 6 is quite general for test statistics in the form of (1) one can construct a generic recipe for consistently conservative resampling algorithms. This builds upon a fundamental result of Anderson (And55) which has a particularly clean interpretation for multivariate Gaussian random variables (Ton90, Theorem 4.2.5).

Theorem 2 (Anderson's Theorem)**.** *Consider two multivariate Gaussian random variables* $\mathcal{X} \sim \mathcal{N}(0, S_{\mathcal{X}})$ *and* $\mathcal{Y} \sim \mathcal{N}(0, S_{\mathcal{Y}})$*. If* $S_{\mathcal{X}} \preceq S_{\mathcal{Y}}$ *in the Loewner partial order then*

$$\mathbb{P}(\mathcal{X} \in B) \geq \mathbb{P}(\mathcal{Y} \in B)$$

for any convex set B *which is mirror symmetric about the origin (i.e.,* $x \in B \iff -x \in B$*).*

Mild conditions on $\mathcal{P}_{\mathscr{J}}$ ensure that $\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{J}}\right)$ obeys a central limit theorem; sufficient conditions are detailed explicitly in our supplementary material. The asymptotic variance of the limit is dependent upon the data generating procedure, specifically $\mathcal{P}_{\mathscr{J}}$; for now we simply write that $\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{J}}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. Assuming that $\hat{\xi}(\mathbf{y}(\mathbf{Z})) \xrightarrow{p} \xi$ it follows that the asymptotic distributional behavior of $T_{\mathscr{J}}(\mathbf{y}(\mathbf{Z}))$ is given by the f_{ξ}-pushforward of the Gaussian measure $\gamma_{0,\Sigma}$. Condition 3 ensures that the preimage of the set $(\infty, t]$ under f_{η} is convex and mirror symmetric for any $t \in \mathbb{R}_{\geq 0}$. Consequently, Anderson's theorem naturally suggests an approach for constructing consistently conservative resampling procedures: design $\mathcal{A}_T(Z)$ such that $\rho_{BL}\left(\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}(\mathcal{X})\right) \xrightarrow{a.s.} 0$ where $\mathcal{L}(\mathcal{X})$ is f_{ξ}-pushforward of the Gaussian measure $\gamma_{0,\tilde{\Sigma}}$ and $\tilde{\Sigma} \succeq \Sigma$.

Example 17 (Gaussian Parametric Bootstrap)**.** Consider the absolute difference in means statistic $\sqrt{N}\left|\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{J}}\right|$ and suppose a variance estimator $\hat{V}$ is available such that $\hat{V} \xrightarrow{a.s.} \tilde{\Sigma}$. Take $\mathcal{A}_T(Z)$ to generate $T^*(y^*(z)) = |\mathcal{X}^{(N)}|$ with $\mathcal{X}^{(N)}$ an independent draw

from $\mathcal{N}\left(0, \hat{V}\right)$. This procedure aligns with the concept of the *parametric bootstrap* (ST95, Example 1.6). Such parametric distributional estimators play a critical role for causal inference via Gaussian prepivoting (CF22).

A detailed analysis in Section 4.14.3 of our supplementary material provides proof that this recipe is indeed rigorously justified. The remainder of this paper is devoted to constructing three resampling procedures in accordance with this program which have more desirable properties than the simple parametric example presented above. From a practical perspective, the parametric bootstrap example proposed above fails to possess the automaticity of Efron's *i.i.d.* bootstrap; this automaticity is itself a clear driver of the pervasive use of the Efron's *i.i.d.* bootstrap in industry and applied sciences. From a theoretical perspective, in classical models the parametric bootstrap lacks the higher-order correctness properties of Efron's *i.i.d.* bootstrap (ST95, Section 3.3), (LR05, Section 15.5).

4.6 The *I.I.D.* Bootstrap At The Superpopulation Level

To use a test statistic $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ in a test of the superpopulation weak null $H_{N,\emptyset}$ it suffices to construct a sufficiently high fidelity estimate of the null distribution of $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$. A natural starting point for such an objective is to estimate

$$\mathcal{L}\left((y_i(0), y_i(1), x_i)\right).$$

However, in an experiment the pairs $(y_i(0), y_i(1))$ are never observed, so the joint structure of $\mathcal{L}\left((y_i(0), y_i(1), x_i)\right)$ is inherently unidentifiable. However, one may leverage the observed data $\{(y_i(Z_i), x_i, Z_i)\}_{i=1}^N$ to estimate the projections $\mathcal{L}\left((y_i(0), x_i)\right)$ and $\mathcal{L}\left((y_i(1), x_i)\right)$.

Given a treatment allocation vector $z \in \Omega$, the Horvitz-Thompson empirical measure is

defined as the pair of empirical measures

$$\hat{F}_N(z) = \left(\hat{F}_N^0(z), \hat{F}_N^1(z)\right) = \left(\frac{1}{n_0}\sum_{i\,:\,z_i=0}\delta_{(y_i(z_i),x_i)},\ \frac{1}{n_1}\sum_{i\,:\,z_i=1}\delta_{(y_i(z_i),x_i)}\right).$$

The measure $\hat{F}_N^1(Z)$ is the observed empirical joint measure of outcomes and covariates in the group of individuals which received treatment, and $\hat{F}_N^0(Z)$ is its control analogue. The Horvitz-Thompson empirical measure is an attempt to form an empirical approximation to the true, but unobserved, population $\mathscr{S}$-conditional measure. Although $\hat{F}_N(z)$ only approximates the projections of $\mathcal{L}\left((y_i(0), y_i(1), x_i)\right)$ we demonstrate that it suffices for inferences based upon test statistics of the form (1). At the superpopulation level this should be unsurprising since the distribution of $\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ depends only on the marginal structure $\mathcal{L}(y_i(0))$ and $\mathcal{L}(y_i(1))$. However, we show further on that resampling based upon the Horvitz-Thompson empirical measure is conservative under fixed covariate and finite population models despite the fact that the distribution of $\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ depends upon the joint structure of $\{(y_i(0), y_i(1))\}_{I=1}^N$ in these models.

Suppose that the test statistic $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ is the simple difference in means between the outcomes of the treated and control groups

$$\sqrt{N}\Bigg(\underbrace{\frac{1}{n_1}\sum_{i\,:\,Z_i=1} y_i(Z_i) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0} y_i(Z_i)}_{\hat{\tau}(\mathbf{y}(\mathbf{Z}))} - \overline{\tau}_{\mathscr{S}}\Bigg) =$$

$$\sqrt{N}\left(\mathbb{E}_{\hat{F}_N^1(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^0(Z)}\left[\Pi_y(y,x)\right] - \overline{\tau}_{\mathscr{S}}\right).$$

Writing F^0 as the projection of the distribution of $(y_1(0), y_1(1), x_1)$ onto the coordinates indexed by I_0 and F^1 as the analogous projection onto the coordinates indexed by I_1 we can rewrite $\overline{\tau}_{\emptyset} = \mathbb{E}_{F^1}\left[\Pi_y\left(y,x\right)\right] - \mathbb{E}_{F^0}\left[\Pi_y\left(y,x\right)\right]$. Then it follows that $T_{\emptyset}(\mathbf{y}(\mathbf{Z}))$ relies upon a plug-in estimator for $\overline{\tau}_{\emptyset}$ where the true distributions F^0 and F^1 are replaced by their

empirical counterparts $\hat{F}_N^0(Z)$ and $\hat{F}_N^1(Z)$, respectively.

To estimate the distribution of $T_\emptyset(\mathbf{y}(\mathbf{Z}))$ we follow the classical two-sample bootstrap approach: take $\{(y_i^*(0), x_i^*)\}_{i=1}^{n_0} \overset{iid}{\sim} \hat{F}_N^0$ and independently take $\{(y_i^*(1), x_i^*)\}_{i=1}^{n_1} \overset{iid}{\sim} \hat{F}_N^1$. Define the bootstrapped Horvitz-Thompson measure as

$$\left(\hat{F}_N^{0,*}(z), \hat{F}_N^{1,*}(z)\right) = \left(\frac{1}{n_0} \sum_{i\,:\,z_i=0} \delta_{\left(y_i^*(z_i), x_i^*\right)},\ \frac{1}{n_1} \sum_{i\,:\,z_i=1} \delta_{\left(y_i^*(z_i), x_i^*\right)}\right). \tag{3}$$

The so-called bootstrap test statistic $T^*(y^*(Z))$ is defined as

$$T^*(y^*(Z)) = \sqrt{N}\Bigg(\underbrace{\left(\mathbb{E}_{\hat{F}_N^{1,*}(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^{0,*}(Z)}\left[\Pi_y(y,x)\right]\right)}_{\textit{Bootstrap Resampling Term}} - \underbrace{\left(\mathbb{E}_{\hat{F}_N^{1}(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^{0}(Z)}\left[\Pi_y(y,x)\right]\right)}_{\textit{Centering by Empirical Means}}\Bigg). \tag{4}$$

The conditional distribution of $T^*(y^*(Z))$ given the original data is

$$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right);$$

this conditional distribution is taken as an empirical proxy for $\mathcal{L}\left(T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))\right)$, the true distribution of $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$. The hope of bootstrap inference is that
$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ provides a sufficiently close approximation to $\mathcal{L}\left(T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))\right)$ so that inferences for $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ can be based upon the computable quantity
$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ instead of the experimentally unknowable quantity
$\mathcal{L}\left(T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))\right)$. Indeed this is classically observed to be the case in the superpopulation two-sample model (LR05, Section 15.4.2).

Through an application of the continuous mapping theorem the result for the difference in means $\sqrt{N}\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right)$ can be enlarged the far wider class of test statistics of the form (1)

where the parameter $\hat{\xi}$ obeys the following condition:

Condition 4. *There exists* $\xi \in \Xi$ *such that, with* $Z \sim$ *Unif*(Ω),

1. $\hat{\xi}(\mathbf{y}(\mathbf{Z})) \xrightarrow{p} \xi$
2. $\hat{\xi}(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z \xrightarrow{p} \xi$ *almost surely with respect to* $\mathbf{y}(\mathbf{Z})$ *and* Z.

For the remainder of our results we assume the test statistic $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ to be of the form (1) and that Conditions 3 and 4 hold. For a test statistic of the form (1) the *i.i.d.* bootstrapped test statistic is

$$T^*(y^*(Z)) = f_{\hat{\xi}^*}\Bigg(\sqrt{N}\Bigg(\Big(\mathbb{E}_{\hat{F}_N^{1,*}(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^{0,*}(Z)}\left[\Pi_y(y,x)\right]\Big) - \Big(\mathbb{E}_{\hat{F}_N^{1}(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^{0}(Z)}\left[\Pi_y(y,x)\right]\Big)\Bigg)\Bigg), \quad (5)$$

where $\hat{\xi}^*$ is the value of $\hat{\xi}$ evaluated on the data $\{(y_i^*(0), x_i^*)\}_{i=1}^{n_0}$ and $\{(y_i^*(1), x_i^*)\}_{i=1}^{n_1}$.

The *i.i.d.* bootstrap procedure of (3) and (5) constructs the bootstrap statistic $T^*\left(y^*(Z)\right)$ as a functional of the bootstrapped empirical processes $\hat{F}_N^{0,*}(Z)$ and $\hat{F}_N^{1,*}(Z)$. This resampling procedure is classically popular, but it contravenes the typical advice that ones' resampling ought to mirror the data-generating process: notably, the *i.i.d.* bootstrap procedure of (3) and (5) captures the model-based *i.i.d.* structure of the superpopulation data generating process for $\{(y_i(0), y_i(1), x_i)\}$ but ignores the design-based aspect that the observed data $\{(y_i(Z_i), x_i)\}_{i=1}^{N}$ arose from a completely randomized experiment. In fact, the *i.i.d.* bootstrap procedure of (3) and (5) can be viewed in the light of a design-based resampling procedure which captures both the *i.i.d.* superpopulation model and the completely randomized design; Algorithm 1 constructs this resampling procedure.

Algorithm 1 exactly enumerates the sampling distribution of the random variable constructed by sampling N draws uniformly with replacement from the observed treated and

Algorithm 1: A "pairs bootstrap"-based distributional estimator.

Input: An observed treatment allocation $Z \in \Omega$.

Result: The bootstrap distributional estimator $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$.

Define

$$C = \{(y_i(Z_i), x_i) \ : \ Z_i = 0\},$$
$$T = \{(y_i(Z_i), x_i) \ : \ Z_i = 1\}.$$

for $(D_0, D_1) \in \mathcal{P}(N, C) \times \mathcal{P}(N, T)$ **do**

Say that

$$D_0 = \{(y_i^*(0), x_{i0}^*)\}_{i=1}^N,$$
$$D_1 = \{(y_i^*(1), x_{i1}^*)\}_{i=1}^N.$$

for $B \in \Omega$ **do**

Generate the "bootstrap experimental observations"

$$\{(y_i^*(0), x_{i0}^*) \ : \ B_i = 0\} \cup \{(y_i^*(1), x_{i1}^*) \ : \ B_i = 1\}.$$

Compute $T_{\mathscr{S}}(\cdot)$ using the bootstrap experimental observations with centering by $\mathbb{E}_{\hat{F}_N^1}[(y, x)] - \mathbb{E}_{\hat{F}_N^0}[(y, x)]$, denote this $T^*_{D_0, D_1, B}$.

end

end

return

$$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right) = \frac{\sum_{D_0, D_1, B} \delta_{T^*_{D_0, D_1, B}}}{\left|\mathcal{P}(N, C) \times \mathcal{P}(N, T) \times \Omega\right|}.$$

control populations, then running a completely randomized experiment on this imputed population by selecting a new independent treatment allocation $B \sim \mathit{Unif}(\Omega)$, and finally computing $T_{\mathscr{S}}(\cdot)$ on the "experimental data" observed under the treatment allocation B. By exploiting the *i.i.d.* nature of the superpopulation model and the independence of the treatment allocation process it follows that Algorithm 1 indeed generates the conditional distribution of (5). However, Algorithm 1 highlights how (5) can be viewed in the light of creating a full imputed population of N individuals and then allocating a new treatment

vector to this population. This insight will play crucially into our subsequent resampling schemes.

Theorem 3. *In the superpopulation model, subject to mild assumptions, the i.i.d. bootstrap of Algorithm 1 is strongly consistent in the sense that for any test statistic of the form* (1) *satisfying Conditions 3 and 4*

$$\rho_{BL}\left(\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}\left(T_{\emptyset}(\mathbf{y}(\mathbf{Z}))\right)\right) \xrightarrow{a.s.} 0.$$

Theorem 3 prompts the natural question: How does the *i.i.d.* bootstrap of (3) and (5) behave in the fixed covariate model and the finite population model? In the superpopulation model, the *i.i.d.* data generating process of the $(y_i(0), y_i(1), x_i)$ is cleanly replicated by the *i.i.d.* resampling used in the formation of (3). However, the fixed covariate model retains independence between the $(y_i(0), y_i(1), \mathbf{x}_i)$ but these are no longer necessarily identically distributed; consequently, *i.i.d.* resampling from the Horvitz-Thompson empirical measure $\hat{F}_N(Z)$ fails to emulate the data generating process. The finite population model demonstrates this divergence even more starkly; there the data generating procedure inherits randomness only though the allocation of treatment $Z \sim \mathit{Unif}(\Omega)$ and *i.i.d.* resampling from $\hat{F}_N(Z)$ resembles nothing of the original generative model. Nevertheless, in both models the *i.i.d.* bootstrap is still convergent in the sense that $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ limits in the ρ_{BL}-metric to a fixed distribution in $\mathcal{M}(\mathbb{R})$; however the limit no longer matches that of $\mathcal{L}\left(T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))\right)$ for $\mathscr{S} = \mathscr{C}$ or $\mathscr{F}$. Despite this disagreement between the limit and the truth, the *i.i.d.* bootstrap remains inferentially useful.

Theorem 4. *Under the fixed covariate and finite population models the i.i.d. bootstrap of* (3) *and* (5) *is strongly consistently conservative for any test statistic of the form* (1) *satisfying Conditions 3 and 4 .*

Remark 1. For some intuition regarding Theorem 4 in the finite population case con-

sider the analogous procedure in the context of survey sampling. Consider a survey conducted by uniformly sampling n units from N individuals without replacement. Write the empirical mean of this sample as $\hat{\bar{X}}$. Under mild conditions, the *i.i.d.* bootstrap conditional distribution for $\sqrt{N}\hat{\bar{X}}$ converges almost surely in the ρ_{BL}-metric to $\mathcal{N}(0, \Sigma_{boot})$ while the true distribution of $\sqrt{N}\hat{\bar{X}}$ converges to $\mathcal{N}(0, \Sigma)$ where generally $\Sigma_{boot} \geq \Sigma$ (MHL16). Consequently, by Anderson's Theorem (Ton90, Theorem 4.2.5), the *i.i.d.* bootstrap conditional distribution for $f_{\xi^*}\left(\sqrt{N}\left(\hat{\bar{X}}^* - \hat{\bar{X}}\right)\right)$ stochastically dominates the true distribution of $f_{\hat{\xi}}\left(\sqrt{N}\left(\hat{\bar{X}} - \mathbb{E}\left[\hat{\bar{X}}\right]\right)\right)$. Thus, in the context of survey sampling the *i.i.d.* bootstrap is strongly consistently conservative.

Theorem 3 reflects the ability to consistently estimate the variance of $\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ under the superpopulation model; in fact, writing $\hat{V}_1$ as the variance of
$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ with $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\emptyset}\right)$ yields a consistent variance estimator for $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ in a superpopulation. Theorem 4 reflects the fact that $\hat{V}_1$ is conservative in a fixed covariate or finite population model.

Theorem 5. *The conditional bootstrap variance estimator $\hat{V}_1$ converges in probability to $\lim_{N\to\infty} \mathbb{V}\left(\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))\right)$ under $\mathcal{P}_{\emptyset}$. Under $\mathcal{P}_{\mathscr{C}}$ and $\mathcal{P}_{\mathscr{F}}$ the variance estimator $\hat{V}_1$ converges in probability to a conservative limit, in the sense of the Loewner partial order, for the variance of $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ in a fixed covariate or finite population model; formally under both $\mathcal{P}_{\mathscr{C}}$ and $\mathcal{P}_{\mathscr{F}}$*

$$plim_{N\to\infty} \hat{V}_1 = V_1 \succeq \lim_{N\to\infty} \mathbb{V}\left(\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))\right).$$

While Theorems 3 and 4 imply that the *i.i.d.* bootstrap can be used for asymptotically valid inference against $H_{N,\mathscr{S}}$ with $\mathscr{S} = \emptyset$, $\mathscr{C}$, or $\mathscr{F}$ Theorem 4 shows that there is room for improvement in the case that $\mathscr{S} = \mathscr{C}$, or $\mathscr{F}$ since it may well be that the limiting behavior of the conditional bootstrap distribution is needlessly overconservative. The same story is told by Theorem 5: $\hat{V}_1$ is consistent in a superpopulation but there may be alternative variance

estimators for $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ which outperform $\hat{V}_1$ in fixed covariate and finite population models.

4.7 A Residual Bootstrap At The Fixed-Covariate Level

In order to refine the results from Section 4.6 we begin by tailoring to the case of $\mathscr{S} = \mathscr{C}$ wherein we take the covariates to be fixed for each individual but allow for randomness in the potential outcomes. As in the superpopulation case, we begin with a resampling scheme for $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$. Motivated by Algorithm 1 we construct an imputed population; however, in this case we leverage the deterministic covariates of the population $\mathscr{C}$-conditional measure to reduce the potential conservativeness of the bootstrapping procedure. We construct a resampling method inspired by the residual bootstrap (Efr79, ST95). Relying upon only the observed data $\{(y_i(Z_i), \mathbf{x}_i)\}_{i=1}^N$ we attempt to glean information of the conditional distribution $P_{\mathbf{x}}$ in the population $\mathscr{C}$-conditional measure. To this end we use a linear approximation of $P_{\mathbf{x}}$, though we make no assumption that the linear model is well-specified; this reflects the model-agnostic framework of (BBB$^+$19, BBK$^+$19). Even under arbitrary misspecification of the linear model our results will hold, demonstrating a surprising robustness to the true latent structure of the population $\mathscr{C}$-conditional measure. In Section 4.9 we provide simulations illustrating the results below where the true feature-outcome relationship is nonlinear.

Regardless of the relationship between outcomes and covariates in the population $\mathscr{C}$-conditional measure one may seek an optimal solution to the population L_2-norm linear

approximation and its empirical approximation; for $z \in \{0, 1\}$

$$\dot{\beta}_z = \underset{\beta \in \mathbb{R}^{d \times (k+1)}}{\arg\min} \sum_{i=1}^{N} \mathbb{E}\left[\left\| y_i(z) - \beta \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \right\|_2^2 \,\middle|\, x_i \right] \tag{6}$$

$$\hat{\beta}_z = \underset{\beta \in \mathbb{R}^{d \times (k+1)}}{\arg\min} \sum_{i \,:\, Z_i = z} \left\| y_i(z) - \beta \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \right\|_2^2 . \tag{7}$$

Under mild conditions these optimization problems are strictly convex and admit closed-form unique optimal solutions. However, when the systems are underdetermined uncountably many solutions exist. In such cases our results proceed without major modification by selecting a canonical representative of this class of solutions, for instance using the Moore-Penrose pseudoinverse in place of the matrix inverses used in defining $\dot{\beta}_z$ and $\hat{\beta}_z$ suffices; for more details of this argument we direct the reader to (CF21, Appendix G).

After computing $\hat{\beta}_z$ predicted potential outcomes $(\hat{\mu}_0(\mathbf{x}_i), \hat{\mu}_1(\mathbf{x}_i))$ are available for each individual, regardless of the original treatment allocation Z, given by

$$\hat{\mu}_z(\mathbf{x}_i) = \hat{\beta}_z \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} ; \tag{8}$$

the imputed outcomes $\hat{\mu}_z(\mathbf{x}_i)$ are the L_2-optimal linear approximations of the true outcomes given the observed outcomes $\mathbf{y}(\mathbf{Z})$ and the covariates $\mathbf{x}_i$. For each observed outcome the imputed potential outcome defines a residual $\epsilon_i(Z_i) = y_i(Z_i) - \hat{\mu}_{Z_i}(\mathbf{x}_i)$; since the counterfactual $y_i(1 - Z_i)$ is never observed only one residual can be computed for each individual in accordance with the assigned treatment allocation Z_i. Consequently, one observes n_z residuals $\epsilon_i(z)$; from these observed residuals form their Horvitz-Thompson empirical measures

$$\hat{F}_N(z) = \left(\hat{F}_N^0(z), \hat{F}_N^1(z) \right) = \left(\frac{1}{n_0} \sum_{i \,:\, z_i = 0} \delta_{(\epsilon_i(z_i), x_i)}, \; \frac{1}{n_1} \sum_{i \,:\, z_i = 1} \delta_{(\epsilon_i(z_i), x_i)} \right).$$

Remark 2. By the Karush-Kuhn-Tucker first order optimality conditions of ordinary least squares regression, the sample mean of the residuals is zero in each coordinate; formally $n_z^{-1} \sum_{i\,:\,Z_i=z} \epsilon_i(z) = 0$ for $z \in \{0,1\}$.

The residual bootstrap approximates the conditional distribution $P_{\mathbf{x}_i}$ by the distribution of $(\hat{\mu}_0(\mathbf{x}_i) + \epsilon_i^*(0), \hat{\mu}_1(\mathbf{x}_i) + \epsilon_i^*(1))$ where

$$\epsilon_i^*(0) = \Pi_{\epsilon(0)}(\epsilon(0), \mathbf{x}_i) \text{ with } (\epsilon(0), \mathbf{x}_i) \sim \hat{F}_N^0(Z),$$
$$\epsilon_i^*(0) = \Pi_{\epsilon(1)}(\epsilon(1), \mathbf{x}_i) \text{ with } (\epsilon(1), \mathbf{x}_i) \sim \hat{F}_N^1(Z).$$

Denote the approximated conditional distribution formed in this manner as $\hat{P}_{\mathbf{x}_i}$. Just as the *i.i.d.* bootstrapped Horvtiz-Thompson measure (3) attempted to approximate the projections of the population $\emptyset$-conditional measure one can approximate the population $\mathscr{C}$-conditional measure via

$$\frac{1}{N}\sum_{i=1}^{N}(P_{\mathbf{x}_i}, \delta_{\mathbf{x}_i}) \approx \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{\mathbf{x}_i}, \delta_{\mathbf{x}_i}).$$

In the spirit of Algorithm 1 we use this distributional approximation to form an imputed population and allocate new treatments over this imputed population; Algorithm 2 details the procedure.

For $z \in \{0,1\}$, using the unobserved quantity $\dot{\beta}_z$ define

$$\dot{\mu}_z(\mathbf{x}_i) = \dot{\beta}_z \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix},$$
$$\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(\mathbf{x}_i).$$

In the terminology of (DFM19) the systematic treatment effect variation is $\dot{y}_i(1) - \dot{y}_i(0)$ while the idiosyncratic treatment effect variation is $\dot{\epsilon}_i(1) - \dot{\epsilon}_i(0)$. Algorithm 2 refines the conservativeness of the *i.i.d.* resampling bootstrap of Algorithm 1 by leveraging the fixed covariates

Algorithm 2: A residual-based distributional estimator.

Input: An observed treatment allocation $Z \in \Omega$.

Result: The bootstrap distributional estimator $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$.

Compute the imputed values $(\hat{\mu}_0(\mathbf{x}_i), \hat{\mu}_1(\mathbf{x}_i))$ according to (8) and define

$$C = \{(\epsilon_i(Z_i)) \ : \ Z_i = 0\},$$
$$T = \{(\epsilon_i(Z_i)) \ : \ Z_i = 1\}.$$

for $(D_0, D_1) \in \mathcal{P}(N, C) \times \mathcal{P}(N, T)$ **do**

Say that

$$D_0 = \{(\epsilon_i^*(0))\}_{i=1}^N,$$
$$D_1 = \{(\epsilon_i^*(1))\}_{i=1}^N.$$

for $B \in \Omega$ **do**

Generate the "bootstrap experimental observations"

$$\{(\hat{\mu}_0(\mathbf{x}_i) + \epsilon_i^*(0), \mathbf{x}_i) \ : \ B_i = 0\} \cup \{(\hat{\mu}_1(\mathbf{x}_i) + \epsilon_i^*(1), \mathbf{x}_i) \ : \ B_i = 1\}.$$

Compute $T_{\mathscr{S}}(\cdot)$ using the bootstrap experimental observations with centering by $\frac{1}{N}\sum_{i=1}^N (\hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i))$, denote this $T^*_{D_0,D_1,B}$.

end

end

return

$$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right) = \frac{\sum_{D_0,D_1,B} \delta_{T^*_{D_0,D_1,B}}}{\left|\mathcal{P}(N, C) \times \mathcal{P}(N, T) \times \Omega\right|}.$$

to predict systematic treatment effect variation, $\hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i)$, while resampling residuals, $\epsilon_i(Z_i)$, to account for variability in treatment effects $y_i(1) - y_i(0)$ due to idiosyncratic variation.

Theorem 6. *In the fixed covariate model, subject to mild assumptions:*

1. *The residual bootstrap of Algorithm 2 is strongly consistently conservative for any test statistic of the form* (1) *satisfying Conditions 3 and 4 regardless of the truth of the linear model.*

2. *If the linear model is asymptotically well-specified in the sense that* $\dot{\epsilon}_i(1)$ *are indeed i.i.d. – and likewise* $\dot{\epsilon}_i(0)$ *are i.i.d. – with* $cov\left(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1)\right) = 0_{d\times d}$*, then the residual bootstrap is also strongly consistent in the sense that*

$$\rho_{BL}\left(\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}\left(T_{\mathscr{C}}(\mathbf{y}(\mathbf{Z}))\right)\right) \xrightarrow{a.s.} 0.$$

The generic conservativeness of the residual bootstrap in the fixed covariate model is not an artefact of Algorithm 2, it arises from the inherent unidenifiability of the idiosyncratic treatment effect variation under the population $\mathscr{C}$-conditional measure. This is driven by the fact that $\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{C}}\right)$ depends upon the joint behavior of $\mathcal{L}\left((\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))\right)$ but only the marginals $\mathcal{L}\left(\dot{\epsilon}_i(0)\right)$ and $\mathcal{L}\left(\dot{\epsilon}_i(1)\right)$ are approximated by $\hat{F}_N^0(Z)$ and $\hat{F}_N^1(Z)$, respectively.

At first glance, Theorem 6 does not show that the residual bootstrap is a clear improvement the *i.i.d.* bootstrap since both are generally consistently conservative under the fixed covariate model. The residual bootstrap of Algorithm 2 is consistent – rather than consistently conservative – under correct specification of the linear model with *i.i.d.* residuals.

Definition 7. Consider two algorithms, resampling procedure A and resampling procedure B, which produce distributional estimators

$$\mathcal{L}_A\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right),$$
$$\mathcal{L}_B\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right),$$

respectively. Resampling procedure A is *(strongly) more conservative* than resampling procedure B if there exists two random variables $\mathcal{A}$ and $\mathcal{B}$ for which

$$\rho_{BL}\left(\mathcal{L}_A\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}\left(\mathcal{A}\right)\right) \xrightarrow{a.s.} 0$$
$$\rho_{BL}\left(\mathcal{L}_B\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}\left(\mathcal{B}\right)\right) \xrightarrow{a.s.} 0,$$

and $\mathcal{A}$ stochastically dominates $\mathcal{B}$.

Theorem 7. *Under the fixed covariate model the i.i.d. bootstrap is strongly more conservative than the residual bootstrap for any test statistic of the form* (1) *satisfying Conditions 3 and 4 regardless of the truth of the linear model.*

Theorem 7 demonstrates that, even though both the *i.i.d.* bootstrap and the residual bootstrap are consistently conservative under the fixed covariate model, there remains a preference for the residual bootstrap since it results in inferences which can be no more conservative than the those arising from the *i.i.d.* bootstrap. Under a broad array of conditions, inferences derived from the residual bootstrap will be strictly less conservative than those derived using the *i.i.d.* bootstrap.

In Section 4.6 we considered the behavior of the *i.i.d* bootstrap in the fixed covariate and finite population models (c.f. Theorem 4), likewise here we examine the behavior of the residual bootstrap in the finite population model.

Theorem 8. *Under the finite population model the residual bootstrap of Algorithm 2 is strongly consistently conservative for any test statistic of the form* (1) *satisfying Conditions 3 and 4 regardless of the truth of the linear model.*

In the same vein as Theorem 5, the variance of the conditional bootstrap distribution from Algorithm 2 for $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)$ with $\mathscr{S} = \mathscr{C}$ or $\mathscr{F}$ serves as a variance estimator for $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ in a fixed covariate model or finite population model, respectively. Denote this variance estimator $\hat{V}_2$.

Theorem 9. *Under $\mathcal{P}_{\mathscr{C}}$ and $\mathcal{P}_{\mathscr{F}}$ the variance estimator $\hat{V}_2$ converges in probability to a conservative limit, in the sense of the Loewner partial order, for the variance of $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ in a fixed covariate or finite population model; formally under both $\mathcal{P}_{\mathscr{C}}$ and $\mathcal{P}_{\mathscr{F}}$*

$$plim_{N\to\infty} \hat{V}_2 = V_2 \succeq \lim_{N\to\infty} \mathbb{V}\left(\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))\right).$$

Furthermore, in both the fixed covariate and finite population models $V_1 \succeq V_2$.

4.8 An Optimal Transport Bootstrap At The Finite Population Level

Theorem 8 suggests that the residual bootstrap of Section 4.7 may result in conservative inferences at the finite population level. Some of this conservativeness may be due to inherent unidentifiablility of idiosyncratic treatment effect variation under the population $\mathscr{F}$-conditional measure. However, just as seen in Theorem 7 there may be resampling procedures which improve upon Algorithm 2 even though the resulting algorithms themselves remain consistently conservative. The main result of this section is to construct an asymptotically optimal procedure which – while still consistently conservative in general – provides an asymptotically sharp bootstrap variance estimator. For technical reasons, we take the potential outcomes to be scalar valued, $d = 1$; in Section 4.10 we discuss the rationale for this restriction.

To begin our analysis, we examine the distributional behavior of the difference in means $\sqrt{N}(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{F}})$ in a finite population. For a given collection of potential outcomes specified by the event $\mathscr{F}$ define

$$\Sigma^{(N)}_{z,\mathscr{F}} = \frac{1}{N-1}\sum_{i=1}\left(\mathbf{y}_i(z) - \frac{1}{N}\sum_{j=1}\mathbf{y}_j(z)\right)^2 \quad \text{for } z \in \{0,1\},$$
$$\Sigma^{(N)}_{01,\mathscr{F}} = \frac{1}{N-1}\sum_{i=1}\left(\mathbf{y}_i(0) - \frac{1}{N}\sum_{j=1}\mathbf{y}_j(0)\right)\left(\mathbf{y}_i(1) - \frac{1}{N}\sum_{j=1}\mathbf{y}_j(1)\right).$$

In (LD17) Li and Ding show that, under mild conditions, $\sqrt{N}(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{F}})$ converges

weakly to a centered Gaussian random variable with variance

$$\begin{aligned} V_{\mathscr{F}} &:= \frac{1-p}{p}\Sigma_{1,\mathscr{F}} + \frac{p}{1-p}\Sigma_{0,\mathscr{F}} + 2\Sigma_{01,\mathscr{F}}, \\ \Sigma_{z,\mathscr{F}} &= \lim_{N\to\infty} \Sigma_{z,\mathscr{F}}^{(N)} \quad \text{for } z \in \{0,1\}, \\ \Sigma_{01,\mathscr{F}} &= \lim_{N\to\infty} \Sigma_{01,\mathscr{F}}^{(N)}. \end{aligned}$$

Both $\Sigma_{0,\mathscr{F}}$ and $\Sigma_{1,\mathscr{F}}$ can be consistently estimated in a completely randomized experiment (Coc77, Theorem 2.4). Without additional model structure $\Sigma_{01,\mathscr{F}}$ is not consistently estimable; however, it may be bounded. A preliminary bound on $\Sigma_{01,\mathscr{F}}$ can be achieved via the Cauchy-Schwarz inequality (Ney90); however, (AGL14) shows that the Cauchy-Schwarz bound is not optimal and constructs a sharp bound on $\Sigma_{01,\mathscr{F}}$ via the celebrated Frechét-Hoeffding copula upper bound; we denote their upper bound as $\hat{V}_{AGL}$. The proof that the $\hat{V}_{AGL}$ provides an asymptotically sharp upper bound on $\Sigma_{01,\mathscr{F}}$ rests upon the work of (Tch80). We provide an alternative construction of $\hat{V}_{AGL}$ via the solution of a Wasserstein metric optimization problem; this rederivation provides key insight into the construction of a general resampling algorithm for inferences under the finite population model.

4.8.1 Interplay Between Sharp Variance Estimation and the Wasserstein Metric

Given two probability measures μ and ν over $\mathbb{R}$, the p-Wasserstein distance between μ and ν for $p \geq 1$ is defined as

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in C(\mu,\nu)} \mathbb{E}_\gamma \left[|\mathcal{X} - \mathcal{Y}|^p \right] \right)^{1/p},$$

where $C(\mu,\nu)$ denotes the collection of all probability measures on $\mathbb{R}^2$ which marginalize to μ and ν under the coordinate projections and $(\mathcal{X},\mathcal{Y}) \sim \gamma$. The set $C(\mu,\nu)$ is known as the

set of couplings of μ and ν.

In the particular case of $p = 2$, W_p^p reduces to

$$\begin{aligned}\inf_{\gamma \in C(\mu,\nu)} \mathbb{E}_\gamma \left[|\mathcal{X} - \mathcal{Y}|^2 \right] &= \inf_{\gamma \in C(\mu,\nu)} \mathbb{E}_\gamma \left[\mathcal{X}^2 + \mathcal{Y}^2 - 2\mathcal{X}\mathcal{Y} \right] \\ &= \inf_{\gamma \in C(\mu,\nu)} \left(\mathbb{E}_\gamma \left[\mathcal{X}^2 \right] + \mathbb{E}_\gamma \left[\mathcal{Y}^2 \right] - 2\mathbb{E}_\gamma \left[\mathcal{X}\mathcal{Y} \right] \right). \end{aligned} \tag{9}$$

Lemma 1. *The optimization problem of* (9) *admits a unique optimal solution in the sense that there exists a single* $\gamma^{opt} \in C(\mu, \nu)$ *for which*

$$\mathbb{E}_{\gamma^{opt}} \left[|\mathcal{X} - \mathcal{Y}|^2 \right] = \inf_{\gamma \in C(\mu,\nu)} \mathbb{E}_\gamma \left[|\mathcal{X} - \mathcal{Y}|^2 \right].$$

The result of Lemma 1 is well-known and can be proven via the calculus of variations; we point the interested reader to (San15, Theorem 2.9) for a specific reference which constructs γ^{opt} directly, see also (BF81, Lemma 8.1). In (9) the term $\mathbb{E}_\gamma [\mathcal{X}^2]$ is determined only by the marginal distribution of $\mathcal{X}$, which is μ under any choice of $\gamma \in C(\mu, \nu)$, so it is a constant with respect to γ; the same reasoning applies to $\mathbb{E}_\gamma [\mathcal{Y}^2]$. Consequently, γ^{opt} which solves (9) also solves

$$\inf_{\gamma \in C(\mu,\nu)} (-2)\mathbb{E}_\gamma \left[\mathcal{X}\mathcal{Y} \right] = \sup_{\gamma \in C(\mu,\nu)} \mathbb{E}_\gamma \left[\mathcal{X}\mathcal{Y} \right]$$

Likewise, because $\mathbb{E}_\gamma [\mathcal{X}]$ and $\mathbb{E}_\gamma [\mathcal{Y}]$ are constant with respect to γ, the same γ^{opt} solves

$$\sup_{\gamma \in C(\mu,\nu)} \mathbb{E}_\gamma \left[\mathcal{X}\mathcal{Y} \right] + \mathbb{E}_\gamma \left[\mathcal{X} \right] \mathbb{E}_\gamma \left[\mathcal{Y} \right] = \sup_{\gamma \in C(\mu,\nu)} \operatorname{cov}_\gamma \left(\mathcal{X}, \mathcal{Y} \right). \tag{10}$$

In total, we are left with the following result.

Lemma 2. *The optimal* 2*-Wasserstein coupling maximizes the covariance of* $(\mathcal{X}, \mathcal{Y})$ *for* $\mathcal{X} \sim \mu$ *and* $\mathcal{Y} \sim \nu$.

Lemma 2 is classical; see, for instance, (San15, Section 1.7) or (PZ20, Section 1). Im-

portantly Lemma 1 proves the existence and uniqueness of a coupling between μ and ν but does not necessarily provide a transport map $\varphi : \mathbb{R} \to \mathbb{R}$ for which $\varphi(\mathcal{X}) \overset{\mathrm{d}}{=} \mathcal{Y}$. Only under stronger continuity assumptions on the support of μ and ν is such a transport map guaranteed to exist (San15, Theorem 2.5). An immediate consequence of Lemma 2 is that one may construct a sharp upper bound on $\Sigma^{(N)}_{01,\mathscr{F}}$.

Lemma 3. *Let $(y(0), y(1), x)$ be distributed according to the population $\mathscr{F}$-conditional measure. Let μ be the marginal distribution of $y(0)$ and ν be the distribution of $y(1)$; then*

$$\Sigma^{(N)}_{01,\mathscr{F}} \leq \frac{N}{N-1} \sup_{\gamma \in C(\mu,\nu)} cov_\gamma \left(\mathcal{X}, \mathcal{Y}\right)$$

and furthermore, γ^{opt} provides a probability measure on $\mathbb{R}^2$ which marginalizes to μ and ν achieving this upper bound. Under mild conditions laid out in (AGL14, Proposition 1):

1. $\lim_{N\to\infty} \sup_{\gamma \in C(\mu,\nu)} cov_\gamma \left(\mathcal{X}, \mathcal{Y}\right)$ *exists and is non-negative,*

2. $\Sigma_{01,\mathscr{F}} \leq \lim_{N\to\infty} \sup_{\gamma \in C(\mu,\nu)} cov_\gamma \left(\mathcal{X}, \mathcal{Y}\right)$,

3. *The sequence of measures* $\arg\max_{\gamma \in C(\mu,\nu)} cov_\gamma \left(\mathcal{X}, \mathcal{Y}\right)$ *converges weakly to a fixed measure $\gamma^{(\infty)}_{opt} \in \mathcal{M}(\mathbb{R}^2)$ as $N \to \infty$.*

Lemma 3 suggests a natural construction of a variance bound estimator in the spirit of $\hat{V}_{AGL}$; namely, let μ be the empirically observed distribution of control outcomes and ν be the empirically observed distribution of control outcomes and compute the maximal $\mathrm{cov}_\gamma \left(\mathcal{X}, \mathcal{Y}\right)$ over all valid couplings of μ and ν. More formally, let $\hat{\mu}$ and $\hat{\nu}$ be the projections of $\hat{F}^0_N$ and

$\hat{F}_N^1$ onto their first coordinates, respectively, and define

$$\hat{V}_{OT} = \frac{1}{N}\Bigg(\left(\frac{n_0}{n_1}\right) \frac{1}{n_1 - 1} \sum_{i\,:\,Z_i=1} \left(\mathbf{y}_i(1) - \frac{1}{n_1} \sum_{j\,:\,Z_j=1} \mathbf{y}_j(1) \right)^2 + \left(\frac{n_1}{n_0}\right) \frac{1}{n_0 - 1} \sum_{i\,:\,Z_i=0} \left(\mathbf{y}_i(0) - \frac{1}{n_0} \sum_{j\,:\,Z_j=0} \mathbf{y}_j(0) \right)^2 + 2 \sup_{\gamma \in C(\hat{\mu},\hat{\nu})} \mathrm{cov}_\gamma\left(\mathcal{X}, \mathcal{Y}\right) \Bigg).$$

Theorem 10. *The variance estimator* $\hat{V}_{OT} = \frac{N-1}{N}\hat{V}_{AGL}$.

Theorem 10 provides more than just a rederivation of $\hat{V}_{AGL}$; it constructs a probability measure on $\mathbb{R}^2$ which achieves the variance estimator in the sense that samples drawn from γ^{opt} have covariance exactly given by $\hat{V}_{OT}$. In Algorithm 3, we construct a resampling algorithm based upon γ^{opt}.

Algorithm 3: Sampling from the Optimal Transport Solution

Input: An observed treatment allocation Z, with observed responses $\{\mathbf{y}_i(Z_i)\}_{i=1}^N$

Result: A sampled population $\{\mathbf{y}_i^*(0), \mathbf{y}_i^*(1), \mathbf{x}_i\}_{i=1}^N$

Step 1: Compute the optimal coupling $\inf_{\gamma \in C(\hat{\mu},\hat{\nu})} \mathbb{E}_\gamma \left[|\mathcal{X} - \mathcal{Y}|^2\right]$**.**

Write,

$$\gamma = \sum_{\substack{i\,:\,Z_i=0 \\ j\,:\,Z_j=1}} p_{ij}\delta_{(\mathbf{y}_i(0),\mathbf{y}_j(1))}. \tag{11}$$

Step 2: Pair to Observed Treated Units

for *j such that* $Z_j = 1$ **do**

Independently sample $\mathbf{y}_i^{'}(0)$ from $\{\mathbf{y}_i(0) \ : \ Z_i = 0\}$ according to the renormalized j^{th} row probabilities

$$\frac{p_{i_1 j}}{\sum_{i\,:\,Z_i=0} p_{ij}}, \ldots, \frac{p_{i_{n_0} j}}{\sum_{i\,:\,Z_i=0} p_{ij}}.$$

end

Step 3: Pair to Observed Control Units

for *i such that* $Z_i = 0$ **do**

Independently sample $\mathbf{y}_j^{'}(1)$ from $\{\mathbf{y}_j(1) \ : \ Z_j = 1\}$ according to the renormalized i^{th} column probabilities

$$\frac{p_{ij_1}}{\sum_{j\,:\,Z_j=1} p_{ij}}, \ldots, \frac{p_{ij_{n_1}}}{\sum_{j\,:\,Z_j=1} p_{ij}}$$

end

return *The sample population* $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1), \mathbf{x}_i)\}_{i=1}^N$ *defined by*

$$\mathbf{y}_i^*(0) = \begin{cases} \mathbf{y}_i(0) & \text{if } Z_i = 0 \\ \mathbf{y}_i^{'}(0) & \text{if } Z_i = 1 \end{cases}, \quad \text{and} \quad \mathbf{y}_i^*(1) = \begin{cases} \mathbf{y}_i(1) & \text{if } Z_i = 1 \\ \mathbf{y}_i^{'}(1) & \text{if } Z_i = 0 \end{cases}.$$

Remark 3. Computing the optimal coupling γ^{opt} is particularly simple in the context of univariate marginals (PC19, San15). Furthermore, highly efficient open-source software exists to provide easy implementation for computing γ^{opt}, e.g., (FCG+21). As a result, Algorithm 3 is computationally efficient; for ease of use we provide an implementation in `python` in our supplementary materials.

Lemma 4. *The sample population* $\{\mathbf{y}_i^*(0), \mathbf{y}_i^*(1), \mathbf{x}_i\}_{i=1}^N$ *produced by Algorithm 3 obeys the following conditional marginal moment equations*

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\,\middle|\,Z\right] = \frac{1}{n_0}\sum_{j\,:\,Z_j=0}\mathbf{y}_j(0),$$

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\,\middle|\,Z\right] = \frac{1}{n_1}\sum_{j\,:\,Z_j=1}\mathbf{y}_j(1),$$

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)^2\,\middle|\,Z\right] = \frac{1}{n_0-1}\sum_{i\,:\,Z_i=0}\left(\mathbf{y}_i(0) - \frac{1}{n_0}\sum_{j\,:\,Z_j=0}\mathbf{y}_j(0)\right)^2,$$

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right)^2\,\middle|\,Z\right] = \frac{1}{n_1-1}\sum_{i\,:\,Z_i=1}\left(\mathbf{y}_i(1) - \frac{1}{n_1}\sum_{j\,:\,Z_j=1}\mathbf{y}_j(1)\right)^2,$$

Moreover, the sample population's treated-control covariance structure is maximal

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)\left(\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right)\,\middle|\,Z\right] = \left(\frac{N}{N-1}\right)\left(\sup_{\gamma\in C(\hat{\mu},\hat{\nu})} cov_\gamma\left(\mathcal{X},\mathcal{Y}\right)\right).$$

The expectations above are taken with respect to randomness in Algorithm 3 but are conditional upon both Z and the finite population $\mathscr{F}$.

Taking the limit as $N \to \infty$ and using a strong law of large numbers in completely randomized experiments, e.g., (WD21, Lemma A3), Lemma 4 demonstrates that the re-sampling procedure of Algorithm 3 produces a population $\{\mathbf{y}_i^*(z)\}_{i=1}^N$ which asymptotically has marginal second moments matching $\Sigma_{z,\mathscr{F}}$ for $z \in \{0,1\}$ in expectation. Moreover, the expected finite-sample covariance of the population $\{\mathbf{y}_i^*(0), \mathbf{y}_i^*(1)\}_{i=1}^N$ limits to a sharp upper bound on $\Sigma_{01,\mathscr{F}}$.

Remark 4. In (IM21) Imbens and Menzel propose a "causal bootstrap" procedure which exploits the Frechét-Hoeffding upper bound to generate a bootstrap sample of N potential outcomes. Under continuity assumptions their proposal, particularly (IM21, Formula 3.4), asymptotically agrees with Algorithm 3. However, if $\mathcal{P}_{\mathscr{F}}$ limits to a measure with atoms the two procedures diverge – even asymptotically. We include a detailed analysis of the differences between (IM21) and Algorithm 3 in the appendix.

4.8.2 Combining Regression and Optimal Transport for Finite Population Inference

In order to improve upon the *i.i.d.* residual resampling of Algorithm 2 in the finite population model we exploit the optimal transport sampling procedure of Algorithm 3 with respect to the residuals $\{\epsilon_i(Z_i)\}_{i=1}^N$. This leverages the optimality of $\hat{V}_{OT}$ from Section 4.8.1 while simultaneously exploiting the regression strategy of Section 4.7.

Theorem 11. *In the finite population model, subject to mild assumptions, for any test statistic of the form* (1) *satisfying Conditions 3 and 4:*

1. *The optimal-transport-based bootstrap of Algorithm 4 is strongly consistently conservative regardless of the truth of the linear model.*

Algorithm 4: An optimal-transport-based distributional estimator.

Input: An observed treatment allocation $Z \in \Omega$.

Result: The bootstrap distributional estimator $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$.

Compute the imputed values $(\hat{\mu}_0(\mathbf{x}_i), \hat{\mu}_1(\mathbf{x}_i))$ according to (8) and define the residuals $\epsilon_i(Z_i)$.

Define the random variable $\{\epsilon_i^*(0), \epsilon_i^*(1), \mathbf{x}_i\}_{i=1}^N$ as the output of Algorithm 3 computed on the observations $\{\epsilon_i(Z_i)\}_{i=1}^N$.

Select an independent permutation $\pi \sim \mathit{Unif}(\mathcal{S}_N)$.

For an independent draw $B \sim \mathit{Unif}(\Omega)$ generate the "bootstrap experimental observations"

$$\left\{ \left(\underbrace{\hat{\mu}_0(\mathbf{x}_i) + \epsilon^*_{\pi(i)}(0)}_{y_i^*(0)}, \mathbf{x}_i \right) \ : \ B_i = 0 \right\} \cup \left\{ \left(\underbrace{\hat{\mu}_1(\mathbf{x}_i) + \epsilon^*_{\pi(i)}(1)}_{y_i^*(1)}, \mathbf{x}_i \right) \ : \ B_i = 1 \right\}.$$

Compute $T_{\mathscr{F}}(\cdot)$ using the bootstrap experimental observations with centering by $\frac{1}{N}\sum_{i=1}^N \left(y_i^*(1) - y_i^*(0)\right)$, denote this random variable as $T^*(y^*(Z))$.

return

$$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right).$$

2. *If the linear model is asymptotically well specified such that the residuals $(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))$ after regression are indeed coupled via the comonotone coupling, then the residual bootstrap is also strongly consistent in the sense that*

$$\rho_{BL}\left(\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right), \mathcal{L}\left(T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z}))\right)\right) \xrightarrow{a.s.} 0.$$

In line with Theorem 7 one can show that under the finite population model the residual bootstrap of Algorithm 2 is strongly more conservative than the optimal-transport-based bootstrap.

Just like in Theorems 5 and 9, the variance of the conditional bootstrap distribution of Algorithm 4 for $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{F}}\right)$ serves as a variance estimator for $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ under a finite population model; denote this variance estimator $\hat{V}_3$.

Theorem 12. *Under $\mathcal{P}_{\mathscr{F}}$ the variance estimator $\hat{V}_3$ converges in probability to a conservative limit for the variance of $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$ in a finite population model; formally under $\mathcal{P}_{\mathscr{F}}$*

$$plim_{N\to\infty} \hat{V}_3 = V_3 \geq \lim_{N\to\infty} \mathbb{V}\left(\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))\right).$$

Furthermore, in the finite population models $V_2 \geq V_3$.

In our supplementary material we demonstrate that the limit V_3 matches the sharp variance upper bound of (DFM19, Section 4.1). This can be viewed as an optimality result for Algorithm 4 as it indeed recovers the sharpest variance bound for the difference in means under the finite population model after accounting for linear regression adjustment.

4.9 Algorithmic Implementation And Simulations

The theoretical results of the preceding sections are center around exact computations of the bootstrap distribution $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$; this aligns with classical literature on resampling methods. However, in all but the smallest cases, exact computation of $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ using Algorithms 1, 2, or 4 is infeasible due to the enormous cardinality of its support. Instead – as is typical – stochastic approximation of $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ via Monte Carlo sampling is performed in place of computing the exact distribution. Standard Glivenko-Cantelli-style arguments show that stochastic approximation in such a manner allows for arbitrarily precise approximation of $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ in the supremum norm (vdV98, Theorem 19.1). Below we highlight several key features of Algorithms 1, 2, and 4 across superpopulation, fixed covariate, and finite population models.

Simulations are based upon a noisy nonlinear model. At the superpopulation level, we take the covariates $x_i = (x_{i0}, x_{i1})^T$ to be distributed as a bivariate centered Gaussian random

variable with zero correlation. The potential outcomes are given by

$$y_i(0) = x_{i0} + \sin(x_{i0}) + a_i,$$
$$y_i(1) = x_{i1} + \cos(x_{i1}) + b_i,$$

where the noise terms a_i and b_i are independent of the covariates and are marginally given by

$$a_i \sim \mathcal{N}\left(0, \frac{9}{16}\right),$$
$$b_i \sim \begin{cases} -(e_i + 3) & ; \text{ with probability} \frac{1}{2}, \\ e_i + 3 & ; \text{ with probability} \frac{1}{2}, \end{cases}$$
$$e_i \sim Exp(1).$$

The joint distribution of (a_i, b_i) is taken to be the product of the marginal distributions. Our simulations take Z according to a completely randomized design where $n_1 = \lfloor pN \rfloor$. For superpopulation inference we simulate by repeatedly drawing $\{(y_i(0), y_i(1), x_i)\}_{i=1}^N$ independently of one another and then drawing an independent treatment allocation vector Z. For fixed covariate inference, we draw one realization of $\{(x_i)\}_{i=1}^N$ and take these values as fixed for the remainder of the simulation while repeatedly drawing new outcomes $\{(y_i(0), y_i(1))\}_{i=1}^N$ and treatment allocations Z. In our finite population simulations we draw a single realization of $\{(y_i(0), y_i(1), x_i)\}_{i=1}^N$ and leave these values fixed for the entirety of the simulation while only drawing new treatment allocations Z.

Figure 4-1 compares the bootstrap distributions formed by Algorithms 1, 2, and 4 for a single realization of the completely randomized experiment described above where the test

statistic is of the form (1) with $\eta = 1$ and f_η taken as Huber's loss function

$$f_\eta(x) = \begin{cases} \frac{x^2}{2} & ; \text{ if} |x| \leq \eta, \\ \eta\left(|x| - \frac{\eta}{2}\right) & ; \text{ if} |x| > \eta. \end{cases}$$

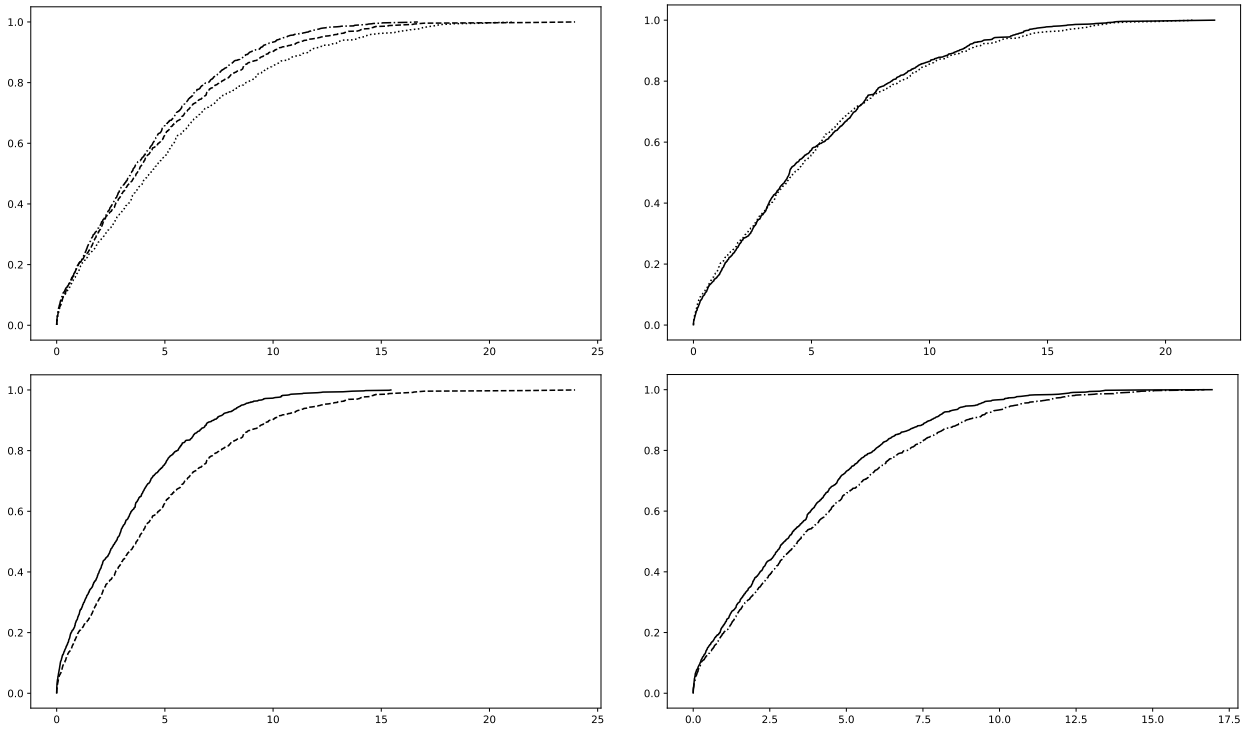

Figure 4-1: (Top-Left) Comparison of bootstrap distributions generated by Algorithm 1 (dot), Algorithm 2 (dash), and Algorithm 4 (dot-dash). (Top-Right) Comparison of the bootstrap distribution of Algorithm 1 (dot) to the superpopulation distribution after centering to enforce $H_{N,\emptyset}$ (solid). (Bottom-Left) Comparison of the bootstrap distribution of Algorithm 2 (dashed) to the fixed covariate model distribution after centering to enforce $H_{N,\mathscr{C}}$ (solid). (Bottom-Right) Comparison of the bootstrap distribution of Algorithm 4 (dot-dash) to the finite population distribution after centering to enforce $H_{N,\mathscr{F}}$ (solid). (Simulation settings: $N = 1000$, $p = .7$, bootstrap distributions formed by 1000 Monte-Carlo samples, true CDFs approximated by 1000 Monte-Carlo samples.)

The top-left panel of Figure 4-1 cleanly demonstrates the decreasing degree of conservativeness in inferences as one ranges from Algorithm 1 to Algorithm 2 and finally to Al-

gorithm 4; the conditional bootstrap cumulative distribution of Algorithm 1 lies uniformly below that of Algorithm 2 which lies uniformly below that of Algorithm 4. Alternatively phrased, the top-left panel of Figure 4-1 shows that the output of Algorithm 1 stochastically dominates that of Algorithm 2 and that the output of Algorithm 2 stochastically dominates that of Algorithm 4. The remaining three panels of Figure 4-1 compare the conditional bootstrap cumulative distribution functions to the true sampling distribution of $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$. All three panels show that the bootstrap distribution aligns with or lies uniformly below the true sampling distribution. Consequently bootstrap inferences in the superpopulation model are valid under Algorithm 1 (cf. Theorem 3); likewise for the fixed covariate model and Algorithm 2 (cf. Theorem 6) and for the finite population model and Algorithm 4 (cf. Theorem 11).

In the top-right panel of Figure 4-1 the close agreement between the superpopulation null distribution of $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ and the output of Algorithm 1 demonstrates the limiting model-free consistency of Algorithm 1 at the superpopulation level (cf. Theorem 3). In the bottom two panels of Figure 4-1 the gaps between the bootstrap distributions of Algorithms 2 and 4 and the fixed covariate and finite population model null distributions, respectively, are due to the fundamental conservativeness inherent in model-agnostic inference for the fixed covariate and finite population models (cf. Theorems 6, 8, and 11).

Figure 4-2 shows the bootstrap distributions formed by Algorithms 1, 2, and 4 for a single realization of the completely randomized experiment described above where the test statistic is simply the difference in means $\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)$. This statistic is of the form (1) with $f_\eta(x) = x$; although the identity function certainly fails Condition 3 this analysis elucidates the root cause of the behavior observed in Figure 4-1.

The top-left panel of Figure 4-2 demonstrates that the bootstrap distribution generated by Algorithm 4 is more concentrated around the origin than that of Algorithm 2 and likewise that the bootstrap distribution generated by Algorithm 2 is more concentrated around the origin than that of Algorithm 1. This peakedness ordering (Ton90, Definition 7.5.1) of

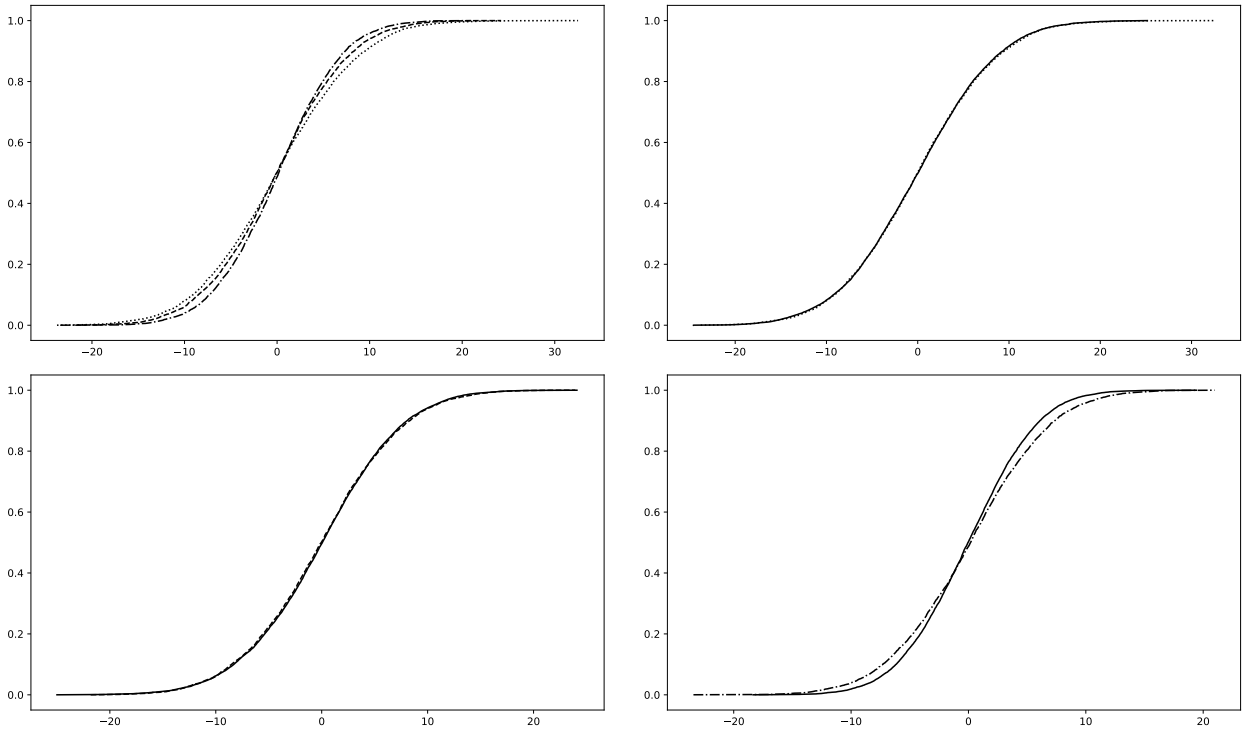

Figure 4-2: (Top-Left) Comparison of bootstrap distributions generated by Algorithm 1 (dot), Algorithm 2 (dash), and Algorithm 4 (dot-dash). (Top-Right) Comparison of the bootstrap distribution of Algorithm 1 (dot) to the superpopulation distribution after centering to enforce $H_{N,\emptyset}$ (solid). (Bottom-Left) Comparison of the bootstrap distribution of Algorithm 2 (dashed) to the fixed covariate model distribution after centering to enforce $H_{N,\mathscr{C}}$ (solid). (Bottom-Right) Comparison of the bootstrap distribution of Algorithm 4 (dot-dash) to the finite population distribution after centering to enforce $H_{N,\mathscr{F}}$ (solid). (Simulation settings: $N = 1000$, $p = .7$, bootstrap distributions formed by 5000 Monte-Carlo samples, true CDFs approximated by 10000 Monte-Carlo samples.).

the output of the three algorithms combined with Anderson's theorem (And55), (Ton90, Theorem 4.2.5) explains the stochastic dominance ordering observed in top-left panel of Figure 4-1 and reflects the discussion in Section 4.5. Similar reasoning applied to the three remaining panels of Figure 4-2 translates back to the corresponding panels of Figure 4-1 as well. The cumulative distribution functions of Figure 4-2 are all approximately Gaussian, and are guaranteed to limit to Gaussian CDFs by appropriate central limit theorems, so their

Table 4.1: Comparison of conditional bootstrap variances to the true sampling variance of $\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)$ in the simulations of Figure 4-2.

	True Variance	Bootstrap Conditional Variance
Superpopulation	51.88	52.54
Fixed Covariate	41.00	42.01
Finite Population	23.08	32.92

shapes are well described by their first two moments. By the centering in Algorithms 1, 2, and 4 all of the bootstrap distributions are guaranteed to have mean zero; this is observed to be true – up to Monte-Carlo noise – in the simulations undergirding Figure 4-2. Table 4.1 compares the bootstrap variance estimators of Theorems 5, 9, and 12 to the true variance of $\sqrt{N}(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}})$ in the same simulation as Figures 4-1 and 4-2. Importantly, the variance estimators are all no less than the true sampling variances, and the variances estimators reflect the trend that $V_1 \geq V_2 \geq V_3$.

4.10 Discussion

We have presented a nested hierarchy of resampling procedures for causal inference which asymptotically control the false rejection probability of Neyman's weak null of no average treatment effect. The resampling procedures are applicable to a wide array of test statistics, are simple to implement, and are computationally efficient. Moreover, the algorithms are model-free in the sense that our results hold without requiring practitioner knowledge of the relationship between features and potential outcomes. Each of our resampling procedures naturally suggests a variance estimator for $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$; these three variance estimators $\hat{V}_1$, $\hat{V}_2$, and $\hat{V}_3$ are guaranteed to be asymptotically conservative regardless of model specification but may be consistent under proper model specification.[2] Furthermore, previous variance estimators proposed in the literature for the difference in means under certain experimen-

[2]The variance estimator $\hat{V}_1$ is consistent at the superpopulation level without any model specification assumptions.

tal designs – including completely randomized designs – may be negative in finite samples (LDR18, Page 9160); such a property presents obvious concern for practical use. Since our estimators $\hat{V}_1$, $\hat{V}_2$, and $\hat{V}_3$ are simply the conditional variance of the bootstrapped $\sqrt{N}$-scaled difference in means they are guaranteed to be non-negative and will be positive except in the most pathological situations.

The algorithms presented above provide substantial capabilities for inference using a wide array of test statistics in completely randomized experiments. An interesting direction of further research would be to apply this framework to other experimental designs. Extensions to Bernoulli designs are likely imminently feasible; however, a more interesting program of research is to explore the relationship between resampling and covariate adaptive designs such as rerandomized designs. In a rerandomized design, the set of allowable treatment allocations $\Omega \subseteq \{0,1\}^N$ is restricted to binary vectors for which the treated individuals and control individuals are sufficiently similar in terms of their aggregate features; inference in rerandomized designs requires accounting for conditioning upon the treatment allocation yielding sufficient covariate balance (MR12). Algorithm 1 can be modified to provide asymptotically valid inference for rerandomized designs subject to general balance criteria, under the definition of (LDR18), by rejecting bootstrap treatment allocations $B \in \{b \in \{0,1\}^N : \sum_i b_i = n_1\}$ which do not yield sufficient covariate balance between the covariates of the resampled treated and control groups. This modification provides valid inference in the superpopulation, fixed covariate, and finite population models. However, analogous modifications to Algorithms 2 and 4 must preserve the joint distributions between the treated outcomes and the covariates (resp. the control outcomes and the covariates). For instance in the formulation of Algorithm 4 the optimization problem 10 would need to include constraints on the set of valid couplings so as to enforce that the coupling maintained the empirical joint distributions between the treated outcomes and the covariates (resp. the control outcomes and the covariates)

The optimal transport bootstrap scheme of Algorithm 4 in Section 4.8 was restricted

to the case of univariate potential outcomes. This contrasts with the *i.i.d.* resampling and residual bootstrap algorithms of Sections 4.6 and 4.7 which are generally applicable for $d \geq 1$. This restriction is due to a fundamental result undergirding the construction of the sharp variance bound of (AGL14). In Section 4.8 we demonstrated that the sharpness of upper bound of (AGL14) is a direct consequence of a variational principle: *given scalar random variables $\mathcal{X}$ and $\mathcal{Y}$ the optimal 2-Wasserstein coupling maximizes the covariance between $\mathcal{X}$ and $\mathcal{Y}$*. Adapting our results of Section 4.8 – or the results of (AGL14) – to $d > 1$ would require some characterization of the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$ which finds a maximal, in the Loewner partial order, covariance matrix for random vectors $\mathcal{X}$ and $\mathcal{Y}$ taking values in $\mathbb{R}^d$. Formally stated, it amounts to finding a $2d \times 2d$ positive semidefinite matrix M so that:

1. M is the covariance matrix of $(\mathcal{X}, \mathcal{Y})$ for some coupling of $\mathcal{X}$ and $\mathcal{Y}$,

2. there exists no other coupling of $\mathcal{X}$ and $\mathcal{Y}$ for which the resulting covariance matrix, $\tilde{M}$, satisfies $\tilde{M} \succeq M$.

The optimal 2-Wasserstein coupling maximizes the trace of the cross-covariance matrix between $\mathcal{X}$ and $\mathcal{Y}$ (PZ20, Section 1), but this comes with no guarantee of the second requirement for $d > 1$. Instead, we suspect that a tractable way forward may be through generalizations of the univariate quantile function given by multivariate statistical depths (CGHH17, DS21, MS22). Such a generalization would correspond to a twofold gain: an analogue of the variance estimator of (AGL14) to multivariate potential outcomes and a version of Algorithm 4 applicable to multivariate potential outcomes.

Supplementary Material

In the following appendices, we formalize a collection of regularity conditions which are sufficient for the results in the preceding sections; these amount to – for the most part – mild moment conditions on the potential outcomes to ensure laws of large numbers and central limit theorems apply. We provide proofs of the results above. Furthermore, we include a detailed analysis of the optimal-transport bootstrapping procedure of Algorithm 4 and contrast its performance with that of (IM21) in the context of limiting marginals with atoms. Code available in `python` implements the methods of the paper and recreates the simulations of Section 4.9; the code can be found at `https://github.com/PeterLCohen`.

4.11 Additional Notation

In order to conveniently write certain positive semidefinite matrices we write $a^{\otimes 2}$ to mean aa^T. Furthermore for a positive semidefinite matrix M we write $M^{1/2}$ to be its matrix square root; i.e., the unique positive semidefinite matrix satisfying $M^{1/2}M^{1/2} = M$. The inverse of $M^{1/2}$, if it exists, is written $M^{-1/2}$. Lastly, we write $\mathcal{A} \mid \mathcal{B} \xrightarrow{d} \mathcal{C}$ to mean that the random variable $\mathcal{A}$ conditioned upon $\mathcal{B}$ converges weakly to $\mathcal{C}$. Often this is not true for all realizations of $\mathcal{B}$, but does apply except for possibly a set of realization of probability zero.

4.12 Regularity Conditions

We begin with an assumption on the design itself which ensures nondegeneracy of the sampling procedure in the limit.

Assumption A.1. *As the experiment size increases the portion of treated units stabilizes to a nondegenerate limit; formally,* $n_1/N \to p$ *for* p *a fixed constant in* $(0, 1)$.

4.12.1 Superpopulation Regularity Conditions

In our superpopulation model we impose two minor moment conditions:

Assumption A.2 (Superpopulation Means and Variances)**.** *The mean vector and covariance matrix of* $(y_i(0), y_i(1), x_i)$ *exist and have finite entries. Furthermore, the covariance matrix of* $(y_i(0), y_i(1), x_i)$ *is positive semidefinite and has non-zero diagonal entries.*

Assumption A.3 (Superpopulation Bounded Fourth Moments)**.** *There exists some* $C < \infty$ *for which* $\mathbb{E}\left[y_i(z)^4\right] < C$ *for* $z = 0, 1$ *and* $\mathbb{E}\left[x_i^4\right] < C$*.*

Assumption A.2 is of fundamental importance to define many of the quantities we examine; of course, assuming finite means and a well-defined non-degenerate covariance matrix are common and non-intrusive assumptions. Furthermore, Assumption A.2 guarantees that the Lindeberg–Lévy central limit theorem applies for empirical means under the superpopulation model (Dur10, Theorem 3.4.1). Typically we use Assumption A.3 to guarantee strong laws of large numbers for empirical means and covariance matrices under $\mathcal{P}_\emptyset$. Assumption A.3 is perhaps stronger than necessary; for instance, control of the absolute $2 + \delta$ moments for any $\delta > 0$ may suffice for many of our results.

4.12.2 Fixed Covariate Regularity Conditions

Assumption A.4 (Fixed-Covariate Limiting Means and Variances)**.** *For* $z = 0, 1$ *there exists a finite limiting value* $\bar{y}(z)_\infty$ *such that* $\lim_{N\to\infty} N^{-1} \sum_{i=1}^N \mathbb{E}\left[y_i(z) \mid \mathbf{x}_i\right] = \bar{y}(z)_\infty$*. Likewise,*

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^N \mathbb{E}\left[\left(\begin{bmatrix} y_i(0) \\ y_i(1) \\ \mathbf{x}_i \end{bmatrix} - N^{-1} \sum_{j=1}^N \mathbb{E}\left[\begin{bmatrix} y_j(0) \\ y_j(1) \\ \mathbf{x}_j \end{bmatrix} \,\middle|\, \mathbf{x}_i\right]\right)^{\otimes 2} \,\middle|\, \mathbf{x}_i\right]$$

exists, is positive semidefinite, and has non-zero diagonal entries.

For $z \in \{0,1\}$ we define

$$\Sigma^{\mathscr{C}}_{y(z)} := \lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} \mathbb{E}\left[\left(y_i(z) - N^{-1} \sum_{j=1}^{N} \mathbb{E}\left[y_j(z) \,\middle|\, \mathbf{x}_j \right] \right)^{\otimes 2} \,\middle|\, \mathbf{x}_i \right]. \tag{12}$$

Assumption A.5 (Fixed-Covariate Bounded Fourth Moments)**.** *There exists some $\varepsilon > 1$ and $C < \infty$ for which, for all $z = 0, 1$ and all N, $N^{-1}\sum_{i=1}^{N} \mathbb{E}\left[|y_i(z)|^{4+\varepsilon} \mid \mathbf{x}_i \right] < C$ and $N^{-1}\sum_{i=1}^{N} \mathbf{x}_i^4 < C$.*

Assumptions A.4 and A.5 mirror Assumptions A.2 and A.3 but account for the different distributions of $(y_i(0), y_i(1)) \sim P_{\mathbf{x}_i}$. In Assumption A.5 the ε term is only of technical importance to guarantee the use of the Marcinkiewicz–Zygmund strong law of large numbers in Lemma A.6; see (Liu88, Lemma 1) for the presentation of this strong law. It is likely not a necessary condition.

Lemma A.5 (Lindeberg Condition)**.** *Under Assumptions A.4 and A.5, the potential outcomes $(y_i(0), y_i(1))$ given $\mathbf{x}_i$ jointly satisfy the conditions of Lindeberg's central limit theorem.*

Proof. Define $s_N^2(z) = \sum_{i=1}^{N} \mathbb{V}\left(y_i(z) \mid \mathbf{x}_i \right)$ where $\mathbb{V}\left(y_i(z) \mid \mathbf{x}_i \right)$ denotes the variance of $y_i(z)$ given the covariates $\mathbf{x}_i$. We will show that Lyapounov's condition (LR05, Equation 11.12) holds at $\delta = 2$ for the potential outcomes; formally, for $z \in \{0,1\}$ and $\delta = 2$

$$\lim_{N\to\infty} \frac{1}{s_N^{2+\delta}(z)} \sum_{i=1}^{N} \mathbb{E}\left[|y_i(z)|^{2+\delta} \mid \mathbf{x}_i \right] = 0. \tag{13}$$

Rewrite (13) as

$$\lim_{N\to\infty} \underbrace{\frac{N}{s_N^{2+\delta}(z)}}_{\text{Term 1}} \cdot \underbrace{\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[|y_i(z)|^{2+\delta} \mid \mathbf{x}_i \right]}_{\text{Term 2}}.$$

Term 2 is bounded above by a finite constant for all N by Assumption A.5, so it suffices to show that Term 1 vanishes as $N \to \infty$. By Assumption A.4 $N^{-1}\sum_{i=1}^{N} \mathbb{V}\left(y_i \mid \mathbf{x}_i \right)$ limits

to a positive constant which we denote $\Sigma_{y(z)}$. Consequently, $Ns_N^{-2} \to \Sigma_{y(z)}^{-1} > 0$ and $s_N^2 = \Theta(N)$. From this, it is immediate that $Ns_N^{-(2+\delta)} \to 0$ as $N \to \infty$ for $\delta = 2$. In total, this establishes (13). Since (13) is sufficient for the Lindeberg condition (LR05, Page 427), the result follows. □

As a consequence of Lemma A.5, the Lindeberg central limit theorem (LR05, Theorem 11.2.5) applies to the empirical mean of treated (or control) units. Furthermore, Assumptions A.4 and A.5 are sufficient for the *Kolmogorov criterion* so that the strong law of large numbers applies to both empirical means of potential outcomes and the entries of the empirical covariance matrices (Fel68, Section 10.7).

4.12.3 Finite Population Regularity Conditions

Assumption A.6 (Finite Population Limiting Means and Variances)**.** *For $z = 0, 1$ there exists a limiting value $\bar{y}(z)_\infty$ such that $\lim_{N\to\infty} N^{-1} \sum_{i=1}^N \mathbf{y}_i(z) = \bar{y}(z)_\infty$. Likewise,*

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} \left(\begin{bmatrix} \mathbf{y}_i(0) \\ \mathbf{y}_i(1) \\ \mathbf{x}_i \end{bmatrix} - N^{-1} \sum_{j=1}^{N} \begin{bmatrix} \mathbf{y}_j(0) \\ \mathbf{y}_j(1) \\ \mathbf{x}_j \end{bmatrix} \right)^{\otimes 2}$$

exists, is positive semidefinite, and has non-zero diagonal entries.

For $z \in \{0, 1\}$ we define

$$\Sigma_{y(z)}^{\mathscr{F}} := \lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{y}_i(z) - N^{-1} \sum_{j=1}^{N} \mathbf{y}_j(z) \right)^{\otimes 2}. \tag{14}$$

Assumption A.7 (Finite Population Bounded Fourth Moments)**.** *There exists some $C < \infty$ for which, for all $z = 0, 1$ and all N, $N^{-1} \sum_{i=1}^N \mathbf{y}_i(z)^4 < C$ and $N^{-1} \sum_{i=1}^N \mathbf{x}_i^4 < C$.*

Assumptions A.6 and A.7 mirror Assumptions A.4 and A.5 but account for the deterministic structure of $\{(\mathbf{y}_i(0), \mathbf{y}_i(1), \mathbf{x}_i)\}_{i=1}^N$ under the finite population model. Assumptions A.6 and A.7 are sufficient for finite population central limit theorems (LD17) and finite population strong laws of large numbers for both empirical means and covariance matrices (WD21, Lemma A3).

Lemma A.6. *Consider a fixed covariate model for which Assumptions A.4 and A.5 hold. Viewing the finite population model as a conditional submodel of the fixed covariate model we have that Assumptions A.6 and A.7 hold for all finite population conditioning events $\mathscr{F}$ up to a set of measure zero under the population $\mathscr{C}$-conditional measure.*

Proof. This follows from the strong law of large numbers applied to second and fourth moments. Strictly speaking, the result for the fourth moments of the potential outcomes relies upon applying the Marcinkiewicz–Zygmund strong law of large numbers to $y_i(z) - \mathbb{E}[y_i(z) \mid \mathbf{x}_i]$; see (Liu88, Lemma 1) for the presentation of this strong law. □

4.13 Linear Regression In Model-Agnostic Contexts

The data generating processes specified under the superpopulation, fixed covariate, and finite population models assign no particular known form to the functional relationship between the potential outcomes and the covariates. Nonetheless, linear regression plays an important role in our analyses despite making no assumptions that the linear model is well-specified. Use of linear regression in such model agnostic contexts has been examined in the finite population context by Lin (Lin13), Freedman (Fre08b, Fre08a), and Ding *et al.* (DFM19). Recent work has also examined linear regression in concert with nonlinear imputation estimators (CF21). We present a self-contained review of some important results from this literature in order to standardize notation.

Define the optimal solution to the population L_2-norm linear approximation and its

empirical approximation as follows; for $z \in \{0, 1\}$

$$\dot{\beta}_z = \underset{\beta \in \mathbb{R}^{d \times (k+1)}}{\arg\min} \sum_{i=1}^{N} \mathbb{E}\left[\left\| y_i(z) - \beta \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \right\|_2^2 \,\middle|\, \mathbf{x}_i \right] \tag{15}$$

$$\hat{\beta}_z = \underset{\beta \in \mathbb{R}^{d \times (k+1)}}{\arg\min} \sum_{i \,:\, Z_i = z} \left\| y_i(z) - \beta \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \right\|_2^2. \tag{16}$$

Under mild conditions these optimization problems are strictly convex and admit closed-form optimal solutions. However, when the systems are underdetermined uncountably many solutions exist but our results proceed without major modification by selecting a canonical representative of this class of solutions, for instance using the Moore-Penrose pseudoinverse in place of the matrix inverses used in defining $\dot{\beta}_z$ and $\hat{\beta}_z$ suffices. A detailed discussion of the rank-deficient case can be found in the appendix of (CF21). In light of this, we proceed as if all of the linear regressions are not rank-deficient with the knowledge that the results hold as well when the regressions are indeed rank-deficient and that the proofs proceed by replacing matrix inverses with Moore-Penrose pseudoinverses in natural locations.

Define the *population fitted values and residuals*, respectively, to be

$$\dot{\mu}_0(\mathbf{x}_i) := \dot{\beta}_0 \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}, \quad \dot{\mu}_1(\mathbf{x}_i) := \dot{\beta}_1 \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}, \tag{17}$$

$$\dot{\epsilon}_i(0) := y_i(0) - \dot{\mu}_0(\mathbf{x}_i), \quad \dot{\epsilon}_i(1) := y_i(1) - \dot{\mu}_1(\mathbf{x}_i). \tag{18}$$

Importantly, in the fixed covariate model the population fitted values are deterministic while the population residuals retain stochasticity due to their dependence upon the potential outcomes.

Remark 5. Since the residuals $(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))$ defined by $\dot{\epsilon}_i(z) = y_i(z) - \dot{\mu}_z(\mathbf{x}_i)$ are deterministic translations of the potential outcomes $(y_i(0), y_i(1))$ in the fixed-covariate model,

Lemma A.5 immediately implies that the residuals $(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))$ jointly satisfy the conditions of Lindeberg's central limit theorem.

In the finite population model, the analogue of (15) is

$$\dot{\beta}_z = \underset{\beta \in \mathbb{R}^{d \times (k+1)}}{\arg\min} \sum_{i=1}^{N} \left|\left| \mathbf{y}_i(z) - \beta \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \right|\right|_2^2; \tag{19}$$

in the finite population model both the population fitted values and population residuals are fully deterministic.

Theorem A.13. *In the fixed covariate model subject to the regularity conditions Assumptions A.4 and A.5:*

1. *The ordinary least squares linear regression coefficient $\dot{\beta}_z$ defined in* (15) *possesses a well-defined limit.*

2. *The empirical least squares coefficients $\hat{\beta}_z$ defined in* (16) *are consistent in the sense that $||\hat{\beta}_z - \dot{\beta}_z||_2 = o_P(1)$.*

In the finite population model subject to the regularity conditions Assumptions A.6 and A.7 the same results hold and – with probability one in the context of the finite population model as a conditional submodel of the fixed covariate model – the limit of $\dot{\beta}_z$ is the same in both models.

Proof. We begin with analysis in the fixed covariate model. Let $\mathbf{x}$ be the design matrix where the i^{th} row is $\mathbf{x}_i^T$ and $y(z)$ be the matrix where the i^{th} row is $y_i(z)^T$. For $Z \in \Omega$ let the submatrix of $\mathbf{x}$ formed by only examining the rows where $Z_i = z$ be denoted by $\mathbf{x}_{[i\,:\,Z_i=z]}$; the analogous notation applies to $y(z)$ as well. Classical linear model theory (BBB$^+$19) dictates

that

$$\dot{\beta}_z = \left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbb{E}\left[y(z) \mid \mathbf{x}\right],$$
$$\hat{\beta}_z = \left(\mathbf{x}^T_{[i\,:\,Z_i=z]}\mathbf{x}_{[i\,:\,Z_i=z]}\right)^{-1}\mathbf{x}^T_{[i\,:\,Z_i=z]}y(z)_{[i\,:\,Z_i=z]}.$$

The closed-form solution for $\dot{\beta}_z$ facilitates direct computation of its limit as $N \to \infty$ which is guaranteed to exist under Assumptions A.4 and A.5. The argument in the proof of Lemma A.3 of (CF21) establishes that

$$\left|\left|\left(\frac{1}{n_z}\mathbf{x}^T_{[i\,:\,Z_i=z]}\mathbf{x}_{[i\,:\,Z_i=z]}\right)^{-1} - \left(\frac{1}{N}\mathbf{x}^T\mathbf{x}\right)^{-1}\right|\right|_2 = o_P(1)$$

and using the triangle inequality, Cauchy-Schwarz inequality, and the strong law of large numbers[3] in the same line of reasoning as Lemma A.3 of (CF21) and (GB20, Appendix pg. 49) to establish that

$$\left|\left|\frac{1}{n_1}\mathbf{x}^T_{[i\,:\,Z_i=z]}y(z)_{[i\,:\,Z_i=z]} - \frac{1}{N}\mathbf{x}^T\mathbb{E}\left[y(z) \mid \mathbf{x}\right]\right|\right|_2 = o_P(1).$$

Combining the previous equations with the closed-form solutions to $\dot{\beta}_z$ and $\hat{\beta}_z$ establishes the consistency of the ordinary least squares coefficients in the fixed covariate model.

In the finite population model, $\hat{\beta}_z$ is defined as before but now classical linear regression theory dictates that

$$\dot{\beta}_z = \left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T y(z).$$

The similar reasoning to that of above again establishes $\dot{\beta}_z$ possesses a well-defined limit under the finite population model subject to Assumptions A.6 and A.7 and the argument of Lemma A.3 of (CF21) establishes the desired consistency of $||\hat{\beta}_z - \dot{\beta}_z||_2 = o_P(1)$.

[3]More precisely, we use the finite population law of large numbers to handle terms including the $\mathbf{x}_i$ and Kolmogorov's strong law of large numbers (Fel68, Section 10.7) to handle terms of the y_i.

Finally, to address the last statement of the theorem we need to show that

$$\left|\left|\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbb{E}\left[y(z)\mid\mathbf{x}\right]-\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^Ty(z)\right|\right|_2=o_P(1).$$

By logic similar to that of above, this amounts to showing that

$$\left|\left|\frac{1}{N}\mathbf{x}^Ty(z)-\frac{1}{N}\mathbf{x}^T\mathbb{E}\left[y(z)\mid\mathbf{x}\right]\right|\right|_2=o_P(1).$$

Under the Assumptions A.4 and A.5 this holds by a coordinate-wise application of Kolmogorov's strong law of large numbers (Fel68, Section 10.7). □

Using the linear regressions (15) and (19) in the fixed covariate and finite population models, respectively, we obtain a decomposition of treatment effect variation in line with that of (DFM19). We begin with a lemma characterizing the expected sample covariance matrix under an independent – but not identically distributed – data generating procedure.

Lemma A.7. *Let $\mathcal{X}_1,\ldots,\mathcal{X}_n$ be independent – but not necessarily identically distributed – random vectors in $\mathbb{R}^d$ where $\mathcal{X}_i$ has ℓ^{th} entry $[\mathcal{X}_i]_\ell$. Let $M(\mathcal{X}_1,\ldots,\mathcal{X}_n)$ be the $d\times d$ matrix with $(\ell,k)^{th}$ entry*

$$\frac{1}{n-1}\sum_{i=1}^{n}cov\left([\mathcal{X}_i]_\ell,[\mathcal{X}_i]_k\right)-\frac{1}{n(n-1)}\sum_{i=1}^{n}cov\left([\mathcal{X}_i]_\ell,[\mathcal{X}_i]_k\right)+$$
$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbb{E}\left[[\mathcal{X}_i]_\ell\right]-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[[\mathcal{X}_j]_\ell\right]\right)\left(\mathbb{E}\left[[\mathcal{X}_i]_k\right]-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[[\mathcal{X}_j]_k\right]\right). \quad (20)$$

Then

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathcal{X}_i-\frac{1}{n}\sum_{j=1}^{n}\mathcal{X}_j\right)^{\otimes 2}\right]=M(\mathcal{X}_1,\ldots,\mathcal{X}_n).$$

Proof. We examine each coordinate separately; say we are interested in the $(\ell,k)^{\text{th}}$ coordi-

nate. Consequently, we are interested in

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left([\mathcal{X}_i]_\ell - \frac{1}{n}\sum_{j=1}^{n}[\mathcal{X}_j]_\ell\right)\left([\mathcal{X}_i]_k - \frac{1}{n}\sum_{j=1}^{n}[\mathcal{X}_j]_k\right)\right]. \tag{21}$$

Write $\mathcal{A}_\ell$ as the vector $([\mathcal{X}_1]_\ell, \ldots, [\mathcal{X}_n]_\ell)^T$ and $\mathcal{B}_k$ as the vector $([\mathcal{X}_1]_k, \ldots, [\mathcal{X}_n]_k)^T$. Simple rearrangement yields that (21) is

$$\frac{1}{n-1}\mathbb{E}\left[\mathcal{A}_\ell^T\Lambda^T\Lambda\mathcal{B}_k\right] = \frac{1}{n-1}\mathbb{E}\left[\mathcal{A}_\ell^T\Lambda\mathcal{B}_k\right] \tag{22}$$

where $\Lambda = I_{d\times d} - n^{-1}(e\otimes e)$ where e is the n-length vector of all 1s and the equality holds because Λ is symmetric and idempotent. Using the cyclic property of the trace we analyze $\mathbb{E}\left[\mathcal{A}_\ell^T\Lambda\mathcal{B}_k\right]$:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{A}_\ell^T\Lambda\mathcal{B}_k\right] &= \mathbb{E}\left[\operatorname{tr}\left(\mathcal{A}_\ell^T\Lambda\mathcal{B}_k\right)\right] \quad \text{(since } \mathcal{A}_\ell^T\Lambda\mathcal{B}_k \text{ is a scalar)}\\
&= \mathbb{E}\left[\operatorname{tr}\left(\Lambda\mathcal{B}_k\mathcal{A}_\ell^T\right)\right] \quad \text{(cyclic property of tr}(\cdot)\text{)}\\
&= \operatorname{tr}\left(\Lambda\mathbb{E}\left[\mathcal{B}_k\mathcal{A}_\ell^T\right]\right) \quad \text{(linearity)}\\
&= \operatorname{tr}\left(\Lambda\left(\operatorname{cov}\left(\mathcal{A}_\ell,\mathcal{B}_k\right) + \mathbb{E}\left[\mathcal{A}_\ell\right]\mathbb{E}\left[\mathcal{B}_k\right]^T\right)\right)\\
&= \operatorname{tr}\left(\Lambda\operatorname{cov}\left(\mathcal{A}_\ell,\mathcal{B}_k\right)\right) + \operatorname{tr}\left(\mathbb{E}\left[\mathcal{A}_\ell\right]\Lambda\mathbb{E}\left[\mathcal{B}_k\right]^T\right) \quad \text{(cyclic property of tr}(\cdot)\text{)}\\
&= \operatorname{tr}\left(\operatorname{cov}\left(\mathcal{A}_\ell,\mathcal{B}_k\right)\right) - n^{-1}e^T\operatorname{cov}\left(\mathcal{A}_\ell,\mathcal{B}_k\right)e + \mathbb{E}\left[\mathcal{A}_\ell\right]\Lambda\mathbb{E}\left[\mathcal{B}_k\right]^T.
\end{aligned}$$

The cross-covariance matrix $\operatorname{cov}\left(\mathcal{A}_\ell,\mathcal{B}_k\right)$ is diagonal since – by assumption – the random vectors $\mathcal{X}_1,\ldots,\mathcal{X}_n$ are independent and so $\operatorname{cov}\left([\mathcal{X}_i]_\ell,[\mathcal{X}_j]_k\right) = 0$ whenever $i \neq j$. Conse-

quently, $\mathrm{tr}\left(\mathrm{cov}\left(\mathcal{A}_\ell, \mathcal{B}_k\right)\right) = \sum_{i=1}^n \mathrm{cov}\left(\left[\mathcal{A}_\ell\right]_i, \left[\mathcal{B}_k\right]_i\right) = \sum_{i=1}^n \mathrm{cov}\left(\left[\mathcal{X}_i\right]_\ell, \left[\mathcal{X}_i\right]_k\right)$. Likewise,

$$n^{-1} e^T \mathrm{cov}\left(\mathcal{A}_\ell, \mathcal{B}_k\right) e = n^{-1} \sum_{i,j=1}^n \mathrm{cov}\left(\left[\mathcal{A}_\ell\right]_i, \left[\mathcal{B}_k\right]_j\right) = \\ n^{-1} \sum_{i=1}^n \mathrm{cov}\left(\left[\mathcal{A}_\ell\right]_i, \left[\mathcal{B}_k\right]_i\right) = n^{-1} \sum_{i=1}^n \mathrm{cov}\left(\left[\mathcal{X}_i\right]_\ell, \left[\mathcal{X}_i\right]_k\right).$$

In the general case we this leaves us with

$$\frac{1}{n-1}\mathbb{E}\left[\mathcal{A}_\ell^T \Lambda \mathcal{B}_k\right] = \frac{1}{n-1}\sum_{i=1}^n \mathrm{cov}\left(\left[\mathcal{X}_i\right]_\ell, \left[\mathcal{X}_i\right]_k\right) - \frac{1}{n(n-1)}\sum_{i=1}^n \mathrm{cov}\left(\left[\mathcal{X}_i\right]_\ell, \left[\mathcal{X}_i\right]_k\right) + \\ \frac{1}{n-1}\sum_{i=1}^n \left(\mathbb{E}\left[\left[\mathcal{X}_i\right]_\ell\right] - \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left[\mathcal{X}_j\right]_\ell\right]\right)\left(\mathbb{E}\left[\left[\mathcal{X}_i\right]_k\right] - \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left[\mathcal{X}_j\right]_k\right]\right). \quad (23)$$

In the special case that $\ell = k$ and we are interested in the diagonal entries of the matrix we simplify (23) to obtain

$$\frac{1}{n-1}\mathbb{E}\left[\mathcal{A}_\ell^T \Lambda \mathcal{A}_\ell\right] = \frac{1}{n-1}\sum_{i=1}^n \mathbb{V}\left(\left[\mathcal{X}_i\right]_\ell\right) - \frac{1}{n(n-1)}\sum_{i=1}^n \mathbb{V}\left(\left[\mathcal{X}_i\right]_\ell\right) + \\ \frac{1}{n-1}\sum_{i=1}^n \left(\mathbb{E}\left[\left[\mathcal{X}_i\right]_\ell\right] - \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left[\mathcal{X}_j\right]_\ell\right]\right)^2. \quad (24)$$

Importantly, (24) generalizes the expectation of the variance estimator V_n^2 presented in (LS95) to the multivariate context. □

Armed with Lemma A.7 (and specifically the notation $M(\cdots)$ of (20)) we now define

some key quantities for the fixed covariate model:

$$\Sigma_{\dot{\mu}_z}^{(N)} := \frac{1}{N-1}\sum_{i=1}^{N}\left(\dot{\mu}_z(\mathbf{x}_i) - \frac{1}{N}\sum_{j=1}^{N}\dot{\mu}_z(\mathbf{x}_j)\right)^{\otimes 2},$$

$$\Sigma_{\dot{\mu}_1-\dot{\mu}_0}^{(N)} := \frac{1}{N-1}\sum_{i=1}^{N}\left((\dot{\mu}_1(\mathbf{x}_i) - \dot{\mu}_0(\mathbf{x}_i)) - \frac{1}{N}\sum_{j=1}^{N}(\dot{\mu}_1(\mathbf{x}_j) - \dot{\mu}_0(\mathbf{x}_j))\right)^{\otimes 2},$$

$$\Sigma_{\dot{\epsilon}_z}^{(N)} := M\left(\dot{\epsilon}_1(z), \ldots, \dot{\epsilon}_N(z)\right), \tag{25}$$

$$\Sigma_{\dot{\epsilon}_1-\dot{\epsilon}_0}^{(N)} := M\left(\dot{\epsilon}_1(1) - \dot{\epsilon}_1(0), \ldots, \dot{\epsilon}_N(1) - \dot{\epsilon}_N(0)\right). \tag{26}$$

In the finite population model, the definition of $\Sigma_{\dot{\mu}_z}^{(N)}$ and $\Sigma_{\dot{\mu}_1-\dot{\mu}_0}^{(N)}$ remains the same, but the residual covariance matrices require adaptation since there is no longer stochasticity in $\dot{\epsilon}_i(z)$. In the finite population model, define the residual covariance matrices as

$$S_{\dot{\epsilon}_z}^{(N)} := \frac{1}{N-1}\sum_{i=1}^{N}\left(\dot{\epsilon}_i(z) - \frac{1}{N}\sum_{j=1}^{N}\dot{\epsilon}_j(z)\right)^{\otimes 2}, \tag{27}$$

$$S_{\dot{\epsilon}_1-\dot{\epsilon}_0}^{(N)} := \frac{1}{N-1}\sum_{i=1}^{N}\left((\dot{\epsilon}_i(1) - \dot{\epsilon}_i(0)) - \frac{1}{N}\sum_{j=1}^{N}(\dot{\epsilon}_j(1) - \dot{\epsilon}_j(0))\right)^{\otimes 2}. \tag{28}$$

The limits of these covariance matrices are guaranteed to exist under Assumptions A.4 and A.5 and Assumptions A.6 and A.7, respectively. We denote their asymptotic limits by dropping the superscripted (N); for instance, $\Sigma_{\dot{\mu}_z} := \lim_{N\to\infty}\Sigma_{\dot{\mu}_z}^{(N)}$.

Finally, define

$$\Sigma_{fitted} := \frac{\Sigma_{\dot{\mu}_1}}{p} + \frac{\Sigma_{\dot{\mu}_0}}{1-p} - \Sigma_{\dot{\mu}_1-\dot{\mu}_0}, \tag{29}$$

$$\Sigma_{resid} := \frac{\Sigma_{\dot{\epsilon}_1}}{p} + \frac{\Sigma_{\dot{\epsilon}_0}}{1-p} - \Sigma_{\dot{\epsilon}_1-\dot{\epsilon}_0}. \tag{30}$$

4.14 Useful Preliminary Results

4.14.1 Central Limit Theorems

Central limit theorems across the superpopulation, fixed covariate, and finite population models form part of the theoretical backbone of our results. Some of the results are classical, while others are novel. For the sake of completeness, we include all of the necessary results here along with references for those whose proofs are common and thus omitted.

Theorem A.14. *In a superpopulation model subject to Assumptions A.2 and A.3*

$$\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\emptyset}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\Sigma_{y(1)}}{p} + \frac{\Sigma_{y(0)}}{1-p}\right),$$

where $\Sigma_{y(z)} = \mathbb{V}\left(y_1(z)\right)$.

Proof. Decompose $\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\emptyset}\right)$ as

$$\underbrace{\sqrt{\frac{N}{n_1}}\sqrt{n_1}\left(\frac{1}{n_1}\sum_{i=1}^{N} Z_i y_i(1) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[y_i(1)\right]\right)}_{\text{Term 1}} -$$

$$\underbrace{\sqrt{\frac{N}{n_0}}\sqrt{n_0}\left(\frac{1}{n_0}\sum_{i=1}^{N}(1-Z_i) y_i(0) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[y_i(0)\right]\right)}_{\text{Term 0}}$$

We first focus on Term 1. By the independence between Z and $y(1)$ under the completely randomized design,

$$\frac{1}{n_1}\sum_{i=1}^{N} Z_i y_i(1) \overset{\mathrm{d}}{=} \frac{1}{n_1}\sum_{i=1}^{n_1} y_i'(1)$$

where $\{y_1'(1), \ldots, y_{n_1}'(1)\}$ are *i.i.d.* draws from the marginal distribution of the treated potential outcomes under the superpopulation model and $\overset{\mathrm{d}}{=}$ denotes equality in law. Consequently,

the Lindeberg–Lévy central limit theorem (Dur10, Theorem 3.4.1) implies that

$$\sqrt{n_1}\left(\frac{1}{n_1}\sum_{i=1}^{N} Z_i y_i(1) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[y_i(1)\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{y(1)}\right).$$

Further more, under Assumption A.1 $\sqrt{N/n_1} \to p^{-1/2}$ and so Term 1 converges in distribution to $\mathcal{N}\left(0, p^{-1}\Sigma_{y(1)}\right)$. Analogous reasoning yields that Term 0 converges in distribution to $\mathcal{N}\left(0, (1-p)^{-1}\Sigma_{y(0)}\right)$. By the independence between the treatment allocation mechanism and the potential outcomes under a completely randomized design and the *i.i.d.* sampling of potential outcomes in the superpopulation model, Term 0 and Term 1 are independent of one another; this independence combined with the fact that the sum of independent Gaussian random variables is itself Gaussian completes the proof. □

Theorem A.15. *In a fixed covariate model subject to Assumptions A.4 and A.5*

$$\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\mathscr{C}}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{\mathscr{C}} + \Sigma_{\tau}^{\mathscr{C}}\right)$$

where, using the $M(\cdots)$ *notation from* (20),

$$\Sigma^{\mathscr{C}} := \lim_{N\to\infty} N\left(\frac{\Sigma_{y(1)}^{(N),\mathscr{C}}}{n_1} + \frac{\Sigma_{y(0)}^{(N),\mathscr{C}}}{n_0} - \frac{\Sigma_{y(1)-y(0)}^{(N),\mathscr{C}}}{N}\right),$$

$$\Sigma_{y(z)}^{(N),\mathscr{C}} := M\left(y_1(z), \ldots, y_N(z)\right), \tag{31}$$

$$\Sigma_{y(1)-y(0)}^{(N),\mathscr{C}} := M\left(y_1(1) - y_1(0), \ldots, y_N(1) - y_N(0)\right). \tag{32}$$

and

$$\Sigma_{\tau}^{\mathscr{C}} := \lim_{N\to\infty} \frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left[\left(y_i(1) - y_i(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[y_j(1) - y_j(0) \mid \mathbf{x}_j\right]\right)^{\otimes 2} \,\middle|\, \mathbf{x}_i\right]. \tag{33}$$

Proof. The proof rests upon the two-phase framework of (RBK05). Specifically, we decom-

pose

$$\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\mathscr{C}}\right) = \underbrace{\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\mathscr{F}}\right)}_{\textit{Design-based Term}} + \underbrace{\sqrt{N}\left(\overline{\tau}_{\mathscr{F}} - \overline{\tau}_{\mathscr{C}}\right)}_{\textit{Model-based Term}}.$$

By Theorem 5.1 of (RBK05) the design-based and model-based terms are asymptotically independent. Furthermore, by the Lindeberg central limit theorem, under Assumptions A.4 and A.5 the model-based term obeys a central limit theorem

$$\sqrt{N}\left(\overline{\tau}_{\mathscr{F}} - \overline{\tau}_{\mathscr{C}}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\tau}^{\mathscr{C}}\right).$$

We examine the design-based term conditionally upon a finite population $\mathscr{F}$. The variance of the design-based term conditioned upon $\mathscr{F}$ is given by

$$\begin{aligned}
\hat{\Sigma}^{\mathscr{C}} &:= N\left(\frac{\hat{\Sigma}_{y(1)}}{n_1} + \frac{\hat{\Sigma}_{y(0)}}{n_0} - \frac{\hat{\Sigma}_{y(1)-y(0)}}{N}\right) \\
\hat{\Sigma}_{y(z)} &:= \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i(z) - \frac{1}{N}\sum_{j=1}^{N} y_j(z)\right)^{\otimes 2} \\
\hat{\Sigma}_{y(1)-y(0)} &:= \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i(1) - y_i(0) - \frac{1}{N}\sum_{j=1}^{N}\left(y_j(1) - y_j(0)\right)\right)^{\otimes 2}.
\end{aligned}$$

The quantity $\hat{\Sigma}^{\mathscr{C}}$ is a random variable under the fixed covariate model. Under Assumptions A.4 and A.5 the Marcinkiewicz–Zygmund strong law of large numbers (see (Liu88, Lemma 1)) ensures that $\hat{\Sigma}^{\mathscr{C}}$ almost surely shares the same limit as $\mathbb{E}\left[\hat{\Sigma}^{\mathscr{C}} \mid \mathbf{x}\right]$. By linearity of expectation and Lemma A.7 it follows that

$$\mathbb{E}\left[\hat{\Sigma}^{\mathscr{C}} \mid \mathbf{x}\right] = N\left(\frac{M\left(y_1(1), \ldots, y_N(1)\right)}{n_1} + \frac{M\left(y_1(0), \ldots, y_N(0)\right)}{n_0} - \frac{M\left(y_1(1) - y_1(0), \ldots, y_N(1) - y_N(0)\right)}{N}\right)$$

where the notation $M(\cdots)$ is defined in (20).

By Lemma A.6 Assumptions A.6 and A.7 hold for all conditioning events $\mathscr{F}$ except for possibly a set of measure zero under the fixed covariate model and so the finite population central limit theorem (LD17) applies to the design based term almost surely in $\mathscr{F}$

$$\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\mathscr{F}}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{\mathscr{C}}\right).$$

Finally, the two central limit theorems are combined via asymptotic independence – which is guaranteed by (RBK05, Theorem 5.1.iii) – to yield the desired result. □

Similar logic to Theorem A.15 allows for the alternative central limit theorem which rests upon the decomposition of variance of (DFM19).

Theorem A.16. *In a fixed covariate model subject to the regularity conditions Assumptions A.4 and A.5*

$$\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\mathscr{C}}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{fitted} + \Sigma^{\mathscr{C}}_{resid} + \Sigma^{\mathscr{C}}_{\tau}\right)$$

where Σ_{fitted} is defined in (29)*, and*

$$\Sigma^{\mathscr{C}}_{resid} := \lim_{N\to\infty} N\left(\frac{\Sigma^{(N)}_{\dot{\epsilon}_1}}{n_1} + \frac{\Sigma^{(N)}_{\dot{\epsilon}_0}}{n_0} - \frac{\Sigma^{(N)}_{\dot{\epsilon}_1-\dot{\epsilon}_0}}{N}\right)$$

where $\Sigma^{(N)}_{\dot{\epsilon}_z}$ and $\Sigma^{(N)}_{\dot{\epsilon}_1-\dot{\epsilon}_0}$ are defined in (25) *and* (26)*, respectively.*

Theorem A.17. *In a finite population model subject to the regularity conditions Assumptions A.6 and A.7 the quantity*

$$\Sigma_{resid,\mathscr{F}} = \lim_{N\to\infty} N\left(\frac{S^{(N)}_{\dot{\epsilon}_1}}{n_1} + \frac{S^{(N)}_{\dot{\epsilon}_0}}{n_0} - \frac{S^{(N)}_{\dot{\epsilon}_1-\dot{\epsilon}_0}}{N}\right)$$

exists and

$$\sqrt{N}\left(\hat{\tau}\left(\mathbf{y}(\mathbf{Z})\right) - \overline{\tau}_{\mathscr{F}}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{fitted} + \Sigma_{resid,\mathscr{F}}\right).$$

An equivalent presentation of the limiting variance is

$$\frac{\Sigma^{\mathscr{F}}_{y(1)}}{p} + \frac{\Sigma^{\mathscr{F}}_{y(0)}}{1-p} - \Sigma^{\mathscr{F}}_{y(1)-y(0)},$$
$$\Sigma^{\mathscr{F}}_{y(1)-y(0)} := \lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{y}_i(1) - \mathbf{y}_i(0) - N^{-1} \sum_{j=1}^{N} \left(\mathbf{y}_j(1) - \mathbf{y}_j(0)\right) \right)^{\otimes 2},$$

where $\Sigma^{\mathscr{F}}_{y(z)}$ *is defined in* (14)

Proof. This result follows from the finite central limit theorem of (LD17) and the decomposition of variance of (DFM19). □

4.14.2 Consistently Conservative Distributional Estimators and Hypothesis Tests

Consider two random variables $\mathcal{X}$ and $\mathcal{Y}$ and two stochastic processes $\{\mathcal{X}^{(N)}\}_{N\in\mathbb{N}}$ and $\{\mathcal{Y}^{(N)}\}_{N\in\mathbb{N}}$ with cumulative distribution functions $F_{\mathcal{X}}$, $F_{\mathcal{Y}}$, $F_{\mathcal{X}^{(N)}}$, and $F_{\mathcal{Y}^{(N)}}$, respectively. Write the quantile function of $\mathcal{X}$ as $F^{-1}_{\mathcal{X}}$; similarly define $F^{-1}_{\mathcal{Y}}$, $F^{-1}_{\mathcal{X}^{(N)}}$, and $F^{-1}_{\mathcal{Y}^{(N)}}$.

Theorem A.18. *Suppose that:*

1. $\mathcal{X}$ *and* $\mathcal{Y}$ *are continuously distributed and their cumulative distribution functions are strictly increasing on their support,*

2. $\mathcal{X}$ *first-order stochastically dominates* $\mathcal{Y}$*,*

3. $\rho_{BL}\left(\mathcal{X}^{(N)}, \mathcal{X}\right) \to 0$ *and* $\rho_{BL}\left(\mathcal{Y}^{(N)}, \mathcal{Y}\right) \to 0$.

Then for any $\alpha \in (0,1)$

$$\lim_{N\to\infty} \mathbb{P}\left(\mathcal{Y}^{(N)} \geq F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right) = \mathbb{P}\left(\mathcal{Y} \geq F^{-1}_{\mathcal{X}}(1-\alpha)\right) \leq \mathbb{P}\left(\mathcal{X} \geq F^{-1}_{\mathcal{X}}(1-\alpha)\right) = \alpha.$$

Proof. Recall that the bounded Lipschitz metric ρ_{BL} metrizes weak convergence (vdVW96, Theorem 1.12.4). Thus, $\mathcal{X}^{(N)} \xrightarrow{d} \mathcal{X}$. By Polya's Theorem (LR05, Theorem 11.2.9) the cumulative distribution function of $\mathcal{X}^{(N)}$ converges uniformly to the cumulative distribution function of $\mathcal{X}$; the proof of uniformity follows by the same argument as in the Glivenko-Cantelli Theorem (vdV98, Theorem 19.1). Because $\mathcal{X}$ is continuously distributed and its cumulative distribution function is strictly increasing on its support it follows that $F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha) \to F^{-1}_{\mathcal{X}}(1-\alpha)$ (LR05, Lemma 11.2.1). By definition

$$\mathbb{P}\left(\mathcal{Y}^{(N)} \geq F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right) = 1 - F_{\mathcal{Y}^{(N)}}\left(F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right) + \mathbb{P}\left(\mathcal{Y}^{(N)} = F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right).$$

Taking limits on both sides and using that $\mathcal{Y}^{(N)}$ limits weakly to a continuous random variable yields

$$\lim_{N\to\infty} \mathbb{P}\left(\mathcal{Y}^{(N)} \geq F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right) = 1 - \lim_{N\to\infty} F_{\mathcal{Y}^{(N)}}\left(F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right).$$

By the uniform convergence of $F_{\mathcal{Y}^{(N)}}$ to $F_{\mathcal{Y}}$, the continuity of $F_{\mathcal{Y}}$, and the pointwise convergence of $F^{-1}_{\mathcal{X}^{(N)}}$ to $F^{-1}_{\mathcal{X}}$

$$1 - \lim_{N\to\infty} F_{\mathcal{Y}^{(N)}}\left(F^{-1}_{\mathcal{X}^{(N)}}(1-\alpha)\right) = 1 - F_{\mathcal{Y}}\left(F^{-1}_{\mathcal{X}}(1-\alpha)\right) = \mathbb{P}\left(\mathcal{Y} \geq F^{-1}_{\mathcal{X}}(1-\alpha)\right)$$

Because $\mathcal{X}$ first-order stochastically dominates $\mathcal{Y}$, for all $\alpha \in (0,1)$

$$\mathbb{P}\left(\mathcal{Y} \geq F^{-1}_{\mathcal{X}}(1-\alpha)\right) \leq \mathbb{P}\left(\mathcal{X} \geq F^{-1}_{\mathcal{X}}(1-\alpha)\right).$$

Finally, by the continuity and strict increasingness of $F_{\mathcal{X}}$

$$\mathbb{P}\left(\mathcal{X} \geq F_{\mathcal{X}}^{-1}(1-\alpha)\right) = \alpha.$$

□

Armed with Theorem A.18 we now prove Theorem 1.

Proof of Theorem 1. Let $\mathcal{X}^{(N)} \sim \mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$ and $\mathcal{Y}^{(N)} \sim \mathcal{L}\left(T_{\mathscr{J}}(\mathbf{y}(\mathbf{Z}))\right)$, then – by Definition 6 – almost surely it holds that

$$\rho_{BL}\left(\mathcal{X}^{(N)}, \mathcal{X}\right) \to 0,$$
$$\rho_{BL}\left(\mathcal{Y}^{(N)}, \mathcal{Y}\right) \to 0,$$

and $\mathcal{X}$ first-order stochastically dominates $\mathcal{Y}$. By the assumptions of Theorem 1 $\mathcal{X}$ and $\mathcal{Y}$ are continuously distributed and their cumulative distribution functions are strictly increasing on their support. Consequently, the conditions of Theorem A.18 hold almost surely, and thus we conclude that the probability of false rejection is indeed bounded above by α. □

In the context of bootstrap hypothesis testing, Theorem A.18 implies that a strongly consistently conservative bootstrap procedure (in the sense of Definition 6) induces almost surely conservative inferences. Specifically, by taking $\mathcal{L}\left(\mathcal{X}^{(N)}\right)$ to be the conditional bootstrap law and $\mathcal{Y}^{(N)}$ to be the random variable of interest, Theorem A.18 demonstrates that the probability that the random variable $\mathcal{Y}^{(N)}$ exceeds the $(1-\alpha)^{\text{th}}$ quantile of the bootstrap distribution is asymptotically no larger than α. In other words, a strongly consistently conservative resampling procedure almost surely controls the Type I error rate at no greater than the nominal level.

Remark 6. Suppose that one considers two bootstrap procedures: resampling procedure A and resampling procedure B. By analogous reasoning to the logic of Theorem A.18, if re-

sampling procedure A is strongly more conservative than resampling procedure B (in the sense of Definition 7) then the Type I error rate of inference under resampling procedure A is almost surely asymptotically no larger than that under resampling procedure B. When resampling procedure B is already strongly consistently conservative, this means that resampling procedure A may be needlessly conservative and so a practitioner ought to prefer resampling procedure B.

4.14.3 Anderson's Theorem: Stochastic Dominance through Conservative Covariance Estimation

Consider two multivariate Gaussian random vectors $\mathcal{X}$ and $\mathcal{Y}$ distributed as

$$\mathcal{X} \sim \mathcal{N}(0, S_{\mathcal{X}}),$$
$$\mathcal{Y} \sim \mathcal{N}(0, S_{\mathcal{Y}}).$$

Tong (Ton90, Theorem 4.2.5) presents a corollary of Anderson's central result from (And55):

Theorem A.19 (Anderson's Theorem)**.** *If $S_{\mathcal{Y}} \preceq S_{\mathcal{X}}$ in the Loewner partial order then*

$$\mathbb{P}(\mathcal{Y} \in B) \geq \mathbb{P}(\mathcal{X} \in B)$$

for any convex set B which is mirror symmetric about the origin (i.e., $x \in B \iff -x \in B$).

For any test statistic $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ of the form (1) the set $B_t = \{a \in \mathbb{R}^d \ : \ f_\xi(a) \leq t\}$ is convex and mirror symmetric about the origin by Condition 3. An immediate consequence of this is the following corollary:

Corollary A.1. *If $S_{\mathcal{Y}} \preceq S_{\mathcal{X}}$ in the Loewner partial order and $f_\xi(\cdot)$ obeys Condition 3 then* $\mathbb{P}(f_\xi(\mathcal{Y}) \leq t) \geq \mathbb{P}(f_\xi(\mathcal{X}) \leq t)$.

Corollary A.1 states that if $S_{\mathcal{Y}} \preceq S_{\mathcal{X}}$ then $f_{\xi}(\mathcal{X})$ first-order stochastically dominates $f_{\xi}(\mathcal{Y})$. Corollary A.1 is a workhorse result in our subsequent analyses. By virtue of Theorems A.14, A.15, and A.17, Conditions 3 and 4, and the continuous mapping theorem the null distribution of $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z}))$ is the f_{ξ}-pushforward of the Gaussian measure with mean zero and covariance matrix $S_{\mathcal{Y}}$ where the specific form of $S_{\mathcal{Y}}$ is determined by the asymptotic variance of the difference in means in the particular model of interest.

Example 18. If $f_{\xi}(t) = ||t||_2^2$ and we take a finite population model then the null distribution of $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z}))$ is the pushforward of $\mathcal{N}(0, \Sigma_{fitted} + \Sigma_{resid,\mathscr{F}})$ under the map $t \mapsto ||t||_2^2$. In the special case that $\Sigma_{fitted} + \Sigma_{resid,\mathscr{F}}$ is the $d \times d$ identity matrix then the null distribution of $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z}))$ is the χ_d^2 distribution.

Corollary A.1 provides a simple method for constructing conservative null distributions: construct a random variable $\mathcal{X}^{(N)}$ which converges in distribution to $\mathcal{X}$ with $S_{\mathcal{Y}} \preceq S_{\mathcal{X}}$ as $N \to \infty$, then the distribution of $f_{\hat{\xi}}\left(\mathcal{X}^{(N)}\right)$ automatically stochastically dominates the asymptotic null distribution of $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z}))$

Example 19. Consider a finite population model and suppose that we construct a random variable $\mathcal{X}^{(N)}$ which converges in distribution to $\mathcal{X}$ where $\mathcal{X} \sim \mathcal{N}(0, \Sigma_{fitted} + V)$ for

$$V = \lim_{N \to \infty} N \left(\frac{S_{\dot{\epsilon}_1}^{(N)}}{n_1} + \frac{S_{\dot{\epsilon}_0}^{(N)}}{n_0} \right).$$

The difference $V - \Sigma_{resid,\mathscr{F}} = \lim_{N \to \infty} S_{\dot{\epsilon}_1 - \dot{\epsilon}_0}^{(N)}$ is guaranteed to be positive semidefinite since $S_{\dot{\epsilon}_1 - \dot{\epsilon}_0}^{(N)}$ is positive semidefinite by construction and the positive semidefinite cone is topologically closed. Consequently, the distribution of $f_{\xi}(\mathcal{X})$ stochastically dominates the asymptotic null distribution of $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z}))$.

Let $\mathcal{A}_T(Z)$ denote a algorithm which takes observed units $\{(y_i(Z_i), x_i)\}_{i=1}^N$ as inputs and generates a random variable conditional upon $\{(y_i(Z_i), x_i)\}_{i=1}^N$. In general, we consider algorithms which involve using the observed units $\{(y_i(Z_i), x_i)\}_{i=1}^N$ to form some new "imputed"

population of N individuals and then apply a test statistic $T(\cdot)$ to this imputed population; as such, the subscript T provides a convenient way to keep track of this dependence on T.

Example 20. The *i.i.d.* bootstrap resampling procedure presented in Algorithm 1 constructs an "imputed population" by *i.i.d.* resampling from the treated observations $\{(y_i(Z_i), x_i)\}_{i\,:\,Z_i=1}$ and control observations $\{(y_i(Z_i), x_i)\}_{i\,:\,Z_i=0}$. If one chooses the test statistic $T_1(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\emptyset}\right)$ then the resampling algorithm $\mathcal{A}_{T_1}(Z)$ produces the random variable (4) conditional upon $\{(y_i(Z_i), x_i)\}_{i=1}^N$. If instead one chooses the test statistic $T_1(\mathbf{y}(\mathbf{Z})) = \left|\left|\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\emptyset}\right)\right|\right|_2^2$ then the resampling algorithm $\mathcal{A}_{T_2}(Z)$ produces the random variable

$$\left|\left|\sqrt{N}\left(\underbrace{\left(\mathbb{E}_{\hat{F}_N^{1,*}(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^{0,*}(Z)}\left[\Pi_y(y,x)\right]\right)}_{\textit{Bootstrap Resampling Term}} - \underbrace{\left(\mathbb{E}_{\hat{F}_N^{1}(Z)}\left[\Pi_y(y,x)\right] - \mathbb{E}_{\hat{F}_N^{0}(Z)}\left[\Pi_y(y,x)\right]\right)}_{\textit{Centering by Empirical Means}}\right)\right|\right|_2^2.$$

Theorem A.20. *Suppose that:*

- *For* $T_1(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)$ *the algorithm* $\mathcal{A}_{T_1}(Z)$ *produces the random variable* $\mathcal{X}^{(N)}$,
- *For all conditioning events up to a set of measure zero, the random variable* $\mathcal{X}^{(N)}$ *conditional upon the observed units* $\{(y_i(Z_i), x_i)\}_{i=1}^N$ *converges in distribution to* $\mathcal{X} \sim \mathcal{N}(0, S_{\mathcal{X}})$.

If $S_{\mathcal{X}} \succeq \lim_{N\to\infty} \mathbb{V}\left(\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)\right)$ *then the algorithm* $\mathcal{A}_T(Z)$ *is strongly consistently conservative for any test statistic* $T(\cdot)$ *of the form* (1) *subject to Conditions 3 and 4.*

Proof. The result follows directly from Corollary A.1 by taking $\mathcal{Y}$ to be the weak limit of $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$. □

Theorem A.20 implies that any resampling algorithm which produces a conservative bootstrap null distribution for the difference in means is automatically consistently conservative for any test statistic $T(\cdot)$ of the form (1) subject to Conditions 3 and 4. This result facilitates proving that a resampling algorithm is consistently conservative by reducing the problem to only considering the behavior of the resampling algorithm in the context of the difference in means. It even further reduces the workload of the analysis by showing that one need only demonstrate two features of the behavior of the resampling algorithm for the difference in means:

1. A conditional central limit theorem holds almost surely for the resampled difference in means,

2. The resulting asymptotic variance of the conditional resampled difference in means exceeds the true asymptotic variance of the difference in means.

By similar reasoning one can compare two resampling algorithms through an analysis only of their behavior with the difference in means.

Theorem A.21. *Consider two resampling algorithms $\mathcal{A}_T(Z)$ and $\mathcal{A}'_T(Z)$ Suppose that:*

- *For $T_1(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)$ the algorithm $\mathcal{A}_{T_1}(Z)$ produces the random variable $\mathcal{X}^{(N)}$ and the algorithm $\mathcal{A}'_{T_1}(Z)$ produces the random variable $\mathcal{X}'^{(N)}$*

- *For all condition events up to a set of measure zero, the random variable $\mathcal{X}^{(N)}$ conditional upon the observed units $\{(y_i(Z_i), x_i)\}_{i=1}^N$ converges in distribution to $\mathcal{X} \sim \mathcal{N}(0, S_{\mathcal{X}})$ and likewise $\mathcal{X}'^{(N)}$ conditionally converges in distribution to $\mathcal{X}' \sim \mathcal{N}(0, S_{\mathcal{X}'})$*

If $S_{\mathcal{X}} \succeq S_{\mathcal{X}'}$ then the algorithm $\mathcal{A}_T(Z)$ is strongly more conservative than $\mathcal{A}'_T(Z)$ for any test statistic $T(\cdot)$ of the form (1) subject to Conditions 3 and 4.

Theorem A.21 provides a similar tool to that of Theorem A.20: it reduces proving that one resampling algorithm is more conservative than another to simply examining their behavior

for the difference in means, showing conditional central limit theorems, and comparing the resulting asymptotic variances.

4.15 Central Limit Theorems For The *I.I.D.* Bootstrap

With Theorems A.20 and A.21 in mind, we turn to analyzing the behavior of the bootstrap distribution of the difference in means formed by Algorithm 1. In this section we prove central limit theorems for the resampled difference in means under Algorithm 1 in the superpopulation, fixed covariate, and finite population models.

Theorem A.22. *Let the bootstrap conditional distribution generated by the i.i.d. resampling procedure of Algorithm 1 applied to the difference in means* $T_{\emptyset}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\emptyset}\right)$ *be denoted by* $\mathcal{L}_{\hat{\tau}^*,\emptyset}(Z)$*. Under the superpopulation model subject to Assumptions A.2 and A.3 the random distribution* $\mathcal{L}_{\hat{\tau}^*,\emptyset}(Z)$ *limits weakly to the law of* $\mathcal{N}\left(0, p^{-1}\Sigma_{y(1)} + (1-p)^{-1}\Sigma_{y(0)}\right)$ *almost surely. Formally,*

$$\rho_{BL}\left(Law_{\hat{\tau}^*,\emptyset}(Z), \gamma_{0,\frac{\Sigma_{y(1)}}{p}+\frac{\Sigma_{y(0)}}{1-p}}\right) \xrightarrow{a.s.} 0.$$

Proof. Algorithm 1 forms the conditional distribution of $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z}))$ under the following resampling procedure:

1. From the observed treated units $\{(y_i(Z_i), x_i)\}_{i\,:\,Z_i=1}$ uniformly at random select with replacement N observations to form a resampled treated population of $\{(y_i^*(1), x_{i1}^*)\}_{i=1}^N$. Analogous resampling is performed on the observed control units $\{(y_i(Z_i), x_i)\}_{i\,:\,Z_i=0}$ to form $\{(y_i^*(0), x_{i0}^*)\}_{i=1}^N$.

2. Generated an independent $B \sim$ *Unif*(Ω) to serve as a treatment allocation vector for this new population of N units.

3. Take the "bootstrap treated units" to be $\{(y_i^*(1), x_{i1}^*) \ : \ B_i = 1\}$ and the "bootstrap control units" to be $\{(y_i^*(0), x_{i0}^*) \ : \ B_i = 0\}$.

It is easily verified by the independence of B that this procedure is equivalent to:

1. From the observed treated units $\{(y_i(Z_i), x_i)\}_{i \, : \, Z_i=1}$ uniformly at random select with replacement n_1 observations to form the "bootstrap treated units" $\{(y_i^*(1), x_{i1}^*)\}_{i=1}^{n_1}$. Similarly from the observed control units $\{(y_i(Z_i), x_i)\}_{i \, : \, Z_i=0}$ uniformly at random select with replacement n_0 observations to form the "bootstrap control units" $\{(y_i^*(0), x_{i0}^*)\}_{i=1}^{n_0}$.

After this reframing it follows that Algorithm 1 is equivalent to Efron's classical bootstrap of *i.i.d.* resampling for the sample mean applied twice independently: once to form a bootstrap null distribution for the treated mean and once to form a bootstrap null distribution for the control mean. Applying (LR05, Theorem 15.4.5) provides yields an analysis of the asymptotic bootstrap distribution for each of these two separate mean-estimation problems. Finally, re-weighting the variances to account for the relative samples sizes of N, n_1, and n_0 and recalling that the sum of independent Gaussian random variables is itself Gaussian yields the desired result that

$$\rho_{BL}\left(Law_{\hat{\tau}^*,\emptyset}(Z), \gamma_{0, \frac{\Sigma_{y(1)}}{p} + \frac{\Sigma_{y(0)}}{1-p}}\right) \xrightarrow{a.s.} 0.$$

□

Remark 7. Theorem 3 follows immediately from Theorems A.14 and A.22 and the fact that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence.

Theorem A.23. *Let the bootstrap conditional distribution generated by the i.i.d. resampling procedure of Algorithm 1 applied to the difference in means* $T_{\mathscr{C}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{C}}\right)$ *be denoted by* $\mathcal{L}_{\hat{\tau}^*,\mathscr{C}}(Z)$. *Under the fixed covariate model subject to Assumptions A.4 and A.5*

the random distribution $\mathcal{L}_{\hat{\tau}^*,\mathscr{C}}(Z)$ *limits weakly to the law of*

$$\mathcal{N}\left(0, \lim_{N\to\infty} N\left(\frac{\Sigma_{y(1)}^{(N),\mathscr{C}}}{n_1} + \frac{\Sigma_{y(0)}^{(N),\mathscr{C}}}{n_0}\right)\right)$$

almost surely where $\Sigma_{y(z)}^{(N),\mathscr{C}}$ *is defined in* (31)*, respectively. Formally,*

$$\rho_{BL}\left(\mathcal{L}_{\hat{\tau}^*,\mathscr{C}}(Z), \gamma_{0,\left(\frac{\Sigma_{y(1)}^{\mathscr{C}}}{p} + \frac{\Sigma_{y(0)}^{\mathscr{C}}}{1-p}\right)}\right) \xrightarrow{a.s.} 0.$$

Proof. The same reframing of Algorithm 1 that was applied in the proof of Theorem A.22 applies equally well in the fixed covariate context since the reframing of Algorithm 1 as two *i.i.d.* bootstrap resampling procedures is not based at all upon the underlying data generating procedure. All that is required to complete the analysis is a suitable replacement for Theorem 15.4.5 of (LR05) that can account for the fact that the potential outcomes are no longer distributed identically.

Consider the subproblem of simply estimating the distribution of the sample mean in the treated population. In the univariate case of $d = 1$ it suffices to apply (LS95, Theorem 1) to establish the L_∞-metric consistency of the conditional bootstrap distribution of the empirical mean to the true sampling distribution of the sample mean. To adapt to the multivariate case of $d \geq 2$ one applies the Cramér–Wold device (LR05, Theorem 11.2.3) and then follows the same line of reasoning.

□

Remark 8. The result of Theorem 4 pertaining to the fixed covariate model follows from Theorems A.15 and A.23, the fact that $\Sigma_{y(1)-y(0)}^{\mathscr{C}} \succeq \Sigma_{\tau}^{\mathscr{C}}$, and the fact that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence.

Theorem A.24. *Let the bootstrap conditional distribution generated by the i.i.d. resampling*

procedure of Algorithm 1 applied to the difference in means $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{F}}\right)$ *be denoted by* $\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z)$.

Under the finite population model subject to Assumptions A.6 and A.7 the random distribution $\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z)$ *limits weakly to the law of* $\mathcal{N}\left(0, p^{-1}\Sigma^{\mathscr{F}}_{y(1)} + (1-p)^{-1}\Sigma^{\mathscr{F}}_{y(0)}\right)$ *almost surely where* $\Sigma^{\mathscr{F}}_{y(z)}$ *is defined in* (14). *Formally,*

$$\rho_{BL}\left(\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z), \gamma_{0, \frac{\Sigma^{\mathscr{F}}_{y(1)}}{p} + \frac{\Sigma^{\mathscr{F}}_{y(0)}}{1-p}}\right) \xrightarrow{a.s.} 0.$$

Proof. By mirroring the arguments from the proofs of Theorems A.22 and A.23 it suffices to show a conditional central limit theorem for *i.i.d.* bootstrap resampling from the observed treated units $\{y_i(Z_i) \ : \ Z_i = 1\}$ almost surely with respect to randomness in the treatment allocations.

Under Assumptions A.6 and A.7 the first two sample moments obey finite population strong laws of large numbers (WD21, Lemma A3) such that – for the treated potential outcomes – the sample mean converges almost surely to the population mean and the sample covariance matrix converges almost surely to the population covariance matrix. Furthermore, Assumptions A.6 and A.7 are sufficient for the finite population central limit theorem of (LD17) and so the assumptions of Lemma A.23 are met. This establishes that the *i.i.d* bootstrapped sample mean of the treated units $\{y_i(Z_i) \ : \ Z_i = 1\}$ obeys a central limit theorem almost surely with respect to randomness in the treatment allocations so that

$$\sqrt{n_1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} y_i^*(1) - \frac{1}{n_1}\sum_{i \,:\, Z_i = 1} y_i(1)\right) \mid Z \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{\mathscr{F}}_{y(1)}\right)$$

almost surely in Z. The analogous result holds for the control quantities. Finally, since the sum of independent Gaussians is Gaussian and ρ_{BL} metrizes weak convergence it follows

that

$$\rho_{BL}\left(\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z), \gamma_{0, \frac{\Sigma^{\mathscr{F}}_{y(1)}}{p} + \frac{\Sigma^{\mathscr{F}}_{y(0)}}{1-p}}\right) \xrightarrow{a.s.} 0.$$

□

Remark 9. The result of Theorem 4 pertaining to the finite population model follows from Theorems A.17 and A.24, the fact that $\Sigma^{\mathscr{F}}_{y(1)-y(0)} \succeq 0$, and the fact that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence.

4.16 Central Limit Theorems For The Residual Bootstrap

In parallel with Algorithm 2 we define Algorithm 5 which uses the population regression coefficients $\dot{\beta}_z$ to form the imputed population and the residuals instead of using the sample regression coefficients $\hat{\beta}_z$ as in the case of Algorithm 2. By the consistency of the empirical residuals and regression coefficients to their population analogues we may analyze Algorithm 2 by instead examining Algorithm 5; such a style of argument is familiar to the regression-adjustment literature; see for example the proofs of (CF21), (GB20), or (Lin13).

Algorithm 5: A population residual-based distributional estimator.

Input: An observed treatment allocation $Z \in \Omega$.

Result: The bootstrap distributional estimator $\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right)$.

Compute the imputed values $\dot{\mu}_0(\mathbf{x}_i)$ and $\dot{\mu}_1(\mathbf{x}_i)$ for each i according to (17) and define

$$\dot{C} = \{(\dot{\epsilon}_i(Z_i)) \ : \ Z_i = 0\},$$
$$\dot{T} = \{(\dot{\epsilon}_i(Z_i)) \ : \ Z_i = 1\}.$$

for $(D_0, D_1) \in \mathcal{P}(N, \dot{C}) \times \mathcal{P}(N, \dot{T})$ **do**

Say that

$$D_0 = \{(\dot{\epsilon}_i^*(0))\}_{i=1}^N,$$
$$D_1 = \{(\dot{\epsilon}_i^*(1))\}_{i=1}^N.$$

for $B \in \Omega$ **do**

Generate the "bootstrap experimental observations"

$$\{(\dot{\mu}_0(\mathbf{x}_i) + \dot{\epsilon}_i^*(0), \mathbf{x}_i) \ : \ B_i = 0\} \cup \{(\dot{\mu}_1(\mathbf{x}_i) + \dot{\epsilon}_i^*(1), \mathbf{x}_i) \ : \ B_i = 1\}.$$

Compute $T(\cdot)$ using the bootstrap experimental observations with centering by

$$\frac{1}{N}\sum_{i=1}^{N}\left(\dot{\mu}_1(\mathbf{x}_i) - \dot{\mu}_0(\mathbf{x}_i)\right) + \left(\frac{1}{n_1}\sum_{i \,:\, Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i \,:\, Z_i=0}\dot{\epsilon}_i(0)\right),$$

denote this $T^*_{D_0,D_1,B}$.

end

end

return

$$\mathcal{L}\left(T^*(y^*(Z)) \mid \mathbf{y}(\mathbf{Z}), Z\right) = \frac{\sum_{D_0,D_1,B}\delta_{T^*_{D_0,D_1,B}}}{\left|\mathcal{P}(N,\dot{C}) \times \mathcal{P}(N,\dot{T}) \times \Omega\right|}.$$

Our analysis of Algorithm 5 decomposes across two separate analyses. In Algorithm 5 the bootstrap population is formed by two independent procedures:

- Forming deterministic predicted outcomes $(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))$ for each individual,
- Using *i.i.d.* resampling to form residuals $(\dot{\epsilon}_i^*(0), \dot{\epsilon}_i^*(1))$.

Next an independent completely randomized experiment is performed on the resampled population formed by $\{(\dot{\mu}_0(\mathbf{x}_i) + \dot{\epsilon}_i^*(0), \dot{\mu}_1(\mathbf{x}_i) + \dot{\epsilon}_i^*(1))\}_{i=1}^N$. When examining the bootstrap resampling for the difference in means under Algorithm 2 we can examine these two steps separately – by their independence – and then recombine the results at the end. First we focus on the bootstrap difference in means in only the context of the predicted outcomes $(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))$. After doing so, we examine the bootstrap difference in means in only the context of the *i.i.d.* resampled residuals $(\dot{\epsilon}_i^*(0), \dot{\epsilon}_i^*(1))$. We will show central limit theorems for both of these quantities, which are then easily combined by leveraging independence.

Remark 10. As mentioned informally above, our results pertain to bootstrapping using the population linear regression coefficients $\dot{\beta}_z$ instead of those that are empirically observed, $\hat{\beta}_z$. However, by the consistency of $\hat{\beta}_z$ for $\dot{\beta}_z$ – formally that $\left|\left|\hat{\beta}_z - \dot{\beta}_z\right|\right|_2 \xrightarrow{a.s.} 0$ – the results proven using the population regression coefficients $\dot{\beta}_z$ apply to the procedures based upon $\hat{\beta}_z$. This is a standard proof technique in the regression adjustment literature; for instance, see (CF21), (GB20), (LD18), and (Lin13) among others. The same logic applies to the use of the population residuals $\dot{\epsilon}_i(z)$ instead of their empirically observed counterparts.

Lemma A.8. *Consider a fixed covariate model subject to Assumptions A.4 and A.5 and let* $B \sim$ *Unif*(Ω) *be draw independently. Then the random variable*

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\mu}_1(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\mu}_0(\mathbf{x}_i)\right) - \overline{\tau}_{\mathscr{C}}\right)$$

converges in distribution to a centered multivariate Gaussian with covariance matrix Σ_{fitted}.

Proof. By the first order condition of the population linear regression (15), linearity of expectation, and the fact that $\mathbb{E}[B_i] = n_1/N$ it follows that

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\mu}_1(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\mu}_0(\mathbf{x}_i)\right] = \overline{\tau}_{\mathscr{C}}.$$

The result then follows from the finite population central limit theorem (LD17) applied to the population $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$ which is justified under the regularity conditions of Assumptions A.4 and A.5. □

Lemma A.9. *Consider a finite population model subject to Assumptions A.6 and A.7 and let $B \sim$ Unif(Ω) be draw independently. Then the random variable*

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\mu}_1(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\mu}_0(\mathbf{x}_i)\right) - \overline{\tau}_{\mathscr{F}}\right)$$

converges in distribution to a centered multivariate Gaussian with covariance matrix Σ_{fitted}.

Proof. The proof mirrors that of Lemma A.8 save for the fact that the first order conditions of linear regression in (19) now imply that

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\mu}_1(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\mu}_0(\mathbf{x}_i)\right] = \overline{\tau}_{\mathscr{F}}.$$

□

Lemma A.10. *Consider a fixed covariate model subject to Assumptions A.4 and A.5 and let $B \sim$ Unif(Ω) be draw independently. Then, conditional upon Z and $\mathbf{y}(\mathbf{Z})$, the random variable*

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\epsilon}_i^*(B_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\epsilon}_i^*(B_i)\right) - \left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right)\right)$$

converges in distribution to a centered multivariate Gaussian with covariance matrix

$$\lim_{N\to\infty} N\left(\frac{\Sigma_{\dot{\epsilon}_1}^{(N)}}{n_1} + \frac{\Sigma_{\dot{\epsilon}_0}^{(N)}}{n_0}\right)$$

almost surely with respect to randomness in the conditioning random variables Z and $\mathbf{y}(\mathbf{Z})$ (where $\Sigma_{\dot{\epsilon}_z}^{(N)}$ is defined in (25)).

Proof. This result follows from applying the same reasoning of Theorem A.23 to the residuals $\dot{\epsilon}_i(z)$ instead of the potential outcomes $y_i(z)$. As before, the main working component of the result is Theorem 1 of (LS95).

□

Lemma A.11. *Consider a finite population model subject to Assumptions A.6 and A.7 and let $B \sim$ Unif(Ω) be draw independently. Then, conditional upon Z, the random variable*

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\epsilon}_i^*(B_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\epsilon}_i^*(B_i)\right) - \left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\dot{\epsilon}_i(1) - \frac{1}{n_0}\sum_{i\,:\,Z_i=0}\dot{\epsilon}_i(0)\right)\right)$$

converges in distribution to a centered multivariate Gaussian with covariance matrix

$$\frac{S_{\dot{\epsilon}_1}^{(N)}}{p} + \frac{S_{\dot{\epsilon}_0}^{(N)}}{1-p}$$

almost surely with respect to randomness in the conditioning random variable Z.

Proof. In the finite population model, the residuals $\dot{\epsilon}_i(z)$ are deterministic. The result follows from repeating the analysis of Theorem A.24 but applied to the finite population with potential outcomes $\{(y_i(0), y_i(1))\}_{i=1}^N$ replaced by $\{(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))\}_{i=1}^N$ instead. That the required assumptions on the population $\{(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))\}_{i=1}^N$ hold is justified by (CF22, Appendix Section 3).

□

Remark 11. In Algorithm 2 the bootstrapped residual quantity is presented as

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\epsilon_i^*(B_i)-\frac{1}{n_0}\sum_{i\,:\,B_i=0}\epsilon_i^*(B_i)\right)\right).$$

However, it is important to recognize that this is equivalent to

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\epsilon_i^*(B_i)-\frac{1}{n_0}\sum_{i\,:\,B_i=0}\epsilon_i^*(B_i)\right)-\left(\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\epsilon_i(1)-\frac{1}{n_0}\sum_{i\,:\,Z_i=0}\epsilon_i(0)\right)\right)$$

because the first order optimality condition of ordinary least squares linear regression ensures that $\frac{1}{n_z}\sum_{i\,:\,Z_i=z}\epsilon_i(z)=0$ for $z\in\{0,1\}$. This observation is important in the setting of Remark 10 when translating the implementation of Algorithm 2 to the context of Lemmas A.10 and A.11 where the sample residual $\epsilon_i(z)$ is replaced with the population residual $\dot{\epsilon}_i(z)$.

Theorem A.25. *Let the bootstrap conditional distribution generated by the residual resampling procedure of Algorithm 5 applied to* $T_{\mathscr{C}}(\mathbf{y}(\mathbf{Z}))=\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z}))-\overline{\tau}_{\mathscr{C}}\right)$ *be denoted by* $\mathcal{L}_{\hat{\tau}^*,\mathscr{C}}(Z)$*. Under the fixed covariate model subject to Assumptions A.4 and A.5 the random distribution* $\mathcal{L}_{\hat{\tau}^*,\mathscr{C}}(Z)$ *limits weakly to the distribution of*

$$\mathcal{N}\left(0,\Sigma_{fitted}+\lim_{N\to\infty}N\left(\frac{\Sigma_{\dot{\epsilon}_1}^{(N)}}{n_1}+\frac{\Sigma_{\dot{\epsilon}_0}^{(N)}}{n_0}\right).\right)$$

almost surely. Formally,

$$\rho_{BL}\left(\mathcal{L}_{\hat{\tau}^*,\mathscr{C}}(Z),\gamma_{0,\Sigma_{fitted}+\left(\frac{\Sigma_{\dot{\epsilon}_1}}{p}+\frac{\Sigma_{\dot{\epsilon}_0}}{1-p}\right)}\right)\xrightarrow{a.s.}0.$$

Proof. The theorem follows by combining Lemmas A.8 and A.10, noting that the sum of independent Gaussian random variables is itself Gaussian, and using that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence. □

Theorem A.26. *Let the bootstrap conditional distribution generated by the residual resampling procedure of Algorithm 5 applied to* $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{F}}\right)$ *be denoted by* $\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z)$*. Under the finite population model subject to Assumptions A.6 and A.7 the random distribution* $\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z)$ *limits weakly to the distribution of*

$$\mathcal{N}\left(0, \Sigma_{fitted} + p^{-1}S_{\dot{\epsilon}(1)} + (1-p)^{-1}S_{\dot{\epsilon}(0)}\right)$$

almost surely. Formally,

$$\rho_{BL}\left(\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z), \gamma_{0,\Sigma_{fitted}+\frac{S_{\dot{\epsilon}(1)}}{p}+\frac{S_{\dot{\epsilon}(0)}}{1-p}}\right) \xrightarrow{a.s.} 0$$

where $S_{\dot{\epsilon}(z)}$ *is defined in* (27)*.*

Proof. The theorem follows by combining Lemmas A.9 and A.11, noting that the sum of independent Gaussian random variables is itself Gaussian, and using that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence.

□

Remark 12. Theorem 6 follows from Theorems A.16 and A.25, the fact that $\Sigma_{\dot{\epsilon}_1-\dot{\epsilon}_0} \succeq \Sigma_{\tau}^{\mathscr{C}}$, and the fact that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence. Theorem 7 follows by comparing the limiting variances of Theorems A.23 and A.25. Likewise, Theorem 8 follows from Theorems A.17 and A.26, the fact that $S_{\dot{\epsilon}_1-\dot{\epsilon}_0} \succeq 0$, and the fact that the bounded Lipschitz metric, ρ_{BL}, metrizes weak convergence.

4.17 Central Limit Theorems For The Optimal Transport Bootstrap

As in our analysis of Algorithm 2, we decompose analysis of Algorithm 4 into two separate lines of reasoning: an examination of the predicted outcomes and an examination of the resampling of residuals according to Algorithm 3. As in Appendix Section 4.16, we exploit independence to combine the two separate analyses together. This final step relies critically upon the selection of an independent permutation π and the formation of the bootstrap population as

$$\left\{ \Big(\underbrace{\hat{\mu}_0(\mathbf{x}_i) + \epsilon^*_{\pi(i)}(0)}_{y_i^*(0)}, \mathbf{x}_i \Big) \right\},$$
$$\left\{ \Big(\underbrace{\hat{\mu}_1(\mathbf{x}_i) + \epsilon^*_{\pi(i)}(1)}_{y_i^*(1)}, \mathbf{x}_i \Big) \right\}.$$

Permuting the residuals by π enforces that a completely randomized experiment simulated on the bootstrap residuals is independent from a completely randomized experiment simulated on the fitted values. Consistent with Remark 10, we prove results for the population residuals and predicted values $\dot{\epsilon}_i(z)$ and $\dot{\mu}_z(\mathbf{x}_i)$ instead of their sample analogues. Lemma A.9, from above, handles the distributional behaviour of the bootstrapping schema based upon the population predicted values $\dot{\mu}_z(\mathbf{x}_i)$. Consequently the remaining analysis of Algorithm 4 rests upon examining the behavior of the difference in means in a completely randomized experiment with the residuals resampled according to Algorithm 3. To ease the presentation of our main results, we defer the proof to Appendix Sections 4.19 and 4.20. Furthermore, in Appendix Section 4.21 we provide an analysis of a bootstrap scheme of (IM21) which – while valid under further assumptions of marginal asymptotic continuity – may behave pathologically when the limiting marginal distributions of the treated or control potential

outcomes possess atoms.

We define Algorithm 6 which uses the population regression coefficients $\dot{\beta}_z$ to form the imputed population and the residuals instead of using the sample regression coefficients $\hat{\beta}_z$ as in the case of Algorithm 4.

Algorithm 6: An optimal-transport-based distributional estimator using population regression coefficients.

Input: An observed treatment allocation $Z \in \Omega$.

Result: The bootstrap distributional estimator $\mathcal{L}\left(T^*(\dot{y}^*(Z)) \mid \mathbf{y}(\mathbf{Z})\right)$.

Compute the imputed values $\dot{\mu}_0(\mathbf{x}_i)$ and $\dot{\mu}_1(\mathbf{x}_i)$ for each i according to (17) and define the residuals $\dot{\epsilon}_i(Z_i)$.

Define the random variables $\{\dot{\epsilon}_i^*(0), \dot{\epsilon}_i^*(1), \mathbf{x}_i\}_{i=1}^N$ as the output of Algorithm 3 computed on the observations $\{\dot{\epsilon}_i(Z_i)\}_{i=1}^N$.

Select an independent permutation $\pi \sim \mathit{Unif}(\mathcal{S}_N)$.

For an independent draw $B \sim \mathit{Unif}(\Omega)$ generate the "bootstrap experimental observations"

$$\left\{ \Big(\underbrace{\dot{\mu}_0(\mathbf{x}_i) + \dot{\epsilon}^*_{\pi(i)}(0)}_{\dot{y}_i^*(0)}, \mathbf{x}_i \Big) \ : \ B_i = 0 \right\} \cup \left\{ \Big(\underbrace{\dot{\mu}_1(\mathbf{x}_i) + \dot{\epsilon}^*_{\pi(i)}(1)}_{\dot{y}_i^*(1)}, \mathbf{x}_i \Big) \ : \ B_i = 1 \right\}.$$

Compute $T(\cdot)$ using the bootstrap experimental observations with centering by $\frac{1}{N}\sum_{i=1}^N \left(\dot{y}_i^*(1) - \dot{y}_i^*(0)\right)$, denote this random variable as $T^*(\dot{y}^*(Z))$.

return

$$\mathcal{L}\left(T^*(\dot{y}^*(Z)) \mid \mathbf{y}(\mathbf{Z})\right).$$

The following lemma examines the behavior of the difference of means in a completely randomized experiment simulated upon the resampled population residuals $\{\dot{\epsilon}_i^*(0), \dot{\epsilon}_i^*(1), \mathbf{x}_i\}_{i=1}^N$ from Algorithm 6.

Write the finite population residual measures

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{\dot{\epsilon}_i(0)} \quad \text{and} \quad \nu = \frac{1}{N}\sum_{i=1}^{N} \delta_{\dot{\epsilon}_i(1)}.$$

Define the *finite population maximal residual variance* $V^H_{resid,N}$ as

$$V^H_{resid,N} = \frac{1}{N}\Bigg(\left(\frac{n_0}{n_1}\right) \frac{1}{N-1}\sum_{i=1}^{N}\left(\dot{\epsilon}_i(1) - \frac{1}{1}\sum_{j=1}^{N}\dot{\epsilon}_j(1)\right)^2 + \left(\frac{n_1}{n_0}\right) \frac{1}{N-1}\sum_{i=1}^{N}\left(\dot{\epsilon}_i(0) - \frac{1}{1}\sum_{j=1}^{N}\dot{\epsilon}_j(0)\right)^2 + 2\sup_{\gamma\in C(\mu,\nu)} \mathrm{cov}_\gamma\left(\mathcal{X},\mathcal{Y}\right)\Bigg).$$

Under Assumptions A.6 and A.7 and the conditions of (AGL14, Proposition 1),

$$\lim_{N\to\infty} N V^H_{resid,N}$$

exists; for notation we write $V^H_{resid} = \lim_{N\to\infty} N V^H_{resid,N}$.

Lemma A.12. *Consider a finite population model with univariate potential outcomes subject to Assumptions A.6 and A.7 and the regularity conditions of (AGL14, Propostion 1). Let $B \sim$ Unif(Ω) be draw independently. Then, conditional upon Z, the random variable*

$$\sqrt{N}\left(\left(\frac{1}{n_1}\sum_{i\,:\,B_i=1}\dot{\epsilon}^*_i(B_i) - \frac{1}{n_0}\sum_{i\,:\,B_i=0}\dot{\epsilon}^*_i(B_i)\right) - \left(\frac{1}{N}\sum_{i=1}^{N}\dot{\epsilon}^*_i(1) - \frac{1}{N}\sum_{i1}^{N}\dot{\epsilon}^*_i(0)\right)\right)$$

converges in distribution to a centered multivariate Gaussian with covariance matrix V^H_{resid} almost surely with respect to randomness in the conditioning random variable Z.

We defer proof of Lemma A.12 to Appendix Section 4.20; the result follows from applying

Theorem A.29 to the population formed by the residuals $\{\dot{\epsilon}_i(0), \dot{\epsilon}_i(1), \mathbf{x}_i\}_{i=1}^N$.

Theorem A.27. *Let the bootstrap conditional distribution generated by the residual resampling procedure of Algorithm 6 applied to* $T_{\mathscr{F}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{F}}\right)$ *be denoted by* $\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z)$. *Under the finite population model subject to Assumptions A.6 and A.7 the random distribution* $\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z)$ *limits weakly to the law of* $\mathcal{N}\left(0, \Sigma_{fitted} + V_{resid}^H\right)$ *almost surely. Formally*

$$\rho_{BL}\left(\mathcal{L}_{\hat{\tau}^*,\mathscr{F}}(Z), \gamma_{0,\Sigma_{fitted}+V_{resid}^H}\right) \xrightarrow{a.s.} 0.$$

Proof. The result follows from combining Lemmas A.9 and A.12, noting that that sum of independent Gaussians is itself Gaussian, and using that ρ_{BL} metrizes weak convergence. □

Remark 13. Theorem 11 follows from Theorems A.17 and A.27. The conservativeness of the limiting conditional bootstrap variance is a direct consequence of (AGL14, Lemma 1) and (DFM19, Section 4.1).

4.18 Variance Estimators

The variance estimators proposed in the main text are constructed by applying the bootstrap algorithms of Sections 4.6, 4.7, and 4.8 to the test statistic $T_{\mathscr{S}}(\mathbf{y}(\mathbf{Z})) = \sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z})) - \overline{\tau}_{\mathscr{S}}\right)$. We now prove that the proposed estimators are indeed consistent or asymptotically conservative as stated in Theorems 5, 9, and 12.

4.18.1 A General Purpose Consistency Theorem

In general, these three theorems – under suitable regularity conditions – are all implied by the following principle: *uniform integrability translates consistent (resp. consistently conservative) bootstrap procedures to consistent (resp. asymptotically conservative) variance estimators.* First, recall that a sequence of random variables $\mathcal{X}^{(N)}$ is *asymptotically uniformly*

integrable if

$$\lim_{M\to\infty}\lim_{N\to\infty}\mathbb{E}\left[\left|\mathcal{X}^{(N)}\right|\mathbb{1}_{\left\{\left|\mathcal{X}^{(N)}\right|>M\right\}}\right]=0.$$

The following theorem – which is simply a combination of Theorem 2.20 and Example 2.21 of (vdV98) – generally dictates that asymptotic uniform integrability is sufficient for weak convergence to imply convergence of variances.

Theorem A.28. *If* $\mathcal{X}^{(N)}\xrightarrow{d}\mathcal{X}$ *and*

$$\limsup_{N\to\infty}\mathbb{E}\left[\left|\mathcal{X}^{(N)}\right|^{\ell}\right]$$

is bounded then $\left\{\left|\mathcal{X}^{(N)}\right|^{\ell'}\right\}_{N\in\mathbb{N}}$ *is asymptotically uniformly integrable and*

$$\mathbb{E}\left[\left(\mathcal{X}^{(N)}\right)^{\ell'}\right]\to\mathbb{E}\left[\mathcal{X}^{\ell'}\right]$$

for all $\ell'<\ell$.

In particular, Assumption A.3 (bounded fourth moments in the superpopulation model) guarantees that the conditional bootstrap random variable of Algorithm 1 applied to the difference in means $T_{\emptyset}(\mathbf{y}(\mathbf{Z}))=\sqrt{N}\left(\hat{\tau}(\mathbf{y}(\mathbf{Z}))-\overline{\tau}_{\emptyset}\right)$ has bounded limit superior of its fourth moment in a superpopulation model almost surely with respect to the conditioning upon the realized observations $\mathbf{y}(\mathbf{Z})$ and treatment allocation Z. Consequently, Theorem A.28 (and the two central limit theorems of Theorem A.14 and A.22) allows us to conclude that $\hat{V}_1$ is indeed consistent at the superpopulation level. Similar analyses leveraging the other bounded fourth moment assumptions (Assumptions A.5 and A.7) allow for control of the conditional bootstrap fourth moments output by Algorithms 1 and 4, then applying Theorem A.28 and comparing the appropriate central limit theorems from Appendix Section 4.14.1 to the bootstrap conditional central limit theorems of Appendix Sections 4.15, 4.16, and 4.17 provides the remainder of the results.

In addition to the argument above based upon uniform integrability, we sketch out some alternative arguments below which may be more intuitive but draw upon more case-by-case analyses. However, we note that Theorem A.28 provides a powerful general-purpose tool to prove such results under other regularity conditions or for other bootstrap schema.

4.18.2 The *I.I.D.* Bootstrap Variance Estimator

To prove Theorem 5 we note that the consistency of $\hat{V}_1$ at the superpopulation level (resp. conservativeness of $\hat{V}_1$ at the fixed covariate level) reduces examining Efron's *i.i.d.* bootstrap variance estimators in the *i.i.d.* data generating model of (LR05, Example 15.4.2) (resp. the independent but not identically distributed data generating model of (LS95)). The consistency of the bootstrap variance estimator of Efron's *i.i.d.* bootstrap in the *i.i.d.* data generating model of (LR05, Example 15.4.2) is guaranteed by (LR05, Theorem 15.4.5) under Assumptions A.2 and A.3. Likewise, the conservativeness of the bootstrap variance estimator of Efron's *i.i.d.* bootstrap in the independent but not identically distributed data generating model of (LS95) is guaranteed by (LS95, Equation 2.2) under Assumptions A.4 and A.5.

In the finite population model, we note that the conservativeness of $\hat{V}_1$ reduces to the fact that $\hat{V}_1$ is exactly Neyman's classical variance estimator. In a finite population model subject to Assumptions A.6 and A.7 Neyman's variance estimator converges in probability to $p^{-1}\Sigma^{\mathscr{F}}_{y(1)} + (1-p)^{-1}\Sigma^{\mathscr{F}}_{y(0)}$ which is conservative for the true variance of $\sqrt{N}\hat{\tau}(\mathbf{y}(\mathbf{Z}))$; see (AGL14, Section 2.1) for a discussion of Neyman's work and (CF22, Lemma E) for a proof of the required convergence in probability.

4.18.3 The Residual Bootstrap Variance Estimator

As discussed in Appendix Section 4.16 the analysis of Algorithm 1 decomposes into an analysis of a completely randomized experiment on the population $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$ and a separate analysis of the *i.i.d.* resampling procedure of Algorithm 1 applied to the residuals

$\{(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))\}_{i=1}^N$. The arguments of Section 4.18.2 handle the component of $\hat{V}_2$ controlled by *i.i.d.* resampling of residuals, so all that remains is an analysis of the component of $\hat{V}_2$ driven by a completely randomized experiment on the population $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$.

In both the fixed covariate and finite population model, the set $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$ is fully deterministic and the variance of $\sqrt{N}$-scaled difference in means for a completely randomized experiment performed on this population is given by

$$N\left(\frac{\Sigma_{\dot{\mu}_1}^{(N)}}{n_1} + \frac{\Sigma_{\dot{\mu}_0}^{(N)}}{n_0} - \frac{\Sigma_{\dot{\mu}_1-\dot{\mu}_0}^{(N)}}{N}\right)$$

which limits, under either Assumptions A.4 and A.5 or Assumptions A.6 and A.7, to Σ_{fitted} as defined in (29). This exactly characterizes the behavior of the component of $\hat{V}_2$ governed by $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$ and the proof of Theorem 9 follows from combining this result with the arguments of Section 4.18.2 analyzing the component of $\hat{V}_2$ controlled by *i.i.d.* resampling of residuals.

4.18.4 The Optimal Transport Bootstrap Variance Estimator

As discussed in Appendix Section 4.17 the introduction of the independent permutation π in Algorithm 4 (and likewise in Algorithm 6) allows one to decompose analysis of $\hat{V}_3$ into an analysis of a completely randomized experiment on the population $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$ and a separate analysis of the optimal transport-based resampling procedure of Algorithm 3 applied to the residuals $\{\dot{\epsilon}_i(Z_i)\}_{i=1}^N$. The argument of Appendix Section 4.18.3 handles the component of $\hat{V}_3$ driven by a completely randomized experiment on the population $\{(\dot{\mu}_0(\mathbf{x}_i), \dot{\mu}_1(\mathbf{x}_i))\}_{i=1}^N$. Consequently, all that remains is to examine the component of $\hat{V}_3$ which is controlled by the optimal transport-based resampling procedure of Algorithm 3 applied to the residuals $\{\dot{\epsilon}_i(Z_i)\}_{i=1}^N$.

Theorem 10 and Lemma 4 assert that the variance estimator formed via the optimal

transport sampling of Algorithm 3 matches the variance upper bound estimator of (AGL14) up to a factor of $N/(N-1)$. Consequently, the portion of $\hat{V}_3$ which is controlled by the optimal transport-based resampling procedure of Algorithm 3 applied to the residuals $\{\dot{\epsilon}_i(Z_i)\}_{i=1}^N$ may be analyzed by examining the N-scaled variance upper bound estimator of (AGL14, Equation 9) applied to the residuals $\{\dot{\epsilon}_i(Z_i)\}_{i=1}^N$. This is exactly the subject of (AGL14, Proposition 1) which guarantees the variance estimator converges in probability to a conservative limit.

4.19 Bootstrap Sampling From The Optimal Coupling

In a slight abuse of notation, we ignore covariates for this section and write

$$\hat{F}_N^0(Z) = \frac{1}{n_0} \sum_{i\,:\,Z_i=0} \delta_{\mathbf{y}_i(0)}$$
$$\hat{F}_N^1(Z) = \frac{1}{n_1} \sum_{i\,:\,Z_i=1} \delta_{\mathbf{y}_i(1)}.$$

We examine sampling from the optimal coupling from the perspectives of Algorithms 3 and 4 directly on a finite population *sans* regression adjustment. This facilitates direct comparison between our results and the existing literature: particularly (AGL14) and (IM21) which do not consider regression-based frameworks. However, our analysis occurs without loss of generality as one can simply apply the results presented below to the finite population residuals $\{(\dot{\epsilon}_i(0), \dot{\epsilon}_i(1))\}_{i=1}^N$ in place of $\{(\mathbf{y}_i(0), \mathbf{y}_i(1))\}_{i=1}^N$.

The random variable underlying Algorithm 4 can be constructed as follows. Given observed outcomes $\mathbf{y}(\mathbf{Z})$, form a new population via probabilistic imputation according to the

optimal coupling γ_{opt}:

Control	*Treated*
$\vdots$	$\vdots$
$\mathbf{y}_i(0)$	?
?	$\mathbf{y}_j(1)$
$\mathbf{y}_k(0)$	?
$\vdots$	$\vdots$

$\xrightarrow[(\textit{using Alg. 3})]{\textit{Sample}}$

Control	*Treated*
$\vdots$	$\vdots$
$\mathbf{y}_i(0)$	$y_i^*(1)$
$y_j^*(0)$	$\mathbf{y}_j(1)$
$\mathbf{y}_k(0)$	$y_k^*(1)$
$\vdots$	$\vdots$

and then draw a realization of the difference in means statistic from a completely randomized experiment on this hypothetical population after centering by difference in means of the sampled population. Denoting this random variable as $T^*(Z, B)$ it follows that the random variable returned by Algorithm 4 is distributed according to the conditional distribution of $T^*(Z, B)$ given Z, and hence also given $\mathbf{y}(\mathbf{Z})$, since – in the finite population model – $\mathbf{y}(\mathbf{Z})$ is fully determined by Z.

4.19.1 Relation to Previous Literature

In (IM21), a bootstrapping procedure related to the Frechét-Hoeffding coupling is proposed for finite populations wherein the limiting marginals are continuous cumulative distribution functions. We now address how this procedure relates to Algorithms 3 and 4; in particular highlighting the differences between the behavior of resampling according to (IM21, Equation 3.4) and sampling according to Algorithm 4 when the limiting continuity assumption is discarded.

We begin by taking a slight detour to establish some historical background. Ranging as far back as Hoeffding, it was known that for any two random variables X and Y with marginal distributions μ and ν, respectively, the coupling $\gamma \in C(\mu, \nu)$ which maximized the

off-diagonal term in the covariance matrix of (X, Y) is the so-called *comonotone coupling*, γ_H (Tch80). Writing the induced cumulative distribution functions of μ and ν as F_μ and F_ν, respectively, the comonotone coupling is defined as the joint distribution corresponding to the multivariate cumulative distribution function

$$\overline{H}(x, y) := \min \{F_\mu(x), F_\nu(y)\} \tag{34}$$

Section 2 of (Tch80) contains a proof of this fact; Section 2.5 of (Nel06) contains further details. An immediate corollary of this is that the infimum in (9) is achieved, and so it may be replaced by a min without loss of rigour (likewise, the suprema of (10) are achieved and can be replaced with max). In (AGL14), the result that $\overline{H}$ achieves the maximal covariance of (X, Y) is leveraged directly to compute an asymptotically sharp variance estimator in completely randomized experiments.

One can sample from the comonotone coupling via

$$\left(F_\mu^{-1}(U), F_\nu^{-1}(U)\right) \quad \text{for } U \sim \text{Unif}(0, 1) \tag{35}$$

where the quantile function F_μ^{-1} denotes the generalized inverse of F_μ and is defined as

$$F_\mu^{-1}(p) := \inf \{x \in \mathbb{R} \ : \ p \leq F_\mu(x)\}$$

and F_ν^{-1} is defined analogously; see Section 2.9 of (Nel06) for justification of this procedure and Proposition 2.2 of (San15) for further theoretical discussion. Crucially, this sampling procedure does not rely upon continuity (or strict monotonicity) of the cumulative distribution functions F_μ and F_ν.

In two special cases (35) can be rewritten:

$$\left(F_{\mu}^{-1}\left(F_{\nu}(Y)\right), Y\right) \text{ for } Y \sim F_{\nu} \text{ if } F_{\nu} \text{ is continuous,} \tag{36}$$

$$\left(X, F_{\nu}^{-1}\left(F_{\mu}(X)\right)\right) \text{ for } X \sim F_{\mu} \text{ if } F_{\mu} \text{ is continuous.} \tag{37}$$

The immediate justification for (36) and (37) is the probability integral transform, which relies critically upon the continuities remarked above. Equations (36) and (37) form the intuition for the bootstrap resampling algorithm of (IM21, Equation (3.4)). Crucially, (36) and (37) are not equivalent to (35) when F_{ν} or F_{μ}, respectively, is discontinuous. This yields an important practical consequence:

Lemma A.13. *When the limiting cumulative distribution function of the potential outcomes under control (or under treatment) is not continuous, the conditional bootstrap distribution variance produced by the resampling procedure of Equation (3.4) in (IM21) need not align with the variance estimator $\hat{V}_N^H$ of (AGL14). In fact, the conditional bootstrap distribution of (IM21, Equation (3.4)) may not converge to any fixed limit.*

In Section 4.21.3 we construct a concrete example of Lemma A.13 in the particular case of binary potential outcomes. Theorem A.31 rigorously analyzes the particular counterexample to demonstrate that the conditional bootstrap distribution of (IM21, Equation (3.4)) fails to converge to a fixed distribution, but instead converges – in a sense made precise below in Section 4.21.4 – to a **random** distribution.

The proof techniques of (AGL14) do not immediately generalize to a desirable resampling procedure. Instead, as motivated in the Section 4.8, one can view Proposition 1 of (AGL14) through the lens of optimal transport. Suppose that one solves the optimal transport problem

(9) for $\mu = \hat{F}_N^1(Z)$ and $\nu = \hat{F}_N^0(Z)$. Direct computation yields

$$\begin{aligned}
\inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \mathbb{E}_\gamma \left[|X - Y|^2 \right] &= \inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \left(\mathbb{E}_\gamma \left[X^2 \right] + \mathbb{E}_\gamma \left[Y^2 \right] - 2\mathbb{E}_\gamma \left[XY \right] \right) \\
&= \mathbb{E}_{\hat{F}_N^1(Z)} \left[X^2 \right] + \mathbb{E}_{\hat{F}_N^0(Z)} \left[Y^2 \right] + \\
&\qquad \inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \left(-2\mathbb{E}_\gamma \left[XY \right] \right) \\
&= \mathbb{E}_{\hat{F}_N^1(Z)} \left[X^2 \right] + \mathbb{E}_{\hat{F}_N^0(Z)} \left[Y^2 \right] - \\
&\qquad \underbrace{2 \sup_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \mathbb{E}_\gamma \left[XY \right]}_{\hat{\sigma}_H}. \qquad (38)
\end{aligned}$$

Since the value of $\inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \mathbb{E}_\gamma \left[|X - Y|^2 \right]$ can easily be computed via open-source optimization software (FCG+21) we can easily compute $\hat{\sigma}_H$ from the observed data via

$$\hat{\sigma}_H = \frac{1}{2} \left(\mathbb{E}_{\hat{F}_N^1(Z)} \left[X^2 \right] + \mathbb{E}_{\hat{F}_N^0(Z)} \left[Y^2 \right] - \inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \mathbb{E}_\gamma \left[|X - Y|^2 \right] \right). \qquad (39)$$

Consequently, denoting the sample variance of the observed treated (resp. control) outcomes as $\hat{\Sigma}_1$ (resp. $\hat{\Sigma}_0$) the variance upper-bound of (AGL14) can be directly rewritten as

$$\begin{aligned}
\hat{V}_N^H &= \frac{1}{N-1} \left(\frac{n_0}{n_1} \hat{\Sigma}_1 + \frac{n_1}{n_0} \hat{\Sigma}_0 + 2\hat{\sigma}_H \right) \\
&= \frac{1}{N-1} \left(\frac{n_0}{n_1} \hat{\Sigma}_1 + \frac{n_1}{n_0} \hat{\Sigma}_0 + \right. \\
&\qquad \left. \left(\mathbb{E}_{\hat{F}_N^1(Z)} \left[X^2 \right] + \mathbb{E}_{\hat{F}_N^0(Z)} \left[Y^2 \right] - \inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \mathbb{E}_\gamma \left[|X - Y|^2 \right] \right) \right).
\end{aligned}$$

Inherently, this equation demonstrates that the formula for $\hat{V}_N^H$ given by Equations (8) and

(9) of (AGL14) essentially relies upon a closed-form solution to

$$\inf_{\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))} \mathbb{E}_\gamma \left[|X - Y|^2 \right].$$

This suggests leveraging the optimal coupling $\gamma \in C(\hat{F}_N^1(Z), \hat{F}_N^0(Z))$ directly to create a resampling algorithm which recovers $\hat{V}_N^H$. This serves exactly as the motivation for Algorithms 3 and 4.

4.20 Analyzing The Optimal Transport Bootstrap Distribution

4.20.1 Some Preliminary Computations

To start, we compute some simple results concerning the moments of the bootstrap distribution.

$$\frac{1}{N} \sum_{j=1}^{N} \mathbf{y}_j^*(1) = \frac{1}{N} \sum_{j \,:\, Z_j = 1} \mathbf{y}_j(1) + \frac{1}{N} \sum_{j \,:\, Z_j = 0} \mathbf{y}_j^*(1).$$

The first term on the right-hand-side is simply $n_1 N^{-1} \mu(\hat{F}_N^1(Z))$ where $\mu(\cdot)$ is the expectation operator (conditional upon Z). We turn attention to the second term; this is a sum over the n_0 individuals who received control, we start by writing it as $n_0 N^{-1} n_0^{-1} \sum_{j \,:\, Z_j = 0} \mathbf{y}_j^*(1)$ and examine $n_0^{-1} \sum_{j \,:\, Z_j = 0} \mathbf{y}_j^*(1)$. Our object of interest is

$$\mathbb{E}\left[n_0^{-1} \sum_{j \,:\, Z_j = 0} \mathbf{y}_j^*(1) \mid Z \right]$$

Consider the following two-stage procedure:

1. Uniformly at random select an index from $\{j \ : \ Z_j = 0\}$; call this index j^*

2. Draw y as $y^*_{j^*}(1)$ (in other words, sample from the j^* column of the optimal γ matrix constructed in Algorithm 3).

By the tower property of conditional expectations, the uniformity of j^* over $\{j \ : \ Z_j = 0\}$, independence, and linearity

$$\begin{aligned}
\mathbb{E}\left[y \mid Z\right] &= \mathbb{E}\left[\mathbb{E}\left[y^*_{j^*}(1) \mid j^*, Z\right] \mid Z\right] \\
&= \sum_{j \,:\, Z_j = 0} \mathbb{P}\left(j^* = j \mid Z\right) \mathbb{E}\left[y^*_{j^*}(1) \mid j^* = j, Z\right] \\
&= \sum_{j \,:\, Z_j = 0} \mathbb{P}\left(j^* = j \mid Z\right) \mathbb{E}\left[\mathbf{y}^*_j(1) \mid Z\right] \\
&= \sum_{j \,:\, Z_j = 0} \frac{1}{n_0} \mathbb{E}\left[\mathbf{y}^*_j(1) \mid Z\right] \\
&= \mathbb{E}\left[n_0^{-1} \sum_{j \,:\, Z_j = 0} \mathbf{y}^*_j(1) \,\middle|\, Z\right]. \qquad (40)
\end{aligned}$$

Consequently, to understand $\mathbb{E}\left[n_0^{-1} \sum_{j \,:\, Z_j = 0} \mathbf{y}^*_j(1) \mid Z\right]$ we only need to understand $\mathbb{E}\left[y \mid Z\right]$. The two-step process listed above produces a pair $(y_{j^*}(0), y^*_{j^*}(1)$ where $y_{j^*}(0) \sim \hat{F}^0_N(Z)$ and $y^*_{j^*}(1)$ is distributed according to the conditional distribution of γ_{opt} conditioned upon the first coordinate being $y_{j^*}(0)$. Since the distribution γ_{opt} is supported on finitely many points, there are no concerns about the well-definedness of this conditional distribution. This procedure automatically generates pairs $(y_{j^*}(0), y^*_{j^*}(1) \sim \gamma_{opt}$; passing to y simply amounts to ignoring the first coordinate. Consequently, $\mathbb{E}\left[y \mid Z\right]$ is just the marginal distribution of γ_{opt} on the second coordinate. By construction, γ_{opt} marginalizes to $\hat{F}^1_N(Z)$ in this coordinate,

so $\mathbb{E}[y \mid Z] = \mu(\hat{F}_N^1(Z))$. Combining this with (40) implies that

$$\mathbb{E}\left[n_0^{-1} \sum_{j\,:\,Z_j=0} \mathbf{y}_j^*(1) \,\middle|\, Z \right] = \mu(\hat{F}_N^1(Z)). \tag{41}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} \mathbf{y}_j^*(1) \mid Z \right] &= \mathbb{E}\left[\frac{1}{N} \sum_{j\,:\,Z_j=0} \mathbf{y}_j(1) \mid Z \right] + \mathbb{E}\left[\frac{1}{N} \sum_{j\,:\,Z_j=0} \mathbf{y}_j^*(1) \right] \\
&= \frac{n_1}{N} \mu(\hat{F}_N^1(Z)) + \frac{n_0}{N} \mu(\hat{F}_N^1(Z)) \\
&= \mu(\hat{F}_N^1(Z)).
\end{aligned} \tag{42}$$

The same reasoning implies that $\mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} \left(\mathbf{y}_j^*(1) \right)^2 \mid Z \right]$ agrees with the second moment of $\hat{F}_N^1(Z)$. We are left with the following result.

Lemma A.14. *Under the sampling procedure of Algorithm 3,*

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0) \mid Z\right] = \mu(\hat{F}_N^0(Z)) = \frac{1}{n_0}\sum_{j\,:\,Z_j=0}\mathbf{y}_j(0),$$

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1) \mid Z\right] = \mu(\hat{F}_N^1(Z)) = \frac{1}{n_1}\sum_{j\,:\,Z_j=1}\mathbf{y}_j(1),$$

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)^2 \mid Z\right] = \frac{n_0}{n_0-1}\mathbb{V}\left(\hat{F}_N^0(Z)\right)$$

$$= \frac{1}{n_0-1}\sum_{i\,:\,Z_i=0}\left(\mathbf{y}_i(0) - \frac{1}{n_0}\sum_{j\,:\,Z_j=0}\mathbf{y}_j(0)\right)^2,$$

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right)^2 \mid Z\right] = \frac{n_1}{n_1-1}\mathbb{V}\left(\hat{F}_N^1(Z)\right)$$

$$= \frac{1}{n_1-1}\sum_{i\,:\,Z_i=1}\left(\mathbf{y}_i(1) - \frac{1}{n_1}\sum_{j\,:\,Z_j=1}\mathbf{y}_j(1)\right)^2.$$

Next we look at the "bootstrap treatment effects" $\mathbf{y}_i^*(1) - \mathbf{y}_i^*(0)$, these are of the form

$$\begin{aligned}\mathbf{y}_i'(1) - \mathbf{y}_i(0) \quad &\textit{if } Z_i = 0\\ \mathbf{y}_i(1) - \mathbf{y}_i'(0) \quad &\textit{if } Z_i = 1.\end{aligned}$$

By linearity

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(1) - \mathbf{y}_i^*(0)\right) \mid Z\right] = \mu(\hat{F}_N^1(Z)) - \mu(\hat{F}_N^0(Z)),$$

so centering the bootstrap distribution by the observed difference in means under Z enforces Neyman's weak null in the bootstrap population.

We now compute the expected variance (conditional upon Z) of the bootstrap treatment

effects

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left((\mathbf{y}_i^*(1)-\mathbf{y}_i^*(0))-\frac{1}{N}\sum_{j=1}^{N}(\mathbf{y}_i^*(1)-\mathbf{y}_i^*(0))\right)^2\right].$$

Rewriting the inner sample variance using the identity that $\mathbb{V}(A-B)=\mathbb{V}(A)+\mathbb{V}(B)-2\text{cov}(A,B)$ yields

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(1)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right)^2 \mid Z\right]+ \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)^2 \mid Z\right]- 2\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)\left(\mathbf{y}_i^*(1)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right) \mid Z\right]$$

The first two terms we already know how to handle thanks to Lemma A.14, this reduces it to

$$\frac{n_1}{n_1-1}\mathbb{V}\left(\hat{F}_N^1(Z)\right)+\frac{n_0}{n_0-1}\mathbb{V}\left(\hat{F}_N^0(Z)\right)- 2\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)\left(\mathbf{y}_i^*(1)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right) \mid Z\right]$$

We rearrange the final term via the identity $\text{cov}(A,B)=\mathbb{E}[AB]-\mathbb{E}[A]\mathbb{E}[B]$

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)\left(\mathbf{y}_i^*(1)-\frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right) \mid Z\right]= \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right]-\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0) \mid Z\right]\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(1) \mid Z\right]$$

Applying Lemma A.14 again yields

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right] - \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0) \mid Z\right]\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(1) \mid Z\right] = \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right] - \mu(\hat{F}_N^0(Z))\mu(\hat{F}_N^1(Z)).$$

Finally, we examine $\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right]$ by breaking up the sum over the observed treated and control indices

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right] &= \mathbb{E}\left[\frac{1}{N-1}\sum_{i\,:\,Z_i=0}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right] + \\
&\qquad \mathbb{E}\left[\frac{1}{N-1}\sum_{i\,:\,Z_i=1}\mathbf{y}_i^*(0)\mathbf{y}_i^*(1) \mid Z\right] \\
&= \mathbb{E}\left[\frac{1}{N-1}\sum_{i\,:\,Z_i=0}\mathbf{y}_i(0)\mathbf{y}_i^*(1) \mid Z\right] + \\
&\qquad \mathbb{E}\left[\frac{1}{N-1}\sum_{i\,:\,Z_i=1}\mathbf{y}_i^*(0)\mathbf{y}_i(1) \mid Z\right] \\
&= \frac{n_0}{N-1}\mathbb{E}\left[\frac{1}{n_0}\sum_{i\,:\,Z_i=0}\mathbf{y}_i(0)\mathbf{y}_i^*(1) \mid Z\right] + \\
&\qquad \frac{n_1}{N-1}\mathbb{E}\left[\frac{1}{n_1}\sum_{i\,:\,Z_i=1}\mathbf{y}_i^*(0)\mathbf{y}_i(1) \mid Z\right].
\end{aligned}$$

We examine the first term using the same intuition that was employed in the analysis of $\mathbb{E}\left[\frac{1}{n_0}\sum_{i\,:\,Z_i=0}\mathbf{y}_i^*(1) \mid Z\right]$. Consider the following two-step procedure:

1. Uniformly at random select an index from $\{j \;:\; Z_j = 0\}$; call this index j^*
2. Let $y = y_{j^*}(0)y_{j^*}^*(1)$.

By the tower property of conditional expectations, the uniformity of j^* over $\{j \;:\; Z_j = 0\}$,

independence, and linearity

$$\begin{aligned}
\mathbb{E}\left[y \mid Z\right] &= \mathbb{E}\left[\mathbb{E}\left[y_{j^*}(0)y^*_{j^*}(1) \mid j^*, Z\right] \mid Z\right] \\
&= \sum_{j\,:\,Z_j=0} \mathbb{P}\left(j^* = j \mid Z\right) \mathbb{E}\left[y_{j^*}(0)y^*_{j^*}(1) \mid j^* = j, Z\right] \\
&= \sum_{j\,:\,Z_j=0} \mathbb{P}\left(j^* = j \mid Z\right) \mathbb{E}\left[y_{j^*}(0)y^*_{j^*}(1) \mid Z\right] \\
&= \sum_{j\,:\,Z_j=0} \frac{1}{n_0} \mathbb{E}\left[\mathbf{y}_j(0)\mathbf{y}^*_j(1) \mid Z\right] \\
&= \mathbb{E}\left[n_0^{-1} \sum_{j\,:\,Z_j=0} \mathbf{y}_j(0)\mathbf{y}^*_j(1) \,\middle|\, Z \right]. \qquad (43)
\end{aligned}$$

Again, since the two-step procedure above generates a single draw from γ_{opt} and then sets y to be the product of the pair, $\mathbb{E}\left[y \mid Z\right]$ equals $\mathbb{E}\left[AB \mid Z\right]$ for $(A, B) \sim \gamma_{opt}$. Analogous reasoning handles $\mathbb{E}\left[\frac{1}{n_1}\sum_{i\,:\,Z_i=1} \mathbf{y}^*_i(0)\mathbf{y}_i(1) \mid Z\right]$. In total this leaves us with

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}^*_i(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}^*_j(0)\right)\left(\mathbf{y}^*_i(1) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}^*_j(1)\right) \mid Z\right] = \frac{N}{N-1}\mathrm{cov}\left(A, B\right)$$

for $(A, B) \sim \gamma_{opt}$. By the optimality of γ_{opt} and Lemma 2 it follows that

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}^*_i(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}^*_j(0)\right)\left(\mathbf{y}^*_i(1) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}^*_j(1)\right) \mid Z\right]$$

coincides with $\frac{N}{N-1}\hat{\sigma}^H_N$ of (AGL14). We record this result in the following lemma.

Lemma A.15. *Under the sampling procedure of Algorithm 4,*

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left((\mathbf{y}_i^*(1)-\mathbf{y}_i^*(0))-\frac{1}{N}\sum_{j=1}^{N}(\mathbf{y}_i^*(1)-\mathbf{y}_i^*(0))\right)^2 \mid Z\right] = \frac{n_1}{n_1-1}\mathbb{V}\left(\hat{F}_N^1(Z)\right)+\frac{n_0}{n_0-1}\mathbb{V}\left(\hat{F}_N^0(Z)\right)-2\frac{N}{N-1}\hat{\sigma}_N^H.$$

4.20.2 The Bootstrap Distribution Conditional Variance

Overall we are interested in $\mathbb{V}\left(T^*(Z,B) \mid Z\right)$, the variance of the bootstrap distribution (conditional on Z) constructed in Algorithm 4. By the law of total variance

$$\mathbb{V}\left(T^*(Z,B) \mid Z\right) = \mathbb{E}\left[\mathbb{V}\left(T^*(Z,B) \mid Z, y^*\right) \mid Z\right] + \mathbb{V}\left(\mathbb{E}\left[T^*(Z,B) \mid y^*, Z\right] \mid Z\right),$$

where conditioning on y^* is shorthand for conditioning upon the bootstrap population $\{\mathbf{y}_i^*(0), \mathbf{y}_i^*(1)\}_{i=1}^N$ produced by Algorithm 3.

We approach the first term on the right. The conditional variance $\mathbb{V}\left(T^*(Z,B) \mid Z, y^*\right)$ is given by the usual Neyman variance formula since the "potential outcomes" $\{\mathbf{y}_i^*(0), \mathbf{y}_i^*(1)\}_{i=1}^N$

are fixed by the conditioning event.[4] Consequently,

$$\begin{aligned}
\mathbb{V}\left(T^*(Z,B) \mid Z, y^*\right) &= \frac{\hat{\Sigma}_{y^*(1)}}{n_1} + \frac{\hat{\Sigma}_{y^*(0)}}{n_0} - \frac{\hat{\Sigma}_{y^*(1)-y^*(0)}}{N}, \\
\hat{\Sigma}_{y^*(0)} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(0) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(0)\right)^2, \\
\hat{\Sigma}_{y^*(1)} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^*(1)\right)^2, \\
\hat{\Sigma}_{y^*(1)-y^*(0)} &= \frac{1}{N-1}\sum_{i=1}^{N}\left((\mathbf{y}_i^*(1) - \mathbf{y}_i^*(0)) - \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{y}_i^*(1) - \mathbf{y}_j^*(0)\right)\right)^2.
\end{aligned}$$

By Lemmas A.14 and A.15 it follows that

$$\mathbb{E}\left[\mathbb{V}\left(T^*(Z,B) \mid Z, y^*\right) \mid Z\right] = \frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)}{n_1 - 1} + \frac{\mathbb{V}\left(\hat{F}_N^0(Z)\right)}{n_0 - 1} - \frac{\frac{n_1}{n_1-1}\mathbb{V}\left(\hat{F}_N^1(Z)\right) + \frac{n_0}{n_0-1}\mathbb{V}\left(\hat{F}_N^0(Z)\right) - 2\frac{N}{N-1}\hat{\sigma}_N^H}{N}. \quad (44)$$

Now we turn to analysing the second term in the decomposition of $\mathbb{V}\left(T^*(Z,B) \mid Z\right)$. First, since $\mathbb{E}\left[B_i\right] = n_1/N$, by linearity it follows that

$$\mathbb{E}\left[T^*(Z,B) \mid y^*, Z\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^*(0) - \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^*(0)\right) = 0.$$

Of course, the variance of a constant is zero, so the term $\mathbb{V}\left(\mathbb{E}\left[T^*(Z,B) \mid y^*, Z\right] \mid Z\right) = 0$. This leaves us with:

Lemma A.16. *The variance of the bootstrap distribution under Algorithm 4 (conditional*

[4]The conditional variance $\mathbb{V}\left(T^*(Z,B) \mid Z, y^*\right)$ is not impacted by the bootstrap centering term $\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^*(1) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^*(0)$ since this term is deterministic after conditioning upon Z and y^*.

upon Z) is

$$\begin{aligned}\mathbb{V}\left(T^*(Z,B) \mid Z\right) &= \frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)}{n_1-1} + \frac{\mathbb{V}\left(\hat{F}_N^0(Z)\right)}{n_0-1} - \\ &\qquad \frac{\frac{n_1}{n_1-1}\mathbb{V}\left(\hat{F}_N^1(Z)\right) + \frac{n_0}{n_0-1}\mathbb{V}\left(\hat{F}_N^0(Z)\right) - 2\frac{N}{N-1}\hat{\sigma}_N^H}{N} \\ &= \frac{1}{N}\left(\frac{(N-n_1)}{n_1-1}\mathbb{V}\left(\hat{F}_N^1(Z)\right) + \frac{(N-n_0)}{n_0-1}\mathbb{V}\left(\hat{F}_N^0(Z)\right) + \right. \\ &\qquad \left. 2\frac{N}{N-1}\hat{\sigma}_N^H\right).\end{aligned}$$

Recall from (AGL14, Equation 9) that the sharp variance estimator $\hat{V}_N^H$ is defined as

$$\begin{aligned}\hat{V}_N^H &= \frac{1}{N-1}\left(\frac{(N-n_1)}{n_1}\hat{\sigma}^2(1) + \frac{(N-n_0)}{n_0}\hat{\sigma}^2(0) + 2\hat{\sigma}_N^H\right), \\ \hat{\sigma}^2(0) &= \frac{N-1}{N(n_0-1)}\sum_{i\,:\,Z_i=0}\left(\mathbf{y}_i(0) - \frac{1}{n_0}\sum_{j\,:\,Z_j=0}\mathbf{y}_j(0)\right)^2 = \frac{N-1}{N}\mathbb{V}\left(\hat{F}_N^0(Z)\right), \\ \hat{\sigma}^2(1) &= \frac{N-1}{N(n_1-1)}\sum_{i\,:\,Z_i=1}\left(\mathbf{y}_i(1) - \frac{1}{n_1}\sum_{j\,:\,Z_j=1}\mathbf{y}_j(1)\right)^2 = \frac{N-1}{N}\mathbb{V}\left(\hat{F}_N^1(Z)\right).\end{aligned}$$

So as long as *plim* $\hat{\sigma}_N^H$ exists as some finite constant Lemma A.16 shows that

$$\left|N(\hat{V}_N^H - \mathbb{V}\left(T^*(Z,B) \mid Z\right))\right| = o_P(1).$$

4.20.3 The Bootstrap Distribution Conditional Mean

Next, we are interested in $\mathbb{E}\left[T^*(Z,B) \mid Z\right]$, the mean of the bootstrap distribution (conditional on Z) constructed in Algorithm 4. By the tower property of conditional expectation

$$\mathbb{E}\left[T^*(Z,B) \mid Z\right] = \mathbb{E}\left[\mathbb{E}\left[T^*(Z,B) \mid y^*, Z\right] \mid Z\right]$$

As noted earlier, since $\mathbb{E}[B_i] = n_1/N$, by linearity it follows that

$$\mathbb{E}[T^*(Z,B) \mid y^*, Z] = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^*(1) - \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^*(0) - \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^*(1) - \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^*(0)\right) = 0,$$

and so $\mathbb{E}[T^*(Z,B) \mid Z] = 0$. Thus, the bootstrap distribution is indeed centered.

4.20.4 Asymptotic Normality of the Conditional Bootstrap Distribution

Given that we now have an understanding of the bootstrap distribution's conditional mean and variance, all that is left is to show that the bootstrap distribution is indeed asymptotically Gaussian. First, we start by examining the bootstrap distribution conditional on **both** Z and y^* so that the only randomness comes in the form of the bootstrap treatment allocations B.

Let $\mathcal{L}(\sqrt{N}T^*(Z,B) \mid y^*, Z)$ be the conditional law of $\sqrt{N}T^*(Z,B)$ given both Z and y^*. This is simply the distribution of

$$\sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^{N} B_i \mathbf{y}_i^*(1) - \frac{1}{n_0}\sum_{i=1}^{N}(1-B_i)\mathbf{y}_i^*(0) - \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^*(1) - \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^*(0)\right)\right)$$

where the $\mathbf{y}_i^*(0)$s and $\mathbf{y}_i^*(1)$s are fixed.

Lemma A.17. *Under the conditions on the potential outcomes and experimental design assumed by Proposition 1 of (AGL14), $\mathcal{L}(\sqrt{N}T^*(Z,B) \mid y^*, Z)$ converges in weakly to the*

centered Gaussian with variance

$$plim_{N\to\infty} N\left(\frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)}{n_1}+\frac{\mathbb{V}\left(\hat{F}_N^0(Z)\right)}{n_0}-\frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)+\mathbb{V}\left(\hat{F}_N^0(Z)\right)-2\hat{\sigma}_N^H}{N}\right)$$

almost surely with respect to randomness in y^* and Z.

Proof. Define the *variance functional* for a random variable A as the map $\mathcal{L}(A)\mapsto\mathbb{V}(A)$; in a slight abuse of notation, we denote this functional directly as $\mathbb{V}(\mathcal{L}(A))$.

Under the conditions on the potential outcomes and experimental design assumed by Proposition 1of (AGL14), $plim N\hat{V}_N^H$ is a finite strictly positive constant, call this constant Σ_∞. Since $\left|N(\hat{V}_N^H-\mathbb{V}(T^*(Z,B)\mid Z))\right|=o_P(1)$ it follows that $\mathbb{V}\left(\mathcal{L}(\sqrt{N}T^*(Z,B)\mid Z)\right)\xrightarrow{p}\Sigma_\infty$. This establishes that the quantity

$$plim_{N\to\infty} N\left(\frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)}{n_1}+\frac{\mathbb{V}\left(\hat{F}_N^0(Z)\right)}{n_0}-\frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)+\mathbb{V}\left(\hat{F}_N^0(Z)\right)-2\hat{\sigma}_N^H}{N}\right). \quad (45)$$

is a well-defined positive constant under the conditions of the lemma. By a suitable strong law of large numbers the conditional variance $\mathbb{V}\left(\mathcal{L}(\sqrt{N}T^*(Z,B)\mid y^*,Z)\right)$ limits almost surely to (45). Furthermore, as mentioned earlier, the conditional mean of $T^*(Z,B)$ given y^* and Z is exactly zero. Consequently, the mean and variance (conditional upon y^* and Z) of the bootstrap distribution scaled by $\sqrt{N}$ match those of the desired limiting distribution.

Since the distribution of interest is the conditional distribution of

$$\sqrt{N}\left(\frac{1}{n_1}\sum_{i=1}^N B_i\mathbf{y}_i^*(1)-\frac{1}{n_0}\sum_{i=1}^N(1-B_i)\mathbf{y}_i^*(0)-\left(\frac{1}{N}\sum_{i=1}^N\mathbf{y}_i^*(1)-\frac{1}{N}\sum_{i=1}^N\mathbf{y}_i^*(0)\right)\right)$$

where the $\mathbf{y}_i^*(0)$s and $\mathbf{y}_i^*(1)$s are fixed it suffices to show that the conditions of a finite population central limit theorem are satisfied for $\{(\mathbf{y}_i^*(0),\mathbf{y}_i^*(1))\}_{i=1}^N$ with probability tending to one as $N\to\infty$. The first two finite population moments of $\{(\mathbf{y}_i^*(0),\mathbf{y}_i^*(1))\}_{i=1}^N$ (conditional

on y^*) converge almost surely to the original finite population moments, and so the conditions of the finite population central limit theorem (LD17) are ensured asymptotically almost surely. □

Lemma A.17 demonstrates that the sampling procedure of Algorithm 4 produces the desired distribution, but the result of Lemma A.17 conditions on both the observed treatment allocation Z and the optimal-transport-based samples $y_*(0)$ and $y_*(1)$. For the analysis of Algorithm 4 the object of interest is the bootstrap distribution conditioned upon the observed treatment allocation Z, but not conditioned upon the samples $y_*(0)$ and $y_*(1)$. We use Lemma A.17 and a familiar deconditioning argument to establish the desired result, presented below.

Theorem A.29. *Under the conditions on the potential outcomes and experimental design assumed by Lemma A.17, $\mathcal{L}(\sqrt{N}T^*(Z,B) \mid Z)$ converges in weakly to the centered Gaussian with variance*

$$plim_{N\to\infty} N\left(\frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)}{n_1} + \frac{\mathbb{V}\left(\hat{F}_N^0(Z)\right)}{n_0} - \frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right) + \mathbb{V}\left(\hat{F}_N^0(Z)\right) - 2\hat{\sigma}_N^H}{N}\right)$$

almost surely with respect to randomness in Z.

Proof. By Lemma A.17 $\mathcal{L}(\sqrt{N}T^*(Z,B) \mid y^*, Z)$ converges weakly to the Gaussian distribution with mean zero and variance

$$plim_{N\to\infty} N\left(\frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right)}{n_1} + \frac{\mathbb{V}\left(\hat{F}_N^0(Z)\right)}{n_0} - \frac{\mathbb{V}\left(\hat{F}_N^1(Z)\right) + \mathbb{V}\left(\hat{F}_N^0(Z)\right) - 2\hat{\sigma}_N^H}{N}\right).$$

By the forward implication of Lemma 4.1 in (DDCZ13) this implies that $\mathcal{L}(\sqrt{N}T^*(Z,B) \mid Z)$ converges weakly to the Gaussian distribution with mean zero and same variance. We remark that this deconditioning argument has been applied fruitfully elsewhere, e.g., (CF22,

Appendix Lemma D) .

□

4.21 Analyzing The Procedure Of Imbens & Menzel For Binary Outcomes

4.21.1 Set-up

Suppose we are interested in binary potential outcomes and we take a treatment allocation $Z \sim Unif(\Omega_{CRE})$ where $\Omega_{CRE} = \{z \in \{0,1\}^N \ : \ \sum_{i=1}^N z_i = n_1\}$; say $n_0 = N - n_1$ and $n_1/N \to p \in (0,1)$. Define the random variables

$$S_0 = \sum_{i \,:\, Z_i = 0} \mathbf{y}_i(0)$$
$$S_1 = \sum_{i \,:\, Z_i = 1} \mathbf{y}_i(1);$$

these just count the number of observed 1's in the control and treated groups, respectively.

Define the empirical cumulative distribution functions of the observed treated and control groups as

$$\hat{F}(t) = \frac{1}{n_0} \sum_{i \,:\, Z_i = 0} \mathbb{1}_{\{\mathbf{y}_i(0) \leq t\}} = \begin{cases} 0 & \text{if } t < 0, \\ \frac{n_0 - S_0}{n_0} & \textit{if } t \in [0,1), \\ 1 & \text{if } t \geq 1. \end{cases}$$

$$\hat{G}(t) = \frac{1}{n_1} \sum_{i \,:\, Z_i = 1} \mathbb{1}_{\{\mathbf{y}_i(1) \leq t\}} = \begin{cases} 0 & \text{if } t < 0, \\ \frac{n_1 - S_1}{n_1} & \textit{if } t \in [0,1), \\ 1 & \text{if } t \geq 1. \end{cases}$$

The corresponding empirical quantile functions are given by

$$\hat{F}^{-1}(p) = \inf\{x \in \mathbb{R} \ : \ \hat{F}(x) \geq p\} = \begin{cases} 0 & \text{if } p \leq \frac{n_0 - S_0}{n_0}, \\ 1 & \text{if } p > \frac{n_0 - S_0}{n_0}. \end{cases}$$

$$\hat{G}^{-1}(p) = \inf\{x \in \mathbb{R} \ : \ \hat{G}(x) \geq p\} = \begin{cases} 0 & \text{if } p \leq \frac{n_1 - S_1}{n_1}, \\ 1 & \text{if } p > \frac{n_1 - S_1}{n_1}. \end{cases}$$

Notice the following:

$$\hat{F}^{-1}\left(\hat{G}(0)\right) = \hat{F}^{-1}\left(\frac{n_1 - S_1}{n_1}\right) = \begin{cases} 0 & \text{if } \frac{n_1 - S_1}{n_1} \leq \frac{n_0 - S_0}{n_0}, \\ 1 & \text{otherwise} \end{cases}$$

$$\hat{G}^{-1}\left(\hat{F}(0)\right) = \hat{G}^{-1}\left(\frac{n_0 - S_0}{n_0}\right) = \begin{cases} 0 & \text{if } \frac{n_1 - S_1}{n_1} \geq \frac{n_0 - S_0}{n_0}, \\ 1 & \text{otherwise} \end{cases}$$

$$\hat{F}^{-1}\left(\hat{G}(1)\right) = \hat{F}^{-1}(1) = 1,$$

$$\hat{G}^{-1}\left(\hat{F}(1)\right) = \hat{G}^{-1}(1) = 1.$$

4.21.2 Variance Analysis

The imputation scheme of (IM21, Equation 3.4) is as follows

$$\mathbf{y}_i^*(0) = \begin{cases} \mathbf{y}_i(0) & \text{if } Z_i = 0 \\ \hat{F}^{-1}\left(\hat{G}(\mathbf{y}_i(1))\right) & \text{otherwise} \end{cases}$$

$$\mathbf{y}_i^*(1) = \begin{cases} \mathbf{y}_i(1) & \text{if } Z_i = 1 \\ \hat{G}^{-1}\left(\hat{F}(\mathbf{y}_i(0))\right) & \text{otherwise} \end{cases}$$

This results in the following three schemes. When $\frac{n_0-S_0}{n_0} < \frac{n_1-S_1}{n_1}$

$$\begin{cases} \textit{Observed treated} \begin{cases} 0 \xrightarrow{\text{imputes to}} 1 \\ 1 \xrightarrow{\text{imputes to}} 1 \end{cases} \\ \textit{Observed control} \begin{cases} 0 \xrightarrow{\text{imputes to}} 0 \\ 1 \xrightarrow{\text{imputes to}} 1. \end{cases} \end{cases}$$

When $\frac{n_0-S_0}{n_0} > \frac{n_1-S_1}{n_1}$

$$\begin{cases} \textit{Observed treated} \begin{cases} 0 \xrightarrow{\text{imputes to}} 0 \\ 1 \xrightarrow{\text{imputes to}} 1 \end{cases} \\ \textit{Observed control} \begin{cases} 0 \xrightarrow{\text{imputes to}} 1 \\ 1 \xrightarrow{\text{imputes to}} 1. \end{cases} \end{cases}$$

When $\frac{n_0-S_0}{n_0} = \frac{n_1-S_1}{n_1}$

$$\begin{cases} \textit{Observed treated} \begin{cases} 0 \xrightarrow{\text{imputes to}} 0 \\ 1 \xrightarrow{\text{imputes to}} 1 \end{cases} \\ \textit{Observed control} \begin{cases} 0 \xrightarrow{\text{imputes to}} 0 \\ 1 \xrightarrow{\text{imputes to}} 1. \end{cases} \end{cases}$$

Consequently, when $\frac{n_0-S_0}{n_0} < \frac{n_1-S_1}{n_1}$ the imputed population $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1))\}_{i=1}^N$:

- has treated outcomes with $S_1 + S_0$ ones,
- has control outcomes with $S_0 + n_1$ ones,

- has treatment effects which are
  - $+1$ for 0 individuals
  - -1 for $n_1 - S_1$ individuals,
  - 0 for $S_1 + S_0 + (n_0 - S_0) = S_1 + n_0$ individuals.

The variance of the difference in means for the imputed population (conditional upon $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1))\}_{i=1}^N$) is

$$\begin{aligned}
\hat{V}_< &= \frac{\Sigma_1^*}{n_1} + \frac{\Sigma_0^*}{n_0} - \frac{\Sigma_\tau^*}{N} \\
\Sigma_1^* &= \left(\frac{S_1 + S_0}{N}\right)\left(1 - \frac{S_1 + S_0}{N}\right) \\
\Sigma_0^* &= \left(\frac{S_0 + n_1}{N}\right)\left(1 - \frac{S_0 + n_1}{N}\right) \\
\Sigma_\tau^* &= \left(\frac{n_1 - S_1}{N}\right)\left(1 - \frac{n_1 - S_1}{N}\right).
\end{aligned}$$

Similar reasoning applies when $\frac{n_0 - S_0}{n_0} > \frac{n_1 - S_1}{n_1}$; in this case the imputed population $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1))\}_{i=1}^N$:

- has treated outcomes with $S_1 + n_0$ ones,
- has control outcomes with $S_0 + S_1$ ones,
- has treatment effects which are
  - $+1$ for $n_0 - S_0$ individuals
  - -1 for 0 individuals,
  - 0 for $S_1 + S_0 + (n_1 - S_1) = S_0 + n_1$ individuals.

The variance of the difference in means for the imputed population (conditional upon $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1))\}_{i=1}^N$) is

$$\begin{aligned}
\hat{V}_> &= \frac{\Sigma_1^*}{n_1} + \frac{\Sigma_0^*}{n_0} - \frac{\Sigma_\tau^*}{N} \\
\Sigma_1^* &= \left(\frac{S_1 + n_0}{N}\right)\left(1 - \frac{S_1 + n_0}{N}\right) \\
\Sigma_0^* &= \left(\frac{S_0 + S_1}{N}\right)\left(1 - \frac{S_0 + S_1}{N}\right) \\
\Sigma_\tau^* &= \left(\frac{n_0 - S_0}{N}\right)\left(1 - \frac{n_0 - S_0}{N}\right).
\end{aligned}$$

Now suppose that $\frac{n_0 - S_0}{n_0} = \frac{n_1 - S_1}{n_1}$; in this case the imputed population $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1))\}_{i=1}^N$:

- has treated outcomes with S_1 ones,
- has control outcomes with S_0 ones,
- has treatment effects which are
  - $+1$ for 0 individuals
  - -1 for 0 individuals,
  - 0 for all N individuals.

The variance of the difference in means for the imputed population (conditional upon $\{(\mathbf{y}_i^*(0), \mathbf{y}_i^*(1))\}_{i=1}^N$) is

$$\begin{aligned}
\hat{V}_= &= \frac{\Sigma_1^*}{n_1} + \frac{\Sigma_0^*}{n_0} \\
\Sigma_1^* &= \left(\frac{S_1}{N}\right)\left(1 - \frac{S_1}{N}\right) \\
\Sigma_0^* &= \left(\frac{S_0}{N}\right)\left(1 - \frac{S_0}{N}\right)
\end{aligned}$$

Remark 14. In the case that $\frac{n_0 - S_0}{n_0} = \frac{n_1 - S_1}{n_1}$ the Imbens & Menzel procedure imputes counterfactuals exactly as one would if one simply assumed that Fisher's sharp null held. In the case of binary outcomes, this imputation coincides with the construction of the comonotone coupling; see Claim 1. Consequently, $\hat{V}_=$ can be immediately analyzed by porting over the analysis of $\hat{V}_N^H$ from (AGL14).

Notice that by the law of large numbers

$$S_0/N = (n_0/N)\frac{1}{n_0}\sum_{i\,:\,Z_i=0} \mathbf{y}_i(0) \xrightarrow{p} (1-p)\overline{y}_\infty(0)$$
$$S_1/N = (n_1/N)\frac{1}{n_1}\sum_{i\,:\,Z_i=1} \mathbf{y}_i(1) \xrightarrow{p} p\overline{y}_\infty(1).$$

Consequently,

$$N\hat{V}_< \xrightarrow{p} \frac{((1-p)\overline{y}_\infty(0) + p\overline{y}_\infty(1))\,(1-(1-p)\overline{y}_\infty(0) - p\overline{y}_\infty(1))}{p} + \frac{((1-p)\overline{y}_\infty(0) + p)\,(1-(1-p)\overline{y}_\infty(0) - p)}{1-p} - (p - p\overline{y}_\infty(1))\,(1-p+p\overline{y}_\infty(1))\,. \quad (46)$$

and

$$N\hat{V}_> \xrightarrow{p} \frac{(p\overline{y}_\infty(1) + (1-p))\,(1-p\overline{y}_\infty(1) - (1-p))}{p} + \frac{((1-p)\overline{y}_\infty(0) + p\overline{y}_\infty(1))\,(1-(1-p)\overline{y}_\infty(0) - p\overline{y}_\infty(1))}{1-p} - ((1-p) - (1-p)\overline{y}_\infty(0))\,(1-(1-p)+(1-p)\overline{y}_\infty(0))\,. \quad (47)$$

In total, we have established the following lemma.

Lemma A.18. *The conditional bootstrap variance of the Imbens & Menzel procedure for the*

$\sqrt{N}$*-scaled difference in means is*

$$\begin{aligned} N\hat{V}_{<} \quad & \textit{if} \quad \frac{n_0 - S_0}{n_0} < \frac{n_1 - S_1}{n_1}, \\ N\hat{V}_{=} \quad & \textit{if} \quad \frac{n_0 - S_0}{n_0} = \frac{n_1 - S_1}{n_1}, \\ N\hat{V}_{>} \quad & \textit{if} \quad \frac{n_0 - S_0}{n_0} > \frac{n_1 - S_1}{n_1}. \end{aligned}$$

Unwinding notation yields that

$$\begin{aligned} \frac{n_0 - S_0}{n_0} - \frac{n_1 - S_1}{n_1} &= 1 - \frac{1}{n_0} \sum_{i \,:\, Z_i = 0} \mathbf{y}_i(0) - 1 + \frac{1}{n_1} \sum_{i \,:\, Z_i = 1} \mathbf{y}_i(1) \\ &= \underbrace{\frac{1}{n_1} \sum_{i \,:\, Z_i = 1} \mathbf{y}_i(1) - \frac{1}{n_0} \sum_{i \,:\, Z_i = 0} \mathbf{y}_i(0)}_{\text{Difference in Means}}. \end{aligned}$$

So,

$$\begin{aligned} \mathbb{P}\left(\frac{n_0 - S_0}{n_0} - \frac{n_1 - S_1}{n_1} > 0 \right) &= \mathbb{P}\left(\frac{1}{n_1} \sum_{i \,:\, Z_i = 1} \mathbf{y}_i(1) - \frac{1}{n_0} \sum_{i \,:\, Z_i = 0} \mathbf{y}_i(0) > 0 \right) \\ &= \mathbb{P}\left(\sqrt{N} \left(\frac{1}{n_1} \sum_{i \,:\, Z_i = 1} \mathbf{y}_i(1) - \frac{1}{n_0} \sum_{i \,:\, Z_i = 0} \mathbf{y}_i(0) \right) > 0 \right). \end{aligned}$$

Assuming that Neyman's null is true so that $\overline{\tau} = 0$ a finite population central limit theorem (LD17) implies that

$$\mathbb{P}\left(\sqrt{N} \left(\frac{1}{n_1} \sum_{i \,:\, Z_i = 1} \mathbf{y}_i(1) - \frac{1}{n_0} \sum_{i \,:\, Z_i = 0} \mathbf{y}_i(0) \right) > 0 \right) \to \Phi_{V_{Neyman}}(0) = \frac{1}{2}$$

where $\Phi_{V_{Neyman}}$ denotes the cumulative distribution function for the centered Gaussian with variance given by Neyman's variance for the $\sqrt{N}$-scaled difference in means. Analogous

reasoning implies that

$$\begin{aligned}
\mathbb{P}\left(\frac{n_0 - S_0}{n_0} < \frac{n_1 - S_1}{n_1}\right) &\to \frac{1}{2}, \\
\mathbb{P}\left(\frac{n_0 - S_0}{n_0} = \frac{n_1 - S_1}{n_1}\right) &\to 0, \\
\mathbb{P}\left(\frac{n_0 - S_0}{n_0} < \frac{n_1 - S_1}{n_1}\right) &\to \frac{1}{2}.
\end{aligned}$$

Putting this all together gives us the following informal result; we make the result precise in Theorem A.31.

Theorem A.30 (Informal)**.** *Suppose that the potential outcomes satisfy the requirements of (LD17) to ensure central limit behaviour; then asymptotically the conditional bootstrap variance of the Imbens & Menzel procedure for the $\sqrt{N}$-scaled difference in means is*

$$\begin{aligned}
&\textit{plim}(N\hat{V}_{>}) \quad \textit{with probability} \quad \frac{1}{2} \\
&\textit{plim}(N\hat{V}_{<}) \quad \textit{with probability} \quad \frac{1}{2},
\end{aligned}$$

where the values of $\textit{plim}(N\hat{V}_{<})$ and $\textit{plim}(N\hat{V}_{>})$ are given by (46) *and* (47)*, respectively.*

4.21.3 A Concrete Counterexample

Claim 1. For a finite populations whose potential outcomes take values in $\{0, 1\}$, satisfy Neymans's weak null for each N, and obey the regularity conditions of (AGL14, Proposition 1), the variance upper-bound V_N^H of (AGL14) is achieved by the permutation of the outcomes which enforces Fisher's sharp null.

Proof of Claim. Because the potential outcomes are binary, the distribution of the control units, defined as $F = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{y}_i(0)}$, puts mass on just two points $\{0, 1\}$. Consequently, F is totally determined by the number of ones in the control population: it puts mass

$\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i(0)$ on 1 and $1-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i(0)$ on 0. The same logic applies to the distribution of treated units $G=\frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{y}_i(1)}$ which puts mass $\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i(1)$ on 1 and $1-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i(1)$ on 0. Enforcing Neyman's null amounts to enforcing that $\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i(0)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i(1)$; so Neyman's null implies $F=G$ in the special case that the potential outcome are binary.

The variance upper-bound of V_N^H of (AGL14) relies upon the solution to

$$\max_{h\in\mathcal{H}}\operatorname{cov}(X,Y)$$

for $X\sim F$, $Y\sim G$, and $\mathcal{H}$ the set of joint measures with marginals F and G. This is known to be maximized at h equal to the comonotone coupling of F and G (Tch80). When $F=G$ and their support is discrete the comonotone coupling amounts to just ranking the support F as $y_{(1)}\leq\ldots,\leq y_{(N)}$ and then forming the joint measure $h_{comonotone}=\frac{1}{N}\sum_{i=1}^{N}\delta_{(y_{(1)},y_{(1)})}$. This matches the joint measure that one would get by permuting the potential outcomes to force Fisher's sharp null to hold. □

Remark 15. Interestingly, Claim 1 seems to have been implicitly known by Robins; see (Rob88, Page 774). Nonetheless, the presentation of Claim 1 in terms of the Frechét-Hoeffding copula bounds – specifically the comonotone copula – is novel.

Consider the following set of binary potential outcomes for $\theta\in(0,1)$:

$$\begin{aligned}(\mathbf{y}_i(0),\mathbf{y}_i(1))&=(1,1)\quad\text{for } i=1,\ldots,\lfloor\theta N\rfloor,\\(\mathbf{y}_i(0),\mathbf{y}_i(1))&=(0,0)\quad\text{for } i=\lfloor\theta N\rfloor+1,\ldots,N.\end{aligned}$$

Fisher's sharp null (and consequently, Neyman's weak null) holds for this population with $\overline{y}_\infty(0)=\theta=\overline{y}_\infty(1)$. Take $p\neq.5$, say 0.8, and $\theta=0.7$, then

$$\text{plim}N\hat{V}_<\approx 0.362\quad\&\quad\text{plim}N\hat{V}_>\approx 1.222.$$

Consequently, the limiting bootstrap conditional variance from the procedure of Imbens & Menzel is not constant; furthermore, both of these values are strictly smaller than the true variance of the $\sqrt{N}$-scaled difference in means (which is 1.313). Moverover, for binary outcomes satisfying Fisher's sharp null, by Claim 1 the upper bound of (AGL14) is consistent and matches that of (Rob88) under Fisher's sharp null; since there is no treatment effect heterogeneity, these two variance estimators asymptotically agree with the classical variance estimator of Neyman. As a result, the limiting bootstrap variances of Imbens & Menzel are both asymptotically anti-conservative with respect to the true variance as well as the three standard variance estimators from previous literature: Neyman's estimator and those of (AGL14) and (Rob88). Figure 4-3 plots $\text{plim}(N\hat{V}_{>})$ and $\text{plim}(N\hat{V}_{<})$ as functions of p and $\overline{y} := \overline{y}_{\infty}(0) = \overline{y}_{\infty}(1)$ as $(p, \overline{y})$ range over $(0,1)^2$.

4.21.4 Bootstrap Distribution Analysis

Theorem A.30 informally examines the limiting conditional bootstrap variance of the Imbens & Menzel procedure. We now examine the limiting conditional bootstrap distribution of the Imbens & Menzel procedure holistically.

For this analysis we will need to be more precise in our notation than is usually necessary, suppose that we work over a single probability space $(\tilde{\Omega}, \mathbb{P})$ which is constructed as the infinite product over individual probability spaces for each experiment of size N; formally $(\tilde{\Omega}, \mathbb{P}) = \bigotimes_{N \in \mathbb{N}} (\tilde{\Omega}^{(N)}, \mathbb{P}^{(N)})$. We superscript with (N) to refer to quantities pertaining to the

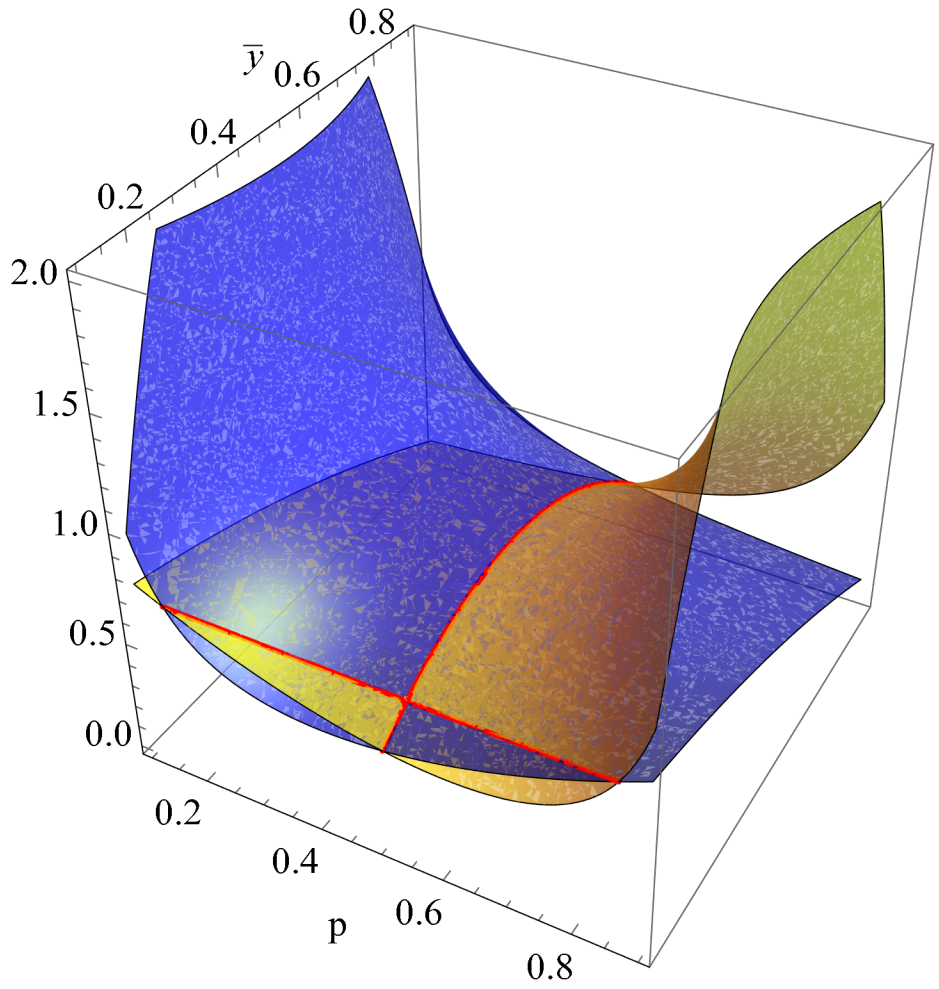

Figure 4-3: The two curves of $\text{plim}(N\hat{V}_>)$ and $\text{plim}(N\hat{V}_<)$ plotted as functions of the design limit p and the potential outcomes' common mean $\overline{y} := \overline{y}_\infty(0) = \overline{y}_\infty(1)$. The lines of intersection between the two curves are plotted to highlight when $\text{plim}(N\hat{V}_>) = \text{plim}(N\hat{V}_<)$. These intersections are the exception and not the rule; in general $\text{plim}(N\hat{V}_>) \neq \text{plim}(N\hat{V}_<)$.

finite population of size N. Define

$$\mathcal{A}_N = \left\{ \omega \in \tilde{\Omega} \ : \ \frac{n_0^{(N)} - S_0^{(N)}(\omega)}{n_0^{(N)}} < \frac{n_1^{(N)} - S_1^{(N)}(\omega)}{n_1^{(N)}} \right\},$$

$$\mathcal{B}_N = \left\{ \omega \in \tilde{\Omega} \ : \ \frac{n_0^{(N)} - S_0^{(N)}(\omega)}{n_0^{(N)}} > \frac{n_1^{(N)} - S_1^{(N)}(\omega)}{n_1^{(N)}} \right\},$$

$$\mathcal{C}_N = \left\{ \omega \in \tilde{\Omega} \ : \ \frac{n_0^{(N)} - S_0^{(N)}(\omega)}{n_{0406}^{(N)}} = \frac{n_1^{(N)} - S_1^{(N)}(\omega)}{n_1^{(N)}} \right\}.$$

For a given $\omega \in \tilde{\Omega}$ take

$$I_{\mathcal{A}}(\omega) = \{N \in \mathbb{N} \,:\, \omega \in \mathcal{A}_N\},$$
$$I_{\mathcal{B}}(\omega) = \{N \in \mathbb{N} \,:\, \omega \in \mathcal{B}_N\},$$
$$I_{\mathcal{C}}(\omega) = \{N \in \mathbb{N} \,:\, \omega \in \mathcal{C}_N\};$$

$I_{\mathcal{A}}(\omega)$ indexes the set of experiments for which $\frac{n_0^{(N)} - S_0^{(N)}(\omega)}{n_0^{(N)}} < \frac{n_1^{(N)} - S_1^{(N)}(\omega)}{n_1^{(N)}}$; $I_{\mathcal{B}}(\omega)$ and $I_{\mathcal{C}}(\omega)$ have analogous interpretations.

Lemma A.19. *For all $\omega \in \tilde{\Omega}$ except for a set of measure zero, $I_{\mathcal{A}}(\omega)$ and $I_{\mathcal{B}}(\omega)$ are of infinite cardinality.*

Proof. By the finite population central limit theorem argument detailed in Section 4.21.2, $\mathbb{P}(\mathcal{A}_N) \to 1/2$ and $\mathbb{P}(\mathcal{B}_N) \to 1/2$. Since $(\tilde{\Omega}, \mathbb{P}) = \bigotimes_{N \in \mathbb{N}} (\tilde{\Omega}^{(N)}, \mathbb{P}^{(N)})$, by the second Borel-Cantelli lemma we have that $\sum_{N \in \mathbb{N}} \mathbb{P}(\mathcal{A}_N) = \infty$ implies that $\mathbb{P}(\limsup \mathcal{A}_N) = 1$. Of course, since $\mathbb{P}(\mathcal{A}_N) \to 1/2$ the tail of $\sum_{N \in \mathbb{N}} \mathbb{P}(\mathcal{A}_N)$ is non-vanishing and so $\sum_{N \in \mathbb{N}} \mathbb{P}(\mathcal{A}_N) = \infty$. Thus, $\mathbb{P}(\limsup \mathcal{A}_N) = 1$; in other words, for all $\omega \in \tilde{\Omega}$ except for some set of measure zero we have that $\omega \in \mathcal{A}_N$ infinitely often. Consequently, $I_{\mathcal{A}}(\omega)$ is of infinite cardinality with probability one. The same logic applies to the analysis of $I_{\mathcal{B}}(\omega)$. □

By Lemma A.19, $I_{\mathcal{A}}(\omega)$ and $I_{\mathcal{B}}(\omega)$ are of infinite cardinality for all $\omega \in \tilde{\Omega}$ except for some set of measure zero. Fix an $\omega \in \tilde{\Omega}$ for which $|I_{\mathcal{A}}(\omega)| = \infty$ and $|I_{\mathcal{B}}(\omega)| = \infty$. Along the subsequence $I_{\mathcal{A}}(\omega) = \{N_1, N_2, \ldots\} \subseteq \mathbb{N}$ we have that the imputed population $\{(\mathbf{y}_i^*(0,\omega), \mathbf{y}_i^*(1,\omega))\}_{i=1}^N$:

- has treated outcomes with $S_1^{(N_j)}(\omega) + S_0^{(N_j)}(\omega)$ ones,
- has control outcomes with $S_0^{(N_j)}(\omega) + n_1^{(N_j)}$ ones,
- has treatment effects which are

- $+1$ for 0 individuals
- -1 for $n_1^{(N_j)} - S_1^{(N_j)}(\omega)$ individuals,
- 0 for $S_1^{(N_j)}(\omega) + S_0^{(N_j)}(\omega) + (n_0^{(N_j)} - S_0^{(N_j)}(\omega)) = S_1^{(N_j)} + n_0^{(N_j)}$ individuals.

The reasoning is discussed in Section 4.21.2. Since $|I_{\mathcal{A}}(\omega)| = \infty$ discussions of asymptotic quantities (e.g., strong laws, distributional limits, etc.) are well-formulated. By the strong law of large numbers for completely randomized experiments (WD21, Lemma A.3) $S_0/n_0 \xrightarrow{a.s.} \overline{y}_\infty(0)$ and $S_1/n_1 \xrightarrow{a.s.} \overline{y}_\infty(1)$, so the numeric sequence $S_z(\omega)/n_z \to \overline{y}_\infty(z)$ for all ω except some set of measure zero. Consequently, without loss of generality we assume that these limits hold for the ω under our consideration. Since the sequence $S_z(\omega)/n_z$ limits to $\overline{y}_\infty(z)$ it certainly holds that along the subsequence $I_{\mathcal{A}}(\omega)$

$$S_z^{(N_j)}(\omega)/n_z^{(N_j)} \to \overline{y}_\infty(z) \quad \text{for } z \in \{0,1\}.$$

Consequently, along the subsequence $I_{\mathcal{A}}(\omega)$ the conditions of a finite central limit theorem (e.g., (LD17, Theorem 5)) are satisfied for the sequence of imputed populations

$$\left\{ \{(\mathbf{y}_i^*(0,\omega), \mathbf{y}_i^*(1,\omega))\}_{i=1}^{N_j} \right\}_{N_j \in I_{\mathcal{A}}(\omega)}.$$

For $B^{(N_j)} \sim \mathit{Unif}(\Omega_{CRE})$ independent of all other random variables we have that the conditional bootstrap distribution of the $\sqrt{N}$-scaled difference in means under the Imbens & Menzel procedure is the distribution of

$$\tau^{*,(N_j)}(B,\omega) = \sqrt{N^{(N_j)}} \left(\frac{1}{n_1^{(N_j)}} \sum_{i\,:\,B_i^{(N_j)}=1} \mathbf{y}_i^*(1,\omega) - \frac{1}{n_0^{(N_j)}} \sum_{i\,:\,B_i^{(N_j)}=0} \mathbf{y}_i^*(0,\omega) \right) - \sqrt{N^{(N_j)}} \left(\frac{1}{N^{(N_j)}} \sum_{i=1}^{N_j} \mathbf{y}_i^*(1,\omega) - \mathbf{y}_i^*(0,\omega) \right).$$

By (LD17, Theorem 5), the strong law of large numbers (WD21, Lemma A.3), Slutsky's Lemma, and the analysis of Section 4.21.2 for all ω except for a set of measure zero $\tau^{*,(N_j)}(B^{(N_j)},\omega)$ converges in distribution to $\mathcal{N}\left(0, V_{<,\infty}\right)$ along the subsequence $N_j \in I_{\mathcal{A}}(\omega)$ where

$$V_{<,\infty} := \frac{\left((1-p)\overline{y}_\infty(0) + p\overline{y}_\infty(1)\right)\left(1-(1-p)\overline{y}_\infty(0) - p\overline{y}_\infty(1)\right)}{p} + \frac{\left((1-p)\overline{y}_\infty(0) + p\right)\left(1-(1-p)\overline{y}_\infty(0) - p\right)}{1-p} - \left(p - p\overline{y}_\infty(1)\right)\left(1-p+p\overline{y}_\infty(1)\right).$$

By analogous reasoning for all ω except for a set of measure zero $\tau^{*,(N_j)}(B^{(N_j)},\omega)$ converges in distribution to $\mathcal{N}\left(0, V_{>,\infty}\right)$ along the subsequence $N_j \in I_{\mathcal{B}}(\omega)$ where

$$V_{>,\infty} := \frac{\left(p\overline{y}_\infty(1) + (1-p)\right)\left(1-p\overline{y}_\infty(1) - (1-p)\right)}{p} + \frac{\left((1-p)\overline{y}_\infty(0) + p\overline{y}_\infty(1)\right)\left(1-(1-p)\overline{y}_\infty(0) - p\overline{y}_\infty(1)\right)}{1-p} - \left((1-p) - (1-p)\overline{y}_\infty(0)\right)\left(1-(1-p)+(1-p)\overline{y}_\infty(0)\right).$$

We collect our results thus far into the following lemma.

Lemma A.20. *Define the event $\mathcal{E}_\infty$ to be the set*

$$\left\{\omega \in \tilde{\Omega} \ : \ |I_{\mathcal{A}}| = \infty \text{ and } |I_{\mathcal{B}}| = \infty\right\}$$

for which

- *$\tau^{*,(N_j)}(B^{(N_j)},\omega)$ converges in distribution to $\mathcal{N}\left(0, V_{<,\infty}\right)$ along the subsequence $N_j \in I_{\mathcal{A}}(\omega)$,*
- *$\tau^{*,(N_j)}(B^{(N_j)},\omega)$ converges in distribution to $\mathcal{N}\left(0, V_{>,\infty}\right)$ along the subsequence $N_j \in$*

$I_{\mathcal{B}}(\omega)$.

The event $\mathcal{E}_\infty$ has probability one in the probability space $(\tilde{\Omega}, \mathbb{P}) = \bigotimes_{N\in\mathbb{N}}(\tilde{\Omega}^{(N)}, \mathbb{P}^{(N)})$.

Let the *bounded-Lipschitz* distance between two probability measures ν_X and ν_Y be defined as

$$\rho_{BL}(\nu_X, \nu_Y) := \sup_{\substack{f\in Lip_1 \\ ||f||_\infty \leq 1}} \left| \int f(x) d\nu_X(x) - \int f(y) d\nu_Y(y) \right|$$

where Lip_1 is the family of functions from $\mathbb{R}$ to $\mathbb{R}$ which Lipschitz continuous with Lipschitz constant at most one. Detailed discussion of ρ_{BL} can be found in (vdVW96, Section 1.12), most importantly, ρ_{BL} metrizes weak convergence of separable probability measures (vdVW96, Theorem 1.12.4).

Our interest is to characterize the convergence of the conditional bootstrap distribution produced by the Imbens & Menzel algorithm. We shall see that – in the binary potential outcomes case – the conditional bootstrap distribution generally fails to converge to a fixed distribution except for very special cases. As such, we leverage tools from the theory of random probability measures to characterize exactly the asymptotic behaviour of the conditional bootstrap distribution produced by the Imbens & Menzel algorithm. Random probability measures have been of substantial interest to probabilists; for two excellent references on the general theory of random measures we point to (Cra02, Kal17). The general theory of such measures requires care to account for topological considerations, these concerns are substantially alleviated since we are only interested in probability measures over $\mathbb{R}$ which is separable and complete under the usual Euclidean metric. We use the following definition from (Cra02, Section 3):

Definition 8. Given a probability space $(\tilde{\Omega}, \mathbb{P})$ and $\mathscr{B}$ the collection of Borel sets of $\mathbb{R}$ a

random probability measure on $\mathbb{R}$ is a map

$$\begin{aligned} \nu : \mathscr{B} \times \tilde{\Omega} &\to [0,1] \\ (B, \omega) &\mapsto \nu_\omega(B) \end{aligned}$$

such that

- for every $B \in \mathscr{B}$, the function $\omega \mapsto \nu_\omega(B)$ is measurable,
- for all $\omega \in \tilde{\Omega}$ except for some set of $\mathbb{P}$-measure zero, $B \mapsto \nu_\omega(B)$ is a Borel probability measure.

Theorem A.31. *For binary potential outcomes satisfying Neyman's weak null and the regularity conditions of (AGL14, Proposition 1) the conditional bootstrap distribution of the $\sqrt{N}$-scaled difference in means generated by the procedure of (IM21) converges almost surely in the bounded-Lipschitz metric to a random probability measure which takes the value $\mathcal{N}(0, V_{<,\infty})$ with probability $1/2$ and the value $\mathcal{N}(0, V_{>,\infty})$ with probability $1/2$.*

Proof. For $\omega \in \tilde{\Omega}$ define the random probability measure $\nu_\omega^{(N)}$ as

$$\nu_\omega^{(N)} = \begin{cases} \mathcal{N}(0, V_{<,\infty}) & \text{if } \omega \in \mathcal{A}_N, \\ \mathcal{N}(0, V_{>,\infty}) & \text{if } \omega \in \mathcal{B}_N, \\ \mathcal{N}(0, V_{=,\infty}) & \text{if } \omega \in \mathcal{C}_N. \end{cases}$$

Consider an $\omega \in \tilde{\Omega}$ such that

- $\omega \in \Big\{\omega \in \tilde{\Omega} \,:\, |I_{\mathcal{A}}| = \infty \text{ and } |I_{\mathcal{B}}| = \infty\Big\}$,
- $\tau^{*,(N_j)}(B^{(N_j)}, \omega)$ converges in distribution to $\mathcal{N}(0, V_{<,\infty})$ along the subsequence $N_j \in I_{\mathcal{A}}(\omega)$,

- $\tau^{*,(N_j)}(B^{(N_j)},\omega)$ converges in distribution to $\mathcal{N}\left(0,V_{>,\infty}\right)$ along the subsequence $N_j \in I_{\mathcal{B}}(\omega)$.

Since, $\tau^{*,(N_j)}(B^{(N_j)},\omega) \xrightarrow{d} \mathcal{N}\left(0,V_{<,\infty}\right)$ along the subsequence $N_j \in I_{\mathcal{A}}(\omega)$, and ρ_{BL} metrizes weak convergence we have that

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N_j)}(B^{(N_j)},\omega)\right),\mathcal{N}\left(0,V_{<,\infty}\right)\right) \to 0 \quad \text{along } N_j \in I_{\mathcal{A}}(\omega),$$

where $\mathcal{L}\left(\tau^{*,(N_j)}(B^{(N_j)},\omega)\right)$ denotes the law of the random variable $\tau^{*,(N_j)}(B^{(N_j)},\omega)$. Simply using the definition of $\nu_{\omega}^{(N_j)}$ we rewrite the limit above as

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N_j)}(B^{(N_j)},\omega)\right),\nu_{\omega}^{(N_j)}\right) \to 0 \quad \text{along } N_j \in I_{\mathcal{A}}(\omega). \tag{48}$$

Analogous reasoning yields

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N_j)}(B^{(N_j)},\omega)\right),\nu_{\omega}^{(N_j)}\right) \to 0 \quad \text{along } N_j \in I_{\mathcal{B}}(\omega). \tag{49}$$

We now break the proof into two cases depending upon the cardinality of $I_{\mathcal{C}}(\omega)$.

Case 1: Assume that $I_{\mathcal{C}}(\omega)$ is of infinite cardinality. Then, taking limits along the sequence $N_j \in I_{\mathcal{C}}(\omega)$ makes sense, and the reasoning used in the analysis of $I_{\mathcal{A}}(\omega)$ and $I_{\mathcal{B}}(\omega)$ implies that – for all but some set of ω with measure zero – $\tau^{*,(N_j)}(B^{(N_j)},\omega)$ converges in distribution to $\mathcal{N}\left(0,V_{=,\infty}\right)$ along the subsequence $N_j \in I_{\mathcal{C}}(\omega)$. Then,

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N_j)}(B^{(N_j)},\omega)\right),\nu_{\omega}^{(N_j)}\right) \to 0 \quad \text{along } N_j \in I_{\mathcal{C}}(\omega). \tag{50}$$

Since $I_{\mathcal{A}}(\omega)$, $I_{\mathcal{B}}(\omega)$, and $I_{\mathcal{C}}(\omega)$ partition $\mathbb{N}$ we use (48), (49), and (50) to conclude that

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N)}(B^{(N)},\omega)\right),\nu_{\omega}^{(N)}\right) \to 0 \quad \text{along } N \in \mathbb{N}.$$

Case 2: Assume that $I_{\mathcal{C}}(\omega)$ is of finite cardinality. In this situation, taking limits along the sequence $N_j \in I_{\mathcal{C}}(\omega)$ does not makes sense, so the argument from the previous case cannot be applied here. However, since $|I_{\mathcal{C}}(\omega)| < \infty$ there exists some $\tilde{N}(\omega)$ for which $n \leq \tilde{N}(\omega)$ for all $n \in I_{\mathcal{C}}(\omega)$. Since the limiting behaviour of $\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N)}(B^{(n)},\omega)\right),\nu_{\omega}^{(n)}\right)$ is insensitive to the behaviour of any finite number of n we can discard all $n \leq \tilde{N}(\omega)$ without changing the limit. Consequently, when $|I_{\mathcal{C}}(\omega)| < \infty$, we can completely ignore $I_{\mathcal{C}}(\omega)$ when examining the asymptotic behaviour of $\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N)}(B^{(n)},\omega)\right),\nu_{\omega}^{(n)}\right)$. Then (48) and (49) imply that

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N)}(B^{(N)},\omega)\right),\nu_{\omega}^{(N)}\right) \to 0 \quad \text{along } N \in \mathbb{N}.$$

Consequently, regardless of the cardinality of $I_{\mathcal{C}}(\omega)$ it holds that

$$\rho_{BL}\left(\mathcal{L}\left(\tau^{*,(N)}(B^{(N)},\omega)\right),\nu_{\omega}^{(N)}\right) \to 0. \tag{51}$$

By Lemma A.20, the set of $\omega \in \tilde{\Omega}$ for which (51) holds is of probability one. The interpretation of (51) is that almost surely the conditional bootstrap distribution – viewed as a random probability measure – can be approximated arbitrarily well by a random probability measure which takes values on the three Gaussian measures $\mathcal{N}(0, V_{<,\infty})$, $\mathcal{N}(0, V_{>,\infty})$, and $\mathcal{N}(0, V_{=,\infty})$

By the central limit theorem, $\mathbb{P}(\mathcal{A}_N) \to 1/2$, $\mathbb{P}(\mathcal{B}_N) \to 1/2$, and $\mathbb{P}(\mathcal{C}_N) \to 0$. So $\nu^{(N)}$ – in the sense of (Kal17, Theorem 4.11) – limits in distribution in the vague topology to the random measure which takes the value $\mathcal{N}(0, V_{<,\infty})$ with probability $1/2$ and the value $\mathcal{N}(0, V_{>,\infty})$ with probability $1/2$. □

Since the Prohorov metric and the bounded Lipschitz metric both metrize weak convergence of measures on $\mathbb{R}$ under the Euclidean metric (vdVW96, Theorem 1.12.4 and Page 77) (51) establishes convergence in the Prohorov metric as well. Consequently, in the sense of

(Kal17, Lemma 4.3) Theorem A.31 establishes that

$$\mathcal{L}\left(\tau^{*,(N)}(B^{(N)})\right) \xrightarrow{w} \nu^{(N)}$$

almost surely.

4.22 Additional Technical Results For Finite Population Inference

Suppose that the set $\{\mathbf{A}_1^{(N)}, \ldots, \mathbf{A}_N^{(N)}\}$ forms a deterministic finite population with each $\mathbf{A}_i^{(N)} \in \mathbb{R}^d$. Assume that the finite populations $\{\mathbf{A}_1^{(N)}, \ldots, \mathbf{A}_N^{(N)}\}$ satisfy the regularity conditions of Assumptions A.6 and A.7. For notation we write the mean and covariance matrix of this finite population as

$$\begin{aligned}
\overline{\mathbf{A}}^{(N)} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{A}_i^{(N)} \\
\mathbf{V}^{(N)} &= \frac{1}{N-1}\sum_{i=1}^{N} \left(\mathbf{A}_i^{(N)} - \overline{\mathbf{A}}^{(N)}\right)\left(\mathbf{A}_i^{(N)} - \overline{\mathbf{A}}^{(N)}\right)^T.
\end{aligned}$$

For $Z^{(N)} \sim \text{Unif}(\Omega)$ define the sample mean and covariance matrix as

$$\begin{aligned}
\overline{\mathbf{a}}^{(N)} &= \frac{1}{n_1}\sum_{i=1}^{N} Z_i^{(N)} \mathbf{A}_i^{(N)} \\
\hat{\mathbf{V}}^{(N)} &= \frac{1}{n_1 - 1}\sum_{i=1}^{N} Z_i^{(N)} \left(\mathbf{A}_i^{(N)} - \overline{\mathbf{a}}^{(N)}\right)\left(\mathbf{A}_i^{(N)} - \overline{\mathbf{a}}^{(N)}\right)^T.
\end{aligned}$$

For a random vector $\chi \in L_2(\mathbb{R}^d)$ denote its covariance matrix by $\text{cov}(\chi)$.

Lemma A.21. *Let $\omega_1, \ldots, \omega_N$ be drawn independently and uniformly from*

$\{\mathbf{A}_1^{(N)}, \ldots, \mathbf{A}_N^{(N)}\}$. *If*

$$cov\left(\overline{\mathbf{a}}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}}^{(N)} - \overline{\mathbf{A}}^{(N)}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right) \tag{52}$$

then

$$cov\left(\frac{1}{N}\sum_{i=1}^{N}\omega_i\right)^{-1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\omega_i - \overline{\mathbf{A}}^{(N)}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right).$$

Proof. We start with the univariate case (ie. $d = 1$). In this case, $\text{cov}\left(\overline{\mathbf{a}}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}}^{(N)} - \overline{\mathbf{A}}^{(N)}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right)$ reduces to

$$\frac{\sqrt{N}\left(\overline{\mathbf{a}}^{(N)} - \overline{\mathbf{A}}^{(N)}\right)}{\sqrt{N\text{cov}\left(\overline{\mathbf{a}}^{(N)}\right)}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right)$$

As noted in (LD17, Section A.1) $\overline{\mathbf{a}}^{(N)}$ is asymptotically normal in the sense of (52) if and only if for all $\varepsilon > 0$

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(S_i^{(N)}\right)^2 \mathbb{1}_{\left\{\left|S_i^{(N)}\right| > \varepsilon\sqrt{n_1(N-n_1)/N}\right\}} \to 0 \tag{53}$$

where $S_i^{(N)} = (\mathbf{A}_i^{(N)} - \overline{\mathbf{A}}^{(N)})/\sqrt{\mathbf{V}^{(N)}}$. Consider selecting $S_1^*, \ldots, S_N^*$ independently and uniformly from $\left\{S_i^{(N)}\right\}_{i=1}^{N}$. The Lindeberg condition (LR05, Equation 11.11) holds for the $S_1^*, \ldots, S_N^*$. We now detail why this claim holds. The Lindeberg condition requires that

$$\frac{1}{\sum_{i=1}^{N}\text{cov}\left(S_i^*\right)}\sum_{i=1}^{N}\mathbb{E}\left[\left(S_i^* - \mathbb{E}\left[S_i^*\right]\right)^2 \mathbb{1}_{\left\{\left|S_i^* - \mathbb{E}\left[S_i^*\right]\right| > \varepsilon\sqrt{\sum_{i=1}^{N}\text{cov}\left(S_i^*\right)}\right\}}\right] \tag{54}$$

converges to zero. Because $\operatorname{cov}(S_i^*) = 1$ and $\mathbb{E}[S_i^*] = 0$, (54) reduces to

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(S_i^*)^2\,\mathbb{1}_{\{|S_i^*|>\varepsilon\sqrt{N}\}}\right] &= \frac{1}{N}\left(N\mathbb{E}\left[(S_1^*)^2\,\mathbb{1}_{\{|S_1^*|>\varepsilon\sqrt{N}\}}\right]\right)\\ &= \sum_{i=1}^{N}\left(S_i^{(N)}\right)^2\mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{N}\right\}}\mathbb{P}\left(S_1^* = S_i^{(N)}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(S_i^{(N)}\right)^2\mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{N}\right\}}.\end{aligned}$$

Since $n_1/N \in (0,1)$ and $0 \le n_1 \le N$, $\sqrt{n_1(N-n_1)/N} \le \sqrt{N}$ for all N. Thus,

$$\mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{n_1(N-n_1)/N}\right\}} \ge \mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{N}\right\}}.$$

Thus,

$$\begin{aligned}0 \le \limsup \frac{1}{N}\sum_{i=1}^{N}&\left(S_i^{(N)}\right)^2\mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{N}\right\}}\\ &\le \limsup\frac{1}{N-1}\sum_{i=1}^{N}\left(S_i^{(N)}\right)^2\mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{n_1(N-n_1)/N}\right\}} = 0\end{aligned}$$

and so

$$\frac{1}{N}\sum_{i=1}^{N}\left(S_i^{(N)}\right)^2\mathbb{1}_{\left\{\left|S_i^{(N)}\right|>\varepsilon\sqrt{N}\right\}} \to 0. \tag{55}$$

This confirms that (54) converges to zero, which verifies the Lindeberg condition for $\{S_i^*\}_{i=1}^N$.

With the univariate case completed, we now generalize to the multivariate case ($d > 1$).

By the Cramér-Wold Device (LR05, Theorem 11.2.3),

$$\begin{gathered}\operatorname{cov}\left(\overline{\mathbf{a}}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}}^{(N)}-\overline{\mathbf{A}}^{(N)}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right) \\ \Updownarrow \\ \operatorname{cov}\left(\langle \mathbf{v}, \overline{\mathbf{a}}^{(N)}\rangle\right)^{-1/2}\left\langle \mathbf{v}, \overline{\mathbf{a}}^{(N)}-\overline{\mathbf{A}}^{(N)}\right\rangle \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 1\right) \quad \forall\, \mathbf{v}\in\mathbb{R}^d. \end{gathered} \tag{56}$$

For any fixed choice of $\mathbf{v}\in\mathbb{R}^d$, (56) is a univariate statement about the asymptotic normality of the sample mean for the population $\{\langle \mathbf{v}, \mathbf{A}_1^{(N)}\rangle, \ldots, \langle \mathbf{v}, \mathbf{A}_N^{(N)}\rangle\}$. Applying the proof from the case of $d=1$, it follows that the sample mean of N *i.i.d.* uniform samples from $\{\langle \mathbf{v}, \mathbf{A}_1^{(N)}\rangle, \ldots, \langle \mathbf{v}, \mathbf{A}_N^{(N)}\rangle\}$ (after suitable centering and scaling) displays central limit behavior. Since this holds for any $\mathbf{v}\in\mathbb{R}^d$, the Cramér-Wold Device establishes asymptotic normality in the sense of

$$\operatorname{cov}\left(\frac{1}{N}\sum_{i=1}^{N}\omega_i\right)^{-1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\omega_i-\overline{\mathbf{A}}^{(N)}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0}, I_{d\times d}\right).$$

□

Lemma A.22. *Assume that* $\overline{\mathbf{a}}^{(N)}-\overline{\mathbf{A}}^{(N)}\xrightarrow{a.s.}\mathbf{0}$ *and* $\hat{\mathbf{V}}^{(N)}-\mathbf{V}^{(N)}\xrightarrow{a.s.}\mathbf{0}_{d\times d}$.

Given $Z^{(N)}$ *define the empirical measure* $\hat{F}=\frac{1}{n_1}\sum_{i=1}^{N}Z_i^{(N)}\delta_{\mathbf{A}_i^{(N)}}$.
Take $\mathbf{a_1^*},\ldots,\mathbf{a_{n_1}^*}\overset{iid}{\sim}\hat{F}$ *and define their sample mean as*

$$\overline{\mathbf{a}^*}^{(N)}=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{a_i}^*.$$

As in Lemma A.21, let $\omega_1,\ldots,\omega_N$ *be drawn independently and uniformly from* $\{\mathbf{A}_1^{(N)},\ldots,\mathbf{A}_N^{(N)}\}$. *If the Lindeberg condition holds for the* ω_i *then conditional on* $Z^{(N)}$

$$cov\left(\overline{\mathbf{a}^*}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}^*}^{(N)}-\overline{\mathbf{a}}^{(N)}\right)\;\Big|\; Z^{(N)}\xrightarrow{d}\mathcal{N}\left(\mathbf{0}, I_{d\times d}\right)$$

almost surely with respect to randomness in $Z^{(N)}$.

Proof. By the same logic that was used in the proof of Lemma A.21, it suffices to show the univariate case and then rely on the Cramér-Wold device to carry through to the case of arbitrary d. As such, fix $d = 1$.

The Lindeberg condition for the $\mathbf{a}_i^* \mid Z^{(N)}$ is that

$$\frac{1}{\sum_{i=1}^N \operatorname{cov}\left(\mathbf{a}_i^*\right)} \sum_{i=1}^N \mathbb{E}\left[\left(\mathbf{a}_i^* - \mathbb{E}\left[\mathbf{a}_i^*\right]\right)^2 \mathbb{1}_{\left\{\left|\mathbf{a}_i^* - \mathbb{E}\left[\mathbf{a}_i^*\right]\right| > \varepsilon \sqrt{\sum_{i=1}^N \operatorname{cov}\left(\mathbf{a}_i^*\right)}\right\}}\right] \tag{57}$$

converges to 0 conditional upon $Z^{(N)}$. By assumption, $\mathbb{E}\left[\mathbf{a}_i^* \mid Z^{(N)}\right] - \overline{\mathbf{A}}^{(N)} \to \mathbf{0}$ for almost all sequences of $Z^{(N)}$. Similarly, by assumption $\operatorname{cov}\left(\mathbf{a}_i^* \mid Z^{(N)}\right) - \mathbf{V}^{(N)} \to \mathbf{0}_{d\times d}$ for almost all sequences of $Z^{(N)}$. Thus, for almost all sequences of $Z^{(N)}$, (57) conditional on $Z^{(N)}$ has the same limit as

$$\begin{aligned}
&\frac{1}{\sum_{i=1}^N \mathbf{V}^{(N)}} \sum_{i=1}^N \mathbb{E}\left[\left(\mathbf{a}_i^* - \overline{\mathbf{A}}^{(N)}\right)^2 \mathbb{1}_{\left\{\left|\mathbf{a}_i^* - \overline{\mathbf{A}}^{(N)}\right| > \varepsilon \sqrt{\sum_{i=1}^N \mathbf{V}^{(N)}}\right\}}\right] \\
&= \frac{1}{N\mathbf{V}^{(N)}} N\mathbb{E}\left[\left(\mathbf{a}_1^* - \overline{\mathbf{A}}^{(N)}\right)^2 \mathbb{1}_{\left\{\left|\mathbf{a}_1^* - \overline{\mathbf{A}}^{(N)}\right| > \varepsilon \sqrt{N\mathbf{V}^{(N)}}\right\}}\right] \\
&= \frac{1}{\mathbf{V}^{(N)}} \sum_{i=1}^N Z_i \left(\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}\right)^2 \mathbb{1}_{\left\{\left|\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}\right| > \varepsilon \sqrt{N\mathbf{V}^{(N)}}\right\}} \mathbb{P}\left(\mathbf{a}_1^* = \mathbf{A}_i \mid Z^{(N)}\right) \\
&= \sum_{i=1}^N Z_i \left(\frac{\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}}{\sqrt{\mathbf{V}^{(N)}}}\right)^2 \mathbb{1}_{\left\{\left|\frac{\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}}{\sqrt{\mathbf{V}^{(N)}}}\right| > \varepsilon \sqrt{N}\right\}} \mathbb{P}\left(\mathbf{a}_1^* = \mathbf{A}_i \mid Z^{(N)}\right) \\
&= N\left(\frac{1}{N}\sum_{i=1}^N Z_i \left(\frac{\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}}{\sqrt{\mathbf{V}^{(N)}}}\right)^2 \mathbb{1}_{\left\{\left|\frac{\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}}{\sqrt{\mathbf{V}^{(N)}}}\right| > \varepsilon \sqrt{N}\right\}} \mathbb{P}\left(\mathbf{a}_1^* = \mathbf{A}_i \mid Z^{(N)}\right)\right) \\
&\leq \left(\frac{N}{n_1}\right)\left(\frac{1}{N}\sum_{i=1}^N \left(\frac{\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}}{\sqrt{\mathbf{V}^{(N)}}}\right)^2 \mathbb{1}_{\left\{\left|\frac{\mathbf{A}_i - \overline{\mathbf{A}}^{(N)}}{\sqrt{\mathbf{V}^{(N)}}}\right| > \varepsilon \sqrt{N}\right\}}\right).
\end{aligned} \tag{58}$$

The term n_1^{-1} in the inequality of (58) follows from

$$\mathbb{P}\left(\mathbf{a}_1^* = \mathbf{A}_i \mid Z^{(N)}\right) = \begin{cases} n_1^{-1}; & \text{if } Z_i^{(N)} = 1 \\ 0; & \text{otherwise.} \end{cases}$$

Since the Lindeberg condition holds for the ω_i the second term of (58) converges to 0 (cf. (55)). By assumption $n_1/N \to p \in (0,1)$. Thus, (58) converges to zero; since the limit superior of (57) is upper bounded by the limit superior of (58) and is lower bounded by zero, it is sandwiched to zero. Thus, (57) limits to zero, thereby verifying the Lindeberg condition for the $\mathbf{a}_i^*$. Since (58) holds for any realization of $Z^{(N)}$ the result holds almost surely with respect to the conditioning variable $Z^{(N)}$. □

Lemma A.23. *Assume that* $\overline{\mathbf{a}}^{(N)} - \overline{\mathbf{A}}^{(N)} \xrightarrow{a.s.} \mathbf{0}$ *and* $\hat{\mathbf{V}}^{(N)} - \mathbf{V}^{(N)} \xrightarrow{a.s.} \mathbf{0}_{d\times d}$.

Given $Z^{(N)}$ *define the empirical measure* $\hat{F} = \frac{1}{n_1}\sum_{i=1}^{N} Z_i^{(N)} \delta_{\mathbf{A}_i^{(N)}}$.
Let $\mathbf{a_1}^*, \ldots, \mathbf{a_{n_1}}^* \overset{iid}{\sim} \hat{F}$ *and*

$$\overline{\mathbf{a}^*}^{(N)} = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{a_i}^*.$$

If $cov\left(\overline{\mathbf{a}}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}}^{(N)} - \overline{\mathbf{A}}^{(N)}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right)$ *then conditional on* $Z^{(N)}$

$$cov\left(\overline{\mathbf{a}^*}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}^*}^{(N)} - \overline{\mathbf{a}}^{(N)}\right) \;\Big|\; Z^{(N)} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I_{d\times d}\right)$$

almost surely with respect to randomness in $Z^{(N)}$.

Proof. The conditions assumed imply Lemma A.21. In turn, Lemma A.21 implies that Lemma A.22 holds, which establishes the desired result.

On an informal level, the argument can be summarized as follows:

1. The conditions required to guarantee a finite sample central limit theorem for

$$\operatorname{cov}\left(\overline{\mathbf{a}}^{(N)}\right)^{-1/2}\left(\overline{\mathbf{a}}^{(N)}-\overline{\mathbf{A}}^{(N)}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},I_{d\times d}\right)$$

   imply a central limit theorem for the sample mean when independently uniformly sampling from the *full* finite population. Furthermore, not only does a central limit theorem hold for the *i.i.d.* samples, but so too does the Lindeberg condition (see Lemma A.21).

2. The Lindeberg condition (and resulting central limit theorem) for the sample mean of *i.i.d.* samples from the full population implies that a Lindeberg condition holds for sampling from a *restricted subset* of the population - specifically those for which $Z_i^{(N)}=1$ (see Lemma A.22).

□

Remark 16. Lemma A.23 can be informally read as: *if a central limit theorem holds for the mean when sampling without replacement from a finite population, then a central limit theorem also holds for the mean when i.i.d. subsampling from the Horvitz-Thompson empirical measure in the finite population model.*

Bibliography

[AGL14] Peter M. Aronow, Donald P. Green, and Donald K. K. Lee. Sharp bounds on the variance in randomized experiments. *Ann. Statist.*, 42(3):850–871, 2014.

[And55] T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.*, 6:170–176, 1955.

[BBB+19] Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: consequences illustrated with linear regression. *Statist. Sci.*, 34(4):523–544, 2019.

[BBK+19] Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as approximations II: A model-free theory of parametric regression. *Statist. Sci.*, 34(4):545–565, 2019.

[BF81] Peter J. Bickel and David A. Freedman. Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6):1196 – 1217, 1981.

[CF21] Peter L. Cohen and Colin B. Fogarty. No-harm calibration for generalized oaxacablinder estimators, 2021.

[CF22] Peter L. Cohen and Colin B. Fogarty. Gaussian prepivoting for finite population causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):295–320, 2022.

[CGHH17] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge-kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

[Coc77] William G. Cochran. *Sampling techniques*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-London-Sydney, third edition, 1977.

[Cra02] Hans Crauel. *Random probability measures on Polish spaces*, volume 11 of *Stochastics Monographs*. Taylor & Francis, London, 2002.

[CT97] Yuan Shih Chow and Henry Teicher. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. Independence, interchangeability, martingales.

[DDCZ13] Lutz Dümbgen and Perla Del Conte-Zerial. On low-dimensional projections of high-dimensional distributions. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 91–104. Inst. Math. Statist., Beachwood, OH, 2013.

[DFM19] Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

[DS21] Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 0(0):1–16, 2021.

[Dur10] Rick Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010.

[Efr79] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[Efr82] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[FCG+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[Fel68] William Feller. *An introduction to probability theory and its applications. Vol. I.* John Wiley & Sons, Inc., New York-London-Sydney, third edition, 1968.

[Fre08a] David A. Freedman. On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.*, 2(1):176–196, 2008.

[Fre08b] David A. Freedman. On regression adjustments to experimental data. *Adv. in Appl. Math.*, 40(2):180–193, 2008.

[GB20] Kevin Guo and Guillaume Basse. The generalized oaxaca-blinder estimator, 2020.

[HW91] Peter Hall and Susan R. Wilson. Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2):757–762, 1991.

[IM21] Guido Imbens and Konrad Menzel. A causal bootstrap. *The Annals of Statistics*, 49(3):1460 – 1488, 2021.

[Kal17] Olav Kallenberg. *Random measures, theory and applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2017.

[LD17] Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.*, 112(520):1759–1769, 2017.

[LD18] Lihua Lei and Peng Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates, 2018.

[LDR18] Xinran Li, Peng Ding, and Donald B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

[Lin13] Winston Lin. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Stat.*, 7(1):295–318, 2013.

[Liu88] Regina Y. Liu. Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, 16(4):1696–1708, 1988.

[LR05] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[LS95] Regina Y. Liu and Kesar Singh. Using i.i.d. bootstrap inference for general non-i.i.d. models. *J. Statist. Plann. Inference*, 43(1-2):67–75, 1995.

[MHL16] Zeinab Mashreghi, David Haziza, and Christian Léger. A survey of bootstrap methods in finite population sampling. *Statist. Surv.*, 10:1–52, 2016.

[MR12] Kari Lock Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *Ann. Statist.*, 40(2):1263–1282, 04 2012.

[MS22] Gilles Mordant and Johan Segers. Measuring dependence between random vectors via optimal transport. *Journal of Multivariate Analysis*, 189:104912, 2022.

[Nel06] Roger B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006.

[Ney90] Jerzy Splawa Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.*, 5(4):465–472, 1990. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed.

[PC19] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*, volume 37 of *Foundations and Trends® in Machine Learning*. now, Norwell, MA, 2019.

[PZ20] Victor M. Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham, [2020] ©2020.

[RBK05] Susana Rubin-Bleuer and Ioana Schiopu Kratina. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789 – 2810, 2005.

[Rob88] James M. Robins. Confidence intervals for causal parameters. *Statistics in Medicine*, 7(7):773–785, 1988.

[San15] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[SS07] Moshe Shaked and J. George Shanthikumar, editors. *Stochastic Orders*. Springer New York, 2007.

[ST95] Jun Shao and Dong Sheng Tu. *The jackknife and bootstrap*. Springer Series in Statistics. Springer-Verlag, New York, 1995.

[Tch80] Andre H. Tchen. Inequalities for distributions with given marginals. *The Annals of Probability*, 8(4):814–827, 1980.

[Ton90] Y. L. Tong. *The Multivariate Normal Distribution*. Springer Series in Statistics. Springer-Verlag, New York, 1990.

[vdV98] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[WD21] Jason Wu and Peng Ding. Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116(536):1898–1913, 2021.