Optimized variance estimation under interference and complex experimental designs

Christopher Harshaw[1], Joel Middleton[2], and Fredrik Sävje[3]

[1] *Massachusetts Institute of Technology.*
[2] *??? FIX*
[3] *Yale University.*

May 26, 2024

Abstract

Unbiased and consistent variance estimators generally do not exist for design-based treatment effect estimators because experimenters never observe more than one potential outcome for any unit. The problem is exacerbated by interference and complex experimental designs. Experimenters must accept conservative variance estimators in these settings, but they can strive to minimize conservativeness. In this paper, we show that the task of constructing a minimally conservative variance estimator can be interpreted as an optimization problem that aims to find the lowest estimable upper bound of the true variance given the experimenter's risk preference and knowledge of the potential outcomes. We characterize the set of admissible bounds in the class of quadratic forms, and we demonstrate that the optimization problem is a convex program for many natural objectives. The resulting variance estimators are guaranteed to be conservative regardless of whether the background knowledge used to construct the bound is correct, but the estimators are less conservative if the provided information is reasonably accurate. Numerical results show that the resulting variance estimators can be considerably less conservative than existing estimators, allowing experimenters to draw more informative inferences about treatment effects.

Keywords: Causal inference, randomized experiments, variance estimation.

This paper builds on a working paper by one of the coauthors (Middleton, 2018). We thank P. M. Aronow, Mateo Díaz, Peng Ding, Jon Erickson, Avi Feller, Erin Hartman, Guido Imbens, Miles Lubin, Cyrus Samii, Daniel Spielman and Sekhar Tatikonda for helpful comments and discussions. Christopher Harshaw was supported by NSF Graduate Research Fellowship (DGE1122492) and Foundations of Data Science Institute (FODSI) NSF grant DMS2023505.

1 Introduction

The design-based, finite population approach to causal inference considers treatment assignment as the only source of randomness. In this framework, the variance of treatment effect estimators depends on aspects of the joint distribution of the potential outcomes. This poses a challenge for variance estimation because experimenters can never observe more than one potential outcomes for each unit, meaning that the observed potential outcomes provide little information about the joint distribution. Without strong assumptions, such as constant treatment effects, it is not possible to consistently estimate the variance of treatment effect estimators.

In this paper, we consider variance estimation in experiments with interference or complex experimental designs. Interference occurs when the treatment assigned to one unit affects other units. The variance estimation problem is particularly difficult in these settings, because interference and complex designs typically introduce strong dependencies between effective treatments, making more aspects of the joint distribution inaccessible. Conventional techniques for constructing conservative variance estimators can therefore not be used, or they produce overly conservative estimators. The purpose of this paper is to address this problem by describing valid variance estimators that minimize conservativeness.

Following the previous literature, we break up the task of constructing a variance estimator by first constructing an upper bound for the variance. An estimator of the bound then acts as a conservative estimator of the variance. However, unlike previous work, we consider the variance estimation under arbitrary interference and arbitrary experimental designs. We also consider a large class of linear point estimators, which includes all commonly used treatment effect estimators.

There are two main contributions of the paper. First, in Section 4, we describe and characterize the variance estimation problem. We define a concept of admissibility for the class of variance bounds that are quadratic forms, allowing us to discard a large set of poorly performing variance bounds. The characterization allows us to understand previous variance bounds in a common framework, and we show that the currently most commonly used type of variance bound is inadmissible.

Second, in Section 5, we reinterpret the task of selecting a variance bound as an optimization problem. We describe two classes of objective functions, which allow experimenters to construct variance bounds based on their level of risk aversion and prior substantive knowledge. The bounds are valid and admissible no matter which class of objective functions is used and no matter if the supplied information is correct, but the resulting bound is less conservative if the information is accurate. The underlying optimization problem is convex, meaning that it is computational tractable.

Supplementary contributions include an investigation in Section 6 of how challenges when estimating a bound affect which bound to select. We highlight that some bounds are easier to estimate than others, meaning that we might prefer a bound that is more conservative if we can estimate it with greater precision. Section 7 reports the results from a simulation exercise based on real-world data examining how the methods we describe in the paper behave in practice.

2 Illustration and Preview of Main Results

To illustrate the central question of the paper, we consider a stylized version of the study by Paluck, Shepherd, and Aronow (2016), which we also use in the simulation exercise in Section 7. The authors investigate whether an anticonflict intervention for students in US middle schools reduces conflict. They were particularly interested in how the effect of the intervention spread through the student peer network, and they investigated this by comparing students directly exposed to the intervention with students only indirectly exposed through their peer networks.

For simplicity in this illustration, we will consider a sample of only two students. The experiment is such that exactly one student, chosen at random with equal probability, will be directly exposed to the anticonflict intervention. The two students are in the same peer network, so the student not directly exposed is considered to be indirectly exposed. The estimator $\widehat{\tau}$ is the difference in observed outcomes of the two students. Let a_i denote the outcome of student $i \in \{1, 2\}$ when directly exposed, and let b_i denote outcome when indirectly exposed. The estimator takes two values with equal probability: $a_1 - b_2$ and $a_2 - b_1$. The variance of the estimator is therefore

$$\operatorname{Var}(\widehat{\tau}) = \frac{1}{4}(a_1^2 + a_2^2 + b_1^2 + b_2^2) + \frac{1}{2}(a_1 b_1 + a_2 b_2 - a_1 a_2 - b_1 b_2 - a_1 b_2 - a_2 b_1).$$

At the heart of the variance estimation problem is that some terms in the variance expression are never observed. We never observe $a_1 b_1$ or $a_2 b_2$, because a student cannot be assigned to both direct and indirect exposure at the same time. Similarly, we never observe $a_1 a_2$ or $b_1 b_2$, because the two students are always assigned to different exposures. The unobserved terms prevent us from constructing an unbiased, or even consistent, estimator of the variance unless we impose strong assumptions on the potential outcomes.

A common way to address this problem is to construct an estimable, upper bound for the variance. An estimator of the bound acts as a conservative estimator of the variance. A simple upper bound in this setting uses the fact that $\operatorname{Var}(\widehat{\tau}) \leq \operatorname{E}[\widehat{\tau}^2]$. Hence, the variance $\operatorname{Var}(\widehat{\tau})$ is upper bounded by

$$B_1 = \operatorname{E}\big[\widehat{\tau}^2\big] = \frac{1}{2}(a_1^2 + a_2^2 + b_1^2 + b_2^2) - (a_1 b_2 + a_2 b_1).$$

Note that this bound holds for any values of the potential outcomes, so no additional assumptions are required for its validity. Furthermore, the bound is estimable because all terms are observed with some positive probability.

The bound we just derived is only one of many possible bounds. A somewhat more intricate bound uses the fact that $(x^2 + y^2)/2$ is an upper bound for the product xy for any real-valued x and y. Applying this inequality to the problematic terms in the variance expression, we arrive at the bound

$$B_2 = \frac{3}{4}(a_1^2 + a_2^2 + b_1^2 + b_2^2) - \frac{1}{2}(a_1 b_2 + a_2 b_1).$$

Both B_1 and B_2 are estimable bounds, so either can be used to construct a variance estimator that is conservative in expectation. Indeed, there are infinitely many estimable bounds in this setting, with infinitely many corresponding conservative variance estimators. While we want a variance estimator that is conservative, which would be achieved by any of these bounds, we want to avoid excessive conservativeness.

The idea we explore in this paper is to use an optimization approach to choose one of these estimable bounds for the variance estimator so as to minimize conservativeness. To make the approach tractable, we focus on bounds that are quadratic forms; the variance itself is a known quadratic form in the potential outcomes, so we find it natural to restrict attention to bounds of the same form. We collect all estimable bounds that are quadratic forms in the set $\mathcal{B}$. Any conservative variance estimator we consider will correspond to an element in this set.

The set $\mathcal{B}$ is always infinite, and it always contains bounds that are overly conservative. We describe a concept of admissibility to characterize bounds that are unnecessarily conservative. A bound is inadmissible if there exists another (valid) bound that is less conservative no matter what the potential outcomes might be, in which case we say that the second bound dominates the first. If an inadmissible bound is used to construct a variance estimator, we say that also the estimator inadmissible. The bias of an inadmissible variance estimator will be larger than the bias of the variance estimator that dominates it, motivating us to never use an inadmissible estimator.

Of the two bounds considered in this section, B_1 dominates B_2, meaning that B_2 is inadmissible. In particular, their difference is

$$B_2 - B_1 = \frac{(a_1 + b_2)^2 + (a_2 + b_1)^2}{4} \geq 0,$$

so a variance estimator based on B_2 will in expectation always be more conservative than an estimator based on B_1.

The admissibility concept allows us to discard many bounds in $\mathcal{B}$, but there will generally still be infinitely many admissible bounds, so admissibility alone does not allow us to select a

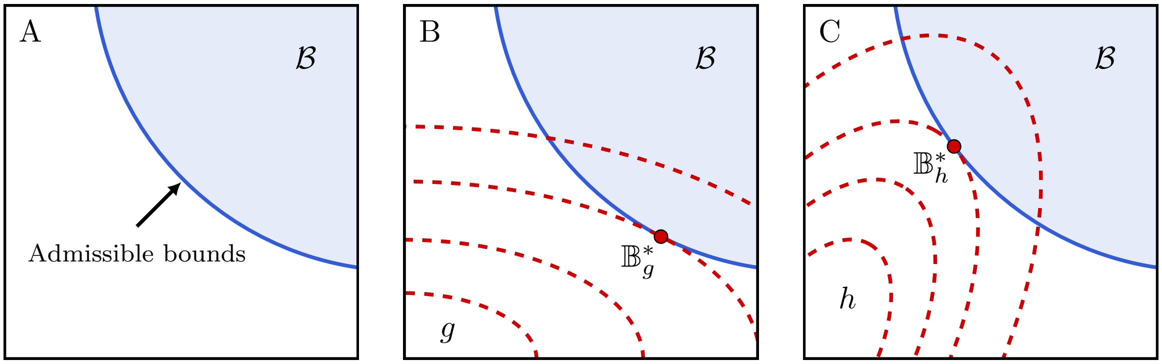


Figure 1: Illustration of selecting of a variance bound using optimization.

bound to use in our experiments. We suggest that experimenters select the admissible variance bound that best conforms with their risk preferences and any background information they might have, as encoded in an objective function $g : \mathcal{B} \to \mathbb{R}$. The selected variance bound is the minimizer of the objective function: $\mathbb{B}^* = \arg\min_{\mathbb{B}\in\mathcal{B}} g(\mathbb{B})$. The properties of the resulting variance estimator will inevitably depend on the choice of the objective function g. However, one central result of the paper is that this approach produces admissible bounds for a large class of objective functions. This result is presented and discussed as Theorem 1 in Section 5.1 below, and it is previewed here for reference. Importantly, the theorem holds no matter if the background information used to construct the objective function g is correct. The approach therefore allows experimenters to target the variance estimator to their setting without risking to inadvertently using an anticonservative or inadmissible variance estimator.

Theorem 1. *If the objective function g is strictly monotone, then the bound based on the minimizer of g in $\mathcal{B}$ is conservative, estimable and admissible.*

Figure 1 provides a graphical illustration of the approach we explore in the paper. Each panel should be interpreted a set of potential estimation targets. The illustration is highly stylized, and the axes do not necessarily correspond to any particular parameterization. In Panel A, the set $\mathcal{B}$ is plotted, where the region shaded in light blue contains all valid and estimable bounds. The boundary of this region, marked in a stronger blue, is the set of all admissible bounds.

Panels B and C show the contour lines of two different objective functions, g and h, where contour lines further from the origin corresponds to higher values. The two minimizers, $\mathbb{B}_g^*$ and $\mathbb{B}_h^*$, are different, but they are both on the boundary containing the admissible bounds.

3 Related Work

Neyman (1990/1923) was first to recognize that the variance of a treatment effect estimator is not directly estimable. He showed that the variance of the difference-in-means estimator under the complete randomization design depends on the covariance of unit-level potential outcomes, which cannot be estimated from the data. Neyman applied the Cauchy–Schwarz inequality followed by the AM–GM inequality to arrive at an estimable upper bound of the variance. He also noted that unbiased variance estimation is possible when treatment effects are constant between units, which sometimes is referred to as strict additivity.

Neyman's approach has been improved and extended in several directions. An important line of work aims to sharpen the bound. Robins (1988) focuses on binary outcomes and derives a variance estimator that extracts all information about the joint distribution of the potential outcomes contained in the marginal distributions. Aronow, Green, and Lee (2014) use Fréchet–Hoeffding-type bounds to generalize the estimator by Robins (1988) to arbitrary outcome variables. Nutz and Wang (2021) provide further improvements under the assumption that all unit-level treatment effects are non-negative. Imbens and Menzel (2021) provide higher-order refinements to these bounds using a bootstrap approach. These bounds are applicable only when experimenters use the difference-in-means estimator under complete randomization; it is unclear whether and how these results generalize to more complex estimators and designs.

Another strand of the literature considers variance estimation under other experimental designs than complete randomization. Early examples include Kempthorne (1955) and Wilk (1955), who studied variance estimation under various blocked designs. These investigations generally impose structural assumptions on the potential outcomes, such as strict additivity, which limits their applicability. A more recent strand of the literature has derived Neyman-type variance estimators for some types of blocked or stratified designs without such assumptions (see, e.g., Abadie & Imbens, 2008; Fogarty, 2018; Gadbury, 2001; Higgins, Sävje, & Sekhon, 2015; Imai, 2008; Pashley & Miratrix, 2021).

A related strand of the literature has derived Neyman-type variance estimators for other point estimators than the difference-in-means estimator. Samii and Aronow (2012) investigate variance estimators for the ordinary least square regression estimator, and Aronow and Middleton (2013) do the same for the Horvitz–Thompson estimator. Mukerjee, Dasgupta, and Rubin (2018) connect both of these strands of the literature and consider variance estimation for unbiased linear estimators of treatment effects for arbitrary experimental designs. These authors use a formulation reminiscent to the one in this paper to weaken the strict additivity assumption employed by Neyman (1990/1923) to obtain an unbiased variance estimator.

All papers mentioned so far in this section have assumed that the experimental units do not interfere with each other. The strand of the literature closest to the current paper considers variance estimation in settings with interference under arbitrary experimental designs. To the best of our knowledge, the only previous result here is due to Aronow and Samii (2013, 2017). They describe a method for constructing a bound for the variance of the Horvitz–Thompson estimator when many pair-wise assignment probabilities are zero, as often is the case under interference. In this paper, we ask whether better bounds exist in this setting. We answer this question in the affirmative. Indeed, as we show in Section 4.6, the Aronow–Samii bound is inadmissible in the class of bounds that we consider.

4 The Variance and Variance Bounds

4.1 Preliminaries

Consider an experiment consisting of n units indexed by $U = \{1, \ldots, n\}$. Each unit $i \in U$ is assigned one of two treatment conditions $z_i \in \{0, 1\}$. We collect the assignments of all units into an assignment vector $\boldsymbol{z} = (z_1, \ldots, z_n) \in \{0, 1\}^n$. The assignments are random, and Z_i denotes the random assignment for unit i. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ denote the random treatment vector that collects all units' assignments. The distribution of $\boldsymbol{Z}$ is the *design* of the experiment, which is taken to be known.

Each unit $i \in U$ has an associated *potential outcome* function $y_i : \{0, 1\}^n \to \mathbb{R}$ that specifies the response of unit i under all possible treatment assignments. Because the function y_i depends on the full assignment vector, the response of unit i is allowed to depend not only on its own treatment, but potentially also on the treatments assigned to other units. This is commonly referred to as interference. The observed outcome of unit i is $Y_i = y_i(\boldsymbol{Z})$, and the vector of all observed outcomes is denoted $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The potential outcome functions themselves are deterministic and the randomness in the observed outcomes arises from the fact that treatment is randomly assigned.

We will use the framework described by Aronow and Samii (2017) to model interference. A related framework is described by Manski (2013). Each unit $i \in U$ has an *exposure mapping* $d_i : \{0, 1\}^n \to \Delta$ that maps each assignment vector to a set of exposures Δ. When two or more assignment vectors map to the same exposure for some unit, those assignments are considered causally equivalent with respect to that unit. The number of exposures $|\Delta|$ is typically small compared to the number of units.

Experimenters using exposure mappings often assume that the mappings are correctly specified, and we will do the same in this paper. The assumption states that a unit's outcome is completely determined by its exposure, in the sense that $d_i(\boldsymbol{z}) = d_i(\boldsymbol{z}')$ implies $y_i(\boldsymbol{z}) = y_i(\boldsymbol{z}')$

for all $\boldsymbol{z}, \boldsymbol{z}' \in \{0,1\}^n$. For each unit $i \in U$, we define $D_i = d_i(\boldsymbol{Z})$ to be the exposure produced by the realized treatment assignment.

The causal quantity of interest τ in this context is typically an average contrast between two exposures. That is, given two exposures $e_1, e_0 \in \Delta$, experimenters aim to estimate

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\big[y_i(e_1) - y_i(e_0)\big],$$

where we have overloaded the notation by writing $y_i(e)$ to denote the outcome of unit i under exposure $e \in \Delta$. Many commonly studied estimands, including total, direct and indirect treatment effects, are of this form (Hudgens & Halloran, 2008). The conventional average treatment effect under a no-interference assumption also takes this form. While experimenters almost exclusively consider estimands that are unweighted averages of contrasts of potential outcomes, all results in this paper generalize to estimands that are arbitrary linear functions of the potential outcomes.

We consider the class of *linear estimators*. An estimator in this class can be written as a random linear combination of the observed outcomes:

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} W_i Y_i, \tag{1}$$

where the coefficients W_i may depend arbitrarily on the treatment assignments $\boldsymbol{Z}$ and characteristics of the units, but they cannot depend on the observed outcomes $\boldsymbol{Y}$. Thus, the coefficients can, and typically will, be random. In Section S1 of the supplement, we show that the class of linear estimators includes all estimators commonly used by experimenters to estimate treatment and exposure effects. This includes Horvitz–Thompson, IPW, difference-in-means, Hájek, OLS-adjusted and AIPW estimators.

4.2 The Variance of Linear Estimators

Any estimator in the class of linear estimators can be written as

$$\widehat{\tau} = \frac{1}{n} \sum_{k \in P} V_k \theta_k = n^{-1} \boldsymbol{V}^{\intercal} \boldsymbol{\theta},$$

where $P = [K]$ is a set of indices, $\boldsymbol{V} = (V_1, \ldots, V_K)$ is a vector of known random variables, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is a vector of unknown non-random potential outcomes. In many cases, the estimator depends only on two types of potential outcomes, in which case $K = 2n$, and the elements of $\boldsymbol{V}$ will take the form $W_i \mathbb{1}[D_i = e]$. In Section S2 of the supplement, we show how to go from the estimator written as in Equation (1) to the current form in a general setting.

The advantage of writing the estimator in the current form is that all randomness is isolated in the coefficient vector $\boldsymbol{V}$. This makes the derivation of the variance of the estimator straightforward, as shown in the following lemma. All proofs appear in Section S5 of the supplement.

Lemma 1. *The variance of a linear estimator* $\widehat{\tau} = n^{-1} \boldsymbol{V}^{\intercal} \boldsymbol{\theta}$ *is* $\operatorname{Var}(\widehat{\tau}) = n^{-2} \boldsymbol{\theta}^{\intercal} \mathbb{A} \boldsymbol{\theta}$*, where* $\mathbb{A} = \operatorname{Cov}(\boldsymbol{V})$ *is the covariance matrix of the coefficient vector* $\boldsymbol{V}$*.*

The lemma is useful because $\mathbb{A} = \operatorname{Cov}(\boldsymbol{V})$ does not depend on the potential outcomes, so it is known. Furthermore, because $\mathbb{A}$ is a covariance matrix, it is positive semidefinite. The variance is thus a known positive semidefinite quadratic form of the potential outcome vector, which makes it conducive to analysis.

A possible complication is that the covariance matrix may be difficult to derive analytically for some estimators and designs. If that turns out to be the case, experimenters can use numerical methods to compute the matrix (Fattorini, 2006). This generally does not cause troubles because experimenters can run the Monte Carlo simulation until the matrix is known to desired precision. However, to avoid distractions from the main ideas and insights of the paper, we will proceed under the assumption that $\mathbb{A}$ is known.

4.3 The Variance Is Not Estimable

The preceding subsection reduced the task of estimating the variance of a linear estimator to a task of estimating a (known) quadratic form in the (unknown) potential outcome vector $\boldsymbol{\theta}$.

For our purposes, the central problem is that some quadratic forms cannot be estimated well. In particular, some pairs of potential outcomes may never, or only very rarely, be observed at the same time, and this makes it difficult or impossible to estimate the quadratic form.

Let S be a random subset of P that collects the indices $k \in P$ of potential outcomes $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ that are observed under the realized treatments $\boldsymbol{Z}$. If $\Pr(k, \ell \in S) = 0$ for some pair $k, \ell \in P$, then the corresponding product $\theta_k \theta_\ell$ is never observed. These unobservable products will be central to our discussion, so we collect all such pairs in a set:

$$\Omega = \big\{(k, \ell) \in P \times P : \Pr(k, \ell \in S) = 0\big\}.$$

As formalized in the following definition and proposition, estimable quadratic forms are those that are compatible with this pattern of observability.

Definition 1. A quadratic form $\boldsymbol{\theta}^\intercal \mathbb{A} \boldsymbol{\theta}$ is *design compatible* if the probability of simultaneously observing θ_k and θ_ℓ is zero only when the corresponding element in $\mathbb{A}$ is zero:

$$\forall k, \ell \in P, \ (k, \ell) \in \Omega \implies a_{k\ell} = 0,$$

where $a_{k\ell}$ is the element in the kth row and ℓth column of $\mathbb{A}$.

Proposition 1. *An unbiased estimator exists for a quadratic form if and only if it is design compatible.*

Because it is impossible to simultaneously observe two potential outcome of the same unit, a quadratic form representing variances will always be design incompatible, no matter the design. The pattern of design incompatibility is such that the bias is large also in large samples, so consistent variance estimation is also impossible. Furthermore, the sign of the bias of a design incompatible quadratic form will generally not be known, meaning that inferences based on such a biased variance estimator could be anti-conservative.

The variance estimation problem is exacerbated by interference and complex experimental designs. When units interfere, the structure of the exposure mappings often prevents certain combinations of exposures to be simultaneously realizable. For example, when units interact with each other in a network, all neighbors of a unit that is treated will necessarily be indirectly exposed to treatment, meaning that they cannot be in a pure control condition if any of their neighbors are treated. A similar problem occurs with complex experimental designs, which often introduce strong dependencies between the treatment assignments of different units. This could either be in an effort to improve precision, such as with the matched-pair design, or because the design is forced on the experimenter by external factors, such as with the cluster-randomized design.

4.4 Conservative Variance Bounds

A variance estimator with bias of unknown sign could lead to misleading conclusions. To address this, experimenters tend to opt for conservative variance estimators that systematically overestimate the variance, providing a pessimistic assessment of the precision of the point estimator. Confidence intervals based on conservative variance estimators err on the side of caution, in the sense that they motivate firm conclusions only under disproportionately strong evidence.

We can understand conservative variance estimators as estimators of an upper bound of the variance. A *variance bound* is a function $\mathrm{VB}\colon \mathbb{R}^K \to \mathbb{R}$ that satisfies $\mathrm{VB}(\boldsymbol{\theta}) \geq \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\tau})$ for all potential outcomes $\boldsymbol{\theta} \in \mathbb{R}^K$. If the function also complies with the structure of simultaneous observability of the potential outcomes, in the sense that it is design compatible, then we can construct an estimator of $\mathrm{VB}(\boldsymbol{\theta})$. This estimator acts a conservative estimator of the variance, because it estimates a quantity that is guaranteed to be larger than the variance.

Implicitly in the previous literature, the focus has primarily been on upper bounds that themselves are positive semidefinite quadratic forms. We do the same in this paper. That is, we consider bounds of the form

$$\mathrm{VB}(\boldsymbol{\theta}) = \frac{1}{n^2}\boldsymbol{\theta}^{\intercal}\mathbb{B}\boldsymbol{\theta} = \frac{1}{n^2}\sum_{k\in P}\sum_{\ell\in P} b_{k\ell}\theta_k\theta_\ell,$$

where $\mathbb{B}$ is a K-by-K positive semidefinite matrix, and $b_{k\ell}$ is the element in the kth row and ℓth column of $\mathbb{B}$. Throughout the remainder of the paper, we will use $\mathbb{B}$ to refer to both the coefficient matrix and the variance bound function $\mathrm{VB}(\boldsymbol{\theta})$.

To serve its role as the basis for a conservative estimator, we require the variance bounds to be both conservative and design compatible. This imposes two types of constraints on the coefficient matrix $\mathbb{B}$. To satisfy design conservativeness, $\mathbb{B}$ must be larger than $\mathbb{A}$, in the sense that $\boldsymbol{\theta}^{\intercal}\mathbb{A}\boldsymbol{\theta} \leq \boldsymbol{\theta}^{\intercal}\mathbb{B}\boldsymbol{\theta}$ for all vectors $\boldsymbol{\theta}$. This is precisely the *Loewner partial order* on symmetric matrices, where $\mathbb{A} \preceq \mathbb{B}$ denotes that $\mathbb{B} - \mathbb{A}$ is positive semidefinite. To satisfy design compatibility, $\mathbb{B}$ must be such that $b_{k\ell} = 0$ for all pairs $(k, \ell) \in \Omega$. We refer to symmetric matrices that satisfy these two conditions as valid variance bounds.

Definition 2. A symmetric matrix $\mathbb{B}$ is a *valid* variance bound for $\mathbb{A}$ if it is larger than $\mathbb{A}$ in the Loewner order and design compatible under the current design. Let $\mathcal{B}$ collect all valid variance bounds: $\mathcal{B} = \big\{\mathbb{B} : \mathbb{A} \preceq \mathbb{B} \text{ and } b_{k\ell} = 0 \text{ for all } (k, \ell) \in \Omega\big\}$.

An alternative, but equivalent, way to characterize the set of variance bounds is to use a slack matrix $\mathbb{S}$. A variance bound is constructed by adding the slack matrix to the variance matrix: $\mathbb{B} = \mathbb{A} + \mathbb{S}$. The resulting variance bound is conservative if and only if $\mathbb{S} = \mathbb{B} - \mathbb{A}$ is positive semidefinite. Thus, the slack captures what we are adding to the variance matrix in order to achieve design compatibility. The set of slack matrices that produces valid variance bounds is $\mathcal{S} = \{\mathbb{S} : 0 \preceq \mathbb{S} \text{ and } s_{k\ell} = -a_{k\ell} \text{ for all } (k, \ell) \in \Omega\}$, where $s_{k\ell}$ is the element in the kth row and ℓth column of $\mathbb{S}$. We can reproduce the set of valid variance bounds as $\mathcal{B} = \{\mathbb{A} + \mathbb{S} : \mathbb{S} \in \mathcal{S}\}$. While the two representations are equivalent, it is often more convenient to work with slack matrices.

4.5 Admissibility

Some valid variance bounds $\mathbb{B} \in \mathcal{B}$ will introduce slack beyond what is required for design compatibility. Such bounds are unnecessarily conservative. Experimenters will typically want to use a variance bound that introduces as little conservativeness, or slack, as possible. The amount of slack introduced will depend on the potential outcomes, so there is no universal ordering of the bounds with respect to conservativeness. Even if there exists no universally best bound, some bounds can be ruled out because they introduce more slack than some other bound no matter what the potential outcomes might be. The following notion of inadmissibility characterizes such bounds.

Definition 3. A variance bound $\mathbb{B} \in \mathcal{B}$ is *inadmissible* if there exists another bound $\mathbb{C} \in \mathcal{B}$ such that $\boldsymbol{\theta}^{\intercal}\mathbb{C}\boldsymbol{\theta} \leq \boldsymbol{\theta}^{\intercal}\mathbb{B}\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \mathbb{R}^K$ and $\boldsymbol{\theta}^{\intercal}\mathbb{C}\boldsymbol{\theta} < \boldsymbol{\theta}^{\intercal}\mathbb{B}\boldsymbol{\theta}$ for at least one $\boldsymbol{\theta} \in \mathbb{R}^K$. Equivalently, $\mathbb{B}$ is inadmissible if there exists a bound $\mathbb{C} \in \mathcal{B}$, distinct from $\mathbb{B}$, such that $\mathbb{C} \preceq \mathbb{B}$. A variance bound that is not inadmissible is said to be *admissible*.

The set of admissible bounds consists exactly of the minimal elements of $\mathcal{B}$ with respect to the Loewner order. Because this is a partial order, there will be many minimal elements, mirroring the fact that there exists no universally best bound. Moreover, there will generally be infinitely many admissible variance bounds.

We say that a procedure for generating variance bounds is admissible if it produces admissible bounds for all input instances. The procedures for deriving variance bounds that we describe in this paper are admissible by construction. However, if a bound is constructed in some other way, it could be inadmissible. In Section S3 of the supplement, we describe a procedure to test whether an arbitrary bound is admissible in the class of quadratic bounds. Experimenters can use this procedure to confirm that the variance estimator they are using is admissible.

Admissibility of variance bounds has not previously been considered in the literature. The previous literature has primarily focused on whether a bound is sharp, meaning that it coincides with the true variance for at least some potential outcomes. As the following definition and proposition show, admissibility is a stronger concept than sharpness.

Definition 4. A variance bound VB: $\mathbb{R}^K \to \mathbb{R}$ is *sharp* if there exists a nonzero vector of potential outcomes $\boldsymbol{\theta} \in \mathbb{R}^K$ such that $\mathrm{VB}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\tau})$.

Proposition 2. *Every admissible variance bound is sharp.*

The procedures described in this paper always yield admissible variance bounds, so the proposition shows that they also are sharp. However, the converse of the proposition is not true; there are sharp bounds that are not admissible. Therefore, it is an open question whether commonly used sharp bounds, such as the Frechet-Hoeffding style bounds described by Robins (1988) and Aronow et al. (2014), are admissible.

4.6 Examples

We can use the formalization of the variance estimation problem described in this section to understand existing variance estimators. Our first example is the variance estimator described by Neyman (1990/1923). In the setting of complete randomization with two equally sized treatment groups, Neyman showed that the variance of the difference-in-means estimator is $\mathrm{Var}(\widehat{\tau}) = (\sigma_1^2 + \sigma_0^2 + 2\rho)/(n-1)$, where σ_1^2 and σ_0^2 are the population variances of the potential outcomes under treatment and control, respectively, among all units in the experiment, and ρ is the covariance between the two potential outcomes. The covariance is not estimable and must be bounded. Neyman's solution was to use the Cauchy–Schwartz inequality followed by the AM-GM inequality on ρ to obtain the upper bound $\mathrm{Var}(\widehat{\tau}) \leq 2(\sigma_1^2 + \sigma_0^2)/(n-1)$. This upper bound can be estimated by the sample variances corresponding to σ_1^2 and σ_0^2.

The Neyman bound can be rewritten in our framework. As above, the variance of the estimator can be written $n^{-2}\boldsymbol{\theta}^{\mathsf{T}}\mathbb{A}\boldsymbol{\theta}$, where the covariance matrix is

$$\mathbb{A} = \mathrm{Cov}(\boldsymbol{V}) = \frac{n}{n-1}\begin{bmatrix} \mathbb{H} & \mathbb{H} \\ \mathbb{H} & \mathbb{H} \end{bmatrix} \qquad \text{and} \qquad \mathbb{H} = \mathbb{I} - \mathbf{1}\mathbf{1}^{\mathsf{T}}/n,$$

and the potential outcome vector is $\boldsymbol{\theta} = \big(y_1(1), \ldots, y_n(1), y_1(0), \ldots, y_n(0)\big)$.

The matrix $\mathbb{A}$ is not design compatible because the diagonal elements in the off-diagonal blocks are nonzero, but the corresponding pairs of potential outcomes are never simultaneously observed. For example, $(1, n+1) \in \Omega$, so the product $\theta_1\theta_{n+1} = y_1(1)y_1(0)$ is never observed, but entry in row 1 and column $n+1$ of $\mathbb{A}$ is one: $a_{1,n+1} = 1$. To address this, the Neyman variance estimator implicitly uses the slack matrix

$$\mathbb{S} = \frac{n}{n-1}\begin{bmatrix} \mathbb{H} & -\mathbb{H} \\ -\mathbb{H} & \mathbb{H} \end{bmatrix}, \quad \text{yielding the variance bound} \quad \mathbb{B} = \frac{2n}{n-1}\begin{bmatrix} \mathbb{H} & 0 \\ 0 & \mathbb{H} \end{bmatrix}.$$

This bound is a valid because $\mathbb{S}$ is positive semidefinite and $s_{k\ell} = -a_{k\ell}$ for all $(k, \ell) \in \Omega$.

Our second example is the class of variance estimators described by Aronow and Samii (2013, 2017). The authors consider variance estimation for the Horvitz–Thompson point estimator under arbitrary experimental designs, and they describe a bound based on Young's inequality for products. The most straightforward version of Young's inequality states that $2xy \leq x^2 + y^2$ for any two real numbers x and y. Recall that the central problem is that the variance expression contains terms $a_{k\ell}\theta_k\theta_\ell$ such that $\theta_k\theta_\ell$ is unobservable and $a_{k\ell}$ is not zero. To address this, the Aronow–Samii bound apply Young's inequality separately on each of these problematic terms: $a_{k\ell}\theta_k\theta_\ell \leq |a_{k\ell}|\big(\theta_k^2 + \theta_\ell^2\big)/2$.

We can use the quadratic form representation to write the Aronow–Samii bound as a slack matrix. Let $\mathbb{M}_{k\ell}$ be a $K \times K$ matrix with zeros entries except in the (k, ℓ)th block, which instead is given by $|a_{k\ell}|$ in the diagonal entries (k, k) and (ℓ, ℓ), and $-a_{k\ell}$ in the off-diagonal entries (k, ℓ) and (ℓ, k). The slack matrix corresponding to the Aronow–Samii bound is $\mathbb{S} = \sum_{(k,\ell)\in\Omega} \mathbb{M}_{k\ell}/2$. This bound is design compatible because $s_{k\ell} = -a_{k\ell}$ by construction for all $(k, \ell) \in \Omega$. Furthermore, because all matrices $\mathbb{M}_{k\ell}$ are positive semidefinite, their sum $\mathbb{S}$ will also be positive semidefinite. Hence, the bound is conservative. However, as the following proposition shows, the bound is not admissible.

Proposition 3. *The Aronow–Samii bounding procedure is inadmissible in the class of quadratic* *bounds.*

The bound B_2 in the illustration in Section 2 is the Aronow–Samii bound, so the fact that bound B_1 dominates B_2 proves the proposition. It is possible to construct similar examples in more involved settings, including with larger sample and more intricate designs, but we omit those in the interest of space.

5 Constructing Variance Bounds

5.1 Variance Bound Programs

There is currently no method that allows experimenters to construct a variance estimators that minimize conservativeness. Indeed, there exists no method to even construct admissible variance bounds for general exposure mappings, designs and estimators. To address this, we describe a computational approach that selects a variance bound from $\mathcal{B}$ using an optimization formulation. For some real-valued function g on symmetric matrices, we aim to find a slack matrix $\mathbb{S} \in \mathcal{S}$ that minimizes g. We refer to this procedure as OPT-VB, which is the following mathematical program:

$$\mathbb{S}^* \in \operatorname*{arg\,min}_{\mathbb{S} \in \mathcal{S}} g(\mathbb{S}). \tag{OPT-VB}$$

The properties of a variance bound constructed in this way and the associated variance estimator will depend on the choice of objective function g. The ideal objective function is $g(\mathbb{S}) = \boldsymbol{\theta}^{\mathsf{T}} \mathbb{S} \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ refers to the true potential outcomes, because then the objective captures the actual conservativeness in the current experiment. But such an objective is infeasible, because it requires exact knowledge of the potential outcomes. Instead, the approach we explore in this paper is to encode in g the experimenter's preferences concerning risk trade-offs and any background knowledge they might have about the potential outcomes.

Unless otherwise noted, all objective functions discussed in this paper are such that they strictly penalize matrices that are weakly larger in the Loewner order, which we refer to as strict monotonicity. Monotonicity ensures that the variance bound produced by OPT-VB using the objective is admissible. This result was previewed in Section 2, and the following definition and proposition present the formal result in full.

Definition 5. A real-valued function g on symmetric matrices is *strictly monotone* if $g(\mathbb{Q}) < g(\mathbb{P})$ whenever $\mathbb{Q} \preceq \mathbb{P}$ and $\mathbb{Q} \neq \mathbb{P}$.

Theorem 1. *If the input objective function g is strictly monotone, then* OPT-VB *returns a variance bound that is conservative, design compatible and admissible.*

Definition 5 differs from the conventional definition of strict monotonicity based on the strict Loewner order. The conventional definition states that a strictly monotone function f satisfies $f(\mathbb{Q}) < f(\mathbb{P})$ whenever $\mathbb{Q} \prec \mathbb{P}$. This definition, though well-motivated in many applications, does not align with our notion of admissibility, necessitating us to extend it slightly.

The OPT-VB program is generally computationally tractable. As a rule of thumb, an optimization program is tractable if it is convex (Boyd & Vandenberghe, 2004; Rockafellar, 1993). The set of slack matrices $\mathcal{S}$ is convex, implying that OPT-VB is a convex problem if g is a convex function. All objective functions considered in this paper are convex, so they admit efficient algorithms for finding optimal solutions, up to desired tolerances.

5.2 Norm Objectives

We will first consider when an experimenter has little or no background knowledge about the potential outcomes. Our goal here is to select a variance bound that is not excessively conservative for most potential outcomes. This corresponds to selecting a quadratic form of small magnitude, as measured by a matrix norm of its coefficient matrix. We will use the family of Schatten p-norms for matrices to make this idea precise.

There is an implicity trade-off between average performance and worst-case performance when selecting a matrix norm, corresponding to the experimenter's risk preference concerning conservativeness. To understand how the Schatten norm captures this risk trade-off, consider the spectral decomposition of the coefficient matrix of a bound $\mathbb{B} \in \mathcal{B}$. Because the coefficient matrix is symmetric, we can write it as $\mathbb{B} = \sum_{k=1}^{K} \lambda_k \boldsymbol{\eta}_k \boldsymbol{\eta}_k^\intercal$, where $\boldsymbol{\eta}_k$ is the kth eigenvector of $\mathbb{B}$ and λ_k is the corresponding kth eigenvalue. This allows us to write a variance bound as

$$\mathrm{VB}(\boldsymbol{\theta}) = \frac{1}{n^2} \boldsymbol{\theta}^\intercal \mathbb{B} \boldsymbol{\theta} = \frac{\|\boldsymbol{\theta}\|^2}{n^2} \sum_{k=1}^{K} w_k \lambda_k,$$

where $w_k = \langle \boldsymbol{\theta}, \boldsymbol{\eta}_k \rangle^2 / \|\boldsymbol{\theta}\|^2$ captures the alignment of the potential outcome vector to the kth eigenvector $\boldsymbol{\eta}_k$. Because $\mathbb{B}$ is positive semidefinite, all eigenvalues are non-negative. By construction, the coefficients w_k are non-negative and sum to one, so they act as weights in a convex combination of the eigenvalues. The conservativeness of the variance bound is therefore determined by the eigenvalues and the alignment of the potential outcomes to the eigenvectors of $\mathbb{B}$. If we make the eigenvalues of the variance bound matrix $\mathbb{B}$ small, we ensure that the bound is not excessively conservative.

The Schatten norms are different ways of measuring the magnitude of the eigenvalues. Formally, a Schatten p-norm of $\mathbb{B}$ is the usual p-norm applied to the vector of singular values of $\mathbb{B}$, which in our case coincide with the eigenvalues:

$$\|\mathbb{B}\|_p = \left(\sum_{k=1}^{K} |\lambda_k|^p \right)^{1/p}.$$

When p is small, the norm tolerates a few large eigenvalues if it means that many other eigenvalues are small. When p is large, the norm is disproportionately affected by large eigenvalues, diminishing the influence of smaller eigenvalues. Therefore, a risk averse experimenter would want to use a Schatten p-norm with a large p, because minimizing such a norm ensures that no eigenvalue is much larger than the others. A risk tolerant experimenter would instead prefer a Schatten p-norm with a smaller p, as this will ensure that the sum of the eigenvalues is small. The following proposition shows that we achieve admissibility no matter the choice of p.

Theorem 2. *For all $p \in [1, \infty)$, the Schatten p-norm objective $g(\mathbb{S}) = \|\mathbb{A} + \mathbb{S}\|_p$ is strictly monotone, ensuring that the variance bound produced by* OPT-VB *using g is admissible.*

The Schatten p-norm coincides with some more familiar matrix norms for particular values of p. If we set $p = 1$, the Schatten p-norm is simply the sum of the absolute values of the eigenvalues. This is the nuclear norm, which also is called the trace norm. Using this norm produces a bound with the best average performance, in the sense that it puts uniform weight on all eigenvalues no matter their magnitude.

At the other extreme, when we let $p \to \infty$, we obtain the operator norm induced by the 2-norm, which in our setting coincides with the maximum eigenvalue of $\mathbb{B}$. An experimenter who is maximally risk adverse would use this norm, as it would trade-off any amount of average conservativeness for even a minute reduction is worst-case conservativeness. The operator norm is not strictly monotone according to Definition 5, so an arbitrary minimizer of an objective function using this norm is not guaranteed to be admissible. This can be addressed by using a large but not infinite p-norm, which will behave like the operator norm for practical purposes. Alternatively, we describe a regularization procedure of the operator norm in Section S4 of the supplement that ensures admissibility.

Finally, we recover the Frobenius norm when $p = 2$. This norm provides an intermediate point in the risk trade-off; it disproportionately penalizes large eigenvalues, making sure that no eigenvalue gets very large, but it does not ignore the smaller eigenvalues completely.

5.3 Targeted Linear Objectives

The norm objectives in the previous subsection cannot encode background knowledge experimenters might have about the potential outcomes. We describe a class of targeted objective functions to fill this role. The prior knowledge the experimenter encodes in the objective function need not be correct, not even approximately, to ensure the validity and admissibility of the resulting variance bound. But if they are able to provide reasonably accurate information, the bound will be less conservative. This idea is related to the model-assisted tradition that originated in the literature on design-based survey sampling (see, e.g., Särndal, Swensson, & Wretman, 1992, and Basse & Airoldi, 2018).

The class of *targeted linear objectives* takes the form

$$g(\mathbb{S}) = \langle \mathbb{S}, \mathbb{W} \rangle,$$

where $\mathbb{S}$ is a slack matrix, $\mathbb{W}$ is a targeting matrix of the same dimensions, and $\langle \cdot, \cdot \rangle$ denotes the trace inner product on matrices: $\langle \mathbb{S}, \mathbb{W} \rangle = \operatorname{tr}(\mathbb{S}\mathbb{W})$. As we discuss in the next section, $\mathbb{W}$ is used to target particular potential outcomes, motivated by prior substantive knowledge. All objective functions in this class are linear in the coefficients of the slack matrix, so the optimization problem underlying OPT-VB becomes a semidefinite program, ensuring computational tractability.

By construction, the bound returned by OPT-VB using a targeted linear objective will be valid. What makes the class of targeted linear objectives stand out compared to the norm objectives is a type of completeness result. Namely, the class of targeted linear objectives completely characterizes the set of all admissible variance bounds.

Theorem 3. *A bound $\mathbb{B}$ is admissible if and only if it can be obtained from* OPT-VB *using the objective function $g(\mathbb{S}) = \langle \mathbb{S}, \mathbb{W} \rangle$ for some positive definite targeting matrix $\mathbb{W}$.*

The proof that every bound returned by OPT-VB using a positive definite targeting matrix is admissible proceeds by showing that every targeted linear objective is strictly monotone and then appeals to Theorem 1. The proof of the opposite direction, that every admissible bound can be obtained as a solution to OPT-VB using some targeted linear objective, is more involved and appeals to the separating hyperplane theorem from convex analysis. The proof is provided in the supplement.

Theorem 3 shows that we always obtain an admissible bound when we use a targeted linear objective with a positive definite targeting matrix. Furthermore, due to the one-to-one correspondence between admissible bounds and targeted linear objectives, the theorem allows us to re-interpret other procedures for constructing variance bounds by showing what matrix they implicitly target, which by extension shows what potential outcomes they implicitly target.

5.4 Choosing Targeting Matrices

Recall that the variance bound using coefficients $\mathbb{B} = \mathbb{A} + \mathbb{S}$ is

$$n^2 \operatorname{VB}(\boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathsf{T}}\mathbb{B}\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathsf{T}}\mathbb{A}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathsf{T}}\mathbb{S}\boldsymbol{\theta}.$$

If the true potential outcomes were known, the experimenter would use the targeting matrix $\mathbb{W} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}$, because it directly targets the conservativeness of the bound: $\langle \mathbb{S}, \mathbb{W} \rangle = \boldsymbol{\theta}^{\mathsf{T}}\mathbb{S}\boldsymbol{\theta}$. The challenge here is, of course, that the potential outcomes are unknown.

Suppose the experimenter has some prior, partial knowledge about the potential outcomes, and they encode that knowledge in a generative model. That is, we would consider $\boldsymbol{\theta}$ as a random variable drawn from some known distribution. Seen from this perspective, the value of the variance bound is random, because the randomness of $\boldsymbol{\theta}$ is passed on to $\operatorname{VB}(\boldsymbol{\theta})$. A natural target in this setting is to minimize the expectation of the variance bound $\operatorname{VB}(\boldsymbol{\theta})$ with respect to the stipulated generative model.

We use a subscripted expectation operator $\operatorname{E}_{\boldsymbol{\theta}}[\cdot]$ to denote the expectation with respect to the imagined distribution of $\boldsymbol{\theta}$, rather than the true randomization distribution induced by the experimental design, as in the rest of the paper. The expected value of a variance bound given by $\mathbb{B} = \mathbb{A} + \mathbb{S}$ is then

$$n^2 \operatorname{E}_{\boldsymbol{\theta}}[\operatorname{VB}(\boldsymbol{\theta})] = \big\langle \mathbb{B}, \operatorname{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}] \big\rangle = \langle \mathbb{A}, \operatorname{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}] \rangle + \langle \mathbb{S}, \operatorname{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}] \rangle.$$

To take advantage of prior knowledge about the potential outcomes, the experimenter would use a targeted linear objective with matrix $\mathbb{W} = \operatorname{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}]$, because this directly minimizes the expected variance bound under the stipulated generative model.

It should be emphasized that the interpretation of $\boldsymbol{\theta}$ as a random variable is simply a convenient way to express prior (partial) knowledge about the potential outcomes. It is not assumed nor required for any of our results that the stipulated generative model accurately reflects how the potential outcomes actually were generated. The resulting bound is valid no matter what distribution one uses for $\boldsymbol{\theta}$, and the resulting bound is admissible as long as $\mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\intercal}]$ is positive definite. However, the bound will be less conservative if the distribution is a good approximation of the true potential outcomes.

Experimenters should take care to ensure that $\mathbb{W} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\intercal}]$ indeed is positive definite. Targeting matrices that are not positive definite disregard some dimensions of the potential outcome vector space, meaning that they do not penalize excessive conservativeness in those dimensions. A simple way to ensure that a targeting matrix is positive definite is to include independent noise in the generative model, as in the following example.

To illustrate how a generative model could be used to construct a targeting matrix, consider when the experimenter knows, or presumes to know, that the potential outcomes can be well-approximated by a linear function of some set of covariates. For simplicity, we consider when there are two exposures, e_1 and e_0, so the potential outcome vector is $\boldsymbol{\theta} = \big(y_1(e_1), \dots, y_n(e_0)\big)$. Letting $\mathbb{X}$ be a n-by-m matrix collecting m covariates for the n units, the generative model for the potential outcomes could be written as $\boldsymbol{\theta} = (\mathbb{X}\boldsymbol{\beta}_{e_1} + \boldsymbol{\varepsilon}_{e_1}, \mathbb{X}\boldsymbol{\beta}_{e_0} + \boldsymbol{\varepsilon}_{e_0})$, where $\boldsymbol{\beta}_{e_1}$ and $\boldsymbol{\beta}_{e_0}$ are coefficient vectors describing how the covariates relate to the potential outcomes, and $\boldsymbol{\varepsilon}_{e_1}$ and $\boldsymbol{\varepsilon}_{e_0}$ describe aspects of the potential outcomes not captured by the covariates.

The covariate matrix $\mathbb{X}$ is observed and fixed, but we might not have a good sense of $(\boldsymbol{\beta}_{e_1}, \boldsymbol{\beta}_{e_0}, \boldsymbol{\varepsilon}_{e_1}, \boldsymbol{\varepsilon}_{e_0})$. We can express our ignorance about these vectors as a distribution. For illustration here, we will consider when we have a good sense of $\boldsymbol{\beta}_{e_1}$ and $\boldsymbol{\beta}_{e_0}$, so they are non-random vectors, and the coordinates of $(\boldsymbol{\varepsilon}_{e_1}, \boldsymbol{\varepsilon}_{e_0})$ are independent and follow a standard normal distribution. With this generating model, the targeting matrix becomes

$$\mathbb{W} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\intercal}] = \begin{bmatrix} \mathbb{X} & 0 \\ 0 & \mathbb{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{e_1} \\ \boldsymbol{\beta}_{e_0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{e_1}^{\intercal} & \boldsymbol{\beta}_{e_0}^{\intercal} \end{bmatrix} \begin{bmatrix} \mathbb{X}^{\intercal} & 0 \\ 0 & \mathbb{X}^{\intercal} \end{bmatrix} + \mathbb{I}.$$

More intricate generative working models generate more elaborate targeting matrices.

5.5 Composite Objectives

There are situations where experimenters want a combination of properties offered by different objective functions. Using the fact that monotonicity is maintained under positive combinations, the following proposition shows that a combination of elementary objectives can be used with OPT-VB.

Proposition 4. *If g is strictly monotone and h is monotone, then the function $g+\gamma h$ is strictly monotone for any $\gamma \geq 0$.*

One situation in which a composite objective is useful is when an experimenter wants to regularize a targeted linear objective, perhaps because they are not very confident in the information encoded in the targeting matrix. They can then use a composite objective that includes one of the norm objectives discussed in Section 5.2. For some Schatten p-norm and coefficient $\gamma > 0$, deciding the relative focus on the two objectives, the composite objective function is

$$g(\mathbb{S}) = \langle \mathbb{S}, \mathbb{W} \rangle + \gamma \|\mathbb{A} + \mathbb{S}\|_p.$$

If $p \in [1, \infty)$, this composite objective is strictly monotone even if $\mathbb{W}$ is not full rank, so the composite objective always yields a bound that is conservative, estimable and admissible.

6 Estimating Variance Bounds

6.1 Precision of Variance Bound Estimator

A quadratic form can be reinterpreted as a linear function of the elements $\theta_k\theta_\ell$ of the outer product $\boldsymbol{\theta}\boldsymbol{\theta}^\intercal$. This means that we can use any estimator in the class of linear estimators to estimate a quadratic variance bound itself once it has been derived, yielding a conservative variance estimator. It is beyond the scope of this paper to investigate which of these estimators is best suited for estimation of quadratic forms. In this section, we will instead consider how the choice of the variance bound itself influences the estimation task.

We restrict our focus to the Horvitz–Thompson estimator of the bounds. This estimator is sufficiently simple so as to not distract from the main ideas and insights we aim to explore. For a variance bound $\mathbb{B}$, the corresponding estimator is

$$\widehat{\mathrm{VB}}(\boldsymbol{\theta}) = \frac{1}{n^2} \sum_{k \in S} \sum_{\ell \in S} \frac{b_{k\ell} \theta_k \theta_\ell}{\Pr(k, \ell \in S)},$$

where, as above, $\Pr(i, j \in S)$ is the probability of simultaneously observing potential outcomes θ_k and θ_ℓ.

The Horvitz–Thompson estimator is unbiased whenever the variance bound is design compatible. However, unbiasedness does not ensure that the estimator is precise. While the precision of the estimator critically depends on the potential outcomes and the experimental design, the experimenter's choice of variance bound also plays a part. To explore this, we define

$$R_{k\ell} = \frac{\mathbb{1}[k, \ell \in S] \times \mathbb{1}[b_{k\ell} \neq 0]}{\Pr(k, \ell \in S)}$$

to be a random variable for each pair $(k, \ell) \in P \times P$, capturing the inverse propensity weighting done by the estimator. We define the ratio of zero and zero to be zero, meaning that $R_{k\ell} = 0$ if potential outcomes $k, \ell \in P$ are never observed simultaneously. Collecting the K^2 variables $R_{k\ell}$ in a vector $\boldsymbol{R}$, we use the covariance matrix $\mathrm{Cov}(\boldsymbol{R})$ to characterize the precision of the variance bound estimator.

Proposition 5. *If the variance bound $\mathbb{B}$ is design compatible, the normalized mean squared error of the Horvitz–Thompson estimator of the variance bound is bounded as*

$$\mathrm{E}\Big[\big(n\widehat{\mathrm{VB}}(\boldsymbol{\theta}) - n\,\mathrm{VB}(\boldsymbol{\theta})\big)^2\Big] \leq \frac{1}{n^2} \|\mathrm{Cov}(\boldsymbol{R})\|_\infty \times \|\boldsymbol{\theta}\|_\infty^2 \times \|\mathbb{B}\|_2^2,$$

where $\|\mathrm{Cov}(\boldsymbol{R})\|_\infty$ is the operator norm of the covariance matrix of the inverse propensity variables, $\|\boldsymbol{\theta}\|_\infty = \max_{k \in P} |\theta_k|$ is the largest magnitude of the potential outcomes, and $\|\mathbb{B}\|_2^2$ is the squared Frobenius norm of coefficient matrix of the bound.

The proposition allows us to consider the design, potential outcomes and variance bound separately when building understanding of the behavior of variance bound estimators. The factor $\|\mathrm{Cov}(\boldsymbol{R})\|_\infty$ captures aspects of the experimental design. This norm will be small for designs that do not induce too much dependence between the exposures. A design that induces highly correlated exposures might make precise estimation impossible even in large samples, which would be reflected in a large $\|\mathrm{Cov}(\boldsymbol{R})\|_\infty$.

Importantly, while properties of the design that facilitate precise point estimation generally coincide with those that facilitate precise variance estimation, they are not exactly the same. In particular, the performance of the point estimator is governed by first-order exposure probabilities, but the construction of $\boldsymbol{R}$ uses second-order probabilities. It is possible that a design makes the first-order probabilities well-behaved but still have many second-order probabilities being close to zero. In such cases, experimenters should consider extending Definition 1 so that design compatibility requires $\Pr(k, \ell \in S) \geq c$ for some constant $c > 0$, rather than just not being zero. This will make the bound more conservative, but one would ensure that $\|\mathrm{Cov}(\boldsymbol{R})\|_\infty$ is well-controlled. For the purpose of this section, we will proceed under the presumption that experimenters have taken the steps necessary to ensure that $\|\mathrm{Cov}(\boldsymbol{R})\|_\infty$ is well-controlled.

The factor $\|\boldsymbol{\theta}\|_\infty$ captures the scale of the potential outcomes. If the potential outcomes are large in magnitude, the estimator will naturally be less precise in absolute terms. For simplicity, we use the uniform norm to measure the scale of the potential outcomes; if the potential outcomes are known to be in some interval, this norm is asymptotically bounded by construction. However, the uniform norm can paint an overly pessimistic picture, and we show in Section S5.8 in the supplement that it is possible to replace the uniform norm with the fourth moment, at the cost of making the mean square error bound more sensitive to outliers among the coefficients in the variance bound $\mathbb{B}$.

The final factor $\|\mathbb{B}\|_2^2$ measures the magnitude of the coefficients in the variance bound. If $\|\mathbb{B}\|_2^2$ is large relative to $\|\mathbb{A}\|_2^2$, then the bound is achieving design compatibility by over-weighting a subset of the potential outcome products, making the variance bound estimators disproportionally sensitive to estimation errors in those terms. The following corollary, which follows directly from Proposition 5, states that control over $\|\mathbb{B}\|_2^2$ ensure that the error of the variance bound estimator is small with high probability in large samples.

Corollary 1. *If $\|\operatorname{Cov}(\boldsymbol{R})\|_\infty$ and $\|\boldsymbol{\theta}\|_\infty$ are asymptotically bounded and $\|\mathbb{B}\|_2^2$ is dominated by n^2, then the variance bound estimator is consistent: $n\widehat{\mathrm{VB}}(\boldsymbol{\theta}) - n\,\mathrm{VB}(\boldsymbol{\theta}) = o_p(1)$.*

The corollary suggests that a useful heuristic to improve precision of the variance bound estimator is to make $\|\mathbb{B}\|_2^2$ small. A way to achieve this is to use the Frobenius norm as the objective in OPT-VB, as discussed in Section 5.2. This will not ensure consistency, as there may be no valid bound with sufficiently small norm. In cases where the potential outcomes are not bounded, experimenters should consider using a Schatten p-norm for some $p > 2$ to account for the fact that $\|\boldsymbol{\theta}\|_\infty$ might not be well-controlled.

6.2 Accuracy With Respect to the True Variance

The previous subsection considered the precision of the variance bound estimator with respect to the variance bound itself. This does not account for the fact that the variance bound potentially could be very conservative, in which case the variance bound estimator would give a misleading picture of the precision of the point estimator even if itself is precise. We can address this using a composite objective, as discussed in Section 5.5.

Consider the normalized mean square error of the variance bound estimator $\widehat{\mathrm{VB}}(\boldsymbol{\theta})$ with respect to the true variance $\operatorname{Var}(\widehat{\tau})$. Using the usual bias–variance decomposition, we can write the error as

$$\mathrm{E}\Big[\big(n\widehat{\mathrm{VB}}(\boldsymbol{\theta}) - n\operatorname{Var}(\widehat{\tau})\big)^2\Big] = \big(n\,\mathrm{VB}(\boldsymbol{\theta}) - n\operatorname{Var}(\widehat{\tau})\big)^2 + \mathrm{E}\Big[\big(n\widehat{\mathrm{VB}}(\boldsymbol{\theta}) - n\,\mathrm{VB}(\boldsymbol{\theta})\big)^2\Big].$$

The first term is the slack introduced to make the variance bound design compatible. This was the focus of the investigation in Section 5. For example, if we are following the model-assisted approach described in Section 5.4, we would use the inner product of $\mathbb{W} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^{\intercal}]$ and $\mathbb{B}$ as a proxy for this first term. The second term is the precision of the variance bound estimator with respect the bound itself, which was the focus of the previous subsection. We could use the bound from Proposition 5 as proxy for this second term. This leads to the following composite objective as a heuristic for the mean square error of the variance estimator:

$$g(\mathbb{S}) = \langle \mathbb{S}, \mathbb{W} \rangle + \gamma \|\mathbb{A} + \mathbb{S}\|_2^2, \qquad \text{where} \qquad \gamma = \|\mathrm{Cov}(\boldsymbol{R})\|_\infty \times \|\boldsymbol{\theta}\|_\infty.$$

It is not generally obvious what the appropriate relative weighting of the two terms is in this objective, and γ effectively functions as a tuning parameter for the composite objective. However, the bound will be valid and admissible as long as $\gamma > 0$.

7 Numerical Illustration

Our simulation exercise uses data collected by Paluck et al. (2016) from a randomized network experiment involving 24,183 students in 56 public middle schools in New Jersey. The purpose of the study was to investigate the effectiveness of various interventions aimed at reducing conflict and bullying among adolescents. In particular, randomly selected students were given an experimental intervention expected to have spill-over effects on the perceived social norms and behavior of their peers. In addition to treatment status and outcome data, the authors recorded the self-reported social network among students as well as numerous covariates. Our goal here is not to perform a re-analysis of this study, but rather to use the original data to construct a set of empirical settings that reasonably reflect a real-world study.

The experimental design used by Paluck et al. (2016) is a two stage randomization process. A set of "seed students," who were well-connected in the social network, was selected in each school. The 56 schools are grouped into 14 blocks and a random half of the schools in each block were selected to receive treatment. Among the schools selected to receive treatment, a random half of the seed students were chosen to receive the intervention.

The exposure mapping is defined by a tuple of binary variables (s_i, z_i, a_i), where s_i indicates whether the school was selected for treatment, z_i indicates whether the student received the intervention, and a_i indicates whether at least one of the student's peers in the network received the intervention. In our simulations, we focus on estimating the direct effect of the intervention, corresponding to the contrast between exposures $e_1 = (1, 1, 0)$ and $e_0 = (0, 0, 0)$. Because only seed students can receive the exposure $e_1 = (1, 1, 0)$ under the design, the effect cannot be unbiasedly estimated across all subjects. Instead, we narrow the subjects of interest to the subset of 2,170 seed students who receive each exposure with probability at least 0.5%.

We consider two types of potential outcomes. In the first setting, the outcomes are generated as a linear function of a set of observed covariates $\boldsymbol{\theta} = (\mathbb{X}\boldsymbol{\beta}_{e_1} + \boldsymbol{\varepsilon}_{e_1}, \mathbb{X}\boldsymbol{\beta}_{e_0} + \boldsymbol{\varepsilon}_{e_0})$, in line with the generative model discussed in Section 5.4. The included covariates are age, height, weight, gender, and grade. We use $\boldsymbol{\beta}_{e_1} = 5/4 \times \mathbf{1}$ and $\boldsymbol{\beta}_{e_0} = 3/4 \times \mathbf{1}$ for the coefficients, and $\boldsymbol{\varepsilon}_{e_1}$ and $\boldsymbol{\varepsilon}_{e_0}$ are independent standard normal. In the second setting, we use reported outcomes from the original study: adoption of an anti-bullying wristband and school-reported disciplinary actions. Each potential outcome in the simulation is one of these reported outcomes, mixing the two type of reported outcomes in the same setting. Hence, this exercise does not capture a causal effect in the original study, but the apporach allows us to use the reported data unaltered, and it retains any peculiarities of the outcome distributions.

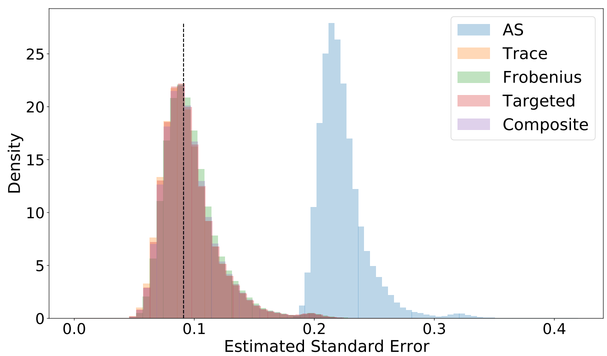
We use the Horvitz–Thompson estimator to estimate the direct effect on the 2,170 students satisfying the first-order positivity condition. We construct the variance matrix $\mathbb{A}$ using a Monte Carlo with 5 million replicates from the experimental design. We define the set of unobservable products Ω as the unit-exposure tuples that are observed with probability less than 0.1%.

We examine variance bounds produced by several objective functions described in the paper: (i) the trace norm, (ii) the Frobenius norm, (iii) a targeted linear objective, and (iv) a composite objective. We also examine the Aronow–Samii bound, which is the only existing bound that can be used in this setting. The matrix used in the targeted linear objective is constructed as described in Section 5.4 with $\boldsymbol{\beta}_{e_1} = \boldsymbol{\beta}_{e_0} = \mathbf{1}$, using the same covariates as above. The targeted linear objective is therefore nearly correctly specified when the outcomes are linearly generated, but likely misspecified for the second setting. The composite objective is the targeted linear objective with Frobenius penalty as described in Section 6.2 with $\gamma = 1$. In order to solve the OPT-VB program, we use JuMP modeling software (Dunning, Huchette, & Lubin, 2017) and the SCS solver (O'Donoghue, Chu, Parikh, & Boyd, 2023). We use the Horvitz–Thompson estimator of the bounds as described in Section 6. We construct Wald-type 95% confidence intervals using the square root of the variance estimator as an estimate for the standard error of the point estimator.

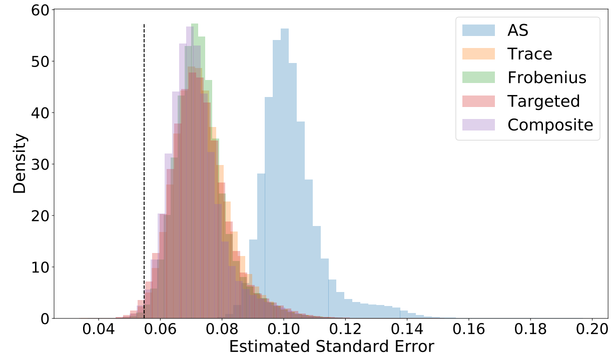
Table 1: Simulation results

	Panel A: Synthetic outcomes				Panel B: Real data outcomes			
	Bias	Precision	Coverage	Width	Bias	Precision	Coverage	Width
AS	3.129	0.862	0.999	0.368	3.869	0.832	1.000	0.418
Trace	1.167	0.407	0.975	0.223	2.196	0.590	0.994	0.314
Frobenius	1.099	0.420	0.961	0.216	1.894	0.409	0.994	0.293
Targeted	1.080	0.470	0.968	0.213	1.858	0.561	0.992	0.288
Composite	1.070	0.451	0.959	0.213	1.771	0.442	0.993	0.282

The simulation results are presented in Table 1, with one panel for each of the two outcomes. The first column in each panel presents the bias of each variance estimator relative to the true variance. We find that the Aronow–Samii variance estimator is more than two to three times larger than the true variance. The four variance estimators based on optimized bounds inflates the variance to a considerably lesser degree, with a relative increase of between 4.3% and 16.7%. The targeted linear objectives (with and without penalty) have the smallest bias in this setting, reflecting the fact that the covariates are informative of the potential outcomes. The next column presents the precision of the variance estimators in the form of their standard errors relative to the standard error of the Aronow–Samii variance estimator. We find that all variance estimators based on optimized bounds have better precision, showing that we do not trade-off bias for imprecision when using the optimized bounds.



(a) Synthetic Outcomes



(b) Real Data Outcomes

Figure 2: Sampling Distributions of Standard Error Estimators

The third column in each panel presents coverage rates for confidence intervals at the 95% nominal level. The variance estimators are conservative by construction, so they all overcover, as expected. However, confidence intervals based on the Aronow–Samii variance estimator stand out by severely overcovering. The last column presents average width of confidence intervals constructed based on each variance estimator. As expected from the reduction in bias, we find that intervals based on optimized bounds are markedly narrower, making the confidence intervals more informative and allowing us to draw sharper inferences.

Figure 2 presents the sampling distributions of the five standard error estimators, obtained from taking the square root of the variance estimators. The dotted black line indicates the true standard error of the point estimator. Across both outcomes, we see that the estimator derived from the Aronow–Samii bound is much further to the right of the estimators derived from OPT-VB. These histograms indicate that the standard error estimates derived from OPT-VB will be smaller than those derived from the Aronow–Samii bound with large probability. This corroborates the statistics in Table 1 which show that the relative bias of the variance estimator and expected width of the interval are reduced using OPT-VB.

The appendix contains several additional simulation results. We investigate several additional outcomes, and we vary the cut-off parameter which determines the set of unobservable products. These additional results are in line with those reported here. double check!!!!

8 Concluding Remarks

Variance estimation for treatment effect estimators is a balancing act. Unbiased and consistent estimators generally do not exist. Experimenters therefore opt for conservative estimators to avoid misleading inferences, but they want to avoid excessive conservativeness. The methods we have described in this paper allow experimenters to construct valid variance estimators that minimize conservativeness. Experimenters can take advantage of background information about the potential outcome to reduce conservativeness by using a targeted linear objective. In case no such information is available, experimenters can use a norm objective to reduce conservativeness for all potential outcomes. No matter the approach, the resulting estimator is guaranteed to be conservative and admissible, even if the experimenters perceived knowledge of the potential outcomes happens to be incorrect.

There are several extensions and open questions that are yet to be explored. We have considered the class of linear point estimators in this paper, and it remains an open question whether our results and methods can be extended to a larger class of estimators. While the class of linear estimators includes almost all conventional treatment effect estimators, recently developed estimators based on machine learning techniques do not fall in this class (see, e.g., Aronow & Middleton, 2013; Chernozhukov et al., 2018; Wager, Du, Taylor, & Tibshirani, 2016; Wu & Gagnon-Bartsch, 2018). The key challenge is that the variance of these estimators are not quadratic forms in the potential outcome vector, so bounds that themselves are quadratic forms will generally not be valid. A possible way forward is to linearize the point estimators, in which case one could construct quadratic bounds that are asymptotically valid. However, extending the finite-sample results in the current paper to this larger class of treatment effect estimators appears to currently be beyond reach.

Relatedly, it remains an open question if the ideas explored in this paper can be extended to a larger class of bounds. Motivated by the fact that the variance itself is a quadratic form, we considered bounds that are quadratic forms. It is possible that tighter bounds exist in a larger class of bounds. One possible route to explore is whether one can construct a class bounds for the general setting inspired by the Fréchet–Hoeffding-type bounds for the difference-in-means estimator under complete randomization mentioned in Section 3. This bound is sharp for comonotonic potential outcomes, which is a fairly large set of potential outcomes. However, it is not currently known how this type of bound trades off the slack in the full set of potential outcome vectors, and it remains to be investigated whether it dominates the class admissible quadratic bounds.

Our discussion about how to select a variance bound to minimize the mean squared error of the variance estimator in Section 6 was based on a bound on the precision of the estimator. This bound will occasionally be loose, so the approach we describe in this paper is best seen as a heuristic. While we believe this heuristic is useful and appropriate in most circumstances, it remains an open question whether one can select the variance bound so as to directly minimize mean squared error.

Finally, our investigation relies on the assumption that the exposures are correctly specified. The interference literature has recently considered estimation of exposure effects when the exposures are misspecified or nonparametric (see, e.g., Auerbach & Tabord-Meehan, 2023; Leung, 2022; Li & Wager, 2022; Sävje, 2024). It is an open question if our results and methods extend to settings with misspecified exposures.

References

Abadie, A., & Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, *91/92*, 175–187.

Aronow, P. M., Green, D. P., & Lee, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Annals of Statistics*, *42*(3), 850–871.

Aronow, P. M., & Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, *1*(1).

Aronow, P. M., & Samii, C. (2013). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology*, *39*(1), 231–241.

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference. *Annals of Applied Statistics*, *11*(4), 1912–1947.

Auerbach, E., & Tabord-Meehan, M. (2023). *The local approach to causal inference under network interference.* (arXiv:2105.03810v4)

Basse, G. W., & Airoldi, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, *105*(4), 849–858.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization.* USA: Cambridge University Press.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Dunning, I., Huchette, J., & Lubin, M. (2017). Jump: A modeling language for mathematical optimization. *SIAM Review*, *59*(2), 295-320.

Fattorini, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, *93*(2), 269–278.

Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 1035–1056.

Gadbury, G. L. (2001). Randomization inference and bias of standard errors. *American Statistician*, *55*(4), 310–313.

Higgins, M. J., Sävje, F., & Sekhon, J. S. (2015). *Blocking estimators and inference under the Neyman–Rubin model.* (arXiv:1510.01103v1)

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, *27*(24), 4857–4873.

Imbens, G., & Menzel, K. (2021). A causal bootstrap. *The Annals of Statistics*, *49*(3), 1460–1488.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, *50*(271), 946–967.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, *90*(1), 267–293.

Li, S., & Wager, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, *50*(4), 2334–2358.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, *16*(1), 1–23.

Middleton, J. A. (2018). *A unified theory of regression adjustment for design-based inference.* (arXiv:1803.06011v1)

Mukerjee, R., Dasgupta, T., & Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, *113*(522), 868–881.

Neyman, J. (1990/1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472. (Original work published 1923.)

Nutz, M., & Wang, R. (2021). The directional optimal transport. *Annals of Applied Probability*, *in press*.

O'Donoghue, B., Chu, E., Parikh, N., & Boyd, S. (2023, November). *SCS: Splitting conic solver, version 3.2.4.*

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571.

Pashley, N. E., & Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, *46*(3), 271–296.

Robins, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, *7*(7), 773–785.

Rockafellar, R. T. (1993). Lagrange multipliers and optimality. *SIAM Review*, *35*(2), 183–238.

Samii, C., & Aronow, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, *82*(2), 365–370.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Sävje, F. (2024). Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, *111*(1), 1–15.

Wager, S., Du, W., Taylor, J., & Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, *113*(45), 12673–12678.

Wilk, M. B. (1955). The randomzation analysis of a generalized randomized block design. *Biometrika*, *42*(1-2), 70–79.

Wu, E., & Gagnon-Bartsch, J. A. (2018). The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation Review*, *42*(4), 458–488.