Journal of Statistical Planning and Inference 211 (2021) 287–314

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Ridge rerandomization: An experimental design strategy in the presence of covariate collinearity

Zach Branson[a,*], Stephane Shao[b]

[a] Department of Statistics and Data Science, Carnegie Mellon University, United States of America
[b] Department of Statistics, Harvard University, United States of America

ARTICLE INFO

Article history:
Received 10 February 2020
Received in revised form 13 June 2020
Accepted 2 July 2020
Available online 10 July 2020

Keywords:
Covariate balance
Experimental design
Mahalanobis distance
Randomization inference
Ridge regression

ABSTRACT

Randomization ensures that observed and unobserved covariates are balanced, on average. However, randomizing units to treatment and control often leads to covariate imbalances in realization, and such imbalances can inflate the variance of estimators of the treatment effect. One solution to this problem is rerandomization – an experimental design strategy that randomizes units until some balance criterion is fulfilled – which yields more precise estimators of the treatment effect if covariates are correlated with the outcome. Most rerandomization schemes in the literature utilize the Mahalanobis distance, which may not be preferable when covariates are high-dimensional or highly correlated with each other. As an alternative, we introduce an experimental design strategy called ridge rerandomization, which utilizes a modified Mahalanobis distance that addresses collinearities among covariates. This modified Mahalanobis distance has connections to principal components and the Euclidean distance, and – to our knowledge – has remained unexplored. We establish several theoretical properties of this modified Mahalanobis distance and our ridge rerandomization scheme. These results guarantee that ridge rerandomization is preferable over randomization and suggest when ridge rerandomization is preferable over standard rerandomization schemes. We also provide simulation evidence that suggests that ridge rerandomization is particularly preferable over typical rerandomization schemes in high-dimensional or high-collinearity settings.

1. Introduction

Randomized experiments are often considered the "gold standard" of scientific investigations because, on average, randomization balances all potential confounders, both observed and unobserved (Krause and Howard, 2003). However, many have noted that randomized experiments can yield "bad allocations", where some covariates are not well-balanced across treatment groups (Seidenfeld, 1981; Lindley, 1982; Papineau, 1994; Rosenberger and Sverdlov, 2008). Covariate imbalance among different treatment groups complicates the interpretation of estimated causal effects, and thus covariate adjustments are often employed, typically through regression or other comparable methods.

However, it would be better to prevent such covariate imbalances from occurring before treatment is administered, rather than depend on assumptions for covariate adjustment post-treatment which may not hold (Freedman, 2008). One common experimental design tool is blocking, where units are first grouped together based on categorical covariates, and then treatment is randomized within these groups. However, blocking is less intuitive when there are non-categorical

[*] Corresponding author.
E-mail address: zach@stat.cmu.edu (Z. Branson).

https://doi.org/10.1016/j.jspi.2020.07.002

covariates. A more recent experimental design tool that prevents covariate imbalance and allows for non-categorical covariates is the rerandomization scheme of Morgan and Rubin (2012), where units are randomized until a prespecified level of covariate balance is achieved. Rerandomization has been discussed as early as R.A. Fisher (e.g., see Fisher, 1992), and more recent works (e.g., Cox, 2009; Bruhn and McKenzie, 2009; Worrall, 2010) recommend rerandomization. Morgan and Rubin (2012) formalized these recommendations in treatment-versus-control settings and was one of the first works to establish a theoretical framework for rerandomization schemes. Since Morgan and Rubin (2012), several extensions have been made. Morgan and Rubin (2015) developed rerandomization for treatment-versus-control experiments where there are tiers of covariates that vary in importance; Branson et al. (2016) extended rerandomization to 2^K factorial designs; and Zhou et al. (2018) developed a rerandomization scheme for sequential designs. Finally, Li et al. (2018) established asymptotic results for the rerandomization schemes considered in Morgan and Rubin (2012, 2015), and Li and Ding (2020) established asymptotic results for regression adjustment combined with rerandomization.

All of these works focus on using an omnibus measure of covariate balance – the Mahalanobis distance (Mahalanobis, 1936) – during the rerandomization scheme. The Mahalanobis distance is well-known within the matching and observational study literature, where it is used to find subsets of the treatment and control that are similar (Rubin, 1974; Rosenbaum and Rubin, 1985; Gu and Rosenbaum, 1993; Rubin and Thomas, 2000). The Mahalanobis distance is particularly useful in rerandomization schemes because (1) it is symmetric in the treatment assignment, which leads to unbiased estimators of the average treatment effect under rerandomization; and (2) it is equal-percent variance reducing if the covariates are ellipsoidally symmetric, meaning that rerandomization using the Mahalanobis distance reduces the variance of all covariate mean differences by the same percentage (Morgan and Rubin, 2012).

However, the Mahalanobis distance is known to perform poorly in matching for observational studies when there are strong collinearities among the covariates or there are many covariates (Gu and Rosenbaum, 1993; Olsen, 1997; Stuart, 2010). One reason for this is that matching using the Mahalanobis distance places equal importance on balancing all covariates as well as their interactions (Stuart, 2010), and this issue also occurs in rerandomization schemes that use the Mahalanobis distance. This issue was partially addressed by Morgan and Rubin (2015), who proposed an extension of Morgan and Rubin (2012) that incorporates tiers of covariates that vary in importance, such that the most important covariates receive the most variance reduction. However, this requires researchers to specify an explicit hierarchy of importance for the covariates, which might be difficult, especially when the number of covariates is large. Furthermore, it is unclear how to conduct current rerandomization schemes if collinearity is so severe that the covariance matrix of covariates is degenerate, and thus the Mahalanobis distance is undefined.

As an alternative, we consider a rerandomization scheme using a modified Mahalanobis distance that inflates the eigenvalues of the covariates' covariance matrix to alleviate collinearities among the covariates, which has connections to ridge regression (Hoerl and Kennard, 1970). Such a quantity has remained largely unexplored in the literature. First we establish several theoretical properties about this quantity, as well as several properties about a rerandomization scheme that uses this quantity. In particular, instead of reducing the variance of all covariates equally, ridge rerandomization increases the variance reduction of the first principal components of the covariate space at the expense of decreasing the variance reduction of the last principal components. We show through simulation that a rerandomization scheme that incorporates this modified criterion can be beneficial in terms of variance reduction when there are strong collinearities among the covariates. We also discuss how this modified Mahalanobis distance connects to other criteria, such as principal components and the Euclidean distance. Because the rerandomization literature has focused almost exclusively on the Mahalanobis distance, this work also contributes to the literature by exploring the use of other criteria besides the Mahalanobis distance for rerandomization schemes.

The remainder of this paper is organized as follows. In Section 2, we introduce the notation that will be used throughout the paper. In Section 3, we review the rerandomization scheme of Morgan and Rubin (2012). In Section 4, we outline our proposed rerandomization approach and establish several theoretical properties of this approach, as well as several theoretical properties about the modified Mahalanobis distance. In Section 5, we provide simulation evidence that suggests that our rerandomization approach is often preferable over other rerandomization approaches, particularly in high-dimensional or high-collinearity settings. In Section 6, we conclude with a discussion of future work.

2. Notation

We use the colon notation $\lambda_{1:K} = (\lambda_1, \dots, \lambda_K) \in \mathbb{R}^K$ for tuples of objects, and we let $f(\lambda_{1:K}) = (f(\lambda_1), \dots, f(\lambda_K))$ for any univariate function $f : \mathbb{R} \to \mathbb{R}$. We respectively denote by $\mathbf{I}_N$ and $\mathbf{1}_N$ the $N \times N$ identity matrix and the N-dimensional column vector whose coefficients are all equal to 1. Given a matrix A, we denote by A_{ij} its (i, j)-coefficient, $A_{i\bullet}$ its ith row, $A_{\bullet j}$ its jth column, $A^\top$ its transpose, and $\mathrm{tr}(A)$ its trace when A is square. Given two symmetric matrices A and B of the same size, we write $A > B$ (resp. $A \geq B$) if the matrix $A - B$ is positive definite (resp. semi-definite).

Let $\mathbf{x}$ be the $N \times K$ matrix representing K covariates measured on N experimental units. Let $W_i = 1$ if unit i is assigned to treatment and 0 otherwise, and let $\mathbf{W} = (W_1 \, \dots \, W_N)^\top$. Unless stated otherwise, we will focus on completely randomized experiments (Imbens and Rubin, 2015, see Definition 4.2) with a fixed number of N_T treated units and $N_C = N - N_T$ control units. For a given assignment vector $\mathbf{W}$, we define $\bar{\mathbf{x}}_T = N_T^{-1} \mathbf{x}^\top \mathbf{W}$ and $\bar{\mathbf{x}}_C = N_C^{-1} \mathbf{x}^\top (\mathbf{1}_N - \mathbf{W})$ as the respective covariate mean vectors within treatment and control.

For completely randomized experiments, the covariance matrix of the covariate mean differences is $\boldsymbol{\Sigma} = \mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}) = N N_T^{-1} N_C^{-1} S_{\mathbf{x}}^2$, where $S_{\mathbf{x}}^2 = (N-1)^{-1}(\mathbf{x} - \mathbf{1}_N \bar{\mathbf{x}}_N)^\top (\mathbf{x} - \mathbf{1}_N \bar{\mathbf{x}}_N)$ is the sample covariance matrix of

$\mathbf{x}$ with $\bar{\mathbf{x}}_N = N^{-1}\mathbf{1}_N^\top \mathbf{x}$ (Morgan and Rubin, 2012). Throughout, we use $\boldsymbol{\Sigma}$ to refer to this fixed covariance matrix, and we assume $\boldsymbol{\Sigma} > 0$. The spectral decomposition ensures that $\boldsymbol{\Sigma}$ is diagonalizable with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_K > 0$. Let $\boldsymbol{\Gamma}$ be the orthogonal matrix of corresponding eigenvectors, so that we may write $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\mathbf{Diag}(\lambda_{1:K})\boldsymbol{\Gamma}^\top$, where $\mathbf{Diag}(\lambda_{1:K})$ denotes the $K \times K$ diagonal matrix whose (k, k)-coefficient is λ_k. Thus, $\boldsymbol{\Sigma}$ and its eigenstructure are available in closed-form, and the latter coincides with the eigenstructure of $S^2_{\mathbf{x}}$ up to a scaling factor.

We let χ^2_K denote a chi-squared distribution with K degrees of freedom, $\mathbb{P}(\chi^2_K \leq a)$ its cumulative distribution function (CDF) evaluated at $a \in \mathbb{R}$, and $q_{\chi^2_K}(p)$ its p-quantile for $p \in (0, 1)$.

3. Review of rerandomization

We follow the potential outcomes framework (Rubin, 1990, 2005), where each unit i has fixed potential outcomes $Y_i(1)$ and $Y_i(0)$, which denote the outcome for unit i under treatment and control, respectively. Thus, the observed outcome for unit i is $y_i^{obs} = W_iY_i(1) + (1 - W_i)Y_i(0)$. Define $\mathbf{y}^{obs} = (y_1^{obs} \; \ldots \; y_N^{obs})^\top$ as the vector of observed outcomes. We focus on the average treatment effect as the causal estimand, defined as

$$\tau = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]. \tag{1}$$

Furthermore, we focus on the mean-difference estimator

$$\hat{\tau} = \bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C \,, \tag{2}$$

where $\bar{\mathbf{y}}_T = N_T^{-1}\mathbf{W}^\top \mathbf{y}^{obs}$ and $\bar{\mathbf{y}}_C = N_C^{-1}(\mathbf{1}_N - \mathbf{W})^\top \mathbf{y}^{obs}$ are the average treatment and control outcomes, respectively. When conducting a randomized experiment, ideally we would like $\bar{\mathbf{x}}_T$ and $\bar{\mathbf{x}}_C$ to be close; otherwise, the estimator $\hat{\tau}$ could be confounded by imbalances in the covariate means.

Morgan and Rubin (2012) focused on a rerandomization scheme using the Mahalanobis distance to ensure that the covariate means are reasonably balanced for a particular treatment assignment. The Mahalanobis distance between the treatment and control covariate means is defined as

$$M = (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C), \tag{3}$$

where the dependence of M on the assignment vector $\mathbf{W}$ is implicit through $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$. Morgan and Rubin (2012) suggest randomizing units to treatment and control by performing independent draws from the distribution of $\mathbf{W} \mid \mathbf{x}$ until $M \leq a$ for some threshold $a \geq 0$. Hereafter, we refer to this procedure of randomizing units until $M \leq a$ as *rerandomization*. The expected number of draws until the first acceptable randomization is equal to $1/p_a$, where $p_a = \mathbb{P}(M \leq a \mid \mathbf{x})$ is the probability that a particular realization of $\mathbf{W}$ yields a Mahalanobis distance M less than or equal to a. Thus, fixing p_a effectively allocates an expected computational budget and induces a corresponding threshold a: the smaller the acceptance probability p_a, the smaller the threshold a and thus the more balanced the two groups, but the larger the expected computational cost of drawing an acceptable $\mathbf{W}$. For example, to restrict rerandomization to the "best" 1% randomizations, one would set $p_a = 0.01$, which implicitly sets a equal to the p_a-quantile of the distribution of M given $\mathbf{x}$. If one assumes $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, then $M \mid \mathbf{x} \sim \chi^2_K$, so that a can be chosen equal to the p_a-quantile of a chi-squared distribution with K degrees of freedom. The assumption $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ can be justified by invoking the finite population Central Limit Theorem (Erdös and Rényi, 1959; Li and Ding, 2017). When the distribution of $M \mid \mathbf{x}$ is unknown, one can approximate it via Monte Carlo by simulating independent draws of $M \mid \mathbf{x}$ and setting a to the p_a-quantile of M's empirical distribution.

Morgan and Rubin (2012) established that the mean-difference estimator $\hat{\tau}$ under this rerandomization scheme is unbiased in estimating the average treatment effect τ, i.e., that $\mathbb{E}\left[\hat{\tau} \mid \mathbf{x}, M \leq a\right] = \tau$. Furthermore, they also established that under rerandomization, if $N_T = N_C$ and $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, then not only are the covariate mean differences centered at 0, i.e., $\mathbb{E}\left[\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M \leq a\right] = 0$, but also they are more closely concentrated around 0 than they would be under randomization. More precisely, Morgan and Rubin (2012) proved that

$$\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M \leq a) = v_a \mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}), \tag{4}$$

$$\text{with} \quad v_a = \frac{\mathbb{P}(\chi^2_{K+2} \leq a)}{\mathbb{P}(\chi^2_K \leq a)} \in (0, 1). \tag{5}$$

Therefore, under their assumptions, rerandomization using the Mahalanobis distance reduces the variance of each covariate mean difference by $100(1 - v_a)$% compared to randomization. Morgan and Rubin (2012) call this last property *equally percent variance reducing* (EPVR). Thus, using the Mahalanobis distance for rerandomization can be quite appealing, but Morgan and Rubin (2012) rightly point out that non-EPVR rerandomization schemes may be preferable in settings with covariates of unequal importance. This is in part addressed by Morgan and Rubin (2015), who developed a rerandomization scheme that incorporates tiers of covariates that vary in importance. However, this requires researchers to specify an explicit hierarchy of covariate importance, which may not be immediately clear, especially when the number of covariates is large. Furthermore, if there are strong collinearities amongst covariates such that $\mathbf{\Sigma}$ is degenerate and thus the M in (3) is undefined, then it is unclear how one should conduct the rerandomization scheme of Morgan and Rubin (2012) and its extensions (Morgan and Rubin, 2015; Branson et al., 2016; Li et al., 2018; Li and Ding, 2020).

4. Ridge rerandomization

As an alternative, we consider a modified Mahalanobis distance, defined as

$$M_\lambda = (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^\top (\mathbf{\Sigma} + \lambda\, \mathbf{I}_K)^{-1} (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \tag{6}$$

for some prespecified $\lambda \geq 0$. Guidelines for choosing λ will be provided in Section 4.2. The eigenvalues of $\mathbf{\Sigma}$ in (6) are inflated in a way that is reminiscent of ridge regression (Hoerl and Kennard, 1970). For this reason, we will refer to the quantity M_λ as the *ridge Mahalanobis distance*. To our knowledge, the ridge Mahalanobis distance has remained largely unexplored, except for Kato et al. (1999), who used it in an application for a Chinese and Japanese character recognition system. Our proposed rerandomization scheme, referred to as *ridge rerandomization*, involves using the ridge Mahalanobis distance in place of the standard Mahalanobis distance within the rerandomization framework of Morgan and Rubin (2012). In other words, one randomizes the assignment vector $\mathbf{W}$ until $M_\lambda \leq a_\lambda$ for some threshold $a_\lambda \geq 0$.

In order to make a fair comparison between rerandomization and ridge rerandomization, we will fix the expected computational cost of ridge rerandomization by calibrating the respective thresholds so that

$$\mathbb{P}(M_\lambda \leq a_\lambda \mid \mathbf{x}) = \mathbb{P}(M \leq a \mid \mathbf{x}) = p_a. \tag{7}$$

Thus, fixing p_a implicitly determines the pair (λ, a_λ), so that for every fixed $\lambda \geq 0$ and $p_a \in (0, 1)$ corresponds to a unique a_λ that satisfies (7).

As we will discuss in Section 4.3, the ridge Mahalanobis distance alleviates collinearity among the covariate mean differences by placing higher importance on the directions that account for the most variation. In that section we also discuss how ridge rerandomization encapsulates a spectrum of other standard rerandomization schemes. But first, in Section 4.1 we establish several theoretical properties of ridge rerandomization for some prespecified (λ, a_λ), and in Section 4.2 we provide guidelines for specifying (λ, a_λ). In Section 4.4, we discuss how to conduct inference for the average treatment effect τ after ridge rerandomization is used to design a randomized experiment.

4.1. Properties of ridge rerandomization

The following theorem establishes that, on average, the covariate means in the treatment and control groups are balanced under ridge rerandomization, and that $\hat{\tau}$ is an unbiased estimator of τ under ridge rerandomization.

Theorem 4.1 (*Unbiasedness Under Ridge Rerandomization*). *Let $\lambda \geq 0$ and $a_\lambda \geq 0$ be some prespecified constants. If $N_T = N_C$, then*

$$\mathbb{E}[\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \leq a_\lambda] = 0$$

and

$$\mathbb{E}[\hat{\tau} \mid \mathbf{x}, M_\lambda \leq a_\lambda] = \tau.$$

Theorem 4.1 is a particular case of Theorem 2.1 and Corollary 2.2 from Morgan and Rubin (2012). Theorem 4.1 follows from the symmetry of M_λ in treatment and control, in the sense that both assignments $\mathbf{W}$ and $(\mathbf{1}_N - \mathbf{W})$ yield the same value of M_λ. From Morgan and Rubin (2012), we even have the stronger result that $\mathbb{E}[\bar{V}_T - \bar{V}_C \mid \mathbf{x}, M_\lambda \leq a_\lambda] = 0$ for any covariate V, regardless of whether V is observed or not. While it may seem stringent to require that $N_T = N_C$, Morgan and Rubin (2012) demonstrate a simple counterexample where rerandomization also yields biased treatment effect estimates when $N_T \neq N_C$. However, Morgan and Rubin (2015, Section 3.2) conjectured that this bias was small for even moderate sample sizes, and Li et al. (2018) formalized this conjecture by showing that $\hat{\tau}$ is asymptotically unbiased under rerandomization even when $N_T \neq N_C$. While asymptotic properties of ridge rerandomization are outside the scope of this work, we can similarly conjecture that the bias of $\hat{\tau}$ under ridge rerandomization will be small for moderate sample sizes, even when $N_T \neq N_C$. We discuss simulation results that validate this conjecture in Section 5.4.

Now we establish the covariance structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ under ridge rerandomization. To do this, we first derive the exact distribution of M_λ. The following lemma establishes that if we assume $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$, then M_λ is distributed as a weighted sum of K independent χ^2_1 random variables, where the sizes of the weights are ordered in the same fashion as the sizes of the eigenvalues of $\mathbf{\Sigma}$.

Lemma 4.1 (*Distribution of M_λ*)**.** *Let $\lambda \geq 0$ be some prespecified constant. If $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$, then*

$$M_\lambda \mid \mathbf{x} \;\sim\; \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda} Z_j^2 \tag{8}$$

where $Z_1, \ldots, Z_K \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and $\lambda_1 \geq \cdots \geq \lambda_K > 0$ are the eigenvalues of $\mathbf{\Sigma}$.

The proof of Lemma 4.1 is provided in the Appendix; see Appendix A.1. Under the Normality assumption, the representation in (8) provides a straightforward way to simulate independent draws of M_λ, despite its CDF being typically intractable and requiring numerical approximations (e.g., see Bodenham and Adams, 2016, and references therein).

We will find that the covariance structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ under ridge rerandomization depends on the conditional expectations $\mathbb{E}[Z_k^2 | \mathbf{x}, M_\lambda \leq a_\lambda]$, where $k = 1, \ldots, K$ and $Z_1, \ldots, Z_K \overset{\text{i.i.d.}}{\sim} N(0, 1)$. The following lemma establishes a property that will be helpful for characterizing these conditional expectations.

Lemma 4.2 (*Conditional Expectations of Constrained Non-Negative Random Variables*)**.** *Let $L_1, \ldots, L_K$ be independent and identically distributed non-negative random variables, let $C_1, \ldots, C_K$ be non-negative constants such that $C_1 \geq C_2 \geq \cdots \geq C_K$, and let $a > 0$ be some constant. Define, for $k = 1, \ldots, K$,*

$$E_k = \mathbb{E}\left[L_k \middle| \sum_{j=1}^{K} C_j L_j \leq a \right] \tag{9}$$

Then, $E_1 \leq E_2 \leq \cdots \leq E_K$.

The proof of Lemma 4.2 is provided in the Appendix; see Appendix A.2. We would like to thank an anonymous reviewer for suggesting a way to prove this result.

Using Lemmas 4.1 and 4.2, we can derive the covariance structure of $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ under ridge rerandomization, as stated by the following theorem.

Theorem 4.2 (*Covariance Structure Under Ridge Rerandomization*)**.** *Let $\lambda \geq 0$ and $a_\lambda \geq 0$ be some prespecified constants. If $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$ and $N_T = N_C$, then*

$$Cov(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \leq a_\lambda) = \mathbf{\Gamma} \boldsymbol{Diag}((\lambda_k\, d_{k,\lambda})_{1 \leq k \leq K}) \mathbf{\Gamma}^\top \tag{10}$$

where $\mathbf{\Gamma}$ is the orthogonal matrix of eigenvectors of $\mathbf{\Sigma}$ corresponding to the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_K > 0$, and for all $k = 1, \ldots, K$,

$$d_{k,\lambda} = \mathbb{E}\left[Z_k^2 \middle| \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda} Z_j^2 \leq a_\lambda \right] \tag{11}$$

with $Z_1, \ldots, Z_K \overset{\text{i.i.d.}}{\sim} N(0, 1)$, where $d_{1,\lambda} \leq d_{2,\lambda} \leq \cdots \leq d_{K,\lambda}$.

The proof of Theorem 4.2 is in Appendix A.3. The quantities $d_{k,\lambda}$ are intractable functions of λ and a_λ and thus need to be approximated numerically, as explained in Section 4.2. Conditioning on $M_\lambda \leq a_\lambda$ in (11) effectively constrains the magnitude of the positive random variables Z_k^2. Since the weights $\lambda_k(\lambda_k + \lambda)^{-1}$ of their respective contributions to M_λ are positive and non-increasing with $k = 1, \ldots, K$, intuitively $0 < d_{1,\lambda} \leq \cdots \leq d_{K,\lambda} < 1$, and this is established by Lemma 4.2.

Using the above results, we can now compare randomization, rerandomization, and ridge rerandomization. Under the assumptions stated in Theorem 4.2, the covariance matrices of $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ under randomization, rerandomization, and ridge rerandomization can be respectively written as

$$\text{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}) = \mathbf{\Gamma} \mathbf{Diag}\big((\lambda_k)_{1 \leq k \leq K}\big) \mathbf{\Gamma}^\top, \tag{12}$$

$$\text{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M \leq a) = \mathbf{\Gamma} \mathbf{Diag}\big((\lambda_k\, v_a)_{1 \leq k \leq K}\big) \mathbf{\Gamma}^\top, \tag{13}$$

$$\text{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \leq a_\lambda) = \mathbf{\Gamma} \mathbf{Diag}\big((\lambda_k\, d_{k,\lambda})_{1 \leq k \leq K}\big) \mathbf{\Gamma}^\top. \tag{14}$$

where (13) follows from Theorem 3.1 in Morgan and Rubin (2012) with $v_a \in (0, 1)$, and (14) follows from Theorem 4.2 with $d_{k,\lambda} \in (0, 1)$ defined in (11). If we define new covariates $\mathbf{x}^*$ as the principal components of the original ones, i.e., $\mathbf{x}^* = \mathbf{x}\mathbf{\Gamma}$, then (13) and (14) respectively yield

$$\text{Var}\big((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \mid \mathbf{x}, M \leq a\big) = v_a\, \text{Var}\big((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \mid \mathbf{x}\big) \tag{15}$$

and

$$\text{Var}\big((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \mid \mathbf{x}, M_\lambda \leq a_\lambda\big) = d_{k,\lambda}\, \text{Var}\big((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \mid \mathbf{x}\big) \tag{16}$$

for all $k = 1, \dots, K$, where $(\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k$ is the kth *principal component mean difference* between the treatment and control groups, i.e., the kth coefficient of $\mathbf{\Gamma}^\top(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$. From (15) we see that rerandomization reduces the variances of the principal component mean differences equally by $100(1 - v_a)\%$ and is thus EPVR for the principal components, as well as for the original covariates, as discussed in Section 3. On the other hand, ridge rerandomization reduces these variances by unequal amounts: the variance of the kth principal component mean difference is reduced by $100(1 - d_{k,\lambda})\%$, and because $0 < d_{1,\lambda} \leq \cdots \leq d_{K,\lambda} < 1$, ridge rerandomization places more importance on the first principal components.

Translating (16) back to the original covariates yields the following corollary, which establishes that ridge rerandomization is always preferable over randomization in terms of reducing the variance of each covariate mean difference.

Corollary 4.1 (*Variance Reduction for Ridge Rerandomization*)**.** *Under the assumptions of Theorem 4.2, ridge rerandomization reduces the variance of the kth covariate mean difference* $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k$ *by* $100\left(1 - v_{k,\lambda}\right)\%$*, where*

$$v_{k,\lambda} = \frac{\left(\mathbf{\Gamma}\boldsymbol{Diag}\left((\lambda_j\, d_{j,\lambda})_{1\leq j\leq K}\right)\mathbf{\Gamma}^\top\right)_{kk}}{\mathbf{\Sigma}_{kk}} \tag{17}$$

satisfies $v_{k,\lambda} \in (0, 1)$*, so that*

$$Var\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x}, M_\lambda \leq a_\lambda\right) \;<\; Var\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x}\right). \tag{18}$$

The proof of Corollary 4.1 is provided in the Appendix; see Appendix A.4. Reducing the variance of the covariate mean differences is beneficial for precisely estimating the average treatment effect if the outcomes are correlated with the covariates. For example, Theorem 3.2 of Morgan and Rubin (2012) establishes that – under several assumptions, including additivity of the treatment effect – rerandomization reduces the variance of $\hat{\tau}$ defined in (2) by $100(1 - v_a)R^2$ percent, where R^2 denotes the squared multiple correlation between the outcomes and the covariates. Now we establish how the variance of $\hat{\tau}$ behaves under ridge rerandomization.

In the rest of this section, we assume—as in Morgan and Rubin (2012)—that the treatment effect is additive. Without loss of generality, for all $i = 1, \dots, N$, we can write the outcome of unit i as

$$Y_i(W_i) = \beta_0 + \mathbf{x}_{i\bullet}\boldsymbol{\beta} + \tau W_i + \epsilon_i \tag{19}$$

where $\beta_0 + \mathbf{x}\boldsymbol{\beta}$ is the projection of the potential outcomes $\mathbf{Y}(0) = (Y_1(0)\dots Y_N(0))^\top$ onto the linear space spanned by $(\mathbf{1}, \mathbf{x})$, and $\epsilon_i \in \mathbb{R}$ captures any misspecification of the linear relationship between the outcomes and $\mathbf{x}$. Let $\bar{\boldsymbol{\epsilon}}_T = N_T^{-1}\mathbf{W}^\top\boldsymbol{\epsilon}$ and $\bar{\boldsymbol{\epsilon}}_C = N_C^{-1}(\mathbf{1}_N - \mathbf{W})^\top\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1 \dots \epsilon_N)^\top$.

Theorem 4.3 establishes that the variance of $\hat{\tau}$ under ridge rerandomization is always less than or equal to the variance of $\hat{\tau}$ under randomization. Thus, ridge rerandomization always leads to a more precise treatment effect estimator than randomization.

Theorem 4.3. *Under the assumptions of Theorem 4.2, if* $(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C)$ *is conditionally independent of* $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ *given* $\mathbf{x}$*, and if there is an additive treatment effect, then*

$$Var(\hat{\tau} \mid \mathbf{x}) - Var(\hat{\tau} \mid \mathbf{x}, M_\lambda \leq a_\lambda) \;=\; \boldsymbol{\beta}^\top\mathbf{\Gamma}\boldsymbol{Diag}\left((\lambda_k\left(1 - d_{k,\lambda}\right))_{1\leq k\leq K}\right)\mathbf{\Gamma}^\top\boldsymbol{\beta}$$

so that we have

$$Var(\hat{\tau} \mid \mathbf{x}, M_\lambda \leq a_\lambda) \;\leq\; Var(\hat{\tau} \mid \mathbf{x}),$$

where the equality holds if and only if $\boldsymbol{\beta} = \mathbf{0}_K$ *in (19).*

The proof of Theorem 4.3 is in the Appendix; see Appendix A.5. The conditional independence assumption was also leveraged in the proof of Theorem 3.2 in Morgan and Rubin (2012). While this independence assumption may seem strong, Li et al. (2018) showed that it is justified asymptotically, which allowed them to establish that rerandomization is preferable over randomization even if treatment effects are not additive. Again, while the asymptotic properties of ridge rerandomization are outside the scope of this work, we conjecture that Theorem 4.3 holds asymptotically even without the conditional independence and additive treatment effects assumptions. Indeed, we find evidence via simulation that ridge rerandomization is still preferable over randomization (and often rerandomization) when treatment effects are heterogeneous, as discussed in Section 5.4.

The fact that ridge rerandomization performs better than randomization is arguably a low bar, because this is the purpose of any rerandomization scheme. The following corollary quantifies how ridge rerandomization performs compared to the rerandomization scheme of Morgan and Rubin (2012).

Corollary 4.2. *Under the assumptions of Theorem 4.3, the difference in variances of* $\hat{\tau}$ *between rerandomization and ridge rerandomization is*

$$Var(\hat{\tau} \mid \mathbf{x}, M \leq a) - Var(\hat{\tau} \mid \mathbf{x}, M_\lambda \leq a_\lambda) \;=\; \boldsymbol{\beta}^\top\mathbf{\Gamma}\boldsymbol{Diag}\left((\lambda_k\left(v_a - d_{k,\lambda}\right))_{1\leq k\leq K}\right)\mathbf{\Gamma}^\top\boldsymbol{\beta}.$$

It is not necessarily the case that $d_{k,\lambda} \leq v_a$ for all $k = 1, \dots, K$, and so it is not guaranteed that ridge rerandomization will perform better or worse than rerandomization in terms of treatment effect estimation. Ultimately, the comparison of rerandomization and ridge rerandomization depends on $\boldsymbol{\beta}$, which is typically not known until after the experiment has been conducted.

However, in Section 5.3, we provide some heuristic arguments for when ridge rerandomization would be preferable over rerandomization, along with simulation evidence that confirms these heuristic arguments. In particular, we demonstrate that ridge rerandomization is preferable over rerandomization when there are strong collinearities among the covariates. We also discuss a "worst-case scenario" for ridge rerandomization, where $\boldsymbol{\beta}$ is specified such that ridge rerandomization should perform worse than rerandomization in terms of treatment effect estimation accuracy.

In order to implement ridge rerandomization, researchers must specify the threshold $a_\lambda \geq 0$ and the regularization parameter $\lambda \geq 0$. The next section provides guidelines for choosing these parameters.

4.2. Guidelines for choosing a_λ and λ

For ridge rerandomization, we recommend starting by specifying an acceptance probability $p_a \in (0, 1)$, which then binds λ and a_λ together via the identity (7). Once p_a is fixed, there exists a uniquely determined threshold $a_\lambda \geq 0$ for each $\lambda \geq 0$ such that $\mathbb{P}(M_\lambda \leq a_\lambda \mid \mathbf{x}) = p_a$. As in Morgan and Rubin (2012), acceptable treatment allocations under ridge rerandomization are generated by randomizing units to treatment and control until $M_\lambda \leq a_\lambda$. Thus, a smaller p_a leads to stronger covariate balance according to M_λ at the expense of computation time.

The only choice that remains after fixing p_a is the regularization parameter $\lambda \geq 0$. The choice of λ is investigated in Section 4.2.1. Once we fix p_a and λ, we can set a_λ equal to the p_a-quantile of the quadratic form Q_λ defined by

$$Q_\lambda = \sum_{k=1}^{K} \frac{\lambda_k}{\lambda_k + \lambda} Z_k^2 \tag{20}$$

where $Z_1, \dots, Z_K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Such a choice of a_λ is a good approximation of the p_a-quantile of M_λ, especially when N is large enough for $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x}$ to be approximately Normal, as motivated by Lemma 4.1. Computing the probability density function (and quantile) of a weighted sum of independent χ_1^2 random variables is a classical topic in computational statistics (Imhof, 1961; Davies, 1980; Bausch, 2013), so the details of computing a_λ are relegated to Appendix A.6.

Similarly, our choice of λ will depend on the $d_{k,\lambda}$'s defined in (11), which involve intractable conditional expectations. However, these $d_{k,\lambda}$'s can be consistently estimated via Monte Carlo in a way that is computationally negligible, as discussed in Appendix A.6. We denote the corresponding estimators as $\hat{d}_{k,\lambda}$. We can then estimate $v_{k,\lambda}$ from Corollary 4.1 consistently for all $k = 1, \dots, K$, by

$$\hat{v}_{k,\lambda} = \frac{\left(\boldsymbol{\Gamma}\mathbf{Diag}\left((\lambda_j \hat{d}_{j,\lambda})_{1\leq j\leq K}\right)\boldsymbol{\Gamma}^\top\right)_{kk}}{\boldsymbol{\Sigma}_{kk}} \tag{21}$$

which will be used to choose λ, as we discuss in the remainder of this section.

4.2.1. Choosing λ

In this section, assume that p_a has been fixed. Note that choosing $\lambda = 0$ corresponds to rerandomization using the Mahalanobis distance. Thus, we would only choose some $\lambda > 0$ if it is preferable over rerandomization, in the following sense. There are many metrics that could be used for comparing rerandomization and ridge rerandomization; for simplicity, we focus on the average percent reduction in variance across covariate mean differences. Arguably, ridge rerandomization is preferable over rerandomization only if it is able to achieve a higher average reduction in variance across covariate mean differences. Recall that, as discussed in Section 3, rerandomization reduces the variance of each covariate mean difference by $100(1 - v_a)$% compared to randomization, where v_a is defined in (5). Meanwhile, as established by Corollary 4.1, ridge rerandomization reduces the variance of the kth covariate mean difference by $100(1 - v_{k,\lambda})$%, where $v_{k,\lambda}$ is defined in (17). Thus, the average variance reduction under ridge rerandomization is greater than that under rerandomization only if

$$K^{-1}\sum_{k=1}^{K}(1 - v_{k,\lambda}) > 1 - v_a \leftrightarrow K^{-1}\sum_{k=1}^{K} v_{k,\lambda} < v_a \tag{22}$$

Proving the existence of some $\lambda > 0$ such that (22) holds is challenging, so we propose the following iterative procedure (see "Procedure for finding a desirable $\lambda \geq 0$") for choosing such a $\lambda > 0$ if it exists. The technical details justifying this procedure are in the Appendix; but at a high-level, our procedure uses the following intuition:

- Ridge rerandomization with $\lambda > 0$ is preferable over rerandomization (i.e., ridge rerandomization with $\lambda = 0$) only if (22) holds.
- Thus, we will iteratively search for $\lambda > 0$ such that (22) holds.
- If we cannot find any $\lambda > 0$ such that (22) holds, then we set $\lambda = 0$. Otherwise, among all the λ's satisfying (22), we set λ such that the conditional covariance structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ is altered the least.

We discuss why we choose a λ that alters the conditional covariance structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ the least in Section 4.3. In the procedure below, we initialize $\lambda = 0$, and then we iteratively increase candidate λ's by increments of δ, which is specified by the user. As a rule-of-thumb, the step size δ can be chosen as a fraction of the smallest strictly positive gap between consecutive eigenvalues, i.e., $\min\{\lambda_k - \lambda_{k-1} : k = 1, \dots, K \text{ such that } \lambda_k > \lambda_{k-1}\}$ with the convention $\lambda_0 = 0$. The stopping point of this iterative search is chosen dynamically in Step 3 of our procedure, and we discuss in Appendix A.7 why this dynamic search is guaranteed to stop in finite time. Finally, as we discuss further in Appendix A.7, the procedure is computationally efficient in the sense that nK auxiliary Normal variables only need to be simulated once and can be reused when testing different values of λ.

Procedure for finding a desirable $\lambda \geq 0$

1. Specify $p_a \in (0, 1)$, $n \geq 1$, $\delta > 0$, and $\varepsilon > 0$.
2. Initialize $\lambda = 0$ and $\Lambda = \emptyset$.
3. While $|(\lambda + \delta)\hat{a}_{\lambda+\delta} - \lambda \hat{a}_{\lambda}| > \varepsilon$:
    - Set $\lambda = \lambda + \delta$.
    - If $\frac{1}{K} \sum_{k=1}^{K} \hat{v}_{k,\lambda} < \frac{\mathbb{P}(\chi^2_{K+2} \leq q_{\chi^2_K}(p_a))}{p_a}$, then set $\Lambda = \Lambda \cup \{\lambda\}$.
4. If $\Lambda = \emptyset$, then return $\lambda = 0$.
   Else, define $c_k = \lambda_k^2 (\sum_{j=1}^{K} \lambda_j^2)^{-1}$ for all $k = 1, \dots, K$, and return:

$$\lambda_\star = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \left(\sum_{k=1}^{K} c_k \, \hat{d}^2_{k,\lambda} - \left(\sum_{k=1}^{K} c_k \, \hat{d}_{k,\lambda} \right)^2 \right). \tag{23}$$

In our procedure, Λ represents the set of λ such that (22) holds. When the set Λ is empty, we return $\lambda = 0$ (which corresponds to typical rerandomization). However, the following heuristic argument illustrates why we would expect the existence of at least one λ such that (22) holds. The rerandomization scheme of Morgan and Rubin (2012) spreads the benefits of variance reduction across all K covariates equally; however, note that the term $v_a = \mathbb{P}(\chi^2_{K+2} \leq q_{\chi^2_K}(p_a))/p_a$ is monotonically increasing in the number of covariates K for a fixed acceptance probability p_a. Thus, the variance reduction under rerandomization, $100(1-v_a)$%, is monotonically *decreasing* in the number of covariates. A consequence of this is that if one can instead determine a smaller set of $K_e < K$ covariates that is most relevant, then that smaller set of covariates can benefit from a greater variance reduction than what would be achieved by considering all K covariates. As we mentioned at the end of Section 3, this idea was partially addressed in Morgan and Rubin (2015), which extended the rerandomization scheme of Morgan and Rubin (2012) to allow for tiers of covariate importance specified by the researcher, such that the most important covariates receive the most variance reduction. Ridge rerandomization, on the other hand, automatically specifies a hierarchy of importance based on the eigenstructure of the covariate mean differences. To provide intuition for this idea, consider a simple case where the smallest $(K - K_e)$ eigenvalues $\lambda_{K_e+1}, \dots, \lambda_K$ are all arbitrarily close to 0. In this case, we can find $\lambda > 0$ such that $\lambda_j(\lambda_j + \lambda)^{-1} \approx 1$ for the K_e largest eigenvalues and $\lambda_j(\lambda_j + \lambda)^{-1} \approx 0$ for the remaining $K - K_e$ eigenvalues, so that M_λ would be approximately distributed as $\chi^2_{K_e}$ with an effective number of degrees of freedom K_e strictly less than K. For some fixed acceptance probability $p_a \in (0, 1)$ and corresponding thresholds $a_e = q_{\chi^2_{K_e}}(p_a)$ and $a = q_{\chi^2_K}(p_a)$, we would then have

$$v_{a_e} = \frac{\mathbb{P}(\chi^2_{K_e+2} \leq q_{\chi^2_{K_e}}(p_a))}{p_a} < \frac{\mathbb{P}(\chi^2_{K+2} \leq q_{\chi^2_K}(p_a))}{p_a} = v_a \tag{24}$$

since p_a is fixed and $K_e < K$. The relative variance reduction for ridge rerandomization would then be $(1 - v_{a_e})$ for the first K_e principal components – which in this simple example make up the total variation in the covariate mean differences – while the relative variance reduction for rerandomization would be $(1 - v_a) < (1 - v_{a_e})$ for the K covariates. Thus, in this case, ridge rerandomization would achieve a greater variance reduction on a lower-dimensional representation of the covariates than typical rerandomization.

This heuristic argument also hints that our method has connections to a principal-components rerandomization scheme, where one instead balances on some lower dimension of principal components rather than on the covariates themselves. We discuss this point further in Section 4.3.

4.3. Connections to other rerandomization schemes

Ridge rerandomization has connections to other rerandomization schemes. Ridge rerandomization requires specifying the parameter λ; thus, consider two extreme choices of λ:

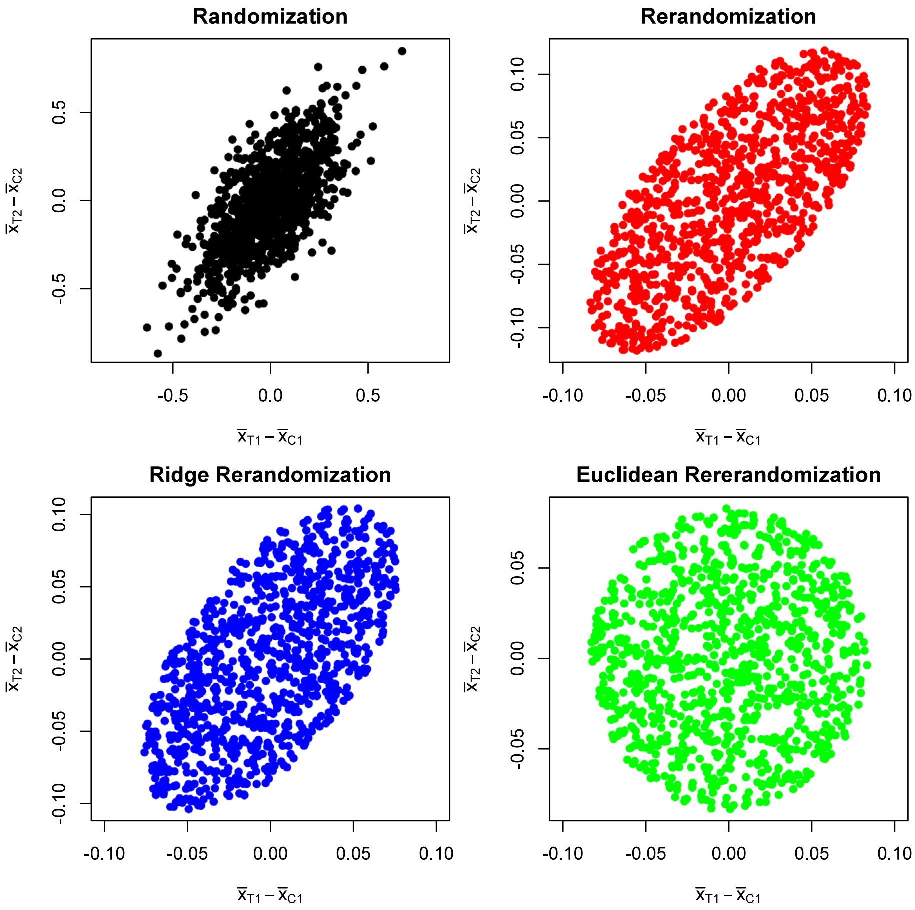


Fig. 1. Distribution of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x}$ under randomization, rerandomization (with $p_a = 0.1$), ridge rerandomization (with $p_a = 0.1$ and $\lambda = 0.005$), and rerandomization using the Euclidean distance. Note the difference in scale for the randomization plot for ease of comparison.

1. $\lambda = 0$: $M_\lambda = M$, i.e., M_λ corresponds to the Mahalanobis distance.
2. $\lambda \to +\infty$: $M_\lambda \approx \lambda^{-1} \|\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C\|^2$, i.e., M_λ tends to a scaled Euclidean distance.

In other words, ridge rerandomization with $\lambda = 0$ is equivalent to rerandomization using the Mahalanobis distance; and for large λ, rerandomization using λM_λ is equivalent to rerandomization using the Euclidean distance. Note, however, that the threshold a_λ will already take the λ^{-1} factor into account when computing the quantile of M_λ, meaning that ridge rerandomization using M_λ for large λ is essentially equivalent to rerandomization using the Euclidean distance.

Thus, for any finite $\lambda > 0$, the distance defined by M_λ can be regarded as a compromise between the Mahalanobis and Euclidean distances. Rerandomization using the Euclidean distance is similar to a rerandomization scheme that places a separate caliper on each covariate, which was proposed by Moulton (2004), Maclure et al. (2006), Bruhn and McKenzie (2009), and Cox (2009). However, Morgan and Rubin (2012) note that such a rerandomization scheme is not affinely invariant and does not preserve the correlation structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ across acceptable randomizations. See Morgan and Rubin (2012) for a full discussion of the benefits of using affinely invariant rerandomization criteria. As discussed in Section 4.2.1, our proposed procedure aims for larger variance reductions of covariate mean differences while mitigating the perturbation of the correlation structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$.

As an illustration, consider a randomized experiment where $N_T = N_C = 50$ units are assigned to treatment and control; and furthermore, where there are two correlated covariates, generated as $x_{1j} \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and $x_{2j} \overset{\text{i.i.d.}}{\sim} N(x_{1i}, 1)$ for $j = 1, \ldots, N$. Fig. 1 shows the distribution of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x}$ across 1000 randomizations, rerandomizations (with $p_a = 0.1$), ridge rerandomizations (with $p_a = 0.1$ and $\lambda = 0.005$), and rerandomizations using the Euclidean distance instead of the Mahalanobis distance.

All three rerandomization schemes reduce the variance of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x}$ for $k \in \{1, 2\}$, compared to randomization; however, rerandomization using the Euclidean distance destroys the correlation structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x}$, while rerandomization and ridge rerandomization largely maintain it. This provides further motivation for Step 4 of the procedure presented in Section 4.2.1.

Furthermore, as discussed in Sections 4.1 and 4.2.1, ridge rerandomization can be regarded as a "soft-thresholding" version of a rerandomization scheme that would focus solely on the first $K_e < K$ principal components of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$. A "hard-thresholding" rerandomization scheme would use a truncated version M_{K_e} of the Mahalanobis distance, defined as

$$M_{K_e} = (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^\top \hat{\boldsymbol{\Sigma}}_{K_e}^{-1} (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$$

with

$$\boldsymbol{\Sigma}_{K_e} = \boldsymbol{\Gamma}\mathbf{Diag}\big((\lambda_1, \ldots, \lambda_{K_e}, 0, \ldots, 0)\big)\boldsymbol{\Gamma}^\top$$

i.e., $\boldsymbol{\Sigma}_{K_e}$ artificially sets the smallest $(K - K_e)$ eigenvalues of $\boldsymbol{\Sigma}$ to 0. This scheme would then be EPVR for the first K_e principal components of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ – although not necessarily EPVR for the original covariates themselves – but would effectively ignore the components associated with the smallest $(K - K_e)$ eigenvalues of $\boldsymbol{\Sigma}$.

Therefore, ridge rerandomization is a flexible experimental design strategy that encapsulates a class of rerandomization schemes, thus making it worth further investigation in future work. We expand on this point in Section 6.

4.4. Conducting inference after ridge rerandomization

Here we outline how to conduct inference for the average treatment effect after ridge rerandomization has been used to conduct an experiment. In general, there are Neymanian, Bayesian, and randomization-based modes of inference for analyzing randomized experiments (Imbens and Rubin, 2015). The Neymanian mode of inference relies on asymptotic approximations for the variance of the mean-difference estimator $\hat{\tau}$; such results are well-established for completely randomized experiments (Neyman et al., 1990), paired experiments (Imai, 2008), blocked experiments (Miratrix et al., 2013; Pashley and Miratrix, 2017), and randomized experiments with stages of random sampling (Branson and Dasgupta, 2020). In a seminal paper, Li et al. (2018) derived many asymptotic results for rerandomized experiments (as discussed in Morgan and Rubin (2012)), thereby establishing Neymanian inference for such experiments. The results therein rely on various properties of the Mahalanobis distance, which – as established by our results – differ from the properties of the ridge Mahalanobis distance. As a consequence, the theory developed in Li et al. (2018) cannot be readily applied to ridge rerandomized experiments, and a promising line of future work is deriving asymptotic results for ridge rerandomized experiments. Asymptotic results could also be used to establish Bayesian inference for such experiments, which would be particularly useful given that one's preference for rerandomization or ridge rerandomization may depend on their prior knowledge of $\boldsymbol{\beta}$, as suggested by Corollary 4.2. Addressing these complications is beyond the scope of this paper. Instead, we focus on randomization-based inference, because it can be readily applied to ridge rerandomization.

Randomization-based inference focuses on inverting sharp null hypotheses that define the relationship between the potential outcomes in terms of treatment effects. The most common null hypothesis is that of an additive treatment effect τ, such that the hypothesis $H_0^\tau : Y_i(1) = Y_i(0) + \tau$ holds for all $i = 1, \ldots, N$. Confidence intervals derived from inverting this hypothesis were first established by Hodges Jr and Lehmann (1963) and have since been popularized for analyzing randomized experiments (e.g., see Rosenbaum, 2002; Imbens and Rubin, 2015). Here we briefly review how to obtain randomization-based confidence intervals for completely randomized experiments, and then we extend them to ridge rerandomized experiments.

As first proposed by Hodges Jr and Lehmann (1963), a valid randomization-based confidence interval is the set of τ such that we fail to reject H_0^τ; such inversion of a hypothesis is a classical way to obtain a confidence set (Kempthorne and Doerfler, 1969). To obtain a valid p-value for H_0^τ, a key insight is that, if H_0^τ holds, then one has full knowledge of the potential outcomes for all units: If we observe the outcome under control for a particular unit, we know that the outcome under treatment for that unit is simply the observed outcome plus τ. As a result, for any hypothetical randomization, a test statistic – such as the mean difference estimator, $\hat{\tau}$ – can be computed. To obtain a p-value for H_0^τ under randomization, one follows this simple three-step procedure:

1. Generate many hypothetical randomizations, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, by permuting the observed treatment indicator.
2. Compute a test statistic $t(\mathbf{w}, \mathbf{x}, \mathbf{y})$, such as the mean-difference estimator, across the randomizations $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ assuming H_0^τ is true.
3. Compute the randomization-based p-value, defined as

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left(|t(\mathbf{w}^{(m)}, \mathbf{x}, \mathbf{y})| > |t^{obs}|\right)}{M + 1} \tag{25}$$

where t^{obs} is the observed test statistic and $\mathbb{1}(\cdot)$ denotes the indicator function. The additional 1 in the numerator and the denominator induces a very small amount of bias in order to validly control the Type 1 error rate and is a standard correction for randomization test p-values (Phipson and Smyth, 2010). Modern statistical software allows one to readily invert H_0^τ after Step 1 is completed (in Section 5, we will use the R package `ri` (Aronow and Samii, 2012) to do this), thereby producing randomization-based confidence intervals. This makes the extension to ridge rerandomization quite straightforward: In Step 1, one generates many hypothetical *ridge* rerandomizations (instead of randomizations), and then proceeds as usual to conduct randomization-based inference. This is identical to the approach discussed in Morgan and Rubin (2012) for obtaining confidence intervals under rerandomization, except using hypothetical ridge rerandomizations instead of hypothetical rerandomizations. This can also be viewed as inverting a *conditional* randomization test, where we condition on the fact that the ridge rerandomization balance criterion has been fulfilled (Hennessy et al., 2016; Branson and Miratrix, 2019). As we shall see in Section 5, confidence intervals for ridge rerandomized experiments are much more precise than intervals for completely randomized experiments, and often more precise than intervals for rerandomized experiments, especially in high dimensional and/or collinearity settings.

5. Simulations

We now provide simulation evidence that supports the heuristic argument presented in Section 4.2 and suggests when ridge rerandomization is an effective experimental design strategy. First, we will consider conducting an experiment where covariates are linearly related with the outcome, treatment effects are additive, and the number of treated units and the number of control units are equal. Then we will consider alternative scenarios. Throughout, we will compare rerandomization and ridge rerandomization in terms of (1) their ability to balance covariates, (2) their ability to produce precise treatment effect estimators, and (3) their ability to produce precise confidence intervals. We find that ridge rerandomization is particularly preferable over rerandomization in high-dimensional or high-collinearity settings.

5.1. Simulation setup

Consider $N = 100$ units, 50 of which are to be assigned to treatment and 50 are to be assigned to control. Let $\mathbf{x}$ be a $N \times K$ covariate matrix, generated as

$$\mathbf{x} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \right) \tag{26}$$

where $0 \leq \rho < 1$. The parameter ρ corresponds to the correlation among the covariates. Furthermore, let $Y_i(1)$ and $Y_i(0)$ be the potential outcomes under treatment and control, respectively, for unit i, generated as

$$\begin{aligned} &Y_i(0) \sim N(\mathbf{x}\boldsymbol{\beta}, 1) \\ &Y_i(1) = Y_i(0) + \tau \end{aligned} \tag{27}$$

For this simulation study, we set the treatment effect to be $\tau = 1$. Across simulations, we consider number of covariates $K \in \{10, \dots, 90\}$ and correlation parameter $\rho \in \{0, 0.1, \dots, 0.9\}$. We discuss choices for $\boldsymbol{\beta}$ in Section 5.3. In Section 5.4, we discuss scenarios where covariates are nonlinearly related with the outcomes, treatment effects are non-additive, and $N_T \neq N_C$; however, the results for these other scenarios are largely the same as those for the above data-generating process, and so for ease of exposition we focus on results for the case where the covariates are generated from (26) and the potential outcomes are generated from (27).

We will consider three experimental design strategies for assigning units to treatment and control:

1. **Randomization**: Randomize 50 units to treatment and 50 to control.
2. **Rerandomization**: Randomize 50 units to treatment and 50 to control until $M \leq a$, where M is the Mahalanobis distance defined in (3).
3. **Ridge Rerandomization**: Randomize 50 units to treatment and 50 to control until $M_\lambda \leq a_\lambda$, where M_λ is the ridge Mahalanobis distance defined in (6).

For each choice of K, ρ, and $\boldsymbol{\beta}$, we ran randomization, rerandomization, and ridge rerandomization 1000 times. For rerandomization and ridge rerandomization, we set $p_a = 0.1$, which corresponds to randomizing within the 10% "best" randomizations according to the Mahalanobis distance and ridge Mahalanobis distance, respectively. Furthermore, for ridge rerandomization, we used the procedure in Section 4.2.1 for selecting λ, with $n = 1000$, $\delta = 0.01$, and $\epsilon = 10^{-4}$. The value $\lambda = 0.01$ was selected for most K and ρ, and occasionally $\lambda = 0.02$ was selected.

First, in Section 5.2, we compare how these three methods balanced the covariates $\mathbf{x}$, and so the $\boldsymbol{\beta}$ parameter in (27) is irrelevant for this section. Then, in Section 5.3, we compare the accuracy of treatment effect estimators and precision of confidence intervals for each method; in this case, the specification of $\boldsymbol{\beta}$ is consequential.

5.2. Comparing covariate balance across randomizations

First, we computed the covariate mean differences across each randomization, rerandomization, and ridge rerandomization. Fig. 2 shows how much rerandomization and ridge rerandomization reduced the variance of $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ (averaged across covariates) compared to randomization for data generated from (26). For rerandomization, the average variance reduction decreases as K increases (an observation previously made in Morgan and Rubin, 2012), and it stays largely the same across values of ρ for fixed K. As for ridge rerandomization, the average variance reduction also decreases as K increases, but the average variance reduction increases as ρ increases, i.e., as there is more collinearity in $\mathbf{x}$. Finally, the

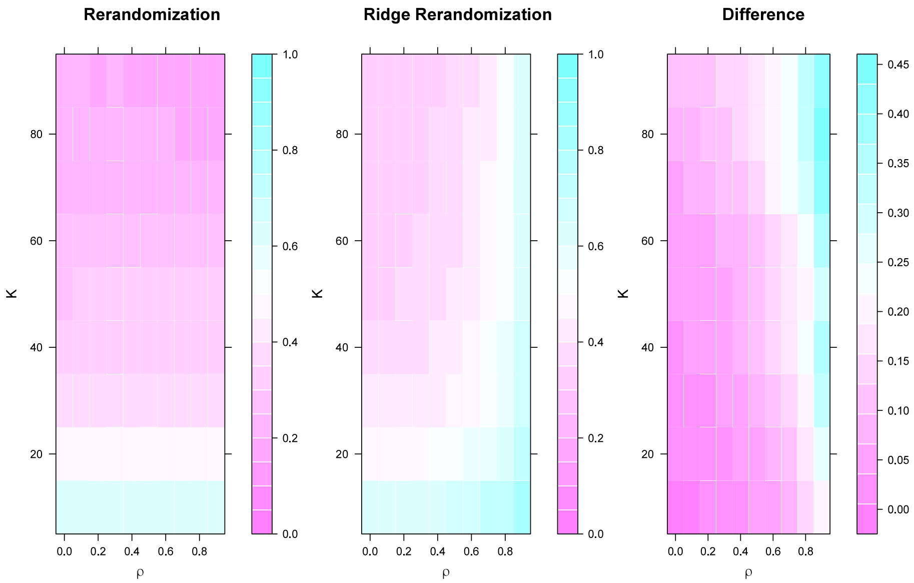


Fig. 2. Variance reduction averaged across covariates for rerandomization and ridge rerandomization, as well as their difference (ridge rerandomization minus rerandomization, i.e., the second plot minus the first).

right-hand plot in Fig. 2 shows that ridge rerandomization has a higher average variance reduction than rerandomization; furthermore, the advantage of ridge rerandomization over rerandomization increases in both K and ρ. This suggests that ridge rerandomization may be particularly preferable over rerandomization in the presence of many covariates and/or high collinearity among covariates, which is intuitive given the motivation of ridge regression (Hoerl and Kennard, 1970).

5.3. Comparing treatment effect estimation accuracy across randomizations

Reducing the variance of each covariate mean difference leads to more precise treatment effect estimates if the covariates are related to the outcome, as in (27). The extent to which the covariates are related to the outcome depends on the $\boldsymbol{\beta}$ parameter. Theorem 4.3 guarantees that ridge rerandomization will improve inference for the average treatment effect, compared to randomization, regardless of $\boldsymbol{\beta}$. However, Corollary 4.2 establishes that $\boldsymbol{\beta}$ dictates whether rerandomization or ridge rerandomization will perform better in terms of treatment effect estimation accuracy. First we will consider a $\boldsymbol{\beta}$ where the covariates are equally related to the outcome, and in this case ridge rerandomization performs better than rerandomization. Then, we will consider a $\boldsymbol{\beta}$ which – according to our theoretical results – should put ridge rerandomization in the worst light as compared to rerandomization.

5.3.1. One choice of $\boldsymbol{\beta}$

Consider $\boldsymbol{\beta} = \mathbf{1}_K$. Because the covariates have been standardized to have the same scale, such a $\boldsymbol{\beta}$ implies that all of the covariates are equally important in affecting the outcome. For each of the 1000 randomizations, rerandomizations, and ridge rerandomizations generated for each $K \in \{10, \dots, 90\}$ and $\rho \in \{0, 0.1, \dots, 0.9\}$, we computed the mean-difference estimator $\hat{\tau}$. Then, we computed the MSE of $\hat{\tau}$ across the 1000 randomizations, rerandomizations, and ridge rerandomizations for each K and ρ. Fig. 3 shows the MSE of rerandomization and ridge rerandomization relative to the MSE of randomization. A lower relative MSE represents a more accurate treatment effect estimator, compared to how that estimator would behave under randomization.

Three observations can be made about Fig. 3. First, both rerandomization and ridge rerandomization reduce the MSE of $\hat{\tau}$ compared to randomization: the relative MSE for both methods is always less than 1. Second, for rerandomization, the relative MSE stays constant across values of ρ and decreases as K decreases. Meanwhile, for ridge rerandomization, the relative MSE decreases as ρ increases and K decreases. Third, for this choice of $\boldsymbol{\beta}$, ridge rerandomization reduces the MSE of the treatment effect estimator more so than rerandomization, especially when K and/or ρ is large. These last two observations reflect the variance reduction behavior observed in Fig. 2.

Meanwhile, for each randomization, rerandomization, and ridge rerandomization, we generated a 95% confidence interval for the average treatment effect using the procedure outlined in Section 4.4. Regardless of the procedure used, coverage was near 95%. This is unsurprising, because these intervals were constructed by inverting randomization tests that are valid for their corresponding assignment mechanism; see Edgington and Onghena (2007) and Good (2013) for classical results on the validity of randomization tests. However, the width of these intervals differed across these three procedures: Fig. 4 compares the relative average interval width (compared to randomization) for rerandomization

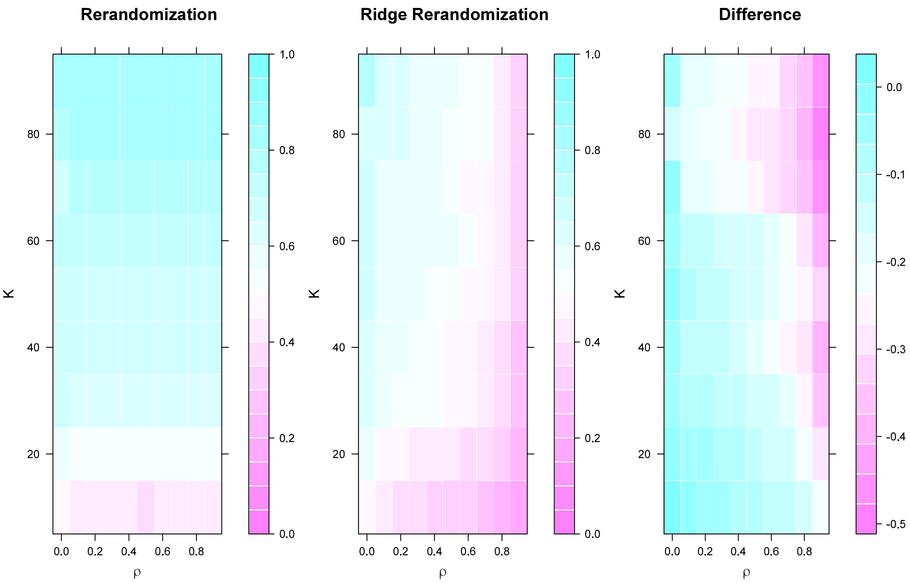


Fig. 3. Relative MSE of $\hat{\tau} = \bar{y}_T - \bar{y}_C$ under rerandomization and ridge rerandomization (relative to randomization) when $\boldsymbol{\beta} = \mathbf{1}_K$ in (27), as well as the difference in relative MSE between the two (i.e., the second plot minus the first).

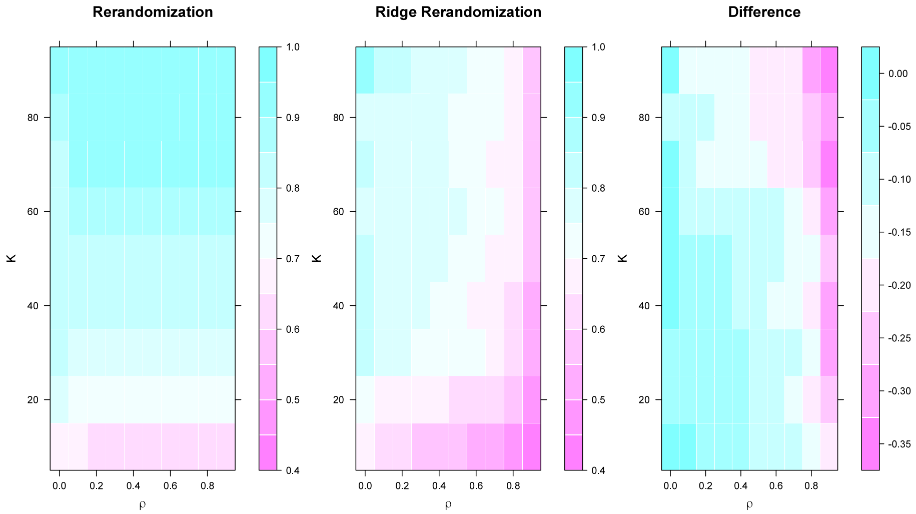


Fig. 4. Relative average 95% confidence interval width under rerandomization and ridge rerandomization (relative to randomization) when $\boldsymbol{\beta} = \mathbf{1}_K$ in (27), as well as the difference between the two (i.e., the second plot minus the first).

and ridge rerandomization. For the first two plots in Fig. 4, a number closer to 1 indicates intervals that are closer in width to intervals under randomization. Meanwhile, for the right-most plot in Fig. 4, a more negative number indicates more narrow confidence intervals for ridge rerandomization, as compared to rerandomization. The qualitative results are identical to the previous results: Ridge rerandomization tends to provide narrower confidence intervals as the covariates' dimension and/or collinearity increases.

5.3.2. A choice of $\boldsymbol{\beta}$ where ridge rerandomization has the least competitive advantage over rerandomization

As can be seen by Corollary 4.2, there may exist $\boldsymbol{\beta}$ where rerandomization performs better than ridge rerandomization. To assess how poorly ridge rerandomization can perform compared to rerandomization, now we will specify a $\boldsymbol{\beta}$ that puts ridge rerandomization in the worst light when comparing it to rerandomization in terms of treatment effect estimation accuracy.

Under the assumptions of Corollary 4.2, the difference in treatment effect estimation accuracy between rerandomization and ridge rerandomization is given by $\Delta = \boldsymbol{\beta}^{\top}\boldsymbol{\Gamma}\mathbf{Diag}\left(\left(\lambda_k\left(v_a - d_{k,\lambda}\right)\right)_{1\leq k\leq K}\right)\boldsymbol{\Gamma}^{\top}\boldsymbol{\beta}$, which can be artificially minimized

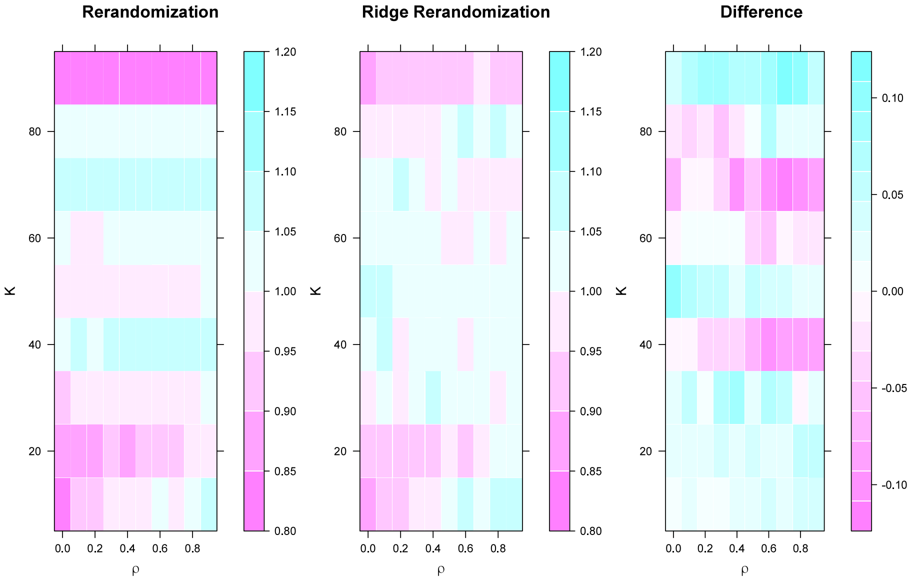


Fig. 5. Relative MSE of $\hat{\tau} = \bar{y}_T - \bar{y}_C$ under rerandomization and ridge rerandomization (relative to randomization) for the $\boldsymbol{\beta}$ such that ridge rerandomization has the least competitive advantage over rerandomization, as well as the difference in relative MSE between the two (i.e., the second plot minus the first).

with respect to $\boldsymbol{\beta}$, subject to some constraint on $\boldsymbol{\beta}$ for the minimum to exist, e.g., $\|\boldsymbol{\beta}\| \leq 1$. If $d_{k,\lambda} < v_a$ for all $k = 1, \dots, K$, then ridge rerandomization dominates rerandomization since $\Delta > 0$ for all $\boldsymbol{\beta} \neq 0$, and these schemes are only tied when $\Delta = 0$ for $\boldsymbol{\beta} = 0$, i.e., the covariates are uncorrelated with the outcomes. In other cases, we can define $\boldsymbol{\beta}^* = \boldsymbol{\Gamma}_{\bullet k^*}$ where $\boldsymbol{\Gamma}_{\bullet k^*}$ is the k^*-th column of $\boldsymbol{\Gamma}$ and $k^* = \text{argmin}_{1 \leq k \leq K}(v_a - d_{k,\lambda})$. We would typically have $k^* = K$, because the $d_{k,\lambda}$'s are non-increasing. By construction, $\boldsymbol{\beta}^*$ minimizes Δ over $\{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta}\| \leq 1\}$ and yields $\Delta < 0$ as negative as possible. This is equivalent to $\boldsymbol{\beta}$ being in the direction that accounts for the least variation in the covariates. While such a case is unlikely, we consider such a $\boldsymbol{\beta}$ to see how much worse ridge rerandomization performs as compared to rerandomization in this scenario.

Fig. 5 shows the relative MSE (as compared to randomization) for rerandomization and ridge rerandomization for this specification of $\boldsymbol{\beta}$. Interestingly, there are occasions where rerandomization and ridge rerandomization have relative MSEs greater than 1, i.e., when they perform worse than randomization in terms of treatment effect estimation accuracy. At first this may be surprising, especially when findings from Morgan and Rubin (2012) guarantee that rerandomization should perform better than randomization. However, in this case, $\boldsymbol{\beta}$ is in the direction of the last principal component of the covariate space, meaning that the covariates have nearly no relationship with the outcomes. Thus, the relative MSE that we see in the first two plots of Fig. 5 is more or less the behavior we would expect if we compared 1000 randomizations to 1000 other randomizations. Furthermore, from the third plot in Fig. 5, we can see that rerandomization occasionally performs better than ridge rerandomization – particularly when K is small – but the differences in relative MSE across simulations are somewhat centered around zero. Meanwhile, Fig. 6 compares the relative average confidence interval width for rerandomization and ridge rerandomization, and the qualitative results are largely the same as the relative MSE results: Rerandomization and ridge rerandomization are fairly comparable, but rerandomization tends to provide slightly narrower confidence intervals for low-dimensional covariates.

Note that this specification of $\boldsymbol{\beta}$ is a unit vector. We could have scaled $\boldsymbol{\beta}$ arbitrarily large, and, as a result, the differences in the last plots of Figs. 5 and 6 could have been made arbitrarily large. Thus, ridge rerandomization can perform much worse than rerandomization when $\boldsymbol{\beta}$ exhibits particularly large effects in the direction of the last principal component of the covariate space, especially when the number of covariates is small. Practically speaking, such a scenario is unlikely, but it is a scenario that researchers should acknowledge and consider when comparing rerandomization and ridge rerandomization.

5.4. Additional simulations: Unequal sample sizes, nonlinearity, heterogeneous treatment effects, and rank deficiency

In the above, we considered scenarios where an equal number of units are assigned to treatment and control, covariates are linearly related with the potential outcomes, and treatment effects are additive. In Appendix A.8, we present simulation results for scenarios where $N_T \neq N_C$, covariates are nonlinearly related with the potential outcomes, and treatment effects are heterogeneous. The results presented therein are very similar to the results presented above: Rerandomization and

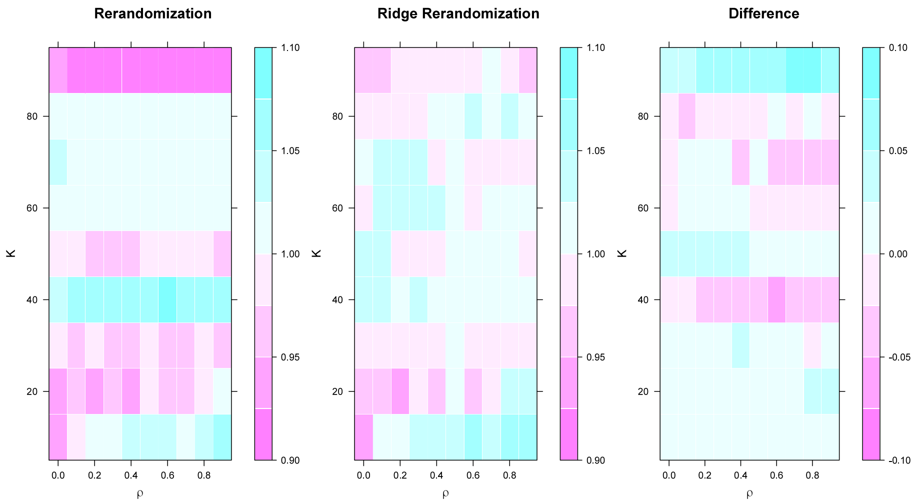


Fig. 6. Relative average 95% confidence interval width under rerandomization and ridge rerandomization (relative to randomization) for the $\boldsymbol{\beta}$ such that ridge rerandomization has the least competitive advantage over rerandomization, as well as the difference between the two (i.e., the second plot minus the first).

ridge rerandomization are still preferable over randomization, and ridge rerandomization is preferable over rerandomization in high-dimensional and/or high-collinearity scenarios. We found that ridge rerandomization's advantage over rerandomization was somewhat diminished when treatment and control sample sizes were highly unequal or when covariates were nonlinearly related with the potential outcomes, but the advantage in high-dimensional and/or high-collinearity scenarios was still clear. Due to the similarity of these results, we relegated these additional simulations to the Appendix.

Finally, note that all of our previous simulation studies focused on the case where $N = 100$ and $K \in \{10, 20, \ldots, 90\}$. In this case, the covariance matrix $\boldsymbol{\Sigma}$ is always invertible, which we have assumed throughout the manuscript. When $N \leq K$, $\boldsymbol{\Sigma}$ is not invertible, the Mahalanobis distance is undefined, and rerandomization cannot be implemented. However, the ridge Mahalanobis distance M_λ in (6) is still defined, and ridge rerandomization can still be implemented. In Appendix A.8, we present simulation results when $N = 100$ and $K = 101$, and we again find that ridge rerandomization is preferable over randomization, especially in high-collinearity scenarios. This suggests that ridge rerandomization may be a viable experimental design strategy when $N \leq K$, and interesting future work would be establishing theoretical results for ridge rerandomization even when $\boldsymbol{\Sigma}$ is not invertible but the ridge Mahalanobis distance is still defined.

5.5. Summary of simulation results

Importantly, the effectiveness of rerandomization or ridge rerandomization in balancing the covariates does not depend on the covariates' relationship with the outcomes. In other words, the variance reduction results in Fig. 2 do not depend on $\boldsymbol{\beta}$, whereas the treatment effect estimation accuracy results in Figs. 3 and 5 and confidence interval results in Figs. 4 and 6 do. From Fig. 2 we see that ridge rerandomization appears to generally be more effective than rerandomization in balancing covariates in high-dimensional or high-collinearity settings, and from Figs. 3 and Fig. 4 we see that this can result in more precise treatment effect estimators and confidence intervals. These results also hold when treatment and control sample sizes are unequal, the outcome is nonlinearly related with the covariates, or when there is treatment effect heterogeneity, as discussed briefly in Section 5.4 and more fully in Appendix A.8. However, from Section 5.3.2, we see that there are cases where rerandomization can perform better than ridge rerandomization in terms of treatment effect estimation. In particular, if the relationship between the covariates and the outcome is strongly in the direction of the last principal component of the covariate space, rerandomization can perform arbitrarily better than ridge rerandomization, especially when there are only a few number of covariates. In general, the comparison between rerandomization and ridge rerandomization depends on the relationship between the covariates and the outcomes, which is typically not known until after the experiment is conducted.

In summary, these simulations suggest that ridge rerandomization is often preferable over rerandomization by targeting the directions that best explain variation in the covariates rather than the covariates themselves. If the covariates are related to the outcomes (linearly or nonlinearly), ridge rerandomization appears to be an appealing experimental design strategy when there are many covariates and/or highly collinear covariates.

6. Discussion and conclusion

The rerandomization literature has focused on experimental design strategies that utilize the Mahalanobis distance. Starting with Morgan and Rubin (2012) and continuing with works such as Morgan and Rubin (2015), Branson et al. (2016), Zhou et al. (2018), and Li et al. (2018), many theoretical results have been established for rerandomization schemes using the Mahalanobis distance. However, the Mahalanobis distance is known to not perform well in high dimensions or when there are strong collinearities among covariates—settings which the current rerandomization literature has not addressed.

To address experimental design settings where there are many covariates or strong collinearities among covariates, we presented a rerandomization scheme that utilizes a modified Mahalanobis distance. This modified Mahalanobis distance inflates the eigenvalues of the covariance matrix of the covariates, thereby increasing the variance reduction of the covariates' first principal components at the expense of decreasing the variance reduction of the last principal components. Such a quantity has remained largely unexplored in the literature. We established several theoretical properties of this modified Mahalanobis distance, as well as properties of a rerandomization scheme that uses it—an experimental design strategy we call ridge rerandomization. These results establish that ridge rerandomization preserves the unbiasedness of treatment effect estimators and reduces the variance of covariate mean differences. If the covariates are related to the outcomes of the experiment, ridge rerandomization will yield more precise treatment effect estimators than randomization. Furthermore, we conducted a simulation study that suggests that ridge rerandomization is often preferable over rerandomization in high-dimensional or high-collinearity scenarios, which is intuitive given ridge rerandomization's connections to ridge regression.

This modified Mahalanobis distance represents a class of rerandomization criteria, which has connections to principal components and the Euclidean distance. This motivates future work for rerandomization schemes that utilize other criteria. In particular, our theoretical results establish that the benefit of our class of rerandomization schemes over typical rerandomization depends on the covariates' relationship with the outcomes, which usually is not known until after the experiment has been conducted. However, if researchers have prior information about the relationship between the covariates and the outcomes, this information may be useful in selecting rerandomization criteria. An interesting line of future work is further exploring other classes of rerandomization criteria, as well as demonstrating how prior outcome information can be used to select useful rerandomization criteria when designing an experiment.

CRediT authorship contribution statement

Zach Branson: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Stephane Shao:** Conceptualization, Formal analysis, Investigation.

Acknowledgments

We would like to thank Espen Bernton and Tirthankar Dasgupta for insightful comments throughout the progress of this work. We would also like to especially thank Evan Greif for discussions that initially motivated this work. This research was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1144152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A

A.1. Proof of Lemma 4.1

Since $\mathbf{\Sigma} > 0$, it is invertible and we can write

$$(\mathbf{\Sigma} + \lambda\, I_K)^{-1} = \mathbf{\Sigma}^{-\frac{1}{2}}(I_K + \lambda\, \mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-\frac{1}{2}}$$

so that

$$M_\lambda = \widetilde{\mathbf{Z}}^\top (I_K + \lambda\, \mathbf{\Sigma}^{-1})^{-1}\widetilde{\mathbf{Z}}$$

where $\widetilde{\mathbf{Z}} = \mathbf{\Sigma}^{-\frac{1}{2}}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$. Thanks to the assumed Normality of $\widetilde{\mathbf{Z}} \mid \mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_K)$, we may write

$$M_\lambda \mid \mathbf{x} \;\sim\; \mathbf{Z}^\top (I_K + \lambda\, \mathbf{\Sigma}^{-1})^{-1}\mathbf{Z}$$

where $\mathbf{Z} = (Z_1 \dots Z_K)^\top \sim \mathcal{N}(0, \mathbf{1}_K)$ marginally and independently of $\mathbf{x}$. The matrix $(I_K + \lambda\, \mathbf{\Sigma}^{-1})^{-1}$ shares the same orthonormal basis x of eigenvectors $\mathbf{\Gamma}$ as $\mathbf{\Sigma}$, with corresponding eigenvalues $\lambda_1(\lambda_1 + \lambda)^{-1}, \dots, \lambda_K(\lambda_K + \lambda)^{-1}$. As a consequence, we have

$$M_\lambda \mid \mathbf{x} \;\sim\; (\mathbf{\Gamma}^\top \mathbf{Z})^\top\, \mathbf{Diag}\left(\left(\frac{\lambda_j}{\lambda_j + \lambda}\right)_{1 \le j \le K}\right)(\mathbf{\Gamma}^\top \mathbf{Z}) \tag{28}$$

Since $(\mathbf{\Gamma}^\top \mathbf{Z}) \sim \mathcal{N}(0, \mathbf{\Gamma}^\top \mathbf{\Gamma}) \sim \mathcal{N}(0, \mathbf{I}_K) \sim \mathbf{Z}$ by orthogonality of $\mathbf{\Gamma}$, we get

$$M_\lambda \mid \mathbf{x} \;\sim\; \mathbf{Z}^\top\, \mathbf{Diag}\left(\left(\frac{\lambda_j}{\lambda_j + \lambda}\right)_{1 \le j \le K}\right)\mathbf{Z} \;=\; \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda} Z_j^2$$

where $Z_1, \dots, Z_K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $\lambda_1 \ge \cdots \ge \lambda_K > 0$ are the eigenvalues of $\mathbf{\Sigma}$. □

A.2. Proof of Lemma 4.2

Without loss of generality, let $K = 2$. Thus, the aim of this proof is to establish that $E_1 \leq E_2$, i.e.,

$$\mathbb{E}\left[L_1 \middle| C_1L_1 + C_2L_2 \leq a\right] \leq \mathbb{E}\left[L_2 \middle| C_1L_1 + C_2L_2 \leq a\right] \tag{29}$$

where L_1 and L_2 are independent and identically distributed non-negative random variables, $C_1 \geq C_2 \geq 0$ are constants, and $a > 0$ is a constant.

First, it will be helpful to note that the event $C_1L_1 + C_2L_2 \leq a$ can be partitioned into two events:

$$A = \{C_1L_1 + C_2L_2 \leq a, C_1L_2 + C_2L_1 \leq a\}$$
$$B = \{C_1L_1 + C_2L_2 \leq a, C_1L_2 + C_2L_1 > a\}$$

In other words, $A \cup B$ is equal to the event $C_1L_1 + C_2L_2 \leq a$. Thus,

$$\mathbb{E}\left[L_1 \middle| C_1L_1 + C_2L_2 \leq a\right] = \mathbb{E}[L_1|A]P(A) + \mathbb{E}[L_1|B]P(B) \tag{30}$$

and analogously for L_2.

Now note that if $C_1L_1 + C_2L_2 \leq a$ and $L_1 \geq L_2$, then $C_1L_2 + C_2L_1 \leq a$ and thus B cannot occur. To see this, note that if $L_1 \geq L_2$, then $C_2(L_1 - L_2) - C_1(L_1 - L_2) \leq 0$, because $C_1 \geq C_2 \geq 0$, and therefore:

$$\begin{aligned} C_1L_2 + C_2L_1 &= C_1L_1 + C_2L_2 + C_2(L_1 - L_2) - C_1(L_1 - L_2) \\ &\leq C_1L_1 + C_2L_2 \\ &\leq a \end{aligned}$$

In other words, B will only occur if $L_1 < L_2$, and therefore $\mathbb{E}[L_1|B] < \mathbb{E}[L_2|B]$.

Meanwhile, due to the symmetry of L_1 and L_2 in the two constraints in A, $\mathbb{E}[L_1|A] = \mathbb{E}[L_2|A]$. Thus, revisiting (30), we have the following:

$$\begin{aligned} \mathbb{E}\left[L_1 \middle| C_1L_1 + C_2L_2 \leq a\right] &= \mathbb{E}[L_1|A]P(A) + \mathbb{E}[L_1|B]P(B) \\ &= \mathbb{E}[L_2|A]P(A) + \mathbb{E}[L_1|B]P(B) \\ &\leq \mathbb{E}[L_2|A]P(A) + \mathbb{E}[L_2|B]P(B) \\ &= \mathbb{E}\left[L_2 \middle| C_1L_1 + C_2L_2 \leq a\right] \end{aligned}$$

which completes the proof. For $K > 2$, the same application of the proof applies, with the only difference being partitioning the event $\sum_{j=1}^{K} C_jL_j \leq a$ into $2(K! - 1)$ events. □

A.3. Proof of Theorem 4.2

Using the same notation and reasoning as for the proof of Lemma 4.1 in Appendix A.1, in particular (28), we can write

$$\begin{aligned} &\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \leq a_\lambda) \\ =\ & \mathrm{Cov}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{Z} \,\middle|\, \mathbf{x}, \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda}(\boldsymbol{\Gamma}^\top\mathbf{Z})_j^2 \leq a_\lambda\right) \\ =\ & \mathrm{Cov}\left(\boldsymbol{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)(\boldsymbol{\Gamma}^\top\mathbf{Z}) \,\middle|\, \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda}(\boldsymbol{\Gamma}^\top\mathbf{Z})_j^2 \leq a_\lambda\right) && (31) \\ =\ & \boldsymbol{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathrm{Cov}\left((\boldsymbol{\Gamma}^\top\mathbf{Z}) \,\middle|\, \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda}(\boldsymbol{\Gamma}^\top\mathbf{Z})_j^2 \leq a_\lambda\right)\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\boldsymbol{\Gamma}^\top \\ =\ & \boldsymbol{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda}Z_j^2 \leq a_\lambda\right)\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\boldsymbol{\Gamma}^\top && (32) \end{aligned}$$

where (31) follows from the definition of $\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\boldsymbol{\Gamma}^\top$ along with the constructed independence of $\mathbf{Z}$ and $\mathbf{x}$ to get rid of the conditioning on $\mathbf{x}$, and (32) follows from $(\boldsymbol{\Gamma}^\top\mathbf{Z}) \sim \mathbf{Z}$ by orthogonality of $\boldsymbol{\Gamma}$ and standard Normality of $\mathbf{Z}$.

All that is left now is to compute the conditional covariance matrix appearing in (32). Starting by its diagonal elements, the symmetry of the Normal distribution ensures that $\mathbf{Z} \sim -\mathbf{Z}$, which implies

$$\begin{aligned}\mathbb{E}\left[Z_k \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right] &= \mathbb{E}\left[-Z_k \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} (-Z_j)^2 \le a_\lambda \right] \\ &= -\mathbb{E}\left[Z_k \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} (Z_j)^2 \le a_\lambda \right]\end{aligned}$$

for all $k = 1, \ldots, K$, so that

$$\mathbb{E}\left[Z_k \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right] = 0$$

Thus, the diagonal elements $d_{k,\lambda}$ of $\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right)$ are given by

$$d_{k,\lambda} = \mathrm{Var}\left(Z_k^2 \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right) = \mathbb{E}\left[Z_k^2 \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right] \tag{33}$$

for all $k = 1, \ldots, K$. Now for the (ℓ, m)-element of $\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right)$ with $\ell \ne m$, we use again the symmetry of the Normal distribution by noticing that $\mathbf{Z} \sim \mathbf{Z}^*$, where we define $Z_i^* = Z_i$ for all $i \ne \ell$ and $Z_\ell^* = -Z_\ell$, so that

$$\begin{aligned}\mathrm{Cov}\left(Z_\ell, Z_m \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right) &= \mathrm{Cov}\left(Z_\ell^*, Z_m^* \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} (Z_j^*)^2 \le a_\lambda \right) \\ &= -\mathrm{Cov}\left(Z_\ell, Z_m \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right)\end{aligned}$$

which leads to

$$\mathrm{Cov}\left(Z_\ell, Z_m \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right) = 0 \tag{34}$$

for all $1 \le \ell, m \le K$ such that $\ell \ne m$. Combining (33) and (34) gives

$$\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right) = \mathbf{Diag}\left((d_{k,\lambda})_{1 \le k \le K} \right) \tag{35}$$

Plugging (35) back into (32) finally yields

$$\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \le a_\lambda) = \mathbf{\Gamma}\mathbf{Diag}((\lambda_k\, d_{k,\lambda})_{1 \le k \le K})\mathbf{\Gamma}^\top.$$

where the $d_{k,\lambda}$'s are given by (33). From the expression of $d_{k,\lambda}$, we immediately have $d_{k,\lambda} > 0$ for all $k = 1, \ldots, K$. By using Equation (13) from Palombi and Toti (2013), we also get

$$\mathbb{E}\left[Z_k^2 \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda \right] < \mathbb{E}\left[Z_k^2 \right] = 1$$

for all $k = 1, \ldots, K$. Therefore, we have $d_{k,\lambda} \in (0, 1)$ for all $k = 1, \ldots, K$. □

A.4. Proof of Corollary 4.1

By definition of $v_{k,\lambda}$ and by Theorem 4.2, we have

$$\begin{aligned}v_{k,\lambda} = \frac{\mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x}, M_\lambda \le a_\lambda \right)}{\mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x} \right)} &= \frac{\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \le a_\lambda)_{kk}}{\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x})_{kk}} \\ &= \frac{\left(\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, d_{j,\lambda})_{1 \le j \le K} \right) \mathbf{\Gamma}^\top \right)_{kk}}{\mathbf{\Sigma}_{kk}}.\end{aligned}$$

Since $\lambda_j(1-d_{j,\lambda}) > 0$ for all $j = 1, \dots, K$, the matrix

$$\mathbf{\Sigma} - \mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, d_{j,\lambda})_{1\le j\le K}\right)\mathbf{\Gamma}^\top = \mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\,(1-d_{j,\lambda}))_{1\le j\le K}\right)\mathbf{\Gamma}^\top$$

is positive definite. This implies that

$$\mathbf{v}^\top\left(\mathbf{\Sigma} - \mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, d_{j,\lambda})_{1\le j\le K}\right)\mathbf{\Gamma}^\top\right)\mathbf{v} \;>\; 0 \tag{36}$$

for all $\mathbf{v} \in \mathbb{R}^K\backslash\{0\}$. In particular, by using (36) with $\mathbf{v}$ chosen to be the kth canonical basis vector of $\mathbb{R}^K$ (whose elements are all 0 except its kth element equal to 1), we get, for all $k = 1, \dots, K$,

$$\mathbf{\Sigma}_{kk} \;>\; \left(\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, d_{j,\lambda})_{1\le j\le K}\right)\mathbf{\Gamma}^\top\right)_{kk}. \tag{37}$$

These terms being strictly positive, this leads to $\upsilon_{k,\lambda} \in (0, 1)$ for all $j = 1, \dots, K$, i.e.

$$\mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x}, M_\lambda \le a_\lambda\right) \;<\; \mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \mid \mathbf{x}\right) \quad \square$$

A.5. Proof of Theorem 4.3

By using (19), we can write

$$\hat{\tau} \;=\; (\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C) \;=\; \tau + \boldsymbol{\beta}^\top(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) + (\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C) \tag{38}$$

By conditional independence of $(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C)$ and $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ given $\mathbf{x}$, we have

$$\begin{aligned}\mathrm{Var}(\hat{\tau} \mid \mathbf{x}) &= \mathrm{Var}(\boldsymbol{\beta}^\top(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x}) + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \mid \mathbf{x}) \\ &= \boldsymbol{\beta}^\top\mathbf{\Sigma}\boldsymbol{\beta} + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \mid \mathbf{x}) \qquad (39)\end{aligned}$$

Conditional on $\mathbf{x}$, M_λ is a deterministic function of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$, thus $(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C)$ is conditionally independent of M_λ given $\mathbf{x}$. This leads to

$$\begin{aligned}\mathrm{Var}(\hat{\tau} \mid \mathbf{x}, M_\lambda \le a_\lambda) &= \mathrm{Var}(\boldsymbol{\beta}^\top(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \mid \mathbf{x}, M_\lambda \le a_\lambda) + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \mid \mathbf{x}, M_\lambda \le a_\lambda) \\ &= \boldsymbol{\beta}^\top\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \mid \mathbf{x}, M_\lambda \le a_\lambda)\boldsymbol{\beta} + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \mid \mathbf{x}) \qquad (40) \\ &= \boldsymbol{\beta}^\top\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_k d_{k,\lambda})_{1\le k\le K}\right)\mathbf{\Gamma}^\top\boldsymbol{\beta} + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \mid \mathbf{x}) \qquad (41)\end{aligned}$$

where (40) follows from the conditional independence of $(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C)$ and M_λ given $\mathbf{x}$, and (41) follows from Theorem 4.2. By plugging (39) into (41), we get

$$\begin{aligned}\mathrm{Var}(\hat{\tau} \mid \mathbf{x}) - \mathrm{Var}(\hat{\tau} \mid \mathbf{x}, M_\lambda \le a_\lambda) &= \boldsymbol{\beta}^\top(\mathbf{\Sigma} - \mathbf{\Gamma}\mathbf{Diag}\left((\lambda_k d_{k,\lambda})_{1\le k\le K}\right)\mathbf{\Gamma}^\top)\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^\top\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_k\left(1 - d_{k,\lambda}\right))_{1\le k\le K}\right)\mathbf{\Gamma}^\top\boldsymbol{\beta}\end{aligned}$$

As explained by (36) in the proof of Corollary 4.1, the positive definiteness of the matrix $\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_k\left(1 - d_{k,\lambda}\right))_{1\le k\le K}\right)\mathbf{\Gamma}^\top$ guarantees that

$$\mathrm{Var}(\hat{\tau} \mid \mathbf{x}, M_\lambda \le a_\lambda) \;\le\; \mathrm{Var}(\hat{\tau} \mid \mathbf{x})$$

for all $\boldsymbol{\beta} \in \mathbb{R}^K$, with equality if and only if $\boldsymbol{\beta} = 0$. $\square$

A.6. Calibration of a_λ and $d_{k,\lambda}$

Here we discuss how to compute the threshold a_λ after the acceptance probability p_a and the regularization parameter λ are set. We also discuss how to approximate the $d_{k,\lambda}$'s in (11) via Monte Carlo.

A.6.1. Estimating a_λ

As discussed in Lemma 4.1 and Section 4.2, the distribution of the ridge Mahalanobis distance M_λ can be approximated as a weighted sum of independent χ^2_1 random variables. Thus, we set a_λ equal to the p_a-quantile of this weighted sum, defined as Q_λ in (20).

Let $F_{Q_\lambda}(q) = \mathbb{P}(Q_\lambda \le q)$ denote the CDF of Q_λ. Since Q_λ is a weighted sum of independent χ^2_1 variables, its characteristic function ϕ_{Q_λ} is given by $\phi_{Q_\lambda}(t) = \prod_{k=1}^K[1 - 2i\lambda_k(\lambda_k + \lambda)^{-1}t]^{-1/2}$, which can then be inverted to yield

$$F_{Q_\lambda}(q) = \lim_{U\to+\infty} F_{Q_\lambda,U}(q)$$

where

$$F_{Q_\lambda,U}(q) = \frac{1}{2} - \frac{1}{\pi}\int_0^U \frac{\sin\left(\frac{1}{2}\left[-t\,q + \sum_{k=1}^K \arctan\left(\frac{\lambda_k}{\lambda_k+\lambda}\,t\right)\right]\right)}{t\,\prod_{k=1}^K\left[1 + \left(\frac{\lambda_k}{\lambda_k+\lambda}\right)^2 t^2\right]^{1/4}}\,dt \tag{42}$$

as detailed in Equation (3.2) of Imhof (1961). In practice, for any fixed $U \geq 0$, $F_{Q_\lambda,U}(q)$ can be computed with arbitrary precision and at a negligible cost by using any (deterministic) univariate numerical integration scheme. We can then approximate $F_{Q_\lambda}(q)$ with $F_{Q_\lambda,U}(q)$ by choosing U large enough. As explained in Imhof (1961), the approximation tends to improve as the number of covariates K increases, and one can guarantee a truncation error of at most $\xi > 0$ in absolute value by choosing $U_\xi = [\xi\, \pi\, (K/2) \prod_{k=1}^{K} \sqrt{\lambda_k(\lambda_k + \lambda)^{-1}}]^{-2/K}$. More recent algorithms for approximating $F_{Q_\lambda}(q)$ include Davies (1980) and Bausch (2013), and computationally cheaper but less accurate alternatives to approximate F_{Q_λ} are discussed in Bodenham and Adams (2016).

Finally, we approximate the p_a-quantile of Q_λ by

$$\hat{a}_\lambda = \inf\{q \in \mathbb{R} : F_{Q_\lambda,U_\xi}(q) \geq p_a\} \tag{43}$$

i.e., the p_a-quantile of $F_{Q_\lambda,U}$. The hat on $\hat{a}_\lambda$ only reflects the distributional approximation of M_λ by Q_λ, whereas the errors due to numerical integration and truncation can be regarded as virtually nonexistent compared to the Monte Carlo errors involved in the later approximations of $v_{k,\lambda}$. In the simulations of Section 5, we use $\xi = 10^{-4}$ by default.

A.6.2. Estimating $d_{k,\lambda}$

As discussed in Section 4.2, choosing λ depends on the $d_{k,\lambda}$'s defined in (11), which involve intractable conditional expectations. By considering n simulated sets of K independent variables $\widetilde{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, K$, the expectations appearing in (11) can be consistently estimated via Monte Carlo, for all $k = 1, \ldots, K$, by

$$\hat{d}_{k,\lambda} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}(M_\lambda^{(i)} \leq \hat{a}_\lambda)} \sum_{i=1}^{n} \widetilde{Z}_{ik}^2\, \mathbb{1}(M_\lambda^{(i)} \leq \hat{a}_\lambda) \tag{44}$$

with $M_\lambda^{(i)} = \sum_{j=1}^{K} \lambda_k(\lambda_k + \lambda)^{-1} \widetilde{Z}_{ij}^2$ and $\hat{a}_\lambda$ defined in (43), where $\mathbb{1}(A)$ denotes the indicator function of an event A.

We regard the computational cost of generating nK independent Normal variables as negligible compared to the expected cost of generating $1/p_a$ successive random assignment vectors and testing the acceptability of each assignment, since the former can be done in parallel at virtually the same cost as generating one single Normal random variable.

A.7. Details on procedure for finding a desirable $\lambda \geq 0$

Here we discuss the details of the procedure outlined in Section 4.2, specifically Steps 3 and 4 of that procedure.

The justification of our proposed procedure stems from the following facts. By definition, we have $\mathbb{P}(M_\lambda \leq a_\lambda \mid \mathbf{x}) = p_a$ for all $\lambda \geq 0$. By taking the limit as $\lambda \to +\infty$ under the assumptions of Lemma 4.1, we get

$$p_a = \lim_{\lambda\to+\infty} \mathbb{P}\left(\sum_{k=1}^{K} \frac{\lambda_k}{\lambda_k + \lambda} Z_k^2 \leq a_\lambda\right) = \lim_{\lambda\to+\infty} \mathbb{P}\left(\sum_{k=1}^{K} \lambda_k Z_k^2 \leq \lambda\, a_\lambda\right)$$

so that

$$\lambda\, a_\lambda \xrightarrow[\lambda\to+\infty]{} q^*(p_a) \tag{45}$$

where $q^*(p_a)$ is the p_a-quantile of the distribution of $\sum_{k=1}^{K} \lambda_k Z_k^2$. This in turn implies that, for all $k = 1, \ldots, K$, we have

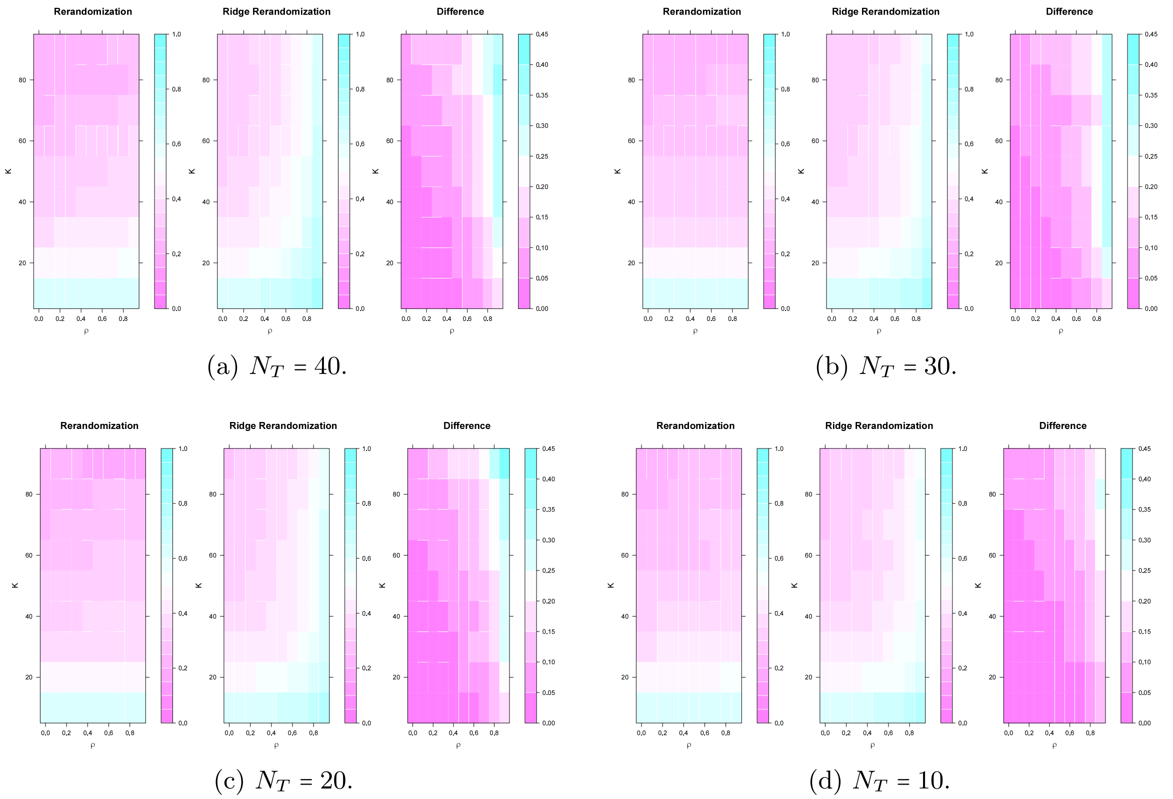
$$v_{k,\lambda} \xrightarrow[\lambda\to+\infty]{} \frac{\left(\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, d_j^*)_{1\leq j\leq K}\right)\mathbf{\Gamma}^\top\right)_{kk}}{\mathbf{\Sigma}_{kk}} \tag{46}$$

where $d_k^* = \mathbb{E}\left[Z_k^2 | \sum_{k=1}^{K} \lambda_k Z_k^2 \leq q^*(p_a)\right]$ for all $k = 1, \ldots, K$. Since the limits in (46) are strictly positive, this shows that increasing λ beyond a certain value will no longer yield any practical gain. This is in line with the intuition that the ridge Mahalanobis distance degenerates to the Euclidean distance when $\lambda \to +\infty$, as discussed in Section 4.3. Thus, in practice, it is sufficient to search for λ only over a bounded range of values. The lower bound $\lambda = 0$ corresponds to rerandomization with the standard Mahalanobis distance; the upper bound is determined dynamically via Step 3, which is guaranteed to stop in finite time by using an argument similar to (45). As mentioned in Section 4.2, the step size δ can be chosen as a fraction of the smallest strictly positive gap between consecutive eigenvalues, i.e., $\min\{\lambda_k - \lambda_{k-1} : k = 1, \ldots, K$ such that $\lambda_k > \lambda_{k-1}\}$ with the convention $\lambda_0 = 0$. Finally, among all the acceptable λ's satisfying (22), Step 4 returns the $\lambda_\star$ that aims at altering the conditional covariance structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ the least, in the sense of minimizing the distance between $\text{Cov}\big(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{x}, M_\lambda \leq \hat{a}_\lambda\big)$ and the linear span of $\mathbf{\Sigma}$, i.e.,

$$\lambda_\star = \underset{\lambda\in\Lambda}{\text{argmin}} \left(\min_{c\in\mathbb{R}} \left\| c\mathbf{\Sigma} - \mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, \hat{d}_{j,\lambda})_{1\leq j\leq K}\right)\mathbf{\Gamma}^\top \right\|\right)$$

where $\|\mathbf{\Sigma}\| = \sqrt{\text{tr}(\mathbf{\Sigma}^\top\mathbf{\Sigma})} = \sum_{k=1}^{K} \lambda_k^2$ stands for the Frobenius norm, and $\hat{a}_\lambda$ and the $\hat{d}_{j,\lambda}$'s are defined in (43) and (44), respectively. The inner minimization can be written as

$$\min_{c\in\mathbb{R}} \left(\sum_{k=1}^{K} \lambda_k^2 \left(c - \hat{d}_{k,\lambda}\right)^2\right)$$



(a) $N_T = 40$.

(b) $N_T = 30$.

(c) $N_T = 20$.

(d) $N_T = 10$.

Fig. 7. Average variance reduction for rerandomization and ridge rerandomization, as well as their difference (ridge rerandomization minus rerandomization, i.e., the second plot minus the first) for $N_T \in \{10, 20, 30, 40\}$. This is analogous to Fig. 2, but for different values of N_T.

which is attained at $c_\star = \sum_{k=1}^{K} c_k \hat{d}_{k,\lambda}$ with $c_k = \lambda_k^2 (\sum_{j=1}^{K} \lambda_j^2)^{-1}$ for all $k = 1, \ldots, K$, thus yielding Eq. (23). The outer minimization is then straightforward since the set Λ of candidates is finite by construction.

Finally, note that our procedure relies on computing $\hat{a}_\lambda$ and the $\hat{d}_{j,\lambda}$'s; these quantities rely on nK auxiliary Normal variables $\widetilde{Z}_{ij}$, which only need to be simulated once and can then be reused when testing different values of λ.

A.8. Additional simulations: Unequal sample sizes, nonlinearity, treatment effect heterogeneity, and rank deficiency

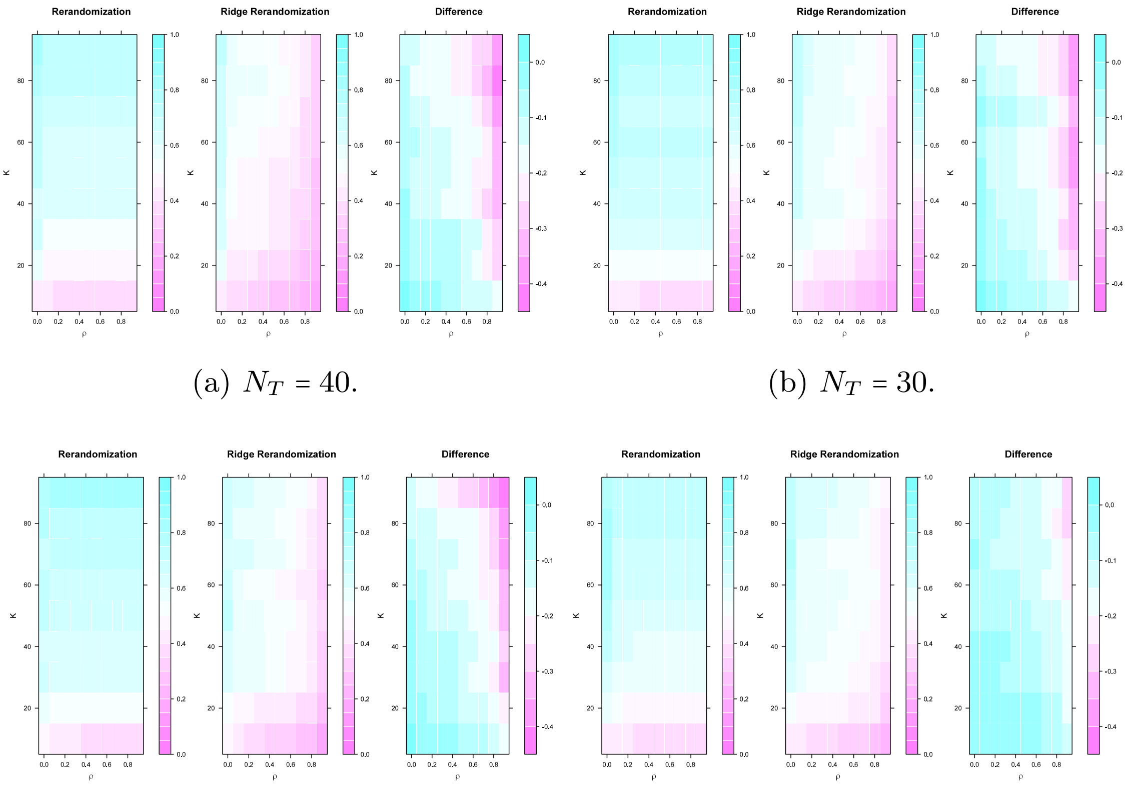
In Section 5 we considered scenarios where an equal number of units are assigned to treatment and control, covariates are linearly related with the potential outcomes, and treatment effects are additive. In this section, we provide additional simulation results for other scenarios. However, the results presented here are largely the same as those presented in Section 5 — i.e., both rerandomization and ridge rerandomization are preferable over randomization, and ridge rerandomization is preferable over rerandomization in high-dimensional and/or high-collinearity scenarios.

A.8.1. Unequal sample sizes

Similar to Section 5, we consider $N = 100$ units to be assigned to treatment and control. For each unit, the covariate matrix $\mathbf{x}$ is still generated with (26) and the potential outcomes are generated with (27), as in Section 5. However, unlike in Section 5, when implementing randomization, rerandomization, and ridge rerandomization, $N_T \neq 50$ units will be assigned to treatment and $100 - N_T$ units will be assigned to control.

We will consider $N_T \in \{10, 20, 30, 40\}$, where smaller N_T denotes more unequal sample sizes between treatment and control. Similar to Section 5, we will consider collinearity $\rho \in \{0, 0.1, \ldots, 0.9, 1.0\}$ for (26), and treatment effect $\tau = 1$ and coefficients $\boldsymbol{\beta} = \mathbf{1}_K$ for (27). We will run randomization, rerandomization, and ridge rerandomization 1000 times for each setting, and then we will compare rerandomization and ridge rerandomization in terms of (1) the average reduction in variance across covariates, (2) relative MSE for the average treatment effect, and (3) relative average 95% confidence interval width for the average treatment effect. Here, "relative" means relative to randomization.

Figs. 7, 8, and 9 show the simulation results for average reduction in variance, relative MSE, and relative average confidence interval width, respectively. These figures are analogous to Section 5 Figs. 2, 3, 4, but for $N_T \neq 50$. The results in these figures are nearly identical to those presented in Section 5: By focusing on the "Difference" plots,



(a) N_T = 40.

(b) N_T = 30.

(c) N_T = 20.

(d) N_T = 10.

Fig. 8. Relative MSE of $\hat{\tau} = \bar{y}_T - \bar{y}_C$ under rerandomization and ridge rerandomization (relative to randomization), as well as the difference between the two (i.e., the second plot minus the first) for $N_T \in \{10, 20, 30, 40\}$. This is analogous to Fig. 3, but for different values of N_T.

we see that ridge rerandomization tends to have (1) a higher average variance reduction, (2) lower relative MSE, and (3) lower relative average confidence interval width, especially in high-dimensional and/or high-collinearity settings, even if the treatment and control sample sizes are unequal. The $N_T = 10$ subfigures suggest that ridge rerandomization's advantage over rerandomization may be slightly dimensioned when N_T and N_C are highly unequal, but nonetheless ridge rerandomization appears preferable when K and/or ρ are large.

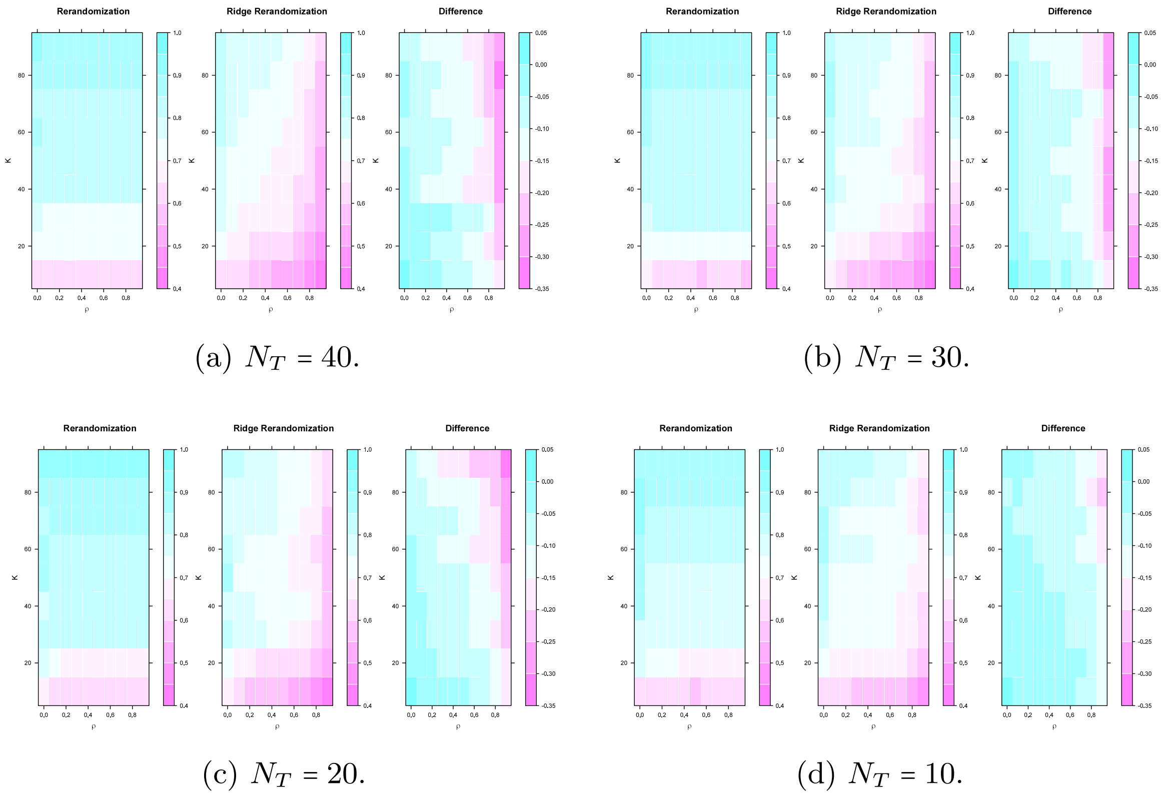
A.8.2. Nonlinearity

Similar to Section 5, we consider $N = 100$ units to be assigned to treatment and control. For each unit, the covariate matrix $\mathbf{x}$ is still generated with (26) and $N_T = N_C = 50$ units will be assigned to treatment and control when implementing randomization, rerandomization, and ridge rerandomization. However, instead of using (27) to generate the potential outcomes, we will use the following model:

$$\begin{aligned} Y_i(0) &\sim N\left(\exp(\mathbf{x})\boldsymbol{\beta}, 1\right) \\ Y_i(1) &= Y_i(0) + \tau \end{aligned} \tag{47}$$

where $\exp(\mathbf{x})$ denotes the matrix of values $e^{\mathbf{x}}$. Again we set $\tau = 1$ and $\boldsymbol{\beta} = \mathbf{1}_K$ and consider $K \in \{10, \ldots, 90\}$ and $\rho \in \{0, 0.1, \ldots, 0.9\}$ when generating the covariates.

Rerandomization and ridge rerandomization only aim to balance the first moments of the covariates, and thus the simulations in Section 5 (where the potential outcomes are linearly related with the covariates) may be considered a "well-specified" scenario, and here we are considering a misspecified scenario where averages across potential outcomes depend on more than just the first moments of covariates. This alternative model for the potential outcomes does not affect rerandomization and ridge rerandomization's ability to balance covariates' first moments, but it does affect their ability to precisely estimate treatment effects. Fig. 10 compares the relative MSE (compared to randomization) of rerandomization and ridge rerandomization, and Fig. 11 does the same for relative average 95% confidence interval width. Although ridge rerandomization does not have as clear of an advantage over rerandomization in this misspecified scenario, it still tends to perform better than rerandomization in high-dimensional and high-collinearity settings. Furthermore,



(a) $N_T = 40$.

(b) $N_T = 30$.

(c) $N_T = 20$.

(d) $N_T = 10$.

Fig. 9. Relative average 95% confidence interval width under rerandomization and ridge rerandomization (relative to randomization), as well as the difference between the two (i.e., the second plot minus the first) for $N_T \in \{10, 20, 30, 40\}$. This is analogous to Fig. 4, but for different values of N_T.

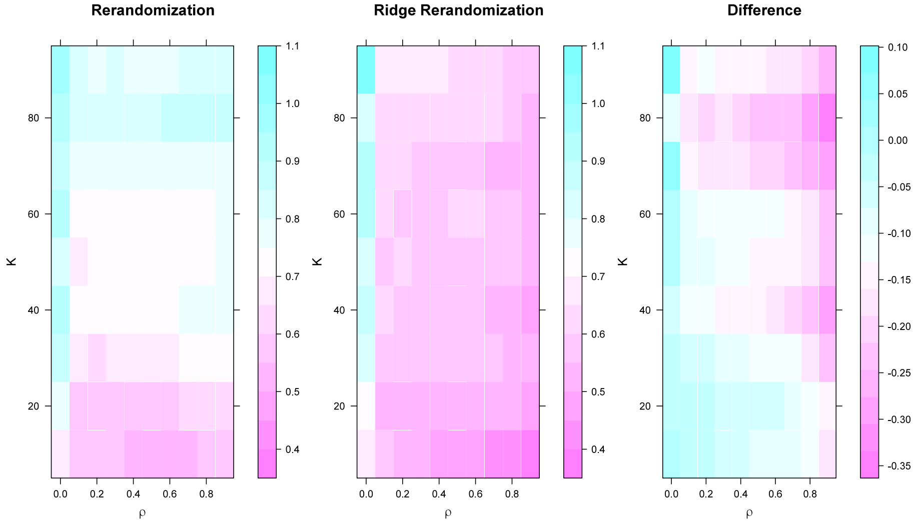


Fig. 10. Relative MSE of $\hat{\tau}$ under rerandomization and ridge rerandomization (relative to randomization) when $\boldsymbol{\beta} = \mathbf{1}_K$ in (47), as well as the difference in relative MSE between the two (i.e., the second plot minus the first).

both rerandomization and ridge rerandomization still provide more precise inference for the average treatment effect compared to randomization, although not as much as when the potential outcomes were generated from a linear model. This is because the covariates still have some linear relationship with the covariates, and thus one can still obtain more

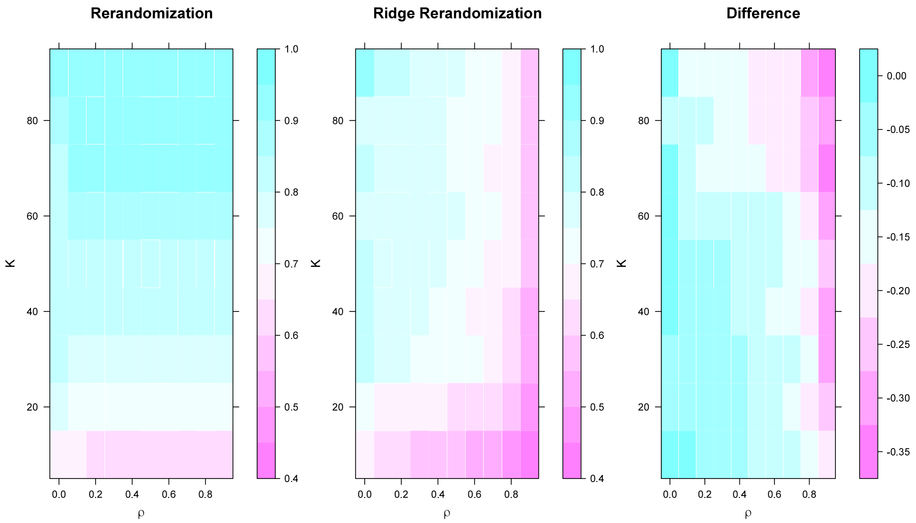


Fig. 11. Relative average 95% confidence interval width under rerandomization and ridge rerandomization (relative to randomization) when $\boldsymbol{\beta} = \mathbf{1}_K$ in (47), as well as the difference between the two (i.e., the second plot minus the first).

precise estimators and intervals for the average treatment effect by balancing the first moments of the covariates (Li et al., 2018). In short, the results presented here are largely the same as those presented in Section 5, where the potential outcomes were linearly related with the covariates.

A.8.3. Treatment effect heterogeneity

Similar to Section 5, we consider $N = 100$ units to be assigned to treatment and control. For each unit, the covariate matrix $\mathbf{x}$ is still generated with (26) and $N_T = N_C = 50$ units will be assigned to treatment and control when implementing randomization, rerandomization, and ridge rerandomization. However, instead of using (27) to generate the potential outcomes, we will use the following model:

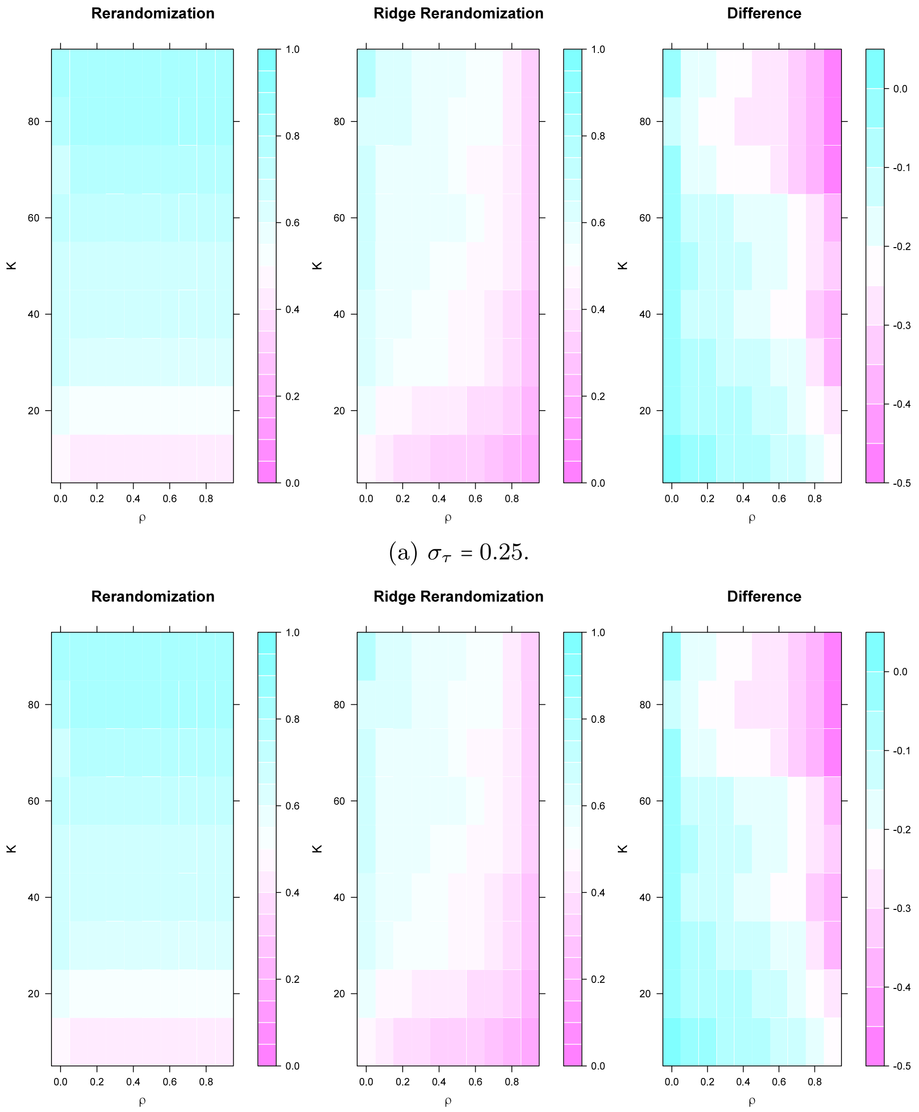
$$\begin{aligned} &Y_i(0) \sim N(\mathbf{x}\boldsymbol{\beta}, 1) \\ &Y_i(1) = Y_i(0) + \tau + \sigma_\tau Y_i(0) \end{aligned} \tag{48}$$

The above setup is similar to the simulation setup used in Ding et al. (2016) for studying treatment effect heterogeneity. In the following simulations, we set $\tau = 1$. When the heterogeneity parameter $\sigma_\tau = 0$, this simulation setup is identical to that used in Section 5, where treatment effects are additive. In this section, we will consider $\sigma_\tau \in \{0.25, 0.5\}$; similar to Ding et al. (2016), $\sigma_\tau = 0.25$ corresponds to moderate treatment effect heterogeneity and $\sigma_\tau = 0.5$ corresponds to strong treatment effect heterogeneity. Furthermore, we again set $\boldsymbol{\beta} = \mathbf{1}_K$ and consider $K \in \{10, \dots, 90\}$ and $\rho \in \{0, 0.1, \dots, 0.9\}$ when generating the covariates.

Thus, the only simulation feature we are changing (compared to Section 5) is the way that the potential outcomes are generated. This will affect the analysis stage but not the design stage, and thus results for the average reduction in variance will be identical to those in Section 5, regardless of the heterogeneity parameter. Thus, in what follows, we will only study the relative MSE and relative average 95% confidence interval width for rerandomization and ridge rerandomization.

We will implement randomization, rerandomization, and ridge rerandomization 1000 times and compute the MSE and average confidence interval width for estimating the average treatment effect. Similar to Section 5, we focus on using the mean-difference estimator $\hat{\tau} = \bar{y}_T - \bar{y}_C$. However, unlike in Section 5, the average treatment effect is no longer simply $\tau = 1$, because each unit now has its own treatment effect $\tau \equiv Y_i(1) - Y_i(0) = \tau + \sigma_\tau Y_i(0)$. Thus, when computing the MSE for randomization, rerandomization, and ridge rerandomization, we compute $\mathbb{E}[(\hat{\tau} - \bar{\tau})^2]$, where $\bar{\tau} = N^{-1} \sum_{i=1}^N \tau_i$.

Fig. 12 compares the relative MSE (compared to randomization) of rerandomization and ridge rerandomization, and Fig. 13 does the same for relative average 95% confidence interval width. Once again, the results in these figures are nearly identical to those presented in Section 5: Ridge rerandomization tends to have a lower relative MSE and lower relative average confidence interval width, especially in high-dimensional and/or high-collinearity settings, regardless of whether treatment effect heterogeneity is moderate ($\sigma_\tau = 0.25$) or large ($\sigma_\tau = 0.5$). We should note that the raw MSE and average confidence interval width (not shown) for randomization, rerandomization, and ridge rerandomization all increased from $\sigma_\tau = 0.25$ to $\sigma_\tau = 0.5$; however, their *relative* performance to each other did not substantially change from moderate



(a) $\sigma_\tau = 0.25$.

(b) $\sigma_\tau = 0.5$.

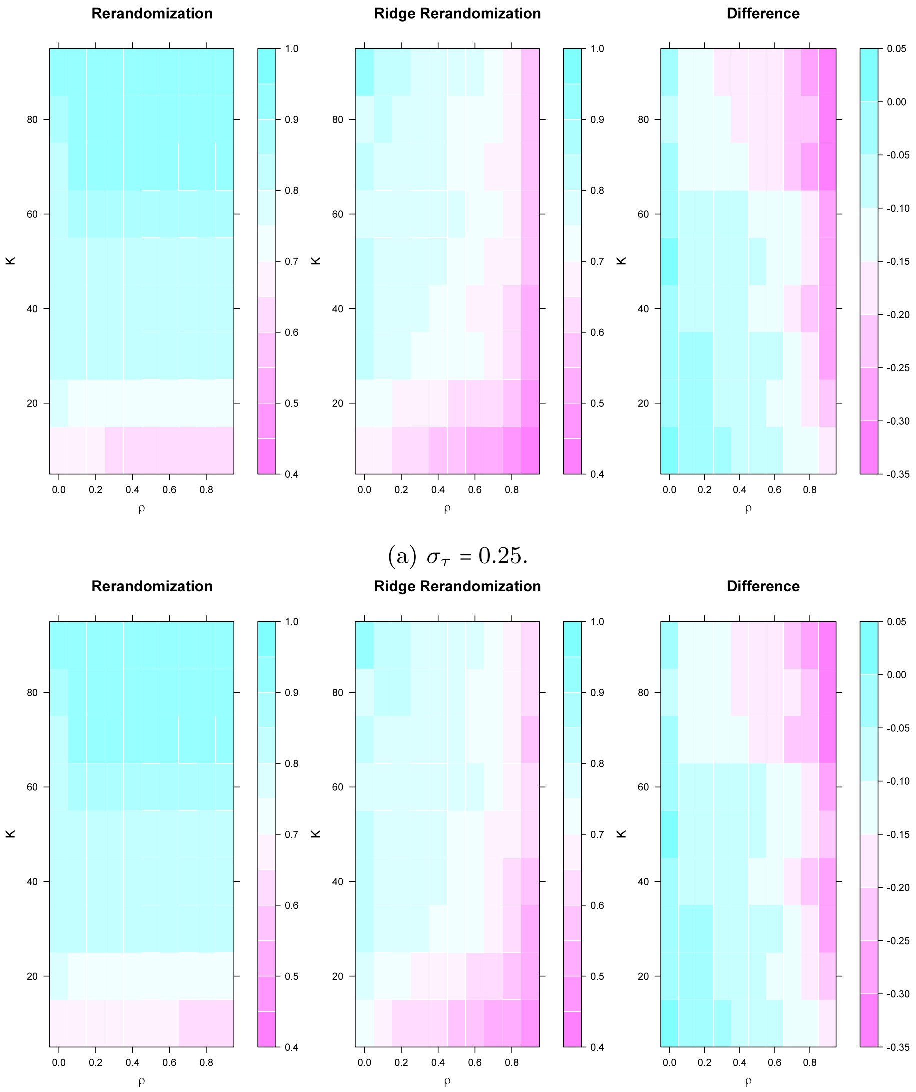
Fig. 12. Relative MSE of $\hat{\tau} = \bar{y}_T - \bar{y}_C$ under rerandomization and ridge rerandomization (relative to randomization), as well as the difference between the two (i.e., the second plot minus the first) for $\sigma_\tau \in \{0.25, 0.5\}$. This is analogous to Fig. 3, but for heterogeneous treatment effects using (48) to generate the potential outcomes.

to strong treatment effect heterogeneity, as shown by Figs. 12 and 13. In short, even though inference becomes more challenging when treatment effect heterogeneity increases, ridge rerandomization still appears to exhibit an advantage over rerandomization in high-dimensional and/or high-collinearity settings.

A.8.4. Rank deficiency

Similar to Section 5, we consider $N = 100$ units where 50 units are assigned to treatment and 50 units are assigned to control. For each unit, the covariate matrix $\mathbf{x}$ is still generated with (26) and the potential outcomes are generated with (27), where $\boldsymbol{\beta} = \mathbf{1}_K$ and $\tau = 1$. Again we consider $\rho \in \{0, 0.1, \ldots, 0.9\}$ when generating the covariates. For this subsection, we will focus on the case where there are $K = 101$ covariates.

When $K = 101$, the covariates' covariance matrix $\boldsymbol{\Sigma}$ is rank-deficient, because $N < K$. In other words, $\boldsymbol{\Sigma}$ is not invertible, the Mahalanobis distance is undefined, and rerandomization cannot be implemented. Morgan and Rubin (2012) noted that when $N \leq K$, the pseudo-inverse for $\boldsymbol{\Sigma}$ can be used when defining the Mahalanobis distance; however, when we attempted this on our simulated data, we found that the resulting Mahalanobis distance was constant across all



(a) $\sigma_\tau = 0.25$.

(b) $\sigma_\tau = 0.5$.

Fig. 13. Relative average 95% confidence interval width under rerandomization and ridge rerandomization (relative to randomization), as well as the difference between the two (i.e., the second plot minus the first) for $\sigma_\tau \in \{0.25, 0.5\}$. This is analogous to Fig. 4, but for heterogeneous treatment effects using (48) to generate the potential outcomes.

randomizations, thereby leaving it uninformative. In our own past exploration of the Mahalanobis distance using the pseudo-inverse (not shown), we have found this to also occasionally occur with real datasets. Interesting future work would be investigating when using the pseudo-inverse for $\mathbf{\Sigma}$ leads to a properly defined Mahalanobis distance.

In any case, the ridge Mahalanobis distance M_λ in (6) is still defined even when $N \leq K$, and we can still assess the benefits of ridge rerandomization over randomization in this case, even if we cannot assess rerandomization. Similar to the previous sections, we implemented randomization and ridge rerandomization 1000 times under this scenario and computed (1) the average reduction in variance across covariates, (2) relative MSE for the average treatment effect, and (3) relative average 95% confidence interval width for the average treatment effect. Fig. 14 shows the results for $\rho \in \{0, 0.1, \ldots, 0.9\}$. Once again, we see that ridge rerandomization reduces the average variance of covariate mean differences compared to randomization, and it also leads to a lower MSE and narrower confidence intervals when estimating the average treatment effect. This is especially the case when collinearity is high. This suggests that ridge rerandomization may be a viable experimental design strategy when $N \leq K$.

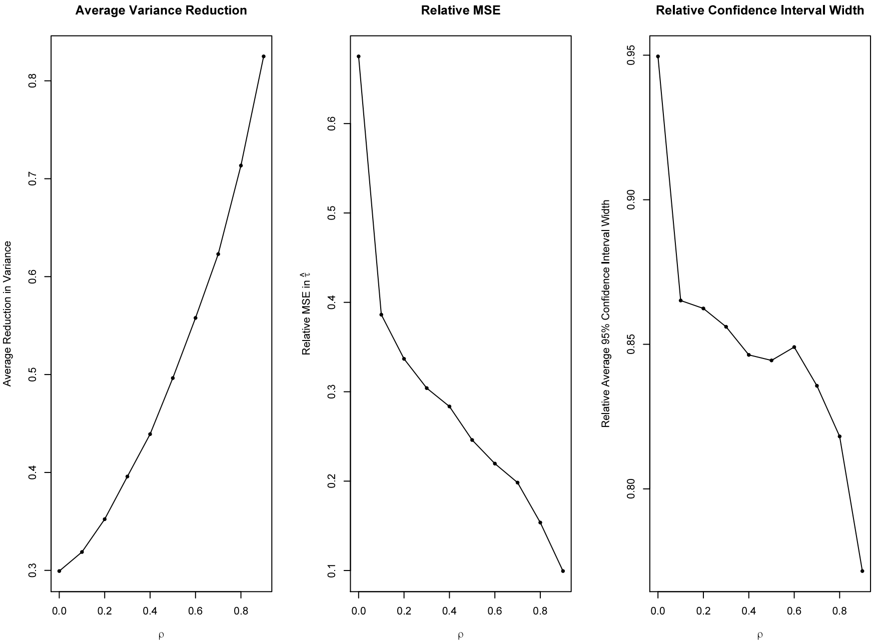


Fig. 14. Average reduction in variance, relative MSE, and relative average confidence interval width for ridge rerandomization (in comparison with randomization) when $K = 101$ for collinearity $\rho \in \{0, 0.1, \ldots, 0.9\}$.

References

Aronow, P.M., Samii, C., 2012. Ri: R package for performing randomization-based inference for experiments.

Bausch, J., 2013. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. J. Phys. A 46 (50), 505202.

Bodenham, D.A., Adams, N.M., 2016. A comparison of efficient approximations for a weighted sum of chi-squared random variables. Stat. Comput. 26 (4), 917–928.

Branson, Z., Dasgupta, T., 2020. Sampling-based randomised designs for causal inference under the potential outcomes framework. Internat. Statist. Rev. 88 (1), 101–121.

Branson, Z., Dasgupta, T., Rubin, D.B., et al., 2016. Improving covariate balance in 2k 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.

Branson, Z., Miratrix, L.W., 2019. Randomization tests that condition on non-categorical covariate balance. J. Causal Inference 7 (1).

Bruhn, M., McKenzie, D., 2009. In pursuit of balance: Randomization in practice in development field experiments. Am. Econ. J.: Appl. Econ. 1 (4), 200–232.

Cox, D., 2009. Randomization in the design of experiments. Internat. Statist. Rev. 77 (3), 415–429.

Davies, R.B., 1980. Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. J. R. Stat. Soc. Ser. C (Appl. Stat.) 29 (3), 323–333.

Ding, P., Feller, A., Miratrix, L., 2016. Randomization inference for treatment effect variation. J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (3), 655–671.

Edgington, E., Onghena, P., 2007. Randomization Tests. Chapman and Hall/CRC.

Erdös, P., Rényi, A., 1959. On the central limit theorem for samples from a finite population. Publ. Math. Inst. Hungar. Acad. Sci. 4, 49–61.

Fisher, R.A., 1992. The arrangement of field experiments. In: Breakthroughs in Statistics. Springer, pp. 82–91.

Freedman, D.A., 2008. On regression adjustments to experimental data. Adv. Appl. Math. 40 (2), 180–193.

Good, P., 2013. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media.

Gu, X.S., Rosenbaum, P.R., 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. J. Comput. Graph. Statist. 2 (4), 405–420.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., Sarkar, P., 2016. A conditional randomization test to account for covariate imbalance in randomized experiments. J. Causal Inference 4 (1), 61–80.

Hodges Jr, J.L., Lehmann, E.L., 1963. Estimates of location based on rank tests. Ann. Math. Stat. 598–611.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

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

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Imhof, J.-P., 1961. Computing the distribution of quadratic forms in normal variables. Biometrika 48 (3/4), 419–426.

Kato, N., Suzuki, M., Omachi, S., Aso, H., Nemoto, Y., 1999. A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance. IEEE Trans. Pattern Anal. Mach. Intell. 21 (3), 258–262.

Kempthorne, O., Doerfler, T., 1969. The behaviour of some significance tests under experimental randomization. Biometrika 56 (2), 231–248.

Krause, M.S., Howard, K.I., 2003. What random assignment does and does not do. J. Clin. Psychol. 59 (7), 751–766.

Li, X., Ding, P., 2017. General forms of finite population central limit theorems with applications to causal inference. J. Amer. Statist. Assoc. (just-accepted).

Li, X., Ding, P., 2020. Rerandomization and regression adjustment. J. R. Stat. Soc. Ser. B Stat. Methodol..

Li, X., Ding, P., Rubin, D.B., 2018. Asymptotic theory of rerandomization in treatment–control experiments. Proc. Natl. Acad. Sci. 115 (37), 9157–9162.

Lindley, D.V., 1982. The role of randomization in inference. In: PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol. 1982. Philosophy of Science Association, pp. 431–446.

Maclure, M., Nguyen, A., Carney, G., Dormuth, C., Roelants, H., Ho, K., Schneeweiss, S., 2006. Measuring prescribing improvements in pragmatic trials of educational tools for general practitioners. Basic Clin. Pharmacol. Toxicol. 98 (3), 243–252.

Mahalanobis, P.C., 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India 1936 49–55.

Miratrix, L.W., Sekhon, J.S., Yu, B., 2013. Adjusting treatment effect estimates by post-stratification in randomized experiments. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (2), 369–396.

Morgan, K.L., Rubin, D.B., 2012. Rerandomization to improve covariate balance in experiments. Ann. Statist. 40 (2), 1263–1282.

Morgan, K.L., Rubin, D.B., 2015. Rerandomization to balance tiers of covariates. J. Amer. Statist. Assoc. 110 (512), 1412–1421.

Moulton, L.H., 2004. Covariate-based constrained randomization of group-randomized trials. Clin. Trials 1 (3), 297–305.

Neyman, J., Dabrowska, D.M., Speed, T., 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statist. Sci. 465–472.

Olsen, S.P., 1997. Multivariate matching with non-normal covariates in observational studies.

Palombi, F., Toti, S., 2013. A note on the variance of the square components of a normal multivariate within a Euclidean ball. J. Multivariate Anal. 122, 355–376.

Papineau, D., 1994. The virtues of randomization. Br. J. Phil. Sci. 45 (2), 437–450.

Pashley, N.E., Miratrix, L.W., 2017. Insights on variance estimation for blocked and matched pairs designs. arXiv preprint arXiv:1710.10342.

Phipson, B., Smyth, G.K., 2010. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Stat. Appl. Genet. Mol. Biol. 9 (1).

Rosenbaum, P.R., 2002. Overt bias in observational studies. In: Observational Studies. Springer, pp. 71–104.

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Amer. Statist. 39 (1), 33–38.

Rosenberger, W.F., Sverdlov, O., 2008. Handling covariates in the design of clinical trials. Statist. Sci. 404–419.

Rubin, D.B., 1974. Multivariate matching methods that are equal percent bias reducing, I: Some examples. ETS Res. Rep. Ser. 1974 (2).

Rubin, D.B., 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statist. Sci. 5 (4), 472–480.

Rubin, D.B., 2005. Causal inference using potential outcomes: Design, modeling, decisions. J. Amer. Statist. Assoc. 100 (469), 322–331.

Rubin, D.B., Thomas, N., 2000. Combining propensity score matching with additional adjustments for prognostic covariates. J. Amer. Statist. Assoc. 95 (450), 573–585.

Seidenfeld, T., 1981. Levi on the dogma of randomization in experiments. In: Henry E. Kyburg, Jr. & Isaac Levi. Springer, pp. 263–291.

Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. Statist. Sci. 25 (1), 1.

Worrall, J., 2010. Evidence: philosophy of science meets medicine. J. Eval. Clin. Pract. 16 (2), 356–362.

Zhou, Q., Ernst, P., Morgan, K.L., Rubin, D., Zhang, A., 2018. Sequential rerandomization. Biometrika 105 (3), 745–752.