Ridge Rerandomization: An Experimental Design Strategy in the Presence of Collinearity

Zach Branson and Stephane Shao*

Department of Statistics, Harvard University

August 18, 2018

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 correlated or vary in importance. As an alternative, we introduce an experimental design strategy called ridge rerandomization, which utilizes a modified Mahalanobis distance that addresses collinearities among covariates and automatically places a hierarchy of importance on the covariates according to their eigenstructure. 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 & Howard, 2003). However, many have noted that randomized experiments can yield "bad allocations,"

*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.

where some covariates are not well-balanced across treatment groups (Seidenfeld, 1981; Lindley, 1982; Papineau, 1994; Rosenberger & 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 covariates. A more recent experimental design tool that prevents covariate imbalance and allows for non-categorical covariates is the rerandomization scheme of Morgan & 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 & McKenzie (2009), and Worrall (2010)) recommend rerandomization. Morgan & 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 & Rubin (2012), several extensions have been made. Morgan & 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. (2017) developed a rerandomization scheme for sequential designs. Finally, Li et al. (2016) established asymptotic results for the rerandomization schemes considered in Morgan & Rubin (2012) and Morgan & Rubin (2015).

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 & Rubin, 1985; Gu & Rosenbaum, 1993; Rubin & 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 & Rubin, 2012).

However, the Mahalanobis distance is known to perform poorly in matching for observational studies when covariates are not ellipsoidally symmetric, there are strong collinearities among the covariates, or there are many covariates (Gu & 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 & Rubin (2015), who proposed an extension of Morgan & 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.

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 & 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. 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. In particular, this rerandomization scheme automatically specifies a hierarchy of importance based on the eigenstructure of the covariates, which can be useful when researchers are unsure about how much importance they should place on each covariate when designing a randomized experiment. 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 & 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 $\lambda_{1:K} = (\lambda_1, ..., \lambda_K) \in \mathbb{R}^K$ for tuples of objects, and we let $f(\lambda_{1:K}) = (f(\lambda_1), ..., f(\lambda_K))$ 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 i-th row, $A_{\bullet j}$ its j-th 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 & 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. Finally, we define the covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x})$ of the covariate mean dif-

ferences with respect to the distribution of $\mathbf{W}$ given $\mathbf{x}$, and we assume $\mathbf{\Sigma} > 0$. The spectral decomposition ensures that $\mathbf{\Sigma}$ is diagonalizable with eigenvalues $\lambda_1 \geq ... \geq \lambda_K > 0$. Let $\mathbf{\Gamma}$ be the orthogonal matrix of corresponding eigenvectors, so that we may write $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{Diag}(\lambda_{1:K})\mathbf{\Gamma}^\top$, where $\mathbf{Diag}(\lambda_{1:K})$ denotes the $K \times K$ diagonal matrix whose (k,k)-coefficient is λ_k.

For completely randomized experiments, we have $\mathbf{\Sigma} = NN_T^{-1}N_C^{-1}S_{\mathbf{x}}^2$, 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 & Rubin, 2012). Thus, $\mathbf{\Sigma}$ and its eigenstructure are available in closed-form, and the latter coincides with the eigenstructure of $S_{\mathbf{x}}^2$ up to a scaling factor. We let χ_K^2 denote a chi-squared distribution with K degrees of freedom, $\mathbb{P}(\chi_K^2 \leq a)$ its cumulative distribution function (CDF) evaluated at $a \in \mathbb{R}$, and $q_{\chi_K^2}(p)$ 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 $\mathbf{y}^{obs} = (y_1^{obs} \ \dots \ y_N^{obs})^\top$ 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 & 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

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

where the dependence of M on the assignment vector $\mathbf{W}$ is implicit through $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$. Morgan & Rubin (2012) suggest randomizing units to treatment and control by performing independent draws from the distribution of $\mathbf{W}\,|\,\mathbf{x}$ 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 draws until the first acceptable randomization is equal to $1/p_a$, where $p_a = \mathbb{P}(M \leq a\,|\,\mathbf{x})$ 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 $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \,|\, \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, then $M \,|\, \mathbf{x} \sim \chi^2_K$, so that a can be chosen equal to the p_a-quantile of a chi-squared distribution with K degrees of freedom. The assumption $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \,|\, \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ can be justified by invoking the finite population Central Limit Theorem (Erdös & Rényi, 1959; Li & Ding, 2017). When the distribution of $M \,|\, \mathbf{x}$ is unknown, one can approximate it via Monte Carlo by simulating independent draws of $M \,|\, \mathbf{x}$ and setting a to the p_a-quantile of M's empirical distribution.

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

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

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

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 & Rubin (2012) call this last property *equally percent variance reducing* (EPVR). Thus, using the Mahalanobis distance for rerandomization can be quite appealing, but Morgan & Rubin (2012) rightly point out that non-EPVR rerandomization schemes may be preferable in settings with covariates of unequal importances. This is in part addressed by Morgan & 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.

4 Ridge Rerandomization

As an alternative, we define a modified Mahalanobis distance as

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

for some prespecified $\lambda \geq 0$. Guidelines for choosing λ will be provided in Section 4.2. The eigenvalues of $\boldsymbol{\Sigma}$ in (6) are inflated in a way that is reminiscent of ridge regression (Hoerl & 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 & Rubin (2012). In other words, one randomizes the assignment vector $\mathbf{W}$ until $M_\lambda \le a_\lambda$ for some threshold $a_\lambda \ge 0$.

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

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

Thus, fixing p_a implicitly determines the pair (λ, a_λ), so that for every fixed $\lambda \ge 0$ and $p_a \in (0,1)$ corresponds 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).

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 $\lambda \ge 0$ and $a_\lambda \ge 0$ be some prespecified constants. If $N_T = N_C$, then*

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

and

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

Theorem 4.1 is a particular case of Theorem 2.1 and Corollary 2.2 from Morgan & 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 & Rubin (2012), we even have the stronger result that $\mathbb{E}[\bar{V}_T - \bar{V}_C \,|\, \mathbf{x}, M_\lambda \le a_\lambda] = 0$ for any covariate V, regardless of whether V is observed or not.

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 $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \,|\, \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, 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 $\boldsymbol{\Sigma}$.

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

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

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

The proof of Lemma 4.1 is provided in the Appendix; see Section 7.1. 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 & Adams, 2016, and references therein).

Using Lemma 4.1, 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 $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\,|\,\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and $N_T = N_C$, then*

$$\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) = \boldsymbol{\Gamma}\mathbf{Diag}((\lambda_k\, d_{k,\lambda})_{1\leq k\leq K})\boldsymbol{\Gamma}^\top \tag{9}$$

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

$$d_{k,\lambda} = \mathbb{E}\left[Z_k^2 \,\middle|\, \sum_{j=1}^{K} \frac{\lambda_j}{\lambda_j + \lambda} Z_j^2 \leq a_\lambda \right] \tag{10}$$

with $Z_1, \ldots, Z_K \overset{i.i.d.}{\sim} N(0,1)$.

The proof of Theorem 4.2 is in the Appendix in Section 7.2. The quantities $d_{k,\lambda}$ are intractable functions of λ and a_λ and thus need to be approximated numerically, as explained in Section 4.2.1. Conditioning on $M_\lambda \leq a_\lambda$ in (10) 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 $k = 1, \ldots, K$, we may conjecture that $0 < d_{1,\lambda} \leq \cdots \leq d_{K,\lambda} < 1$. Possible directions for a proof may make use of Proposition 2.1 from Palombi & Toti (2013) and Equation (A.1) from Palombi et al. (2017).

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

$$\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x}) = \boldsymbol{\Gamma}\mathbf{Diag}((\lambda_k)_{1\leq k\leq K})\boldsymbol{\Gamma}^\top, \tag{11}$$

$$\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x}, M \leq a) = \boldsymbol{\Gamma}\mathbf{Diag}((\lambda_k\, v_a)_{1\leq k\leq K})\boldsymbol{\Gamma}^\top, \tag{12}$$

$$\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) = \boldsymbol{\Gamma}\mathbf{Diag}((\lambda_k\, d_{k,\lambda})_{1\leq k\leq K})\boldsymbol{\Gamma}^\top. \tag{13}$$

where (12) follows from Theorem 3.1 in Morgan & Rubin (2012) with $v_a \in (0,1)$, and (13) follows from Theorem 4.2 with $d_{k,\lambda} \in (0,1)$ defined in (10). If we define new covariates $\mathbf{x}^*$ as the principal components of the original ones, i.e., $\mathbf{x}^* = \mathbf{x}\boldsymbol{\Gamma}$, then (12) and (13) respectively yield

$$\mathrm{Var}((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \,|\, \mathbf{x}, M \leq a) = v_a\, \mathrm{Var}((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \,|\, \mathbf{x}) \tag{14}$$

and

$$\mathrm{Var}((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) = d_{k,\lambda}\, \mathrm{Var}((\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k \,|\, \mathbf{x}) \tag{15}$$

for all $k = 1, \ldots, K$, where $(\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_C^*)_k$ is the k-th *principal component mean difference* between the treatment and control groups, i.e., the k-th coefficient of

$\mathbf{\Gamma}^\top(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$. From (14) 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 k-th principal component mean difference is reduced by $100(1-d_{k,\lambda})\%$, and because typically $0 < d_{1,\lambda} \leq \cdots \leq d_{K,\lambda} < 1$, ridge rerandomization places more importance on the first principal components.

Translating (15) 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 k-th covariate mean difference $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k$ by $100\,(1 - v_{k,\lambda})\,\%$, where*

$$v_{k,\lambda} = \frac{\left(\mathbf{\Gamma}\mathbf{Diag}\left((\lambda_j\, d_{j,\lambda})_{1\leq j\leq K}\right)\mathbf{\Gamma}^\top\right)_{kk}}{\mathbf{\Sigma}_{kk}} \tag{16}$$

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

$$\mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \,|\, \mathbf{x}, M_\lambda \leq a_\lambda\right) \;<\; \mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \,|\, \mathbf{x}\right). \tag{17}$$

The proof of Corollary 4.1 is provided in the Appendix; see Section 7.3. 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 & 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 & Rubin (2012)—that the treatment effect is additive. Without loss of generality, for all $i = 1, \ldots, N$, we can write the outcome of unit i as

$$Y_i(W_i) = \beta_0 + \mathbf{x}_{i\bullet}\boldsymbol{\beta} + \tau W_i + \epsilon_i \tag{18}$$

where $\beta_0 + \mathbf{x}\boldsymbol{\beta}$ is the projection of the potential outcomes $\mathbf{Y}(0) = (Y_1(0) \ldots Y_N(0))^\top$ 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 $\boldsymbol{\epsilon} = (\epsilon_1 \ldots \epsilon_N)^\top$.

Theorem 4.3 below 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*

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

so that we have

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

where the equality holds if and only if $\boldsymbol{\beta} = 0$ *in* (18).

The proof of Theorem 4.3 is in the Appendix; see Section 7.4. The conditional independence assumption was also leveraged in the proof of Theorem 3.2 in Morgan & Rubin (2012).

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 & 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*

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

It is not necessarily the case that $d_{k,\lambda} \le v_a$ for all $k = 1, \ldots, K$, 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 $a_\lambda \ge 0$ and the regularization parameter $\lambda \ge 0$. 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 $a_\lambda \ge 0$ for each $\lambda \ge 0$ such that $\mathbb{P}(M_\lambda \le a_\lambda \,|\, \mathbf{x}) = p_a$. As in Morgan & Rubin (2012), acceptable treatment allocations under ridge rerandomization are generated by randomizing units to treatment and control until $M_\lambda \le a_\lambda$. Thus, a smaller p_{a_λ} 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 $\lambda \ge 0$. Section 4.2.1 details how a_λ is automatically calibrated once we fix p_a and λ. The choice of λ is investigated in Section 4.2.3, after discussing how to assess the performance of ridge rerandomization in Section 4.2.2.

4.2.1 Calibration of a_λ

Given p_a and λ, we can choose to set a_λ equal to the p_a-quantile of the quadratic form Q_λ defined by

$$Q_\lambda = \sum_{k=1}^{K} \frac{\lambda_k}{\lambda_k + \lambda} Z_k^2 \tag{19}$$

where $Z_1, ..., Z_K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. 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 $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \,|\, \mathbf{x}$ to be approximately Normal, as motivated by Lemma 4.1. Let $F_{Q_\lambda}(q) = \mathbb{P}(Q_\lambda \leq q)$ denote the CDF of Q_λ. Since Q_λ is a weighted sum of independent χ_1^2 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

$$F_{Q_\lambda, U}(q) = \frac{1}{2} - \frac{1}{\pi} \int_0^U \frac{\sin\left(\frac{1}{2}\left[-t\,q + \sum_{k=1}^{K} \arctan\left(\frac{\lambda_k}{\lambda_k + \lambda}\, t\right)\right]\right)}{t \prod_{k=1}^{K} \left[1 + \left(\frac{\lambda_k}{\lambda_k + \lambda}\right)^2 t^2\right]^{1/4}} \, dt \tag{20}$$

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 $\tau > 0$ in absolute value by choosing $U_\tau = [\tau\, \pi\, (K/2) \prod_{k=1}^{K} \sqrt{\lambda_k (\lambda_k + \lambda)^{-1}}]^{-2/K}$. Computationally cheaper but less accurate alternatives to approximate F_{Q_λ} are discussed in Bodenham & Adams (2016).

Finally, we approximate the p_a-quantile of Q_λ by

$$\hat{a}_\lambda = \inf\{q \in \mathbb{R} : F_{Q_\lambda, U_\tau}(q) \geq p_a\} \tag{21}$$

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 will use $\tau = 10^{-4}$ by default.

4.2.2 Approximation of $d_{k,\lambda}$ and $v_{k,\lambda}$

We will use Corollary 4.1 and Theorem 4.2 as a proxy for how ridge rerandomization improves the variance of each covariate mean difference as compared to rerandomization. We would like to set (λ, a_λ) so that the $d_{k,\lambda}$'s defined in (10) are small, in a sense to be made precise in the next section. To achieve this, we would need to compute $d_{k,\lambda}$ for all $k = 1, ...K$, which involves 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 $i = 1, .., n$ and $j = 1, .., K$, the expectations appearing in (10)

can be consistently estimated via Monte Carlo, for all $k = 1, ..., K$, by

$$\hat{d}_{k,\lambda,n} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{\{M_\lambda^{(i)} \le \hat{a}_\lambda\}}} \sum_{i=1}^{n} \widetilde{Z}_{ik}^2 \mathbb{1}_{\{M_\lambda^{(i)} \le \hat{a}_\lambda\}} \tag{22}$$

with $M_\lambda^{(i)} = \sum_{j=1}^{K} \lambda_k (\lambda_k + \lambda)^{-1} \widetilde{Z}_{ij}^2$ and $\hat{a}_\lambda$ defined in (21), where $\mathbb{1}_A$ denotes the indicator function of a set A. Using (22), we can then estimate $v_{k,\lambda}$ from Corollary 4.1 consistently as $n \to +\infty$, for all $k = 1, ..., K$, by

$$\hat{v}_{k,\lambda,n} = \frac{\left(\mathbf{\Gamma Diag}\left((\lambda_j \hat{d}_{j,\lambda,n})_{1 \le j \le K}\right) \mathbf{\Gamma}^\top\right)_{kk}}{\mathbf{\Sigma}_{kk}} \tag{23}$$

For simplicity, we will 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.

4.2.3 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, one rerandomization scheme is preferable over another if it is able to achieve a higher average reduction in variance across covariates. Thus, ideally, we would only choose a particular $\lambda > 0$ if $K^{-1} \sum_{k=1}^{K} v_{k,\lambda} < v_a$. In practice, we will use the criterion

$$\frac{1}{K} \sum_{k=1}^{K} \hat{v}_{k,\lambda,n} < v_a \tag{24}$$

where v_a and $\hat{v}_{k,\lambda,n}$ are respectively defined in (5) and (23), with a being set to $q_{\chi^2_K}(p_a)$, i.e., the choice of a as recommended by Morgan & Rubin (2012). Proving the existence of some $\lambda > 0$ such that (24) holds is challenging, so we propose the following iterative procedure for choosing such a $\lambda > 0$ if it exists. The procedure relies on (5), (21), and (23), where the auxiliary Normal variables $\widetilde{Z}_{ij}$ only need to be simulated once and can then 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 $\Lambda = \varnothing$.
3. While $|(\lambda+\delta)\hat{a}_{\lambda+\delta} - \lambda \hat{a}_\lambda| > \varepsilon$:
   - Set $\lambda = \lambda + \delta$.
   - If $\dfrac{1}{K}\sum_{k=1}^{K} \hat{v}_{k,\lambda,n} < \dfrac{\mathbb{P}(\chi^2_{K+2} \leq q_{\chi^2_K}(p_a))}{p_a}$, then set $\Lambda = \Lambda \cup \{\lambda\}$.
4. If $\Lambda = \varnothing$, then return $\lambda = 0$.
   Else, define $c_k = \lambda_k^2 \,(\sum_{j=1}^{K} \lambda_j^2)^{-1}$ for all $k = 1, ..., K$, and return:

$$\lambda_\star = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \left(\sum_{k=1}^{K} c_k \, \hat{d}^2_{k,\lambda,n} - \left(\sum_{k=1}^{K} c_k \, \hat{d}_{k,\lambda,n} \right)^2 \right). \tag{25}$$

The justification of our proposed procedure stems from the following facts. By definition, we have $\mathbb{P}(M_\lambda \leq a_\lambda \,|\, \mathbf{x}) = p_a$ 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

$$\lambda \, a_\lambda \quad \xrightarrow[\lambda \to +\infty]{} \quad q^*(p_a) \tag{26}$$

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 $k = 1, ..., K$, we have

$$v_{k,\lambda} \quad \xrightarrow[\lambda \to +\infty]{} \quad \frac{\left(\mathbf{\Gamma} \mathbf{Diag}\left((\lambda_j \, d_j^*)_{1 \leq j \leq K} \right) \mathbf{\Gamma}^\top \right)_{kk}}{\mathbf{\Sigma}_{kk}} \tag{27}$$

where $d_k^* = \mathbb{E}\left[Z_k^2 \middle| \sum_{k=1}^{K} \lambda_k Z_k^2 \leq q^*(p_a) \right]$ for all $k = 1, ..., K$. Since the limits in (27) 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 further 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 (26). The step size δ can be chosen as a fraction of the smallest strictly positive gap between consecutive eigenvalues, i.e., $\min\{\lambda_k - \lambda_{k-1} : k = 1, ..., K \text{ such that } \lambda_k > \lambda_{k-1}\}$ with the convention $\lambda_0 = 0$. Finally, among all the acceptable λ's satisfying (24), Step 4 returns the $\lambda_\star$ that aims at altering the covariance structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$ the least, in the sense of minimizing the distance between $\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{x}, M_\lambda \leq \hat{a}_\lambda)$ and the linear span

of $\mathbf{\Sigma}$, i.e.,

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

where $\|\mathbf{\Sigma}\| = \sqrt{\operatorname{tr}(\mathbf{\Sigma}^\top \mathbf{\Sigma})} = \sum_{k=1}^K \lambda_k^2$ stands for the Frobenius norm. The inner minimization can be written as

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

which is attained at $c_\star = \sum_{k=1}^K c_k \hat{d}_{k,\lambda,n}$ with $c_k = \lambda_k^2 \left(\sum_{j=1}^K \lambda_j^2 \right)^{-1}$ for all $k = 1, ..., K$, thus yielding (25). The outer minimization is then straightforward since the set Λ of candidates is finite by construction.

When the set Λ is empty, we simply return $\lambda = 0$, although the following heuristic argument illustrates why we would expect the existence of at least one λ such that (24) holds. The rerandomization scheme of Morgan & Rubin (2012) spreads the benefits of variance reduction across all K covariates equally; however, note that the term $v_a = \mathbb{P}(\chi^2_{K+2} \le q_{\chi^2_K}(p_a))/p_a$ is monotonically increasing in the number of covariates K for a fixed acceptance probability p_a. 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 & Rubin (2015), which extended the rerandomization scheme of Morgan & 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 $\lambda_{K_e+1}, ..., \lambda_K$ 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

$$v_{a_e} = \frac{\mathbb{P}(\chi^2_{K_e+2} \le q_{\chi^2_{K_e}}(p_a))}{p_a} < \frac{\mathbb{P}(\chi^2_{K+2} \le q_{\chi^2_K}(p_a))}{p_a} = v_a \tag{28}$$

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:

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.

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 & McKenzie (2009), and Cox (2009). However, Morgan & 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 randomizations. See Morgan & Rubin (2012) for a full discussion of the benefits of using affinely invariant rerandomization criteria. As discussed in Section 4.2.3, our proposed procedure aims for larger variance reductions of important 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 $j = 1, ..., N$. Figure 1 shows the distribution of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\,|\,\mathbf{x}$ 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 $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k\,|\,\mathbf{x}$ for $k \in \{1, 2\}$, compared to randomization; however, rerandomization using the Euclidean distance destroys the correlation structure of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\,|\,\mathbf{x}$, while rerandomization and ridge rerandomization largely maintain it. This provides further motivation for Step 4 of the procedure presented in Section 4.2.3.

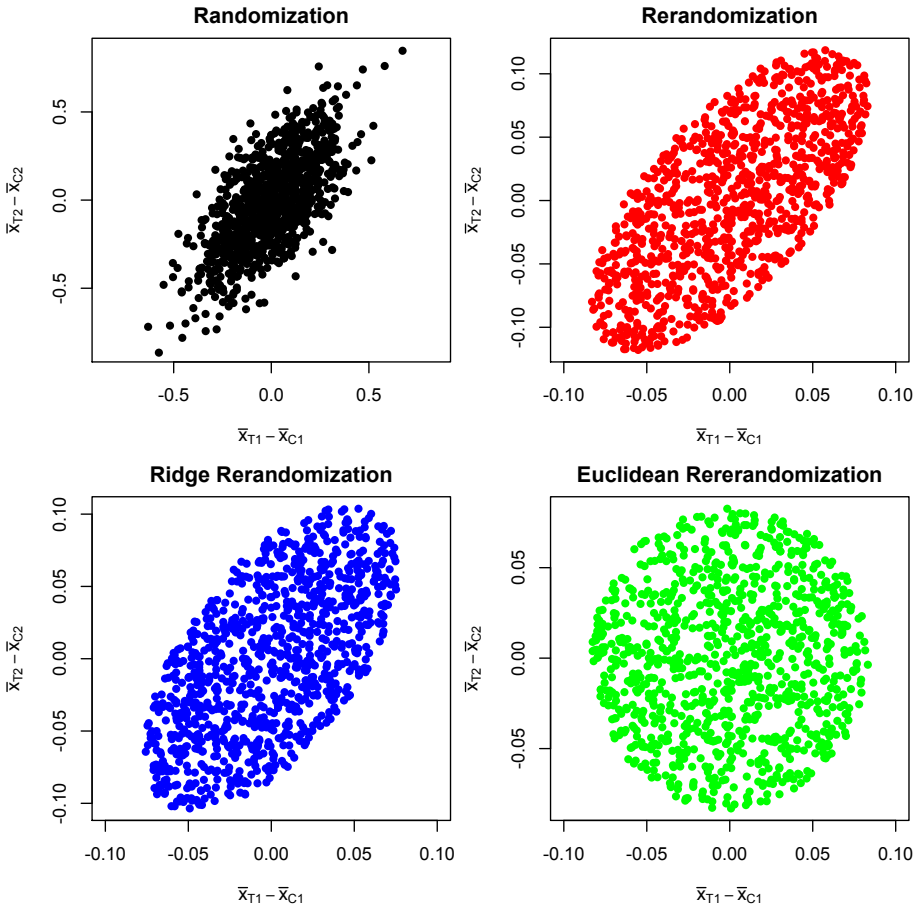

Figure 1: Distribution of $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\,|\,\mathbf{x}$ under randomization, rerandomization (with $p_a = 0.1$), ridge rerandomization (with $p_a = 0.1$ and $\lambda = 0.005$), and rerandomization using the Euclidean distance. Note the difference in scale for the randomization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

Furthermore, as discussed in Sections 4.1 and 4.2.3, 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

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

with

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

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.

5 Simulations

We now provide simulation evidence that supports the hueristic argument presented in Section 4.2 and suggests when ridge rerandomization is an effective experimental design strategy. 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

$$\mathbf{x} \sim \mathcal{N}\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1\end{pmatrix}\right) \tag{29}$$

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

$$\begin{aligned} Y_i(0) &\sim N\left(\mathbf{x}\boldsymbol{\beta}, 1\right) \\ Y_i(1) &= Y_i(0) + \tau \end{aligned} \tag{30}$$

For this simulation study, we set the treatment effect to be $\tau = 1$. Across simulations, we consider number of covariates $K \in \{10, \ldots, 90\}$ and correlation parameter $\rho \in \{0, 0.1, \ldots, 0.9\}$. We discuss choices for $\boldsymbol{\beta}$ in Section 5.3. We also considered data-generating processes where covariances varied among covariates and where there are an uneven number of units assigned to treatment and control (i.e., unbalanced designs). However, the results for these other scenarios were 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 (29) and the potential outcomes are generated from (30).

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 4.2.3 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 (30) is irrelevant for this section. Then, in Section 5.3, we compare the accuracy of the resulting treatment effect estimator $\hat{\tau} = \bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C$ 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. Figure 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 (29). For rerandomization, the average variance reduction decreases as K increases (an observation previously made in Morgan & 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 right-hand plot in Figure 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 & Kennard, 1970).

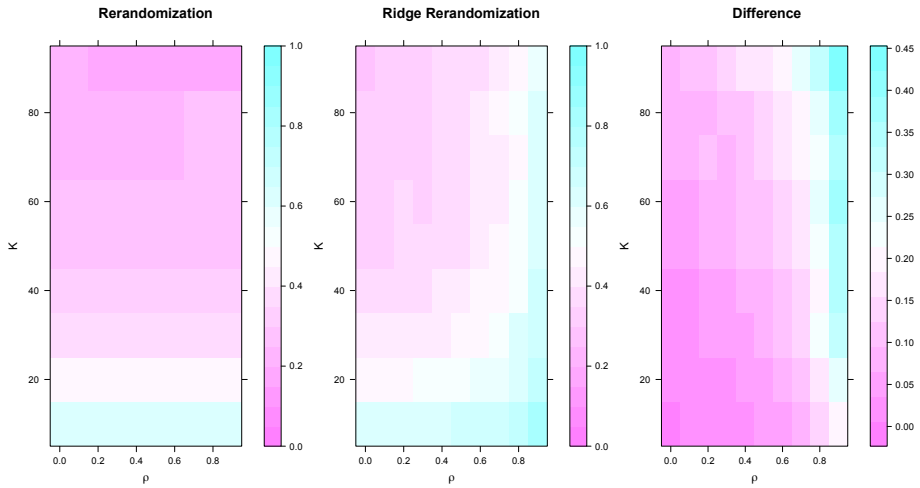

Figure 2: Variance reduction averaged across covariates for rerandomization and ridge rerandomization, as well as their difference (ridge rerandomization minus rerandomization, i.e., the second plot minus the first).

5.3 Comparing Accuracy of Treatment Effect Estimation 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 (30). 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 $K \in \{10, \ldots, 90\}$ and $\rho \in \{0, 0.1, \ldots, 0.9\}$, 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. Figure 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.

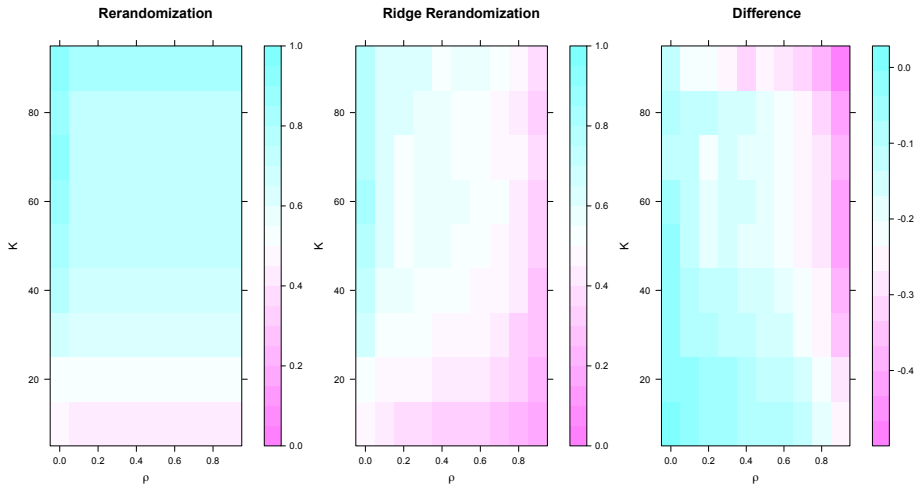

Figure 3: Relative MSE of $\hat{\tau} = \bar{y}_T - \bar{y}_C$ under rerandomization and ridge rerandomization (relative to randomization) when $\boldsymbol{\beta} = \mathbf{1}_K$ in (30), as well as the difference in relative MSE between the two (i.e., the second plot minus the first).

Three observations can be made about Figure 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 Figure 2.

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 $\Delta = \boldsymbol{\beta}^\top \boldsymbol{\Gamma} \mathbf{Diag}\left(\left(\lambda_k\left(v_a - d_{k,\lambda}\right)\right)_{1\le k\le K}\right)\boldsymbol{\Gamma}^\top\boldsymbol{\beta}$, which can be artificially minimized with respect to $\boldsymbol{\beta}$, subject to some constraint on $\boldsymbol{\beta}$ for the minimum to exist, e.g. $\|\boldsymbol{\beta}\| \le 1$. If $d_{k,\lambda} < v_a$ for all $k = 1, ..., K$, then ridge rerandomization dominates rerandomization since $\Delta > 0$ for all $\boldsymbol{\beta} \ne 0$, 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 $k^* = \text{argmin}_{1\le k\le K}(v_a - d_{k,\lambda})$. We would typically have $k^* = K$ when the $d_{k,\lambda}$'s are non-increasing. By construction, $\boldsymbol{\beta}^*$ minimizes Δ over $\{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta}\| \le 1\}$

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.

Figure 4 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 & 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 Figure 4 is more or less the behavior we would expect if we compared 1000 randomizations to 1000 other randomizations. Furthermore, from the third plot in Figure 4, we can see that rerandomization occasionally performs better than ridge rerandomization—particularly when K is large—but the differences in relative MSE across simulations are somewhat centered around zero.

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 plot of Figure 4 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.

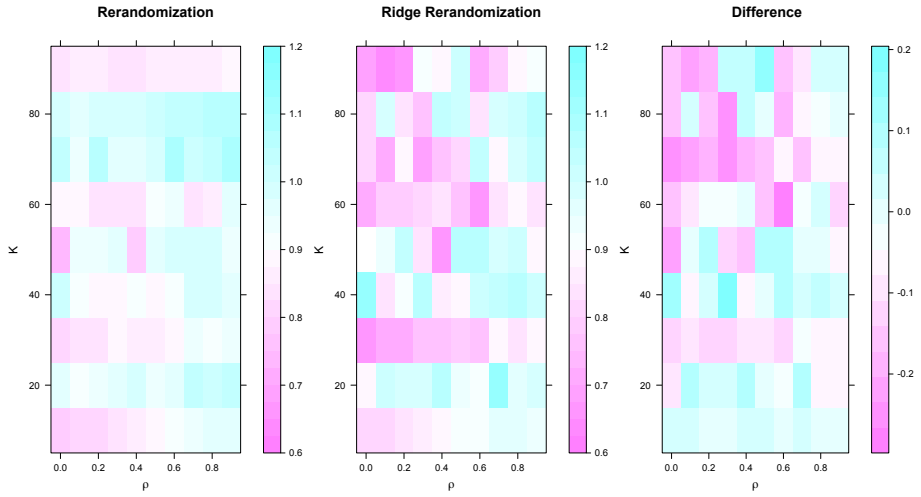

Figure 4: Relative MSE of $\hat{\tau} = \bar{y}_T - \bar{y}_C$ under rerandomization and ridge rerandomization (relative to randomization) for the $\boldsymbol{\beta}$ such that ridge rerandomization has the least competitive advantage over rerandomization, as well as the difference in relative MSE between the two (i.e., the second plot minus the first).

5.4 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 Figure 2 do not depend on $\boldsymbol{\beta}$, whereas the treatment effect estimation accuracy results in Figures 3 and 4 do. From Figure 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 Figure 3 we see that this can result in more precise treatment effect estimation. However, from Figure 4 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, 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 & Rubin (2012) and continuing with works such as Morgan & Rubin (2015), Branson et al. (2016), Li et al. (2016), and Zhou et al. (2017), 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 automatically placing a hierarchy of importance among the covariates according to their 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 randomiza-

tion. 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. To our knowledge, this is the first work to establish theoretical results for rerandomization criteria besides the Mahalanobis distance, which 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.

References

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

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. *The Annals of Applied Statistics*, *10*(4), 1958–1976.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, *1*(4), 200–232.

Cox, D. (2009). Randomization in the design of experiments. *International Statistical Review*, *77*(3), 415–429.

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*, (pp. 82–91). Springer.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180–193.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.

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

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 transactions on pattern analysis and machine intelligence*, *21*(3), 258–262.

Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, *59*(7), 751–766.

Li, X., & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, (just-accepted).

Li, X., Ding, P., & Rubin, D. B. (2016). Asymptotic theory of rerandomization in treatment-control experiments. *arXiv preprint arXiv:1604.00698*.

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, (pp. 431–446). Philosophy of Science Association.

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 & clinical pharmacology & toxicology*, *98*(3), 243–252.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 1936*, (pp. 49–55).

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263–1282.

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, *110*(512), 1412–1421.

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

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. *Journal of Multivariate Analysis*, *122*, 355–376.

Palombi, F., Toti, S., & Filippini, R. (2017). Numerical reconstruction of the covariance matrix of a spherically truncated multinormal distribution. *Journal of Probability and Statistics*, *2017*.

Papineau, D. (1994). The virtues of randomization. *The British journal for the philosophy of science*, *45*(2), 437–450.

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

Rosenberger, W. F., & Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*, (pp. 404–419).

Rubin, D. B. (1974). Multivariate matching methods that are equal percent bias reducing, i: Some examples. *ETS Research Report Series*, *1974*(2).

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

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585.

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

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Worrall, J. (2010). Evidence: philosophy of science meets medicine. *Journal of evaluation in clinical practice*, *16*(2), 356–362.

Zhou, Q., Ernst, P., Morgan, K. L., Rubin, D., & Zhang, A. (2017). Sequential rerandomization. *arXiv preprint arXiv:1706.04182*.

7 Appendix

7.1 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 $\widetilde{\mathbf{Z}} \,|\, \mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_K)$, we may write

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

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 $\lambda_1(\lambda_1+\lambda)^{-1}, ..., \lambda_K(\lambda_K+\lambda)^{-1}$. As a consequence, we have

$$M_\lambda \,|\, \mathbf{x} \;\sim\; (\mathbf{\Gamma}^\top \mathbf{Z})^\top \,\mathbf{Diag}\left(\left(\frac{\lambda_j}{\lambda_j+\lambda}\right)_{1\le j\le K}\right)(\mathbf{\Gamma}^\top \mathbf{Z}) \tag{31}$$

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

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

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

7.2 Proof of Theorem 4.2

Using the same notation and reasoning as for the proof of Lemma 4.1 in Section 7.1, in particular (31), we can write

$$\begin{aligned}
&\mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x}, M_\lambda \le a_\lambda) \\
&= \mathrm{Cov}\left(\mathbf{\Sigma}^{1/2}\mathbf{Z} \,\middle|\, \mathbf{x}, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda}(\mathbf{\Gamma}^\top\mathbf{Z})_j^2 \le a_\lambda\right) \\
&= \mathrm{Cov}\left(\mathbf{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)(\mathbf{\Gamma}^\top\mathbf{Z}) \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda}(\mathbf{\Gamma}^\top\mathbf{Z})_j^2 \le a_\lambda\right) \qquad (32) \\
&= \mathbf{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathrm{Cov}\left((\mathbf{\Gamma}^\top\mathbf{Z}) \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda}(\mathbf{\Gamma}^\top\mathbf{Z})_j^2 \le a_\lambda\right)\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathbf{\Gamma}^\top \\
&= \mathbf{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \le a_\lambda\right)\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathbf{\Gamma}^\top \qquad (33)
\end{aligned}$$

where (32) follows from the definition of $\mathbf{\Sigma}^{1/2} = \mathbf{\Gamma}\mathbf{Diag}\left(\sqrt{\lambda_{1:K}}\right)\mathbf{\Gamma}^\top$ along with the constructed independence of $\mathbf{Z}$ and $\mathbf{x}$ to get rid of the conditioning on $\mathbf{x}$, and (33) follows from $(\mathbf{\Gamma}^\top\mathbf{Z}) \sim \mathbf{Z}$ by orthogonality of $\mathbf{\Gamma}$ and standard Normality of $\mathbf{Z}$. All that is left now is to compute the conditional covariance matrix appearing in (33). 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 $k = 1, ..., K$, 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 $\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \leq a_\lambda\right)$ are given by

$$d_{k,\lambda} = \mathrm{Var}\left(Z_k^2 \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \leq a_\lambda\right) = \mathbb{E}\left[Z_k^2 \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \leq a_\lambda\right] \tag{34}$$

for all $k = 1, ..., K$. Now for the (ℓ, m)-element of $\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \leq a_\lambda\right)$ with $\ell \neq m$, 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 $i \neq \ell$ and $Z_\ell^* = -Z_\ell$, so that

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

which leads to

$$\mathrm{Cov}\left(Z_\ell, Z_m \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \leq a_\lambda\right) = 0 \tag{35}$$

for all $1 \leq \ell, m \leq K$ such that $\ell \neq m$. Combining (34) and (35) gives

$$\mathrm{Cov}\left(\mathbf{Z} \,\middle|\, \sum_{j=1}^K \frac{\lambda_j}{\lambda_j+\lambda} Z_j^2 \leq a_\lambda\right) = \mathbf{Diag}\left((d_{k,\lambda})_{1\leq k\leq K}\right) \tag{36}$$

Plugging(36) back into (33) finally yields

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

where the $d_{k,\lambda}$'s are given by (34). From the expression of $d_{k,\lambda}$, we immediately have $d_{k,\lambda} > 0$ for all $k = 1, ..., K$. By using Equation (13) from Palombi & Toti (2013), we also get

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

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

7.3 Proof of Corollary 4.1

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

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

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

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

is positive definite. This implies that

$$\mathbf{v}^\top \left(\mathbf{\Sigma} - \mathbf{\Gamma Diag}\left((\lambda_j \, d_{j,\lambda})_{1 \leq j \leq K} \right) \mathbf{\Gamma}^\top \right) \mathbf{v} \;>\; 0 \tag{37}$$

for all $\mathbf{v} \in \mathbb{R}^K \backslash \{0\}$. In particular, by using (37) with $\mathbf{v}$ chosen to be the k-th canonical basis vector of $\mathbb{R}^K$ (whose elements are all 0 except its k-th element equal to 1), we get, for all $k = 1, ..., K$,

$$\mathbf{\Sigma}_{kk} \;>\; \left(\mathbf{\Gamma Diag}\left((\lambda_j \, d_{j,\lambda})_{1 \leq j \leq K} \right) \mathbf{\Gamma}^\top \right)_{kk}. \tag{38}$$

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

$$\mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \,|\, \mathbf{x}, M_\lambda \leq a_\lambda \right) \;<\; \mathrm{Var}\left((\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)_k \,|\, \mathbf{x} \right)$$

□

7.4 Proof of Theorem 4.3

By using (18), we can write

$$\hat{\tau} \;=\; (\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C) \;=\; \tau + \boldsymbol{\beta}^\top (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) + (\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C) \tag{39}$$

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

$$\begin{aligned} \mathrm{Var}(\hat{\tau} \,|\, \mathbf{x}) \;&=\; \mathrm{Var}(\boldsymbol{\beta}^\top (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \,|\, \mathbf{x}) + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \,|\, \mathbf{x}) \\ &=\; \boldsymbol{\beta}^\top \mathbf{\Sigma} \boldsymbol{\beta} + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \,|\, \mathbf{x}) \end{aligned} \tag{40}$$

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

$$\begin{aligned} \mathrm{Var}(\hat{\tau} \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) \;&=\; \mathrm{Var}(\boldsymbol{\beta}^\top (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) \\ &=\; \boldsymbol{\beta}^\top \mathrm{Cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C \,|\, \mathbf{x}, M_\lambda \leq a_\lambda) \boldsymbol{\beta} + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \,|\, \mathbf{x}) \qquad (41) \\ &=\; \boldsymbol{\beta}^\top \mathbf{\Gamma Diag}\left((\lambda_k d_{k,\lambda})_{1 \leq k \leq K} \right) \mathbf{\Gamma}^\top \boldsymbol{\beta} + \mathrm{Var}(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C \,|\, \mathbf{x}) \qquad (42) \end{aligned}$$

where (41) follows from the conditional independence of $(\bar{\boldsymbol{\epsilon}}_T - \bar{\boldsymbol{\epsilon}}_C)$ and M_λ given $\mathbf{x}$, and (42) follows from Theorem 4.2. By plugging (40) into (42), we get

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

As explained by (37) in the proof of Corollary 4.1, the positive definiteness of the matrix $\mathbf{\Gamma Diag}\left((\lambda_k \, (1 - d_{k,\lambda}))_{1 \leq k \leq K} \right) \mathbf{\Gamma}^\top$ guarantees that

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

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