

PCA Rerandomization

Hengtao ZHANG¹, Guosheng YIN^{1,2*} , and Donald B. RUBIN^{3,4,5}

¹Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

²Department of Mathematics, Imperial College London, London, UK

³Yau Mathematical Sciences Center, Tsinghua University, Beijing, China

⁴Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, Pennsylvania, U.S.A.

⁵Department of Statistics, Harvard University, Cambridge, Massachusetts, U.S.A.

Key words and phrases: Covariate Balance; experimental design; Mahalanobis distance; principal component analysis; randomization.

MSC 2020: Primary 62K99; secondary 62P99.

Abstract: Mahalanobis distance of covariate means between treatment and control groups is often adopted as a balance criterion when implementing a rerandomization strategy. However, this criterion may not work well for high-dimensional cases because it balances all orthogonalized covariates equally. We propose using principal component analysis (PCA) to identify proper subspaces in which Mahalanobis distance should be calculated. Not only can PCA effectively reduce the dimensionality for high-dimensional covariates, but it also provides computational simplicity by focusing on the top orthogonal components. The PCA rerandomization scheme has desirable theoretical properties for balancing covariates and thereby improving the estimation of average treatment effects. This conclusion is supported by numerical studies using both simulated and real examples.

Résumé: La distance de Mahalanobis entre les moyennes des covariables de groupes traités et non traités est souvent utilisée comme critère d'équilibre lors de la mise en œuvre d'une stratégie de re-randomisation. Cela dit, ce critère peut ne pas fonctionner correctement pour les cas à grande dimension car il équilibre toutes les covariables orthogonalisées de manière égale. Les auteurs de ce travail proposent de recourir à l'analyse en composantes principales (ACP) afin d'identifier les sous-espaces appropriés dans lesquels la distance de Mahalanobis devrait être calculée. L'ACP peut non seulement réduire efficacement la dimensionnalité pour les covariables de grande dimension, mais elle offre également une simplicité de calcul en se concentrant sur les composantes orthogonales les plus importantes. Ce schéma de re-randomisation basé sur l'ACP possède des avantages théoriques intéressants pour équilibrer les covariables et, par conséquent, améliorer l'estimation des effets moyens du traitement. Les auteurs appuient leur conclusion par des études numériques utilisant à la fois des simulations et des exemples concrets.

1. INTRODUCTION

Randomized experiments have long been regarded as the gold standard to measure the effect of an intervention, because randomization can reduce the potential bias of estimates by balancing the covariate distributions between treatment groups on average. However, when pure (complete) randomization is implemented in practice, it often yields unbalanced allocations, so that

* Corresponding author: gyin@hku.hk

[Correction added April 20, 2023 after first online publication: minor copyediting errors were corrected throughout this article.]

the groups should be rerandomized before the experiment is actually conducted. Although rerandomization has been discussed earlier (Fisher, 1926; Cox, 2009; Worrall, 2010), its formal theoretical framework was not well established until the publication of Morgan & Rubin (2012). Using Mahalanobis distance as the balance criterion for treatment–control experiments, rerandomization was shown to improve the covariate balance and the precision of estimated treatment effects. Following the work of Morgan & Rubin (2012), effort has been made to extend or modify such rerandomization schemes. For example, Morgan & Rubin (2015) proposed a rerandomization strategy for covariates with different tiers of anticipated importance with respect to the outcome variable. The extension of rerandomization to a 2^K factorial design was developed by Branson, Dasgupta & Rubin (2016) based on a real example with educational data. Zhou et al. (2018) considered rerandomization for experiments with sequentially enrolled units. Li, Ding & Rubin (2018, 2020) investigated the asymptotic properties of the standard treatment effect estimator for the treatment–control settings and 2^K factorial designs, respectively. Li & Ding (2020) further established asymptotic properties for the combination of rerandomization and regression adjustment. Wang, Wang & Liu (2021) studied the statistical properties of stratification with rerandomization. Zhang & Yin (2021) incorporated the response information into rerandomization for ethical concerns in clinical trials. Yang, Qu & Li (2021) proposed rerandomization for survey experiments, using asymptotic theories.

All the aforementioned methods use Mahalanobis distance as the covariate balance measure, due to its several appealing characteristics. First, it is invariant to any affine transformation of the original covariates. Second, Morgan & Rubin (2012) showed that not only can Mahalanobis distance preserve the unbiasedness of the treatment effect estimator as well as the balance of covariate means between equal-sized treatment and control groups, but it also reduces an equal percent of sampling variance for each covariate. Apart from rerandomization, Mahalanobis distance is also widely applied in matching methods for observational studies (Rubin, 1973a,b, 1979, 1980; Rosenbaum & Rubin, 1985; Stuart, 2010).

Despite the advantages discussed above, the full-rank Mahalanobis distance may not work well for rerandomization with high-dimensional data (Branson & Shao, 2021), because it is difficult to balance equally a large number of covariates with different magnitudes of sampling variances. In related work, Morgan & Rubin (2015) proposed balancing the covariates hierarchically using prespecified tiers of importance of covariates related to the outcome. Branson & Shao (2021) pointed out that it might be difficult to specify the relative importance for a large number of covariates *a priori*. They proposed including a ridge term in Mahalanobis distance, which puts more emphasis on the top principal components of the covariate space after a principal component analysis (PCA). However, the ridge rerandomization of Branson & Shao (2021) relies on complicated Monte Carlo integration and constraint optimization to determine the value of the ridging parameter. Rather than using Mahalanobis distance, Johansson & Schultzberg (2020) proposed a different rank-based balance measure for rerandomization, but their heuristic metric is designed for longitudinal data where the pre-experimental outcomes are available to estimate the relative importance of each covariate. Moreover, the theoretical properties have not yet been developed under their proposed balance metric.

Following a PCA, we propose using only the top principal components to calculate Mahalanobis distance in the associated subspace and then perform rerandomization. Our PCA rerandomization can be viewed as using a lower dimensional alternative to the full-rank Mahalanobis distance. Because PCA rerandomization only reduces the variance of the selected top components, it imposes more shrinkage on them relative to full-rank rerandomization, given the same acceptance probability. Moreover, the lower dimensional orthogonality of top principal components simplifies the covariance matrix into a diagonal matrix and thus improves the computational efficiency when calculating Mahalanobis distance. We establish a theory for PCA rerandomization, including the sampling distribution of the modified balance criterion and the

variance reduction properties for the treatment effect estimator and covariate mean differences compared with complete randomization. Practically, despite using PCA, our method is as easy to implement as the original rerandomization, and delivers desirable performance without cumbersome parameter specification or increased computation as required in ridge rerandomization.

In Section 2, we review the rerandomization framework based on Mahalanobis distance (Morgan & Rubin, 2012). We present the details of PCA rerandomization and its theoretical properties in Section 3. Section 4 reports the results of numerical experiments, which show the potential desirable performance of our proposed method compared with other randomization schemes. Section 5 concludes with a discussion.

2. RERANDOMIZATION WITH MAHALANOBIS DISTANCE

Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times d}$ be the covariate matrix representing n trial participants with $\mathbf{x}_i \in \mathbb{R}^d$. Assume that all $\mathbf{x}_i$ s are standardized to have zero mean and unit variance: $\mathbf{X}^\top \mathbf{1}_n = \mathbf{0}$, where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector of length n with all components equal to 1. We focus on treatment-versus-control experiments and define $\mathbf{W} = (W_1, \dots, W_n)^\top \in \{0, 1\}^n$ to be the random vector of indicators for allocation, where $W_i = 1$ if the i th unit is assigned to the treatment group and $W_i = 0$ otherwise. For simplicity, we further impose the constraint on $\mathbf{W}$ such that the treatment and control groups are of the same size,

$$\sum_{i=1}^n W_i = \sum_{i=1}^n (1 - W_i) = n/2. \quad (1)$$

Given $\mathbf{W}$ and $\mathbf{X}$, let

$$\bar{\mathbf{x}}_T = \frac{2}{n} \mathbf{X}^\top \mathbf{W} \quad \text{and} \quad \bar{\mathbf{x}}_C = \frac{2}{n} \mathbf{X}^\top (\mathbf{1}_n - \mathbf{W}), \quad (2)$$

which represent the mean vectors of the covariates for the treatment and control groups, respectively. Under the potential outcome framework (Neyman, 1923; Rubin, 1974; Imbens & Rubin, 2015), each unit is associated with two potential outcomes $y_i(1)$ or $y_i(0)$ corresponding to the situation where the unit is assigned to the treatment group or the control group. Only one outcome can be observed after the allocation, written as $y_i = W_i y_i(1) + (1 - W_i) y_i(0)$. The goal of causal inference is to estimate the sample average treatment effect (SATE),

$$\tau = \frac{1}{n} \sum_{i=1}^n \{y_i(1) - y_i(0)\}.$$

The standard estimator of SATE is the mean difference of the observed outcomes between the two groups,

$$\hat{\tau} = \bar{y}_T - \bar{y}_C = \frac{2}{n} \{ \mathbf{W}^\top \mathbf{y} - (\mathbf{1}_n - \mathbf{W})^\top \mathbf{y} \}, \quad (3)$$

where $\mathbf{y} = (y_1, \dots, y_n)^\top$. In randomized experiments, balancing covariates across treatment and control groups generally leads to a more precise estimator $\hat{\tau}$. Complete randomization only balances covariates between the treatment and control groups on average, and thus a particular realized allocation can be unbalanced, thereby adversely affecting statistical inference.

Morgan & Rubin (2012) formally established the rerandomization framework using Mahalanobis distance, M , as the balance measure, where

$$M = (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C), \quad (4)$$

and $\Sigma = \text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}) = 4\text{cov}(\mathbf{X})/n$ is the covariance matrix of $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ with respect to all $\mathbf{W}$ satisfying (1), and $\text{cov}(\mathbf{X}) = \mathbf{X}^\top \mathbf{X}/(n-1)$ is the sample covariance matrix of $\mathbf{X}$. When Σ is singular, the pseudo-inverse is adopted. Rather than performing one single randomization, rerandomization continues generating feasible allocation vectors $\mathbf{W}$ until the corresponding M is smaller than a predefined threshold a ($a > 0$). The first $\mathbf{W}$ with $M \leq a$ is chosen for the actual allocation in the experiment, where the threshold a can be determined by controlling the acceptance probability p_a with $\Pr(M \leq a | \mathbf{X}) = p_a$. If $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where the randomness comes from all possible allocations, the distribution of Mahalanobis distance $M | \mathbf{X}$ follows a χ_d^2 distribution (Morgan & Rubin, 2012). The normality $\mathcal{N}(\mathbf{0}, \Sigma)$ for $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}$ can be derived asymptotically with respect to all possible allocations based on the finite population central limit theorem as shown in Section 3 and Example 8 in Li & Ding (2017). Therefore, the distribution of the discrete Mahalanobis statistic can be approximated by a continuous chi-squared distribution when the sample size n is large enough. For ease of exposition, we use the approximate χ_d^2 distribution for the Mahalanobis statistic $M | \mathbf{X}$ throughout.

Rerandomization using Mahalanobis distance has several attractive properties. First, rerandomization preserves the unbiased estimation of τ ; that is, $\mathbb{E}(\hat{\tau} | \mathbf{X}, M \leq a) = \tau$, because it balances the mean difference of any observed and unobserved covariate x : $\mathbb{E}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M \leq a) = \mathbf{0}$. Second, it reduces the variance of each covariate by the same proportion, $100(1 - v_a)\%$ relative to complete randomization,

$$\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M \leq a) = v_a \text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}), \quad (5)$$

where $v_a = \Pr(\chi_{d+2}^2 \leq a) / \Pr(\chi_d^2 \leq a) \in (0, 1)$. Similarly, if there exists a linear relationship between the outcome y and covariate $\mathbf{x}$, the variance reduction for the estimated treatment effect, $\text{var}(\hat{\tau} | \mathbf{X}, M \leq a) = \{1 - (1 - v_a)R^2\} \text{var}(\hat{\tau} | \mathbf{X})$, where R^2 is the multiple squared correlation between y and $\mathbf{x}$.

Branson & Shao (2021) developed ridge rerandomization by including a ridge term when calculating Mahalanobis distance,

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

where $\mathbf{I}_d$ denotes the d -dimensional identity matrix; they showed that such ridging automatically assigns more weight to the top components and can provide better balance in high-dimensional/high-collinearity cases. We propose using the top principal components rather than the original covariates (corresponding to using all principal components) to perform the rerandomization. Our lower dimensional method inherits most of the theoretical properties of rerandomization and can be more convenient to implement without possibly cumbersome specifications of tuning parameters as in ridge rerandomization.

3. PCA RERANDOMIZATION

3.1. Rerandomization using Principal Components

Define $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ to be the singular value decomposition of $\mathbf{X}$, where $\mathbf{U} \in \mathbb{R}^{n \times p}$ and $\mathbf{V} \in \mathbb{R}^{d \times p}$ correspond to the matrices of the left and right singular vectors, with $\mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}_p$ and $p = \min\{n, d\}$; $\mathbf{D} = \text{diag}\{\sigma_1, \dots, \sigma_p\}$ is a diagonal matrix composed of non-negative singular (i.e., eigen) values $\sigma_1 \geq \dots \geq \sigma_p > 0$. We focus on the common case with $p = d$ (i.e., $n > d$) so that $\mathbf{Z} = (z_{ij}) = \mathbf{U}\mathbf{D}$ are the principal components of $\mathbf{X}$. Furthermore, those components are obtained from the correlation matrix of covariates because $\mathbf{X}$ is assumed to be standardized in Section 2. It is common to perform standardization in the PCA literature (Singh & Harrison, 1985; Jolliffe & Morgan, 1992; Baxter, 1995), to avoid the influence of different scales among covariates

(Jolliffe, 2002, Section 2.3). Let $\mathbf{Z}_k = \mathbf{U}_k \mathbf{D}_k = (\mathbf{z}_1^{(k)}, \dots, \mathbf{z}_n^{(k)})^\top \in \mathbb{R}^{n \times k}$ denote the matrix of the top k ($k \leq d$) principal components of $\mathbf{X}$, where $\mathbf{z}_i^{(k)} \in \mathbb{R}^{k \times 1}$ contains the first k elements of the i th row of $\mathbf{Z}$, and $\mathbf{U}_k \in \mathbb{R}^{n \times k}$ is the first k columns of $\mathbf{U}$ and $\mathbf{D}_k \in \mathbb{R}^{k \times k}$ is the top k -dimensional submatrix of $\mathbf{D}$. Similarly, let $\tilde{\mathbf{Z}}_{d-k} = \tilde{\mathbf{U}}_{d-k} \tilde{\mathbf{D}}_{d-k} \in \mathbb{R}^{n \times (d-k)}$ denote the last $d-k$ principal components, where $\tilde{\mathbf{U}}_{d-k} \in \mathbb{R}^{n \times (d-k)}$ is the last $d-k$ columns of $\mathbf{U}$ and $\tilde{\mathbf{D}}_{d-k} \in \mathbb{R}^{(d-k) \times (d-k)}$ is the corresponding $(d-k)$ -dimensional submatrix of $\mathbf{D}$. We calculate Mahalanobis distance based on the top k principal components,

$$M_k = \left(\bar{\mathbf{z}}_T^{(k)} - \bar{\mathbf{z}}_C^{(k)} \right)^\top \boldsymbol{\Sigma}_z^{-1} \left(\bar{\mathbf{z}}_T^{(k)} - \bar{\mathbf{z}}_C^{(k)} \right), \quad (6)$$

where $\bar{\mathbf{z}}_T^{(k)}$ and $\bar{\mathbf{z}}_C^{(k)}$ are defined similarly following (2),

$$\bar{\mathbf{z}}_T^{(k)} = \frac{2}{n} \mathbf{Z}_k^\top \mathbf{W} \quad \text{and} \quad \bar{\mathbf{z}}_C^{(k)} = \frac{2}{n} \mathbf{Z}_k^\top (\mathbf{I}_n - \mathbf{W}),$$

and $\boldsymbol{\Sigma}_z = C_n \mathbf{Z}_k^\top \mathbf{Z}_k = C_n \mathbf{D}_k^2$ with $C_n = 4/(n^2 - n)$. For selecting a treatment assignment, we proceed by generating $\mathbf{W}$ until the criterion $M_k \leq a_k$ is reached. This is referred to as PCA rerandomization.

For a particular randomized allocation, let $\bar{z}_{T,j}$ and $\bar{z}_{C,j}$ be the mean values of the j th principal component for the treatment and control groups respectively. Let s_j^2 denote the sample variance of the j th component. It can be shown that $s_j^2 = (z_{1j}^2 + z_{2j}^2 + \dots + z_{nj}^2)/(n-1) = \sigma_j^2/(n-1)$, because $\mathbf{Z}$ is also centred, $\mathbf{1}_n^\top \mathbf{Z} = \mathbf{1}_n^\top \mathbf{XV} = \mathbf{0}$. Thus, we can rewrite Equation (6) as

$$M_k = \frac{n}{4} \sum_{j=1}^k \left(\frac{\bar{z}_{T,j} - \bar{z}_{C,j}}{s_j} \right)^2,$$

which is the sum of the standardized mean differences of the top k principal components. Because $\mathbf{Z}$ is obtained from the affine transformation on covariates $\mathbf{Z} = \mathbf{XV}$, it can be shown that

$$M = \frac{n}{4} \sum_{j=1}^d \left(\frac{\bar{z}_{T,j} - \bar{z}_{C,j}}{s_j} \right)^2 \quad \text{and} \quad M_\lambda = \frac{n}{4} \sum_{j=1}^d \frac{\sigma_j^2}{\sigma_j^2 + \lambda/C_n} \left(\frac{\bar{z}_{T,j} - \bar{z}_{C,j}}{s_j} \right)^2,$$

where the former is the original Mahalanobis distance and the latter corresponds to ridge rerandomization. Therefore, PCA rerandomization is a truncated version of the original rerandomization, that is, $M_k \leq M$ because of $k \leq d$ where the equality holds only if $k = d$. Furthermore, M_λ is a weighted version of M , with weight $\sigma_j^2/(\sigma_j^2 + \lambda/C_n)$ attached to all components, that is, there is no dimension reduction. Therefore, ridge rerandomization can be viewed as a smooth counterpart to PCA rerandomization, which uses a binary weight,

$$1_{j \leq k}(j) = \begin{cases} \sigma_j^2/(\sigma_j^2 + 0) = 1, & \text{if } j \leq k, \\ \sigma_j^2/(\sigma_j^2 + \infty) = 0, & \text{if } j > k. \end{cases}$$

Although there are two tuning parameters in PCA rerandomization, they are easy to specify. In the PCA (Jolliffe, 2002, Chapter 6), the number of top components k can be readily determined using the variation explained criterion, the Kaiser rule (Kaiser, 1960) or the scree plot (Cattell, 1966).

The variation explained criterion selects the components based on the percent of total variation that they account for,

$$k = \min_j \left\{ j \left| \frac{\sum_{i=1}^j \sigma_i^2}{\sum_{i=1}^d \sigma_i^2} \geq \gamma_k \right. \right\}, \quad (7)$$

where $\gamma_k \in (0, 1)$ is a prespecified constant. Kaiser's rule maintains the top components whose variations are larger than the average value $(\sigma_1^2 + \cdots + \sigma_d^2)/d$. In the scree plot, the number of components is chosen visually by finding the elbow point in the plot of the proportion of variation explained $\sigma_i^2/(\sigma_1^2 + \cdots + \sigma_d^2)$ versus the component index i . Given k , we determine the threshold a_k via the acceptance probability p_{a_k} analogous to the full-rank rerandomization, $\Pr(M_k \leq a_k | \mathbf{X}) = p_{a_k}$, where we use the χ_k^2 distribution as an approximation to that of $M_k | \mathbf{X}$.

3.2. Theoretical Properties

Given k and a_k , several statistical properties can be derived for PCA rerandomization, and we defer the corresponding technical details to the Appendix. PCA rerandomization balances the covariates between the treatment and control groups on average, and additionally leads to unbiased estimation of τ .

Theorem 1. *Given a constant $a_k > 0$,*

$$\mathbb{E}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M_k \leq a_k) = \mathbf{0} \quad \text{and} \quad \mathbb{E}(\hat{\tau} | \mathbf{X}, M_k \leq a_k) = \tau.$$

According to the definition of M_k , it has the same value for both allocations $\mathbf{W}$ and $\mathbf{1}_n - \mathbf{W}$ for any given threshold $a_k > 0$. Assuming that Equation (1) holds in PCA rerandomization, the unbiasedness for $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ and $\hat{\tau}$ follows Theorem 2.1 and Corollary 2.2 in Morgan & Rubin (2012).

The covariate balance in Theorem 1 can be further extended to the unobserved covariates, as implied by Corollary 2.2 of Morgan & Rubin (2012). In addition to removing the conditional bias, PCA rerandomization tends to make the difference of covariate means and $\hat{\tau}$ more concentrated. Following the rerandomization literature (Morgan & Rubin, 2012, 2015; Branson & Shao, 2021), we adopt the normal approximation $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for theoretical analysis of the covariance structure of $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ and the variance of $\hat{\tau}$ for PCA rerandomization.

Theorem 2. *Let k denote the number of selected top principal components. Under the normal approximation $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $M_k | \mathbf{X}$ follows a chi-squared distribution with the degrees of freedom being the number of top principal components selected, that is, $M_k | \mathbf{X} \sim \chi_k^2$.*

The original rerandomization yields $M | \mathbf{X} \sim \chi_d^2$, because it uses all d components. One major application of Theorem 2 is to specify the threshold a_k via the probability p_{a_k} such that $p_{a_k} = \Pr(\chi_k^2 \leq a_k)$. Given the same acceptance probability $p_a = p_{a_k}$, fewer degrees of freedom yield a smaller threshold, that is, $a_k < a$. Moreover, it is easier for our method to determine a_k compared with ridge rerandomization, which has to specify the threshold from a mixture distribution.

Theorem 3. *Given the top k principal components and the threshold $a_k > 0$, under the normal approximation $(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$,*

$$\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M_k \leq a_k) = C_n \mathbf{V} \begin{pmatrix} v_{a_k} \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix} \mathbf{V}^\top,$$

where $C_n = 4/(n^2 - n)$ and $v_{a_k} = \Pr(\chi_{k+2}^2 \leq a_k) / \Pr(\chi_k^2 \leq a_k)$.

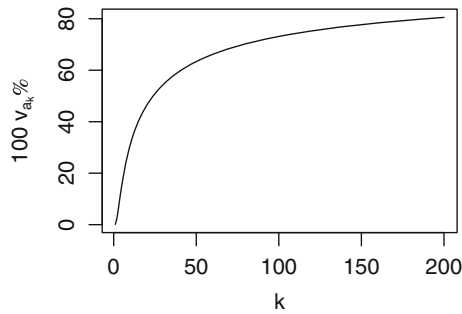


FIGURE 1: The percent reduction in variance $100v_{a_k}\%$ of PCA rerandomization versus the number of principal components k given the acceptance probability $p_{a_k} = 0.05$.

Theorem 3 states the obvious fact that PCA rerandomization only reduces the variation of the top k selected principal components, for which the percent reduction in variance (PRV) is $100(1 - v_{a_k})\%$. This strategy automatically identifies and balances the most variable subspace of original covariates. Furthermore, the value v_{a_k} in PCA rerandomization is at most its rerandomization counterpart v_a given the same acceptance probability, as evidenced by Figure 1.

From Morgan & Rubin (2012), the covariance reduction for rerandomization is

$$\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M \leq a) = C_n \mathbf{V} \begin{pmatrix} v_a \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & v_a \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix} \mathbf{V}^\top. \quad (8)$$

Therefore, smaller v_{a_k} means that we reduce more variance along the top principal axes than the original rerandomization. In contrast to ridge rerandomization, our scheme only reduces the same percent of variance for the selected k components, whereas ridge rerandomization reduces for all components but with different percents, and the top components receive more shrinkage. That is,

$$\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M_\lambda \leq a_\lambda) = C_n \mathbf{V} \text{diag}(\xi_{\lambda,1} \sigma_1^2, \dots, \xi_{\lambda,d} \sigma_d^2) \mathbf{V}^\top,$$

where $0 < \xi_{\lambda,1} \leq \dots \leq \xi_{\lambda,d}$ are constants defined in Theorem 4.2 of Branson & Shao (2021) for a given λ .

Corollary 1. *Given the conditions in Theorem 3 and the (j, j) th element of Σ , $\Sigma_{jj} > 0$, for $j \in \{1, \dots, d\}$, the PRV of $\bar{x}_{T,j} - \bar{x}_{C,j}$ is $100(1 - v_{a_{k,j}})\%$ for $j \in \{1, \dots, k\}$, where*

$$v_{a_{k,j}} = \frac{\left(C_n \mathbf{V} \text{diag} \left\{ v_{a_k} \mathbf{D}_k^2, \tilde{\mathbf{D}}_{d-k}^2 \right\} \mathbf{V}^\top \right)_{jj}}{\Sigma_{jj}} \in (0, 1).$$

To assess how PCA rerandomization can improve the estimation of τ , we follow Morgan & Rubin (2012) to assume that the treatment effect is additive,

$$y_i(W_i) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \tau W_i + \varepsilon_i, \quad (9)$$

where $\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}$ is the projection of the outcome y_i onto the subspace spanned by $(\mathbf{1}, \mathbf{X})$, and ε_i is the residual of y_i orthogonal to the linear subspace spanned by $\mathbf{X}$.

Theorem 4. *Given the conditions in Theorem 3, if the data arise from Equation (9) and $\hat{\tau}$ is normally distributed for given $\mathbf{X}$, then for any $\boldsymbol{\beta} \in \mathbb{R}^d$,*

$$\text{var}(\hat{\tau}|\mathbf{X}) - \text{var}(\hat{\tau}|\mathbf{X}, M_k \leq a_k) = C_n \boldsymbol{\beta}^\top \mathbf{V} \begin{pmatrix} (1 - v_{a_k}) \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V}^\top \boldsymbol{\beta} \geq 0,$$

where the equality holds only when the first k components of $\mathbf{V}^\top \boldsymbol{\beta}$ are all zeros.

Using PCA rerandomization, the precision of the treatment effect estimator $\hat{\tau}$ never deteriorates in comparison to complete randomization. As shown in the proof in the Appendix, the sampling variance reduction of $\hat{\tau}$ is closely related to that of covariates for any rerandomization scheme,

$$\text{var}(\hat{\tau}|\mathbf{X}) - \text{var}(\hat{\tau}|\mathbf{X}, M \leq a) = \boldsymbol{\beta}^\top \{ \text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C|\mathbf{X}) - \text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C|\mathbf{X}, M \leq a) \} \boldsymbol{\beta}. \quad (10)$$

Therefore, the variance reduction of $\hat{\tau}$ is governed by all principal components for the original rerandomization and ridge rerandomization due to the fact that $\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C|\mathbf{X}) > \text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C|\mathbf{X}, M \leq a)$ under both schemes, whereas PCA rerandomization only considers the first k principal components.

Once the PCA rerandomization obtains an acceptable allocation vector $\mathbf{W}$, the trial starts assigning treatment to participants and outcomes $\mathbf{y}$ would be observed. The point estimator $\hat{\tau}$ can then be calculated following Equation (3) for the treatment effect τ and the randomization test can be readily performed to further analyze the data (Morgan & Rubin, 2012). In the randomization test, we keep the observed outcomes fixed and regenerate many acceptable allocations from PCA rerandomization to create the null distribution of a predefined test statistic. The P -value can then be obtained by comparing the observed test statistic with its empirical null distribution. The confidence interval of the treatment effect can also be derived by inverting the randomization test. One may refer to Rosenbaum (2010, Section 2.4) and Imbens & Rubin (2015, Chapter 5) for more detail on randomization-based inference.

4. NUMERICAL STUDIES

4.1. Simulation Settings

Our Monte Carlo study can be compactly described as a $4 \times 4 \times 3 \times 6 \times 2 \times 2 \times 2$ factorial design, where the seven factors with their respective levels are shown in Table 1. The factors can be divided into two categories at the design stage of causal inference following the setups in Rubin (1979) and Gutman & Rubin (2013, 2015, 2017). Known factors, including the characteristics of covariate distributions, sample size, and rerandomization schemes, are explicitly known to the investigator or can be estimated without using the outcome data. Unknown factors, including the variance of residuals and response surfaces, cannot be estimated without outcome data, and are generally unknown at the design stage.

We generate $n \in \{100, 200, 500, 1000\}$ samples from a d -dimensional multivariate normal distribution, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, (1 - \rho)\mathbf{I}_d + \rho\mathbf{1}_d\mathbf{1}_d^\top)$, where the correlation coefficient $\rho \in \{0.1, 0.5, 0.9\}$ and the dimension $d \in \{10, 50, 90, 180\}$. The sample size (n) and the covariate dimension (d) are factors known at the design stage, whereas the correlation ρ is an estimable factor that is also essentially known to the investigator. Half of the n units are assigned to the treatment group and the other half to the control group, where the allocation of units is generated by one of the three randomization schemes under comparison. Given the covariates $\mathbf{x}$ and allocation variable W , the outcome is simulated from a model with an additive treatment effect τ ,

$$y = g(\mathbf{x}, \boldsymbol{\beta}) + \tau W + \varepsilon, \quad (11)$$

TABLE 1: Seven factors and the corresponding levels in the simulation study.

Factors	Levels	Descriptions
n	{ 100, 200, 500, 1000 }	Sample size
d	{ 10, 50, 90, 180 }	Dimension of covariates
ρ	{ 0.1, 0.5, 0.9 }	Correlation coefficient of covariates
Scheme	{ ReR, Ridge-ReR, PCA-ReR $_{\gamma_k=0.5,0.7,0.9;\text{Kaiser}}$ }	Rerandomization (ReR) scheme
$g(\mathbf{x}, \boldsymbol{\beta})$	{ $\mathbf{x}^\top \boldsymbol{\beta}, \exp(\mathbf{x})^\top \boldsymbol{\beta}$ }	Response surface
$\boldsymbol{\beta}$	{ $\mathbf{1}_d, (\mathbf{1}_{d/2}^\top, 2 \times \mathbf{1}_{d/2}^\top)^\top$ }	Coefficient vector in the response surface
σ_ϵ^2	{ 0.5, 1 }	Residual variance

where $g(\mathbf{x}, \boldsymbol{\beta})$ represents the response surface function that is unknown to the investigator but estimable to some extent, and the residual follows a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, also unknown to the investigator. We adopt two different functions for the response surface: $g(\mathbf{x}, \boldsymbol{\beta}) \in \{\mathbf{x}^\top \boldsymbol{\beta}, \exp(\mathbf{x})^\top \boldsymbol{\beta}\}$ with $\boldsymbol{\beta} \in \{\mathbf{1}_d, (\mathbf{1}_{d/2}^\top, 2 \times \mathbf{1}_{d/2}^\top)^\top\}$. The residual variance σ_ϵ^2 takes values from {0.5, 1}, and we set $\tau = 1$.

We compare PCA rerandomization (PCA-ReR) with three other randomization schemes: (a) complete randomization (CR) with constraint (1), (b) original rerandomization (ReR), and (c) ridge rerandomization (Ridge-ReR). For a fair comparison, we set the same acceptance probability for all rerandomization-based methods, that is, $p_{a_k} = p_{a_\lambda} = p_a = 0.05$, corresponding to PCA-ReR, Ridge-ReR, and ReR. For PCA-ReR, we choose the top principal components using the variation explained criterion with $\gamma_k \in \{0.5, 0.7, 0.9\}$ in (7) and Kaiser’s rule, respectively. The optimal ridge coefficient λ is determined by the procedure given in Branson & Shao (2021). It is possible that ReR may never find an acceptable allocation for a large d because the high-dimensional covariate mean difference $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ may not well approximate the normal distribution under a given sample size and thus the threshold a may be too strict. The number of randomizations required for an acceptable allocation of any rerandomization scheme ideally follows a geometric distribution with the expectation as the inverse of acceptance probability, which is $1/0.05 = 20$ in our setting. Hence, we conduct at most 10,000 randomizations for each rerandomization method, for fair comparisons, and we use the assignment corresponding to the minimum Mahalanobis distance if no allocation can meet the threshold.

Three criteria are adopted for performance evaluation: the covariate balance, the estimation precision of treatment effect, and the number of randomizations. Because all approaches yield balanced covariates on average, we adopt the average variance of the mean difference $\bar{\mathbf{x}}_{T,j} - \bar{\mathbf{x}}_{C,j}$ across all covariates $j \in \{1, \dots, d\}$, denoted by $\bar{\sigma}^2$, to evaluate the empirical balance. For the treatment effect, we choose the mean squared error (MSE) of $\hat{\tau}$ as the evaluation metric. To normalize these metrics, we compute the reduction percents of $\bar{\sigma}^2$ and MSE for each rerandomization scheme relative to CR, denoted by $r_{\bar{\sigma}^2}$ and r_{MSE} respectively. We also record the number of randomizations for generating a feasible allocation under each rerandomization method. In addition, we consider ways of selecting the number of top components k for PCA-ReR to explore its performances under different settings.

Given a triplet of (n, d, ρ) , we simulate 2000 independent covariate matrices. For each covariate matrix, four randomization schemes are used to generate allocations and then the responses are simulated from model (11) for all combinations of $\{g(\mathbf{x}, \boldsymbol{\beta}), \boldsymbol{\beta}, \sigma_\epsilon^2\}$. Following Rubin (1979), we use the same covariate matrix when comparing different rerandomization methods, and the covariate matrices with smaller n and d , such as $(n, d) = (100, 10)$, are obtained

TABLE 2: ANOVA results of the 13 most influential factors for r_{σ^2} , which is the reduction percent of the average empirical variance of $\bar{x}_{T,j} - \bar{x}_{C,j}$ across all covariates ($j = 1, \dots, d$) relative to complete randomization.

Factors	DF	MS	F-ratio
ρ	2	50.16	57,913
d	3	10.64	12,289
Scheme	5	3.20	3695
$\rho \times \text{scheme}$	10	2.53	2917
$d \times \rho$	6	0.76	874
$d \times \text{scheme}$	15	0.57	656
$d \times \rho \times \text{scheme}$	30	0.07	81
$n \times \rho$	6	0.04	49
$n \times \text{scheme}$	15	0.04	41
$n \times \rho \times \text{scheme}$	30	0.02	24
$n \times d \times \text{scheme}$	45	0.01	16
n	3	0.01	9
$n \times d$	9	0.01	8
Residuals	2592	0.001	

Note: DF is the degrees of freedom, MS represents the between-configuration mean square, and the last row “Residuals” refers to the within-configuration mean square.

as the first 100 rows and 10 columns of the covariate matrices with $(n, d) = (1000, 180)$ given the same ρ . This nested design strategy minimizes the number of random samples and correlates the results of different rerandomization methods, which make comparisons more precise.

All evaluation metrics are calculated under each configuration of factors based on 2000 replications. We divide 2000 replications into 10 separate groups of size 200 to create repeated metrics for calculating the within-configuration mean square in ANOVA. For each group, we compute the three metrics: r_{σ^2} , r_{MSE} , and the number of randomizations. Under each rerandomization scheme, we can obtain $4 \times 4 \times 3 \times 10 = 480$ different values for both r_{σ^2} and the number of randomizations based on the levels of n , d , and ρ , as well as $4 \times 4 \times 3 \times 2^3 \times 10 = 3840$ different values for r_{MSE} with respect to all factors other than the scheme.

4.2. Comparisons of Rerandomization Schemes

Following Rubin (1979) and Gutman & Rubin (2013), we begin with three separate ANOVAs corresponding to three evaluation metrics based on the seven-factor design, to identify the most influential factors. In the ANOVA, we consider the full factorial design with the main effects of the factors as well as all the interactions. The relative importance of factors and their interactions are evaluated by the F -ratios or, equivalently, the ANOVA mean squares.

Table 2 presents the ANOVA results of r_{σ^2} for known factors, where the covariate dimension (d) and correlation (ρ) strongly influence the covariate balance metric and the first three factors (ρ , d , and scheme) account for around 77% of the total sum of squares. We only report the top 13 factorial effects because the rest have much smaller mean squares. The residual mean square, corresponding to the within-configuration variability, is rather small, which demonstrates the stability of our results. The ANOVA results of r_{MSE} in Table 3 only display the top 20 influential

TABLE 3: ANOVA results of the 20 most influential factors based on the ANOVA F -ratio for r_{MSE} , which is the reduction percent of mean squared error (MSE) of $\hat{\tau}$ relative to complete randomization.

Factors	DF	MS	F -ratio
$g(\mathbf{x}, \boldsymbol{\beta})$	1	201.57	47,945
ρ	2	114.11	27,142
Scheme	5	82.88	19,713
d	3	61.91	14,725
$\rho \times g(\mathbf{x}, \boldsymbol{\beta})$	2	28.29	6728
$\rho \times \text{scheme}$	10	15.18	3611
$d \times \rho$	6	5.92	1408
$d \times g(\mathbf{x}, \boldsymbol{\beta})$	3	3.43	816
$d \times \text{scheme}$	15	2.63	626
Scheme $\times g(\mathbf{x}, \boldsymbol{\beta})$	5	2.29	544
$d \times \rho \times \text{scheme}$	30	0.96	228
$n \times \text{scheme}$	15	0.79	188
$\rho \times \text{scheme} \times g(\mathbf{x}, \boldsymbol{\beta})$	10	0.79	187
$n \times \rho$	6	0.65	155
$d \times \rho \times g(\mathbf{x}, \boldsymbol{\beta})$	6	0.41	97
$n \times \rho \times \text{scheme}$	30	0.32	77
$n \times d \times \text{scheme}$	45	0.16	39
$n \times g(\mathbf{x}, \boldsymbol{\beta})$	3	0.16	39
n	3	0.13	32
$n \times d$	9	0.11	27
Residuals	20,736	0.004	

Note: DF is the degrees of freedom, MS represents the between-configuration mean square, and the last row “Residuals” refers to the within-configuration mean square.

factorial factors based on the F -ratios. The covariate dimension (d), the correlation coefficient (ρ), and the type of response surface $g(\mathbf{x}, \boldsymbol{\beta})$ are the most important factors in addition to the scheme. The top four factors explain around 68% of the total sum of squares, and the residuals have a small mean square. From Table 4, we can identify the scheme, d , and n as the most influential factors for the number of randomizations, and about 96% of the total sum of squares can be explained by the first seven factorial effects. After identifying the crucial factors, we take the average of each metric over the levels of all the unselected factors.

The first panel of Table 5 compares different rerandomization schemes in terms of $r_{\hat{\sigma}^2}$ with respect to ρ and d . To demonstrate the degree of dimension reduction by PCA, Figure 2 shows the number of top components selected under different k -specification strategies. Although PCA-ReR uses a smaller number of principal components, it can generally result in more balanced covariates than ReR and the improvement is substantial in the cases with large ρ and d . The main reason is that a smaller k is typically chosen in such cases and that would lead to a much smaller variance shrinkage coefficient v_{a_k} for PCA-ReR as revealed by Figure 1. Compared with Ridge-ReR, which takes all principal components into account and puts more emphasis on the

TABLE 4: ANOVA results of known or estimable factors for the number of randomizations with respect to each rerandomization scheme.

Factors	DF	MS	F-ratio
d	3	73,375,319	51,133
Scheme	5	70,703,755	49,271
n	3	61,289,811	42,711
$n \times d \times \text{scheme}$	45	58,552,564	40,803
$n \times \text{scheme}$	15	38,165,186	26,596
$d \times \text{scheme}$	15	35,322,605	24,615
$n \times d$	9	32,583,133	22,706
ρ	2	1,370,398	955
$d \times \rho$	6	1,215,879	847
$n \times \rho$	6	1,185,041	826
$\rho \times \text{scheme}$	10	1,175,609	819
$d \times \rho \times \text{scheme}$	30	1,129,375	787
$n \times \rho \times \text{scheme}$	30	1,118,027	779
$n \times d \times \rho$	18	1,114,145	776
$n \times d \times \rho \times \text{scheme}$	90	1,077,482	751
Residuals	2592	1435	

Note: DF is the degrees of freedom, MS represents the between-configuration mean square, and the last row “Residuals” refers to the within-configuration mean square.

top ones, PCA-ReR only uses the most variable subspace to perform rerandomization, and its empirical performance can be superior to Ridge-ReR when the top few components explain most of the variations, as shown by the cases with $\rho = 0.9$.

The lower two panels of Table 5 present r_{MSE} for each rerandomization scheme under different configurations of the influential factors $g(\mathbf{x}, \boldsymbol{\beta})$, ρ , and d . PCA-ReR yields more precise estimation of τ than ReR in all situations. Similar to r_{σ^2} , PCA-ReR demonstrates evident advantages over Ridge-ReR for a large ρ , and the superiority can also be preserved in other combinations of (ρ, d) with properly chosen principal components ($\gamma_k = 0.5$). In some cases, such as $(\rho, d) = (0.1, 10)$, PCA-ReR can outperform Ridge-ReR even with less balanced covariates, because the estimation precision of $\hat{\tau}$ under PCA-ReR and Ridge-ReR also involves the coefficient $\boldsymbol{\beta}$ as revealed by Theorem 4 and Theorem 4.3 in Branson & Shao (2021). All methods yield smaller r_{MSE} for the nonlinear surface $\exp(\mathbf{x})^\top \boldsymbol{\beta}$ in comparison with the linear surface $\mathbf{x}^\top \boldsymbol{\beta}$, because these approaches only balance the first moments of covariates. Moreover, all three rerandomization schemes perform better for small d given the same ρ and $g(\mathbf{x}, \boldsymbol{\beta})$, because it is more difficult to simultaneously balance a larger number of covariates.

Table 6 summarizes the number of randomizations for three rerandomization schemes, which shows that PCA-ReR dominates the efficiency in most cases. Using a computer with Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, it takes around 1–75 s for implementing Ridge-ReR, particularly for a large d , whereas PCA-ReR takes at most 0.2 s under all situations. The reason

TABLE 5: The evaluation metrics $r_{\bar{\sigma}^2} \times 100$ and $r_{\text{MSE}} \times 100$ for three rerandomization (ReR) schemes under different combinations of ρ and d , where $r_{\bar{\sigma}^2}$ and r_{MSE} denote the reduction percents of the average empirical variance of $\bar{X}_{T,j} - \bar{X}_{C,j}$ across all covariates ($j = 1, \dots, d$) and mean squared error (MSE) of $\hat{\tau}$ relative to complete randomization, respectively. The number of top principal components for PCA rerandomization is based on the variation explained criterion (7) with the parameter $\gamma_k \in (0, 1]$ and Kaiser's rule.

Scheme	$d = 10$			$d = 50$			$d = 90$			$d = 180$		
	$\rho = 0.1$	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
$r_{\bar{\sigma}^2} \times 100$												
ReR	69.1	69.2	69.4	36.1	35.9	34.8	26.6	27.6	27.3	14.0	14.1	14.0
Ridge-ReR	69.3	75.4	89.8	40.4	50.8	78.2	33.1	45.4	74.3	27.1	40.0	71.6
PCA-ReR $_{\gamma_k=0.5}$	48.1	55.0	90.8	28.8	50.9	90.0	23.5	49.9	89.9	18.1	49.0	89.8
PCA-ReR $_{\gamma_k=0.7}$	58.5	64.6	90.8	33.2	43.4	90.0	26.6	36.3	89.9	20.0	30.0	89.8
PCA-ReR $_{\gamma_k=0.9}$	66.7	68.7	90.7	36.1	39.6	88.1	28.2	31.3	86.5	21.0	24.3	84.4
PCA-ReR $_{\text{Kaiser}}$	40.0	54.8	90.8	30.8	50.3	90.0	25.6	45.1	89.9	19.9	37.3	89.8
$r_{\text{MSE}} \times 100$ for linear response surface												
ReR	67.5	68.8	69.1	37.1	35.7	34.7	26.8	28.4	27.6	14.4	14.3	14.2
Ridge-ReR	69.3	81.6	92.0	58.3	63.6	82.9	51.7	59.9	79.2	47.4	54.7	76.9
PCA-ReR $_{\gamma_k=0.5}$	82.8	98.4	99.2	59.1	98.7	99.8	51.0	97.9	99.9	41.9	96.5	99.9
PCA-ReR $_{\gamma_k=0.7}$	75.7	88.3	99.2	48.7	62.7	99.8	42.5	53.7	99.9	32.7	45.2	99.9
PCA-ReR $_{\gamma_k=0.9}$	69.8	74.4	98.9	40.0	44.9	97.5	33.9	36.0	95.8	24.3	28.7	93.6
PCA-ReR $_{\text{Kaiser}}$	87.4	98.4	99.2	54.4	92.9	99.8	42.1	80.7	99.9	32.1	64.6	99.9
$r_{\text{MSE}} \times 100$ for nonlinear response surface												
ReR	49.9	52.3	41.7	31.5	27.6	21.9	24.0	21.2	18.3	13.3	11.1	8.8
RidgeReR	49.1	60.3	56.9	48.9	50.6	52.3	45.8	45.8	49.8	44.2	44.0	49.2
PCA-ReR $_{\gamma_k=0.5}$	60.7	72.2	62.0	50.5	76.0	62.2	46.2	75.3	62.3	38.9	74.9	62.3
PCA-ReR $_{\gamma_k=0.7}$	57.0	66.3	62.0	41.6	48.1	62.2	37.9	42.5	62.3	30.5	34.8	62.3
PCA-ReR $_{\gamma_k=0.9}$	51.7	56.0	62.1	33.7	34.4	61.4	29.2	28.9	59.4	22.4	22.2	58.8
PCA-ReR $_{\text{Kaiser}}$	63.9	72.2	61.9	45.8	70.6	62.2	38.3	62.7	62.3	29.9	49.7	62.3

for the long computation time of Ridge-ReR is due to the complexity of specifying its optimal values of λ and α_λ as well as a considerable number of randomizations in generating a feasible allocation. When d is close to n , such as $(n, d) = (100, 90)$ and $(200, 180)$, ReR takes a much longer time (18 and 91 s) than for cases with $n \gg d$ (at most 0.7 s), to generate an acceptable allocation. This is due to the long rejection period for a given threshold a , because it is more difficult to balance a large number of covariates with limited sample size. Furthermore, a is theoretically determined under a multivariate normal approximation, and the normality may not be satisfied when d is close to n , so the given threshold level a may be too restrictive for ReR. This result reflects the practical advantage of using PCA to accelerate the search for a feasible allocation.

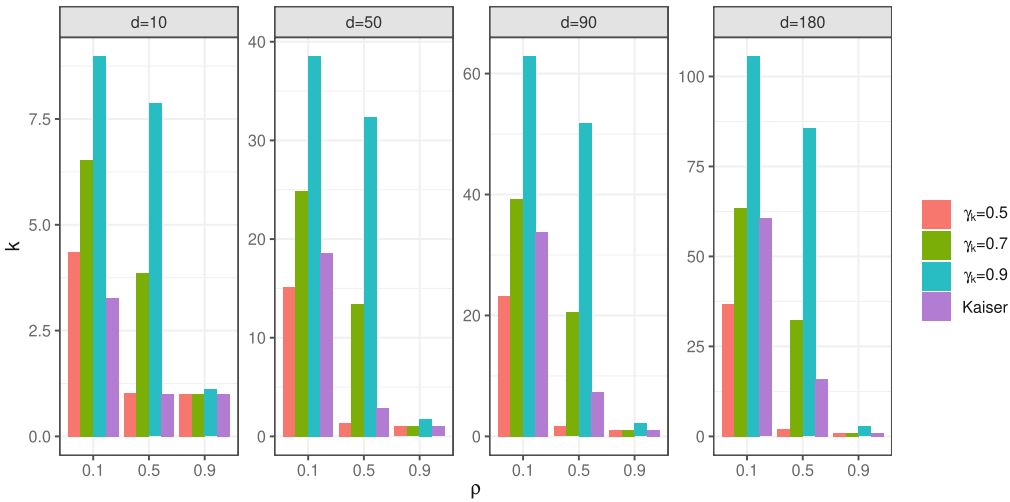


FIGURE 2: The number of selected top principal components for PCA rerandomization with respect to the variation explained criterion ($\gamma_k \in \{0.5, 0.7, 0.9\}$) defined in Equation (7), and Kaiser's rule, under different dimensions $d = 10, 50, 90, 180$ and correlations $\rho = 0.1, 0.5, 0.9$.

In Table 6, ReR displays unusual results when $d > n$, that is, $(n, d) = (100, 180)$. In such a case, Mahalanobis distance of ReR reduces to a constant,

$$M = \frac{n-1}{n} (2\mathbf{W} - \mathbf{1}_n)^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top (2\mathbf{W} - \mathbf{1}_n) = n - 1,$$

where $(\mathbf{X}^\top \mathbf{X})^-$ represents the pseudo-inverse of $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{U} \in \mathbb{R}^{n \times n}$ is composed of the left singular vectors of $\mathbf{X}$ with $\mathbf{X} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top = \mathbf{U} \mathbf{U}^\top = \mathbf{I}_n$. We further find that the threshold a of ReR is larger than $n - 1$ for $(n, d) = (100, 180)$, so that ReR reduces to CR in such situations, and thus the number of randomizations is reduced to 1.

Tables 5 and 6 finally demonstrate the differences between the variation explained criterion and Kaiser's rule for selecting top components in PCA-ReR. In general, evident improvements can be observed for PCA-ReR in terms of r_{MSE} and r_{σ^2} by adopting $\gamma_k = 0.5$ or Kaiser's rule, where both strategies tend to select a smaller number of components as revealed in Figure 2. For the variation explained criterion, a larger value of γ_k leads to closer performances between PCA-ReR and ReR in terms of the estimation precision and covariate balance, due to the increasing number of k as shown in Figure 2. Furthermore, it takes a larger number of randomizations to find a feasible allocation with a large γ_k , because more components should be considered to be balanced.

4.3. Guideline for Rerandomization

Based on the simulation results, we provide a guideline for selecting a proper scheme under different settings of the factors. Tables 2–4 show that d and ρ are jointly the most influential factors in terms of the F -ratios. When both d and ρ are small, all three rerandomization schemes show similar performance in terms of r_{σ^2} and r_{MSE} . When either d or ρ is large, PCA-ReR and Ridge-ReR should be considered because they leverage the importance of principal components to improve covariate balance as well as estimation precision. Note that β and $g(\mathbf{x}, \beta)$ are unknown to the investigator during the design stage, and Table 5 implies that both methods could yield better performance than ReR, but PCA-ReR can be superior to Ridge-ReR in many cases. We hence recommend using PCA-ReR not only due to its empirical desirability, but also because

TABLE 6: The number of randomizations required for generating a feasible allocation with respect to three rerandomization (ReR) schemes under different combinations of n and d . The number of top principal components for PCA rerandomization is determined based on the variation explained criterion (7) with the parameter $\gamma_k \in (0, 1]$ and Kaiser's rule.

Scheme	$n = 100$				$n = 200$			
	$d = 10$	50	90	180	10	50	90	180
ReR	23	72	6178	1	21	33	63	8109
Ridge-ReR	23	59	200	7873	21	31	50	192
PCA-ReR $_{\gamma_k=0.5}$	20	22	23	24	20	21	21	22
PCA-ReR $_{\gamma_k=0.7}$	21	24	26	31	21	22	23	26
PCA-ReR $_{\gamma_k=0.9}$	22	31	43	82	21	25	30	47
PCA-ReR $_{\text{Kaiser}}$	20	23	29	31	20	21	22	29

Scheme	$n = 500$				$n = 1000$			
	$d = 10$	50	90	180	10	50	90	180
ReR	21	24	27	45	20	22	23	28
Ridge-ReR	21	23	27	40	20	22	26	26
PCA-ReR $_{\gamma_k=0.5}$	20	20	20	20	20	20	20	21
PCA-ReR $_{\gamma_k=0.7}$	20	21	21	23	20	20	21	22
PCA-ReR $_{\gamma_k=0.9}$	20	22	23	27	20	20	21	23
PCA-ReR $_{\text{Kaiser}}$	20	20	20	22	20	20	20	21

of its fast implementation as well as ease of dealing with high-dimensionality. The sufficient dimension reduction of PCA allows the investigator to try expanding the dimension of the covariate matrix by incorporating different combinations of nonlinear features (e.g., second moments or interactions of covariates), which may further boost the performance by using nonlinear response surfaces. As for the number of components in PCA-ReR, we recommend Kaiser's rule or simply setting $\gamma_k = 0.5$ due to their objectivity and empirical desirability. One can further leverage the scree plot as the reference to make the final decision between the two strategies.

4.4. Real Application

For illustration, we compare different rerandomization methods using a real dataset from the Infant Health and Development Program (IHDP), which is provided in Hill (2011). This programme aimed at improving cognitive development for low-birth-weight and premature infants by providing high-quality child care and home visits from trained specialists for the infants in the treatment group. The conclusion is that this intervention successfully promoted cognitive test scores for the treated children compared with the control group. After removing the missing values, the dataset consisted of 908 participants with continuous test scores and the 25 covariates selected by Hill (2011), among which six covariates are continuous and the others are binary. To mimic the real study process, we used all observed data to fit a sparse linear model via LASSO using 25 main effect covariates and all their $\binom{25}{2} = 300$ two-way interactions, which

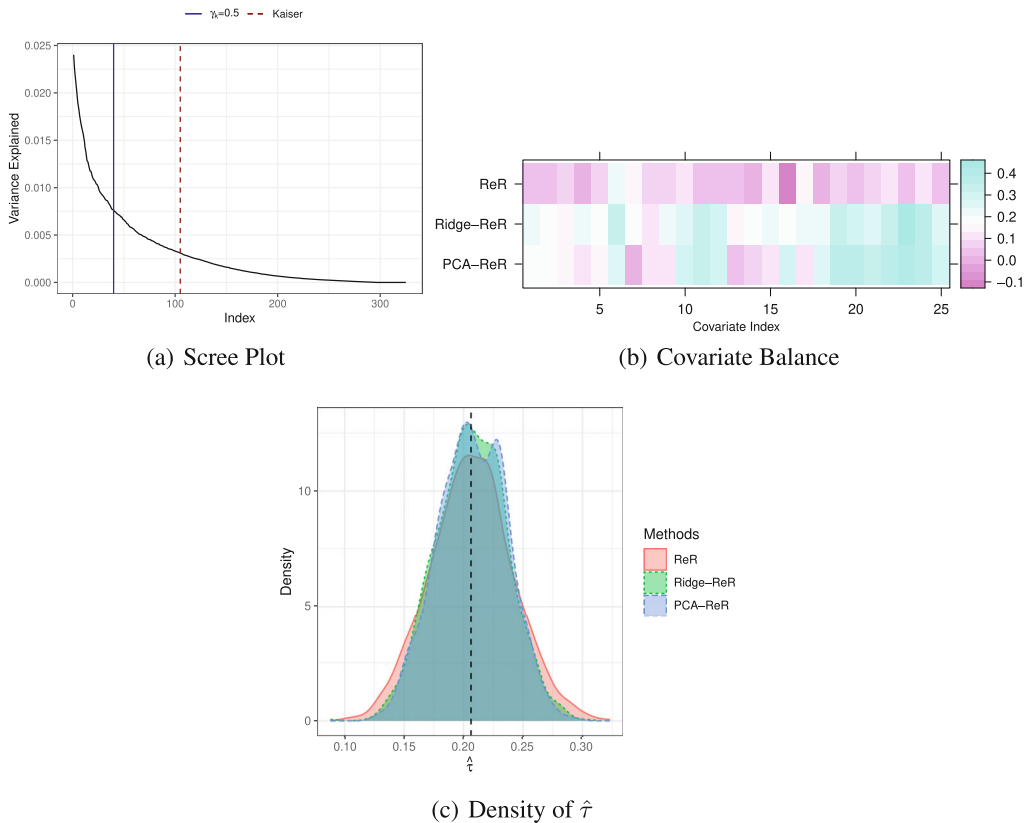


FIGURE 3: The scree plot and the results of different rerandomization approaches on the IHDP dataset. (a) The scree plot versus the number of components k chosen by setting $\gamma_k = 0.5$ in (7) and Kaiser's rule respectively; (b) the reduction percent of empirical variance for $\bar{x}_{T,j} - \bar{x}_{C,j}$ ($j = 1, \dots, 25$) relative to complete randomization; (c) density plots of the estimated treatment effect, where the black vertical line denotes the estimated value in the linear model.

were standardized before the model fitting. The tuning parameter of the LASSO penalty was chosen to be 0.03 from cross-validation, and 82 covariates were finally selected. The fitted model with the estimated treatment effect 0.21 was used to generate responses given the allocation during rerandomization. We constructed the standardized covariate matrix $\mathbf{X} \in \mathbb{R}^{908 \times 325}$ by all instances, and assigned an equal number of units to the treatment and control groups. Different rerandomization approaches were then applied to generate 1000 independent allocations, from which we calculated the evaluation metrics. For PCA-ReR, we first calculated the number of components based on $\gamma_k = 0.5$ and Kaiser's rule, which led to 40 and 105, respectively. The scree plot shown in Figure 3a helped us to finalize $k = 40$, because it was close to the elbow of the scree curve.

Figure 3b shows the reduction percent of the empirical variance for the mean difference of $\bar{x}_{T,j} - \bar{x}_{C,j}$ ($j \in \{1, \dots, 25\}$) relative to CR. We observe that PCA-ReR using 40 principal components achieves more balanced covariates than ReR and its performance is comparable to Ridge-ReR using all 325 components. We further present the density plot of the treatment effect estimator $\hat{\tau}$ based on 1000 allocations for ReR, Ridge-ReR, and PCA-ReR in Figure 3c, where the distribution of $\hat{\tau}$ under PCA-ReR is more concentrated than that under ReR and demonstrates a comparable concentration to Ridge-ReR. We further observe that the overall reduction percentages r_{σ^2} are 0.06, 0.23, and 0.18 for ReR, Ridge-ReR, and PCA-ReR, respectively, and

the corresponding reduction percentages r_{MSE} are 0.09, 0.32, and 0.37. Therefore, PCA-ReR outperforms both Ridge-ReR and ReR for the estimation of treatment effect. In terms of the computational time for generating 1000 feasible allocations, PCA-ReR (27 s) runs 46 times faster than Ridge-ReR (1229 s) and 4 times faster than ReR (116 s). Moreover, PCA-ReR needs 23 randomizations on average to generate a feasible allocation, whereas 3 and 38 randomizations are required for ReR and Ridge-ReR, respectively. Thus, ReR takes a longer time to generate a single randomization due to the calculation of Mahalanobis distance, but the threshold α for ReR is relatively large, which leads to a small number of randomizations.

5. DISCUSSION

We propose a PCA rerandomization scheme to allocate participants in randomized experiments where there are high-dimensional or high-collinearity covariates. Compared with the original rerandomization (Morgan & Rubin, 2012), we use the top principal components rather than the original covariates to calculate Mahalanobis distance. Not only does PCA rerandomization share most theoretical characteristics with the original rerandomization, but it also demonstrates empirical advantages with high-dimensional or highly correlated covariates in terms of covariate balance, the precision of treatment effect estimation as well as the number of randomizations and computational time. Future work is warranted on the extension to 2^K factorial designs, following Branson, Dasgupta & Rubin (2016), as well as developing the asymptotic property of the treatment effect estimator under PCA rerandomization following Li, Ding & Rubin (2018). The computer code and data are accessible from <https://github.com/BobZhangHT/PCARer>.

ACKNOWLEDGEMENTS

We thank the Associate Editor and the referee for their careful reviews of our manuscript and many insightful comments that have significantly improved the paper. This research was partially supported by funding from the Research Grants Council of Hong Kong (17308321).

REFERENCES

- Baxter, M. J. (1995). Standardization and transformation in principal component analysis, with applications to archaeometry. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4), 513–527.
- Branson, Z., Dasgupta, T., & Rubin, D. B. (2016). Improving covariate balance in 2^K 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.
- Branson, Z. & Shao, S. (2021). Ridge rerandomization: An experimental design strategy in the presence of covariate collinearity. *Journal of Statistical Planning and Inference*, 211, 287–314.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
- Cox, D. (2009). Randomization in the design of experiments. *International Statistical Review*, 77(3), 415–429.
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.
- Gutman, R. & Rubin, D. (2017). Estimation of causal effects of binary treatments in unconfounded studies with one continuous covariate. *Statistical Methods in Medical Research*, 26(3), 1199–1215.
- Gutman, R. & Rubin, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Statistics in Medicine*, 32(11), 1795–1814.
- Gutman, R. & Rubin, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in Medicine*, 34(26), 3381–3398.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240.
- Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, Cambridge, UK.

- Johansson, P. & Schultzberg, M. (2020). Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. *Journal of Computational and Graphical Statistics*, 29(4), 798–813.
- Jolliffe, I. T. & Morgan, B. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, 1(1), 69–95.
- Jolliffe, I. T. (2002). *Principal Component Analysis*, 2nd ed., Springer, New York, USA.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–151.
- Li, X. & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520), 1759–1769.
- Li, X. & Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1), 241–268.
- Li, X., Ding, P., & Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37), 9157–9162.
- Li, X., Ding, P., & Rubin, D. B. (2020). Rerandomization in 2^K factorial experiments. *The Annals of Statistics*, 48(1), 43–63.
- 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.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. *Statistical Science*, 5, 465–471 (*Roczniki Nauk Rolniczych Tom X* (in Polish)).
- Rosenbaum, P. R. (2010). *Design of Observational Studies*, 1st ed., Springer, New York, USA.
- 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.
- Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, 29(1), 159–183.
- Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1), 185–203.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.
- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366), 318–328.
- Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics*, 36(2), 293–298.
- Singh, A. & Harrison, A. (1985). Standardized principal components. *International Journal of Remote Sensing*, 6(6), 883–896.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21.
- Wang, X., Wang, T., & Liu, H. (2021). Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.2021.1990767>
- Worrall, J. (2010). Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice*, 16(2), 356–362.
- Yang, Z., Qu, T., & Li, X. (2021). Rejective sampling, rerandomization, and regression adjustment in survey experiments. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.2021.1984926>
- Zhang, H. & Yin, G. (2021). Response-adaptive rerandomization. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(5), 1281–1298.
- Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., & Zhang, A. (2018). Sequential rerandomization. *Biometrika*, 105(3), 745–752.

APPENDIX

Proof of Theorem 2. From (2), we have

$$\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C = \frac{2}{n} \mathbf{X}^\top (2\mathbf{W} - \mathbf{1}_n) = \frac{2}{n} \mathbf{V} \mathbf{Z}^\top (2\mathbf{W} - \mathbf{1}_n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Furthermore, it is easy to see that

$$\mathbf{Z} = (\mathbf{Z}_k, \tilde{\mathbf{Z}}_{d-k}) \quad \text{and} \quad \mathbf{V}^\top \boldsymbol{\Sigma} \mathbf{V} = C_n \mathbf{D}^2 = C_n \begin{pmatrix} \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix}, \quad (\text{A1})$$

where $C_n = 4/(n^2 - n)$. Consequently,

$$\bar{\mathbf{z}}_T^{(k)} - \bar{\mathbf{z}}_C^{(k)} = \frac{2}{n} \mathbf{Z}_k^\top (2\mathbf{W} - \mathbf{1}_n) = \frac{2}{n} (\mathbf{I}_k, \mathbf{0}) \mathbf{V}^\top \mathbf{X}^\top (2\mathbf{W} - \mathbf{1}_n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_z), \quad (\text{A2})$$

where $\boldsymbol{\Sigma}_z = \mathbf{Z}_k^\top \mathbf{Z}_k = C_n \mathbf{D}_k^2$. This concludes that $M_k | \mathbf{X} \sim \chi_k^2$ according to the property of a multivariate normal distribution. ■

Proof of Theorem 3. For simplicity, let $\boldsymbol{\eta} = \mathbf{V}^\top (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C) = (\boldsymbol{\eta}_k, \tilde{\boldsymbol{\eta}}_{d-k})^\top$ where $\boldsymbol{\eta}_k = (\eta_1, \dots, \eta_k)^\top \in \mathbb{R}^k$ and $\tilde{\boldsymbol{\eta}}_{d-k} = (\eta_{k+1}, \dots, \eta_d)^\top \in \mathbb{R}^{d-k}$. According to (A1) and (A2), we know that $\boldsymbol{\eta} | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, C_n \mathbf{D}^2)$, $\boldsymbol{\eta}_k | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, C_n \mathbf{D}_k^2)$ and $\tilde{\boldsymbol{\eta}}_{d-k} | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, C_n \tilde{\mathbf{D}}_{d-k}^2)$. Thus, M_k can be written in terms of $\boldsymbol{\eta}_k$ and singular values $\sigma_1 \geq \dots \geq \sigma_k > 0$, i.e., $M_k = \eta_1^2 / (C_n \sigma_1^2) + \dots + \eta_k^2 / (C_n \sigma_k^2)$.

Note that we have

$$\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M_k \leq a_k) = \mathbf{V} \text{cov}(\boldsymbol{\eta} | \mathbf{X}, M_k \leq a_k) \mathbf{V}^\top. \quad (\text{A3})$$

Let $a \stackrel{d}{=} b$ denote two variables a and b having the same distribution. First, one can obtain that $\mathbb{E}(\eta_j | \mathbf{X}, M_k \leq a_k) = 0$ through the symmetry of a normal distribution, that is, $-\eta_j \stackrel{d}{=} \eta_j \sim \mathcal{N}(0, C_n \sigma_j^2)$, for all $j \in \{1, \dots, k\}$ (Branson & Shao, 2021). Specifically,

$$\begin{aligned} \mathbb{E}(\eta_j | \mathbf{X}, M_k \leq a_k) &= \mathbb{E} \left(\eta_j \middle| \mathbf{X}, \sum_{j=1}^k \frac{\eta_j^2}{C_n \sigma_j^2} \leq a_k \right) \\ &= \mathbb{E} \left(-\eta_j \middle| \mathbf{X}, \sum_{j=1}^k \frac{(-\eta_j)^2}{C_n \sigma_j^2} \leq a_k \right) \\ &= -\mathbb{E}(\eta_j | \mathbf{X}, M_k \leq a_k), \end{aligned} \quad (\text{A4})$$

which leads to $\mathbb{E}(\eta_j | \mathbf{X}, M_k \leq a_k) = 0$. Now, we focus on the covariance, $\text{cov}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k)$ with $i \neq j$ and $i, j \in \{1, \dots, k\}$. Similarly to (A4), we only need to flip the sign of η_j to show $\mathbb{E}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) = 0$ and thus further derive $\text{cov}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) = \mathbb{E}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) = 0$. On the other hand, $\eta_j^2 / (C_n \sigma_j^2)$ are independent and identically follow χ_1^2 for $j \in \{1, \dots, k\}$. Combining the above results, the variance of η_j for $j \in \{1, \dots, k\}$ is

$$\begin{aligned} \text{var}(\eta_j | \mathbf{X}, M_k \leq a_k) &= \mathbb{E}(\eta_j^2 | \mathbf{X}, M_k \leq a_k) \\ &= C_n \sigma_j^2 \mathbb{E} \left(\frac{\eta_j^2}{C_n \sigma_j^2} \middle| \mathbf{X}, \sum_{j=1}^k \frac{\eta_j^2}{C_n \sigma_j^2} \leq a_k \right) \\ &= \frac{C_n \sigma_j^2}{k} \mathbb{E}(M_k | \mathbf{X}, M_k \leq a_k), \end{aligned}$$

where the last equation follows the exchangeability of i.i.d. $\{\eta_j^2 / (C_n \sigma_j^2)\}_{j=1}^k$. Because $M_k | \mathbf{X} \sim \chi_k^2$, we have $\mathbb{E}(M_k | \mathbf{X}, M_k \leq a_k) / k = v_{a_k} = \Pr(\chi_{k+2}^2 \leq a_k) / \Pr(\chi_k^2 \leq a_k)$ following the proof of Theorem 3.1 in Morgan & Rubin (2012). Therefore, the variance is $\text{var}(\eta_j | \mathbf{X}, M_k \leq a_k) = C_n v_{a_k} \sigma_j^2$.

It is evident that for $j = k + 1, \dots, d$, we have

$$\mathbb{E}(\eta_j | \mathbf{X}, M_k \leq a_k) = \mathbb{E}(\eta_j | \mathbf{X}) = 0,$$

$$\text{var}(\eta_j | \mathbf{X}, M_k \leq a_k) = \text{var}(\eta_j | \mathbf{X}) = C_n \sigma_j^2,$$

$$\text{cov}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) = \text{cov}(\eta_i \eta_j | \mathbf{X}) = 0, \quad \forall i \neq j, \quad i \in \{k + 1, \dots, d\}.$$

Finally, we show $\text{cov}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) = 0$ when $i \in \{1, \dots, k\}$ and $j \in \{k + 1, \dots, d\}$ as follows,

$$\begin{aligned} \text{cov}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) &= \mathbb{E}(\eta_i \eta_j | \mathbf{X}, M_k \leq a_k) \\ &= \mathbb{E} \{ \eta_i \mathbb{E}(\eta_j | \eta_i, \mathbf{X}, M_k \leq a_k) | \mathbf{X}, M_k \leq a_k \} \\ &= \mathbb{E} \{ \eta_i \mathbb{E}(\eta_j | \mathbf{X}) | \mathbf{X}, M_k \leq a_k \} \\ &= 0. \end{aligned}$$

Considering all the aforementioned results of the variance and covariance, we have

$$\text{cov}(\boldsymbol{\eta} | \mathbf{X}, M_k \leq a_k) = C_n \begin{pmatrix} v_{a_k} \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix}.$$

The theorem can be proved after plugging the above equation into (A3). ■

Proof of Corollary 1. According to the definition of PRV, we can obtain the expression of $v_{a_{k,j}}$ as the ratio of the j th diagonal entry of the covariance matrix for PCA-ReR and $\boldsymbol{\Sigma}$; that is,

$$v_{a_{k,j}} = \frac{\left(C_n \mathbf{V} \text{diag} \left\{ v_{a_k} \mathbf{D}_k^2, \tilde{\mathbf{D}}_{d-k}^2 \right\} \mathbf{V}^\top \right)_{jj}}{\boldsymbol{\Sigma}_{jj}}.$$

Here, we show that $v_{a_{k,j}} \in (0, 1)$. Define $\mathbf{V} = (\mathbf{V}_k, \tilde{\mathbf{V}}_{d-k})$ where $\mathbf{V}_k \in \mathbb{R}^{d \times k}$ and $\tilde{\mathbf{V}}_{d-k} \in \mathbb{R}^{d \times (d-k)}$ correspond to the first k and last $d - k$ components. Let $\boldsymbol{\alpha}_j = (\mathbf{0}_{j-1}^\top, 1, \mathbf{0}_{d-j}^\top)^\top$ be a d -dimensional vector of 0s except for the j th element taking a value of 1. We have that

$$\begin{aligned} \boldsymbol{\Sigma}_{jj} &= \boldsymbol{\alpha}_j^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_j \\ &= C_n \boldsymbol{\alpha}_j^\top \mathbf{V} \begin{pmatrix} \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix} \mathbf{V}^\top \boldsymbol{\alpha}_j \\ &= C_n \boldsymbol{\alpha}_j^\top \mathbf{V}_k \mathbf{D}_k^2 \mathbf{V}_k^\top \boldsymbol{\alpha}_j + C_n \boldsymbol{\alpha}_j^\top \tilde{\mathbf{V}}_{d-k} \tilde{\mathbf{D}}_{d-k}^2 \tilde{\mathbf{V}}_{d-k}^\top \boldsymbol{\alpha}_j > 0, \end{aligned}$$

and

$$\left(C_n \mathbf{V} \text{diag} \left\{ v_{a_k} \mathbf{D}_k^2, \tilde{\mathbf{D}}_{d-k}^2 \right\} \mathbf{V}^\top \right)_{jj} = v_a C_n \boldsymbol{\alpha}_j^\top \mathbf{V}_k \mathbf{D}_k^2 \mathbf{V}_k^\top \boldsymbol{\alpha}_j + C_n \boldsymbol{\alpha}_j^\top \tilde{\mathbf{V}}_{d-k} \tilde{\mathbf{D}}_{d-k}^2 \tilde{\mathbf{V}}_{d-k}^\top \boldsymbol{\alpha}_j$$

since $v_{a_k} = \Pr(\chi_{k+2}^2 \leq a_k) / \Pr(\chi_k^2 \leq a_k) \in (0, 1)$ for any $a_k > 0$. We can conclude that

$$0 < \left(C_n \mathbf{V} \text{diag} \left\{ v_{a_k} \mathbf{D}_k^2, \tilde{\mathbf{D}}_{d-k}^2 \right\} \mathbf{V}^\top \right)_{jj} < \Sigma_{jj},$$

and thus $v_{a_{k,j}} \in (0, 1)$ because both $\mathbf{V}_k \mathbf{D}_k^2 \mathbf{V}_k^\top$ and $\tilde{\mathbf{V}}_{d-k} \tilde{\mathbf{D}}_{d-k}^2 \tilde{\mathbf{V}}_{d-k}^\top$ are positive definite. ■

Proof of Theorem 4. Define $\bar{\epsilon}_T = 2\mathbf{W}^\top \epsilon / n$ and $\bar{\epsilon}_C = 2(\mathbf{1} - \mathbf{W})^\top \epsilon / n$, where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$. According to (9), $\hat{\tau}$ can be written as

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

Leveraging the orthogonality between the first and last terms, we have that

$$\begin{aligned} \text{var}(\hat{\tau}|\mathbf{X}) &= \text{var} \left\{ (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^\top \boldsymbol{\beta} | \mathbf{X} \right\} + \text{var}(\bar{\epsilon}_T - \bar{\epsilon}_C | \mathbf{X}) \\ &= \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} + \text{var}(\bar{\epsilon}_T - \bar{\epsilon}_C | \mathbf{X}) \\ &= C_n \boldsymbol{\beta}^\top \mathbf{V} \begin{pmatrix} \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix} \mathbf{V}^\top \boldsymbol{\beta} + \text{var}(\bar{\epsilon}_T - \bar{\epsilon}_C | \mathbf{X}). \end{aligned} \quad (\text{A5})$$

Furthermore, the conditional normality on $\hat{\tau}$ and $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ leads to the conditional independence between $\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C$ and $\bar{\epsilon}_T - \bar{\epsilon}_C$, because these two terms are uncorrelated. Therefore, M_k is also conditionally independent of $\bar{\epsilon}_T - \bar{\epsilon}_C$, and we have that

$$\begin{aligned} \text{var}(\hat{\tau}|\mathbf{X}, M_k \leq a_k) &= \text{var} \left\{ (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^\top \boldsymbol{\beta} | \mathbf{X}, M_k \leq a_k \right\} + \text{var}(\bar{\epsilon}_T - \bar{\epsilon}_C | \mathbf{X}, M_k \leq a_k) \\ &= \boldsymbol{\beta}^\top \text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C | \mathbf{X}, M_k \leq a_k) \boldsymbol{\beta} + \text{var}(\bar{\epsilon}_T - \bar{\epsilon}_C | \mathbf{X}) \\ &= C_n \boldsymbol{\beta}^\top \mathbf{V} \begin{pmatrix} v_{a_k} \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_{d-k}^2 \end{pmatrix} \mathbf{V}^\top \boldsymbol{\beta} + \text{var}(\bar{\epsilon}_T - \bar{\epsilon}_C | \mathbf{X}). \end{aligned} \quad (\text{A6})$$

Combining (A5) and (A6), it can be shown that

$$\text{var}(\hat{\tau}|\mathbf{X}) - \text{var}(\hat{\tau}|\mathbf{X}, M_k \leq a_k) = C_n \boldsymbol{\beta}^\top \mathbf{V} \begin{pmatrix} (1 - v_{a_k}) \mathbf{D}_k^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V}^\top \boldsymbol{\beta} \geq 0,$$

where the non-negativity arises from the positive semidefiniteness of the matrix. It is easy to obtain $\text{var}(\hat{\tau}|\mathbf{X}) = \text{var}(\hat{\tau}|\mathbf{X}, M_k \leq a_k)$ when the first k elements of $\mathbf{V}^\top \boldsymbol{\beta}$ are zeros. ■

Received 27 December 2021

Accepted 9 November 2022