Rerandomization in Stratified Randomized Experiments

Xinhe Wang\textsuperscript{a}, Tingyu Wang\textsuperscript{b}, and Hanzhong Liu\textsuperscript{c}

\textsuperscript{a}Department of Mathematical Sciences, Tsinghua University, Beijing, China; \textsuperscript{b}Department of Physics, Tsinghua University, Beijing, China; \textsuperscript{c}Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

ABSTRACT
Stratification and rerandomization are two well-known methods used in randomized experiments for balancing the baseline covariates. Renowned scholars in experimental design have recommended combining these two methods; however, limited studies have addressed the statistical properties of this combination. This article proposes two rerandomization methods to be used in stratified randomized experiments, based on the overall and stratum-specific Mahalanobis distances. The first method is applicable for nearly arbitrary numbers of strata, strata sizes, and stratum-specific proportions of the treated units. The second method, which is generally more efficient than the first method, is suitable for situations in which the number of strata is fixed with their sizes tending to infinity. Under the randomization inference framework, we obtain the asymptotic distributions of estimators used in these methods and the formulas of variance reduction when compared to stratified randomization. Our analysis does not require any modeling assumption regarding the potential outcomes. Moreover, we provide asymptotically conservative variance estimators and confidence intervals for the average treatment effect. The advantages of the proposed methods are exhibited through an extensive simulation study and a real-data example.

ARTICLE HISTORY
Received November 2020
Accepted October 2021

KEYWORDS
Blocking; Causal inference; Covariate adjustment; Randomization inference; Stratification

1. Introduction

The application of randomized experiments has recently gained increasing popularity in various fields, including industry, social sciences, and clinical trials (e.g., Box, Hunter, and Hunter 2005; Gerber and Green 2012; Rosenberger and Lachin 2015). Often, there are covariates that are likely to be unbalanced in completely randomized experiments (Fisher 1926; Senn 1989; Morgan and Rubin 2012). Fisher (1926) first recognized this issue and introduced the use of blocking, or stratification, for balancing discrete covariates. In stratified randomized experiments, units are divided into strata according to the discrete covariates and complete randomization is conducted within each stratum. Appropriate stratification improves the covariate balance and inference efficiency; see Imai (2008), Miratrix, Sekhon, and Yu (2013), and Imbens and Rubin (2015) for an overview.

Whereas stratification balances only discrete covariates, rerandomization is a more powerful tool that excludes allocations causing covariate imbalance. Covariate balance can be measured by a predetermined criterion, and only the allocations that meet this criterion are accepted (Morgan and Rubin 2012). Morgan and Rubin (2012) used the Mahalanobis distance of the sample means of the covariates in the treatment and control groups for measuring covariate balance and set a threshold in advance to rule out unsatisfactory allocations. The authors showed that the difference-in-means average treatment effect estimator remains unbiased under the symmetric balance criterion for the treatment and control groups, and that rerandomization enhances efficiency when the treatment effect is additive (i.e., all the units have the same treatment effect) and the covariates are correlated with the potential outcomes. For more general situations, Li, Ding, and Rubin (2018) obtained an asymptotic distribution of the difference-in-means estimator under rerandomization, and developed a method to construct large-sample confidence intervals for the average treatment effect.

Renowned scholars, such as R. A. Fisher, have recommended combining the rerandomization and stratification methods. This design strategy was summarized by D. B. Rubin as “Block what you can and rerandomize what you cannot.” Recently, Schultzberg and Johansson (2019) developed a stratified rerandomization design where stratification on binary covariates was followed by rerandomization on continuous covariates. They demonstrated that for binary covariates, stratification is equivalent to rerandomization, and that stratified rerandomization enhances both inference and computation efficiencies under equal-sized treatment and control groups and an additive treatment effect, or the Fisher sharp null hypothesis. However, when the sizes of the treatment and control groups are not equally sized, or the treatment effect is not additive, especially when the number of strata tends toward infinity, the efficient strategy of stratified rerandomization and its statistical behavior are unknown.
The present article proposes two rerandomization strategies in stratified randomized experiments and establishes their asymptotic theory by using the Neyman–Rubin potential outcomes model (Rubin 1974; Neyman, Dabrowska, and Speed 1990) and randomization inference framework (Kempthorne 1955; Li and Ding 2017; Zhao et al. 2018), without any modeling assumption regarding the potential outcomes. The proposed methods are termed the overall strategy and the stratum-specific strategy. Both use the Mahalanobis distance for measuring covariate imbalance. However, the first computes the overall covariate imbalance and rerandomizes over the entire strata together, and the second computes the stratum-specific covariate imbalance and rerandomizes within each stratum independently. The overall strategy is flexible and applicable to nearly arbitrary numbers of strata and their sizes, and does not require the same propensity scores (proportions of the treated units) across different strata. This strategy is essential for stratified experiments with strata containing single treated or control unit and hybrid experiments with both large and small strata. These scenarios can easily appear in modern social science experiments. For instance, multisite trials in education often have several sites, including only a few schools on each site (e.g., Wills et al. 2018). The stratum-specific strategy is a straightforward extension of Li, Ding, and Rubin (2018), and is generally more efficient than the first method; however, it requires the number of strata to be fixed with their sizes tending to infinity.

We prove that under mild conditions, the stratified difference-in-means estimators are asymptotically unbiased and truncated-normal distributed under both stratified rerandomization strategies. In addition, we show that stratified rerandomization improves, or at least does not degrade, the precision as compared to stratified randomization (SR). We further provide asymptotically conservative estimators for the variances and confidence intervals under both strategies. Finally, we illustrate the performances of the proposed methods through an extensive simulation study and a real-data example.

2. Framework, Notation, and Stratified Rerandomization

In stratified randomized experiments with \( n \) units, \( p_0 \) discrete covariates and \( p \) additional (discrete or continuous) covariates are collected before the physical implementation of randomization. The units are divided into \( K \) strata according to the \( p_0 \) discrete covariates, each having \( n_k \) \( (k = 1, \ldots, K) \) units, such that \( n = n_{11} + \cdots + n_{K} \). Let \( X \in \mathbb{R}^{n \times p} \) denote an additional covariate matrix, whose \( i \)th row, denoted by \( X_i \), indicates the observations of the additional covariates of unit \( i \). In stratum \( k \), \( n_{1k} = p_k n_k \) units are randomly selected and assigned to the treatment group, and the remaining units \( n_{0k} = (1 - p_k) n_k \) units are assigned to the control group, where \( p_k \in (0, 1) \) is called the propensity score. The total numbers of treated and control units are \( n_1 = \sum_{k=1}^{K} n_{1k} \) and \( n_0 = \sum_{k=1}^{K} n_{0k} \), respectively. For each unit \( i = 1, \ldots, n \), let \( Z_i \) be the treatment assignment indicator, where \( Z_i = 1 \) if it is assigned to the treatment group and \( Z_i = 0 \) if it is assigned to the control group. We use \( i \in [k] \) to denote the indices taken over the stratum \( k \). Let \( Y_i(z) \) be the potential outcomes for unit \( i \) under the treatment arm \( z \in \{0, 1\} \), where \( z = 1 \) indicates treatment and \( z = 0 \) indicates control. The unit level treatment effect is defined as \( \tau_i = Y_i(1) - Y_i(0) \), and the average treatment effect is defined as \( \tau = n^{-1} \sum_{k=1}^{K} \sum_{i \in [k]} \tau_i = \sum_{k=1}^{K} \pi_k \tau_k \), where \( \pi_k = n_k / n \) is the proportion of stratum size and \( \tau_k = n_{1k}^{-1} \sum_{i \in [k]} \tau_i \) is the stratum-specific average treatment effect in stratum \( k \) \( (k = 1, \ldots, K) \).

In stratum \( k \), the stratum-specific means of covariates and potential outcomes are denoted as \( \bar{X}_k \) and \( \bar{Y}_k(z) = n_{1k}^{-1} \sum_{i \in [k]} Y_i(z), z = 0, 1 \), and the stratum-specific variances and covariances are denoted as follows:

\[
S_{[k]Y}(z) = \frac{1}{n_k - 1} \sum_{i \in [k]} (Y_i(z) - \bar{Y}_k(z))^2,
\]

\[
S_{[k]XX} = \frac{1}{n_k - 1} \sum_{i \in [k]} (X_i - \bar{X}_k)(X_i - \bar{X}_k)^T,
\]

\[
S_{[k]XY}(z) = \frac{1}{n_k - 1} \sum_{i \in [k]} (X_i - \bar{X}_k)(Y_i(z) - \bar{Y}_k(z)),
\]

\[
S_{[k]r} = \frac{1}{n_k - 1} \sum_{i \in [k]} (\tau_i - \bar{r}_k)^2.
\]

Under the stable unit treatment value assumption (Rubin 1980), for any realized value of \( Z_i \), the observed outcome of unit \( i \) is \( Y_i^o = Z_i Y_i(1) + (1 - Z_i) Y_i(0) \). For the treatment arm \( z = 1 \), the observed stratum-specific means of the potential outcomes and covariates are denoted as \( \bar{Y}_k(1) = n_{1k}^{-1} \sum_{i \in [k]} Z_i Y_i(1) \) and \( \bar{X}_k(1) = n_{1k}^{-1} \sum_{i \in [k]} Z_i X_i \). Similarly, we define \( \bar{Y}_k(0) \) and \( \bar{X}_k(0) \) for the control arm \( z = 0 \). The stratified difference-in-means estimator of the average treatment effect is

\[
\tilde{\tau} = \sum_{k=1}^{K} \pi_k \left[ \bar{Y}_k^o - \bar{Y}_k^o \right] = \sum_{k=1}^{K} \pi_k \tilde{\tau}_k \tag{1}
\]

where \( \tilde{\tau}_k = \bar{Y}_k^o - \bar{Y}_k^o \) is the difference-in-means estimator of \( \tau_k \).

This article proposes two stratified rerandomization criteria, one based on the overall Mahalanobis distance and the other based on the stratum-specific Mahalanobis distance.

1. Stratified rerandomization based on the overall Mahalanobis distance. Because covariates can be viewed as potential outcomes that are unaffected by the treatment assignment with zero treatment effect, the Mahalanobis distance of the stratified sample means of the covariates under two treatment arms can be used to measure the covariate imbalance. More specifically, denote

\[
\tilde{\tau}_X = \sum_{k=1}^{K} \pi_k \left[ \bar{X}_k^o - \bar{X}_k^o \right] = \sum_{k=1}^{K} \pi_k \tilde{\tau}_k X_k.
\]

where \( \tilde{\tau}_k X_k = \bar{X}_k^o - \bar{X}_k^o \) indicates the difference-in-means of the covariates in stratum \( k \). The overall Mahalanobis distance is defined as \( M_{\tilde{\tau}_X} = \tilde{\tau}_X^T \text{cov}(\tilde{\tau}_X)^{-1} \tilde{\tau}_X \). Here, a random assignment is accepted only when \( M_{\tilde{\tau}_X} < a \), where \( a \) is a predetermined threshold.

2. Stratified rerandomization based on the stratum-specific Mahalanobis distance. When each stratum comprises a large
number of units, rerandomizing within each stratum separately and independently can be more efficient than the overall rerandomization. Thus, we use this rerandomization criterion in our study, where the stratum-specific Mahalanobis distance is defined as $M_{ik} = (\bar{t}_{ik|x})^T \text{cov}(\bar{t}_{ik|x})^{-1} \bar{t}_{ik|x}$, $k = 1, \ldots, K$. Here, a random assignment is accepted only when $M_{ik} < a_k$, where $a_k$ is a predetermined threshold for the stratum $k$.

To investigate the asymptotic properties of the above two stratified rerandomization strategies and obtain valid inferences for the average treatment effect, we first establish the joint asymptotic normality of the stratified difference-in-means estimator for vector potential outcomes. Our analysis is conducted under the randomization inference framework, where both $Y_i(z)$ and $X_i$ are fixed quantities, and randomness originates only from the treatment assignment $Z_i$.

3. Joint Asymptotic Normality of Stratified Difference-in-Means Estimator

Let us consider (fixed) $d$-dimensional potential outcomes $R_i(z) = (R_{i1}(z), \ldots, R_{id}(z))^T$, $i = 1, \ldots, n$, $z = 0, 1$. In what follows, $R_i(z)$ can take the form of $Y_i(z), X_i$, or $(Y_i(z), X_i)^T$.

Similar to the definitions established in Section 2, we can define the vector-form average treatment effect $\bar{\tau}_R$, its stratified difference-in-means estimator $\hat{\tau}_R$, and the covariances of $R_i(z)$ and $\bar{\tau}_R$.

Proposition 1. Under SR, the covariance of $n^{1/2}(\hat{\tau}_R - \bar{\tau}_R)$ is

$$
\Sigma_R = \sum_{k=1}^{K} \pi_{[k]} \left[ \frac{S_{[k]}^{2}(1)}{p_{[k]}} + \frac{S_{[k]}^{2}(0)}{1 - p_{[k]}} - S_{[k]}^{2} \right].
$$

Because covariances can be considered potential outcomes with no treatment effect, we can apply Proposition 1 to $R_i(z) = (Y_i(z), X_i)^T$ and obtain the following proposition.

Proposition 2. Under SR, the covariance of $n^{1/2}(\hat{\tau} - \bar{\tau}, \hat{\tau}_R^T)$ is

$$
\Sigma = \left( \begin{array}{cc}
\Sigma_{\tau \tau} & \Sigma_{\tau x} \\
\Sigma_{x \tau} & \Sigma_{xx}
\end{array} \right)
= \sum_{k=1}^{K} \pi_{[k]} \left( \frac{S_{[k]}^{2}(1)}{p_{[k]}} + \frac{S_{[k]}^{2}(0)}{1 - p_{[k]}} - S_{[k]}^{2} \right) \left( \frac{S_{[k]}^{2}(1)}{p_{[k]}} + \frac{S_{[k]}^{2}(0)}{1 - p_{[k]}} - S_{[k]}^{2} \right).
$$

To establish the joint asymptotic normality of $\hat{\tau}_R$, the following conditions need to be satisfied. Without further explanation, limits are taken as $n$ tends to infinity with no restriction on $K$ and $n_{ik}$. Let $\| \cdot \|_\infty$ denote the infinity norm of a vector, and let $N(\mu, \Sigma)$ denote a normal distribution with mean $\mu$ and covariance matrix $\Sigma$.

Condition 1. For $k = 1, \ldots, K$, there exist constants $p_{k}^{\infty}$ and $c \in (0, 0.5)$ such that $p_{k}^{\infty} \in (c, 1 - c)$ and $\max_{k=1,\ldots,K} \|R_{[k]} - R_{[k]}(z)\|_\infty / n \to 0$.

Condition 2. For $z = 0, 1$, $\max_{k=1,\ldots,K} \max_{i\in[k]} \|R_{i}(z) - R_{[k]}(z)\|_\infty / n \to 0$.

Condition 3. The following three matrices have finite limits:

$$
\sum_{k=1}^{K} \pi_{[k]} \frac{S_{[k]}^{2}(1)}{p_{[k]}}, \sum_{k=1}^{K} \pi_{[k]} \frac{S_{[k]}^{2}(0)}{1 - p_{[k]}}, \sum_{k=1}^{K} \pi_{[k]} \frac{S_{[k]}^{2}}{\tau_{[k]}^R},
$$

and the limit of $\Sigma_R$, denoted as $\Sigma_R^{\infty}$, is (strictly) positive definite.

Remark 1. Condition 1 assumes that the propensity scores for all strata uniformly converge to limits between zero and one. Condition 2 requires that the maximum squared distance between each component of the potential outcomes and its stratum-specific means, divided by $n$, tends to zero. When $K = 1$, Condition 2 reduces to that proposed in Li and Ding (2017) for establishing the finite-population central limit theorem for simple randomization. Condition 3 is a technical condition. When $d = 1$, Conditions 2 and 3 reduce to those proposed in Liu and Yang (2019) for analyzing the properties of regression adjustments in stratified randomized experiments.

Theorem 1. Under Conditions 1–3 and SR, $n^{1/2}(\hat{\tau}_R - \bar{\tau}_R)$ converges in distribution to $N(0, \Sigma^{\infty}_R)$ as $n$ tends to infinity.

Theorem 1 provides a normal approximation for the distribution of $\hat{\tau}_R$. It generalizes the asymptotic normality of the stratified difference-in-means estimator from one-dimensional outcomes (Liu and Yang 2019) to $d$-dimensional vector outcomes, as well as the result of Li and Ding (2017) from simple randomization to SR. The generalization is straightforward in case of a fixed $K$ with each $n_{ik}$ tending to infinity, but novel for an asymptotic regime where both $K$ and $n_{ik}$ can tend to infinity, including the special cases of paired randomized experiments, finely stratified randomized experiments (Fogarty 2018), and threshold blocking design (Higgins, Säve, and Sekhon 2016).

To construct a large-sample confidence set of $\tau_R$ using Theorem 1, we propose an asymptotically conservative estimator of $\Sigma_R$. When there are at least two treated and two control units in each stratum, we can replace the stratum-specific population covariances, $S_{[k]}^{2}(1)$ and $S_{[k]}^{2}(0)$, by their sample analog and ignore the term $S_{[k]}^{2}$ to conservatively estimate $\Sigma_R$. When there is only one treated or control unit in some strata, we can generalize the variance estimator proposed by Ashley and Miratrix (2020) to handle $d$-dimensional potential outcomes. Define $\hat{\tau}_R^{2}(z)$ as the sample covariance of $R_{ik}$ in stratum $k$ under treatment arm $z$ when $n_{ik} \geq 2, z = 0, 1$. Let $A_{ss} = \{k: n_{ik} = 1 \lor n_{ik} = 0\}$ be the set of small strata, where "ss" stands for "small strata". Let $n_{ss} = \sum_{k \in A_{ss}} n_{ik}$ be the total number of units in small strata, and let $\tau_{ss} = \sum_{k \in A_{ss}} (n_{ik}/n_{ss}) \tau_{[k]}$ and $\hat{\tau}_{ss} = \sum_{k \in A_{ss}} (n_{ik}/n_{ss}) \hat{\tau}_{[k]}$. Throughout the article, we assume that $n_{ik} \leq n_{ss}/2$ for all $k \in A_{ss}$. Then an asymptotically conservative estimator of $\Sigma_R$ is

$$
\hat{\Sigma}_R = \sum_{k \not\in A_{ss}} \pi_{[k]} \left[ \frac{\hat{\tau}_R^{2}(1)}{p_{[k]}} + \frac{\hat{\tau}_R^{2}(0)}{1 - p_{[k]}} \right] + \left( \frac{n_{ss}}{n} \right)^2 \sum_{k \in A_{ss}} (n_{ss} - 2n_{ik}) (n_{ss} + \sum_{h \in A_{ss}} n_{ih} - 2n_{ih}) (\hat{\tau}_{[k]} - \hat{\tau}_{ss}) (\hat{\tau}_{[k]} - \hat{\tau}_{ss})^T.
$$
**Condition 4.** There exists a constant $C$ such that $n^{-1} \sum_{i=1}^{n} R_i(z)^T R_i(z) \leq C$, $z = 0, 1$.

**Theorem 2.** Under Conditions 1–4 and SR, $E(\hat{\Sigma}_R) = \Sigma_R$ and $\hat{\Sigma}_R - \Sigma_R$ converges to zero in probability, where

$$\hat{\Sigma}_R = \Sigma_R + \sum_{k \neq A_k} n_k^2 \pi_k S_{[k]}^T \tau \hat{\Sigma}_R$$

$$+ \frac{n_k^2}{k \neq A_k (n_{ss} - 2n_k)} \frac{n_k^2}{\tau_{[k]} - \tau_{ss}} (\tau_{[k]} - \tau_{ss})^T$$

with $\hat{\Sigma}_R - \Sigma_R$ being positive semidefinite.

Next, we apply Theorem 1 to $R_i(z) = (Y_i(z), X_i^T)^T$. The following conditions should be met.

**Condition 5.** For each treatment arm $z = 0, 1$,

$$\max_{k=1, \ldots, K} \max_{i \in [k]} \| Y_i(z) - \tilde{Y}_i(z) \|^2 / n \rightarrow 0, \quad \text{and} \quad \max_{k=1, \ldots, K} \max_{i \in [k]} \| X_i - \tilde{X}_i \|_\infty / n \rightarrow 0.$$

**Condition 6.** The following two matrices have finite limits:

$$\sum_{k=1}^{K} \pi_k [S_{[k]}^2(1) S_{[k]}^T]_{XYY}(1),$$

$$\sum_{k=1}^{K} \pi_k [1 - p_k] \left( S_{[k]}^2(0) S_{[k]}^T_{XYY}(0) \right),$$

$$\sum_{k=1}^{K} \pi_k S_{[k]}^T \tau \hat{\Sigma}_R$$

has a limit, and the limit of $\Sigma_t$ denoted by $\Sigma_\infty$, is (strictly) positive definite.

**Corollary 1.** Under SR, if Conditions 1, 5, and 6 hold, then $n^{1/2}(\hat{\tau} - \tau, \hat{\tau}_{X}^T)^T$ converges in distribution to $N(0, \Sigma_\infty)$ as $n$ tends to infinity.

## 4. Asymptotics of Stratified Randomization

### 4.1. Stratified Randomization Based on the Overall Mahalanobis Distance

According to Proposition 2, the overall Mahalanobis distance $M_{\hat{\tau}} = (\hat{\tau}_{X})^T \Sigma_{XX}^{-1} \hat{\tau}_{X}$, where $\Sigma_{XX} = \sum_{k=1}^{K} \pi_k S_{[k]} S_{[k]}^T / (p_k (1 - p_k))$, is the lower right block matrix of $\Sigma$ known at the design stage of the experiment. Denote $\hat{M}_{\hat{\tau}} = \{ (Z_1, \ldots, Z_n) : \hat{M}_{\hat{\tau}} < a \}$ as an event that an assignment is accepted under the stratified randomization based on the overall Mahalanobis distance $M_{\hat{\tau}}$, which is abbreviated as SRoM.

**Proposition 3.** Under SRoM, if Conditions 1, 5, and 6 hold, then the asymptotic probability of accepting a random assignment is $p_a = \Pr(\chi^2_\rho < a)$, where $\chi^2_\rho$ represents a chi-squared distribution with $\rho$ degrees of freedom.

The asymptotic distribution of $n^{1/2}(\hat{\tau} - \tau, \hat{\tau}_{X})^T$ can be derived from Corollary 1. Let $R^2 = \text{cov}(\hat{\tau}, \hat{\tau}_{X}) / \text{var}(\hat{\tau}_{X})^{-1}$

$$\text{cov}(\hat{\tau}_{X}, \hat{\tau}) / \text{var}(\hat{\tau}_{X})^{-1} = \Sigma_{tX} \Sigma_{XX}^{-1} \Sigma_{XX} / \Sigma_{tX}$$

be the squared multiple correlation between $\hat{\tau}$ and $\hat{\tau}_{X}$ under SR. Let us denote independent random variables as $e_0 \sim N(0, 1)$ and $L_{p,a} \sim (D_1 | D' D < a)$, where $D = (D_1, \ldots, D_p)^T$ is a $p$-dimensional $N(0, I)$ distributed random vector. In what follows, the notation $\tilde{\cdot}$ will be used for two sequences of random vectors converging to the same distribution as $n$ tends to infinity.

**Theorem 3.** Under SRoM, if Conditions 1, 5, and 6 hold, then

$$n^{1/2}(\hat{\tau} - \tau) | M_{\hat{\tau}_{X}} \sim \hat{\Sigma}_{\hat{\tau}_{X}} \left\{ (1 - R^2)^{1/2} e_0 + (R^2)^{1/2} L_{p,a} \right\}.$$

When the number of strata is fixed with their sizes tending to infinity, Theorem 3 becomes a direct extension of the asymptotic result of rerandomization in completely randomized experiments (Li, Ding, and Rubin 2018), and can also be obtained from the asymptotic theory of rerandomization for tiers of covariates (Morgan and Rubin 2015; Li, Ding, and Rubin 2018). The novelty of this theorem lies in the fact that it makes few restrictions on the number of strata and their sizes, allowing the number of strata to tend to infinity with their sizes fixed.

According to Theorem 3, the asymptotic distribution of the stratified estimator under SRoM is a truncated-normal, which has the same formula as that of the difference-in-means estimator under merely rerandomization; however, both $\Sigma_{tX}$ and $R^2$ have distinct meanings due to different sources of randomness.

**Theorem 3** implies the asymptotic unbiasedness and improvement in the efficiency of stratified rerandomization, as summarized in the next corollary. Let $\nu_{p,a} = \Pr(\chi^2_{p+2} < a)/\Pr(\chi^2_{p} \leq a)$ denote the variance of $L_{p,a}$.

**Corollary 2.** Under SRoM, if Conditions 1, 5, and 6 hold, then $\hat{\tau}$ is an asymptotically unbiased estimator of $\tau$. The asymptotic variance of $n^{1/2}(\hat{\tau} - \tau)$ under SRoM is the limit of $\Sigma_{tX} \{ 1 - (1 - \nu_{p,a}) R^2 \}$, whereas the percentage of reduction in asymptotic variance compared to SR is the limit of $(1 - \nu_{p,a}) R^2$.

**Remark 2.** Schultzberg and Johansson (2019) proposed a stratified rerandomization strategy using the Mahalanobis distance $M_{\hat{\tau}} = (T_{X})^T \Sigma_{XX}^{-1} T_{X}$, where $T_{X} = (1/n_1) \sum_{i=1}^{n_1} Z_{X} - (1/n_0) \sum_{i=n_1+1}^{n} Z_{X}$ is the difference-in-means estimator of the covariates. They showed that there is no guarantee that the overall difference-in-means estimator is unbiased (non-asymptotically) unless $n_{[k]} = n_{[k]}$ in each stratum. We generalize this result using asymptotic theory in the Supplementary Material. We show that the overall difference-in-means estimator can be asymptotically biased if the propensity scores differ across strata. The aim of Schultzberg and Johansson (2019) was to show that substantial computational and efficiency gains can be obtained by first stratifying and then finding the “optimal” allocations within each stratum. For this procedure, the asymptotic theory in Li, Ding, and Rubin (2018) has not yet been proved to be valid, which is why Schultzberg and Johansson (2019) suggested a Fisher randomization test for inferences.

As the threshold $a$ tends to 0, the asymptotic variance $\Sigma_{tX} \{ 1 - (1 - \nu_{p,a}) R^2 \}$ tends to its minimum value $\Sigma_{tX} (1 - R^2) = n \times \min_{\gamma} E(\hat{\tau} - \tau - \hat{\tau}_{X}^T \gamma)^2$, which is equal to the
variance of the errors in the linear projection of $\sqrt{n}(\hat{\tau} - \tau)$ onto $\sqrt{n} \hat{T}_X$. Let us define the projection coefficient vector as $\gamma_{opt} = \arg \min_{\gamma} E \left( \sqrt{n}(\hat{\tau} - \tau - \hat{T}_X^T \gamma)^2 \right) = \Sigma_{XX}^{-1} \Sigma_{X\tau}$. This motivates us to consider a covariate-adjusted estimator $\tau_{adj} = \hat{\tau} - \hat{T}_X^T \hat{\gamma}_{opt}$, which has the smallest asymptotic variance among the adjusted estimators of the same form. Asymptotically, the efficiency of $\tau_{adj}$ under SR is equal to that of $\hat{\tau}$ under SRRoM with a threshold $a \rightarrow 0$. That is, stratified rerandomization can be viewed as covariate/regression adjustment in design. In practice, however, $\gamma_{opt}$ is usually unknown because it depends on the unknown potential outcomes. We need to derive a consistent estimator of $\gamma_{opt}$. Define $s_{X|Y}(1)$ as the sample covariance between $X_i$ and $Y_i(1)$ in stratum $k$ under the treatment when $n_{k[i]} \geq 2$ and define it as $\{n_{k[i]}/(n_{k[i]} - 1)\} \sum_i I_{k[i]}(Z_i - \bar{X}_i[k])Y_i^{obs}$ when $n_{k[i]} = 1$. The intuition comes from the fact that, when $n_{k[i]} = 1$,

$$E \left[ \frac{n_{k[i]} - 1}{n_{k[i]} - 1} \sum_i I_{k[i]}(Z_i - \bar{X}_i[k])Y_i^{obs} \right] = \frac{1}{n_{k[i]} - 1} \sum_i (X_i - \bar{X}_i[k])Y_i(1) = s_{X|Y}(1).$$

This relies on our knowledge of $\bar{X}_i[k]$. Similarly, we can define $s_{X|Y}(0)$.

A consistent estimator of $\gamma_{opt}$ is

$$\hat{\gamma}_{opt} = \Sigma_{XX}^{-1} \sum_{k=1}^K \pi_k \left( \frac{s_{X|Y}(1)}{p[k]} + \frac{s_{X|Y}(0)}{1 - p[k]} \right).$$

**Condition 7.** There exists a constant $C$ such that $n^{-1} \sum_{i=1}^n Y_i^2(z) \leq C, z = 0, 1$.

**Proposition 4.** Under SR or SRRoM, if **Conditions 1 and 5–7** hold, then $\hat{\gamma}_{opt} - \gamma_{opt}$ converges to zero in probability.

**Remark 4.** Let $I_{k[i]}$ be the stratification indicator. With unequal propensity scores, we cannot obtain $\tau_{adj}$ by running a single regression of $Y_i$ on $Z_i, I_{k[i]}, X_i$, and/or their interactions, while with equal propensity scores, let $p[k] = \pi$, we have $\gamma_{opt} = (1 - \pi) \beta(1) + \pi \beta(0)$, where $\beta(z) = (\sum_{k=1}^K \pi_k s_{X|Y}(z) \Sigma_{XX}^{-1} \sum_{k=1}^K \pi_k s_{X|Y}(z))$ is the projection coefficient of $X_i$ in the weighted projection of $Y_i(1)$ onto $I_{k[i]}$ and $X_i$ with weights $\pi_k/(n_{k[i]} - 1)$. When $n_{k[i]} \geq 2$ for all $k$ and $z$, we can estimate $\beta(z)$ using the sample, that is, by running a weighted regression of $Y_i(1)$ on $I_{k[i]}$ and $X_i$ under treatment arm $z$ with weights $\pi_k/(n_{k[i]} - 1)$. Let $\beta(z)$ be the OLS estimator of the coefficient of $X_i$. Then we can derive another consistent estimator of $\gamma_{opt}$:

$$\tilde{\tau}_{adj} = \hat{\tau} - \hat{T}_X \hat{\gamma}_{opt} = \sum_{k=1}^K \pi_k \left[ \left( \tilde{Y}_i^{obs}[k] - \bar{X}_i[k] \right) \hat{\beta}(1) \right] - \left( \tilde{Y}_i^{obs}[k] - \bar{X}_i[k] \right) \hat{\beta}(0).$$

As shown by Liu and Yang (2019), $\tilde{\tau}_{adj}$ is the OLS estimator of $Z_i$ in a weighted regression of $Y_i$ on $Z_i, I_{k[i]}, X_i, Z_i \times (I_{k[i]} - \pi[k])$, and $Z_i \times (X_i - \bar{X}_i[k]).$

**Remark 5.** In completely randomized experiments, rerandomization followed by regression adjustment can further improve the estimation efficiency if the analyzer has access to more covariates than the designer (Li and Ding 2020). This property is inherited in stratified randomized experiments. Under SRRoM, the regression-adjusted estimator $\hat{\tau}_{adj}$ using all the covariates that are used in the design stage and additional covariates that are collected in the analysis stage can further improve the efficiency.

Next, we compare the quantile ranges of $n^{1/2}(\hat{\tau} - \tau)$ under SRRoM and SR. Let $v_p(R^2, p_a, p)$ be the $\xi$th quantile of the random variable $(1 - R^2)^{1/2} \alpha_p + R^2)^{1/2} \alpha_p$, then under SRRoM, the asymptotic $(1 - \alpha)$ quantile range of $n^{1/2}(\hat{\tau} - \tau)$ is the limit of

$$\left[ \Sigma_{1} \tau \Sigma_{1} v_{p/2}(R^2, p_a, p), \Sigma_{1} \tau v_{\alpha/2}(R^2, p_a, p) \right],$$

for the length of which we present the following corollary.

**Corollary 3.** If **Conditions 1, 5, and 6** hold, then the length of the $(1 - \alpha)$ quantile range of the asymmetric distribution of $n^{1/2}(\hat{\tau} - \tau)$ under SRRoM is less than or equal to that under SR; this length is nonincreasing in $R^2$ and nondecreasing in $p_a$ and $p$.

**Remark 6.** This result is similar with Theorem 2 in Li, Ding, and Rubin (2018) where they compared the lengths of the quantile ranges of $n^{1/2}(\hat{\tau} - \tau)$ under rerandomization using Mahalanobis distance and complete randomization, while we compare the lengths under SRRoM and SR. **Corollary 3** suggests that a smaller value of $p_a$ leads to better improvement; however, setting $p_a$ to a very small value can be problematic if very few assignments are acceptable, which renders little power to randomization inference. Thus, how to choose the value of $p_a$ remains an open issue and should be investigated in the future. In practice, we suggest to choose a small value of $p_a$, for example, $p_a = 0.001$.

As the experimental results yield only part of the potential outcomes, the precise variance of $\hat{\tau}$ and the theoretical confidence interval of $\tau$ are unknown; however, we can construct asymptotically conservative estimators. According to **Proposition 4**, a consistent estimator of $\Sigma_{1} \tau$ is

$$\hat{\Sigma}_{1} \tau = \hat{\Sigma}_{1} \tau = \sum_{k=1}^K \pi_k \left[ \frac{s_{X|Y}(1)}{p[k]} + \frac{s_{X|Y}(0)}{1 - p[k]} \right].$$

Applying **Theorem 2** to $Y_i(z)$, we can estimate $\Sigma_{1} \tau$ as follows. Define $\bar{s}_{i[k]}(z)$ as the sample variance of $Y_i(z)$ in stratum $k$ under treatment arm $z$ when $n_{k[i]} \geq 2, z = 0, 1$. Let $\hat{\tau}_n = \sum_{k \in A_a} (n_{k[i]}/n_a) \hat{\tau}_n$. Then a conservative estimator of $\Sigma_{1} \tau$ is

$$\hat{\Sigma}_{1} \tau = \sum_{k \in A_a} \pi_k \left[ \frac{\bar{s}_{i[k]}(z)}{p[k]} + \frac{\bar{s}_{i[k]}(0)}{1 - p[k]} \right] \left( \frac{n_a}{n} \right)^2 \left( \hat{\tau}_n - \hat{\tau}_n \right)^2,$$

$$\sum_{k \in A_a} \left( n_a - n_{k[i]} \right) \left( n_a + \sum_{h \in A_a} \pi_{k[h]}^2 - n_{k[i]} \right) \left( \hat{\tau}_n - \hat{\tau}_n \right)^2,$$
Let $\hat{R}^2 = \hat{\Sigma}_{\tau\tau}/\hat{\Sigma}_{x\tau}$ and $v_0(\hat{R}^2, p_\alpha, p)$ be the $\xi$th quantile of $(1 - \hat{R}^2)^{1/2}/\varepsilon_0 + (R^2)^{1/2}L_{p,\alpha}$.

**Theorem 4.** Under SRRoM, if Conditions 1 and 5–7 hold, then
$$\hat{\Sigma}_{\tau\tau}(1 - (1 - v_{p,\alpha})\hat{R}^2)$$
is an asymptotically conservative estimator for the asymptotic variance of $n^{1/2}(\hat{\tau} - \tau)$ and
$$[\hat{\tau} - (\hat{\Sigma}_{\tau\tau}/n)^{1/2}v_{1-\alpha/2}(\hat{R}^2, p_\alpha, p), \hat{\tau} - (\hat{\Sigma}_{\tau\tau}/n)^{1/2}v_{1/2}(\hat{R}^2, p_\alpha, p)]$$
is an asymptotically conservative $(1 - \alpha)$ confidence interval of $\tau$.

**Remark 7.** Based on Pasley and Miratrix (2020), we obtain another conservative variance estimator, which is applicable when there exist at least two strata of each (small) stratum size; see the supplementary material for detailed discussions.

**Remark 8.** The above results can be applied to paired randomized experiments, finely stratified randomized experiments (Fogarty 2018), and threshold blocking designs (Higgins, Sävé, and Sekhon 2016). Moreover, when there are at least two treated and two control units in each stratum, we do not require Condition 7.

### 4.2. Stratified Rerandomization Based on the Stratum-Specific Mahalanobis Distance

In the special case where $K$ is fixed and all $n_{k[i]}$’s tend to infinity, we can rerandomize in each stratum separately and independently. Let $M_{i} = \{Z_{i1}, \ldots, Z_{iN}: M_{[k]} < a_k, k = 1, \ldots, K\}$ denote an event in which an assignment is accepted under the stratified rerandomization based on the stratum-specific Mahalanobis distance $M_{[k]}$, which is abbreviated as SRRoM. In this section, we assume that $K$ is fixed and $n_{k[i]} \to \infty$, $k = 1, \ldots, K$ as $n \to \infty$ unless stated otherwise.

Since $n^{1/2}(\hat{\tau} - \tau) = \sum_{k=1}^{K} p_{k[i]} n_{k[i]}(\hat{\tau}_{k[i]} - \tau_{k[i]})$, and each stratum is rerandomized independently under SRRoM, we can simply apply the asymptotic distribution of $n^{1/2}(\hat{\tau}_{k[i]} - \tau_{k[i]})$ under complete rerandomization (Li, Ding, and Rubin 2018) to derive the asymptotic distribution of $n^{1/2}(\hat{\tau} - \tau)$.

**Condition 8.** For each $k = 1, \ldots, K$, as $n_{k[i]} \to \infty$, $S_{[k]}^2(z) (z = 0, 1)$ and $S_{[k]}^2(z)$ have finite limits; the limit of $\text{var}(n_{k[i]}^{1/2}(\hat{\tau}_{k[i]} - \tau_{k[i]}))$ is positive; $S_{[k]}^{x}XX$ converges to a (strictly) positive definite matrix; and $S_{[k]}^{xXY}(z) (z = 1)$ converges to finite limits.

**Proposition 5.** Under SRRoM, if Conditions 1, 5–6, and 8 hold, then the asymptotic probability of accepting a random assignment is
$$\prod_{k=1}^{K} p_{k[i]} = \prod_{k=1}^{K} \text{pr}(\chi^2_{p_\alpha} < a_k).$$

Let us denote the covariance matrix of $n_{k[i]}^{1/2}(\hat{\tau}_{k[i]} - \tau_{k[i]}), (\hat{\tau}_{k[i]}X - \tau_{k[i]}X)^{T}$ as
$$\begin{pmatrix}
S_{k\tau\tau}^{1} & S_{k\tau\tau}^{2} & S_{k\tau\tau}^{3} \\
S_{k\tau\tau}^{2} & S_{k\tau\tau}^{4} & S_{k\tau\tau}^{5} \\
S_{k\tau\tau}^{3} & S_{k\tau\tau}^{5} & S_{k\tau\tau}^{6}
\end{pmatrix}
,$$and let $R^2_{[k]} = \Sigma_{[k]\tau\tau}^{-1} \Sigma_{[k]\tau\tau}^{-1} / \Sigma_{[k]\tau\tau}$ be the squared correlation between $\hat{\tau}_{k[i]}$ and $\tau_{k[i]}$ under SR. Let $\varepsilon_0$ be a $N(0, 1)$ distributed random variable and let $L_{p,\alpha}^{1}, \ldots, L_{p,\alpha}^{K}$ be independent and $L_{p,\alpha}^{k} \sim \text{Gamma} k = 1, \ldots, K$, where $L_{p,\alpha}^{k}$ is defined in Section 4.1. Suppose that $\varepsilon_0$ and $L_{p,\alpha}^{1}, \ldots, L_{p,\alpha}^{K}$ are independent.

**Theorem 5.** Under SRRsM, if Conditions 1, 5–6, and 8 hold, then

$$n^{1/2}(\hat{\tau} - \tau) \bigg| M_{i} \sim \left\{ \sum_{k=1}^{K} \left( \tau_{[k]} - \tau \right) \right\} \right)^{1/2} \varepsilon_0$$

$$+ \sum_{k=1}^{K} \left( \tau_{[k]} - \tau \right)^{1/2} L_{p,\alpha}^{k}.$$ (4)

The asymptotic unbiasedness of $\hat{\tau}$, asymptotic variance of $n^{1/2}(\hat{\tau} - \tau)$ under SRRsM, and variance reduction are summarized in the following corollary.

**Corollary 4.** Under SRRsM, if Conditions 1, 5–6, and 8 hold, then $\hat{\tau}$ is an asymptotically unbiased estimator of $\tau$, and the asymptotic variance of $n^{1/2}(\hat{\tau} - \tau)$ is the limit of $\sum_{k=1}^{K} \left( \tau_{[k]} - \tau \right)^{1/2} \left( \tau_{[k]} - \tau \right)^{1/2} L_{p,\alpha}^{k}$, and the percentage of reduction in asymptotic variance compared to SR is the limit of $\sum_{k=1}^{K} \left( \tau_{[k]} - \tau \right)^{1/2} \left( \tau_{[k]} - \tau \right)^{1/2} L_{p,\alpha}^{k} / \Sigma_{\tau\tau}$.

SRRoM is also applicable in this case, whereas intuitively, SRRsM achieves better covariance balance because it balances covariates in each stratum. According to Propositions 3 and 5, asymptotically, the proportions of all possible assignments $p_{\alpha}$ and $\prod_{k=1}^{K} p_{k[i]}$ are acceptable under SRRoM and SRRsM, respectively. Therefore, if we use identical thresholds, that is, $a_1 = \cdots = a_K = a$, SRRsM appears stricter than SRRoM because $\prod_{k=1}^{K} p_{k[i]} = (p_{\alpha})^K < p_{\alpha}$.

**Theorem 6.** When the thresholds $a_1, \ldots, a_K$ and $a$ are identical or tend to 0, the asymptotic variance of $n^{1/2}(\hat{\tau} - \tau)$ under SRRsM is smaller than or equal to that under SRRoM. Particularly, $\sum_{k=1}^{K} \left( \tau_{[k]} - \tau \right)^{1/2} \left( \tau_{[k]} - \tau \right)^{1/2} L_{p,\alpha}^{k}$, where the equality holds if and only if $\Sigma_{[k]}^{-1} \Sigma_{[k]}^{X\tau}$ for $k = 1, \ldots, K$.

**Theorem 6** implies that SRRsM improves the efficiency of SRRoM in the situation where there are only a few large strata and the thresholds $a_1, \ldots, a_K$ and $a$ are identical or tend to 0. The only exception (that is, they have the same efficiency) is the case that the strata are homogeneous in the sense that the stratum-specific projection coefficients $\Sigma_{[k]}^{-1} \Sigma_{[k]}^{X\tau}$ for $k = 1, \ldots, K$ are the same as the overall projection coefficients $\Sigma_{xx}^{-1} \Sigma_{x\tau}$, when projecting the treatment effect onto the covariates. In other situations, the relative reduction in asymptotic variance is related, in a complicated form, to $\tau_{p,\alpha} \tau_{p,\alpha}$ and the covariances matrices defined in Equations (2) and (3). In our simulation study, the SRRsM with $p_{k[i]} = (p_{\alpha})^{1/K}$ (which ensures the same acceptance probabilities) performs better than the SRRoM when there are a few heterogeneous strata. In contrast, when there exist many small strata, SRRsM performs worse than SRRoM, even with $p_{k[i]} = p_{\alpha}$. 
Now we compare the quantile ranges of $n^{1/2}(\hat{\tau} - \tau)$ under SRRsM and SR. Denote $q_{\alpha}(R_{[1]}^2, \ldots, R_{[K]}^2, p_{a_1}, \ldots, p_{a_K}, p)$ as the $\xi$th quantile of the random variable on the right-hand side of Equation (4), then the asymptotic $(1 - \alpha)$ quantile range of $n^{1/2}(\hat{\tau} - \tau)$ under SRRsM is the limit of

$$\left[q_{\alpha/2}(R_{[1]}^2, \ldots, R_{[K]}^2, p_{a_1}, \ldots, p_{a_K}, p), q_{1-\alpha/2}(R_{[1]}^2, \ldots, R_{[K]}^2, p_{a_1}, \ldots, p_{a_K}, p)\right].$$

**Corollary 5.** Under SRRsM, if Conditions 1, 5–6, and 8 hold, then the length of the $(1 - \alpha)$ quantile range of the asymptotic distribution of $n^{1/2}(\hat{\tau} - \tau)$ is less than or equal to that under SR, with the length nonincreasing in $R_{[1]}^2, \ldots, R_{[K]}^2$ and non-decreasing in $p_{a_1}, \ldots, p_{a_K}$ and $p$.

To obtain a valid inference of $\tau$ based on $n^{1/2}(\hat{\tau} - \tau)$ under SRRsM, we need to estimate the asymptotic variance and quantile range. To achieve this, we follow Li, Ding, and Rubin (2018). Let

$$s_{[k]|X}^2 = \{s_{[k]|Y}(1) - s_{[k]|Y}(0)\}^T (S_{[k]|X}^{-1}) \{s_{[k]|Y}(1) - s_{[k]|Y}(0)\}$$

be an estimator of the variance of the linear projection of $\tau$ on $X$ in stratum $k$. Then, $\Sigma_{[k]|X}$ is estimated by $\hat{\Sigma}_{[k]|X} = \hat{p}_{[k]}^{-1} s_{[k]|X}^2 (1 - \hat{p}_{[k]})^{-1} s_{[k]|X}^2 - \hat{\Sigma}_{[k]|X}$. Let $s_{[k]|Y|X}(z) = s_{[k]|Y}(z) S_{[k]|X}^{-1} s_{[k]|Y}(z)$ be the sample variance of linear projection of $Y$ on $X$. Then, the estimator of $R_{[k]}^2$ is

$$\hat{R}_{[k]}^2 = \hat{\Sigma}_{[k]|X}^{-1} \left[ \frac{s_{[k]|Y|X}(1)}{\hat{p}_{[k]} - \hat{\Sigma}_{[k]|X}^{-1} s_{[k]|X}^2 (1 - \hat{p}_{[k]})^{-1} s_{[k]|X}^2} \right],$$

which is set to 0 if the right-hand side is negative.

With the above-constructed estimators, the asymptotic distribution of $n^{1/2}(\hat{\tau} - \tau)$ can be estimated conservatively by

$$\left\{ \sum_{k=1}^K \pi_k (\hat{\tau}_{[k]|X} - \hat{R}_{[k]}^2) \right\}^{1/2} \epsilon_0 + \sum_{k=1}^K \pi_k (\hat{\tau}_{[k]|X} - \hat{R}_{[k]}^2)^{1/2} \epsilon_0, \pi_{a_k}.$$

Let $\hat{q}_\xi$ be the $\xi$th quantile of the above conservative distribution.

**Theorem 7.** Under SRRsM, if Conditions 1, 5–6, and 8 hold, then $\sum_{k=1}^K \pi_k \hat{\Sigma}_{[k]|X}^{-1} \left\{ 1 - \left( 1 - \pi_{a_k} \right) \hat{R}_{[k]}^2 \right\}^{1/2} \epsilon_0$ is an asymptotically conservative estimator for the asymptotic variance of $n^{1/2}(\hat{\tau} - \tau)$ and $\left\{ \hat{\tau} - n^{-1/2} \hat{q}_{1-\alpha/2}, \hat{\tau} - n^{-1/2} \hat{q}_{\alpha/2} \right\}$ is an asymptotically conservative $(1 - \alpha)$ confidence interval of $\tau$.

5. Simulation Study

We conduct a simulation study to evaluate the finite-sample performance of the point and interval estimators for the average treatment effect under stratified randomization strategies, SRRoM and SRRsM, and compare them with those under SR. Data are generated from the model: $Y_i(z) = X_i^T \beta_{z1} + \exp(X_i^T \beta_{z2} + e_i(z), i = 1, \ldots, n, z = 0, 1$, where the covariate vectors $X_i$’s are eight-dimensional vectors drawn independently from normal distribution with mean zero and covariance matrix $\Sigma$, whose entries $\Sigma_{ij} = 0.5^{\min(i,j)}, i, j = 1, \ldots, 8$, and the disturbances $e_i(z)$ are normally distributed with mean zero and variance 10. The jth components of the coefficients are generated independently from the distributions: $\beta_{11j} \sim t_3, \beta_{12j} \sim 0.1t_3, \beta_{21j} \sim t_3, \beta_{22j} \sim 0.1t_3, j = 1, \ldots, 8$, where $t_3$ denotes $t$ distribution with three degrees of freedom.

The number of strata $K$ and strata sizes $n_{[k]}$ are set in four cases: Case 1, there are many small strata, with $K = 25, 50, 100$, and $n_{[k]} = 10$; Case 2, there are many small strata and two large strata, with $K = 10 + 2, 20 + 2, 50 + 2, \text{and } n_{[k]} = 10$ for small strata and $n_{[k]} = 100$ for the two large strata; Case 3, there are two large homogeneous strata, with $K = 2$ and $n_{[k]} = 100, 200,$ and $500$; Case 4, there are two large heterogeneous strata where the coefficients $\beta_{z1}$ and $\beta_{z2}$ are generated independently for each stratum, with $K = 2$ and $n_{[k]} = 100, 200,$ and $500$. For the given $K$ and $n_{[k]}$, first, we generate the covariates and potential outcomes, and randomly assign $n_{[k]}$ units in stratum $k$ to the treatment group, where the propensity scores are equal, $p_{[k]} = 0.5$ ($k = 1, \ldots, K$), or unequal, $p_{[k]} = 0.4$ ($k \leq K/2, 0.6$ ($k > K/2$). If the covariate balance criterion is not met, then we generate new assignments until we find an assignment that meets the criterion. Then, based on the assignment, we compute the stratified difference-in-means estimator and the 95% confidence interval. In SR, we use $\hat{\tau} = (\Sigma_{[1]}^{-1})^{1/2} 2_{1-\alpha/2}, \hat{\tau} - (\Sigma_{[1]}^{-1})^{1/2} 2_{\alpha/2}$ as the conservative confidence interval of $\tau$, where $2_p$ is the $p$th quantile of a standard normal distribution. The preceding process of allocation and computation is repeated for $10^4$ times to evaluate the bias, standard deviation, root mean squared error (RMSE), mean confidence interval length, and empirical coverage probability under SR and stratified randomization. The threshold for SRRoM is set such that $p_a = 0.001$, and the thresholds for SRRsM are set such that $p_{a_k} = (0.001)^{1/K}$ for a fair comparison or $p_{a_k} = 0.001$ for an unfair comparison, $k = 1, \ldots, K$. For a stratum of size 10, there are only 252 possible assignments, and SRRsM sometimes rejects all possible assignments under an unfair comparison. In this case, we perform SR instead of SRRsM.

Figure 1 shows the results for equal propensity scores. Additional results are available in the Supplementary Material. Our findings are summarized as follows. First, the treatment effect estimators under all assignment mechanisms have small finite-sample biases, which are more than 10 times smaller than the standard deviations. Second, compared to SR, SRRoM always reduces the RMSEs and confidence interval lengths, regardless of the stratum numbers and sizes. The percentages of reduction are 2.6%–56.0% and 3.6%–40.7%, respectively. Third, when there exist small strata (Cases 1 and 2), fair SRRsM performs similarly to, or slightly better than, SR in terms of RMSE, and is less efficient than SRRoM. In this setting, fair SRRsM can still reduce the confidence interval lengths (3.4%–11.6%) compared to SR because it uses less conservative variance estimators. Fourth, when unfair SRRsM can be implemented (Cases 3 and 4), it is generally better than fair SRRsM because it uses stricter thresholds. When there are two large homogeneous strata (Case 3), fair SRRsM is less efficient than SRRoM, and unfair SRRsM is comparable to SRRoM in terms of RMSE but gives slightly longer confidence intervals. In contrast, when there are two large heterogeneous strata (Case 4), fair SRRsM is better than SRRoM, with percentages of reduction being 27.6%–38.9% in RMSEs and 13.5%–25.5% in confidence interval lengths. Finally, all the interval estimators are conservative, with the empirical coverage probabilities being larger than the confidence level.

In general, we recommend SRRoM when there exist small strata and SRRsM when there are only a few large strata. We
also conduct simulations for paired randomized experiments and finely stratified randomized experiments, and the results are given in the supplementary material.

6. Application

In this section, we analyze the “Opportunity Knocks” experiment data (Angrist, Oreopoulos, and Williams 2014) using two stratified rerandomization methods and compare them with SR. The Opportunity Knocks data are obtained from an experiment that aims at evaluating the influence of a financial incentive demonstration plan on college students’ academic performance. The research subjects of this experiment included first- and second-year students of a large Canadian commuter university who applied for financial aid. Stratification was conducted according to the year, sex, and high school GPA quintile. Students were randomly assigned to the treatment and control groups within each stratum, and those who fell in the treated group had peer advisors and received cash bonuses for attaining the given grades. Students with missing outcomes or covariates were excluded, resulting in 16 strata, a treatment group of size 382, and a control group of size 821. The propensity scores varied from 0.22 to 0.51.

The outcome of interest is the average grade for the semester right after the experiment (2008 fall). From the original dataset, we cannot determine the true gains of stratified rerandomization. To evaluate the repeated sampling properties of SRRoM and SRRsM, we generate a synthetic dataset, with the missing values of the potential outcomes imputed by a linear model of regressing the observed outcomes on the treatment indicator, average grade in 2008 spring, gender, and treatment-by-covariate interactions. The resulting average treatment effect is 0.205. To conduct stratified rerandomization, we select seven covariates: high school grade, average grade in 2008 spring, number of college graduates in the family, whether the first/second question in the survey is correctly answered, whether the mother tongue is English, and credits earned in 2008 fall. We center the covariates at their stratum-specific means. Stratified rerandomization is conducted under the same stratification and propensity scores as the original dataset, with acceptance probability \( p_a = 0.001 \) for SRRoM and \( p_{ak} = (p_a)^{1/16} \) for SRRsM for a fair comparison, and \( p_{ak} = 0.001 \) for an unfair comparison, \( k = 1, \ldots, K \). We repeat the stratified rerandomization for \( 10^4 \) times and compute the bias, standard deviation, RMSE, mean confidence interval length, and empirical coverage probability of different methods.

Table 1 and Figure 2 show the results, where the bias of each method is more than 10 times smaller than the standard deviation. Among the considered methods, SRRoM performs similarly to unfair SRRsM, and both of them are better than the other two methods. They reduce the RMSE of...
the stratified difference-in-mean estimator by approximately 26% when compared to SR. Fair SRRsM is less efficient than SRRoM and it does not substantially improve efficiency compared to SR. Moreover, all confidence intervals are conservative, with the coverage probabilities being larger than the confidence level.

7. Conclusion

In this article, we propose two rerandomization strategies based on the overall and stratum-specific Mahalanobis distances in stratified randomized experiments. We derive the asymptotic distributions of the stratified difference-in-means estimator under both strategies. We demonstrate the advantages of the proposed strategies through theoretical investigations, a simulation study, and a real data analysis. In addition, we provide asymptotically conservative estimators for the variances and asymptotic distributions, which can be used to construct large-sample conservative confidence intervals for the average treatment effect. Our work provides statistical support for the recommendation of renowned scholars, such as R. A. Fisher and D. B. Rubin, to "Block what you can and rerandomize what you cannot."

In theory, when there are only a few large strata, the stratum-specific stratified rerandomization strategy is more efficient than the overall stratified rerandomization strategy. Thus, when there are a few large strata and many small strata, it might be more efficient to pool small strata into large ones, and then, use the stratum-specific stratified rerandomization. It would be interesting to study how to efficiently pool small strata together and investigate the theoretical properties of the stratum-specific stratified rerandomization after pooling. Moreover, this article focuses on stratified rerandomization for a binary treatment and continuous potential outcomes. It would be interesting to generalize our results to stratified rerandomization with multiple treatments (including factorial experiments) or binary potential outcomes. In addition, we assume that the number of the additional covariates is fixed; in practice, however, the number of the additional covariates can be large, even larger than the sample size. It is worthy of further investigation to develop stratified rerandomization method using high-dimensional covariates.

Funding

The authors gratefully acknowledge the Tsinghua University Initiative Scientific Research Program and the National Natural Science Foundation of China grant 11701316 and 12071242.

Supplementary Material

The Supplementary Material provides a detailed discussion on Schultzberg and Johansson’s stratified rerandomization criterion, another conservative variance estimator, additional simulation results, and proofs.

References

Angrist, J., Oreopoulos, P., and Williams, T. (2014), “When Opportunity Knocks, Who Answers? New Evidence on College Achievement Awards,” Journal of Human Resources, 49, 572–610. [1302]
Athey, S., and Imbens, G. W. (2017), “The State of Applied Econometrics: Causality and Policy Evaluation,” Journal of Economic Perspectives, 31, 3–32. [1295]
Box, G. E. P., Hunter, J. S. and Hunter, W. G. (2005), Statistics for Experimenters: Design, Innovation and Discovery, Hoboken, NJ: Wiley-Interscience. [1295]
Fisher, R. A. (1926), “The Arrangement of Field Experiments,” Journal of the Ministry of Agriculture of Great Britain, 33, 503–513. [1295]
Fogarty, C. B. (2018), “On Mitigating the Analytical Limitations of Finely Stratified Experiments,” Journal of Royal Statistical Society, Series B, 80, 1035–1056. [1297,1300]
Gerber, A. S., and Green, D. P. (2012), Field Experiments: Design, Analysis and Interpretation, New York: WW Norton. [1295]
Higgins, M. J., Sävje, F., and Sekhon, J. S. (2016), “Improving Massive Experiments With Threshold Blocking,” Proceedings of the National Academy of Sciences of the United States of America, 113, 7369–7376. [1297,1300]
Imai, K. (2008), “Variance Identification and Efficiency Analysis in Randomized Experiments Under the Matched-Pair Design,” Statistics in Medicine, 27, 4857–4873. [1295]
Imbens, G. W., and Rubin, D. B. (2015), Causal Inference for Statistics, Social, and Biomedical Sciences, Cambridge: Cambridge University Press. [1295]
Kemphorner, O. (1955), “The Randomization Theory of Experimental Inference,” Journal of American Statistical Association, 50, 946–967. [1296]
Li, X. and Ding, P. (2017), “General Forms of Finite Population Central Limit Theorems With Applications to Causal Inference,” Journal of American Statistical Association, 112, 1759–1769. [1296,1297]
——— (2020), “Rerandomization and Regression Adjustment,” Journal of the Royal Statistical Society, Series B, 82(1), 241–268. [1299]
Li, X., Ding, P., and Rubin, D. B. (2018), “Asymptotic Theory of Rerandomization in Treatment-Control Experiments,” Proceedings of the National
Liu, H. and Yang, Y. (2019), "Regression-Adjusted Average Treatment Effect Estimators in Stratified Randomized Experiments," *Biometrika*, 107, 935–948. [1297,1299]

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013), "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments," *Journal of Royal Statistical Society*, Series B, 75, 369–396. [1295]

Morgan, K. L., and Rubin, D. B. (2012), "Rerandomization to Improve Covariate Balance in Experiments," *Annals of Statistics*, 40, 1263–82. [1295]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [1296]

——— (1980), "Randomization Analysis of Experimental Data: the Fisher Randomization Test Comment," *Journal of American Statistical Association*, 75, 591–593. [1296]

Schultzberg, M., and Johansson, P. (2019), "Re-randomization: A Complement or Substitute for Stratification in Randomized Experiments?" Working Paper, Department of Statistics, Uppsala University, 4. [1295,1298]

Neyman, J., Dabrowska, D. M., and Speed, T. (1990), 'On the Application of Probability Theory to Agricultural Experiments," *Statistical Science*, 5, 465–472. [1296]

Pashley, N. E., and Miratrix, L. W. (2020), "Insights on Variance Estimation for Blocked and Matched Pairs Designs," *Journal of Educational and Behavioral Statistics*. [1297,1300]

Rosenberger, W. F., and Lachin, J. M. (2015), *Randomization in Clinical Trials: Theory and Practice*, 2nd ed. Hoboken, NJ: Wiley. [1295]

Zhao, A., Ding, P., Mukerjee, R. and Dasgupta, T. (2018), "Randomization-Based Causal Inference From Split-Plot Designs," *Annals of Statistics*, 46, 1876–1903. [1296]