UPPSALA
UNIVERSITET

Working Paper 2019:3

Department of Statistics

On optimal re-randomization designs

Per Johansson, Donald Rubin and
Mårten Schultzberg

Working Paper 2019:3
April 2019
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Title: On optimal re-randomization designs

Author: Per Johansson, Donald Rubin and Mårten Schultzberg
E-mail: per.johansson@statistik.uu.se

On Optimal Rerandomization Designs

Per Johansson

Uppsala University, Uppsala, Sweden.

E-mail: per.johansson@statistics.uu.se

Donald B. Rubin

Tsinghua University, China and Fox/Temple University, US

Mårten Schultzberg

Uppsala University, Uppsala, Sweden.

Abstract. Blocking is commonly used in randomized experiments to increase efficiency of estimation. A generalization of blocking is to remove allocations with imbalance in covariates between treated and control units, and then randomize within the set of allocations with balance in these covariates. This idea of *rerandomization* was formalized by [5], who suggested using the affinely invariant Mahalanobis distance between treated and control covariate means as the criterion for removing unbalanced allocations. [3] proposed reducing the set of balanced allocations to the minimum. Here we discuss the implication of such an 'optimal' rerandomization design for inferences to the units in the sample and to the population from which the units in the sample were randomly drawn. We argue that, in general, it is a bad idea to seak the optimal design for an inference to the population because that inference typically only reflects uncertainty from the usually hypothetical random sampling, and not the randomization of treatment versus control.

Keywords: Rerandomization, Optimal Design, Causal inference, Mahalanobis distance, Exact inference

1. Introduction

Randomized controlled trials (RCT) are traditionally seen as the gold standard for causal inference. Even though the estimators from well-conducted experiments are unbiased in expectation, the estimates from any single experiment may still be far from being unbiased (in the colloquial sense) due to an unlucky, albeit random, allocation. Traditionally, blocking has been used to eliminate some bad allocations. Blocking (i.e., stratification) partitions experimental units into groups (strata) based on their similarity on covariates, and then randomization is performed within each stratum, thereby ensuring that units within each stratum will be 'fairly' represented in both the treatment and control groups, see e.g., Imbens and Rubin [2] for a recent overview. An alternative, or complement, to blocking that has received attention lately is to utilize modern computational capabilities by using rerandomization (see e.g., [5] and Kallus [3]).

More specifically, the idea is to remove, from consideration, allocations with imbalance in observed covariates between treated and control units and then randomize within the set of allocations with balance on these covariates; call the set of all allocations $\mathcal{A}$ and the set of acceptable allocations $\mathcal{A}_a$. Such a procedure is a rerandomization design with a pre-specified imbalance criterion, $\mathcal{A} \rightarrow \mathcal{A}_a$, where different criteria give rise to different rerandomization designs. The idea of rerandomization, which, although previously mentioned in the literature, apparently was first formalized by Morgan and Rubin [5], who used the Mahalanobis distance as the criterion for defining $\mathcal{A}_a$ with the aim to make inference about the Sample Average Treatment Effect (SATE), typically estimated by the sample difference in means of treated and control units.

Kallus [3] suggested finding the allocation that minimizes the estimated sampling variance of the SATE estimator, call the set of allocations that achieve this minimum $\mathcal{A}_{Opt}$. The primary focus of this paper is to discuss the implication of using such 'optimal' rerandomization designs for inferences to the population average treatment effect (PATE) and the SATE. Furthermore, by making use of the example in Kallus [3], the idea of Mahalanobis-based rerandomization as proposed in [5] is clarified. Finally, a small Monte Carlo study illustrates the problem with using standard asymptotic results with rerandomization designs, such as proposed by Kallus (2018), and compares the resulting inferences with the exact answer based on Fisher randomization inference for the SATE and with the asymptotic inference that correctly accounts for the restriction to $\mathcal{A}_a$ when drawing inference about SATE or PATE.

Section 2 reviews the background for Mahalanobis-based rerandomization. Section 3 discusses inference to the SATE and PATE and how to conduct inference given an optimal design. Section 4 provides the results from our small Monte Carlo study. The paper concludes with a discussion in Section 5.

2. Basic result from Morgan and Rubin (2012) on Mahalanobis-based rerandomization

Morgan and Rubin [5] suggested rerandomization based on the affinely invariant Mahalanobis distance between the treatment and control covariate means. Following the notation in that paper, consider a RCT with n units in the sample, indexed by i, with n_1 assigned to treatment and n_0 assigned to control, initially for simplicity, with $n_1 = n_0$. Let $W_i = 1$ or $W_i = 0$ if unit i is assigned treatment or control, respectively, and define $\mathbf{W} = (W_1, ..., W_n)'$. Furthermore, let $\mathbf{X}$ be the $n \times K$ matrix of fixed covariates in the sample $(\mathbf{x}_i, i = 1, ..., n)$, with sample covariance $cov(\mathbf{X})$.

Because $n_1 = n_0$, there are $\binom{n}{n_1} = A$ possible treatment allocation (assignment) vectors labelled $\mathbf{W}^j = (W_1^j, ..., W_n^j)'$, $j = 1, ..., A$, where card$(\mathcal{A}) = A$, i.e., the cardinality of the set $\mathcal{A}$. The Mahalanobis distance for allocation j is

$$M(\mathbf{W}^j, \mathbf{X}) \propto \hat{\tau}_X^j{}' cov(\mathbf{X})^{-1} \hat{\tau}_X^j, \; j = 1, ..., A,$$

where

$$\hat{\tau}_X^j = \frac{1}{n_1} \sum_{i=1}^{n_1} W_i^j \mathbf{x}_i' - \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - W_i^j) \mathbf{x}_i' = \overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j.$$

Morgan and Rubin (2012) proposed accepting the jth allocation when its treatment assignment vector $\mathbf{W}^j$ satisfies

$$M(\mathbf{W}^j, \mathbf{X}) \leq a,$$

where a is a positive constant.

By the central limit theorem, and by some experience with real examples for moderate n, the sample means of the covariates will be normally distributed across random samples, so that $M(\mathbf{W}^j, \mathbf{X}) \sim \chi_K^2$ (Morgan and Rubin, 2012). Letting

$$p_a = \Pr(\chi_K^2 \leq a) \simeq \Pr(M(\mathbf{W}^j, \mathbf{X}) \leq a), \tag{1}$$

we see that a is determined from the choice of p_a. Because the number of rerandomizations is geometrically distributed, the expected number of randomizations needed to obtain an acceptable allocation is $1/p_a$, which means that, for instance, with $p_a = 0.001$, the average number of randomizations before drawing an allocation that fulfils the criterion is $1,000$.

Morgan and Rubin [5] show that when $M(\mathbf{W}^j, \mathbf{X}) \sim \chi^2_K$,

$$Cov(\overline{\mathbf{X}}^j_T - \overline{\mathbf{X}}^j_C|\mathbf{X}, M(\mathbf{W}^j, \mathbf{X}) < a) = \nu_a Cov(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C|\mathbf{X}), \qquad (2)$$

with

$$\nu_a = \frac{\Pr(\chi^2_{(K+2)} \leq a)}{\Pr(\chi^2_K \leq a)};\ 0 < \nu_a < 1. \qquad (3)$$

This result implies that the variance in the covariate mean differences across allocations in $\mathcal{A}_a$ is reduced relative to its variance across the allocations in $\mathcal{A}$ by the factor ν_a, and the percent reduction in variance of each of the covariates in $\mathbf{X}$ (or any linear combination of them) is equal to $100(1 - \nu_a)$.

Let $Y_i(w)$ be the potential outcome under treatment w for individual i. Under the Stable Unit Treatment Value Assumption (SUTVA, Rubin 1980), the observed outcome when i is assigned W_i is equal to $Y_i = (1 - W_i)Y_i(0) + W_i Y(1)$. The mean difference estimator is defined as

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0 \qquad (4)$$

where $\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} W_i Y_i(1)$ and $\overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0}(1 - W_i)Y_i(0)$.

Let $\hat{\tau}^{CR}$ and $\hat{\tau}^{RR}$ be the estimators defined in (4) under complete randomization (i.e. when the W_i are randomly drawn from $\mathcal{A}$) and Mahalanobis-based rerandomization ((i.e. when the W_i are randomly drawn from $\mathcal{A}_a$), respectively. These estimators are unbiased for the estimation of the SATE and also of the PATE under random random sampling of the n units from the population. For the exposition of this paper it is convenient to also define the corresponding estimators for a specific sample s: $\hat{\tau}^{RR}_s$ and $\hat{\tau}^{RR}_s$, respectively.

Let $\mathbf{Y}(w) = (Y_1(w), Y_2(w), ..., Y_n(w))', w = 0, 1$, and let R^2 be the squared multiple correlation of $\mathbf{Y}(0)$ on $\mathbf{X}$. Under the assumption that (i) the residual in the linear projection of $\mathbf{Y}(0)$ on $\mathbf{X}$ is normally distributed and that (ii) treatment effects are additive (so that R^2 is also the squared multiple correlation of $\mathbf{Y}(1)$ on $\mathbf{X}$)) it holds that percentage reduction in variance (PRIV) of $\hat{\tau}^{RR}_s$ and $\hat{\tau}^{RR}$ against the corresponding estimators under complete randomization is

$$PRIV = \frac{V(\hat{\tau}_s^{CR}) - V(\hat{\tau}_s^{RR})}{V(\hat{\tau}_s^{CR})} = \frac{V(\hat{\tau}^{CR}) - V(\hat{\tau}^{RR})}{V(\hat{\tau}^{CR})} = 100 \times R^2(1-\nu_a). \tag{5}$$

where $V(.)$ denotes the variance of the estimators (Morgan and Rubin, 2012). From this expression together with Equations 1 and 3, it becomes clear that the variance reduction from Mahalanobis-based rerandomization relative to complete randomization is decreasing in p_a, the strictness of the rerandomization criterion, and non-increasing in K, the dimension of $\mathbf{X}$.

3. The special case of minimizing the variance

The mean difference *estimate* for sample s and allocation j is

$$\hat{\tau}_s^j = \overline{Y}_{1s}^j - \overline{Y}_{0s}^j \tag{6}$$

where $\overline{Y}_{1s}^j = \frac{1}{n_1}\sum_{i=1:i\in s}^{n_1} W_i^j Y_i(1)$ and $\overline{Y}_{0s} = \frac{1}{n_0}\sum_{i=1:i\in s}^{n_0}(1-W_i^j)Y_i(0)$. The sample average treatment effect for sample s is

$$SATE_s = \frac{1}{n}\sum_{i\in s}(Y_i(1) - Y_i(0)) = \frac{1}{A}\sum_{j\in\mathcal{A}}\hat{\tau}_s^j = \frac{1}{A_a}\sum_{j\in\mathcal{A}_a}\hat{\tau}_s^j,$$

where $A_a = \text{card}(\mathcal{A}_a)$ and thus estimators $\hat{\tau}_s^{CR}$ and $\hat{\tau}_s^{RR}$ are unbiased for $SATE_s$ [2, 5]. For sample s the variance of the estimators, $V_n(\hat{\tau}_s^{CR})$ and $V_n(\hat{\tau}_s^{RR})$, only depend on the treatment assignment mechanisms, i.e., the experimental designs.

For the super population, as assumed in [3],

$$PATE = E[Y(1) - Y(0)].$$

The variance of the estimators for inference to PATE are

$$V_n(\hat{\tau}^{CR}) = EV_n(\hat{\tau}_s^{CR}) + Var(SATE). \tag{7}$$

and

$$V_n(\hat{\tau}^{RR}) = EV_n(\hat{\tau}_s^{RR}) + Var(SATE),$$

respectively. The first term of the variance decomposition is the expected variance of the estimators, and the second term is simply the variance of

SATE across random samples. Clearly, only the first term differs between the two designs. From these results it follows that, in line with [3], an optimal rerandomization design for the inferences to PATE should minimize the first term in 7.

Using Mahalanobis-based rerandomization, the optimal design would be obtained by minimizing the variance in the observed covariates $\mathbf{X}$ (Equation 2), which is obtained by letting $a \equiv \min_{\mathcal{A}} M(\mathbf{W}^j, \mathbf{X})$, which, in large samples, implies $\nu_a \simeq 0$. With an additive treatment effect, this criterion implies that the the PRIV of $\widehat{\tau}^{RR}$ is equal to $100 \times R^2$.

3.1. Inference to SATE

Example 1 in Kallus [3] serves as an illustration for his unusual perspective on statistical inference. The example is used to illustrate Theorem 1 in [3, pp 89], or Theorem 1 in [8], stating that complete randomization is minimax if the covariates $\mathbf{X}$ are independent of the outcome.

In this example, X and Y are deterministically generated according to (using the notation of [3])

$$\begin{aligned}
\mathbf{X} \approx \text{round}(\{X_1, ..., X_n\}) &= \{-1, -2, -4, ..., -2^{2^{b-1}-1}, 1, 2, 4, ..., 2^{2^{b-1}-1}\} \\
Y_i(0) &= (-1)^i = (-1)^{\log_2(\text{round}(|x_i|)} - \tau/2 \\
Y_i(1) &= Y_i(0) + \tau
\end{aligned}$$

where $i = 1, ..., n$, $n = 2^b$ and $b \in \mathbb{Z}^+$.† Note that with this data generating process, $\mathbf{X}$ have different distributions for each n, i.e., a larger n does not only give a larger sample but the larger sample arises from a different distribution with larger variance.

The variance of the mean difference estimator under complete randomization for this data generating process can be shown to be $V_n(\widehat{\tau}^{CR}) = V_n(\widehat{\tau}_s^{CR}) = 4/(n-1)$, see the Appendix for details. In the comparison of $V_n(\widehat{\tau}^{CR})$ against the variance under Mahalanobis-based rerandomization, [3] only includes $\mathbf{X}$ (no transformations) in the Mahalanobis distance criterion and uses the minimal limiting acceptance criterion, i.e., $p_a = 0$, by restricting variance calculations to allocations with $M^j = 0$ only. Let

†Our understanding is that the results in [3] are based on rounded $\mathbf{X}$ and not on the specified data generating process for $\mathbf{X}$. Due to the ambiguous notation in [3], we were also not able to recreate the suggested data generating process for $\mathbf{X}$.

$V_n(\hat{\tau}^{RR}|p_a = 0)$ be the variance of the mean difference estimator given the sample size n and with $M^j = 0$.

[3] incorrectly states that $V_n(\hat{\tau}^{RR}|p_a = 0) \equiv 4$ for all n. This mistake stems from the incorrect assumption that the allocation $\mathbf{W} = (0, 1, 0, 1, ..., 0, 1)'$ *uniquely* minimizes the Mahalanobis distance for all n, and therefore the variance calculations are, incorrectly, based solely on this allocation. In fact, an experiment with $n_1 = n_0$, using the Mahalanobis distance balance measure, $\text{card}(\mathcal{A}_{Opt}) \geq 2$, i.e., *there exist at least two allocations, not a unique smallest Mahalanobis distance.* This follows because, using the Mahalanobis distance to measure balance, every allocation always has a mirror allocation with 1's and 0's exchanged but with the same imbalance. Thus, the minimum number of allocations with the smallest imbalance is two (a pair of mirror allocations).

As is shown in the Appendix, with $n = 2$ the only pair of allocations have $M^1 = M^2 = 4$, and therefore $V_2(\hat{\tau}^{RR}|p_a = 0)$ is not defined for $n = 2$. With $n = 4$ there exists one pair of allocations with $M^j = 0$, for which $V_4(\hat{\tau}^{RR}|p_a = 0) = 4$. For $n > 4$ there is more than one pair of allocations that have Mahalanobis distances equal to zero, and the variance decreases in n. With $n = 8$ and $n = 16$, $V_8(\hat{\tau}^{RR}|p_a = 0) = 4/3$ and $V_{16}(\hat{\tau}^{RR}|p_a = 0) = 4/7$, respectively. The convergence rate is, thus, slower than for $V_n(\hat{\tau}^{CR})$ but it is in contrast to the claim $V_n(\hat{\tau}^{RR}|p_a = 0) \equiv 4$ for all n.

In this special case, the only source of randomness is the randomization mechanism for assigning treatment. Under a sharp null hypothesis, e.g., the treatment effect is zero for all units, the value of any test-statistic is known for all possible random allocations. The exact p-value associated with any test statistic is simply the percentile of the observed test-statistic in the histogram of the statistics' value across all possible allocations. This implies that theoretical asymptotic variances are only helpful tools for comparing efficiency in designs where treatment assignment is randomized within a set of allocations sufficiently large for drawing such inferences. In this case, for inference with level $\alpha = 0.05$, comparisons of variances is only meaningful with $n \geq 8$. With $n = 8$ it follows that $\text{card}(\mathcal{A}) = \binom{8}{4} = 70$, which implies that the lowest possible p-value in a two-sided test is less than 0.05. A variance comparison between, e.g., rerandomization and complete randomization should be performed for a rerandomization design where a is chosen such that $2/\text{card}(\mathcal{A}_a) \leq \alpha$, where α is the desired level of the inference. This is the reason for why Morgan and Rubin (2012) states that one should not use too small a value of p_a.

Kallus [3, p. 94] refers to the Morgan and Rubin (2012) rerandomization design as the "historically haphazard practice of rerandomization". To the best of our understanding, the argument for this statement seems to be based on the belief that the Mahalanobis-based rerandomization minimizes the linear projection of $\mathbf{Y}$ on $\mathbf{X}$ due to a structural assumption of a linear relation between $\mathbf{Y}$ on $\mathbf{X}$. However, as pointed out in [5], by including interactions and polynomials of covariates in the Mahalanobis distance, non-linear dependencies can be considered also in Mahalanobis-based rerandomization. To exemplify the potential importance of including relevant transformations, we introduce both X and X^2 into the Mahalanobis distance criterion in the example above.

With both X and X^2 in the Mahalanobis distance criterion, there are no allocations with $M^j = 0$ for $n \leq 16$, and the optimal criterion is instead $p^* \equiv \min p_a : \text{card}(\mathcal{A}_{Opt}) > 0$. With $p_a = p^*$ as the criterion, $\text{card}(A_{Opt}) = 2$ for $n \leq 16$, and for these single pairs of allocations, $V_2(\widehat{\tau}^{RR}|p_a = p^*) = 4$, $V_4(\widehat{\tau}^{RR}|p_a = p^*) = V_8(\widehat{\tau}^{RR}|p_a = p^*) = V_{16}(\widehat{\tau}^{RR}|p_a = p^*) = 0$. Thus, this single set of the 'optimal' design have smaller variance than all the other designs. However, restricting randomization to one single pair, implies that randomization inference has no power. To enable inference, the rerandomization criterion must be increased, allowing for 'non-optimal' allocations. For example, when the inclusion criterion is set to $p_a = 0.1$, we get the number of allowed allocations to $2, 8$, and $1,288$ for $n = 4, 8$, and 16, respectively. For these values of n, we get $V_4(\widehat{\tau}^{RR}|p_a = 0.1) = 0$, $V_8(\widehat{\tau}^{RR}|p_a = 0.1) = 0.5$, and $V_{16}(\widehat{\tau}^{RR}|p_a = 0.1) = 0.216$. As $V_8(\widehat{\tau}^{CR}) = 0.571$ and $V_{16}(\widehat{\tau}^{CR}) = 0.267$ the PRIV for $n = 8$ and 16 is 12% and 19%, respectively. For $n = 16$ there are sufficient number of allocations $(1,288)$ for valid inferences to SATE and PATE with $p_a = 0.1$.

3.2. Inference to the Population

Li et al. [4] derives the asymptotic results for Mahalanobis-based rerandomization. It is shown that the asymptotic distribution of the SATE and PATE (under random sampling) estimators after rerandomization is generally non-normal. Instead, the asymptotic distribution is a linear combination of a normal distributed variable and a truncated normal variable. Furthermore the asymptotic sampling variances and quantile ranges of the mean difference estimator are reduced relative to when this the estimation is based on complete randomization.

Kallus (2018) exemplifies the usefulness of his algorithms by comparing the empirical variances of the different designs using a series of examples with simulated and real data. Given that the asymptotic distributions were not known to Kallus (2018), except for the Mahalanobis criterion, it is not obvious that these comparisons of empirical variances are valid procedures for evaluating the relative efficiency in practice. Furthermore, for the Mahalanobis distance metric, Kallus (2018) only allows the raw covariates in the Mahalanobis criterion. It is likely that by allowing for interactions and polynomials of the covariates, the variances under the Mahalanobis-based rerandomization would have been reduced as was the case in the example above.

[7] shows that when the experimental units are randomly sampled from the population, it is possible to draw inference to the units of the population when choosing the best pair of allocations despite no possibility of drawing inference to the SATE (as illustrated in the example above). Moreover, if the Mahalanobis criterion is used to find the best allocations, the asymptotic sampling distribution is known. However, in the real world, the situation with randomly sampled experimental units is rare with people. Most often in experiments on people, they choose whether or not to participate or they are selectively chosen, which means that when valid inference is the goal, it is a bad idea to choose the best pair of allocations because it only reflects uncertainty from potential random sampling. Instead, designs 'optimal' for the inference to the units of the sample described earlier, should also be used when making inference to the units of the population in the exceptional case of random sampling of the experiment. In other words, the ability to introduce a fully known stochastic mechanism in the design, on which exact inference can always be based, should not be sacrificed for the usually small, often negligible, gain in efficiency achieved by choosing the best pair of allocations, rather than choosing randomly from a small set of the nearly best allocations.

The next section provides a simple simulation study to emphasis two simple points. First, it illustrates the problem with using standard asymptotic inference results with a rerandomization design. Second, it illustrates the gain in power when using more appropriate techniques. The simulation is not designed to be a complete investigation of these two points.

Table 1. The factors of the Monte Carlo simulation study and their corresponding levels.

Factor	Levels
Design and Inference	Complete-Student's t,Rerandomized-Student's t, Rerandomized-Exact, Rerandomized-LDR
n	50, 100, 200 400
Treatment effect	0; 0.3 (standard deviations of Y)
X-Distribution	Normal; Exponential
Coefficent of determination, R^2	0.2; 0.5.

Notes: The exact p-value is approximated by Monte Carlo when calculated for complete randomization. LDR is the asymptotic inference under rerandomization [4]

4. Monte Carlo Simulation Study

This section compares the 'small' sample performance under complete randomization and Mahalanobis-based rerandomization. With the rerandomization, we use (i) standard asymptotic inference (standard student's t-statistic), (ii) exact inference, and (iii) the [4] (henceforth LDR) asymptotic inference.

The study can be compactly described as a $4 \times 2 \times 2 \times 4 \times 2$ factorial study with one summary value (the percentage of replications rejecting the null) in each cell of the 128 cells. The factors in this study are given in Table 1. The first factor is the type of design and inference, which compares the procedures of interest, whereas factors two to five are the parameters in the data generating process (DGP), all of which are of least partly unknowns to the investigator.. Data are generated as

$$Y_i(0) = x_{i1} + x_{i2} + x_{i3} + \epsilon_i \tag{8}$$

where $x_{ij}, j = 1, 2, 3$ and $i = 1, ..., n$ are independent and identically (iid) and either normally distributed with mean 0 and variance 1 (i.e., $x_{ij} \sim N(0, 1), \forall j, i)$ or exponentially distributed with rate 1 (i.e., $x_{ij} \sim \exp(1), \forall j, i)$. The sampling error, $\epsilon_i, i = 1, ..., n$, is iid and $\mathcal{N}(0, \sigma^2)$, where σ^2 is chosen to obtain $R^2 = 0.2$ and 0.5 in the data generating process (8). The sample size is varied and set to $n = 50, 100, 200$ and 400, because we are focused on how quickly the correct asymptotics are achieved.

The difference in performance of factor 1 is examined by testing for differences in outcome means of the active and controls under null and under alternative hypotheses using the nominal level of 5%. The effect is generated through the potential outcome as

$$Y_i(1) = Y_i(0) + 0.3\frac{\sqrt{50}}{\sqrt{N}}\sqrt{\frac{Var(\mathbf{Y}(0))}{N}},$$

For $N = 50$ this implies a treatment effect of 0.3 standard deviations which is often considered a moderately strong effect [1], for $N > 50$ the effect decreases with rate $\sqrt{N}$.

For the rerandomization inference, the Mahalanobis distance criterion, p_a, is set to 0.01, i.e., only allocations belonging to the 1% allocations with the smallest Mahalanobis distances are included. The p-value distribution from the Fisher exact test is inversely related to the number of unique values of the test statistic over all allocations. We randomly sampled 800 allocations fulfilling the criterion, without replacement, which means that, given no ties of the test statistic, the lowest possible p-value of the two-sided hypothesis test is 0.0025 (= 2/800). The treatment allocation used is randomly drawn from these 800 possible allocations. With $p_a = 0.01$ and $K = 3$ $\nu_a = 0.023$, the expected reduction in sampling variance of the $\widehat{\tau}$ from rerandomization are 19.5% and 48.8% with $R^2 = 0.2$ and 0.5, respectively (see section 2 for detials).

The number of replications is 2,000 for the $n = 400$ sample. For each simulated replicate, the sample is randomly split into two samples of size $n = 200$, four samples of size 100 and eight samples of size 50, which implies essentially the same sampling variation for all levels of n. In addition, because the treatment effect is homogeneous, this simulation design implies the same sampling variation within the eight ($n\times$treatment effect) cells. Furthermore, to reduce the sampling variance across the factor R^2, we add a sampling error when generating data under $R^2 = 0.2$ to the sampling error used to generate data under $R^2 = 0.5$.‡ Adding two independent sampling errors increases the variation from sampling between the two levels of this factor. With two different designs on the distribution of the covariates and with four different designs this leaves us with the total of 128 cells with a minimum of Monte Carlo sampling variability.§

The complete results from the experiment are displayed in Table 2

‡To be clear, let the DGP under factor 1 (i.e, $R^2 = 0.5$) be $Y_i^1(0) = x_{i1} + x_{i2} + x_{i3} + \epsilon_{i1}$, then the DGP under factor 2 (i.e., $R^2 = 0.2$) is $Y_i^2(0) = Y_i^1(0) + \epsilon_{i2}$

§For a detailed discussion on the design of efficient Monte Carlo simulation see Rubin [6].

and Figure 1. Table 2 displays the level of the test statistics of each of the 64 cells of the experiment. The top panel displays the level when $R^2 = 0.2$ and the second when $R^2 = 0.5$ across the two distributions of the covariates. The columns display the proportion rejected in the R replications for the four procedures. The results are very similar across distributions of the covariates and across the values of R^2. As expected, the proportion rejected (i.e. real level) is not statistically different from the nominal level with complete randomization. For $n = 50, 100, 200$ and 400 the 95% confidence intervals are [4.66 -5.34], [4.52-5.48], [4.32-5.67], and [4.05-5.95], respectively. The same test with rerandomization is displayed in column (2) from which we can see, what was stated in Morgan and Rubin (2012), that the inference is highly conservative. However, as can be seen from column (3) the proportion rejected is not statistically different from the nominal level using the Fisher exact test and when using the correct asymptotics for $n = 400$ (see column (4)). Furthermore, already with $n = 100$ the level of the asymptotic test is very close to the nominal level.

Figure 1 displays the results of the power comparisons of the four inference strategies. Instead of displaying the results from all 64 cells we, for clarity, show the percentage improvements in the proportion rejected as compared to the proportion rejected using the standard t-test under complete randomization. Using the correct inference procedures, we can see that for all cells, the proportion rejected is higher with rerandomization than with complete randomization. Due to the slight size distortion for the LDR test for n less than 400 the proportion rejected is not a measure of power for n less than 400. However, for the exact test, the proportion rejected is a measure of power for all n. From the lower right panel of the figure we can see that with $n = 50$, and $R^2 = 0.5$ there is an almost 75% increase in power with normally distributed covariates. The lower left panel shows that the power increases to around 60% with exponential distributed covariates. The power is, as expected, decreasing in R^2. With $R^2 = 0.2$ we see an about $20 - 25\%$ improvement in power for the two corresponding cells (upper right and upper left panels).

Table 2. Empirical size of the test of mean difference using the nominal level 5%. Asymptotic tests in columns (1), (3) and (4) and Monte Carlo approximated exact tests in columns (3).

		Complete	Rerandomized	Rerandomized	Rerandomized
Dist	n	Students t	Students t	Exact	LDR
		(1)	(2)	(3)	(4)
		$R^2 = 0.2$			
Normal	50	5.2	2.5	5.2	7.6
	100	5.0	2.3	4.8	5.8
	200	4.9	2.4	5.2	5.7
	400	4.7	2.6	5.5	5.6
Exponential	50	4.6	2.5	5.0	7.3
	100	4.9	2.4	4.9	5.9
	200	5.5	2.9	5.4	6.1
	400	5.2	1.9	4.6	4.8
		$R^2 = 0.5$			
Normal	50	5.0	0.7	5.0	7.4
	100	4.9	0.6	5.1	6.2
	200	5.4	0.6	5.3	5.6
	400	5.0	0.5	4.2	4.8
Exponential	50	4.6	0.7	5.0	7.1
	100	5.2	0.7	5.2	6.2
	200	4.7	0.6	5.2	5.6
	400	4.5	0.4	4.6	4.6

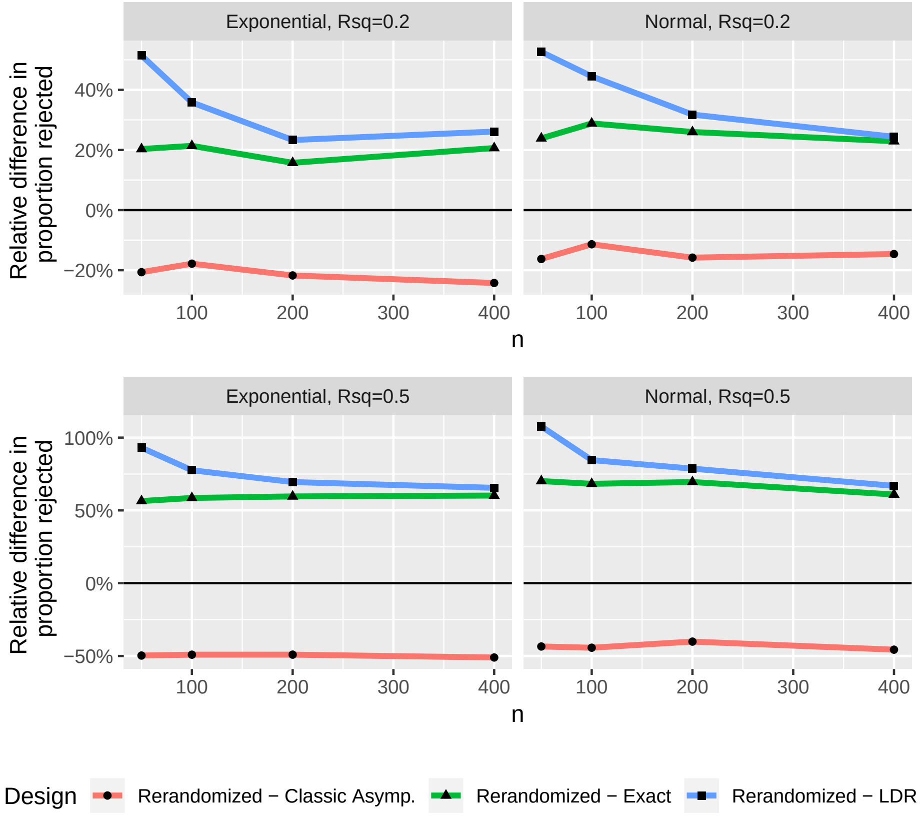

Figure 1. The relative difference in proportion of replicates rejecting the null as compared to complete randomization with asymptotic inference, given for rerandomization with different inference. Note that the rows have different y-axis.

5. Discussion

This paper discusses the implication of an ‘optimal’ rerandomization design for inferences to the Sample Average Treatment Effect (SATE) or the Population Average Treatment Effect (PATE). The idea with rerandomization is to remove, from consideration, allocations with imbalance in observed covariates between treated and control units and then randomize within the set of allocations with balance on these covariates; call the set of all allocations $\mathcal{A}$ and the set of acceptable allocations $\mathcal{A}_a$. Morgan and Rubin [5] were first in formalizing a procedure for rerandomization by suggesting the Mahalanobis distance as the criterion for defining $\mathcal{A}_a$ for the inferences to the SATE. Kallus [3] suggests alternative criteria with the aim of finding the 'optimal' design for the inferences to the PATE.

One of the advantages with the Mahalanobis criterion is that the SATE estimator (i.e., mean difference of treated and control units outcomes) is asymptotic unbiased with a known distribution ([4]). [3] optimal designs is obtained by minimizing the maximum conditional variance of the mean difference estimator under the assumption that conditional means of the outcomes under treatment and control can be estimated. Under this assumption the *pure strategy optimal design* (PSOD) enables finding a single optimal allocation, that is, the allocation in the second stage is deterministic. This means that all variation in treatment allocation stems from the random sampling conditional on the included covariates.

We show that for inference to SATE the cardinality of $\mathcal{A}_a$, card$(\mathcal{A}_a)$, should be large enough to allow the exact Fisher randomization test (FRT) to have power. A variance comparison between, e.g., rerandomization and complete randomization is only meaningful when $2/\text{card}(\mathcal{A}_a) \leq \alpha$, where α is the desired level of the inference. For asymptotic inference to SATE, the only rerandomization criterion, to the best of our knowledge, for which the asymptotic sampling distribution has been derived is the Mahalanobis criterion. With random sampling to the experiment asymptotic inferences to the PATE is in theory possible when card$(\mathcal{A}_a) = 1$. However, again the only criterion with known sampling distribution is the Mahalanobis criterion. Also, it is in general a bad idea to use rerandomization designs with a minimum rerandomization criterion and/or select the final assignment deterministically, as suggested in [3], as such designs only reflect uncertainty from potential random sampling. Instead, designs 'optimal' for the inference to SATE should also be used for inferences to the units of the population. In other words, the ability to introduce a fully known stochastic mechanism in the design, under which exact inference can always be based, should not be sacrificed for the, often negligible, gain in efficiency achieved by choosing the best allocation(s), rather than choosing randomly from a smaller set of the nearly best allocations, as implied by a well-chosen rerandomization criterion.

As a illustration of the problem with standard asymptotic theory with rerandomization designs and the potential with the Fisher randomization test and correct asymptotics ([4]) a small Monte Carlo study is conducted. We find that given correct inferential methods substantial gains in power can be made using rerandomization in comparison to complete randomization

References

[1] Cohen, J. (1992) A Power Primer. *Psychological Bulletin*, **112**, 155–9.

[2] Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

[3] Kallus, N. (2018) Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **80**, 85–112.

[4] Li, X., Ding, P. and Rubin, D. B. (2018) Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences*, **115**, 9157 – 9162.

[5] Morgan, K. L. and Rubin, D. B. (2012) Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, **40**, 1263–1282.

[6] Rubin, D. B. (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, **74**, 318–328.

[7] Schultzberg, M. and Johansson, P. (2019) Asymptotic inference for optimal rerandomization designs. Working paper.

[8] Wu, C.-F. (1981) On the Robustness and Efficiency of Some Randomized Designs. *The Annals of Statistics*, **9**, 1168–1177.

A. Appendix - Details of example 1

A.1. The variance under complete randomization

Due to the homogogenous treatement effect

$$V_n(\widehat{\tau}^{CR}) = \frac{S^2_{Y(1)}}{n_1} + \frac{S^2_{Y(0)}}{n_0},$$

where $S^2_{Y(w)} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(w) - \overline{Y}(w))^2$ and $\overline{Y}(w)$ the mean given w. As $Y_i(0) = (-1)^i$ and $Y_i(1) = Y_i(0) + \tau$ is

$$S^2_{Y(0)} = \frac{1}{n-1}\sum_{i=1}^{n}((-1)^i)^2 = \frac{n}{n-1}$$

and

$$S^2_{Y(1)} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(0) + \tau - \tau)^2 = S^2_{Y(0)}.$$

As $n_1 = n_0 = n/2$ and by the fact that $S^2_{Y(1)} = S^2_{Y(0)}$ we get

$$V_n(\widehat{\tau}^{CR}) = \frac{4}{n-1}.$$

A.2. The variance under Mahalanobis based rerandomization

Using

$$V_n(\widehat{\tau}^{RR}_s) = \frac{1}{A_a}\sum_{j\in\mathcal{A}_a}(\widehat{\tau}^j_s - \tau)^2,$$

we can calculate the variance under Mahalanobis based rerandomization.

With $n = 2$ it follows that $Y(0) = (-1, 1)$, $Y(1) = (1 + \tau, -1 + \tau)$ and $X = (-1, 1)$. The Mahalanobis distances between the covariate means for these two allocations are $M^1 = M^2 = 4$. This means that $V_2(\widehat{\tau}^{RR}|p_a = 0)$ has no solution. Let $\widehat{\tau}_s = (\widehat{\tau}^1_s, ..., \widehat{\tau}^{N_A}_s)'$ and $\mathbf{M} = (M^1, M^2, ...M^{N_A})'$. With $n = 4$, $N_A = 6$, $Y(0) = (-1, 1, -1, 1)$, $Y(1) = (-1+\tau, 1+\tau, -1+\tau, 1+\tau)$ and $X = (-1, -2, 1, 2)$. Here $\widehat{\tau}_s = (\tau, -2 + \tau, \tau, \tau, 2 + \tau, \tau)'$ and $\mathbf{M} = (3.6, 0, 0.4, 0.4, 0, 3.6)'$. Thus for $n = 4$ we have

$$V_4(\widehat{\tau}^{RR}|p_a = 0) = \frac{1}{2}\left((-2)^2 + 2^2\right) = \frac{8}{2} = 4 = V_2(\widehat{\tau}^{CR}_s)$$

Using the same definition for the variance under complete randomization we get

$$V_4(\widehat{\tau}^{CR}_s) = \frac{1}{6}\left(0^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 0^2\right) = \frac{8}{6} = \frac{4}{n-1}.$$

For $n = 8$ $(b = 3)$, $Y(0) = (-1, 1, -1, 1, -1, 1, -1, 1)$, $Y(1) = (1 + \tau, -1+\tau, 1+\tau, -1+\tau, 1+\tau, -1+\tau, 1+\tau, -1+\tau)$ and $X = (-1, -2, -4, -8, 1, 2, 4, 8)$.

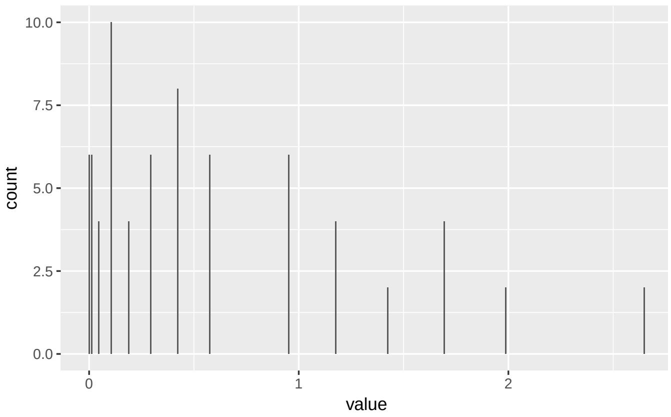

Figure 2. Mahalanobis distance of the 70 allocations.

There are 70 possible treatment allocations, not listed here with either $(\widehat{\tau}_s^j - \tau) = 0$, ± 1, or ± 2. Figure 2 displays the distribution of the $\mathbf{M}$ for all allocations. There are 6 allocations for which $M^j = 0$, which gives

$$V_8(\widehat{\tau}^{RR}|p_a = 0) = \frac{1}{6}\left(0^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 0^2\right) = \frac{4}{3} = V_4(\widehat{\tau}_s^{CR}).$$

For $n = 16$ we get $V_{16}(\widehat{\tau}^{RR}|p_a = 0) = 4/7 = V_8(\widehat{\tau}_s^{CR})$. That is, the variance under rerandomization with the minimum criterion decreases with n, as illustrated here for $n = 4, 8, 16$.