Research Article

Mårten Schultzberg* and Per Johansson

Asymptotic Inference for Optimal Rerandomization Designs

Abstract: Recently a computational-based experimental design strategy called rerandomization has been proposed as an alternative or complement to traditional blocked designs. The idea of rerandomization is to remove, from consideration, those allocations with large imbalances in observed covariates according to a balance criterion, and then randomize within the set of acceptable allocations. Based on the Mahalanobis distance criterion for balancing the covariates, we show that asymptotic inference to the population, from which the units in the sample are randomly drawn, is possible using only the set of best, or 'optimal', allocations. Finally, we show that for the optimal and near optimal designs, the quite complex asymptotic sampling distribution derived by Li et al. (2018), is well approximated by a normal distribution.

Keywords: computational-based design, experimental design, Mahalanobis distance, normal approximation, randomized control trail

1 Introduction

In randomized experiments, the treatment assignment is unconfounded (Rubin 1978; Imbens and Rubin 2015) or equivalently strongly ignorable (Rosenbaum and Rubin 1983), which permits valid inference to a large collection of well-defined estimands. For this reason, among others, randomized experiments are seen as the gold standard for causal inference. However, in a single realized randomized allocation, the imbalance in both observed and unobserved covariates can be substantial. This can lead to poor precision and low efficiency in the inference. To reduce these potential imbalances, blocking (also called stratification) on observed covariates has been used in experimental design, especially with a few discrete covariates. An alternative, or complement, to blocking that has received attention lately is to utilize modern computational capabilities in finding the experimental design (see e.g., Morgan and Rubin (2012); Bertsimas et al. (2015); Kallus (2018); Lauretto et al. (2017); Krieger et al. (2019); Kapelner et al. (2020); Johansson and Schultzberg (2020)). The idea in all the above strategies is to use the computer to discard allocations with imbalance in the observed covariates or, alternatively, to find allocations with balance in observed covariates. The paper by Kallus (2018) proposed algorithms for finding 'optimal' designs for the estimation of the population average treatment effect (PATE).

Bertsimas et al. (2015); Lauretto et al. (2017) are not discussing inference in their designs. Morgan and Rubin (2012); Krieger et al. (2019); Kapelner et al. (2020); Johansson et al. (2020) all used Fisher randomization tests to test for an average effect among the units in the experiment while Kallus (2018) suggest using bootstrap inference to the PATE but without any formal proof of its validity.

This aim of this paper is to clarify the concept of an 'optimal' design for *inferences* to the PATE based on the mean-difference estimator. In addition we suggest a strategy for asymptotic inference to the PATE in an optimal design based on the Mahalanobis-based rerandomization strategy suggested in Morgan and Rubin (2012). Here, the operating characteristics, i.e., the expectation, variance, and asymptotic distribution of the mean-difference estimator are known (Li et al. 2018); the 'optimal' designs suggested by Kallus (2018)

---
*Corresponding Author: Mårten Schultzberg: Uppsala University, Sweden; Email: marten.schultzberg@statistik.uu.se
Per Johansson: Uppsala University, Sweden; Tsinghua University, China; Email: per.johansson@statistik.uu.se

includes rerandomization based on the Mahalanobis distance as a special case when the relation between the covariates and the potential outcomes is linear (Kallus 2018, Section 2.3.3).

Under complete randomization the mean-difference estimator is an unbiased estimator of the sample average treatment effect (SATE), that is an average effect among the units in the experiment. Under random sampling of units from the population to the experiment the estimator is also an unbiased estimator of the PATE. Standard normal asymptotic inference are possible to either SATE and PATE.

With the computer based designs the asymptotic distribution of the mean-difference estimator will not in general be normal. In order to provide an intuition to the problem with (asymptotic) inference in these designs we need to discuss the underlying assumptions used to derive the asymptotic distributions for inferences to SATE and PATE, respectively. The normal asymptotic inference to SATE is derived using finite sample central limit theorems (see e.g. Li and Ding (2017)). The underlying assumption is that of replicated randomization and that the sample size n goes to infinity. The only thing stochastic is thus the allocation of treatments. In an optimal design the number of allocation is reduced to a minimum. The implication, is that we cannot derive the asymptotic distribution for the SATE estimand. For instance, in a deterministic design (i.e. one possible allocation) the resulting distribution has zero variance. Note that with a limited set of allocations a Fisher randomization test cannot either be conducted. The assumption deriving the asymptotic distribution of the mean-difference estimator as an estimator of PATE is that of random sampling to the experiment from a population of size N ($N > n$), or from a superpopulation. When interest is on conducting inference to PATE there is no lower limit on the number of possible allocations as the asymptotic distribution is derived under the assumption of random sampling to the experiment only. The consequence is that we, in theory, can have a deterministic design and then conduct inference to PATE, however no inference is possible regarding an effect in the experiment. This is an anomaly, but it is a consequence of the idea behind Neyman-Pearson inference.

We show that when the cardinality of the set of allocations fulfilling the Mahalanobis-distance covariance-balance criterion is close to its minimum, the asymptotic distribution of the mean-difference estimator for inference to the PATE is, as is the case with complete randomization, normally distributed with known variance. Furthermore, the difference in efficiency compared to using the 'optimal' set is typically very small which means that, using a slightly larger 'near optimal' set, admits non-degenerate inference to both SATE and PATE without substantially decreasing efficiency of estimation to PATE. Lastly, the large sample asymptotic distribution of the mean-difference estimator is well approximated by a normal distribution also when a larger 'near optimal' set is used. That is, as long as the criterion is selected such that it is small enough according to evaluation strategies suggested in this paper, standard z-tests can be be applied for inference. The implication of this results is important as the asymptotic inference after Mahalanobis-based rerandomization is simplified in contrast to what is suggested in Li et al. (2018).

The next section discusses rerandomization using the Mahalanobis metric. Section 3 provides the main results concerning asymptotic inference and Section 4 concludes.

2 Rerandomization based on the Mahalanobis criterion

In line with Morgan and Rubin (2012), consider a trial with n units, with n_1 assigned to treatment and n_0 assigned to control. Let $W_i = 1$ or $W_i = 0$ if unit i is assigned treatment or control, respectively, and define $\mathbf{W} = (W_1, \ldots, W_n)'$. Furthermore let $\mathbf{x}_i$, $i = 1, \ldots, n$, be $K \times 1$ vectors of fixed covariates in the sample and let $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$. Define the sample covariance matrix

$$cov(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^2,$$

where $\bar{\mathbf{x}} = \frac{1}{n} \sum_i^n \mathbf{x}_i$. In a balanced experiment, i.e., $n_1 = n_0 = n/2$, there are $\binom{n}{n_1} = n_A$ possible treatment allocation (assignment) vectors, thus $\mathbf{W}^j$, $j = 1, \ldots, n_A$ and $\mathbb{W} = (\mathbf{W}^1, \ldots, \mathbf{W}^{n_A})$ the complete set of allocations. Note that this set by construction has the 'mirror property', i.e., it can be enumerated as a set of 'mirror allocations' $\mathbf{W}^j$ and $\mathbf{1} - \mathbf{W}^j$.

The Mahalanobis distance for allocation j is

$$M(\mathbf{W}^j, \mathbf{X}) = \frac{n}{4} \widehat{\tau}_X^{j\prime} cov(\mathbf{X})^{-1} \widehat{\tau}_X^j, \; j = 1, \dots, n_A$$

where

$$\widehat{\tau}_X^j = \frac{1}{n_1} \sum_{i=1}^{n_1} W_i^j \mathbf{x}_i - \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - W_i^j)\mathbf{x}_i, = \overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j, \; j = 1, \dots, n_A.$$

Following Li et al. (2018), denote $\widehat{\tau}_X$ the estimator of the covariate mean-difference vector over complete randomization. Morgan and Rubin (2012) suggested accepting the treatment assignment vector $\mathbf{W}^j$ only when

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

where a is a positive constant.

Due to the asymptotic normality of difference in means implied by the Central Limit Theorem (CLT), $M(\mathbf{W}^j, \mathbf{X})$ follows a χ_K^2 distribution asymptotically. This means that a quantile of this distribution can be used as an allocation inclusion/exclusion criterion. Let $P(\chi_K^2 \leq a) = p_a$, then to randomize within the set of the 0.01% best balanced allocations implies setting a so that $p_a = 0.0001$.

The minimum number of allocations in the set of allocations with the smallest Mahalanobis distance (i.e. $M(\mathbf{W}^j, \mathbf{X}) \simeq 0$ or $p_a \simeq 0$) in an experiment with $n_1 = n_0$ is two.[1] Because the Mahalanobis distance of an allocation ($\mathbf{W}^j$) is always exactly the same as for its mirror $(1 - \mathbf{W}^j)$. This minimal set, containing only the allocations with the smallest Mahalanobis distance across all possible allocations, contains the optimal set in terms of covariate balance. Thus, by the mirror property, assuming at least one continuous covariate, there are always $\binom{n}{n/2}/2$ unique Mahalanobis values over the set of all $\binom{n}{n/2}$ allocations. This means that the large sample rerandomization criterion that gives the minimum set can be written as

$$a^{min} := \left\{ a : p_a^{min} = \frac{1}{\frac{\binom{n}{n/2}}{2}} \right\} = \frac{2}{\binom{n}{n/2}}.$$

We refer to this inclusion criterion as the 'best allocation inclusion criteria' (BAIC) and the set of allocations fulfilling BAIC is denoted $\mathcal{M}_{BAIC}$. However with ties in the minimum of the Mahalanobis distance, as could happen with discrete data, $\mathcal{M}_{BAIC}$ would have more than two elements. If $\mathcal{M}_{BAIC}$ is large enough, non-degenerate inference to the SATE is possible (Johansson et al. 2020). However, because the cardinality of $\mathcal{M}_{BAIC}$ is not restricted by design, BAIC does not generally allow for non-degenerate inference to SATE. If there are too few allowed allocations, probabilistic inference to SATE is not helpful (Johansson et al. 2020). Note that in a real experiment it will not be possible to find $\mathcal{M}_{BAIC}$ within a reasonable time limit. The reason is that the number of allocations is exponential increasing in n. Thus with $n > 30$ this is a NP-hard problem.

3 Asymptotic Theory

Let $Y_i(w)$ denote the potential outcome when unit i is exposed to w. The sample average treatment effect is defined as

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

Define the sample means $\overline{Y(w)} = \frac{1}{n} \sum_{i=1}^{n} Y_i(w)$, $w = 0, 1$, then the sample variance and the variance of the treatment effect are equal to

$$S_{Y(w)}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(w) - \overline{Y(w)})^2, w = 0, 1,$$

---

1 In an experiment with $n_1 \neq n_0$ the minimum set could consist of just one allocation. See (Morgan and Rubin 2012, p. 9) for an example with $n = 3$ and $n_1 \equiv 2$.

and

$$S_{Y(1)Y(0)} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(1) - Y_i(0) - (\overline{Y(1)} - \overline{Y(0)}))^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (\tau_i - \tau)^2 = S_\tau^2.$$

Note that these three variances are fixed in the sample. Define the sample means of treatments and controls

$$\overline{Y}_1 = \frac{1}{n_1} \sum_{i:W_i=1}^{n_1} Y_i \text{ and } \overline{Y}_0 = \frac{1}{n_0} \sum_{i:W_i=0}^{n_0} Y_i.$$

Under **SUTVA** (no interference between individuals and the same treatment, Rubin (1980)) the observed Y_i is equal to $Y(W_i)$, then the mean-difference estimator

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0,$$

is an unbiased estimator of τ (Neyman 1923).

Using CLTs where the n units in the sample is embedded into an infinite sequence of finite populations with increasing sizes one can show (see e.g Li and Ding (2017)) that under complete randomization

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V_\tau),$$

where

$$V_\tau = \frac{n}{n_1} S_{Y(1)}^2 + \frac{n}{n_0} S_{Y(0)}^2 - S_\tau^2. \tag{1}$$

The population average treatment effect (PATE) is defined as

$$\tau = \mu_1 - \mu_0,$$

where $\mu_w = E(Y(w))$, $w = 0, 1$. Given random sampling from the population the $\hat{\tau}$ is an unbiased and the third term in Equation (1) vanish (as treated and controls are sampled independently). Based on the standard CLT under random sampling

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V),$$

where $V = \frac{n}{n_1} S_{Y(1)}^2 + \frac{n}{n_0} S_{Y(0)}^2$.

For later use define $\mathbf{Y}(w) = (Y_1(w), \ldots, Y_n(w))'$ and $\tau = (\tau_1, \ldots \tau_n)'$ and let $S_{Y(w)|\mathbf{X}}^2$ and $S_{\tau|\mathbf{X}}^2$ be the variances of the linear projection of $\mathbf{Y}(w)$, $w = 0, 1$ and τ on $\mathbf{X}$, respectively. Using the same conditions as in Li and Ding (2017), Li et al. (2018) showed that the asymptotic distribution of $\hat{\tau}$ after randomly choosing an allocation from the set $\mathcal{A}_a = \{\mathbb{W}|M(\mathbf{W}^j, \mathbf{X}) \le a\}$ is

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \sqrt{V_\tau} Q, \tag{2}$$

where

$$Q = \sqrt{(1 - R^2)} \varepsilon_0 + \sqrt{R^2} L_{K,a}. \tag{3}$$

Here, ε_0 is a standard normal variable (for Y in the space orthogonal to the covariates), $L_{K,a}$ is the projection of Y into the space of covariates and is thus affected by the rerandomization, and

$$R^2 = \frac{\frac{n}{n_1} S_{Y(1)|\mathbf{X}}^2 + \frac{n}{n_0} S_{Y(0)|\mathbf{X}}^2 - \mathbf{s}_{\tau|\mathbf{X}}^2}{V_\tau}$$

Li et al. (2018) show that R^2 can be consistently estimated. Under homogeneous treatment effects, i.e., $S_\tau^2 = 0$ and $S_{\tau|\mathbf{X}}^2 = 0$, it follows that $R^2 = s_{Y(0)|\mathbf{X}}^2 / S_{Y(0)}^2$. This implies that R^2 can be estimated by a linear projection of the outcomes of the treated and control units values on $\mathbf{X}$. The second part of Q has the following distribution

$$L_{K,a} \sim \chi_{K,a} S \sqrt{\beta_K},$$

where $\chi_{K,a} = \chi_K^2 | \chi_K^2 \le a$, S a random variable taking values ± 1 with probability $1/2$, and $\beta_K \sim \beta(1/2, (K-1)/2)$ degenerating to a point mass at 1 when $K = 1$.

3.1 Asymptotic theory for the optimal design

It is only the second term of Equation (3), $\sqrt{R^2}L_{K,a}$, that depends on a, where $\sqrt{R^2}$ is a constant that can be estimated consistently. Thus, from the assumption that the n units are randomly sampled from the population, the asymptotic results when $n \to \infty$ in Li et al. (2018) can be used to derive the asymptotic distribution in the situation when a approaches 0.

$L_{K,a}$ is symmetric and unimodal around zero, with variance

$$\mathrm{Var}(L_{K,a}) = P(\chi^2_{K+2,a} < a)/P(\chi^2_{K,a} < a). \tag{4}$$

For fixed K, $\mathrm{Var}(L_{K,a}) \to 0$ as $a \to 0$. That is, the $L_{K,a}$ distribution will converge to point mass at zero when a, or equivalently p_a, goes to zero. This implies that for large n using the minimum criterion a^{min},

$$\sqrt{n}(\widehat{\tau} - \tau)|\widehat{\tau}_X \stackrel{d}{\to} \sqrt{V_\tau}\left(\sqrt{(1 - R^2)}\varepsilon_0\right). \tag{5}$$

The intuition of the result is that randomizing the treatment assignment within the set of allocations containing only the very best allocations will, in large samples, result in a realized treatment allocation with a Mahalanobis distance close to zero, which means that essentially all variation in $\widehat{\tau}$ that is explained by group differences in $\mathbf{X}$ is removed, that is, all the variance in the linear projection of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ on $\mathbf{X}$ is eliminated. Thus, non-degenerate inference to PATE is in general possible using BAIC given random sampling to the experiment.

Under Theorem 1 in (Li et al. 2018, p. 8) and using Equation (5), the test statistic is thus

$$Z = \frac{\widehat{\tau} - \tau}{\sqrt{\frac{V(1-R^2)}{n}}} \text{ for } \mathbf{W} \in \mathcal{M}_{BAIC} \tag{6}$$

where $Z \sim N(0, 1)$ and V is the variance under random sampling. All the well established asymptotic results for the standard normal distribution apply. The sampling distribution is well defined for $R^2 < 1$. In the situation where $R^2 = 1$ which is impossible in practice, i.e., that the variation in $\mathbf{X}$ explains *all* variation in both $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ and *all* variation in $\mathbf{X}$ is removed by rerandomization, the sampling distribution is degenerated to a point mass at zero.

For the inference to PATE the calculation of the R^2 is simplified as we can can neglect the heterogeneity in the calculation of the variance. We thus separately regress Y_i on $\mathbf{x}_i$ for the $W_i = 1$ and $W_i = 0$ samples, then $s^2_{Y(1)|\mathbf{X}}$ and $s^2_{Y(0)|\mathbf{X}}$ are estimated as

$$s^2_{Y_1|\mathbf{X}} = \frac{1}{n_1 - K} \sum_{i:W_i=1}^{n_1} (Y_i - \mathbf{x}'_i\widehat{\beta}_1)^2 \text{ and } s^2_{Y_0|\mathbf{X}} = \frac{1}{n_0 - K} \sum_{i:W_i=0}^{n_0} (Y_i - \mathbf{x}'_i\widehat{\beta}_0)^2,$$

where $\widehat{\beta}_w$, $w = 0, 1$ is the OLS estimates. The Neyman estimator is used to estimate V, thus

$$\widehat{V(\widehat{\tau})} = \frac{s^2_{Y_1}}{n_1} + \frac{s^2_{Y_0}}{n_0}, \tag{7}$$

where

$$s^2_{Y_1} = \frac{1}{n_1 - 1} \sum_{i:W_i=1}^{n_1} (Y_i - \overline{Y}_1)^2 \text{ and } s^2_{Y_0} = \frac{1}{n_0 - 1} \sum_{i:W_i=0}^{n_0} (Y_i - \overline{Y}_0)^2.$$

This means that

$$\widehat{R}^2 = \frac{\frac{n}{n_1}s^2_{Y_1|\mathbf{X}} + \frac{n}{n_0}s^2_{Y_0|\mathbf{X}}}{\widehat{V(\widehat{\tau})}},$$

and that the estimator of $V(1 - R^2)$ is

$$\widehat{V(\widehat{\tau})} - \left(\frac{n}{n_1}s^2_{Y_1|\mathbf{X}} + \frac{n}{n_0}s^2_{Y_1|\mathbf{X}}\right).$$

By Slutsky's theorem, plugging in these consistent estimators into Equation 6 yields a statistic that fully computable from the sample.

To illustrate that the sampling distribution can be approximated by a normal distribution when p_a is sufficiently small, a small Monte Carlo simulation is conducted. We test for the mean difference between treated and control with an equal number of treated and controls with a sample size $n = 12$, using the corresponding BAIC. This implies $p_a = 2/\binom{12}{6} \approx 0.002$. The sampling distibution of the test-statistics given in Equation (6) is compared to standard normal. We also calculate sampling distributions of the test statistic in Li et al. (2018)

$$\frac{\widehat{\tau} - \tau}{\sqrt{\frac{V_\tau(1-(1-v_a)R^2)}{n}}}, \tag{8}$$

where $v_a = \Pr(\chi^2_{(K+2)} \leq a)/\Pr(\chi^2_K \leq a)$. This distribution is compared to the Q-distribution.

Data are generated [2] as

$$Y_i = x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i} + \varepsilon_i, i = 1, \ldots, N \tag{10}$$

where $x_{ki} \sim N(2, 2)$, $k = 1, \ldots 5$ and $\varepsilon_i \sim N(0, 6)$. This implies $R^2 = 0.5$. 10,000 independent realizations of size $n = 12$ are generated from each distribution, including the theoretical ones. The true R^2 and V_τ are used to exclude the small sample sampling variation of the corresonding estimators, both shown to be consistent estimators in Li et al. (2018). Figure (1) displays the density plots of the empirical sampling distributions and
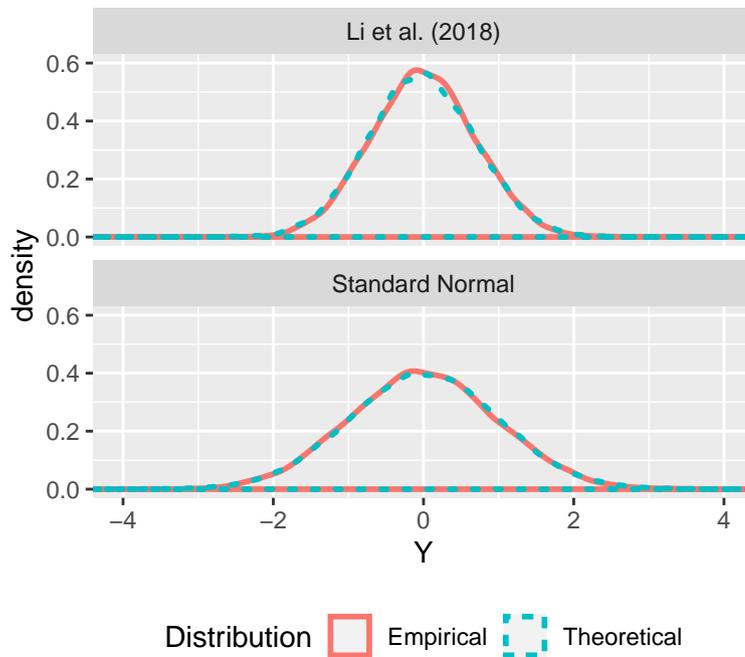
Figure 1: Empirical sampling distribution compared to the theoretical one implied by Li et al. (2018) and standard normal.

their corresponding theoretical distributions. It is clear that both empirical sampling distributions are well

---

2 One can think of the population regression as

$$y = \mathbf{z}\boldsymbol{\beta}_z \tag{9}$$

with $\mathbf{z}$ a $\varkappa \times 1$ vector of covariates. with a subspace of observed covariates (i.e. $\mathbf{x} \subset \mathbf{z}$). Let $\mathbf{x}^c$ be the complement such that $\mathbf{x} \cup \mathbf{x}^c = \mathbf{z}$ then $\varepsilon_i = \mathbf{x}_i^c \boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_x = (1, 1, 1, 1, 1)'$ in $y_i = \mathbf{x}_i \boldsymbol{\beta}_z + \varepsilon_i$. When using Mahalanobis-based rerandomization with BAIC, for large n, the variation in y that is common with $\mathbf{x}$ is removed.

described by their corresponding theoretical distributions. A one-sample Kolmogorov-Smirnov test for the empirical distribution of the statistic given in Equation 6 gives the test statistic D = 0.0096, with p-value = 0.3079 for a two-sided test against a standard normal distribution. It should be noted that in this case the covariates and error term was generated normal such that the asymptotic results are valid for small sample sizes in order to be able to do the simulation study, for non-normal covariates and error term these results require large n to be valid.

To summarize, Table (1) displays the sampling distribution of the $\hat{\tau}$ under Mahalanobis-based rerandomization for different regions of p_a. Under complete randomization ($p_a = 1$), the distribution is normal, for p_a in the interval $(0, 1)$, the sampling distribution is Q, and for BAIC ($p_a \to 0$) the sampling distribution is again normal but with a scaled variance. These results have the important implication that the large sample

Table 1: The sampling distribution of the $\hat{\tau}$ under under Mahalanobis-based rerandomization for different values of p_a.

	Complete rand.	Rerendomization	Optimal design
Reran. crit	$p_a = 1$	$0 < p_a < 1$	$p_a \to 0$
$Var(\sqrt{n}(\hat{\tau} - \tau))$	V_τ	$V_\tau(1 - (1 - v_a)R^2)$	$V(1 - R^2)$
$\sqrt{n}(\hat{\tau} - \tau) \sim$	$N(0, V_\tau)$	$Q(V_\tau, R^2, K, a)$	$N(0, V(1 - R^2))$
note: $v_a = Pr(\chi^2_{(K+2)} \le a)/Pr(\chi^2_K \le a);\ 0 < v_a < 1$			

sampling distribution under Mahalanobis-based rerandomization with sufficiently small p_a is simply normal with mean zero but with variance $V(1 - R^2)$ instead of V as under complete randomization.

As the inference is simplified when conducting inference to PATE given small enough a, a relevant question is whether we in the design phase can choose a p_a such that the (scaled) standard normal asymptotic inferences can be used in the analysis. The issue is discussed in the next section.

3.2 A Mahalanobis-based rerandomization design for simplified asymptotic inference

The variance of Q is equal to

$$(1 - R^2)Var(\epsilon_0) + R^2 Var(L_{K,a}) = (1 - R^2) + R^2 Var(L_{K,a})$$

where the equality comes from $\epsilon_0 \sim N(0, 1)$. Because the asymptotic variance of $L_{K,a}$ is known (cf. Equation 4), this allows us to calculate the relative importance of the second term which goes to zero (by Equation 4) for a specific R^2. The variance ratio (VR) of the second term to the overall variance of the estimator under Mahalanobis based rerandomization given R^2 and a, equals

$$VR = \frac{R^2 Var(L_{K,a})}{R^2 Var(L_{K,a}) + (1 - R^2)} = \frac{P(\chi^2_{K+2,a} < a)/P(\chi^2_{K,a} < a)}{P(\chi^2_{K+2,a} < a)/P(\chi^2_{K,a} < a) + (\frac{1}{R^2} - 1)}. \tag{11}$$

To illustrate, consider a setting with $R^2 = 0.2$ which is often realistic in practice, four covariates, and a small inclusion criterion, e.g. $a = 10^{-4}$ which corresponds to $p_a \approx 1.25 \times 10^{-9}$. It follows that

$$VR = \frac{Var(L_{4,10^{-4}})}{Var(L_{4,10^{-4}}) + (1/0.2 - 1)}$$

$$= \frac{\frac{P(\chi^2_6 \le 10^{-4})}{P(\chi^2_4 \le 10^{-4})}}{\frac{P(\chi^2_6 \le 10^{-4}))}{P(\chi^2_4 \le 10^{-4})} + 4} \tag{12}$$

$$\approx 4.16 \times 10^{-5},$$

which means that around 4.16/1000 of a percent of the variance of Q arises from the variance of the second term. Clearly, including or excluding this term has no practical importance for inferences in this case. For a

large sample, e.g., $n = 100$, there would still be around 1.26×10^{20} ($= 1.25 \times 10^{-9} \times \binom{100}{50}$) allocations fulfilling this rerandomization criterion. Thus choosing a to be 10^{-4} would also enable inference to the units of the sample while simultaneously allowing standard methods to make inference to the units of the population.

In real experiments, R^2 is not known in the design phase, which complicates this type of analysis. However, the VR can be calculated for various hypothetical values of R^2 that are larger than the expected empirical R^2 to create an upper bound for the VR. Figure 2 displays the VR as a function of a for $n = 100$, $K = 5$, and R^2 between 0.05 and 0.95. It is clear that, in this case, even if R^2 is large and the sample size is only moderately large, there are still enough allocations fulfilling the criterion. For example, if a is set to 10^{-4}, there are still 7.88×10^{20} allocations fulfilling the criterion. That this small sample size allows for the simplification even for very large R^2 illustrates that these results often can be applied in typical experimental settings.
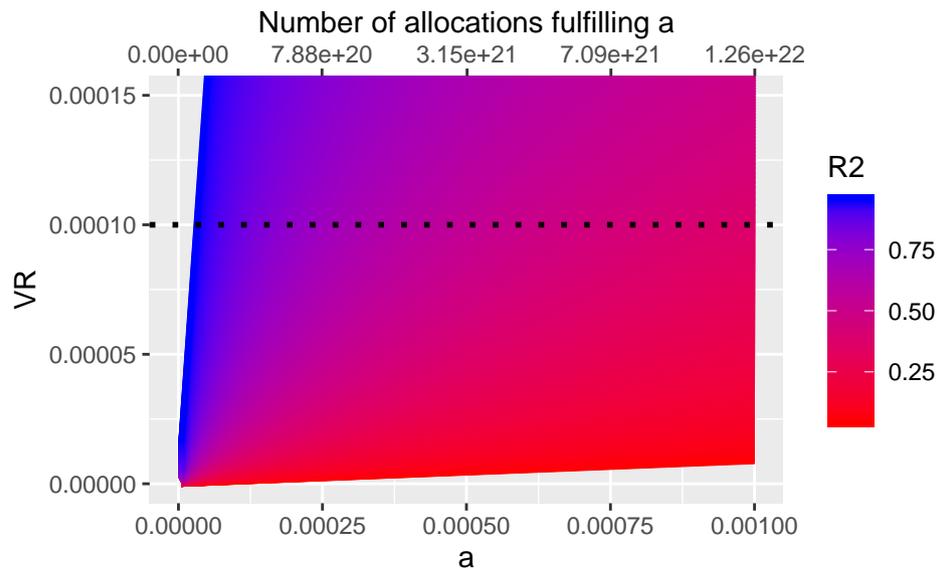
Figure 2: The variance ratio as a function of the Mahalanobis inclusion criterion a, for $K = 5$, $n = 100$ and different values of R^2. The values on the top of the abscissa are the number of allocations satisfying the criterion and the dotted line indicates the VR which implies that variance in the sampling distribution due to the second term in Equation 3 is smaller than 0.01%.

Based on a limited set of simulations, there seem to be no problem with inference, using the normal approximation, if the ratio of the variance (VR) in the sampling distribution due to the second term in Equation 3 is smaller than 0.01%. In the example above, using $VR < 0.01\%$ would allow for the (scaled) standard asymptotic inferences for all $R^2 < 0.95$. These simple calculations can be performed for any experiment to quickly evaluate the possibilities of using the simplified results presented in this paper, or, to select a such that, for the largest plausible R^2, the VR is smaller than the rule of thumb.

4 Discussion

The asymptotic sampling distribution for the mean-difference estimator, $\hat{\tau}$, for inferences to the PATE under Mahalanobis-based rerandomization design is investigated, specifically, how the sampling distribution is affected by letting the Mahalanobis-based rerandomization inclusion criterion approach the 'optimal' design.

The $\hat{\tau}$ under complete randomization is asymptotically normal. However, removing allocations associated with large covariates-differences between treated and controls, as with 'rerandomization designs', affects

the properties of this estimator in ways that depend on the balance criterion used to discard unbalanced allocations. Morgan and Rubin (2012) and Li et al. (2018) used the well known properties of the affinely invariant Mahalanobis distance to derive properties of the $\hat{\tau}$ after randomizing in the set of allocations with a Mahalanobis distance smaller than a specified inclusion criterion.

The $\hat{\tau}$ has a non-degenerate sampling distribution for repeated sampling inference to the population. However, deterministically choosing an 'optimal' design can lead to a degenerate sampling distribution for inference to the units in the experiment. Based on the results in Li et al. (2018), we show that the asymptotic sampling distribution under Mahalanobis-based rerandomization simplifies to a normal distribution when the inclusion criterion is small. When the sample size is moderately large, there will be large number of allocations that fulfills even very restrictive inclusion criteria, which enables inference to the PATE or the sample estimand (SATE) (Johansson et al. 2020).

For these reasons, when using Mahalanobis-based rerandomization, it can be advisable to set the inclusion criterion for admissible allocations slightly smaller than suggested in Li et al. (2018) so that standard asymptotic inference can be used for PATE, and slightly larger than the minimum so that inference is possible to both SATE and PATE. To this end we suggest a simple-to-use rule of thumb for when the simplified asymptotics can be used that makes it possible to choose the Mahalanobis criterion accordingly.

References

Bertsimas, D., Johnson, M. and Kallus, N. (2015), 'The power of optimization over randomization in designing experiments involving small samples', *Operations Research* **63**(4), 868–876. https://doi.org/10.1287/opre.2015.1361

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.

Johansson, P., Rubin, D. B. and Schultzberg, M. (2020), 'On Optimal Rerandomization Designs', *To appear in Journal of the Royal Statistical Society B* .

Johansson, P. and Schultzberg, M. (2020), 'Re-randomization strategies for balancing covariates using pre-experimental longitudinal data'. To appear in *Journal of Computational and Graphical Statistics /doi/full/10.1080/10618600.2020.1753531*.

Kallus, N. (2018), 'Optimal a priori balance in the design of controlled experiments', *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **80**(1), 85–112.

Kapelner, A., Krieger, A. M., Sklar, M., Shalit, U. and Azriel, D. (2020), 'Harmonizing optimized designs with classic randomization in experiments', *The American Statistician* **0**(0), 1–12. https://doi.org/10.1080/00031305.2020.1717619

Krieger, A. M., Azriel, D. and Kapelner, A. (2019), 'Nearly random designs with greatly improved balance', *Biometrika* **106**(3), 695–701.

Lauretto, M. S., Stern, R. B., Morgan, K. L., Clark, M. H. and Stern, J. M. (2017), 'Haphazard intentional allocation and rerandomization to improve covariate balance in experiments', *AIP Conference Proceedings* **1853**(June).

Li, X. and Ding, P. (2017), 'General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference', *Journal of the American Statistical Association* **112**(520), 1759–1769. https://doi.org/10.1080/01621459.2017.1295865

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**(37), 9157 – 9162.

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

Neyman, J. (1923), 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.', *Translated in Statistical Science(1990)* **5**(4), 465–472. http://www.jstor.org/stable/2245382

Rosenbaum, P. R. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rubin, D. B. (1978), 'Bayesian Inference for Causal Effects: The Role of Randomization', *The Annals of Statistics* **6**(1), 34–58.

Rubin, D. B. (1980), 'Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment', *Journal of the American Statistical Association* **75**(371), 591–593.