UPPSALA
UNIVERSITET

Working Paper 2018:4

Department of Statistics

Re-randomization strategies for balancing covariates using pre-experimental longitudinal data

Per Johansson and Mårten Schultzberg

Working Paper 2018:4
January 2019
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Working papers can be downloaded from www.statistics.uu.se

Title: Re-randomization strategies for balancing covariates using pre-experimental longitudinal data

Author: Per Johansson and Mårten Schultzberg
E-mail: marten.schultzberg@statistik.uu.se

Rerandomization strategies for balancing covariates using pre-experimental longitudinal data

Per Johansson*
Department of Statistic, Uppsala University
and
Mårten Schultzberg*
Department of Statistic, Uppsala University

January 24, 2019

Abstract

This paper considers experimental design based on the strategy of rerandomization to increase the efficiency in experiments. Two aspects of rerandomization are addressed. First, we propose a two-stage allocation sample scheme for randomization inference to the units of the experiments in balanced experiments that guarantees that the difference-in-mean estimator is an unbiased estimator of SATE for any experiment, conserves the exactness of randomization inference, and halves the time consumption of the rerandomization design. Second, we propose a rank-based covariate-balance measure which can take into account the estimated relative weight of each covariate. Several strategies for estimating these weights using pre-experimental data are proposed. Using Monte Carlo simulations, the proposed strategies are compared to complete randomization and Mahalanobis-based rerandomization. An empirical example is given where the power of a mean difference test of the electricity consumption of 54 households is increased by 99%, in comparison to complete randomization, using one of the proposed designs based on high frequency longitudinal electricity consumption data.

Keywords: Experimental design, Fisher exact test, High frequency Longitudinal data, Mahalanobis distance criterion, Mirror allocation sampling, Rerandomization

*The authors thanks Rauf Ahmad, Bengt Muthén and Mattias Nordin, seminar participants at the statistics department Uppsala University, and the audience at the The First Beijing Symposium in Biostatistics and Data Science November 16-17 2018.

1 Introduction

There is today a substantial literature on how to design randomized experiments to increase the efficiency as compared to complete randomization. All designs try to improve the similarity in the outcome of the groups of comparison by, in different ways, making the groups balanced in covariates that are observed before the experimental design is decided. The most common design used to improve balance is stratified or blocked randomization. The idea of stratified randomization is to divide units into strata (i.e. groups or blocks) based on similarity on covariates and then perform complete randomization within each strata. In this way, units from all strata will be represented in both the treatment and control groups[1] and thereby imbalance in any of these covariates are avoided, see e.g. Imbens and Rubin (2015) for a recent overview.

An alternative design is rerandomization which was originally suggested by Fisher in the early twentieth century but was first written up and implemented in Morgan and Rubin (2012). As the name suggests, rerandomization consists of redoing the randomization until some pre-specified balance criterion on the observed covariates is met. That is, the randomization is restricted to a subset of admissible allocations that fulfil the rerandomization covariate balance criterion. Based on the affinely invariant Mahalanobis distance covariate balance criterion, Morgan and Rubin (2012) show that rerandomization can decrease the variance in the effect estimate substantially as compared to complete randomization. Morgan and Rubin (2015) extend the results in Morgan and Rubin (2012) to deal with the case when large numbers of covariates are available and when some covariates can be a priori defined as more important than others.

The strategy of rerandomization is especially useful when continuous covariates are available, as in principle, even a single continuous covariate implies infinitely many strata and therefore must be discretized with information loss as a consequence. Compared to stratification, rerandomization is computationally demanding. However, with today's computers it provides a very interesting and powerful alternative design. As also Morgan and Rubin

[1]The concepts in this paper extend to any number of treatment groups. To simplify discussions, the number of treatment groups is restricted throughout this paper to two, referred to as treatment and control.

(2012) point out, rerandomization is not a design strategy that replaces stratification, rather a researcher should block on what covariates are possible and then use rerandomization on remaining covariates within these strata.

One potential caveat with rerandomization is that common test-statistics are no longer asymptotic normally distributed. As is shown in Li et al. (2018), the specific asymptotic distribution under rerandomization depends on which covariate balance measure is used. Li et al. (2018) derive the asymptotic sampling distribution of the estimated difference-in-mean estimator under rerandomization based on the Mahalanobis distance between the covariate-mean vectors of the treated and control in the potential allocations as the balance measure, as proposed in Morgan and Rubin (2012).

One appealing alternative to asymptotic inference was given in Morgan and Rubin (2012), namely, to restrict the inference to the units in the experiments and to base the analysis on exact (randomization) inference (Fisher, 1935). Due to the assumption-free nature of the randomization inference, this strategy is valid for all well-behaved (discussed in section 3) balance criteria. It is advantageous if the subset of allocations from all admissible allocations is of a moderate size when conducting the exact inference. A formal strategy of choosing the 'best' sub-set from the admissible is, to our knowledge, not available in the literature.

The contribution of this paper is two-fold. First, a sampling, or allocation, scheme for choosing the 'best' subset from admissible allocations is proposed. In addition to providing the exact level for the exact test, the sampling scheme: (i) guarantees that the difference-in-mean estimator is unbiased for the sample average treatment effect (SATE) and (ii) reduces the computational time for the design by half. Second, we develop a rerandomization covariate balance measure that is easy to use when pre-experimental outcome data (possibly high frequency longitudinal) are available, a situation that has not previously been addressed in the literature.

The paper should be of broad interest as the situation where the pre-treatment outcome is observed at many time periods is becoming more common. The last few decades' technological development of personal electronic devices like smart phones, smart watches,

fitness trackers, and the "Internet of Things", has made the collection of high frequency longitudinal data substantially simplified and cheaper. This development has also led to an increased interest in what kinds of research questions these data might help us answer, see e.g. Hamaker and Wichers (2017) and Hamaker et al. (2018). The present paper points out yet another possibility that these data brings, namely that of improving designs, enabling informative causal analysis also in relatively small experiments. In addition, we present practical guidelines for any rerandomization design that should be useful for any practitioner that want to use this strategy.

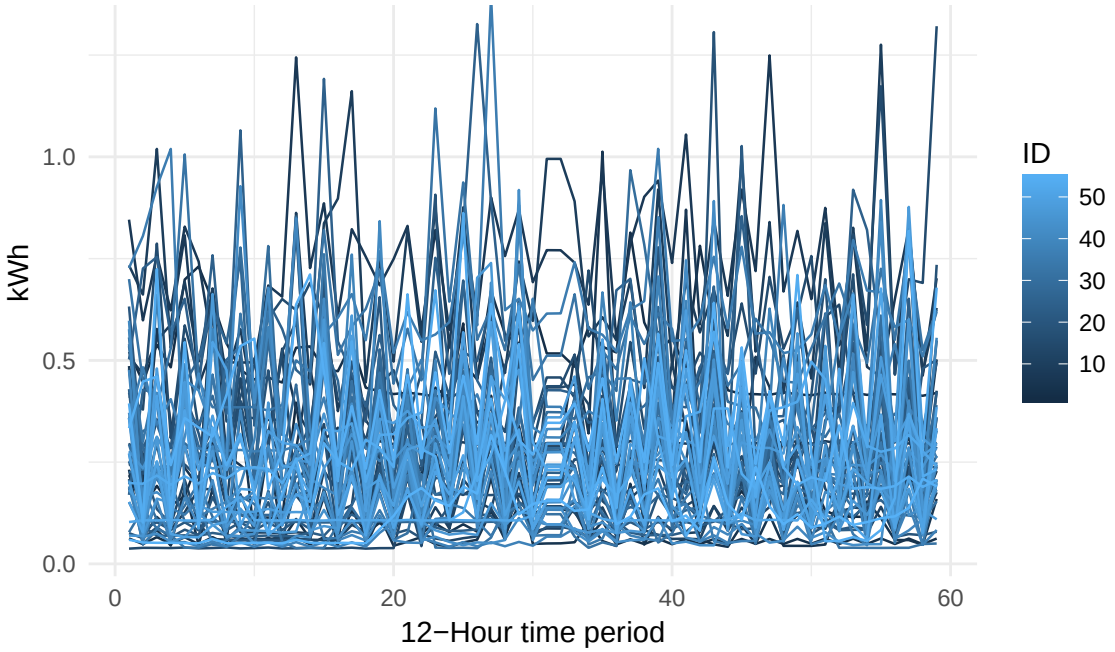

Figure 1: The electricity consumption (kWh) during the pre-treatment month for the 54 households included in the Experiment.

As a motivating example, data from an electricity consumption experiment are considered. In the design stage, repeated measurements of the outcome (kWh), displayed in Figure 1, are available for a sample of 54 households for the month before the treatment assignment. The average consumption for every 12 hour period, i.e. 60 measurements per month is observed. Clearly, there are large variations in several aspects (level, variation, etc.) of the consumption behavior the month before the treatment assignment. The

interest in this particular study is to see how user's electricity consumptions behavior is affected by information campaigns about money saving consumption behaviors. Using one of the strategies proposed in this paper, the power of a mean difference test with equal sized groups at the first time period after the pre-treatment period, is increased by 99% as compared to complete randomization. All the details of this study are presented in Section 5.

The rest of the paper is structured as follows. Section 2 introduces the concept of rerandomization and the Mahalanobis distance rerandomization criterion specifically. Section 3 introduces the formalized allocation sample scheme to select the subset of admissible allocations for exact inference. Section 4 introduces the new rerandomization covariate balance measure based on the ranks of the mean differences in the covariates, and gives Monte Carlo simulation comparisons to Mahalanobis-based rerandomization and complete randomization. Section 5 provides the empirical analysis of electricity consumption data illustrating the proposed procedure and balance measure. Section 6 contains a discussion and concludes the paper.

2 Stratification and Rerandomization

To provide a better understanding of the underpinnings of the rerandomization framework, it is here compared to complete randomization and classical stratified design.

The technical difference between complete and stratified randomization is that allocations that are possible in complete randomization are excluded in the stratified randomization. More precisely, the allocations from complete randomization associated with imbalances in the stratification covariates are excluded. For example, consider a study where two equal sized treatment and control groups will be compared. A sample of 10 males and 10 females are randomly sampled from the population. Under complete randomization there are $\binom{20}{10} = 184,756$ possible treatment allocations. If instead, randomization is stratified on sex, i.e. 5 males and 5 females are allocated to treatment the number of possible treatment allocations are reduced to $\binom{10}{5}\binom{10}{5} = 63,504$, i.e., the $121,252$ $(= 184,756 - 63,504)$ allocations that are unbalanced on sex are excluded. In order to stratify, the covariate

space needs to be partitioned into finite sets. With a set of a few categorical covariates this strategy is easy to implement, at least if there is more than one individual within each strata. However, as mentioned above, if one would like to include continuous covariates (e.g., pre-experiment outcomes) in the design, simple stratification methods run into problems as continuous covariates must be discretized (Hu and Hu, 2012).

Rerandomization is similar to stratified randomization in the sense that certain allocations are excluded, the main difference is the exclusion criterion. To formally introduce the concept of rerandomization, we start by describing the basic idea as it is outlined in Morgan and Rubin (2012) which also forms the basis for the extensions in Morgan and Rubin (2015), Zhou et al. (2018) and Li et al. (2018).

Consider a trial with N individuals of which $N/2 = N_1 = N_0$ are assigned to treatment and control, respectively. Let $\mathbf{x}$ be a fixed $N \times K$ covariate matrix, and let $W_i = 1$ if individual i is treated and $W_i = 0$ if not. Let $\mathbf{W}$ be the matrix of all $\binom{N}{N_1} = N_A$ possible random assignments (i.e. before treatment groups are assigned). For a given allocation $j, j = 1, ..., N_A$ the Mahalanobis distance between the covariate mean vectors of those assigned to treatment (T) and control (C), respectively, is defined as

$$M(\mathbf{x}, \mathbf{W}^j) = M^j = \frac{N}{4}(\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j)' cov(\mathbf{x})^{-1}(\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j),\ j = 1, ..., N_A, \tag{1}$$

where $\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j$ is the difference in mean vectors which is a $K \times 1$ stochastic vector as it depends on the allocation. Morgan and Rubin (2012) suggest randomizing within the set

$$\{\mathbf{W} | M(\mathbf{x}, \mathbf{W}^j) - a \leq 0\}, \tag{2}$$

where a is a constant. This means that instead of randomly choosing one of the N_A possible allocations, a smaller set of allocations with small Mahalanobis distances is considered. If the means are normally distributed then $M^j \sim \chi^2(K)$. This means that a can be indirectly determined by setting p_a in

$$p_a = \Pr(\chi^2(K) \leq a).$$

As the number of rerandomizations is geometrically distributed with expected value $1/p_a$, the expected number of rerandomizations before drawing a randomization fulfilling the criterion with, e.g., $p_a = 0.001$, is 1000. As this holds for any N, the time it takes to find

the allocations from which to finally make the randomization is independent of N for a fixed p_a.

If $\mathbf{x}$ is elepsiodially symmetric then, as a consequence of Mahalanobis metric being multivariate affinely invariant, the variation reduction is equally large for each covariate and, given $M^j \sim \chi^2(K)$, equal to

$$Cov(\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j | \mathbf{x}, M^j \leq a) = \nu_a Cov(\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j | \mathbf{x}), \tag{3}$$

where

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

This implies that the variance of the covariates in the subset of allocations is reduced in comparison to the complete randomization. The equal percent variance reduction (EPVR) of the included covariates is equal to

$$100(1 - \nu_a). \tag{4}$$

ν_a can be written as

$$\frac{2}{K} \times \frac{\gamma(K/2+1, a/2)}{\gamma(K/2, a/2)},$$

where $\gamma(b, c) = \int_0^c y^{b-1} e^{-y} dy$. From this expression it is clear that the variance reduction is increasing in a and decreasing in K.

Let $Y_i(0)$ be the the potential outcome if not treated and let R^2 be the coefficient of determination of a regression of $\mathbf{Y}(0)$ on $\mathbf{x}$, where $\mathbf{Y}(0) = (Y_1(0), ..., Y_N(0))'$. Under the assumptions of conditionally normally distributed outcomes and additive treatment effects Morgan and Rubin (2012) show that the percent reduction in variance of the differences in mean estimator is equal to

$$100R^2(1 - \nu_a). \tag{5}$$

In the two following sections, the two main contributions of this paper are presented, respectively. First, the practical strategy for implementing rerandomization for a large class of rerandomization is proposed, after which the new covariate balance measure is presented.

3 Sampling scheme for exact inference under rerandomization

This section presents a formal strategy for how to implement rerandomization with exact inference in practice. This strategy applies for a large class of rerandomization balance measures and is used throughout this paper

The exact p-value (Fisher, 1935) of a test-statistic for a randomly selected allocation is obtained from the percentile of the histogram of the observed values of the test-statistic for all possible allocations. In complete randomization there are N_A possible allocations. To calculate the exact p-value, the test statistic is calculated for all $j = 1, ...N_A$ allocations to create the histogram. The implication of this procedure is that the smallest possible p-values under the null hypothesis will be restricted by the size N and whether there are ties in the test statistic. If the number of possible allocations is restricted further, as when using rerandomization, the set of allocations fulfilling the rerandomization criteria must be kept large enough to calculate the desired percentiles. This means that, randomization within a set of 'best' allocations is valid only as long as the set of allocations is large enough to obtain the desired resolution of the exact p-value. The resolution of the exact p-value from a two-sided hypothesis test is $2/(\#$ unique $S)$, where S is the set of test-statistic values.[2] This means that, for a resolution, r, on the hundredth in the two-sided p-value, 200 allocations must be selected, given no ties in the statistic. For continuous outcomes and a statistic that is a smooth function of the outcome, e.g. the difference-in-means, all allocations are likely to be unique.

For most sample sizes, the number of allocations fulfilling even very strict rerandomization criteria is usually large why too low resolution of the p-value is not usually an issue. However, too large number of allocations means that it becomes intractable to calculate the exact p-value over *all* admissible allocations. One alternative is to randomly draw an allocation from the admissible allocations, and then Monte Carlo approximate the exact p-value. Another alternative is to first choose a subset of H allocations from all admissible

[2]Given no ties in the test statistic the resolution of the p-values from a two-sided hypothesis test is equal to $2/\binom{N}{N/2}$.

allocations and then randomly draw one of these H allocations. The exact p-value can then be calculated only over the subset of the H allocations. This procedure will in general require sampling a much less number of allocations than with Monte Carlo approximation and provides, by definition, the correct level of the hypothesis test for an effect. In the following sections we discuss in detail how the H allocations should be chosen in practice.

3.1 Mirror allocations and unbiased estimators of SATE under rerandomization

Let Y_i be the outcomes and define $\overline{Y}_w = \frac{1}{N_w} \sum_{i \in (w_i = w)} Y_i$. For a balanced design under complete randomization, the randomization distribution of the difference-in-mean, or SATE, estimator

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

is symmetric around SATE, why it is an unbiased estimator over the randomization distribution, thus $\mathrm{E}(\widehat{\tau}_y) = SATE$. Let $\varphi(\mathbf{x}, \mathbf{W}^j)$ be a row-exchangeble scalar criterion of $\mathbf{x}$ and $\mathbf{W}^j$

$$\varphi(\mathbf{x}, \mathbf{W}^j) = \begin{cases} 1, & \text{if } \mathbf{W}^j \text{ is an acceptable randomization,} \\ 0, & \text{if } \mathbf{W}^j \text{ is not an acceptable randomization} \end{cases} \tag{6}$$

such that $\varphi(\mathbf{x}, \mathbf{W}^j) = \varphi(\mathbf{x}, \mathbf{1} - \mathbf{W}^j)$, and let $\mathbf{W}^\varphi$ be the set with $\varphi(\mathbf{x}, \mathbf{W}^j) \equiv 1$, then $\mathrm{E}(\widehat{\tau}_y | \mathbf{x}, \mathbf{W}^\varphi) = SATE$ (Morgan and Rubin (2012, pp 1279) theorem 2.1). Note that, for the Mahalanobis distance $\varphi(\mathbf{x}, \mathbf{W}) = \mathbb{1}[M(\varphi(\mathbf{x}, \mathbf{W}^j)) \leq a]$, where $\mathbb{1}[A]$ is an indicator function taking the value 1 if A is true and zero otherwise.

One way to understand why the SATE estimator is unbiased using Mahalanobis-based rerandomization is to study the implication of $\varphi(\mathbf{x}, \mathbf{W}^j) = \varphi(\mathbf{x}, \mathbf{1} - \mathbf{W}^j)$. The allocations $\mathbf{W}^j$ and $\mathbf{1} - \mathbf{W}^j$ can be denoted as *mirror allocations*. The mirror allocation of any allocation is simply its opposite, i.e., all experimental units that were assigned treatment in the original allocation are instead assigned control in the mirror allocation and vice versa. This means that any subset of allocations containing only, one or several, pairs of mirror allocations, gives unbiased estimates of SATE. Thus, if we always include mirror allocations, i.e. use symmetric balance measures in the criterion function such that $\varphi(\mathbf{x}, \mathbf{W}^j) = \varphi(\mathbf{x}, \mathbf{1} -$

$\mathbf{W}^j$), the minimum number of allocations included in $\mathbf{W}^\varphi$ is two. If the set $\mathbf{W}^\varphi$ is large, exact inference is in general only possible using Monte Carlo approximations. Instead, it is possible to chose a subset of allocations, $\mathbf{W}^*$, of size H, for which the exact test can be performed. If a subset is chosen, only pairs of mirror allocations should be included to guarantee unbiasedness of the SATE estimator.

3.2 A mirror allocation based sample scheme

In addition to ensuring the exactness of randomization inferences, the inclusion of mirror allocations (pairs of mirror allocations) has the advantage of saving computational time. The time it takes to compare the balance measure of any number of allocations is reduced by a factor of 1/2, i.e. twice as many allocations can be considered during the same amount of time. This follows since for any symmetric balance measure, if an allocation is admissible, so is the corresponding mirror allocation and thus only the covariate balance of one allocation in each pair of mirror allocations has to be evaluated.[3]

As the number of possible allocations grows very fast in N (also when using the mirror allocation strategy), it is the computational time that limits the possibility of finding the globally best allocations for most N in practice. We argue that the goal with any rerandomization design is to find, within a certain time limit, the best possible set of allocations for which exact inference can be conducted with the highest possible efficiency. Therefore, instead of rerandomizing until some pre-specified criterion is met, we propose the following procedure to obtain the approximate **best set** $\mathbf{W}^*$ for large N, the properties for this set are discussed below.

The proposed allocation sample scheme, using the mirror-allocation strategy, is presented in detailed in Allocation sample scheme 1.

Allocation sample scheme 1 *Let N_A be the total number of possible allocations under*

[3]Note that in a balanced factorial experiment with G treatments, the time it takes to compare the balance measure of any number of allocations is reduced by a factor of $1/G!$ when using this mirror allocation design.

complete randomization. Assuming that the allocations are ordered in lexicographic order[4]*, the first* $N_A/2$ *allocations contains no pairs of mirror allocations. Call the set containing the first* $N_A/2$ *allocations* C_M*. Choose a desired level of resolution* r*. Assuming no ties, this gives the number of allocation* $H = 2/r$*. For* $H^* \geq H/2$

1. *From* C_M*, randomly sample a set,* $\mathbf{A}_1$*, containing* H^* *allocations, without replacement.*

2. *Calculate the balance measure for all allocations in* $\mathbf{A}_1$*, store the* $H/2$ *allocation with the smallest imbalance from* $\mathbf{A}_1$ *in the set* $\mathbf{W}_s^*$*.*

3. *In the* i*th iteration* ($i = 2, ..., I$)*, sample a new set of allocations,* $\mathbf{A}_i$*, of size* H^* *from the* $N_A/2 - i \times H^*$ *remaining allocations in* C_M*.*

4. *Calculate the balance measure for all allocations in* $\mathbf{A}_i$*,. Replace the set* $\mathbf{W}_s^*$ *with the* $H/2$ *allocations with smallest imbalances in the set* $\{\mathbf{W}_s^*, \mathbf{A}_i\}$*.*

5. *Repeat steps 3-4 until* $i = I$ *or* $i \times H^* = N_A$ *(all allocations have been considered).*

6. *As the final subset of admissible allocations save* $\mathbf{W}^* = \{\mathbf{W}_s^*, \mathbf{W}_{Mirror}^*\}$*, where* $\mathbf{W}_{Mirror}^*$ *contains all mirror allocations for the allocations in* $\mathbf{W}_s^*$

Note that $H^*/2$ can be tuned in accordance with the memory capacity of the computer to gain speed. As the sampling of allocations are random, $\mathbf{W}^*$ is an approximation of the set containing the globally best allocations. This approximation will be better the longer the sampling of allocations is allowed to continue. The number of iterations can for example be decided based on computational time. Making use of this procedure in the empirical example, presented in detail in Section 5, the allocation sampling scheme was left to work for 11 hours after which about one billion allocations had been considered and the $H = 800$ best had been retrieved. If the sample size is small enough for the available computational power to go through all $N_A/2$ allocations in C_M, the H globally best allocations should of course be selected.

[4]Most programming language have functions for generating combinations in lexicographic order, e.g. the `RcppAlgos`-package in R.

For clarity we here discuss the properties of the proposed sampling scheme using the Mahalanobis balance measure. In summary, there are two parameters that governs the proposed sample scheme. The number of included allocations H (implied by the desired resolution of the exact p-value) and the number of considered allocations ($I \times H^*$). The optimal set of allocations for exact inference, with resolution as implied by H, is defined as the set $\{M^{[1]}, ..., M^{[H]}\}$. That is, the allocations with the H first order statistics of the Mahalanobis distance in the set of all $\binom{N}{N_1}$ allocations. This corresponds to $a = M^{[H]}$ in $p_a = \Pr(\chi^2(K) \leq a)$.

Let $M^{[H_{obs}]}$ be the H:th order statistic in the set of considered allocations. The sample scheme then implies that $p_{a^*} = \Pr(\chi^2(K) < M^{[H_{obs}]})$ is unknown and stochastically dependent on $I \times H^*$. This is in contrast to Morgan and Rubin (2012) where the inclusion criterion p_a is fixed and the number of included allocations is random. If the number of considered allocations is small, i.e. $I \times H^* << \binom{N}{N_1}$, then $M^{[H_{obs}]}$, and thus p_{a^*}, may be large. However, as $I \times H^* \to \binom{N}{N_1}$ it follows that $M^{[H_{obs}]} \to M^{[H]}$ and $p_{a^*} \to \Pr(\chi^2(K) \leq M^{[H]})$. Furthermore, given that the Mahalanobis distance is chi-square distributed the distribution of the H:th order statistic is known for any $I \times H^*$. This means we can calculate an 'upper bound' of p_{a^*} for $M^{[H_{obs}]}$ for any given $I \times H^*$. As an example, Figure 2 displays the CDF of the 800:th ($H = 800$) order statistic for a $\chi^2(3)$ distribution as a function of p_a for different $I \times H^*$. It is clear from the figure that, with probability close to one, $\Pr(\chi^2(K) \leq M^{800}) = 0.009$ when $I \times H^* = 100,000$. Thus, with probability close to one, the set $\mathbf{W}^*$ only contains allocations from the 1% of the globally best allocations, or better. With $I \times H^* = 1\ 000,000$ the corresponding probability is $\Pr(\chi^2(K) \leq M^{800}) = 0.001$, and the set $\mathbf{W}^*$ only contains allocations from the 0.1% of the globally best allocations, or better.

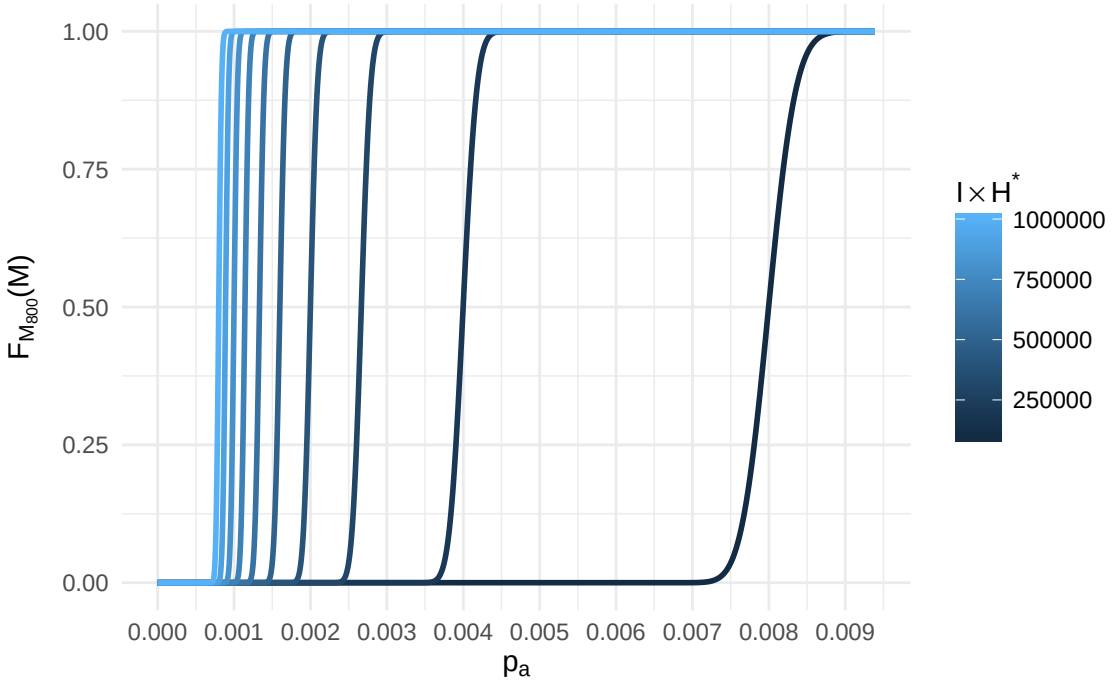

Figure 2: The cumulative distribution function for the Mahalanobis distance (df=3) of the 800th order statistic in the random sample of allocations for numbers of allocations beteen 100,000 and 1000,000.

4 A rank based balance measure

In this section we propose the second contribution of this paper, an alternative rerandomization balance measure based on the rank of the covariate mean difference. The suggested balance measure is marginal affinely invariant and avoids the problem with inverting potentially large and singular covariance matrices that might occur using the Mahalanobis distance, especially with highly correlated covariates. In addition, the suggested balance measure provides a convenient alternative to Morgan and Rubin (2015) if the researcher has a priori information on the relative importance of the observed covariates.

Again, let H be the number of allocations giving the desired resolution, r, of the exact p-value in the final test, and let $\mathbf{w}$ be the vector of the de facto ex post assignment. Based on the rank of the mean difference of the covariates, define the best allocation among all

N_A allocation as

$$\mathbf{w}_1 = \min_{\mathbf{W}} \sum_{k=1}^{K} Rank(|\overline{\mathbf{X}}_{kT}^{j} - \overline{\mathbf{X}}_{kC}^{j}|), \ j = 1, ..., N_A.$$

That is, the 'best' allocation is the allocation with the smallest sum over the ranks of the individual mean differences of the K covariates. Define the $H : th$ best allocation as

$$\mathbf{w}_H = \min_{\mathbf{W}_{H-1}} \sum_{k=1}^{K} Rank(|\overline{\mathbf{X}}_{kT}^{j} - \overline{\mathbf{X}}_{kC}^{j}|), \quad j = 1, ..., (N_A - (H-1))$$

where $\mathbf{W}_{H-1} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{H-1})$ and $\mathbf{W}_H = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{H-1}, \mathbf{w}_H) = \mathbf{W}^*$.

As the exact p-value from a two-sided mean difference test for any covariate is a monotonous transformation of the rank of the absolute differences, the p-values can equally well be used in finding $\mathbf{W}^*$. Let p_{jk} denote the exact p-value of the two-sided mean difference test for allocation j and covariate k, and define the balance measure

$$\varrho(\mathbf{x}, \mathbf{W}^j) = \varrho^j = \left(\sum_{k=1}^{K} \omega_k p_{jk} \right), \tag{7}$$

where ω_k is the weight given to covariate k, with $\sum_{k=1}^{K} \omega_k = 1$. The H allocations with the largest $\varrho^j, j = 1, ...N_A$ gives the set $\mathbf{W}^*$. This follows since the largest p-value implies the smallest mean difference. For this balance measure the criterion is $\varphi(\mathbf{x}, \mathbf{W}^j) = 1(Rank(\varrho(\mathbf{x}, \mathbf{W}^j)) < H) = \varphi(\mathbf{x}, 1-\mathbf{W}^j))$. Furthermore, as we in the sampling only includes mirror allocations is $\widehat{\tau}_y$ from randomization within the set $\mathbf{W}^*$ an unbiased estimator for SATE by design.

With one single covariate, i.e. $K = 1$, the H allocations with the smallest imbalances, i.e. the allocations with the largest ϱ^j among $j = 1, ...N_A$, are the same allocations that have the H smallest Mahalanobis distances. This implies that with one covariate the rerandomization based on this new balance measure yields identical variance reductions as in Morgan and Rubin (2012) and, hence, that the percent reduction in variance of the estimator under normality of the outcome is given in Equation (5). However, for $K > 1$ the two balance measure may give different sets of allocations.

Throughout the remainder of this paper we denote the balance measure using the procedure given in Equation (7) R, while the procedure in Equation (1) using the Mahalanobis

distance balance measure is denoted M. R with uniform weights are simply denoted R while if the R balance measure is based on non-uniform weights it is denoted $R(\omega, o)$, where o is a generic term defining the method being used to set these weights.

Morgan and Rubin (2015) address the case when covariates vary in a priori importance and suggest rerandomization based on Mahalanobis distance within tiers of covariate importance. The R balance measure is a complement to the Morgan and Rubin (2015) which is simple to implement and allows the weights $\omega_k, k = 1, ..., K$, to be based on a priori information of their importance. In the following section we discuss strategies for estimating the weights in the situation where (at least) one pre-measured outcome is observed at the design stage.

4.1 Estimating the weights from pre-treatment data

We first discuss the case with one pre-treatment outcome and then the case when there are no covariates other than the pre-treatment outcome observed for several time periods, i.e. only pre-treatment time-series outcome data are observed.

4.1.1 One pre-treatment outcome

If the outcome Y is observed T time periods before the treatment assignment, and that these pre-treatment outcomes can be assumed to be associated with the post-treatment outcome $Y_{t=T+1}(0)$, the weights $\omega = (\omega_1, ..., \omega_K)'$ can be estimated from data. With $T = 0$, the most obvious strategy is to estimate the weights using the partial correlations of the covariates and the pre-treatment outcome. That is, estimate

$$\widetilde{y}_{i0} = \sum_{k=1}^{K} \beta_k \widetilde{x}_{ik} + \varepsilon_{iT}$$

using ordinary least square (OLS). Here $\sim$ indicate a standardized variables, i.e. $\widetilde{z}_i = (Z_i - \overline{Z})/\sqrt{\widehat{Var(Z_i)}}$, where Z is either Y or any of the X variables.

With K covariates and one pre-treatment outcome we have $Q = K + 1$ variables to base rerandomization on.[5] Denote ω_0 the weight for $\widetilde{y}_{i0}$ then $\omega = (\omega_0, \omega_1, ..., \omega_K)'$. We suggest

[5]The set of covariates can, of course, be extended to include transformations of the originally observed

these weights to be estimated as $\widehat{\omega}'=(\frac{|\widehat{\beta}_1|}{\delta}, ..., \frac{|\widehat{\beta}_K|}{\delta}, \frac{1}{\delta})$, where $\delta = \sum_{k=1}^{K} |\widehat{\beta}_k| + 1$. As all variables are standardized to have unit standard deviation the coefficients are therefore bounded in theory, i.e. $0 \leq |\widehat{\beta}_j| \leq 1 \ \forall \ j = 1, ..., K$, and it follows that $\frac{1}{\delta} \geq \frac{|\widehat{\beta}_j|}{\delta}$. This means that the pre-treatment outcome will always have the largest weight, which is natural given that the weights are estimated under the assumption that the pre-treatment outcome is associated with the outcome at the time period of the experiment. This strategy is denoted as $R(\omega, O)$ throughout the rest of this paper.

If there are many covariates and/or the sample size is small in comparison to K, it might be useful to estimate the partial correlations with some regularization estimator, e.g., LASSO (Tibshirani, 1996). Using LASSO tuned with cross-validation, some covariates can be given zero weights, which might substantially reduce the noise in the weight estimation with many highly correlated covariates. This strategy, using leave-one-out cross-validation, is denoted as $R(\omega, L)$.

In the situation where the outcome is observed at several time periods pre-treatment, i.e. $T > 0$, LASSO is still useful to estimate the weights for the R balance measure, however, for larger numbers of time periods other strategies for performing the rerandomization might be preferable as is discussed in the next section.

4.1.2 Several pre-treatment outcomes

Because the p-values must be calculated for each pre-treatment time period with non-zero estimated weight, the time consumption of estimating $R(\omega, L)$ increases drastically with the number of pre-treatment time periods. An alternative in this situation is to predict the outcome value at the time period of the experiment by fitting a time series model to the pre-treatment outcomes of each individual and then use the one step forecast to base the rerandomization on. This strategy, using the `auto.arima` function in the `forecast` R-package to illustrate, is denoted $R(\omega, F)$. Since there is only one forecast value for each individual, M based on the forecast will give the same allocations as with $R(\omega, F)$.

Note that the $R(\omega, F)$ strategy is not only time saving, it also allows for heterogeneity

covariates.

across experimental units in a natural way. The two strategies $R(\omega, L)$ and $R(\omega, O)$ are likely to work well if the processes across individuals are homogeneous. If the processes are heterogeneous across the individuals, e.g. with differences in the 'memory' or time-dependency in the outcomes, the $R(\omega, F)$ may be more efficient in reducing the variance as it handles this heterogeneity, although, at the cost of estimating more parameters. Thus, if T is large and heterogeneity is present (which is possible to detect using the pre-treatment data) $R(\omega, F)$ should likely be preferred over $R(\omega, L)$ and $R(\omega, O)$.

The following section provides Monte Carlo studies where the different strategies are evaluated.

4.2 Monte Carlo simulation studies

Throughout the Monte Carlo studies, complete randomization will serve as a benchmark for the reduction of variance of the covariates (including pre-treatment outcomes) and the estimated treatment effect under the null for the different rerandomization strategies.

In traditional Neyman-Person asymptotic inference, a test with a small variance of the estimator will be asymptotically more powerful than a test based on an estimator with large variance given that both test are based on consistent estimators of the effect and the variance. As the power of the Fisher Randomized Test (FRT) is based on shifts in the rank due to the shift under the alternative it is not possible to evaluate the power of two strategies based only on the variance of the estimators under null.[6] For this reason, the relative power under the alternative is also presented.

[6]The distribution of the FRT test is only known empirically (i.e. the histogram). Under the Fisher null (i.e. homogeneous treatment effects and the same variance of treated and controls and no treatment effect) the asymptotic variance is equal to $\widehat{V} = Ns^2/(N_1 N_0)$ where s^2 is the sample variance and

$$\frac{\widehat{\tau} - 0}{\sqrt{\widehat{V}}} \xrightarrow{d} N(0, 1)$$

Under these assumptions the t-test (i.e.Neyman Pearson inference) and FRT have the same size asymptotically. However the t-test will be more powerful under the alternative Ding (2017). He show that the difference in $\widehat{V}$ and the Neyman-Pearson variance is positive as a consequence of the squared difference in the homogeneous treatment effect under the alternative.

All evaluations of variance are made on the data for the period after the allocation is made, that is if the treatment assignment is performed at period T, the power and variance is calculated at $T+1$. Denote complete randomization with c, and the different rerandomization strategies with $d = M, R, R(\omega, O), R(\omega, L)$ and $R(\omega, F)$. For each arm of the study, 4000 replications (N_{rep}) are considered.

Let $w_{ir} = 1$ or 0 if unit i is treated or control in replicate r and let z_{rqi} be value of covariate q for unit i in replication r and define

$$\overline{\mathbf{z}}_{rqT} = \frac{\sum_{i=1}^{N} w_{ir} z_{rqi}}{\sum_{i=1}^{N} \mathbb{1}(w_{ir}=1)}, \quad \overline{\mathbf{z}}_{rqC} = \frac{\sum_{i=1}^{N} (1-w_{ir}) z_{rqi}}{\sum_{i=1}^{N} \mathbb{1}(w_{ir}=0)}, \; r = 1, ..., N_{rep},$$

$$\overline{\mathbf{Z}}_{qT} = \begin{bmatrix} \overline{\mathbf{z}}_{1qT} \\ \vdots \\ \overline{\mathbf{z}}_{N_{rep}qT} \end{bmatrix}, \text{and} \quad \overline{\mathbf{Z}}_{qC} = \begin{bmatrix} \overline{\mathbf{z}}_{1qC} \\ \vdots \\ \overline{\mathbf{z}}_{N_{rep}qC} \end{bmatrix}.$$

The relative (compared to complete randomization) change in variance in the mean difference between the treated and controls in covariate/pre-treatment outcome q, at time period T, using design d is then defined as

$$\text{VC}(q|d) \equiv \frac{Var(\overline{\mathbf{Z}}_{qT} - \overline{\mathbf{Z}}_{qT}|d) - Var(\overline{\mathbf{Z}}_{qT} - \overline{\mathbf{Z}}_{qT}|c)}{Var(\overline{\mathbf{Z}}_{qT} - \overline{\mathbf{Z}}_{qT}|d)}, \tag{8}$$

where

$$q = 1, ..., K, \mathbb{1}(T>0)(K+1), ...\mathbb{1}(T>0)(Q-K),$$

The treatment effect estimate for each replicate, r, is defined

$$\hat{\tau}_{rc} = \frac{1}{\sum_{i=1}^{N} \mathbb{1}(w_{ir}=0)} \sum_{i=1}^{N} w_{ir} Y_{iT+1} - \frac{1}{\sum_{i=1}^{N} \mathbb{1}(w_{ir}=0)} \sum_{i=1}^{N} (1-w_{ir}) Y_{iT+1}$$

$$\hat{\tau}_{dr} = \frac{1}{\sum_{i=1}^{N} \mathbb{1}(w_{ir}=0)} \sum_{i=1}^{N} w_{ir} Y_{iT+1} - \frac{1}{\sum_{i=1}^{N} \mathbb{1}(w_{ir}=0)} \sum_{i=1}^{N} (1-w_{ir}) Y_{iT+1} \Big| W \in \mathbf{W}_d^*,$$

The emperical variance under rerandomization and complete randomiztion is then defined

$$Var(\hat{\tau}_d) = \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} (\hat{\tau}_{rd} - \bar{\tau}_d)^2 \text{ and } Var(\hat{\tau}_c) = \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} (\hat{\tau}_{rc} - \bar{\tau}_c)^2$$

where $\bar{\tau}_d$ and $\bar{\tau}_c$ are the average estimated treatment effects across the replications. The relative change in variance of the treatment effect of strategy d is defined

$$\text{VC}_\tau(d) \equiv \frac{Var(\hat{\tau}_d) - Var(\hat{\tau}_c)}{Var(\hat{\tau}_c)}. \tag{9}$$

The exact p-value for the complete randomization and the rerandomization strategies are defined as

$$\pi_{rc} = \Pr(|\widehat{\tau}_{rc}(\mathbf{W}, \mathbf{Y})|) \geq |\widehat{\tau}_{rc}| \text{ and } \pi_{rd} = \Pr(|\widehat{\tau}_{rd}(\mathbf{W}_d^*, \mathbf{Y})|) \geq |\widehat{\tau}_{rd}|.$$

Here $\widehat{\tau}_{rc}(\mathbf{W}, \mathbf{Y})$ is the distribution of estimates over all allocations in replication r and $\widehat{\tau}_{rd}(\mathbf{W}_d^*, \mathbf{Y})$ is the corresponding distribution under rerandomization. The relative power is evaluated with τ being varied from $0.4\sigma_Y$ to $2\sigma_Y$ in steps of $0.4\sigma_Y$, and estimated as

$$\text{Power}(\tau, d) \equiv \frac{p_{\tau d} - p_{\tau c}}{p_{\tau c}}, \tau = 0.4, 0.8, ..., 2.00, \tag{10}$$

where

$$p_{\tau c} = \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} \mathbb{1}(\pi_{rc} \leq 0.05)$$

and

$$p_{\tau d} = \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} \mathbb{1}(\pi_{rd} \leq 0.05).$$

4.3 Cross section data and one pretreatment outcome

Consider the data generating process

$$Y_{it} = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_{it}, i = 1, ...N, t = 0, 1. \tag{11}$$

where $\mathbf{x}_i$ is a $K \times 1$ vector of normal distributed variables with mean 2 and covariance matrix $\boldsymbol{\Sigma}$**,** i.e. $\mathbf{x}_i \sim N(\mathbf{2}, \boldsymbol{\Sigma})$ and $\epsilon_{it} = 0.3 \times \epsilon_{it-1} + \zeta_{it}$, where ζ_{it} is independent and identical distributed (iid) and normal, i.e., $\zeta_{it} \sim N(0, \sigma^2)$.[7] Due to independence the marginal variance of the outcome is equal to

$$\text{Var}(Y) = \boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}' + \frac{\sigma^2}{1 - 0.3^2}. \tag{12}$$

We compare the proposed rerandomization balance measure, R, with M for $T = 0$ ($K = 3$) and $T = 1$ ($K = 3$ and $K = 10$) under different specifications of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. σ^2 is chosen to obtain $R^2 = 0.25$ in expectation.

[7]The results from the Monte Carlo simulations with regards to power are not sensitive with respect to the choice of distributions of the covariates or the error term. The covariates and the error terms are chosen to be normally distributed only in order to compare the results from the Monte Carlo to the theoretical expected variance reductions given in equations (4) and (5) for small N.

4.3.1 Cross section

This section serves to compare the proposed balance measure with the balance measure presented in Morgan and Rubin (2012) in the setting presented there for completeness. That is, the weights of the covariates are assumed uniform, i.e. not estimated, which implies that we expect no improvement using our balance measure as compared to the Mahalanobis distance. As this setting is not primary focus , the results are restricted to $N = 14$, for which the 800 allocations with the (globally) smallest value of the balance measure are used. This implies that $p_a = 800/\binom{14}{7} = 0.233$.

Two data generating processes (DGPs) are considered. In the first, (A), we let $\boldsymbol{\beta} = (1, 1, 1)$ and $\boldsymbol{\Sigma} = \text{diag}(1, 1, 1)$ and in the second, (B), we let $\boldsymbol{\beta} = (1, -1, 1)$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.8 & 0 \\ -0.8 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Table 1: Relative variance change in the covariates and the effect estimate as compared to complete randomization (Eq. 8 and 9) for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The left and right panels display the results for DGPs A and B, respectively.

	A		B	
balance measure	R	M	R	M
$\text{VC}(X_1\|.)$	-0.71	-0.74	-0.83	-0.74
$\text{VC}(X_2\|.)$	-0.71	-0.75	-0.83	-0.75
$\text{VC}(X_3\|.)$	-0.70	-0.75	-0.73	-0.75
$\text{VC}_\tau(.)$	-0.20	-0.19	-0.22	-0.20

Table 1 displays the relative variance reduction in the covariates and the estimated treatment effect. The variance reduction of the covariates for the M balance measure is, as expected, of the same magnitude for all covariates and around 75%. Given that $\nu_a = 0.213$ $(= P(\chi^2_5 \leq 0.233)/P(\chi^2_3 \leq 0.233))$ this variance reduction is close to the one theoretically

expected of 78.7% (cf. equation 4). The variance reductions of the covariates using the R balance measure is around 71% under DGP A. Under DGP B the variance reduction is 83% for the two correlated covariates but only 73% for the independent covariates. The variance reduction of the treatment effect under the null is around 20% for all strategies. Given that $\nu_a = 0.213$ and $R^2 = 0.25$ this is in line with the theoretically expected variance reduction using the M balance measure of 19.7% (cf.equation 5). Figure 3 displays the

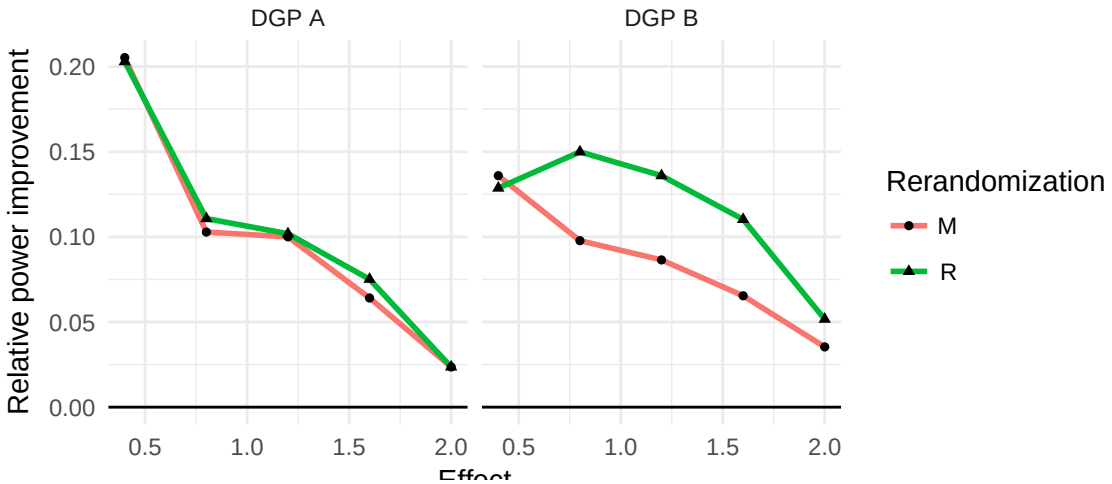

Figure 3: Relative power as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The left and right figures display the results for DGPs A and B, respectively.

relative power gain of the rerandomization as compared to complete randomization for the two DGPs A and B. From the left panel, displaying the result under DGP A, one can see that both criteria increase the power by 20% for small treatment effects. Given that the covariates are uncorrelated, the similarity of the results with M and R with uniform weights is expected. The results displayed in panel B show that the criteria have similar power gains for small effect sizes. However, for larger effect sizes, R gives substantially larger power gains.

4.3.2 Results with one pretreatment outcome

Turning to the case with $T = 1$, two new DGPs are considered. The first, (A), with $K = 3$ we let $\beta = (0, 0.25, 0.75)$ and $\mathbf{\Sigma}$ =diag$(1, 1, 1)$. In the second, (B), with $K = 10$ we let $\beta = (0, 0, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7)$ and $\mathbf{\Sigma}$ =diag$(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$. For both DGP's the importance of the covariates is increasing in k. In DGP A, x_1 does not contribute to the variation of the outcome, and in DGP B, x_1, x_2 and x_3 do not contribute to the variation of the outcome. These two DGPs illustrate how the criteria M, R , $R(\omega, O)$, and $R(\omega, L)$ are affected by the number of included covariates and how well the criteria incorporates the relative importance of the covariates. The considered sample sizes are $N = 14, 50$ and 100. Including one pre-treatment outcome observation implies $Q = 4$ and $Q = 11$ covariates. In the $Q = 11$, $N = 14$ setting, the M balance measure is expected to perform poorly, as the covariance matrix is very large in comparison to the sample size.

When $N = 14$, the 800 globally best allocations for a given sample are used for each criteria. As $p_a = 0.233$ for the Mahalanobis balance measure this implies $\nu_a = 0.29$ $(= P(\chi^2_6 \leq 0.233)/P(\chi^2_4 \leq 0.233))$ with $Q = 4$. For $Q = 11$ we get $\nu_a = 0.51$ $(= P(\chi^2_{13} \leq 0.233)/P(\chi^2_{11} \leq 0.233))$. This means that, using the M balance measure, the expected variance reductions in the covariates are 71% and 49% for $Q = 4$ and $Q = 11$, respectively. The corresponding variance reduction in the variance of the effect estimate are 17.99% and 12.25% percent.

To limit the computational time for $N = 50$ and $N = 100$, we randomly sampled $4,000$ from all possible allocations. To get the same resolution of the exact p-value as in the $N = 14$ case, the rerandomization is performed by selecting the 800 best allocations according to the different criteria, within this set of $4,000$ allocations. This implies that the 20% best allocations among these 4,000 are selected. Since p_a is a little bit smaller than with $N = 14$ the expected variance reductions is a few percent larger. However, as a consequence of the initial random sampling of the 4,000 allocations the theoretical results of the variance reductions for the M balance measure may be less exact for $N = 50$ and 100.

Table 2 displays the change in variance in the covariates and the effect estimates for

the four criteria compared to complete randomization for both DGP's. Figure 4 displays the results with respect to relative power of the four criteria.

From the top panel, displaying the results with $Q = 4$, we can see that the variance reduction of the R and M balance measure are very similar both in the covariates and the effect estimates. For $N = 14$, the variance reductions are a little bit lower than what is theoretically expected (71% in the covariates and 18% in the effect estimate). With $N = 50$ and 100 the variance reduction is a little bit higher than what is theoretically expected. For the $R(\omega, O)$ and $R(\omega, L)$ criteria, the variance reductions in the covariates are the highest for the pre-treatment outcome and monotonically increasing for the covariates which is in accordance with the increasing importance of the covariates in the DGPs. With regard to variance reduction of the estimated treatment effect, the $R(\omega, O)$ balance measure is performing best overall.

From the bottom panel, displaying the results with $Q = 11$, it is clear that the variance reductions of the M balance measure and the R balance measure are performing similar but only when $N \geq 50$. The variance reductions are close to the theoretically expected of 49% and 12% for the M balance measure. However with $N = 14$ the variance reduction is much smaller than the theoretical value as expected given the small number of observations to estimate the 10×10 covariance matrix. It is interesting to note that the variance reduction for the R balance measure with $N = 14$ is on the same levels as with $N = 50$ and 100. Turning to the two criteria using the estimated weights $R(\omega, L)$ and $R(\omega, O)$, the pattern as with $Q = 4$ is repeated. Both criteria are reducing the variance of the lagged outcome the most and are both successfully picking up the relative importance of the covariates in the DGP, even though the $R(\omega, L)$ balance measure are better. Except for $N = 14$, the $R(\omega, O)$ balance measure is giving the largest variance reduction in the estimated treatment effect. With $N = 14$, the $R(\omega, L)$ balance measure performs better than the $R(\omega, O)$ balance measure.

The left panels of Figure 4 displays the results for the relative power when $Q = 4$. From this figure it is clear that all the rerandomization criteria increase power as compared to complete randomization. The power gain for the smallest effect sizes is around $15 - 25\%$

and increasing with N. Both the $R(\omega, O)$ and $R(\omega, L)$ criteria have higher power than the M and R criteria. With $N = 14$, the power gain using the $R(\omega, L)$ balance measure is substantial also compared to the $R(\omega, O)$ balance measure. Turning to the right panels

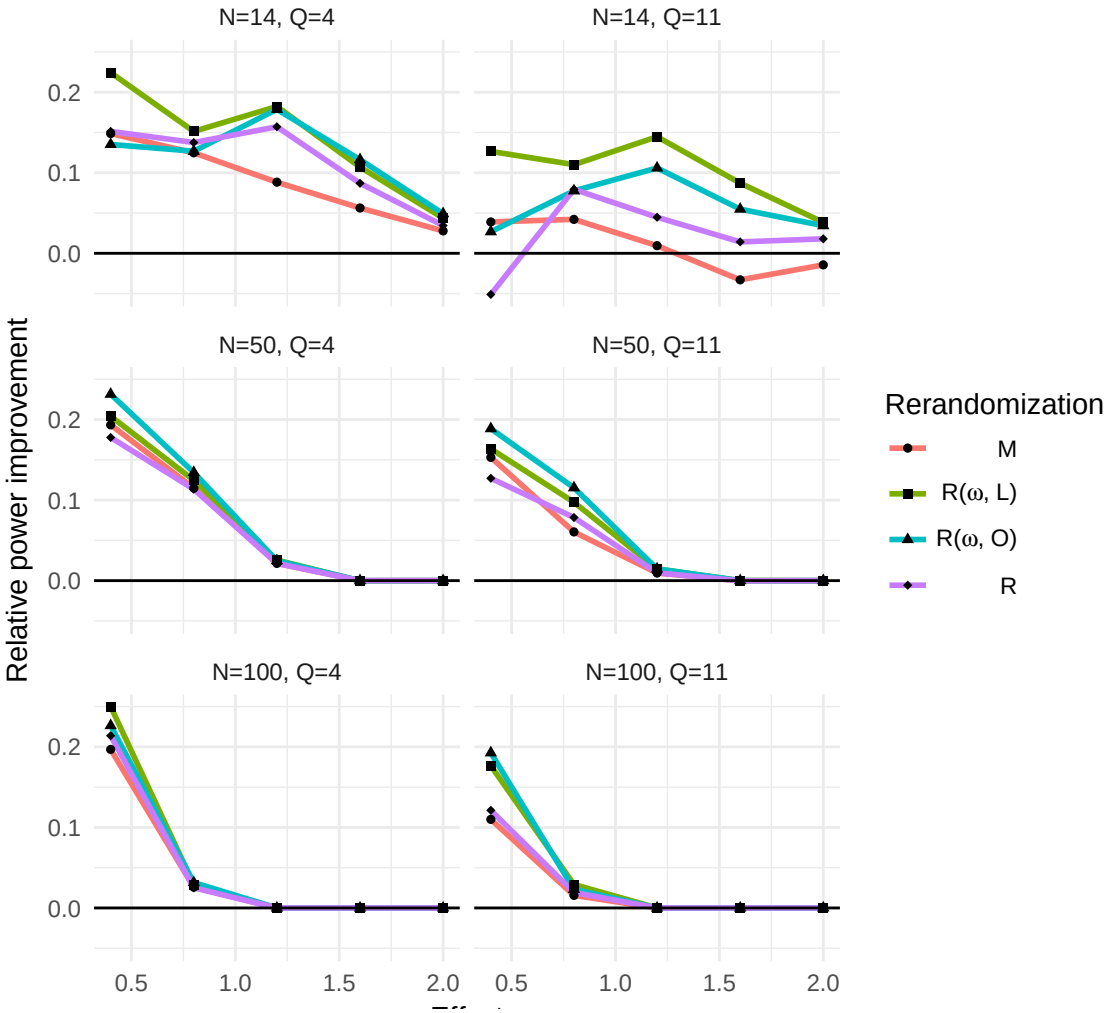

Figure 4: Relative power as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The left and right panels display the results including four (Q=4) and eleven (Q=11) covariates, respectively.

displaying the results with $Q = 11$, the power is increasing in N and that the $R(\omega, O)$ and $R(\omega, L)$ criteria are more efficient than the M and R criteria. Once again, with $N = 14$, the power gain using the $R(\omega, L)$ balance measure is substantial in comparison with the other criteria. This results displays the advantage of using penalizing in the situation of

many covariates in combination with small samples. With $N = 100$ the relative power increase is almost twice as good with the $R(\omega, O)$ and $R(\omega, L)$ criteria as compared to the M and R criteria.

To summarize, for both data generating processes the power was substantially improved by using estimated weights. This result most likely stems from giving more weight to the covariates most correlated with the outcome. Furthermore, with many covariates and small N it is important to penalize the number of covariates to be included in the criteria in order to obtain real power gains.

Table 2: Relative change in variance of the covariates and in the effect estimate as compared to complete randomization (Eq. 8 and 9) for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The top and bottom panels display the results including four (Q=4) and eleven (Q=11) covariates, respectively.

	$R(\omega, L)$			$R(\omega, O)$			R			M		
N	14	50	100	14	50	100	14	50	100	14	50	100
	Q=4											
VC($Y_{t=T}$|.)	-0.94	-0.96	-0.96	-0.92	-0.95	-0.95	-0.66	-0.71	-0.73	-0.64	-0.73	-0.74
VC(X_1|.)	-0.07	-0.09	-0.05	-0.25	-0.15	-0.09	-0.62	-0.66	-0.67	-0.65	-0.73	-0.74
VC(X_2|.)	-0.17	-0.16	-0.12	-0.33	-0.23	-0.15	-0.63	-0.69	-0.66	-0.67	-0.74	-0.73
VC(X_3|.)	-0.31	-0.55	-0.63	-0.51	-0.61	-0.65	-0.67	-0.71	-0.72	-0.63	-0.73	-0.73
VC$_\tau$(.)	-0.18	-0.20	-0.22	-0.20	-0.22	-0.24	-0.17	-0.20	-0.21	-0.16	-0.20	-0.20
	Q=11											
VC($Y_{t=T}$|.)	-0.93	-0.96	-0.96	-0.68	-0.93	-0.94	-0.39	-0.50	-0.46	-0.21	-0.48	-0.47
VC(X_1|.)	-0.04	-0.08	-0.05	-0.24	-0.16	-0.09	-0.37	-0.43	-0.41	-0.21	-0.47	-0.49
VC(X_2|.)	-0.04	-0.07	-0.03	-0.31	-0.15	-0.13	-0.43	-0.42	-0.44	-0.25	-0.47	-0.49
VC(X_3|.)	-0.05	-0.04	-0.08	-0.25	-0.18	-0.12	-0.36	-0.43	-0.42	-0.21	-0.48	-0.50
VC(X_4|.)	0.03	-0.01	-0.07	-0.24	-0.11	-0.14	-0.36	-0.43	-0.45	-0.17	-0.46	-0.49
VC(X_5|.)	-0.04	-0.04	-0.08	-0.29	-0.17	-0.14	-0.37	-0.41	-0.43	-0.23	-0.46	-0.50
VC(X_6|.)	-0.02	-0.10	-0.07	-0.26	-0.18	-0.17	-0.40	-0.45	-0.42	-0.21	-0.47	-0.48
VC(X_7|.)	-0.08	-0.10	-0.16	-0.24	-0.24	-0.23	-0.36	-0.45	-0.43	-0.17	-0.50	-0.48
VC(X_8|.)	-0.13	-0.11	-0.18	-0.30	-0.28	-0.28	-0.41	-0.46	-0.42	-0.23	-0.46	-0.52
VC(X_9|.)	-0.11	-0.19	-0.31	-0.30	-0.35	-0.34	-0.39	-0.44	-0.44	-0.20	-0.49	-0.51
VC(X_{10}|.)	-0.15	-0.22	-0.32	-0.33	-0.34	-0.40	-0.42	-0.44	-0.45	-0.23	-0.47	-0.52
VC$_\tau$(.)	-0.17	-0.16	-0.17	-0.15	-0.21	-0.22	-0.12	-0.12	-0.13	-0.05	-0.16	-0.11

4.4 Longitudinal data

In this section two different time series DGP are considered with $T = 10, 100$ and $N = 14, 50, 100$. In the first DGP the times series process, denoted Homogeneous in the following, is set to be the same for all units

$$Y_{it} = 0.5 \times Y_{it-1} + \zeta_{it}, i = 1, ...N, t = 1, ...T + 1, \tag{13}$$

where ζ_{it} iid $N(0, 1)$. In the second DGP, denoted Heterogeneous in the following, the time series processes differ across four strata according to

$$Y_{it} = \boldsymbol{\phi}_j \boldsymbol{Y}_{i,lag} + \epsilon_{it} \ , i = 1, ..N, t = 1, ..., T + 1, j = 1, ..4, \tag{14}$$

where $\boldsymbol{Y}_{i,lag} = (y_{it-1}, y_{it-2}, y_{it-3})'$, $\boldsymbol{\phi}_1 = (0.5, 0, 0), \boldsymbol{\phi}_2 = (0, 0.5, 0), \boldsymbol{\phi}_3 = (0, 0, 0.5)$, $\boldsymbol{\phi}_4 = (0.39, 0.32, 0)$ and ϵ_{it} iid $N(0, 1)$. For both DGPs $R^2 = 0.25$.

When $T = 10$, the $R(\omega, O)$, $R(\omega, L)$, and $R(\omega, F)$ criteria are used. In this context, the Mahalanobis-based rerandomization is difficult to apply due to singular covariance matrices and is therefore excluded. Given the large number of correlated pre-treatment outcomes with $T = 100$, also the $R(\omega, O)$ balance measure runs into singularity problems in the estimation in that setting and is therefore excluded in that case.

In this setting with only outcome data, the OLS estimated weights are estimated as

$$\widetilde{y}_{iT} = \beta_0 + \sum_{t=1}^{T-1} \beta_t \widetilde{y}_{it} + \varepsilon_i$$

using ordinary least square (OLS). Denote ω_τ the weight for $\widetilde{y}_{i\tau}$ then $\omega = (\omega_1, ..., \omega_T)'$ and the estimated $\widehat{\omega}' = \left(\frac{|\widehat{\beta}_1|}{\delta}, ..., \frac{|\widehat{\beta}_{T-1}|}{\delta}, \frac{1}{\delta} \right)$ where $\delta = \sum_{t=1}^{T-1} |\widehat{\beta}_t| + 1$. This implies $\omega_T \geq \omega_t \, \forall \, t = 1, ..., T - 1$ in expectation.

4.4.1 Results with T=10

The top panel of Table 3 displays the results under the Homogeneous DGP. With $N \geq 50$, all criteria are successful in giving the latter time periods larger weights which is in line with the DGP. The variance reduction for $R(\omega, O)$ varies more across different N, than the two other criteria. The advantage of using penalized regression for small N is confirmed also in

this setting. It is clear that the $R(\omega, L)$ and $R(\omega, O)$ criteria give larger reductions in the variance of the lagged outcomes and the estimated treatment effect under null than $R(\omega, F)$. The variance reduction in the estimated effects is significantly better for the $R(\omega, L)$ and $R(\omega, O)$ criteria than for $R(\omega, F)$ balance measure. The bottom panel of Table 3 displays

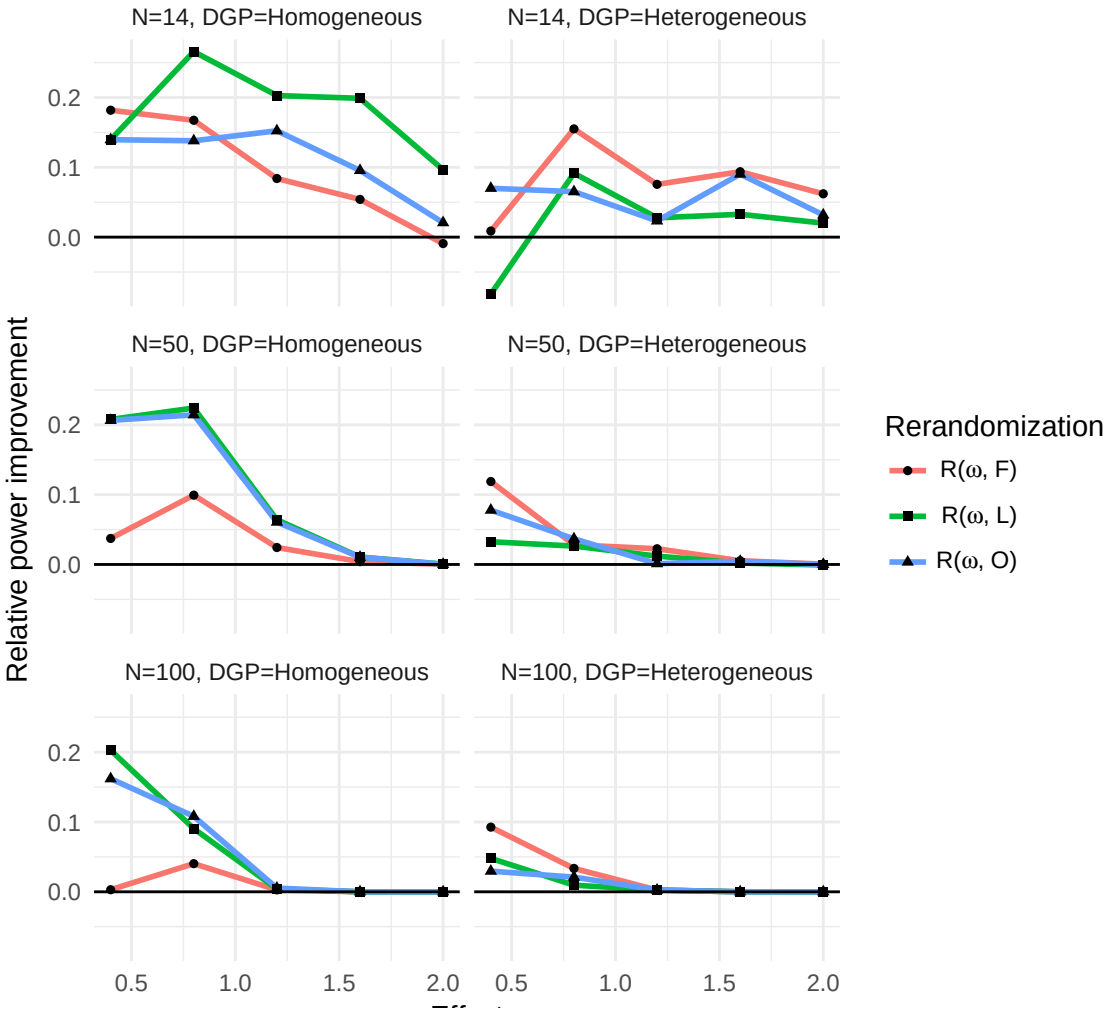

Figure 5: Relative power as compared to complete randomization for the ranked P-values (R) rerandomization designs for T=10. The left panel and right panel display the results from the Homogenous and the Heterogeneous DGPs, respectively.

the results with the Heterogeneous DGP. Also in this case, the variance reduction increases with t when $N \geq 50$ for all criteria. Once again we see that the variance reduction is of similar magnitudes for the $R(\omega, F)$ and $R(\omega, L)$ criteria but that the variance reduction of

$R(\omega, O)$ with $N = 14$ differs to a large extent from the variance reduction with larger N. The variance reduction in the effect estimate is of similar magnitude for $N = 14$ and 50. With $N = 100$, the $R(\omega, F)$ balance measure gives the largest variance reduction.

Figure 5 displays the relative power of the different rerandomization strategies as opposed to complete randomization under the Homogeneous (left panel) and Heterogeneous (right panel) DGPs. From the left panel one can see that for small effects the power gain is around 20% for the $R(\omega, O)$ and $R(\omega, L)$ criteria for all N. Furthermore, these criteria are superior to the $R(\omega, F)$ balance measure. From the right hand panels one can see that the power gains for $N = 50$ and 100 is around 10% for the $R(\omega, F)$ balance measure. For these sample sizes, this balance measure gives almost 100% larger relative improvement than the two other criteria. With $N = 14$ it is hard to get any improvements for any of these criteria in comparison to the complete randomization, at least with $R^2 \leq 0.25$.

Table 3: Variance reduction of the covariates and in the effect estimate as compared to complete randomization (Eq. 8 and 9) for the ranked P-values (R) rerandomization designs. The top and bottom panels display the results for the Homogeneous and Heterogeneous DGPs, respectively.

Rerandomization	$R(\omega, F)$			$R(\omega, L)$			$R(\omega, O)$		
N	14	50	100	14	50	100	14	50	100
	DPG=Homogeneous								
VC($Y1$|.)	-0.01	0.04	0.01	-0.07	-0.02	-0.01	-0.33	-0.17	-0.12
VC($Y2$|.)	-0.01	-0.02	-0.00	-0.09	-0.04	-0.05	-0.37	-0.25	-0.18
VC($Y3$|.)	-0.00	0.01	-0.04	-0.08	-0.03	-0.07	-0.39	-0.24	-0.25
VC($Y4$|.)	-0.01	-0.02	-0.06	-0.05	-0.03	-0.07	-0.42	-0.26	-0.20
VC($Y5$|.)	-0.08	-0.09	-0.08	-0.06	-0.11	-0.07	-0.42	-0.30	-0.16
VC($Y6$|.)	-0.12	-0.08	-0.09	-0.10	-0.10	-0.11	-0.41	-0.29	-0.22
VC($Y7$|.)	-0.17	-0.14	-0.16	-0.08	-0.08	-0.10	-0.39	-0.26	-0.25
VC($Y8$|.)	-0.28	-0.23	-0.25	-0.20	-0.24	-0.21	-0.45	-0.34	-0.32
VC($Y9$|.)	-0.37	-0.31	-0.32	-0.42	-0.62	-0.69	-0.51	-0.68	-0.73
VC($Y10$|.)	-0.39	-0.42	-0.39	-0.92	-0.95	-0.95	-0.69	-0.92	-0.93
VC_τ(.)	-0.10	-0.10	-0.13	-0.19	-0.27	-0.21	-0.14	-0.25	-0.24
	DPG=Heterogeneous								
VC($Y1$|.)	0.01	0.00	-0.03	-0.01	-0.03	-0.08	-0.29	-0.18	-0.17
VC($Y2$|.)	-0.03	-0.03	0.02	-0.10	-0.07	0.03	-0.33	-0.23	-0.10
VC($Y3$|.)	-0.05	-0.01	-0.11	-0.05	-0.01	-0.06	-0.30	-0.18	-0.17
VC($Y4$|.)	-0.08	-0.06	-0.05	-0.10	-0.05	-0.06	-0.35	-0.22	-0.18
VC($Y5$|.)	-0.10	-0.11	-0.08	-0.06	-0.04	-0.05	-0.35	-0.19	-0.17
VC($Y6$|.)	-0.14	-0.10	-0.15	-0.06	-0.05	-0.11	-0.36	-0.22	-0.20
VC($Y7$|.)	-0.19	-0.19	-0.21	-0.08	-0.15	-0.24	-0.36	-0.28	-0.28
VC($Y8$|.)	-0.19	-0.21	-0.16	-0.14	-0.25	-0.29	-0.38	-0.39	-0.34
VC($Y9$|.)	-0.29	-0.27	-0.24	-0.14	-0.09	-0.13	-0.35	-0.24	-0.20
VC($Y10$|.)	-0.31	-0.27	-0.30	-0.93	-0.96	-0.97	-0.71	-0.94	-0.95
VC_τ(.)	-0.06	-0.08	-0.10	-0.04	-0.07	-0.05	-0.06	-0.09	-0.06

4.4.2 Results with T=100

Table 4 displays the change in variance in the effect estimate for the two criteria. With the Homogeneous DGP, the $R(\omega, F)$ and $R(\omega, L)$ criteria give similar variance reductions in the range $19\% - 25\%$. With the Heterogeneous DGP the variance reduction obtained with the $R(\omega, F)$ balance measure is of the same magnitude as with the Homogeneous DGP. The variance reduction using the $R(\omega, L)$ balance measure is now only around 10%. The variance reduction in the 100 lags show a pattern (not displayed) very similar to the pattern presented in Table 3. The first 90 time periods have close to zero weights for both strategies and the last 10 have increasing weights. Figure 6 displays the relative power of the $R(\omega, F)$

Table 4: Relative change in variance in the estimated treatment effect as compared to complete randomization (Eq. 8 and 9) for the ranked P-values (R) rerandomization designs with forecast-based and LASSO estimated weights, respectively. The top and bottom panels display the results for the Homogeneous and Heterogeneous DGPs, respectively.

Rerandomization	$R(\omega, F)$			$R(\omega, L)$		
	Homogeneous					
$\text{VC}_\tau(.)$	-0.19	-0.25	-0.21	-0.22	-0.25	-0.24
	Heterogeneous					
$\text{VC}_\tau(.)$	-0.23	-0.26	-0.22	-0.07	-0.11	-0.06

and $R(\omega, L)$ criteria under the two DGPs. In the lefts panels, displaying the results from the Homogeneous DGP we can see that with the smallest effects size the increase in power is around 10% with $N = 14$ and around 20% with $N = 50$ and 100. For $N = 50$ and 100 the $R(\omega, L)$ balance measure perform better than the $R(\omega, F)$ balance measure. From the right hand panels, displaying the results with the Heterogeneous DGP, we can see that the $R(\omega, F)$ balance measure with $N = 50$ and 100 is in the range $20\% - 15\%$ in comparison to the complete randomization. With $N = 14$ the increase in power is only 5%. The $R(\omega, L)$ balance measure gives hardly any improvement in power in comparison to the complete randomization. In summary, when the number of time periods and sample size is small and the DGPs are heterogeneous, it is difficult to improve the power by the rerandomization

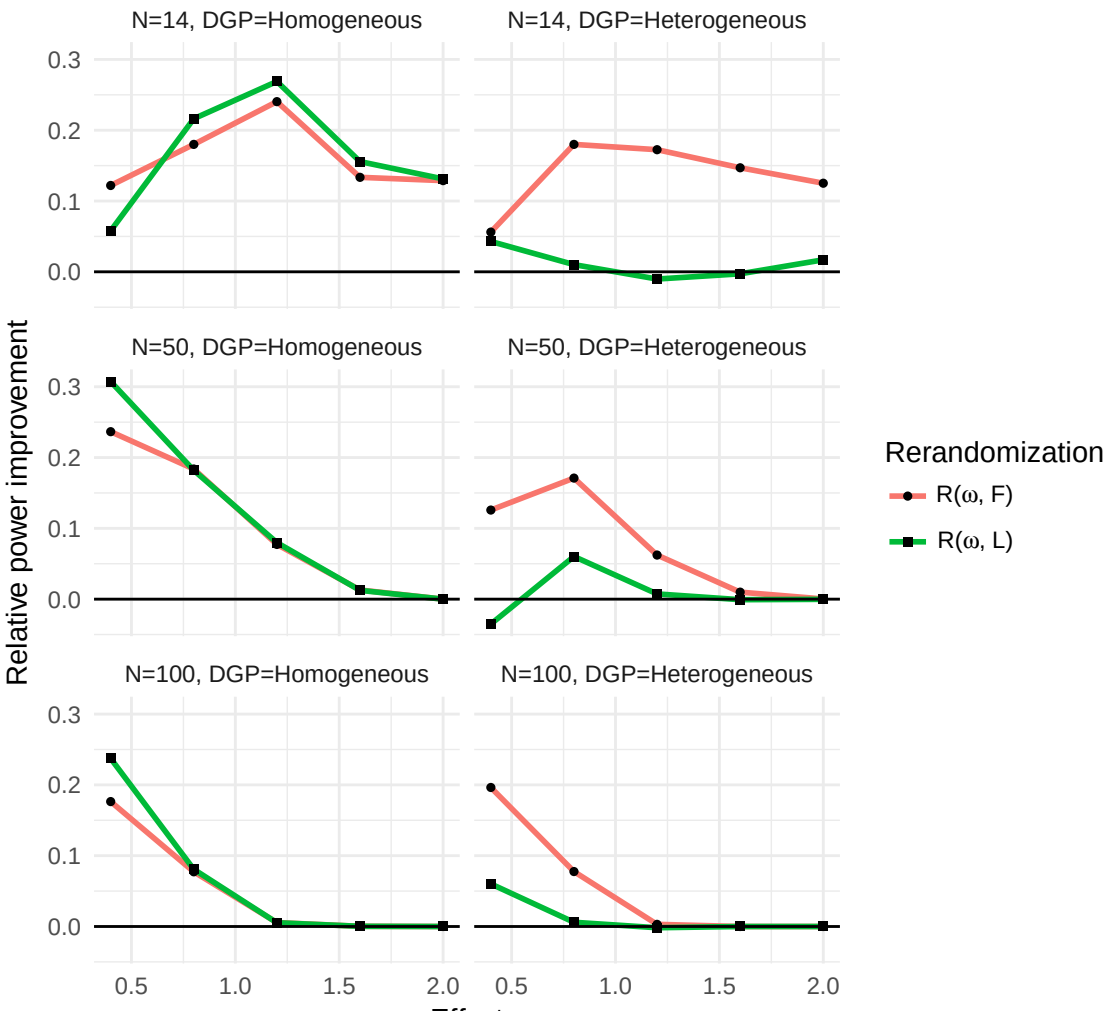

Figure 6: Relative power as compared to complete randomization the ranked P-values (R) rerandomization designs with forecast-based and LASSO estimated weights, respectively, for T=100. The left panel and right panel display the results from the Homogenous and the Heterogeneous DGPs, respectively.

strategies presented in this section. If however, either the sample size and the number of time periods increase, there are gains to be made using the strategies presented here. With long time series of pretreatment outcomes the are large gains for both considered DGPs. Of course, with large T, the level of heterogeneity can be evaluated by pre-analysis.

5 Empirical example - An information experiment on electricity consumption

This section illustrates how the proposed design strategies can be applied in practice in a small sample experiment. The pre-experimental data used in this example originates from Öhrlund et al. (2018), where a small randomized experiment is conducted. The interest is in how electricity consumption behavior can be affected towards a behavior more suitable for solar energy by providing information on energy consumption. Several outcomes are of interest in the original study, here we focus on one of them, the mean consumption difference between the group that received the information and the group that did not. The sample

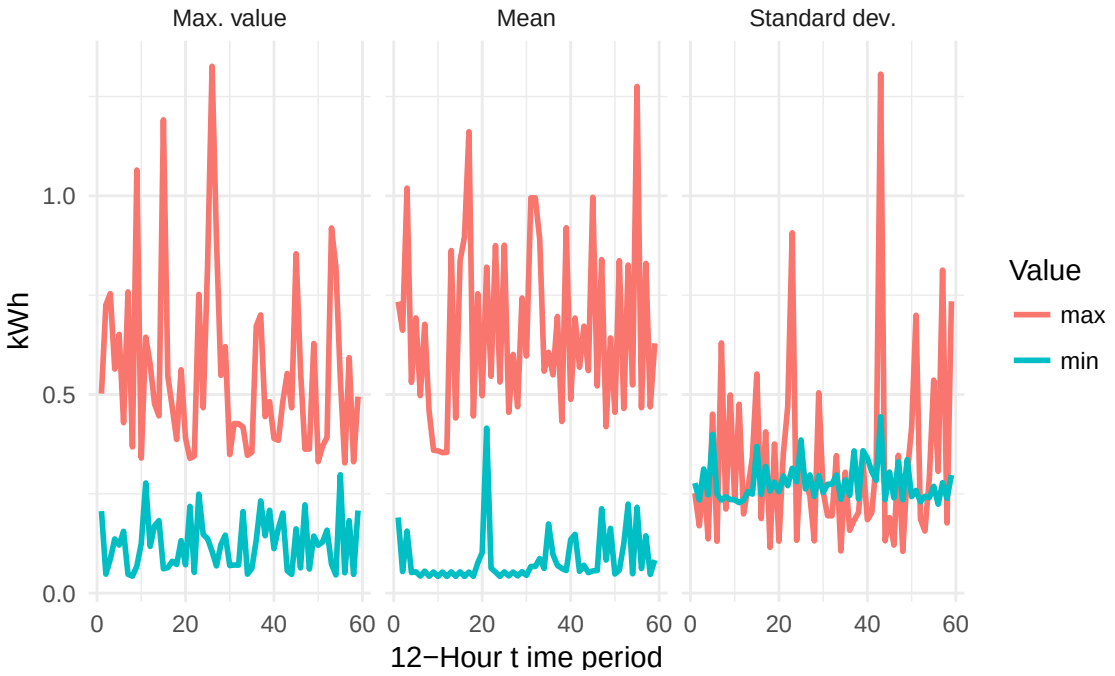

Figure 7: The electricity consumption (kWh) for the households with the largest (max) and smallest (min) maximum consumption, mean consumption, and standard deviation in their consumption, respectively.

consists of 54 households for which the electricity consumption is observed for each hour for the 4 months before treatment assignment. To increase the efficiency, a complex stratified randomization was used to assign treatment in the original study using all pre-treatment

data. Here, as an illustration, only the first 30 days are used, where consumption data are aggregated to 60 distinct 12-hour periods. The last period ($T + 1 = 60$) is left out from the design stage and is used to evaluate the design.

All the pre-treatment outcome time series are presented in Figure 1 . Figure 7 displays the heterogeneity in the pre-treatment outcome by showing the households with the smallest and largest maximum consumption value, the smallest and largest household mean, and the smallest and largest standard deviation over the pre-treatment time periods, in the panels from left to right, respectively. It is clear that there are quite large differences in several aspects of the electricity consumption between the households during the pre-treatment period and it is clearly not trivial to find a balanced design.

Since the pre-treatment data are measured with high frequency and no other covariates are available, the two strategies presented in the latter part of Section 4.4 are used, that is select the best allocations according the $R(\omega, L)$ and $R(\omega, F)$ criteria. Since the number of possible allocations equals $\binom{54}{27} = 1.946939\text{e}15$, the globally best allocations cannot be found and instead the procedure presented in Section 3 is applied. We chose here a resolution of $1/400$ implying $H = 800$, i.e., allocations were sampled randomly without replacement and the best 800 was kept. The procedure was left working for 11 hours which in this case meant that a random sample of one billion allocations were considered. As a benchmark to the rerandomization strategies, complete randomization was conducted. The exact p-value for the complete randomization was Monte Carlo approximated by a random draw of 40,000 allocations from the considered billion.

To evaluate the potential power gains under the proposed designs as compared to complete randomization, hypothetical homogeneous treatment effects where added to the treated group. That is, for each of the 800 selected allocations, the hypothetical treatment effect was added to Y_{60} for the treatment group and the exact p-value calculated. The same procedure was applied to the 40,000 randomly selected allocations (complete randomization), and the relative number of allocations that had exact p-values less than $\alpha = 0.05$ were compared by calculating the relative probability of drawing an allocation that rejects the null under the alternative when using $R(\omega, L)$ or $R(\omega, F)$ as compared to complete

randomization.

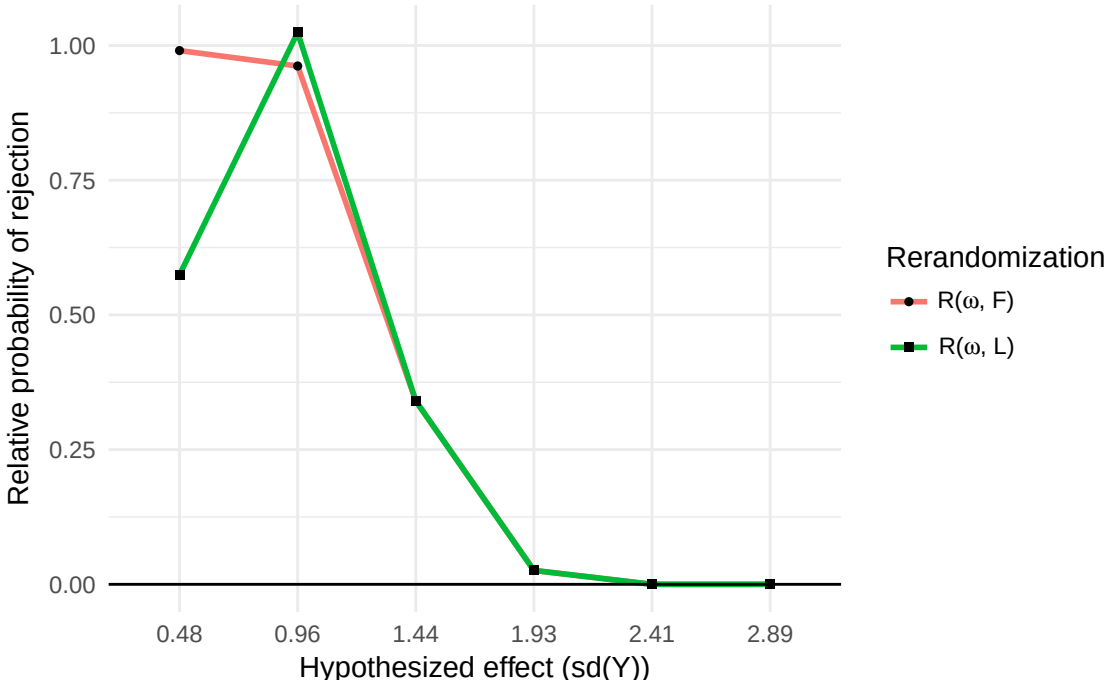

Figure 8: The relative probability as compared to complete randomization of randomly selecting an allocation that gives a significant result for two different rerandomization strategies given different hypothesized treatment effects.

Figure 8 displays the relative probability of rejecting the null under the alternative when random assignment is restricted to the set of allocation defined by the two criteria as compared to random assignment in the complete set of allocations.[8] From the figure we can see that for a half standard deviation (of the outcome) effect, the probability of obtaining a statistically significant effect after making a random draw from the 800 allocations determined by the $R(\omega, F)$ and $R(\omega, L)$ criteria is around 99% and 57% higher, respectively, than when making a random draw from the 40.000 randomly chosen allocations. It is worth noting that the forecast procedure, that is to use $R(\omega, F)$ balance measure is less computationally demanding than the $R(\omega, L)$ balance measure. The $R(\omega, L)$ takes around T time longer than the $R(\omega, F)$ balance measure to calculate [9].

[8]Note that there is no difference in rejection rate under the null as all tests are based on exact inference.

[9]This means that a more fair comparison might let $R(\omega, F)$ sample and consider T-times as many

An alternative way of displaying the difference between the complete randomization and the rerandomization strategies is to look at the empirical variance of the effect estimate under these designs. The variance of the effect estimate is thus obtained by estimating the effect for the restricted set of 800 allocations under the two criteria and for all 40,000 allocations for the complete randomization. Table 5 displays the percentage change in variance in the effect estimate compared to complete randomization. It is clear that, in comparison to complete randomization the variance using the $R(\omega, F)$ and $R(\omega, L)$ criteria is reduced by 56 and 61 percent, respectively.

Table 5: Relative variance change in the effect estimates across the possible allocation using $R(\omega, F)$ and $R(\omega, L)$ as compared to complete randomization (Eq. 9). Complete randomization is here Monte Carlo approximated by 40,000 random allocations.

Rerandomization	$R(\omega, F)$	$R(\omega, L)$
$\text{VC}_\tau(.)$	-0.56	-0.61

6 Discussion

Based on the results in Morgan and Rubin (2012), this paper develops strategies for rerandomization as a means to increase efficiency in randomized experiments. Morgan and Rubin (2012) suggest randomization based on the Mahalanobis distance balance measure of the mean difference vector between potential treated and controls. This paper proposes two things. First, a strategy to sample from set of admissible allocations fulfilling an implicit rerandomization balance measure to find the best possible design, given a balance measure, within a certain time limit. With the proposed sampling strategy and a symmetric balance measure, the sample average treatment effect (SATE) estimator is unbiased and the Fisher test is exact by design. Second, a new covariate balance measure is proposed as an alternative to the Mahalanobis distance. The balance measure differs from the Mahalanobis distance in several ways; it has computational advantages over the Mahalanobis balance

allocations.

measure in the situation of a large set of highly correlated covariates, and importantly, as the proposed balance measure expresses the weights of each covariate explicitly, it enables for various strategies of estimating the weights from data. For given a priori weights, the strategy can be considered an alternative to the strategy of Morgan and Rubin (2015) whom suggest rerandomization within tiers of importance. The proposed criterion balance measure is especially useful with one pre-treatment outcome or longitudinal pre-treatment outcome data. In this situation, the correlation structure of the data can be estimated using data to give different weights to the covariates and the pre-treatment outcomes accordingly.

The Monte Carlo simulations show: (i) that with traditional cross section data (i.e. only covariates) the suggested criterion has similar performance as the Mahalanobis criterion, (ii) an advantage with the new strategy to the Mahalanobis criterion when one or several pre-treatment outcomes are available. Finally (iii), two Monte Carlo simulations with only pre-treatments observations (as in the empirical illustration) shows advantages with the new strategies in comparison to complete randomization.

Taking use of a sample of 54 households with electricity consumption over 60 time periods, it is shown that the power of a mean difference test in a balanced randomized experiment can be increased by up to 100% using one of the proposed rerandomization strategies as compared to complete randomization.

References

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Hamaker, E. L., T. Asparouhov, A. Brose, F. Schmiedek, and B. Muthén (2018). At the Frontiers of Modeling Intensive Longitudinal Data: Dynamic Structural Equation Models for the Affective Measurements from the COGITO Study. *Multivariate Behavioral Research 3171*, 1–22.

Hamaker, E. L. and M. Wichers (2017). No Time Like the Present. *Current Directions in Psychological Science 26*(1), 10–15.

Hu, Y. and F. Hu (2012). Balancing treatment allocation over continuous covariates: A new imbalance measure for minimization. *Journal of Probability and Statistics 2012*.

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

Li, X., P. Ding, and D. B. Rubin (2018, sep). Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences 115*(37), 9157 LP – 9162.

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

Morgan, K. L. and D. B. Rubin (2015). Rerandomization to Balance Tiers of Covariates. *Journal of the American Statistical Association 110*(512), 1412–1421.

Öhrlund, I., B. Stikvoort, and M. Schultzberg (2018). Taking the sun in our on hands: Implications of shared residential solar installations on pre-environmental behaviours (Memeo).

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso.

Zhou, Q., P. A. Ernst, K. L. Morgan, D. B. Rubin, and A. Zhang (2018, jun). Sequential rerandomization. *Biometrika*.