UPPSALA
UNIVERSITET

Working Paper 2018:4

Department of Statistics

Experimental design using longitudinal data

Per Johansson and Mårten Schultzberg

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

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

Title: Experimental design using longitudinal data
Author: Per Johansson and Mårten Schultzberg
E-mail: marten.schultzberg@statistik.uu.se

Experimental design using longitudinal data

Per Johansson and Mårten Schultzberg

November 5, 2018

Abstract

This paper considers rerandomization designs that increase the efficiency in experiments. The focus is on finding rerandomization criteria utilizing pre-treatment outcome data, especially longitudinal outcome data. The proposed criteria have computational advantages with a large set of covariates and provides a convenient way to give different weights to different covariates. In addition, the proposed criteria are especially useful when the pre-treatment outcome data can be used to estimate the importance of different covariates in the design. Using Monte Carlo simulations, the proposed strategies building on the new criteria are compared to complete randomization and Mahalanobis-based rerandomization. Using high frequency longitudinal electricity consumption data on 54 households, the power of a mean difference test is increased by 80%, in comparison to complete randomization, using one of the proposed design strategies.

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[1] 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[2] 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 suggested by Fisher in the early twentieth century and has since been mentioned in the literature 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. This strategy 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. It should be stressed that, as also Morgan and Rubin (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 available covariates within these strata.

One potential caveat with rerandomization is that common test-statistics are no longer asymptotic normally distributed, often yielding highly conservative tests and confidence intervals (Morgan and Rubin, 2012, 2015; Li et al., 2018). Li et al. (2018) derive the asymptotic sampling distribution of the estimated difference-in-mean statistic, using the affinely invariant Mahalanobis distance between the covariate-mean vectors of the treated and control in the potential allocations as the balance criterion for rerandomization. One simple, although computationally demanding, alternative,

[1] Covariates can include pre-experiment measurements of the outcome.

[2] 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.

is to restrict the inference to the units in the experiments and to base the analysis on exact (randomization) inference (Fisher, 1935). This is the inference strategy used in Morgan and Rubin (2012). Based on the Mahalanobis distance criterion, Morgan and Rubin (2012) show that rerandomization decreases 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. Zhou et al. (2018) extend this framework further for the situation when sequential randomization is required.

The primary focus of this paper is to develop rerandomization strategies that are easy to use for (possibly high frequency) longitudinal data, 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 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 hence 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.

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 all 54 households in the sample 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 of the consumption behaviour 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. 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 80% as compared to complete randomization. All the details of this study are presented in Section 5.

[Figure 1 about here.]

The focus throughout this paper is, as in Morgan and Rubin (2012, 2015), and Zhou et al. (2018), on exact inference to the units of the experiment. The assumption-free nature of randomization tests make them highly attractive for rerandomization strategies, especially considering today's software and computer power. With a rerandomization design, the randomization test is performed across the allocations fulfilling the criterion only, implying that the randomization tests performed after the rerandomization is less computationally intensive than would be the case with complete randomization. However, as will be discussed in detail, finding the allocations fulfilling the criteria is often computationally demanding.

The following section introduces the concept of rerandomization. Section 3 introduces a new rerandomization criterion based on the ranks of the mean differences in the covariates. This criterion has computational advantages to the Mahalanobis-distance criterion in the situation of a large set of covariates, and in addition, it provides a convenient way of giving different weights to different covariates according to a priori information of their relative importance, which could be considered as an alternative to the strategy suggested in Morgan and Rubin (2015). Based on the proposed criterion, a number of different strategies to estimate the weights for the covariates using per-treatment outcome data are suggested. Section 4 provides Monte Carlo simulations where the different strategies are compared with complete randomization and the Mahalanobis-based rerandomization criterion (when applicable). Section 5 provides the empirical analysis of electricity consumption data. Section 6 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 extension in Morgan and Rubin (2015), Zhou et al. (2018) and Li et al. (2018).

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

$$M = \frac{N}{2}(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C)' cov(\mathbf{x})^{-1}(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C),$$

where $\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C$ is the difference in mean vectors which is a $K \times 1$ stochastic vector as it depends on all possible allocations. Morgan and Rubin (2012) suggest randomizing within a subset of all allocations, say $\mathbf{W}^*$, for which $M \leq a$. If the means are normally distributed $M \sim \chi^2(K)$,

$$Cov(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C|\mathbf{x}, M \leq a) = \nu_a Cov(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C|\mathbf{x}), \tag{1}$$

where

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

which implies that the variance of the covariates in the subset of allocations is reduced in comparison to the complete randomization. The percent reduction in variance of all of the (equally weighted) included covariates is equal to

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

Let R^2 be the coefficient of determination of a regression of $\mathbf{Y}(0)$ on $\mathbf{x}_i$ then under the further assumptions of conditionally normally distributed outcomes and additive treatment effects the percent reduction in variance on the treatment effect is equal to

$$R^2(1-\nu_a). \tag{3}$$

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 descreasing in K. Given

$$p_a = \Pr(M \leq a),$$

it is evident that a can be indirectly determined by setting p_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 this means that the time it takes to find the allocations from which to finally make the randomization is independent of N for a fixed p_a.

2.1 Chosing the best set of allocations

The exact p-value 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 allocations. In complete randomization there are $\binom{2N}{N}$ possible allocations. To calculate the exact p-value, the test statistic is calculated for all $j = 1, ...\binom{2N}{N}$ allocations to create the histogram. The implication of this procedure is that the smallest possible p-values 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[3] 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. This means that, for a resolution

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

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 mean difference, all allocations are likely to be unique.

As the number of possible allocations grows very fast in N, it is in practice the computational time that limits the possibility of finding the globally best allocations for most N. We argue that the goal with any rerandomization design is to find, within a certain time limit, the best possible set of allocations to obtain the most efficient experiment. Therefore, instead of rerandomizing until some pre-specified criterion is met, we propose the following procedure to obtain the approximate best allocations with large N. Consider a balanced experiment as an example: Decide an approximate level of resolution $r = 2/(\#$ unique $S)$. With no ties this give the number of best allocation $H = 2/r$, e.g., with an aim of having a resolution of $1/400$ we should chose the 800 best allocations. From the complete set of allocations, randomly sample a set containing $H^* \geq H$ allocations, without replacement. Store the H best allocation from $\mathbf{A}_1$ (according to the chosen rerandomization criteria) in the set $\mathbf{W}^*$. Sample a new set of allocations, $\mathbf{A}_2$, of size H^* from the $\binom{2N}{N} - H^*$ remaining allocations. Replace the set $\mathbf{W}^*$ with the H best allocations in the set $\{\mathbf{W}^*, \mathbf{A}_2\}$. Sample a new set of allocations, $\mathbf{A}_3$, of size H^* from the $\binom{2N}{N} - 2H^*$ remaining allocations. Replace the set $\mathbf{W}^*$ with the H best allocations in the set $\{\mathbf{W}^*, \mathbf{A}_3\}$. Repeat this I times or until all $\binom{2N}{N}$ allocations have been considered. The total number of considered allocations is $H^*_{Tot} = I \times H^*$, where I is the number of iterations. The H allocations stored in $\mathbf{W}^*$ in the final iteration are the best allocations within the set $\{\mathbf{A}_1, ..., \mathbf{A}_I\}$. 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 some computational time restriction. In the empirical example, presented in detail in Section 5, the allocation sampling algorithm was left to work overnight (11 hours) after which about one billion allocations had been considered and the 800 best had been retrieved.

To contrast our procedure of finding allocations, using the procedure proposed in Morgan and Rubin (2012), all the allocations within the set of included allocations have Mahalanobis distances smaller than or equal to the criterion a. This implies that the expected variance reduction is an upper bound. If for example, $a = 0.3$ during the sampling of allocations but all the H included allocations have Mahalanobis distances smaller than 0.3, the expected variance reduction is underestimated. If instead the Mahalanobis distance is used in the procedure proposed above, the actual

criterion a used in practice is directly given by

$$a = \max_{\mathbf{W}^*}(M|w \in \mathbf{W}^*).$$

That is, instead of being pre-specified by the researcher, the criterion a is given from the included allocations ensuring that the expected variance reduction estimate is as accurate as possible. In practice, the difference in the expected variance reduction estimate will be very small when a is small in both procedures.

If the sample size is small enough for the available computational power to go through all $\binom{2N}{N}$ allocations, the H globally best allocations should of course be selected.

2.1.1 Mirror allocation

When finding the set of the approximately best allocations, it is a good idea to always include the corresponding mirror allocations. To clearly define a mirror allocation, consider the following example. Consider an experiment where $n = 4$, and two units are randomized into treatment and control, respectively. Say that, e.g., unit 1 and 2 are assigned to treatment and units 3 and 4 to control. The corresponding mirror allocation is then the allocation where unit 3 and 4 is assigned treatment, and unit 1 and 2 to control. This concept extends to more than two treatment groups. The number of mirror allocations are given by $G! - 1$, where G is the number of treatment groups.

There are several reasons for including mirror allocations in the set of best allocations. The first reason is that is saves time; if an allocation has a small rerandomization criterion, the mirror allocation(s) has exactly the same criterion by definition, which means that the time it takes to compare the criterion of H^*_{Tot} allocations is reduced by a factor of $1/G!$. The second reason is symmetry of the balance improvement implied by the rerandomization. The distribution of the test-statistic value around the value implied under the null is symmetric by definition using complete randomization (due to the mirror allocations). However, when rerandomization is used this is not guaranteed unless the procedure finds $H/2$ non-mirror allocations and then include the $H/2$ mirror allocations. If the H globally best allocations are found this is guaranteed. A symmetric balance improvement implies that the probability of randomly drawing an allocations with a test-statistic value with certain deviation from the null is equally large for a positive and a negative deviation. When many allocations are considered the included allocations will be balanced approximately, as they are a random dray from a symmetrical distribution, however, given the time saving aspect

this might as well be fixed in the design. In the empirical example, described in detail in Section 5, the set of the H best allocations contain H/2 unique allocations and their corresponding mirror allocations.

3 Rerandomization using the ranks of the mean differences in covariates

This section presents an alternative rerandomization criterion based on the rank of the covariate mean difference. The suggested criterion 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 criterion 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 and $\mathbf{w}$ 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 $\binom{N}{N/2}$ allocation as

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

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

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

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

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

$$R_j = \left(\sum_{k=1}^{K} \omega_k p_{jk} \right), \tag{4}$$

where ω_k is the weight given to covariate k, with $\sum_{k=1}^{K} \omega_k = 1$. The H allocations with the largest

$R_j, j = 1, ...\binom{N}{N/2}$ gives the set $\mathbf{W}_H$. This follows since the largest p-value implies the smallest mean difference.

With one single covariate, i.e. $K = 1$, the H allocations with the smallest imbalances, i.e. the allocations with the largest R_j among $j = 1, ...\binom{N}{N/2}$, are the same allocations that have the H smallest Mahalanobis distances. This implies that with one covariate the rerandomization based on R 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 (3). However, for $K > 1$ the two criterion may give different sets of allocations.

Note that in the calculation of R_j the correlation structure of the covariates are incorporated. For two positively or negatively correlated covariates, the p-values of the standardized mean difference across the allocations will be positively correlated. This follows since any large mean difference will give a small p-value and vice versa, regardless of the sign of the mean difference due to the absolute sign in Equation 4.

Throughout the remainder of this paper we denote the new criterion R, while the Mahalanobis distance criterion is denoted M. R with uniform weights are simply denoted R while if the R criterion is based on non-uniform weights 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 criterion is a complement to the Morgan and Rubin (2015) which is simple to implement, allowing the weights $\omega_k, k = 1, ..., K$, be based on a priori information on 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.

3.1 Estimating the weights from pre-treatment data

This section considers ways to estimate the weights for the R criterion using pre-treatment outcomes. 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.

3.1.1 One pre-treatment outcome

If the outcome Y is observed before the treatment assignment, and the pre-treatment outcome, $Y_{t=T}$, 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. The most obvious strategy is to estimate the weights using the partial correlations of the covariates and the pre-treatment outcome. That is, estimate

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

using ordinary least square (OLS). Here $\sim$ indicate a standardized variables, i.e. $\widetilde{z}_i = (Z_i - \overline{Z})/\sqrt{\widehat{Var(Z_i)}}$.

With K covariates and one pre-treatment outcome we have $Q = K + 1$ variables to base rerandomization on.[4] Denote ω_0 the weight for $\widetilde{y}_{i0}$ then $\omega = (\omega_0, \omega_1, ..., \omega_K)'$ and the estimated $\widehat{\omega}' = (\frac{|\widehat{\beta}_1|}{\delta}, ..., \frac{|\widehat{\beta}_K|}{\delta}, \frac{1}{\delta})$ 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 practice, 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, LASSO is still useful to estimate the weights for the R criterion, however, for larger numbers of time periods other strategies for performing the rerandomization might be preferable as is discussed in the next section.

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

3.1.2 Several pre-treatment outcomes

If only repeated pre-treatment outcomes are observed, $R(\omega, L)$ is still useful. However, as the p-values must be calculated for each pre-treatment time period with non-zero esitmated weight, the time consumption may increases drastically with the number of pre-treatment time periods. An alternative in this situation is to instead 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, 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 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 Monte Carlo simulation studies

Throughout this Monte Carlo study, 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.[5] For this reason, the relative power under the alternative is also presented.

[5]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

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 (e.g. $d = (M, R)$ in the first set of Monte Carlo experiments). For each arm of the study, 4000 replications (N_{rep}) are considered.

The relative variance reduction of all the covariates and pre-treatment outcomes in time period T in comparison to complete randomization is then defined:

$$\text{Relvar}_r(q) = \frac{Var(\overline{\mathbf{z}}_{rqT} - \overline{\mathbf{z}}_{rqC}|d) - Var(\overline{\mathbf{z}}_{rqT} - \overline{\mathbf{z}}_{rqC}|c)}{Var(\overline{\mathbf{z}}_{rqT} - \overline{\mathbf{z}}_{rqC}|c)}, \tag{5}$$
$$q = 1, ..., K, \mathbb{1}(T > 0)(K + 1), ...\mathbb{1}(T > 0)(Q - K)$$

where $r = 1...N_{rep}$, $\overline{\mathbf{z}}_{lT} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_{il} w_i$ and $\overline{\mathbf{z}}_{lC} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_{il}(1-w_i)$, and w_i being equal to one or zero if treated or control, i.e. the realized treatment allocation of unit i. $\mathbb{1}$ is an indicator functions taking the value 1 if the condition is fulfilled and zero otherwise. In the results, the average Relvar across the N_{rep} replicates are presented, i.e.

$$\bar{\text{Relvar}}(q) = \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} \text{Relvar}_r(q) \tag{6}$$

Let

$$\hat{\tau}_c = \frac{1}{N}\sum_{i=1}^{N} w_i y_i - \frac{1}{N}\sum_{i=1}^{N}(1 - w_i) y_i$$

and

$$\hat{\tau}_d = \frac{1}{N}\sum_{i=1}^{N} w_i y_i - \frac{1}{N}\sum_{i=1}^{N}(1 - w_i) y_i | \mathbf{w} \in \mathbf{W}^*.$$

The relative variance reduction of the treatment effect of strategy d is

$$\text{Relest}(d) = \frac{Var(\hat{\tau}_d) - Var(\hat{\tau}_c)}{Var(\hat{\tau}_c)} \tag{7}$$

---

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.

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

$$\pi_{r,c} = \Pr(|\widehat{\tau}_{rd}(\mathbf{W}, \mathbf{Y})|) \geq |\widehat{\tau}_{rd}|$$

and

$$\pi_{r,d} = \Pr(|\widehat{\tau}_{d}(\mathbf{W}, \mathbf{Y}|\mathbf{W} \in \mathbf{W}_d^*)) \geq |\widehat{\tau}_{rd}|$$

, respectively. $\widehat{\tau}_{rc}$ and $\widehat{\tau}_{rd}$ are the realized estimates of the complete and rerandomization designs in replication r, respectively. $\widehat{\tau}_{rc}(\mathbf{W}, \mathbf{Y})$ is the distribution of estimates for all allocations in replication r. The corresponding distribution for the subset $\mathbf{W}_d^*$ containing the allocations selected by rerandomization is denoted $\widehat{\tau}_{rd}(\mathbf{W}, \mathbf{Y}|\mathbf{W} \in \mathbf{W}_d^*)$. The relative power is evaluated with the true treatment effect being varied from $0.4\sigma_Y$ to $2\sigma_Y$ in steps of $0.4\sigma_Y$, and estimated as

$$\text{Power}(\tau, d) = \frac{p_{\tau,d} - p_{\tau,c}}{p_{\tau,c}}, \tau = 0.4, 0.8, ..., 2.00, \tag{8}$$

where

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

and

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

for $d = M, R, R(\omega, O), R(\omega, L), R(\omega, F)$.

4.1 Cross section data and one pretreatment outcome

Consider the data generating process

$$Y_{it} = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_{it}, i = 1, ...N, t = 1, ..., T + 1. \tag{9}$$

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., $\zeta_{it} \sim N(0, \sigma^2)$.[6] Due to independence the marginal variance of the outcome

[6]The results from the Monte Carlos 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 (2) and (3) for small N.

is equal to

$$\text{Var}(Y) = \boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}' + \frac{\sigma^2}{1-\phi^2}. \tag{10}$$

We compare the proposed rerandomization criterion, 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.

4.1.1 Cross section

This section serves to compare the proposed criterion with the criterion presented in Morgan and Rubin (2012). As this is not the main focus of this paper, the results are limited to the case with $N = 14$. With $N = 14$, the 800 allocations with the (globally) smallest value of the criterion 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 about here.]

[Figure 2 about here.]

Table 1 displays the relative variance reduction in the covariates and the estimated treatment effect. The variance reduction of the covariates for the M criterion 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 2). The variance reductions of the covariates using the R criterion is around 71% under data generating process (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 criterion of 19.7% (cf. 3).

Figure 2 displays the 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.

4.1.2 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 data generating process (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 incorprates 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 criterion 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 are used for each criteria. As $p_a = 0.233$ for the Mahalanobis criterion 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 criterion, the expected variance reductions in the covariates are 71% and 49% for $Q = 4$ and $Q = 11$, respectively (cf. Equation 2). The corresponding variance reduction in the variance of the effect estimate are 17.99% and 12.25% percent (cf. Equation 3).

When N is equal to 50 and 100, the $4,000$ allocations are randomly sampled from all possible allocations to limit the computational time. 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 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 given in (2) and (3) for the M criterion may be less exact for $N = 50$ and 100.

Table 2 displays the variance reduction in the covariates and the effect estimates for the four

criteria compared to complete randomization for both DGP's. Figure 3 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 criterion 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 $R(\omega, O)$ criterion is performing best overall.

From the lower panel displaying the results with $Q = 11$, it is clear that the variance reductions of the M criterion and the R criterion 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 criterion. However with $N = 14$ the variance reduction is much worse 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 criterion 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 same 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 $R(\omega, L)$ criterion are better. Except for $N = 14$, the $R(\omega, O)$ criterion is giving the largest variance reduction in the estimated treatment effect. With $N = 14$, the $R(\omega, L)$ criterion performs better than the $R(\omega, O)$ criterion.

[Table 2 about here.]

The left panels of Figure 3 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 $R(\omega, L)$ criterion is substantial also compared to the $R(\omega, O)$ criterion.

Turning to the right panels 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 $R(\omega, L)$ criterion 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 100% better 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.

[Figure 3 about here.]

4.2 Longitudinal data

In this section two different time series DGPs 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

$$Y_{it} = 0.5 \times Y_{it-1} + \zeta_{it}, i = 1, ...N, t = 1, ...T + 1. \tag{11}$$

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

$$Y_{it} = \boldsymbol{\phi}_j \boldsymbol{Y}_{i,lag} + \epsilon_{it} \ , i = 1, ..N, t = 1, ..., T + 1, j = 1, ..4, \tag{12}$$

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 $R(\omega, F)$[7] criteria are used. Given the large number of correlated pre-treatment outcomes the $R(\omega, O)$ criterion runs into singularity problems in the estimation with $T = 100$ and is therefore excluded in this case.

[7]The `auto.arima` in the `Forecast` package in R is used for all forecasts.

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_1 the weight for $\widetilde{y}_{i1}$ 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 $\delta = \sum_{t=1}^{T-1} |\widehat{\beta}_t| + 1$. That is, as in the previous sections, the weight of the time point before treatment assignment is fixed to one which implies $\omega_T \geq \omega_t \, \forall \, t = 1, ..., T-1$ in expectation. Note that there are of course many alternatives for setting the weight of the last time period, however, this matter is not discussed further here.

4.2.1 Results with T=10

[Table 3 about here.]

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 $R(\omega, F)$ criterion.

The bottom panel of Table 3 displays 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 $R(\omega, F)$ criterion gives the largest variance reduction.

Figure 4 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 $R(\omega, F)$ criterion. From the right hand panels one can see that the power gains for $N = 50$ and 100 is around 10% for the $R(\omega, F)$ criterion. For these sample sizes, this criterion gives almost 100% larger rela-

tive 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 $R^2 \leq 0.25$.

[Figure 4 about here.]

4.2.2 Results with T=100

Table 4 displays the variance reduction 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 $R(\omega, F)$ criterion is of the same magnitude as with the Homogeneous DGP. The variance reduction using the $R(\omega, L)$ criterion 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.

[Table 4 about here.]

Figure 5 displays the relative power of the $R(\omega, F)$ 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 $N = 50$ and around 20% with $N = 50$ and 100. For $N = 50$ and 100 the $R(\omega, L)$ criterion perform better than the $R(\omega, F)$ criterion. From the right hand panels, displaying the results with the Heterogeneous DGP, on can see that the $R(\omega, F)$ criterion 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 $R(\omega, L)$ criterion gives hardly any improvement in power in comparison to the complete randomization

In summary, when the number of time points and sample size is small and the DGPs are heterogeneous, it is difficult to improve the power by the rerandomization 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.

[Figure 5 about here.]

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 data is originates from **?**, 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 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 6 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.

[Figure 6 about here.]

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 4.2 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 2.1 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 overnight (11 hours) which for this setting 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.

Figure 7 displays the relative probability of randomly drawing an allocation that rejects the null under the considered treatment effect, using rerandomization as compared to complete randomization. This means that, if for a given rerandomization strategy and a given treatment effect the relative probability is 0.2, the probability of randomly drawing an allocation from the 800 for which the null is rejected is 20% larger than randomly drawing an allocation from the 40.000 for which the null is rejected. For a zero treatment effect there is no difference in rejection rate by definition as both are based on exact inference which guarantees the size.

From Figure 7 it is clear that for a half standard deviation (of the outcome) effect, the probability of drawing an allocation using the $R(\omega, F)$ and $R(\omega, L)$ criteria is around 80% and 70% higher, respectively, than with complete randomization. It is worth noting that the forecast procedure, that is to use $R(\omega, F)$ criterion is less computationally demanding than the $R(\omega, L)$ criterion. The $R(\omega, L)$ takes around T time longer than the $R(\omega, F)$ criterion to calculate [8].

An alternative way of displaying the difference between the complete randomization and the rerandomization strategies is to look at the variance of the effect estimate under these designs. The variance of the effect estimate is obtained by estimating the effect for all 800 allocations when doing the rerandomization and for all 40,000 allocations for the complete randomization. Table 5 displays the percentage variance reduction 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. Note that this is not the same variance reduction measure as that presented in the Monte Carlo simulations result tables.

[Table 5 about here.]

[Figure 7 about here.]

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

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. Based on the Mahalanobis distance of mean difference between potential treated and controls Morgan and Rubin (2012) suggest randomization from a limited (instead of the full set as in complete randomization) set of allocations fulfilling a given criterion on the Mahalanobis distance of the covariates means for potential treated and controls. This paper proposes an alternative criterion to the Mahalanobis distance that is based the ranks of the mean differences in covariates of potential treated and controls. This new criterion has computational advantages over the Mahalanobis criterion in the situation of a large set of covariates. Importantly, as the proposed criterion expresses the weights of each covariate explicitly, the criterion enables for various strategies of estimating the weights from data. With given a priori weights, the strategy can be considered an alternative to the strategy of Morgan and Rubin (2015) who suggest rerandomization within tiers of importance. The proposed criteria, and the associated strategies, are 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 outcome accordingly.

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 80% using one of the proposed rerandomization strategies as compared to complete randomization.

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 is available. Finally (iii), two Monte Carlo simulations with only pre-treatments observations (as in the empirical illustration) shows the advantage with the new strategies in comparison to complete randomization.

References

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

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., and Muthén, B. (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 Wichers, M. (2017). No Time Like the Present. *Current Directions in Psychological Science*, 26(1):10–15.

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

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

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic Theory of Rerandomization in Treatment-Control Experiments. To appear in PNAS.

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

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

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

Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., and Zhang, A. (2018). Sequential rerandomization. *Biometrika*, pages asy031–asy031.

List of Figures

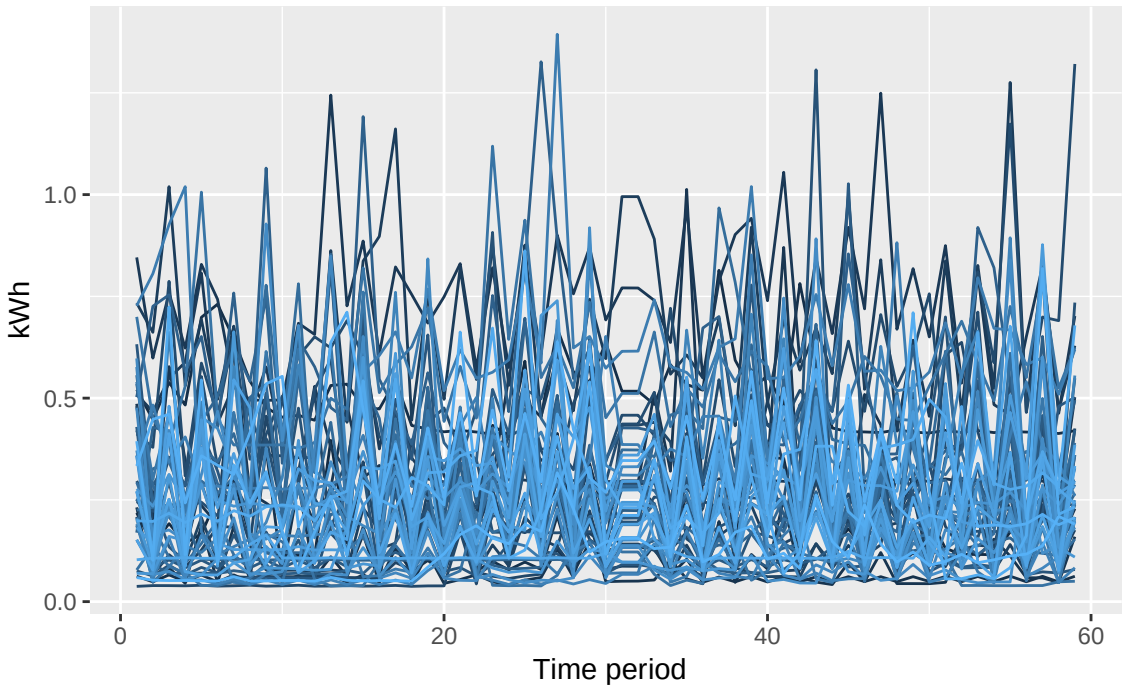

Figure 1: The electricity consumption (kWh) during the pre-treatment month for the 54 households included in the Experiment.

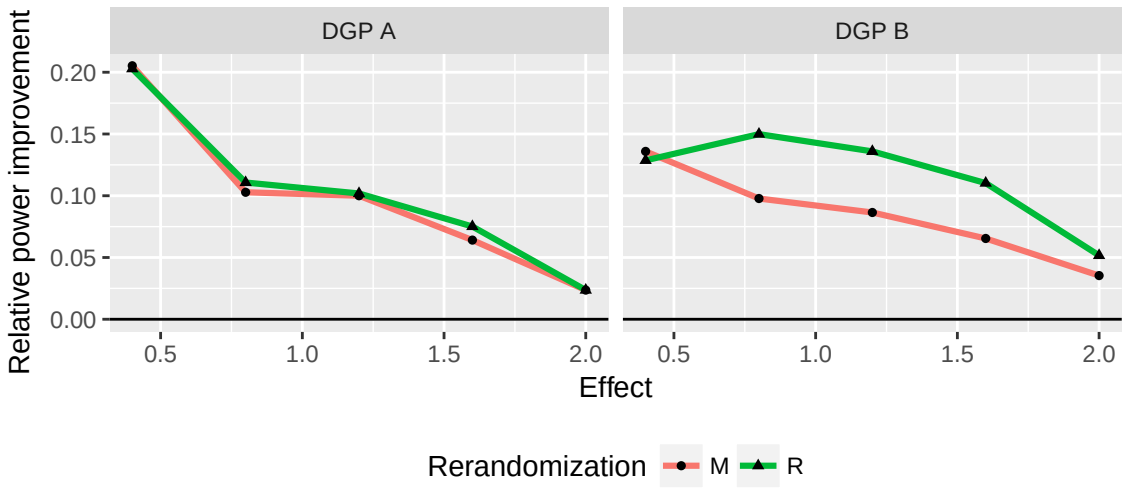

Figure 2: Relative power as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The left and right figures display the results for DGPs A and B, respectively.

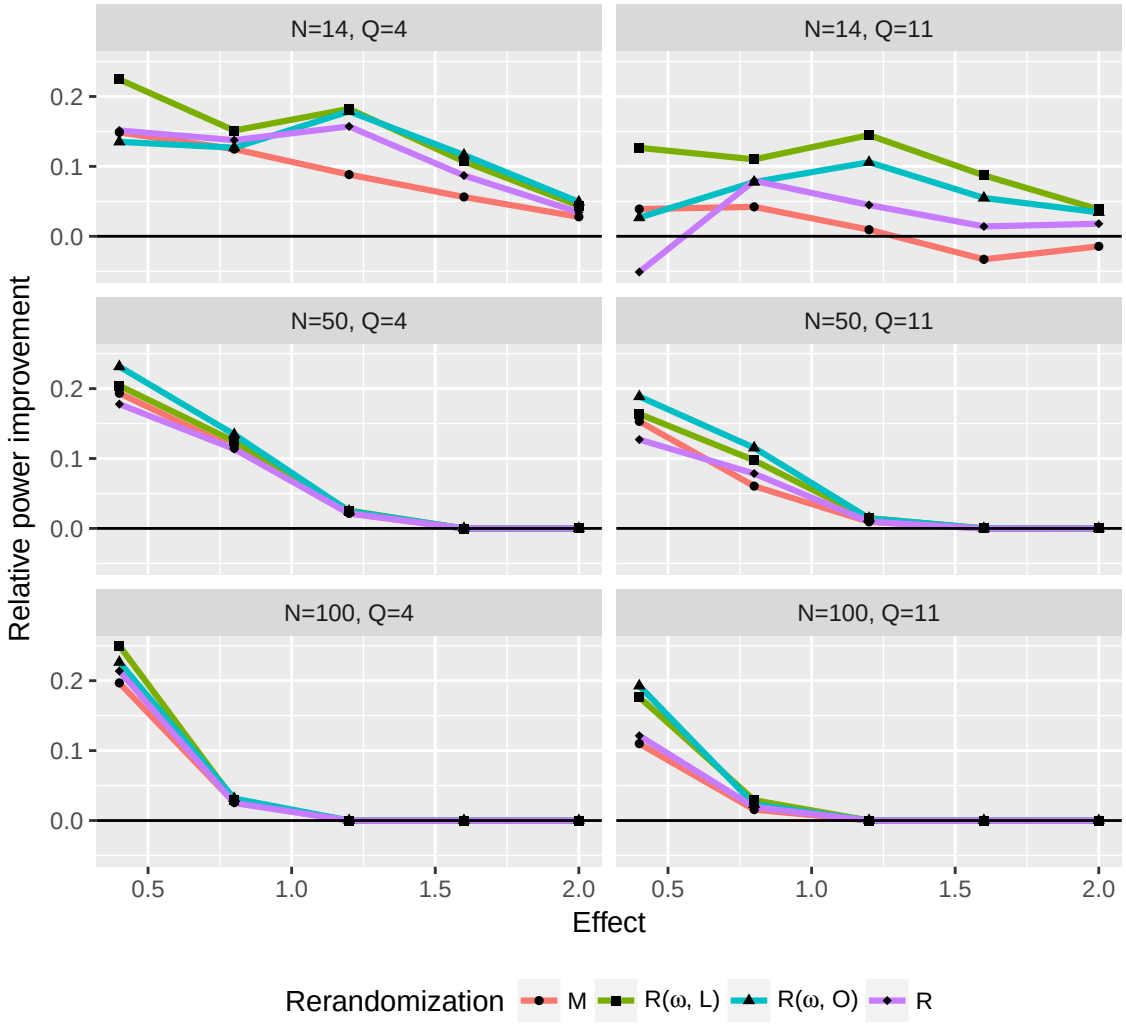

Figure 3: Relative power as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The left and right panels display the results including four(Q=4) and eleven (Q=11) covariates, respectively.

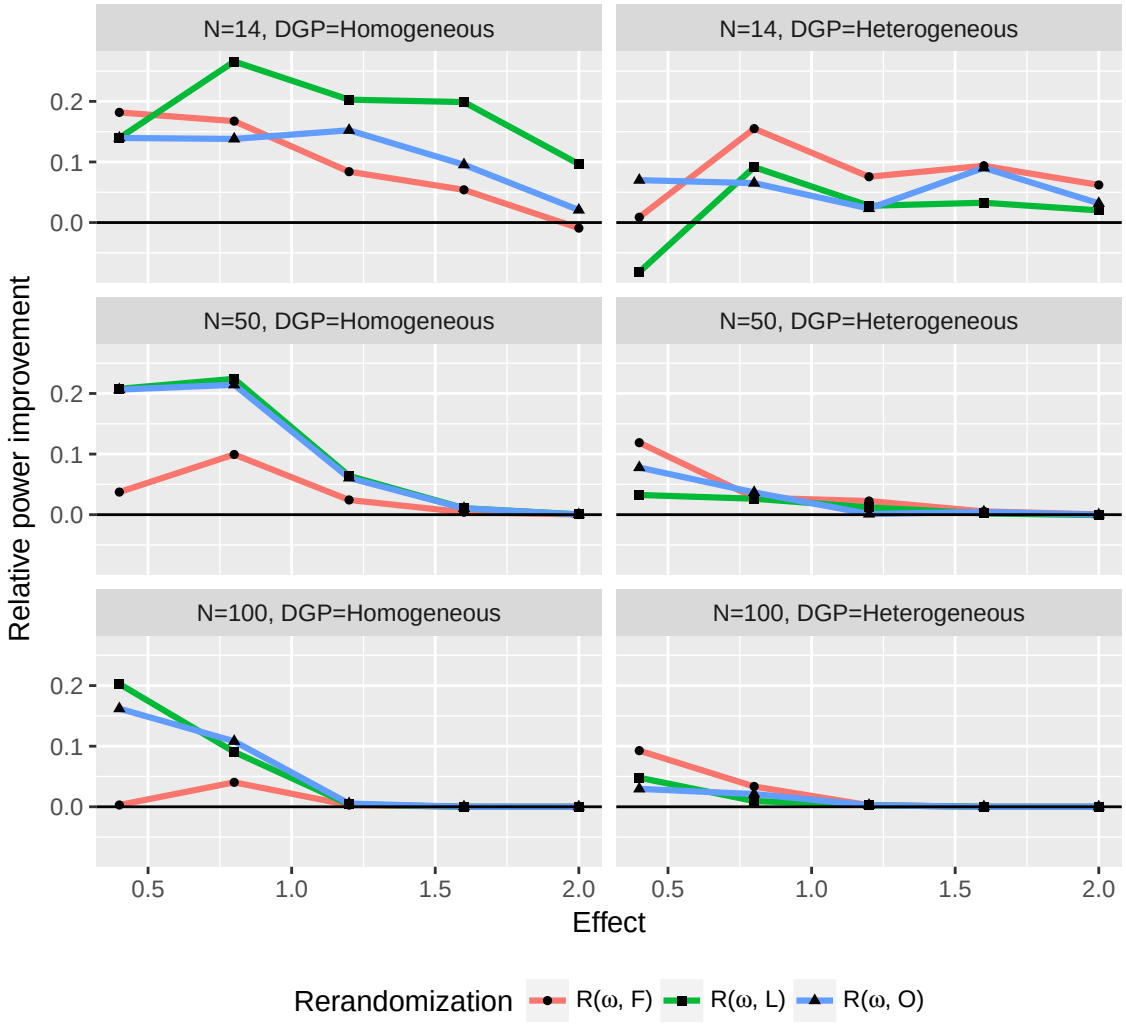

Figure 4: Relative power as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs for T=10. The left panel and right panel display the results from the Homogenous and the Heterogeneous DGPs, respectively.

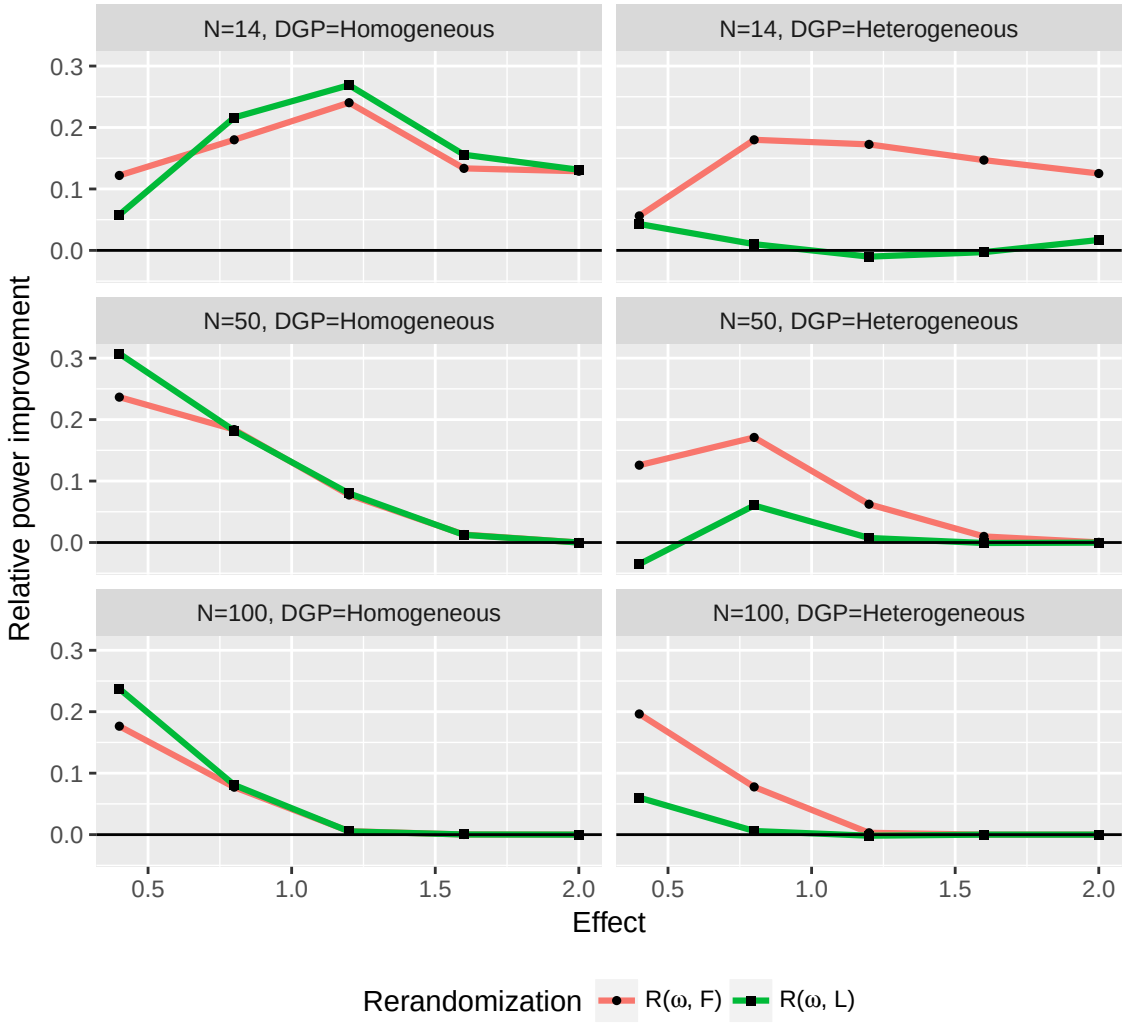

Figure 5: Relative power as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs for T=100. The left panel and right panel display the results from the Homogenous and the Heterogeneous DGPs, respectively.

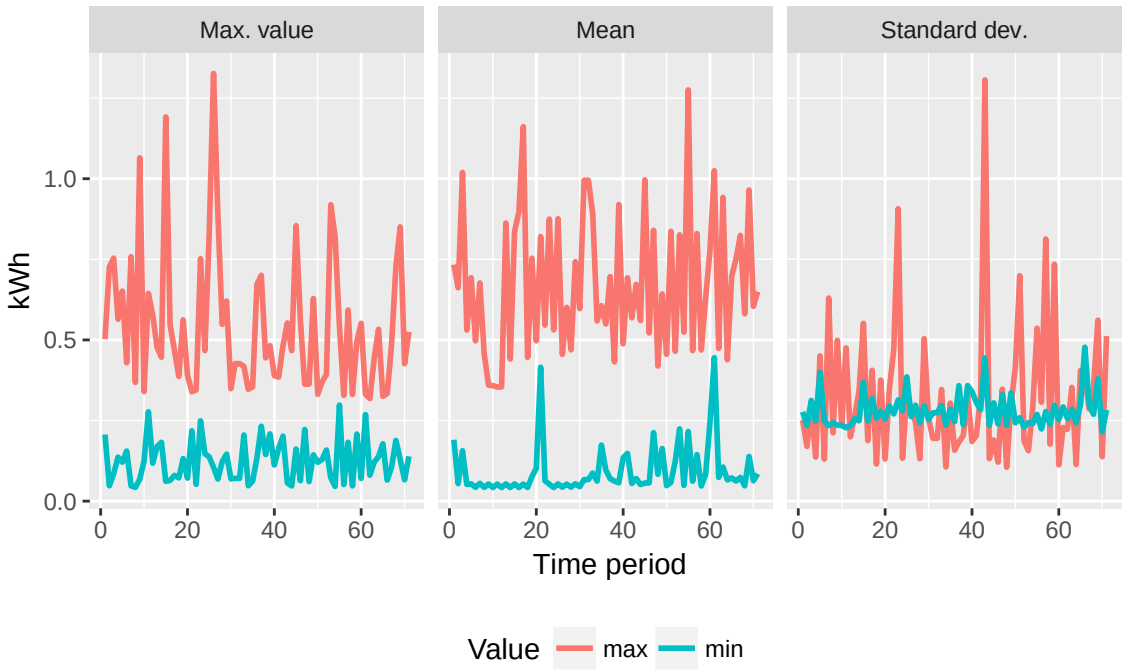

Figure 6: The electricity consumption (kWh) for the households with the largest (max) and smallest (min) maximum consumption, mean consumption, and standard deviation in their consumption, respectively.

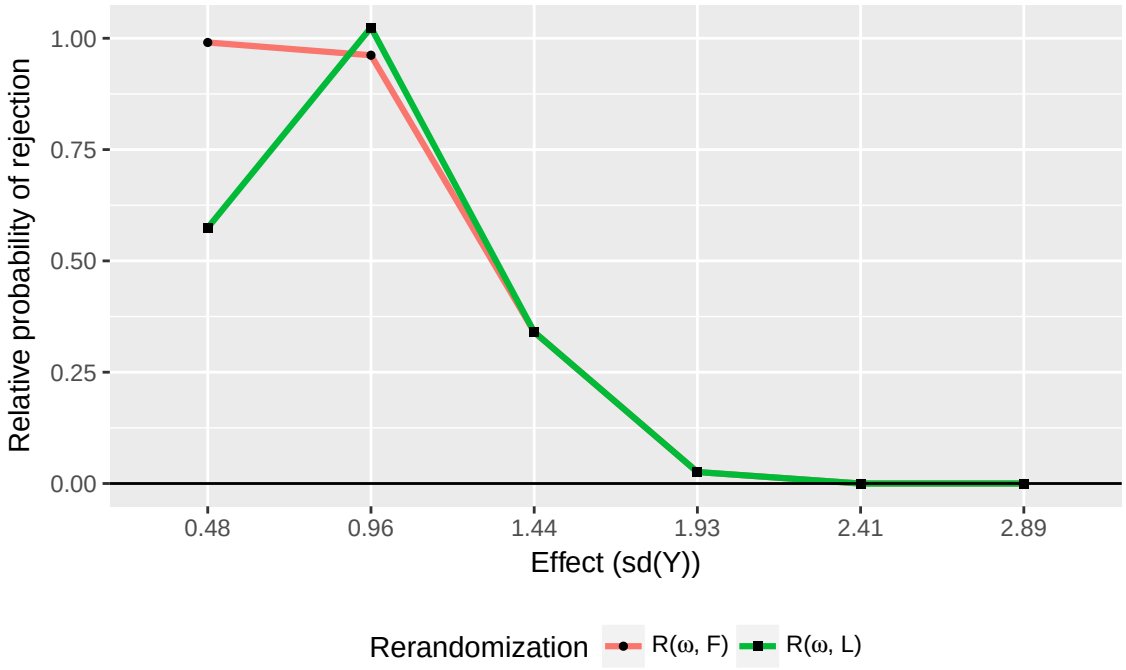

Figure 7: The relative probability as compared to complete randomization of randomly selecting an allocation that gives a significant result for two different rerandomization strategies given different hypothesized treatment effects.

List of Tables

Table 1: Relative variance reduction in the covariates and the effect estimate as compared to complete randomization 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	
Criterion	R	M	R	M
X_1	-0.71	-0.74	-0.83	-0.74
X_2	-0.71	-0.75	-0.83	-0.75
X_3	-0.70	-0.75	-0.73	-0.75
$\hat{\tau}$	-0.20	-0.19	-0.22	-0.20

Table 2: Relative variance reduction of the covariates and in the effect estimate as compared to complete randomization 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											
$Y_{t=T}$	-0.94	-0.96	-0.96	-0.92	-0.95	-0.95	-0.66	-0.71	-0.73	-0.64	-0.73	-0.74
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
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
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
$\hat{\tau}$	-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											
$Y_{t=T}$	-0.93	-0.96	-0.96	-0.68	-0.93	-0.94	-0.39	-0.50	-0.46	-0.21	-0.48	-0.47
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
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
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
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
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
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
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
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
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
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
$\hat{\tau}$	-0.17	-0.16	-0.17	-0.15	-0.21	-0.22	-0.12	-0.12	-0.13	-0.05	-0.16	-0.11

Table 3: Variance reduction of the covariates and in the effect estimate as compared to complete randomization for Mahalanobis distance (M) and 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								
Y1	-0.01	0.04	0.01	-0.07	-0.02	-0.01	-0.33	-0.17	-0.12
Y2	-0.01	-0.02	-0.00	-0.09	-0.04	-0.05	-0.37	-0.25	-0.18
Y3	-0.00	0.01	-0.04	-0.08	-0.03	-0.07	-0.39	-0.24	-0.25
Y4	-0.01	-0.02	-0.06	-0.05	-0.03	-0.07	-0.42	-0.26	-0.20
Y5	-0.08	-0.09	-0.08	-0.06	-0.11	-0.07	-0.42	-0.30	-0.16
Y6	-0.12	-0.08	-0.09	-0.10	-0.10	-0.11	-0.41	-0.29	-0.22
Y7	-0.17	-0.14	-0.16	-0.08	-0.08	-0.10	-0.39	-0.26	-0.25
Y8	-0.28	-0.23	-0.25	-0.20	-0.24	-0.21	-0.45	-0.34	-0.32
Y9	-0.37	-0.31	-0.32	-0.42	-0.62	-0.69	-0.51	-0.68	-0.73
Y10	-0.39	-0.42	-0.39	-0.92	-0.95	-0.95	-0.69	-0.92	-0.93
$\hat{\tau}$	-0.10	-0.10	-0.13	-0.19	-0.27	-0.21	-0.14	-0.25	-0.24
	DPG=Heterogeneous								
Y1	0.01	0.00	-0.03	-0.01	-0.03	-0.08	-0.29	-0.18	-0.17
Y2	-0.03	-0.03	0.02	-0.10	-0.07	0.03	-0.33	-0.23	-0.10
Y3	-0.05	-0.01	-0.11	-0.05	-0.01	-0.06	-0.30	-0.18	-0.17
Y4	-0.08	-0.06	-0.05	-0.10	-0.05	-0.06	-0.35	-0.22	-0.18
Y5	-0.10	-0.11	-0.08	-0.06	-0.04	-0.05	-0.35	-0.19	-0.17
Y6	-0.14	-0.10	-0.15	-0.06	-0.05	-0.11	-0.36	-0.22	-0.20
Y7	-0.19	-0.19	-0.21	-0.08	-0.15	-0.24	-0.36	-0.28	-0.28
Y8	-0.19	-0.21	-0.16	-0.14	-0.25	-0.29	-0.38	-0.39	-0.34
Y9	-0.29	-0.27	-0.24	-0.14	-0.09	-0.13	-0.35	-0.24	-0.20
Y10	-0.31	-0.27	-0.30	-0.93	-0.96	-0.97	-0.71	-0.94	-0.95
$\hat{\tau}$	-0.06	-0.08	-0.10	-0.04	-0.07	-0.05	-0.06	-0.09	-0.06

Table 4: Relative variance reduction in the estimated treatment effect as compared to complete randomization for Mahalanobis distance (M) and ranked P-values (R) rerandomization designs. The top and bottom panels display the results for the Homogeneous and Heterogeneous DGPs, respectively.

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

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

Rerandomazation	$R(\omega, F)$	$R(\omega, L)$
$\hat{\tau}$	-0.56	-0.61