Published in final edited form as:

J Am Stat Assoc. 2015 ; 110(512): 1412–1421. doi:10.1080/01621459.2015.1079528.

Rerandomization to Balance Tiers of Covariates

Kari Lock Morgan [Assistant Professor] and

Department of Statistics, Penn State University, University Park, PA 16802 (klm47@psu.edu)

Donald B. Rubin* [John L. Loeb Professor]

Department of Statistics, Harvard University, Cambridge, MA 02138

Abstract

When conducting a randomized experiment, if an allocation yields treatment groups that differ meaningfully with respect to relevant covariates, groups should be rerandomized. The process involves specifying an explicit criterion for whether an allocation is acceptable, based on a measure of covariate balance, and rerandomizing units until an acceptable allocation is obtained. Here we illustrate how rerandomization could have improved the design of an already conducted randomized experiment on vocabulary and mathematics training programs, then provide a rerandomization procedure for covariates that vary in importance, and finally offer other extensions for rerandomization, including methods addressing computational efficiency. When covariates vary in a priori importance, better balance should be required for more important covariates. Rerandomization based on Mahalanobis distance preserves the joint distribution of covariates, but balances all covariates equally. Here we propose rerandomizing based on Mahalanobis distance within tiers of covariate importance. Because balancing covariates in one tier will in general also partially balance covariates in other tiers, for each subsequent tier we explicitly balance only the components orthogonal to covariates in more important tiers.

Keywords

randomization; experimental design; covariate balance; treatment allocation; causal inference; Mahalanobis distance

1 Introduction

Randomized experiments balance all covariates between treatment groups in expectation, yet chance imbalances often exist. If a randomization yields groups that are unbalanced, which is noticed *before* conducting the experiment, groups should be rerandomized, which can continue until a pre-specified desired level of covariate balance is achieved. Morgan and Rubin (2012) provide theoretical results supporting rerandomization. The general procedure for rerandomization is as follows:

1. Select units for the comparison of treatments, and collect covariate data on all units.

* (rubin@stat.harvard.edu).

2. Define an explicit criterion for covariate balance.
3. Randomize units to treatment groups.
4. Check covariate balance and return to Step 3 if the allocation is unacceptable according to the criterion specified in Step 2; continue until the balance is acceptable.
5. Conduct the experiment.
6. Perform inference (using a randomization test that follows exactly steps 2-4).

This framework as stated requires that covariate data be available for all units to be randomized before randomization begins, so is most applicable when all units are randomized simultaneously, although extensions to sequential assignment are discussed in another manuscript.

Step 2, defining acceptable balance, requires definitions for measuring balance and thresholds for acceptable balance. Morgan and Rubin (2012) recommend choosing an *affinely invariant* measure of covariate balance; that is, a criterion such that any affine transformation of the covariate matrix $\mathbf{x}$, $c + \mathbf{b}\mathbf{x}$, yields the same rerandomization acceptance decisions. Affinely invariant measures of covariate balance are desirable because they (a) reflect the joint distribution of the covariates, (b) provide equal balance improvement for each covariate and for any linear combination of covariates, and (c) yield unbiased estimates for any linear combination of $\mathbf{x}$ if $\mathbf{x}$ is ellipsoidally symmetric (Morgan and Rubin, 2012). Mahalanobis distance, further explored in Section 3, is one common example of an affinely invariant criterion for balance. Although equal balance improvement for all covariates is theoretically appealing, in practice we may wish to impose more stringent criteria for covariates a priori thought to be more important, and be more lenient for less important covariates; Section 4 provides a new rerandomization criterion that does this. The threshold for acceptable balance for large samples involves a trade-off between better covariate balance and computational time; Section 5.1 provides a method that can often improve computational efficiency. Section 5.2 offers an alternative rerandomization procedure: generating a fixed number of allocations and choosing the best, according to a pre-defined measure of balance.

2 Example

Here we evaluate the benefits of rerandomization using data from an actual randomized experiment, conducted by Shadish et al. (2008)¹. The study was designed to examine whether observational studies can be analyzed to yield valid estimates of causal effects. Participants, undergraduate psychology students at a particular college, were first randomized to be in one of two arms: a randomized experiment ($n_r = 235$) or an observational study ($n_o = 210$). Those in the randomized experiment were randomized to take either a vocabulary ($n_v = 116$) or mathematics ($n_m = 119$) training course, and those in

¹The second author of the current manuscript was provided the data as a discussant of Shadish et al. (2008)

the observational study were allowed to choose one of the identical courses. The causal effects being estimated in each are the effect that the vocabulary training course has on subsequent vocabulary test scores, and analogously for mathematics. The goal was to see whether the estimates of the causal effects obtained from the observational study (using methods such as propensity score subclassification) were close to the estimates found from the randomized experiment.

We start by considering the ten covariates we believe to be most relevant: scores on the vocabulary and mathematics pre-tests, number of math courses taken, how much the students like math and literature, whether they prefer math or literature, ACT score, college grade point average (GPA), age, and sex. The balance of these covariates resulting from the actual randomization to experiment or observational study is displayed in Figure 1.

The means for the number of math classes taken, how much they like math, and GPA are all significantly lower for the students randomized to participate in the randomized experiment than for those in the observational study. Because the two groups differ substantially on these important background covariates, differences in estimated causal effects between the experimental group and the observational study group could be because one was an observational study, or because the groups differed at baseline.

Rerandomization provides a way to prevent these chance imbalances, yielding more informative randomized experiments, a point made by Rubin (2008a) in his discussion. We use rerandomization with this example in two ways. In Section 3.2 we balance the ten primary covariates in Figure 1 by rerandomizing based on Mahalanobis distance. In Section 4.3, we consider a richer collection of covariates, twenty-six main effects, two quadratic terms, and sixteen interactions, and use rerandomization involving tiers of covariates of varying importance.

3 Mahalanobis Distance

3.1 Rerandomization based on Mahalanobis Distance: Review of Theory from Morgan and Rubin (2012)

Let $\mathbf{x}$ be the fixed $n \times k$ covariate matrix, k covariates for n experimental units, let $\mathbf{W}$ be the n -component treatment assignment column vector, where $W_i = 1$ indicates unit i is assigned the active treatment and $W_i = 0$ indicates unit i is assigned the control condition. Let *pure randomization* denote complete randomization into groups with no restrictions except for total sample sizes, $n_T = \sum_{i=1}^n W_i$ and $n_C = \sum_{i=1}^n (1 - W_i)$. That is, the classic completely randomized experiment where n units are randomly partitioned into two groups of sizes n_T and n_C . Let $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$ be the k -dimensional vector of the difference in covariate means between the treatment and control groups:

$$\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C = \frac{\mathbf{x}'\mathbf{W}}{n_T} - \frac{\mathbf{x}'(\mathbf{1} - \mathbf{W})}{n_C}. \quad (1)$$

A common scalar measure of multivariate balance is Mahalanobis distance, which we present here standardized by a constant involving sample sizes:

$$M \equiv \frac{n_T n_C}{n} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)' \text{cov}(\mathbf{x})^{-1} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C). \quad (2)$$

Mahalanobis distance effectively calculates simple Euclidean distance after transforming all variables to an orthonormal canonical form; that is, where each transformed covariate has mean zero and variance one, and they are all uncorrelated. For rerandomization with a threshold a , we rerandomize if $M > a$. When $n_T = n_C$, $\mathbb{E}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C | \mathbf{x}, M \leq a) = \mathbf{0}$. Also from Morgan and Rubin (2012), when (a) $n_T = n_C$, (b) covariate means are normally distributed, and (c) units are purely randomized and rerandomized when $M > a$,

$$\text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C | \mathbf{x}, M \leq a) = v_a \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C | \mathbf{x}), \quad (3)$$

where

$$v_a \equiv \frac{2}{k} \times \frac{\gamma\left(\frac{k}{2} + 1, \frac{a}{2}\right)}{\gamma\left(\frac{k}{2}, \frac{a}{2}\right)} = \frac{P(\chi_{k+2}^2 \leq a)}{P(\chi_k^2 \leq a)}, \quad (4)$$

and γ denotes the incomplete gamma function: $\gamma(b, c) \equiv \int_0^c y^{b-1} e^{-y} dy$. Thus, under these conditions, this rerandomization maintains the correlation matrix of $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$, and for each covariate leads to equal improvement in balance in the sense of variance reduction.

Percent Reduction in Variance (PRIV) is defined as the percentage by which rerandomization reduces the variance of the difference in means, for each covariate, x_j , relative to pure randomization:

$$PRIV = 100 \left(\frac{\text{var}(\bar{\mathbf{X}}_{j,T} - \bar{\mathbf{X}}_{j,C} | \mathbf{x}) - \text{var}(\bar{\mathbf{X}}_{j,T} - \bar{\mathbf{X}}_{j,C} | \mathbf{x}, \mathbf{W} \text{ acceptable})}{\text{var}(\bar{\mathbf{X}}_{j,T} - \bar{\mathbf{X}}_{j,C} | \mathbf{x})} \right). \quad (5)$$

The higher the PRIV, the more the actual differences in covariate means will be concentrated around 0. Under the above conditions, the PRIV for each covariate, and for any linear combination of these covariates, is

$$1 - v_a. \quad (6)$$

Fewer covariates and smaller thresholds, a , lead to larger PRIV. In particular, greater computational power (lower proportion of accepted randomizations) is needed to achieve the same degree of balance for a larger number of covariates.

Let $y_i(W_i)$ denote the i th unit's potential outcome under treatment assignment W_i , and assume the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980): the potential outcomes are fixed and do not change with different $\mathbf{W}$. Define τ to be the true average treatment effect in the sample,

$$\tau \equiv \frac{\sum_{i=1}^n y_i(1) - y_i(0)}{n} \quad (7)$$

and define $\hat{\tau}$ to be the simple estimate of the average treatment effect:

$$\hat{\tau} \equiv \frac{\sum_{i=1}^n W_i y_i(1)}{n_T} - \frac{\sum_{i=1}^n (1 - W_i) y_i(0)}{n_C}. \quad (8)$$

Suppose (i) $n_T = n_C$, (ii) the covariate means are normally distributed, (iii) the outcome means are normally distributed given x , (iv) the treatment effect is additive (i.e. $y_i(1) - y_i(0)$ is the same for all i), and (v) units are rerandomized when $M > a$, then the PRIV for $\hat{\tau}$ is

$$(1 - v_a) R^2, \quad (9)$$

where R^2 is the squared multiple correlation between $\mathbf{x}$ and $\mathbf{y}(0)$. Expression (9) also gives the PRIV for any unobserved covariates. Figure 3 in Morgan and Rubin (2012) shows this PRIV as a function of k , the proportion of randomizations accepted, and R^2 . By (9), when $v_a \rightarrow 0$, rerandomization has the potential to provide the same increase in precision that would be achieved by increasing the sample size by the factor $(1 - R^2)^{-1}$, which implies that rerandomization can be equivalent to doubling the sample size if $R^2 = 0.5$, and for $R^2 = 0.9$, rerandomization can give the same power as pure randomization, but with only a tenth of the sample size! If cost or availability of experimental units are issues, this could be hugely beneficial.

Because rerandomization increases precision of estimators when covariates are correlated with the outcome(s), the traditional two sample t-test and corresponding interval will be too “conservative”, resulting in overly wide confidence intervals and less powerful tests (Morgan and Rubin, 2012). Inference using randomization tests, following the rerandomization procedure actually used, reflects this increase in precision, and maintains the Type I error rate and valid frequentist properties (Morgan and Rubin, 2012). For additive treatment effects, interval estimates can be generated as the sets of values that would not be rejected by a randomization test.

3.2 Experimental Data: Mahalanobis using Ten Covariates

We first focus on the initial randomization of the Shadish et al. (2008) data, randomizing 445 participants into the randomized experiment ($n_C = 235$) and the observational study ($n_T = 210$), matching the sample sizes obtained in the actual study. Because the sample sizes are not exactly equal in the two treatment groups, the theoretical results given in Section 3.1 are approximations, but we will see that nonetheless the empirical results closely match these theoretical values. In this section we balance only the ten primary covariates, displayed in Figure 1, and balance these covariates equally by using Mahalanobis distance to measure balance.

Let p_a denote the proportion of all randomizations considered acceptable by the rerandomization criterion, so here, $p_a = P(M \leq a)$. Often in practice p_a will be chosen first, and a chosen correspondingly. If a is to be specified first, care should be taken to ensure that the corresponding p_a is not too small, and in particular that some randomizations will yield $M \leq a$. The number of rerandomizations required to obtain an acceptable allocation follows a Geometric distribution with parameter p_a , so equals $1/p_a$ in expectation. We assume the analysis will use a randomization test, so also relevant is the computational time needed to generate enough (here assumed to be 1000) acceptable randomizations to perform a randomization test. For this data set, generating and checking balance for 1000 pure randomizations takes about 1.1 seconds, so it should take about $1.1/p_a$ seconds to generate 1000 acceptable randomizations using programs in *R* (R Development Core Team, 2011), and computations performed on a Lenovo X200s.

Based on the estimate of computational time, $p_a = 0.001$ or even $p_a = 0.0001$ seem reasonable. With $k = 10$ covariates, using (6) we calculate the PRIV to be about 93% for $p_a = 0.0001$ and 88% for $p_a = 0.001$. In this illustrative example, we decide that reducing the variance by an extra 5% is not worth the additional three hours of computational time, and so use $p_a = 0.001$.

With a sample size of $n = 445$ and these covariates, we can safely assume normality of covariate means, so $M \sim \chi^2_{10}$ under pure randomization. If this were not to be the case, we could simulate an empirical distribution of M , and find a from the empirical distribution.

Thus we set the threshold a such that $P(\chi^2_{10} \leq a) = p_a = 0.001$, yielding $a = 1.48$, so we rerandomize any allocation with $M > 1.48$. The value of M in the actual experiment was $M = 21.0$, and thus the particular randomization used is at the 98th percentile for M under pure randomization, and thus was an unusually unlucky allocation.

To explore the effects of rerandomization, we simulate 10,000 pure randomizations and 10,000 acceptable randomizations with $M \leq 1.48$. Density curves for the difference in means for the first covariate (score on the vocabulary pre-test) are given in Figure 2 under both pure randomization and rerandomization. As expected, rerandomization decreases the variance, thereby eliminating extreme differences in means and creating more allocations with differences in means that are close to zero.

Because rerandomizing using M theoretically yields the same variance reduction for each covariate, Figure 2 should look essentially identical for every covariate. Empirical distributions for the standardized difference in covariate means, the number of standard errors (based on pure randomization) between the treatment group mean and the control group mean, for each covariate are shown in Figure 3. We also include a linear combination of covariates at the bottom of Figure 3, to illustrate that the effects of rerandomization on any linear combination of covariates are also approximately the same as on any individual covariate. This linear combination was not included when calculating M . For each covariate, the empirical PRIV is displayed on the right of Figure 3; note these are all very close to the theoretical value of 88%, despite the fact that the 88% is calculated assuming exact normality of covariate means and equal sample sizes.

We also examine the effect of rerandomization on estimates of the treatment effect. Rather than considering methods for analyzing observational studies, we shift attention to the second stage of randomization in Shadish et al. (2008), where in the experimental arm students are randomized into the vocabulary or mathematics training course. With this randomization, we balance the same ten covariates, but restrict our sample to the 235 participants who were initially randomized to the experimental arm, of which 116 were randomized to vocabulary training and 119 to mathematics. In this study, we are estimating two causal effects: the effects of vocabulary and math training courses on the corresponding test scores. For the purpose of this illustrative example, we fill in the missing potential outcomes assuming the null hypothesis is true, that is, assuming no course effect for any subject. Therefore, in this simulation, we should see causal effect estimates centered around 0.

We again use rerandomization with an acceptance probability of $p_a = 0.001$, so $a = 1.48$ and the PRIV for each covariate is 88%. By (9), the extent to which rerandomization affects the estimates of the treatment effect depends on the R^2 between the outcome and the covariates, approximated by 0.11 for the vocabulary outcome and 0.32 for the mathematics outcome. These are rather low, so we do not expect rerandomization to have a large impact on the sampling variance of the estimated average treatment effects. We compute the theoretical PRIVs of the estimated treatment effect using (9): Vocabulary: $88\% \times R^2 = 88\% \times 0.11 = 9.8\%$, and Mathematics: $88\% \times R^2 = 88\% \times 0.32 = 28\%$. We can also calibrate using increase in precision, or equivalently, in effective sample size. Theoretically, the precision of the estimated vocabulary treatment effect is increased by a factor of $1/(1-0.097) = 1.11$ with rerandomization, and the estimated mathematics treatment effect by a factor of $1/(1-0.28) = 1.39$. For mathematics, this means the effective sample size with rerandomization is $1.39 \times 235 = 327$, rather than the actual 235.

We again simulate 10,000 pure randomizations and rerandomizations, and for each calculate the estimated treatment effects. The empirical PRIVs are 4% for vocabulary and 27% for mathematics, close to the theoretical values, but not exactly the same, due to the approximations involved.

4 Covariates of Varying Importance

Measuring balance with M balances all covariates equally, but when some covariates are known to be substantively more important than others, that is, thought to be more strongly correlated with the outcomes, better balance improvement for them is preferred. This desire is particularly strong when the number of covariates is large, requiring more computational power to achieve a certain degree of equal balance for each. Also, when the sample size is small, ensuring close to perfect balance for all covariates may be impossible, may yield a deterministic randomization, or may restrict the number of acceptable allocations to a number too small for a randomization test. With many covariates, or when balancing interactions and functions of main effects in addition to raw covariates, computational or sample size limits often prevent rerandomizing from achieving identical joint distributions in the treatment and control groups. Rather than either compromising on balance for important main effects, or ignoring balance for less important covariates or interactions, we provide a more flexible framework that allows for differing levels of balance across covariates

4.1 Individual Thresholds

When covariates vary in importance, a natural choice is to impose balance restrictions on each covariate individually, as suggested by Moulton (2004), Maclure et al. (2006), and Cox (2009). We illustrate the sub-optimal nature of this with two simple examples, the first with two uncorrelated covariates, and the second with two correlated covariates, both shown in Figure 4. Restricting each covariate individually forms a rectangular acceptance region, defined by the two calipers, whereas restricting Mahalanobis distance forms an ellipsoidally shaped acceptance region, or spherically shaped with orthogonal covariates. In both situations, the acceptance probability is set to $p_a = 1/2$. Because Mahalanobis distance takes the joint distribution into account, the combinations of covariate difference in means accepted by Mahalanobis distance have greater average probability density than those accepted by restricting each covariate individually, as is clear from the figures. Restricting each covariate individually also fails to preserve the joint distribution of the covariate difference in means across rerandomizations, as pointed out in Rubin (1976).

This problem becomes more substantial as k increases or p_a decreases. Even with independent covariates, if p_a is set to be the same for both methods, the proportion of accepted randomizations falling in the corners after restricting covariates individually (randomizations that would not be accepted by restricting M) goes to 1 as k increases and p_a decreases. With 20 independent covariates restricted equally and $p_a = 0.01$, more than half of the randomizations accepted by the intersection caliper method would be rejected using M . If these covariates were correlated, or there were more covariates, or a more stringent acceptance level were chosen, this problem would become even more severe.

4.2 Tiers of Covariates

We partition the covariates into $\mathcal{T}$ tiers of importance, with k_t covariates in tier t . Tier 1 includes the most important covariates, and subsequent tiers include covariates of decreasing importance. The ordering of covariates within a tier is irrelevant. Let $\mathbf{x}_t = (x_{t1}, \dots, x_{tk_t})$ represent the covariates in tier t . If we were to balance only the covariates in tier 1,

covariates in tier 2 would also be partially balanced to the extent that they are correlated with the covariates in tier 1, by (9). Rather than unnecessarily balancing all of the tier 2 covariates again, we balance only the portion of tier 2 covariates that are orthogonal to the tier 1 covariates.

Let $\mathbf{e}_1 \equiv \mathbf{x}_1$, and for $t > 1$ define e_{tj} to be the residuals for each covariate x_{tj} after regressing on all covariates in lower numbered tiers:

$$x_{tj} = \beta_0 + \sum_{i=1}^{t-1} \beta_i' \mathbf{x}_i + e_{tj}, t=2, \dots, \mathcal{T}. \quad (10)$$

For $1 < t \leq \mathcal{T}$, define $\mathbf{e}_t$ to be the matrix of residuals for covariates in tier t , so $\mathbf{e}_t = (e_{t1}, \dots, e_{tK_t})$. Let M_t be Mahalanobis distance between the treatment and control groups for $\mathbf{e}_t$:

$$M_t = \frac{n_T n_C}{n} (\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C})' \text{cov}(\mathbf{e}_t)^{-1} (\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C}). \quad (11)$$

There are different ways to conduct rerandomization using $M_1, \dots, M_{\mathcal{T}}$, for example placing separate thresholds on each M_t or placing a threshold on a weighted linear combination (e.g., see Lock (2011)). Here we explore only the former option, because it is more computationally efficient. The computationally intensive aspect of rerandomization is assessing balance for each randomization, and with a separate threshold for each tier, if tier 1 does not have adequate balance, there is no need to calculate $M_2, \dots, M_{\mathcal{T}}$. Because early tiers will typically have the more stringent acceptance probabilities, this may avoid many calculations.

Define p_{a_t} to be the acceptance probability for tier t , set the corresponding threshold a_t such that $P(M_t \leq a_t) = p_{a_t}$ and rerandomize if any $M_t > a_t$, which again leads to covariate means with the same expectation in treatment and control groups. Also, we have the following result:

Theorem 4.1—If the covariate means are normally distributed, and if units are purely randomized with $n_T = n_C$ and rerandomized when $M_t > a_t$ for $t=1 \dots \mathcal{T}$, then

$$\mathbb{E}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C | \mathbf{x}, M_t \leq a_t \quad \forall t) = \mathbf{0}, \quad (12)$$

$$\text{cov}(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C} | \mathbf{x}, M_t \leq a_t \quad \forall t) = v_{a_1} \text{cov}(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C} | \mathbf{x}), \quad (13)$$

and the PRIV for the j^{th} covariate in tier t , x_{tj} , is

$$1 - v_{a_1} \quad \text{if } t=1, \quad (14)$$

$$1 - \left[v_{a_1} R_{1,t_j}^2 + v_{a_2} (1 - R_{1,t_j}^2) \right] \quad \text{if } t=2, \quad (15)$$

$$1 - \left[v_{a_1} R_{1,t_j}^2 + \sum_{l=2}^{t-1} v_{a_l} (R_{l,t_j}^2 - R_{l-1,t_j}^2) + v_{a_t} (1 - R_{t-1,t_j}^2) \right] \quad \text{if } t > 2, \quad (16)$$

where R_{l,t_j}^2 denotes the squared multiple correlation between covariate x_{lj} and all covariates in tiers 1 to l , and v_{a_l} is defined by equation (4) with $k = k_l$ and $a = a_l$.

The proof of Theorem 4.1 is in Appendix A.

Theorem 4.1 states that the results for tier 1 covariates are equivalent to rerandomizing based on Mahalanobis distance for only these covariates: the correlation matrix for differences in means is retained and balance improvement is the same for all covariates in tier 1. Note that for tiers $t > 1$, PRIV will differ for covariates within the same tier, depending on how correlated each covariate is with covariates in more important tiers. Note that $1 - v_{a_t}$ is the PRIV for orthogonalized covariates, $\mathbf{e}_t$, in each tier, and the PRIVs given in Theorem 4.1 apply to the original, unorthogonalized covariates and hence can differ within a tier.

To parse the PRIVs given in Theorem 4.1, we first consider covariates in tier $t = 2$. If $R_{1,2j}^2 = 1$, PRIV would be $1 - v_{a_1}$, because the residuals would all be 0. If $R_{1,2j}^2 = 0$, PRIV would be $1 - v_{a_2}$, because in this case balancing covariates in tier 1 will have no impact on the balance of x_{2j} . In general, PRIV is 1 minus a weighted mean of v_{a_1} and v_{a_2} with the weights determined by the correlation between x_{2j} and the covariates in tier 1.

Next, for illustrative purposes, consider PRIV for covariate x_{4j} in tier 4. By Theorem 4.1, this is

$$1 - \left[v_{a_1} R_{1,4j}^2 + v_{a_2} (R_{2,4j}^2 - R_{1,4j}^2) + v_{a_3} (R_{3,4j}^2 - R_{2,4j}^2) + v_{a_4} (1 - R_{3,4j}^2) \right]. \quad (17)$$

Each component in (17) arises from restricting the balance on a particular tier. For $t < 4$, the v_{a_t} for tier t is weighted by the additive contribution of $\mathbf{e}_t$ to $R_{t,4j}^2$. For example, if R_3^2 is only slightly larger than R_2^2 , then the covariates in tier 3 contribute little to R_3^2 , so balancing $\mathbf{e}_3$ will have little impact on the balance of x_{4j} and the weight on v_{a_3} is correspondingly small. In contrast, if x_{4j} is strongly correlated with $\mathbf{e}_3$ but not with $\mathbf{e}_1$ or $\mathbf{e}_2$, then the weight on v_{a_3} will be larger.

Define $\mathbf{p}_a = (p_{a_1}, p_{a_2}, \dots, p_{a_{\mathcal{T}}})$, $\mathbf{a} = (a_1, \dots, a_{\mathcal{T}})$, and $\mathbf{v}_a = (v_{a_1}, v_{a_2}, \dots, v_{a_{\mathcal{T}}})$. Because covariates in earlier tiers are more important, we expect $v_{a_1} < v_{a_2} < \dots < v_{a_{\mathcal{T}}}$. A simple way to ensure this is to choose the tiers such that $k_1 < k_2 < \dots < k_{\mathcal{T}}$, and then set all the acceptance probabilities to be the same. However if this is not possible or if more control is desired, each p_{a_t} can be altered to create the desired level of balance for that tier. The overall acceptance probability, p_a , satisfies $p_a \geq \prod_{t=1}^{\mathcal{T}} p_{a_t}$, where the inequality approaches equality as the number of covariates in each tier increases, because $\chi_{k_t}^2$ approaches a normal distribution as $k_t \rightarrow \infty$, and so orthogonality implies independence.

Theorem 4.2—If the covariate means are normally distributed, and if units are purely randomized with $n_T = n_C$ and rerandomized when $M_t > a_t$ for $t=1 \dots \mathcal{T}$, and if the treatment effect is additive, then

$$\mathbb{E}(\hat{\tau} | \mathbf{x}, M_t \leq a_t \quad \forall t) = \tau, \quad (18)$$

and the PRIV for the estimated treatment effect, $\hat{\tau}$, is

$$(1 - v_{a_1}) R_{1,y}^2 \quad \text{if } t=1, \quad (19)$$

$$(1 - v_{a_1}) R_{1,y}^2 + \left[\sum_{t=2}^{\mathcal{T}} (1 - v_{a_t}) (R_{t,y}^2 - R_{1:(t-1),y}^2) \right] \quad \text{if } t > 1, \quad (20)$$

where $R_{t,y}^2$ is the squared multiple correlation between $\mathbf{x}_t$ and $y(0)$, and $R_{1:t,y}^2$ is the squared multiple correlation between $\mathbf{x}_{1:t}$ and $y(0)$.

The proof of Theorem 4.2 is in Appendix A.

4.3 Shadish Experimental Data: Tiers of Covariates

Here we return to the Shadish experimental data, but now utilize tiers of covariates. This section pertains only to the second stage of the Shadish et al. (2008) randomization, randomizing units to either the vocabulary or mathematics training class within the experimental arm of the study. Here we balance twenty-six main effects, two quadratic terms, and sixteen interactions. If we were to compute Mahalanobis distance on all forty-nine of these covariates, maintaining the acceptance probability used in Section 3.2, $p_a = 0.001$, rerandomization would yield an expected 54% reduction in variance for each covariate. However, for this dataset, some covariates, such as vocabulary and mathematics pre-test scores, are clearly more important than others, and for these covariates, better balance is desired.

We place vocabulary and mathematics pre-test scores in the top tier. In the next tier, we place the number of mathematics courses taken, whether the person likes math, whether the person likes literature, ACT score, and college GPA, because these Kve covariates are likely to be related to the outcomes. The third tier contains nineteen covariates deemed less important, but still worth balancing. The fourth tier contains quadratic terms for both pre-test scores to balance variances of these distributions as well as their means, and interactions between all covariates in the top two tiers, in case any of these are important. This gives $\mathcal{T}=4$ and $\mathbf{k} = (2, 5, 19, 18)$.

To maintain consistency with Section 3.2, we keep the overall acceptance probability at 0.001. Because the tiers have increasing numbers of covariates, one option is to set each $p_{a_t} = (0.001)^{1/4} = 0.178$. Calculating each v_{a_t} by (4), this gives $\mathbf{v}_a = (0.09, 0.29, 0.58, 0.61)$. If this is satisfactory, we would proceed with equal acceptance probabilities. However, here we decide to improve balance for the top two tiers, at the expense of the bottom two tiers. For various vectors $\mathbf{p}_a$, we calculate $\mathbf{v}_a$ for each, and decide on $\mathbf{p}_a = (0.1, 0.1, 0.2, 0.5)$, which means expected PRIVs for the residuals (orthogonalized covariates) will be 95% for tier 1, 78% for tier 2, 41% for tier 3, and 23% for tier 4. Recall that the residuals within a tier will receive the same variance reduction, although the balance improvement for the original covariates may differ, even within a tier. PRIVs for the original covariates can be calculated using Theorem 4.1, and will depend on correlations with covariates in lower numbered tiers.

Simulating 1000 acceptable randomizations took 20.1 minutes. Balancing only the ten covariates in Section 3.2 took 18.3 minutes for the same overall acceptance probability; even though we are now balancing more covariates and potentially checking balance on multiple tiers for each randomization, the computational time is approximately the same because so many randomizations (about 95%) only require calculating Mahalanobis distance for tier 1, which has only two covariates.

Empirical PRIVs for the residuals and covariates in each tier are given in Figure 5. The theoretical values for the residuals of each tier are depicted with vertical dotted lines. The empirical values for the residuals are scattered randomly around these lines, lying close to the theoretical values. The covariates, however, have varying PRIVs, even within each tier. The covariates in tier 3 are only weakly correlated with the covariates in tiers 1 and 2, and thus have PRIVs close to theoretical value for the residuals in tier 3, $1 - v_{a3}$, and because tier 4 is entirely composed of functions (quadratics and interactions) of covariates in tiers 1 and 2, the PRIVs are closer to the theoretical values for the residuals in earlier tiers.

We also explore the effects of this rerandomization scheme on the outcomes, vocabulary and mathematics test scores. The empirical PRIVs are 15% for the estimated treatment effect for vocabulary and 25% for the estimated treatment effect for mathematics. This should be contrasted with the 4% for vocabulary and 27% for mathematics achieved in Section 3.2 by balancing all ten covariates equally. Here we improve upon the estimation of the vocabulary treatment effect quite substantially, but gain essentially nothing for the mathematics outcome. The theoretical PRIVs, as calculated by Theorem 4.2 are 15.9% and 29.7%, for vocabulary and mathematics, respectively. The empirical mean estimated treatment effect is -0.016 and -0.011 for vocabulary and mathematics, respectively. These values are very

close to 0, as expected, because this simulation was conducted under the null hypothesis of no treatment effect.

5 Extensions

5.1 Computational Efficiency

When using Mahalanobis distance, either on all covariates or within tiers, in many situations computational time can be dramatically reduced by first transforming the covariates to an orthonormal basis. Let $e_1 = (x_1 - \bar{x}_1)/s_1$, where $\bar{x}_1$ and s_1 are the sample mean and standard deviation, respectively, of x_1 , and for $j > 2$ define e_j by regressing x_j on $(x_1, \dots, x_{j-1})$:

$$x_j = \beta_0 + \sum_{i=1}^{j-1} \beta_i x_i + \epsilon_j, \quad (21)$$

and standardizing so $e_j = (\epsilon_j - \bar{\epsilon}_j)/s_{\epsilon_j}$. Define $\mathbf{e} = (e_1, \dots, e_k)$, and balance $\mathbf{e}$, rather than $\mathbf{x}$. Because M is affinely invariant, M will be the same for $\mathbf{e}$ as for $\mathbf{x}$, and the order of the covariates is irrelevant. Because $\mathbf{e}$ is an orthonormal basis, $\text{cov}(\mathbf{e}) = \mathbf{I}$, so

$$M = \frac{n_T n_C}{n} \sum_{j=1}^k \left(\bar{E}_{j,T} - \bar{E}_{j,C} \right)^2, \quad (22)$$

where $\bar{E}_{j,T}$ and $\bar{E}_{j,C}$ are the treatment and control means, respectively, for e_j . This allows us to check balance sequentially, adding in one covariate at a time, and when the acceptance probability, p_a , is small, then most randomizations will be rejected after checking only the first few covariates. For the experimental data considered in Section 3.2 with $n = 445$ and $k = 10$ covariates, use of this method resulted in a 60% reduction in computational time.

The standardization of the variables to have mean 0 can also improve computational efficiency, if taken advantage of. We rewrite the difference in means as follows:

$$\begin{aligned} \bar{E}_{j,T} - \bar{E}_{j,C} &= \frac{\mathbf{e}' \mathbf{W}}{n_T} - \frac{\mathbf{e}' (\mathbf{1} - \mathbf{W})}{n_C}, \\ &= \frac{n}{n_T n_C} \mathbf{e}' \mathbf{W} - \frac{n_T}{n_T n_C} \mathbf{e}' \mathbf{1}, \end{aligned} \quad (23)$$

$$= \frac{n}{n_T n_C} \mathbf{e}' \mathbf{W}, \quad (24)$$

where (24) follows from (23) because each e_j has mean 0. For the experimental data from Section 3.2, use of (24) reduced computational time by 75% (in R, as opposed to using $(\text{mean}(\mathbf{e}[\mathbf{W}==1, j]) - \text{mean}(\mathbf{e}[\mathbf{W}==0, j]))$ when computing (22).

Although these methods improve computational efficiency for a variety of datasets, they will not always do so. Using (22) and checking covariates sequentially will be most advantageous with fewer covariates and lower p_a , when it is more likely that only a small number of covariates will need to be checked before the threshold is exceeded. Using (24) is more important for larger sample sizes. In practice we recommend measuring time to check balance for a specified number, say 1000, randomizations and using the method and code that is most efficient for the situation being faced.

5.2 Choosing the Best

An attractive alternative to specifying a threshold for acceptable balance may be to generate a fixed number of allocations, and choose the “best” according to a pre-specified criterion, such as Mahalanobis distance. In practice we expect to see little difference between rerandomizing until an allocation with $M = a$ is obtained, and choosing the allocation with the smallest M from $1/p_a$ randomizations, where $p_a = P(M = a)$. However, with non-scalar measures of balance, such as Mahalanobis distance computed within tiers of covariates, the definition of “best” is not obvious. For choosing the best with tiers of covariates, rather than restricting each tier with a separate threshold, the balance measures for each tier can be combined into a scalar measure with a weighted linear combination: $\mathbf{w}'M_t$.

5.3 Covariance Adjustment

Although alternatives such as blocking or covariance adjustment may work in simple situations with few covariates, incorporating many covariates and functions of covariates is generally difficult. For example, a priori specifying covariance adjustment with many covariates and interactions into a protocol for a randomized experiment could be problematic. Rerandomization is intended to improve covariate balance via changing the *design* of the experiment, and so obviate the need to rely on model-based adjustment for imbalance in the analysis phase. There are many reasons for preventing problems by design rather than trying to fix problems by analysis, as stated in Rubin (2008b), with two of the primary reasons being the subjectivity that can be introduced when model-fitting using the outcome data and the sensitivity of estimates to the specifications of the model imposed. In addition to being a design-based method, rerandomization has the potential to improve precision more than typical covariance adjustment (Morgan and Rubin, 2012). If covariance adjustment is going to be used, Cox (1982) proposed using rerandomization first, to further improve the precision of the covariance adjusted estimator. Also, an experiment with better covariate balance, such as can be achieved through rerandomization, will yield estimators less sensitive to model misspecification. Moreover, the tiered approach for balancing covariates of varying importance introduced here allows for better balance of some covariates over others, whereas with standard covariance adjustment a variable must be either included or excluded from the model.

6 Conclusion

When covariates vary in importance, a rerandomization criterion should be set so that balance is better for more important covariates. However, balancing covariates unequally comes at the price that fully affinely invariant measures of covariate balance can no longer

be used. We compromise by using Mahalanobis distance within tiers of covariates, maintaining affine invariance where possible, while providing the flexibility to achieve better balance for more important covariates.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF SES-0550887 and NSF IIS-1017967) and the National Institutes of Health (NIH-R01DA023879).

Appendix A

Proof of Theorem 4.1

Equation (12) follows from Corollary 2.2 in Morgan and Rubin (2012). Equation (13) follows from (3), because also balancing on components orthogonal to $\mathbf{x}_1$ will not impact the balance of $\mathbf{x}_1$.

Define $\mathbf{x}_{1:(t-1)} = (\mathbf{x}_1, \dots, \mathbf{x}_{t-1})$, all covariates in tiers 1 to $t-1$. By (10),

$$\bar{X}_{tj,T} - \bar{X}_{tj,C} = \beta' (\bar{\mathbf{X}}_{1:(t-1),T} - \bar{\mathbf{X}}_{1:(t-1),C}) + (\bar{e}_{tj,T} - \bar{e}_{tj,C}). \quad (25)$$

Because the two terms on the right are uncorrelated, we have

$$\text{var}(\bar{X}_{tj,T} - \bar{X}_{tj,C}) = \beta' \text{cov}(\bar{\mathbf{X}}_{1:(t-1),T} - \bar{\mathbf{X}}_{1:(t-1),C}) \beta + \text{var}(\bar{e}_{tj,T} - \bar{e}_{tj,C}). \quad (26)$$

Let $\sigma_{x_{tj}}^2$ denote the variance of the covariate x_{tj} and $\sigma_{e_{tj}}^2$ the variance of the residuals e_{tj} so $\sigma_{e_{tj}}^2 = \sigma_{x_{tj}}^2 (1 - R_{t-1,tj}^2)$. Therefore

$$\text{var}(\bar{e}_{tj,T} - \bar{e}_{tj,C} | \mathbf{x}) = \frac{n}{n_T n_C} \sigma_{e_{tj}}^2 \quad (27)$$

$$= \frac{n}{n_T n_C} \sigma_{x_{tj}}^2 (1 - R_{t-1,tj}^2), \quad (28)$$

and

$$\beta' \text{cov}(\bar{\mathbf{X}}_{1:(t-1),T} - \bar{\mathbf{X}}_{1:(t-1),C} | \mathbf{x}) \beta = \text{var}(\bar{\mathbf{X}}_{tj,T} - \bar{\mathbf{X}}_{tj,C} | \mathbf{x}) - \text{var}(\bar{e}_{tj,T} - \bar{e}_{tj,C} | \mathbf{x}) \quad (29)$$

$$= \frac{n}{n_T n_C} (\sigma_{x_{tj}}^2 - \sigma_{e_{tj}}^2) \quad (30)$$

$$= \frac{n}{n_T n_C} \sigma_{x_{tj}}^2 R_{t-1,tj}^2. \quad (31)$$

By (3), because $\mathbf{e}'_t \perp \mathbf{e}''_t$ for $t' \neq t''$, we have

$$\text{cov}(\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C} | \mathbf{x}, \quad M_t \leq a_t \quad \forall t) = v_{a_t} \text{cov}(\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C} | \mathbf{x}). \quad (32)$$

Moreover, the differences in residual means will remain orthogonal across tiers.

For each covariate x_{2j} in tier 2, by (13), (32), (28), and (31), we have

$$\begin{aligned} \text{var}(\bar{X}_{2j,T} - \bar{X}_{2j,C} | \mathbf{x}, \quad M_1 \leq a_1, \quad M_2 \leq a_2) &= \text{var}[\beta'(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C}) + (\bar{e}_{2j,T} - \bar{e}_{2j,C}) | \mathbf{x}, \quad M_1 \leq a_1, \quad M_2 \leq a_2] \\ &= \beta' \text{cov}(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C} | M_1 \leq a_1, \quad M_2 \leq a_2) \beta \\ &+ \text{var}(\bar{e}_{2j,T} - \bar{e}_{2j,C} | \mathbf{x}, \quad M_1 \leq a_1, \quad M_2 \leq a_2) \\ &= v_{a_1} \beta' \text{cov}(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C}) \beta + v_{a_2} \text{var}(\bar{e}_{2j,T} - \bar{e}_{2j,C}) \\ &= \frac{n}{n_T n_C} (v_{a_1} \sigma_{x_{2j}}^2 R_{1,2j}^2 + v_{a_2} \sigma_{x_{2j}}^2 (1 - R_{1,2j}^2)) \\ &= \frac{n}{n_T n_C} \sigma_{x_{2j}}^2 (v_{a_1} R_{1,2j}^2 + v_{a_2} (1 - R_{1,2j}^2)) \\ &= (v_{a_1} R_{1,2j}^2 + v_{a_2} (1 - R_{1,2j}^2)) \text{var}(\bar{X}_{2j,T} - \bar{X}_{2j,C}), \end{aligned}$$

establishing (15).

For tiers $t > 2$, by expressing (10) in terms of $\mathbf{e}_{t-1}$ instead of $\mathbf{x}_{t-1}$, we have

$$\mathbf{x}_{tj} = \beta_0^* + \beta_x' \mathbf{x}_{1:(t-2)} + \beta_e' \mathbf{e}_{t-1} + \mathbf{e}_{tj}. \quad (33)$$

Expressing $\mathbf{x}_{tj}$ as either equation (10) or (33), the residuals $\mathbf{e}_{tj}$ remain the same, as does the coefficient of determination. Therefore we have

$$\bar{X}_{tj,T} - \bar{X}_{tj,C} = \beta_x' (\bar{\mathbf{X}}_{1:(t-2),T} - \bar{\mathbf{X}}_{1:(t-2),C}) + \beta_e' (\bar{\mathbf{e}}_{t-1,T} - \bar{\mathbf{e}}_{t-1,C}) + (\bar{e}_{tj,T} - \bar{e}_{tj,C}). \quad (34)$$

Thus

$$\begin{aligned}
\text{var} \left(\beta'_e \left(\bar{\mathbf{e}}_{t-1,T} - \bar{\mathbf{e}}_{t-1,C} \right) \right) &= \text{var} \left(\bar{X}_{tj,T} - \bar{X}_{tj,C} \right) - \text{var} \left(\beta'_x \left(\bar{\mathbf{X}}_{1:(t-2),T} - \bar{\mathbf{X}}_{1:(t-2),C} \right) \right) - \text{var} \left(\bar{e}_{tj,T} - \bar{e}_{tj,C} \right) \\
&= \frac{n}{n_T n_C} \left(\sigma_{x_{tj}}^2 - \sigma_{x_{tj}}^2 R_{t-2}^2 - \sigma_{x_{tj}}^2 (1 - R_{t-1}^2) \right) \\
&= \frac{n}{n_T n_C} \sigma_{x_{tj}}^2 (R_{t-1}^2 - R_{t-2}^2).
\end{aligned}
\tag{35}$$

Note that we can also rewrite (33) as

$$\mathbf{x}_{tj} = \beta_0^* + \beta'_x \mathbf{x}_1 + \sum_{l=2}^{t-1} \beta'_l \mathbf{e}_l + \mathbf{e}_{tj}.$$
(36)

The coefficients will change, but again the residuals, $\mathbf{e}_{tj}$, will remain the same. Therefore for covariate x_{tj} in tier $t > 2$,

$$\begin{aligned}
\text{var} \left(\bar{X}_{tj,T} - \bar{X}_{tj,C} | \mathbf{x}, M_t \leq a_t \forall t \right) &= \text{var} \left[\beta'_1 \left(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C} \right) + \sum_{l=2}^{t-1} \beta'_l \left(\bar{\mathbf{e}}_{l,T} - \bar{\mathbf{e}}_{l,C} \right) + \left(\bar{e}_{tj,T} - \bar{e}_{tj,C} \right) | \mathbf{x}, M_t \leq a_t \forall t \right] \\
&= \beta'_1 \text{cov} \left(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C} | M_t \leq a_t \forall t \right) \beta_1 + \sum_{l=2}^{t-1} \beta'_l \text{cov} \left(\bar{\mathbf{e}}_{l,T} - \bar{\mathbf{e}}_{l,C} | \mathbf{x}, M_t \leq a_t \forall t \right) \beta_l \\
&\quad + \text{var} \left(\bar{e}_{tj,T} - \bar{e}_{tj,C} | \mathbf{x}, M_t \leq a_t \forall t \right) \\
&= v_{a1} \beta'_1 \text{cov} \left(\bar{\mathbf{X}}_{1,T} - \bar{\mathbf{X}}_{1,C} \right) \beta_1 + \sum_{l=2}^{t-1} v_{al} \beta'_l \text{cov} \left(\bar{\mathbf{e}}_{l,T} - \bar{\mathbf{e}}_{l,C} \right) \beta_l + v_{at} \text{var} \left(\bar{e}_{tj,T} - \bar{e}_{tj,C} \right) \\
&= \frac{n}{n_T n_C} \sigma_{x_{tj}}^2 \left(v_{a1} R_1^2 + \sum_{l=2}^{t-1} v_{al} \left(R_l^2 - R_{l-1}^2 \right) + v_{at} \left(1 - R_{t-1}^2 \right) \right) \\
&= \left(v_{a1} R_1^2 + \sum_{l=2}^{t-1} v_{al} \left(R_l^2 - R_{l-1}^2 \right) + v_{at} \left(1 - R_{t-1}^2 \right) \right) \text{var} \left(\bar{X}_{tj,T} - \bar{X}_{tj,C} \right),
\end{aligned}
\tag{37}$$

where (37) follows from orthogonality. Thus, PRIV for the difference in means for covariate x_{tj} in tier $t > 2$ is as given in equation (20).

Proof of Theorem 4.2

Equation (18) follows directly from Theorem 2.1 of Morgan and Rubin (2012). Without loss of generality, under additivity we can write the outcomes as a linear function of the orthogonalized covariates with added noise:

$$y_i(W_i) = \beta_0 + \beta'_i \mathbf{e}_i + \tau W_i + \varepsilon_i, \tag{38}$$

where $\beta_0 + \beta' \mathbf{e}_i$ is the projection of y_i onto the space spanned by $(\mathbf{1}, \mathbf{e}_i)$ and ε_i a residual that captures any deviations from the linear model. Then the estimated treatment effect can be calculated by

$$= \beta_0 + \sum_{t=1}^{\mathcal{T}} \beta'_t \mathbf{e}_{t,i} + \tau W_i + \varepsilon_i, \quad (39)$$

Because each of these components are orthogonal, we can compute the variance of $\hat{\tau}$ as

$$\hat{\tau} = \sum_{t=1}^{\mathcal{T}} \beta'_t (\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C}) + \tau + (\bar{\varepsilon}_T - \bar{\varepsilon}_C). \quad (40)$$

Rerandomization reduces each $\text{cov}(\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C})$ by the factor v_{a_t} so

$$\text{var}(\hat{\tau}) = \sum_{t=1}^{\mathcal{T}} \beta'_t \text{cov}(\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C}) \beta_t + \text{var}(\bar{\varepsilon}_T - \bar{\varepsilon}_C). \quad (41)$$

Define $\sigma_y^2 = \text{var}(y(1)) = \text{var}(y(0))$ and $R_{e_t,y}^2$ to be the squared multiple correlation between $\mathbf{e}_t$ and $y(1)$ and $y(0)$. Then

$$\text{var}(\hat{\tau} | \mathbf{x}, M_t \leq a_t \forall t) = \sum_{t=1}^{\mathcal{T}} v_{a_t} \beta'_t \text{cov}(\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C} | \mathbf{x}) \beta_t + \text{var}(\bar{\varepsilon}_T - \bar{\varepsilon}_C | \mathbf{x}). \quad (42)$$

Moreover, because the tiers are orthogonal, the squared multiple correlation for all covariates and $Y(0)$ is $\sum_{t=1}^{\mathcal{T}} R_{e_t,y}^2$, so

$$\beta'_t \text{cov}(\bar{\mathbf{e}}_{t,T} - \bar{\mathbf{e}}_{t,C} | \mathbf{x}) \beta_t = R_{e_t,y}^2 \left(\frac{\sigma_y^2 n}{n_T n_C} \right). \quad (43)$$

Therefore,

$$\text{var}(\bar{\varepsilon}_T - \bar{\varepsilon}_C | \mathbf{x}) = \left(1 - \sum_{t=1}^{\mathcal{T}} R_{e_t,y}^2 \right) \left(\frac{\sigma_y^2 n}{n_T n_C} \right). \quad (44)$$

The PRIV is then

$$\text{var}(\hat{\tau}|\mathbf{x}, M_t \leq a_t \forall t) = \left(1 + \sum_{t=1}^{\mathcal{T}} (1 - v_{a_t}) R_{e_t,y}^2\right) \left(\frac{\sigma_y^2 n}{n_T n_C}\right). \quad (45)$$

$$PRIV = \frac{\frac{\sigma_y^2 n}{n_T n_C} - \left(1 + \sum_{t=1}^{\mathcal{T}} (1 - v_{a_t}) R_{e_t,y}^2\right) \left(\frac{\sigma_y^2 n}{n_T n_C}\right)}{\frac{\sigma_y^2 n}{n_T n_C}} \quad (46)$$

$$= \sum_{t=1}^{\mathcal{T}} (1 - v_{a_t}) R_{e_t,y}^2. \quad (47)$$

Note that for $t > 1$, $R_{e_t,y}^2 = R_{t,y}^2 - R_{1:(t-1),y}^2$, because the tiers are orthogonal, completing the proof.

References

- Cox DR. Randomization and concomitant variables in the design of experiments. *Statistics and Probability: Essays in Honor of CR Rao*. 1982:197–202.
- Cox DR. Randomization in the Design of Experiments. *International Statistical Review*. 2009; 77(3): 415–429.
- Lock, KF. PhD thesis. Harvard University; Cambridge, MA: 2011. Rerandomization to Improve Covariate Balance in Randomized Experiments..
- Maclure M, Nguyen A, Carney G, Dormuth C, Roelants H, Ho K, Schneeweiss S. Measuring prescribing improvements in pragmatic trials of educational tools for general practitioners. *Basic & clinical pharmacology & toxicology*. 2006; 98(3):243–252. [PubMed: 16611198]
- Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. *Ann. Statist*. 2012; 40(2):1263–1282. ISSN 0090-5364. doi: 10.1214/12-AOS1008.
- Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*. 2004; 1(3):297. [PubMed: 16279255]
- R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2011. URL <http://www.R-project.org>. ISBN 3-900051-07-0
- Rubin DB. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*. 1976:109–120.
- Rubin DB. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*. 1980; 75(371):591–593.
- Rubin DB. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*. 2008a; 103(484):1350–1353.
- Rubin DB. For objective causal inference, design trumps analysis. *Annals*. 2008b; 2(3):808–840.
- Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*. 2008; 103(484):1334–1344. ISSN 0162-1459.

Figure 1.

Love plots for the standardized difference in covariate means for ten covariates in the Shadish et al. (2008) study: randomized experiment versus observational study.

Figure 2. Empirical densities for difference in means between treatment groups for the Vocabulary Pre-Test, each based on 10, 000 simulated randomizations, pure and rerandomized.

Figure 3.

Empirical distributions for the standardized difference in means for each covariate, each based on 10,000 simulated randomizations, pure and rerandomized. Empirical *PRIVs* are shown on the right. The diamonds show the balance from the actual experiment.

Figure 4.

Points displayed are $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$ for two covariates, under orthogonality (left) and correlation (right), under pure randomization of units into two equally sized treatment groups.

Acceptance regions are shown for two different rerandomization methods, both with $p_a = 0.5$. The squares are the acceptance regions restricting each covariate individually and the ellipses are the acceptance regions based on M .

Figure 5.

Empirical PRIV for residuals and covariates based on 1000 simulated rerandomizations and 10,000 simulated pure randomizations. Covariates are ordered by tier. Theoretical PRIV values for the residuals of each tier, $1 - v_{a_r}$, are displayed at the top and shown with vertical dotted lines.