University of Pennsylvania
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

6-2014

Matching On-the-Fly: Sequential Allocation with Higher Power and Efficiency

Adam Kapelner

Abba Krieger
University of Pennsylvania

Matching On-the-Fly: Sequential Allocation with Higher Power and Efficiency

Abstract

We propose a dynamic allocation procedure that increases power and efficiency when measuring an average treatment effect in fixed sample randomized trials with sequential allocation. Subjects arrive iteratively and are either randomized *or* paired via a matching criterion to a previously randomized subject and administered the alternate treatment. We develop estimators for the average treatment effect that combine information from both the matched pairs and unmatched subjects as well as an exact test. Simulations illustrate the method's higher efficiency and power over several competing allocation procedures in both simulations and in data from a clinical trial.

Keywords

alternatives to randomization, clinical traits, matching, sequential allocation, squared Mahalanobis distance

Disciplines

Business | Business Analytics | Management Sciences and Quantitative Methods | Statistics and Probability

Matching on-the-fly in Sequential Experiments for Higher Power and Efficiency

Adam Kapelner and Abba Krieger

May 23, 2013

Abstract

We propose a dynamic allocation procedure that increases power and efficiency when measuring an average treatment effect in sequential randomized trials. Subjects arrive iteratively and are either randomized *or* paired via a matching criterion to a previously randomized subject and administered the alternate treatment. We develop estimators for the average treatment effect that combine information from both the matched pairs and unmatched subjects as well as an exact test. Simulations illustrate the method's higher efficiency and power over competing allocation procedures in both controlled scenarios and historical experimental data.

1 Introduction

The gold standard of experimentation, randomization, is only golden with large sample size. With small sample size, the empirical distributions of relevant covariates can be different across treatments possibly masking an effect by creating bias and inflating variance. Some improvements over completely randomized design are rerandomization (Morgan and Rubin, 2012), *a priori* matching (Raudenbush et al., 2007), and *adaptive design* which involves any change to the experiment or statistical procedures while the experiment is underway (Chow and Chang, 2008), even assigning subjects to different treatments upon failure or noncompliance (Lei et al., 2012).

We limit our focus to *sequential experiments*, where subjects enter iteratively over time and the experimental condition is administered upon entrance. We develop a new adaptive design for sequential experiments whose goal is to elucidate an average treatment effect (ATE) between two treatments, which we call treatment (T) and control (C). Sequential experiments are very popular in both clinical trials and recently, crowdsourced-Internet experimentation (Horton et al., 2011; Chandler and Kapelner, 2013).

Our design is a new type of *dynamic allocation*, a means of assigning T/C to newly-arrived subjects based on decisions about previous assignments, covariates, and/or responses (Hu and Rosenberger, 2006). Proposals of dynamic allocation began with Efron (1971)'s biased coin design. Here, a coin is biased in favor of the treatment with the fewest subjects, hence leading

to better balance in treatment allocation. Wei (1977)'s urn design generalizes this biased coin procedure, but both methods disregard the subjects' covariates.

The first line of defense to balance covariates is stratification (or "blocking"), a classic strategy dating back to Fisher's agricultural experiments. Stratification becomes quickly impractical when the number of total covariate levels is large relative to sample size. Taves (1974), Wei (1978) and Begg and Iglewicz (1980) tackle these shortcomings by "minimizing" the imbalance between the treatments among all levels of covariates present. The most popular and widely implemented among these methods is Pocock and Simon (1975) whose procedure involves picking a few arbitrary functions to tailor the imbalances. The selection of these functions is still an ongoing area of research (for instance, see Han et al., 2009). Concerned by this arbitrariness, Atkinson (1982) posits a method solidly rooted in linear model theory using D_A optimality.

If the end goal of experiments is to find an effect, then the primary concern is estimator efficiency and test power. Stratification and minimization methods rely on the logic that greater balance among the covariates implies greater efficiency which is mathematically true *only* in homoskedastic linear models (Rosenberger and Sverdlov, 2008). D_A optimality iteratively maximizes efficiency assuming the linear model without explicitly focusing on balance. Thus, we see one of the fundamental problems in previous allocation procedures is the reliance on the homoskedastic linear model, an assumption that is rarely true in practice. We wish to develop a dynamic allocation which is robust when the covariates combine non-linearly and with interactions to produce the response and no worse than current methods when the linear model holds.

The seminal guidebook, Cook and Campbell (1979), states "whenever possible, it is recommended to minimize error terms" and recommend *matching* subjects before randomization on covariates to create better stratification. It was not a novel idea; Student (1931) commented on the $n = 20,000$ children Lanarkshire Milk Experiment proposing the experiment should be performed exclusively on 50 identical twin pairs which would be randomly assigned T/C.

We propose matching iteratively, *on-the-fly*. As subjects walk in the door or engage a survey online they should be matched with people "similar" to them who came in previously, a procedure which Whitehead (1997) believes to be "especially difficult." Imagine the following scenario of a trial testing whether a pill decreases blood pressure. The investigators determine that age, height, weight, and race should be collected as covariates as they are known to be related to blood pressure. Bob, 28, 5'10", 180lb and white enters and is determined to fit the requirements of the study. By the flip of a coin, he receives the pill. The next day, Grace, 45, 5'2", 105lb and Asian shows up. Based on inspection of this demographic data, she is clearly different from Bob; thus she is also randomized. Soon after, Joe, 29, 5'11", 185lb and white enters. We determine that he is similar to Bob, pair them, and deterministally administer to him the placebo. The trial continues and Mary would then await someone to be matched with, which may or may not occur.

The algorithm is simple: incoming subjects are either randomized and placed in a holding pool, called the "reservoir," or if they're found to match a subject already in the reservoir, they're matched and given their match's alternate treatment. The matches and the reservoir form two independent samples yielding two different estimators which are combined to estimate

the ATE.

The closest idea we find in the literature is in Raghavarao (1980) who computes the Maha-lanobis distances between a new subject's covariates and the average covariates in the different treatment groups. The allocation to the treatment is then made via a biased coin with proba-bilities proportional to these distances. We use the idea of Mahalanobis distance which creates robustness to collinearity in the covariates, but we use it to match individual subjects together in pairs.

We layout our scenario assumptions, explain our algorithm and develop testing procedures in section 2. We demonstrate our procedure's improvements over previous procedures via simulations in section 3. Our method performs particularly well in the case where the model is non-linear, performs respectably with linear models, and also performs respectively when the covariates do not inform the response. We then demonstrate higher efficiency using historical data from randomized controlled trials (RCT's) in section 4, where the covariate-response model was unknown but most likely non-linear. We discuss and describe future directions in section 5.

2 The Algorithm, Estimation, and Testing

2.1 Problem Formulation

Subjects arrive sequentially and their covariates, denoted by $\boldsymbol{x}_i := [x_{i1}, \ldots, x_{ip}]$ which are either continuous or binary, are immediately recorded. The subjects must then be assigned to a treatment on-the-spot. We develop our method for allocating two treatments, T or C, denoted by the treatment indicator $\mathbb{1}_{T,i}$. The response y_i is continuous and can be collected at any time after allocation. We assume the following model with independent observations, an additive treatment effect, a possibly non-linear covariate effect, normal and homoskedastic errors, fixed covariate design, and sample size n fixed in advance:

$$Y_i = \beta_T \mathbb{1}_{T,i} + z_i + \mathcal{E}_i, \qquad z_i := f(\boldsymbol{x}_i), \qquad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_e^2\right), \qquad i \in \{1, \ldots, n\}. \qquad (1)$$

We wish to develop a dynamic allocation method followed by an unbiased estimator for β_T with higher efficiency and thereby more powerful when testing a null hypothesis than previous approaches.

2.2 The Algorithm

The first few subjects enter the experiment and are randomized to T or C with the flip of a coin. These subjects comprise the "reservoir." After a certain point, we would like to potentially match an incoming subject with subjects in the reservoir. We would like to match them on $f(\boldsymbol{x})$, which is latent, so we match on what we consider is the next best thing, the

$\boldsymbol{x}$'s themselves. We hope that $\boldsymbol{x}_1 \approx \boldsymbol{x}_2$ implies $f(\boldsymbol{x}_1) \approx f(\boldsymbol{x}_2)$ which is true if the function is sufficiently smooth.

We match using squared Mahalanobis distance which gives a convenient scalar distance between points in $\mathbb{R}^p$ adjusting for collinearities. This metric has a long implementation history in matching applications dating back to Rubin (1979). Matching using Mahalanobis distance and then randomizing the pairs to T/C has been demonstrated to result in better balance and higher power (Greevy et al., 2004). Further, the assumption of normal covariates seems to work well with real data even when the covariates are non-normal (see section 4). After matches are produced, we do not make use of the normality assumption of the $\boldsymbol{x}$'s further in the delvelopment of the estimator. Improvements to the matching machinery that may be more robust to real-world covariate distributions are also discussed in section 5.

Thus, the new subject enters and the squared Mahalanobis distance between its covariate vector, $\boldsymbol{x}_{\mathrm{new}}$, and each of the previous subject covariate vectors, the $\boldsymbol{x}_{\mathrm{old}}$'s, are calculated. Denote $\boldsymbol{S}$ as the covariates' sample variance-covariance matrix calculated with all subjects including the new subject. Assuming normal covariates, the squared Mahalanobis distance then has a scaled F distribution given below:

$$D_M^2 := (\boldsymbol{X}_{\mathrm{new}} - \boldsymbol{X}_{\mathrm{old}})^\top \boldsymbol{S}^{-1} (\boldsymbol{X}_{\mathrm{new}} - \boldsymbol{X}_{\mathrm{old}}), \qquad \frac{n-p}{2p(n-1)} D_M^2 \sim F_{p,n-p} \qquad (2)$$

We then take the minimum of the squared Mahalanobis distances between the new observation and each observation in the reservoir and calculate its probability. Let the minimum distance squared come from the previous subject, $\boldsymbol{x}_{\mathrm{old}}^*$. If the probability is less than λ, a pre-specified hyperparameter, then $\boldsymbol{x}_{\mathrm{new}}$ and $\boldsymbol{x}_{\mathrm{old}}^*$ are matched together; if it's not, $\boldsymbol{x}_{\mathrm{new}}$ is randomized and added to the reservoir. If $\boldsymbol{x}_{\mathrm{new}}$ is matched, it is not added to the reservoir and $\boldsymbol{x}_{\mathrm{old}}$ is removed from the reservoir. $\mathbb{1}_{T,\mathrm{new}}$ is then assigned to be $1 - \mathbb{1}_{T,\mathrm{old}^*}$, i.e. the opposite treatment. The process is repeated until the nth entrant. We left out other implementation details in this discussion but make them explicit in algorithm 1. Note that our proposed procedure is considered a form of *covariate-adaptive randomization* (Rosenberger and Sverdlov, 2008, section 2) because we are using the covariates to determine the dynamic allocation.

Note that upon matching, the treatment indicator is assigned *deterministically* to be the opposite of its match. This can cause selection bias if the investigator is not properly blinded. In defense of our decision to make allocation deterministic for almost half the subjects, note that our algorithm is sufficiently complicated that a duplicitous investigator would not be able to guess whether the entering subject will be assigned T or C based on previous information during a clinical trial. We agree with Begg and Iglewicz (1980) that "the idea that responsible investigators, even if they knew all the allocations to date, would spend their time playing games to try to guess a relatively complicated deterministic procedure... [is not] appealing." Additionally, McEntegart (2003) discusses how this type of machination is unrealistic even in multi-center block permuted designs, which is much simpler than the allocation strategy proposed here. Further, if the procedure is implemented in an Internet-based experiment, the algorithm would be hard-coded and would not be subject to human tampering.

Algorithm 1 The sequential matching algorithm for subjects entering the experiment. The algorithm requires λ to be prespecified, which controls the ease of creating matches.

1: **for** $t \leftarrow \{1, \ldots, n\}$ **do**   ▷ n is the total sample size, fixed a priori
2:     **if** $t \leq p$ **or** reservoir empty **then**
3:        $\mathbb{1}_{T,t} \leftarrow \mathrm{Bern}\left(\frac{1}{2}\right)$ and $[\boldsymbol{x}_t, \mathbb{1}_{T,t}]$ is *added* to the reservoir   ▷ randomize
4:     **else**
5:        $\boldsymbol{S}_t^{-1}$ is calculated using $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t$ ▷ Estimate the true var-cov matrix
6:        $T^{2*} \leftarrow \frac{p(t-1)}{t-p} F^*_{\lambda, p, t-p}$ ▷ F^* is the critical cutoff from the F distribution quantile
7:        **for all** $\boldsymbol{x}_r$ in the reservoir **do**
8:           $T_r^2 \leftarrow \frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{x}_r)^\top \boldsymbol{S}_t^{-1}(\boldsymbol{x}_t - \boldsymbol{x}_r)$
9:        **end for**
10:       $T_{r^*}^2 \leftarrow \min\limits_{r}\left\{T_r^2\right\}, \quad r^* \leftarrow \arg\min\limits_{r}\left\{T_r^2\right\}$ ▷ arbitrate ties if they exist
11:       **if** $T_{r^*}^2 \leq T^{2*}$ **then**
12:          $\mathbb{1}_{T,t} \leftarrow 1 - \mathbb{1}_{T,r^*}$ ▷ assign subject t the opposite of r^*'s assignment
13:          $[\boldsymbol{x}_{r^*}, \mathbb{1}_{T,r^*}]$ is *removed* from the reservoir
14:          record $\langle \boldsymbol{x}_t, \boldsymbol{x}_{r^*} \rangle$ as a new match
15:       **else**
16:          $\mathbb{1}_{T,t} \leftarrow \mathrm{Bern}\left(\frac{1}{2}\right)$ and $[\boldsymbol{x}_t, \mathbb{1}_{T,t}]$ is *added* to the reservoir   ▷ randomize
17:       **end if**
18:     **end if**
19: **end for**

2.3 Estimation and Hypothesis Testing

We first assume the covariate design matrix is fixed. If so, the subjects ultimately matched and the subjects ultimately found in the reservoir are fixed as well. Thus conditioning on the design is the equivalent to conditioning on the sigma field given below:

$$\mathcal{F} = \sigma(\underbrace{\langle \boldsymbol{x}_{T,1}, \boldsymbol{x}_{C,1} \rangle, \langle \boldsymbol{x}_{T,2}, \boldsymbol{x}_{C,2} \rangle, \ldots, \langle \boldsymbol{x}_{T,m}, \boldsymbol{x}_{C,m} \rangle}_{\text{matched pairs}}, \underbrace{\boldsymbol{x}_{R,1}, \boldsymbol{x}_{R,2}, \ldots, \boldsymbol{x}_{R,n_R}}_{\text{reservoir}}).$$

Upon completion of the experiment, there are m matched pairs and n_R subjects in the reservoir, both quantities fixed since the sample size and the design are fixed ($n = 2m + n_R$). In our development of estimators and testing procedures, we always assume conditioning on $\mathcal{F}$, thus this notation is withheld going forward.

We focus on testing the classic hypotheses $H_0 : \beta_T = \beta_0$ versus $H_a : \beta_T \neq \beta_0$. We consider testing under three model assumptions (a) the response has normal noise and may possibly depend on covariates but we do not wish to model their effect (b) the response has normal noise and depends linearly on covariates (c) the response has mean-centered noise and depends on covariates through an unknown model. We develop testing procedures for each situation: (a) a modification to the classic $\bar{Y}_T - \bar{Y}_C$ in section 2.3.1, (b) a modification to ordinary least squares regression in section 2.3.2 and (c) an exact permutation test in section 2.3.3.

5

2.3.1 The Classic Test

We define $\bar{D}$ as the estimator for the average of the differences of the m matched pairs (treatment response minus control response) and $\bar{Y}_{R,T}$, $\bar{Y}_{R,C}$ as the estimators for the averages of the treatments and controls in the reservoir. We combine the estimators using a weight parameter, $B_T := w\bar{D} + (1-w)\left(\bar{Y}_{R,T} - \bar{Y}_{R,C}\right)$. We can find w to minimize variance to obtain:

$$B_T = \frac{\sigma_R^2 \bar{D} + \sigma_{\bar{D}}^2 \left(\bar{Y}_{R,T} - \bar{Y}_{R,C}\right)}{\sigma_R^2 + \sigma_{\bar{D}}^2}, \qquad \mathbb{V}\mathrm{ar}\left[B_T\right] = \frac{\sigma_R^2 \sigma_{\bar{D}}^2}{\sigma_R^2 + \sigma_{\bar{D}}^2}. \tag{3}$$

B_T is unbiased because $\bar{D}$ and $\bar{Y}_{R,T} - \bar{Y}_{R,C}$ are unbiased. Standardizing B_T gives a standard normal due to the assumption of normal noise. To create a usable test statistic, note that the true variances are unknown, so we plugin $S_{\bar{D}}^2$, the matched pairs sample variance estimator, and S_R^2, the pooled two-sample reservoir variance estimator:

$$S_{\bar{D}}^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(D_i - \bar{D}\right)^2,$$

$$S_R^2 = \frac{\sum_{i=1}^{n_{R,T}} \left(Y_{R,T,i} - \bar{Y}_{R,T}\right)^2 + \sum_{i=1}^{n_{R,C}} \left(Y_{R,C,i} - \bar{Y}_{R,C}\right)^2}{n_R - 2} \left(\frac{1}{n_{R,T}} + \frac{1}{n_{R,C}}\right).$$

We use the notation $n_{R,T}$ and $n_{R,C}$ to be the number of treatments and controls in the reservoir ($n_R = n_{R,T} + n_{R,C}$). In practice, $n_{R,T}$ is a random quantity, a binomial distribution with size n_R and probability of one half. Howevere, we will assume $n_{R,T}$ and $n_{R,C}$ are fixed as an approximation. A more careful calculation could include the randomness of $n_{R,T}$ and $n_{R,C}$.

Equation 4 shows the resulting statistic which has an asymptotically standard normal distribution since $S_{\bar{D}}^2 \xrightarrow{p} \sigma_{\bar{D}}^2$ and $S_R^2 \xrightarrow{p} \sigma_R^2$. Also, by the assumption of additive and normal noise, the estimator is also unbiased for finite n.

$$\frac{B_T - \beta_0}{\mathrm{SE}\left[B_T\right]} \approx \frac{\dfrac{S_R^2 \bar{D} + S_{\bar{D}}^2 \left(\bar{Y}_{R,T} - \bar{Y}_{R,C}\right)}{S_R^2 + S_{\bar{D}}^2} - \beta_0}{\sqrt{\dfrac{S_R^2 S_{\bar{D}}^2}{S_R^2 + S_{\bar{D}}^2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{4}$$

Note that in the case where there are no matched pairs, we default to the classic estimator and in the case where there are less than two treatments or controls in the reservoir, we default to the matched pairs estimator.

When is this estimator more efficient than the standard classic estimator, $\Delta\bar{Y} := \bar{Y}_T - \bar{Y}_C$? In other words, when is $\sigma_{\Delta\bar{Y}}^2 / \sigma_{\bar{D}}^2 > 1$? Assuming perfect balance in its treatment allocation ($n_T = n_C = \frac{n}{2}$) for the classic estimator and taking the expectation over both noise and treatment allocation, it can be shown that the variances are:

6

$$\sigma_{\bar{D}}^2 = \frac{1}{m^2} \sum_{k=1}^{m} (z_{T,k} - z_{C,k})^2 + \frac{2}{m} \sigma_e^2, \qquad \sigma_{\Delta\bar{Y}}^2 \approx \frac{4}{n^2} \sum_{i=1}^{n} z_i^2 + \frac{4}{n} \sigma_e^2. \tag{5}$$

This means that the better the matching, the smaller $\sum_{k=1}^{m} (z_{T,k} - z_{C,k})^2$ will be, the smaller the variance becomes, and the higher the power. If we further allow $n_R = 0$ (*all* the subjects matched), then it's clear that $\sigma_{\Delta\bar{Y}}^2 / \sigma_{\bar{D}}^2 > 1$ if and only if $\sum_{k=1}^{m} z_{T,k} z_{C,k} > 0$. Note that the approximation in the last expression is due to ignoring covariance terms which do not exist when conditioning on n_T and n_C.

2.3.2 The Least Squares Test

To construct a test when the response is linear in the covariates or when we wish to make linear adjustments, we extend the idea in the previous section where we combined an effect estimate from the matched pairs data and an effect estimate from the reservoir data to regression models. Consider the following model for the response differences among the matched pairs:

$$D_k = \beta_{0,D} + \beta_{1,D} \Delta x_{1,k} + \ldots + \beta_{p,D} \Delta x_{p,k} + \mathcal{E}_{k,D}, \qquad \mathcal{E}_{k,D} \overset{iid}{\sim} \mathcal{N}\left(0, \tau_D^2\right), \qquad k \in \{1, \ldots, m\}.$$

The parameter of interest is the intercept, $\beta_{0,D}$, with OLS estimator $B_{0,D}$, the analogue of $\bar{D}$ in the previous section. $\Delta x_{1,k}, \ldots, \Delta x_{p,k}$ are the differences between treatment and control within match k for each of the p covariates respectively. $\beta_{1,D}, \ldots, \beta_{p,D}$ are nuisance parameters that adjust for linear imbalances in the covariate differences not accounted for in the matching procedure.

For the responses in the reservoir, consider the model:

$$Y_i = \beta_{T,R} \mathbb{1}_{T,R,i} + \beta_{0,R} + \beta_{1,R} x_{1,i} + \ldots + \beta_{p,R} x_{p,i} + \mathcal{E}_{i,R}, \qquad \mathcal{E}_{i,R} \overset{iid}{\sim} \mathcal{N}\left(0, \tau_R^2\right), \qquad i \in \{1, \ldots, n_R\}.$$

The parameter of interest is the additive effect of the treatment, $\beta_{T,R}$, with OLS estimator $B_{T,R}$, the analague of $\bar{Y}_{R,T} - \bar{Y}_{R,C}$ in the previous section. $\beta_{0,R}, \beta_{1,R}, \ldots, \beta_{p,R}$ are nuisance parameters that adjust for a mean offset and linear imbalances in the covariates.

Using the parallel construction in equations 3 and 4, our modified OLS estimator has the form

$$\frac{B_{T,OLS} - \beta_0}{\text{SE}\left[B_{T,OLS}\right]} \approx \frac{\dfrac{S_{B_{T,R}}^2 B_{0,D} + S_{B_{0,D}}^2 B_{T,R}}{S_{B_{T,R}}^2 + S_{B_{0,D}}^2} - \beta_0}{\sqrt{\dfrac{S_{B_{T,R}}^2 S_{B_{0,D}}^2}{S_{B_{T,R}}^2 + S_{B_{0,D}}^2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{6}$$

where $S_{B_{T,R}}^2$ is the sample variance of $B_{T,R}$ and $S_{B_{0,D}}^2$ is the sample variance of $B_{0,D}$.

2.3.3 The Permutation Test

An application of Fisher's exact test is straightforward. For the matched pairs component of the data, we examine the 2^m configurations (each match can have T-C or C-T) to compute all $\bar{d}$'s. For the reservoir portion of the estimator, we condition on $n_{R,T}$[1] and examine every possible arrangement of the treatment vector to compute every $\bar{y}_{R,T} - \bar{y}_{R,C}$. For each arrangement, we also compute s_D^2 and s_R^2 to create values of the test statistic in equation 3. In practice, the 2-sided p-value is approximated bv comparing the observed b_T from the true sample data to Monte-Carlo samples from the space of all possible test statistics. A similar exact test is available using the modified OLS estimates.

2.4 Properties of the Matching Algorithm

We wish to gain insight about how λ and n affect n_R. Assume for now that we only have one covariate, x (which may also be the largest principal component of a collection of covariates). Mahalanobis distance matches on standardized distance. For this illustration, assume we match if the two x's sample quantiles are within λ of each other. For example, the latest subject in the experiment had a sample quantile of 0.96, they would be matched to the closest subject in the reservoir with quantile between 0.91 and 1 at $\lambda = 0.10$.

Consider dividing the unit interval into $K := 1/\lambda$ intervals of equal length. Two items in one interval qualify to be matched. Assume that K is even (similar results follow for K odd). Consider the Markov process that transitions after each pair of subjects. Let s be the state that $2m$ of the K cells are occupied for $s \in \{0, 1, \ldots, K/2\}$. It is straightforward that $P_{i,j}$, the transition probability of pairs from state i to j, satisfies:

$$
P_{s,j} = \begin{cases}
\frac{2s(2s-1)}{K^2}, & j = s-1, \quad s \neq 0 \\
\frac{K(4s+1)-8s^2}{K^2}, & j = s \\
\frac{K^2-K(4s+1)+2s(2s+1)}{K^2}, & j = s+1, \quad s \neq \frac{K}{2}
\end{cases}
$$

Note the inherent symmetry: $P_{s,j} = P_{K/2-s,K/2-j}$. Hence, the steady-state probabilities are symmetric about $s = K/4$. Therefore, the mean number of items in the reservoir goes to $K/2 = (2\lambda)^{-1}$ as n grows arbitrarily large. For example, $\lim_{n\to\infty} \mathbb{E}[N_R \mid \lambda = 0.10] = 5$.

3 Simulation Studies

We demonstrate our method's performance by simulating in three scenarios: covariates affect the response non-linearly (the "NL" scenario), covariates affect the response linearly (the "LI" scenario) and covariates do not affect the response (the "ZE" scenario). These scenarios were simulated via the settings found in table 1. In practice, we simulated many settings for the NL and LI scenarios with similar results.

---

[1]This is known as the "conditional" exact test (Rosenberger and Lachin, 2002).

Scenario	Y_i
NL	$\beta_T \mathbb{1}_{T,i} + x_{1,i} + x_{2,i} + x_{1,i}^2 + x_{2,i}^2 + x_{1,i}x_{2,i} + \mathcal{E}_i$
LI	$\beta_T \mathbb{1}_{T,i} + 2x_{1,i} + 2x_{2,i} + \mathcal{E}_i$
ZE	$\beta_T \mathbb{1}_{T,i} + \mathcal{E}_i$

Table 1: The response models for the three scenarios proposed. The covariates were $X_{1,i} \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $X_{2,i} \overset{iid}{\sim} \mathcal{N}(0, 1)$ and the errors were $\mathcal{E}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$.

We set the treatment effect to be $\beta_T = 1$. n and λ were varied over a grid found in table 2. We then used σ_e^2 to modulate the resolution in our comparisons. We chose $\sigma_e^2 = 3$ to be a good balance because even at $n = 200$ comparisons were clear.

Parameter	Simulated Values
n	$\{50, 100, 200\}$
λ	$\{0.05, 0.075, 0.10, 0.20\}$

Table 2: Parameters for simulation

In choosing which competitor dynamic allocation methods to simulate against, we wanted to pick methods that are in use in sequential trials. According to Scott et al. (2002), stratification is very popular and Efron's biased coin has been used in a few studies. Most popular is minimization which has been used in over 1,000 trials (McEntegart, 2003) while D_A optimality is not known to the authors to have been implemented to date and according to the simulations in Atkinson (1999), it does not perform dramatically better than minimization even though it rests on more solid theoretical grounds.

Thus, we choose to compare our method against complete randomization (CR), stratification,[2] Efron's biased coin design (BCD),[3] and minimization.[4]

There are three scenarios (NL, LI, and ZE) and four competitors. Naturally, we want to gauge performance if we assumed the correct underlying model, but we also want to ensure we are robust if the model is misspecified. Therefore, we simulate each of these under the three model assumptions discussed in section 2.3. For the classic estimator, all competitors employed $\bar{Y}_T - \bar{Y}_C$; for the linear estimator, all competitors employed OLS; and for the exact test, all competitors employed the standard conditional permutation test.

We hypothesize that in the case of no effects (the ZE scenario), we will slightly underperform against competitors under all three testing procedures because of the loss of power due to the lower effective sample size when analyzing paired differences. If the effects are linear, we

---

[2]Both x_1 and x_2 were stratified into three levels based on the 33.3%ile and 66.6%ile of the standard normal distributions. Thus, we create nine blocks. Within blocks we alternate T / C in order to coerce $n_T \approx n_C$ so no power is lost on allocation imbalance. We then ran OLS using the nine blocks as well as x_1 and x_2 as covariates.

[3]We use the bias parameter of $\alpha = 2/3$ which is Efron's "personal favorite."

[4]We used the same blocks as stratification. For the "D" function, Begg and Iglewicz (1980) compared Pocock and Simon (1975)'s range and variance methods using simulations and found the variance method performed slightly better, thus we implement variance as the "D" function and sum as the "G" function. We set $p = 1$ for deterministic assignments in order to force $n_T \approx n_C$ so no power is lost on allocation imbalance.

hypothesize to do slightly worse against the OLS procedures due to lower sample size. Under all other scenarios and models, we expect to do better.

We simulated each scenario 1,000 times and for exact testing, we Monte-Carlo sampled 1,000 times within each simulation.[5] The levels of λ only minorly affect the comparisons against the four competitors. We empirically observed $\lambda = 10\%$ to be optimal, so we display those results.

n	Allocation Method	Balance	Sample Relative Efficiency Over Competitors					
			Scenario NL		Scenario LI		Scenario ZE	
			Classic	Linear	Classic	Linear	Classic	Linear
	CR	0.816	1.954	1.721	2.572	0.756	0.983	0.898
	Efron's BCD	0.815	2.048	1.576	2.344	0.752	0.915	0.805
50	Stratification	0.423	1.484	1.236	1.198	0.613	0.930	0.834
	Minimization	0.395	1.694	1.426	1.141	0.692	0.961	0.842
	Seq. Matching	**0.587**	—	—	—	—	—	—
	CR	0.798	2.545	1.721	2.748	1.020	0.879	0.903
	Efron's BCD	0.808	2.687	1.576	2.904	0.956	1.001	0.878
100	Stratification	0.390	1.893	1.236	1.234	0.875	0.887	0.911
	Minimization	0.369	2.111	1.426	1.220	0.907	0.970	0.881
	Seq. Matching	**0.497**	—	—	—	—	—	—
	CR	0.823	2.649	1.942	2.947	1.029	0.980	0.852
	Efron's BCD	0.812	2.541	1.948	3.155	1.003	1.033	0.970
200	Stratification	0.379	1.585	1.451	1.475	0.905	0.951	0.913
	Minimization	0.369	1.821	1.750	1.159	0.892	0.977	0.995
	Seq. Matching	**0.419**	—	—	—	—	—	—

Table 3: Balance results and relative sample efficiency results of sequential matching versus competitors by scenario and testing procedure. Efficiencies in red indicate our algorithm performed worse than a competitor via an F-test with 1% significance level and efficiencies in green indicate our algorithm performed better. Note that we did *not* adjust for multiple comparisons. Balance results are averages over all scenarios and two model assumptions (6,000 simulations). Exact tests are not shown because they do not admit a standard error calculation.

Our main metric for comparison is power, the proportion of the times the null was rejected under the Type I error rate of $\alpha = 5\%$. We also record standard error of the estimate (when the estimator was parametric) as well as balance.[6] Results for power against the null of no treatment effect are illustrated in figure 1 and results for balance and relative efficiency vis-a-vis other methods are found in table 3.

In the NL scenario, our sequential matching procedure dominates competitors in power and efficiency, sometimes doubling power and nearly tripling efficiency. Even at $n = 200$, there

---

[5]Original data and source code for reproduction can be found at `github.com/kapelner/sequential_matching_simulations`

[6]By "balance" we mean the maximum standardized difference in the averages of covariates between treatment and control samples: $\max_{j \in \{1,2\}} \{(\bar{x}_{j,T} - \bar{x}_{j,C})/s_j\}$.
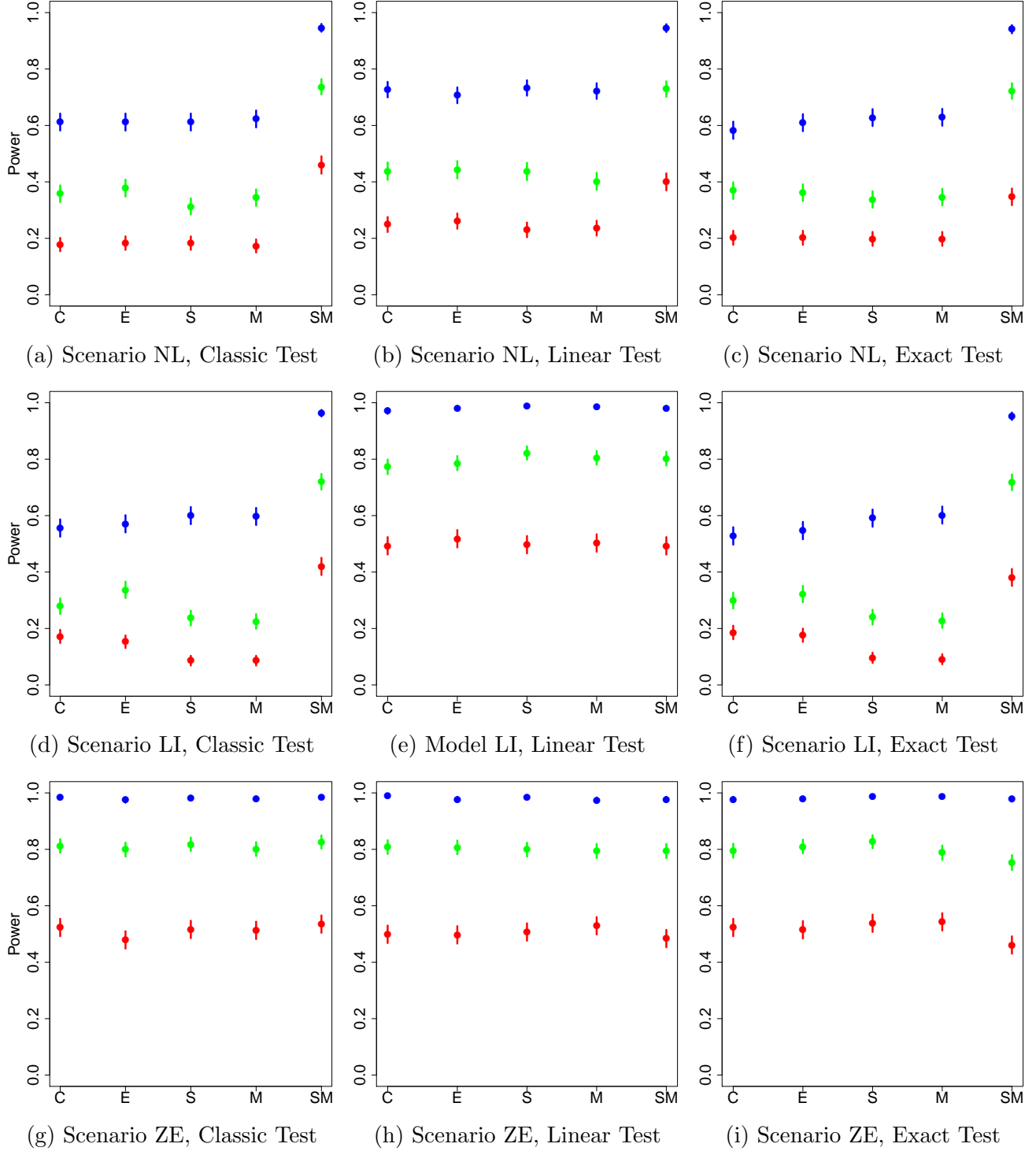
Figure 1: Power illustrated for the three scenarios by the three testing procedures, all five allocation methods (C: Complete Randomization, E: Efron's BCD, S: Stratification, M: Minimization, SM: our sequential matching algorithm) and sample sizes (red illustrates results for $n = 50$, green for $n = 100$ and blue for $n = 200$). Plotted points represent the sample proportion of null hypothesis rejections in 1,000 simulations and segments represent 95% confidence intervals. Matching parameter $\lambda = 10\%$.

are still gains. Regression adjustment helps the competitors, but it cannot adjust for the non-linear portion of the quadratic terms and interaction term; they will appear as higher noise.

In the LI scenario, sequential matching dominates competitors in the classic and the exact test because the competitors do not use the covariate information. In the linear assumption, sequential matching performs similarly in power but has a lower efficiency at $n = 50$. This loss is due to a lower effective sample size due to using matched pairs. The loss does not continue at $n > 50$. This can be explained that there are benefits to matching even when employing regression adjustment. As Greevy et al. (2004) explain, better balance reduces collinearity resulting in a smaller standard error for the estimate. Balance is improved over competitors that do not allocate based on the covariates and this better balance implies higher power and efficiency. Parenthetically, we note that as n increases, it appears as if balance is approaching levels observed in both stratification and minimization. This is expected and is an added bonus of our procedure.

In the ZE scenario, our approach is most severely impacted by lower effective sample size. However, power is not as low as expected. Efficiency seems to be lost for all simulated n but most significantly when $n = 50$.

All in all, sequential matching performs well in the scenario of the response being non-linear in the covariates which is the most realistic case in practice. If the covariate model is truly linear, we only do worse when OLS is employed but our inefficiency is only for small sample sizes. In the case when covariates do not matter at all, we do worse for low sample size, but about equal when $n \geq 100$.

All in all, sequential matching shines in the case of non-linear covariate models which is the most realistic case in practice. If the covariate model is truly linear, sequential matching does worse when OLS is employed but our relative inefficiency is only observed for small sample sizes. In the case when covariates do not matter at all, we begin to perform about equally with competitors when $n \geq 100$. This is an important result in practice because investigators sometimes choose useless covariates which do not affect the outcome measure.

A possible criticism of the high power achieved is we assume our n was large enough for the estimators in equations 4 and 6 to converge. To assuage this concern, we simulated the size of the tests in table 4. For the classic estimator, even at $n = 50$, the size is about 7-8% and by $n = 100$ it's close to the Type I error rate of 5%. For the OLS estimator, convergence is a tad slower. In unshown simulations, we have observed that the modified classic estimator can be approximated by a T_{m-1} distribution which can be used to compute a more conservative significance level. Even under this conservative approximation, the simulated power in figure 1 do not change dramatically (unshown). Other anomalies observed in this table are discussed in section 5.

n	Allocation Method	Scenario NL			Scenario LI			Scenario ZE		
		Classic	Linear	Exact	Classic	Linear	Exact	Classic	Linear	Exact
	CR	0.039	0.038	0.058	0.041	0.057	0.051	0.048	0.060	0.042
	Efron's BCD	0.056	0.049	0.056	0.046	0.051	0.049	0.049	0.049	0.052
50	Stratification	0.019	0.026	0.016	0.005	0.010	0.048	0.048	0.051	0.048
	Minimization	0.036	0.039	0.043	0.007	0.004	0.061	0.049	0.049	0.067
	Seq. Matching	0.075	0.078	0.051	0.071	0.051	0.080	0.073	0.068	0.044
	CR	0.041	0.039	0.035	0.043	0.054	0.048	0.049	0.066	0.052
	Efron's BCD	0.059	0.059	0.062	0.062	0.063	0.052	0.035	0.040	0.055
100	Stratification	0.023	0.040	0.023	0.006	0.045	0.002	0.056	0.038	0.051
	Minimization	0.016	0.036	0.043	0.003	0.049	0.008	0.046	0.041	0.040
	Seq. Matching	0.066	0.076	0.047	0.058	0.058	0.058	0.063	0.071	0.044
	CR	0.042	0.042	0.042	0.047	0.048	0.056	0.054	0.052	0.045
	Efron's BCD	0.056	0.046	0.041	0.047	0.044	0.055	0.054	0.053	0.060
200	Stratification	0.020	0.033	0.023	0.001	0.050	0.005	0.045	0.042	0.050
	Minimization	0.025	0.044	0.051	0.001	0.045	0.001	0.059	0.064	0.056
	Seq. Matching	0.050	0.062	0.051	0.046	0.059	0.056	0.048	0.065	0.065

Table 4: Simulated size of tests for all scenarios, competitors, and all tests at $\lambda = 10\%$. Numbers in red indicate they are different from the purported 5% size at a Bonferroni-corrected significance level (135 comparisons). Numbers in orange indicate they are different from the purported 5% size without Bonferroni correction.

4 Demonstration Using Real Data

We now examine sequential experiment data from two real applications: one from a behavioral study on the Internet and one from a double-blind drug trial. We simulate the subjects being dynamically allocated using the sequential matching procedure by first assuming the entering subjects do not exhibit any time trend; this will allow us to permute their order. During the iterative procedure, all subjects assigned to the reservoir keep whichever assignment they had during the experiment. During matching, if the subject happened to have been assigned the treatment which SM allocates, they are kept in the subject pool; if not, they are discarded (this is illustrated in figure 2). Thus, during our simulations, we result in a *subset* of the data we began with. Note that we only show results for the classic estimator versus the modified estimator in equation 4, not the OLS modified estimator whose results we suspect to be similar.

4.1 Behavioral Experiment

Kapelner and Chandler (2010) ran experiments using the Amazon Mechanical Turk platform, a global outsourcing website for small one-off tasks that can be completed anonymously on the Internet. They focused on measuring subjects' stated preference for a beer price when the beer came from different purchasing locations (an online replication of Thaler, 1985's demonstration of the "framing effect," a cognitive bias). The treatment involved subtle text manipulations: the same beer came from either a *fancy resort* or a *run-down grocery store*.

```
nsim: 1   .............o.xx....o....xxx.ooxx.o.xxooxooxx.xo.  (37)
nsim: 2   ...............xxx..x.xxoo.xxx.o.o.oo..ox.xxxxx.o  (35)
nsim: 3   ...............x.x..oo.xxxo.x.xxxxxxxo.x.ox.oox.o  (34)
nsim: 4   ...........o...o...ox.o.o.xo.ox.o....xxooox.oox.xo  (42)
nsim: 5   ...............x.xoo.o.o..xxoxx.o.x.x.oo.o.xoxx..  (39)
nsim: 6   ...............x..xx.x..xx.oo.xoxoxxx.ooxo.o.xxxo  (35)
nsim: 7   ............x.o..oo..oxx.x..o.oxo.xxox.o.xoxxxoxo  (37)
nsim: 8   ...........xx.oxoxxo..x.xxx.x..oxoo.o.xx.x..xox.o  (34)
nsim: 9   ..............o...xxoo.xo.ooo.....o.ooxox.oxx..o.x  (42)
```

Figure 2: Running an $n = 50$ subset of historical data through the sequential procedure. The dots represent a subject being placed into the reservoir. The "o" signifies that the subject was matched and that their treatment allocation was *opposite* of their matching partner. The "x" signifies that the subject was matched but their treatment allocation was the *same* as their matching partner, resulting in the subject being discarded. The number in parenthesis at the end of the line is the sample size retained of the purported 50.

In their control wing ($n = 168$), no tricks were employed to ensure the subjects were paying attention to the text. Thus, in this wing, the subtle text manipulation did not seem to affect the subjects' stated beer prices. The effect may have been real, but the data was either too noisy or there was insufficient sample size to find it. We demonstrate here that if our sequential matching procedure was employed, the effect estimator would have been more efficient.

For matching, we first used most of the covariates found in the original dataset: age, gender, level of earnings, number of weekly hours spent doing one-off tasks, level of multitasking when performing tasks, stated motivation level, passing the "instructional manipulation check" (Oppenheimer et al., 2009) and a survey gauging the subject's "need for cognition" (Cacioppo and Petty, 1982).

We note that R^2 under OLS was about 18.7%. We then run two off-the-shelf machine learning algorithms that are designed to find interactions and non-linearities in the response function. The in-sample pseudo-R^2 using Chipman et al. (2010)'s Bayesian Additive Regressive Trees (BART) was 42.4% and Breiman (2001)'s Random Forests (RF) was 70.4%. Although this is not a formal test, it is pretty compelling evidence that the covariates do not combine strictly linearly to inform beer price. Thus, as demonstrated in figure 1a and column 4 of table 3, our method should be more powerful and more efficient than using previous dynamic allocation strategies with a classic estimator. The results for 200 simulations at $\lambda = 0.10$ are shown in table 5. Many of the covariates are binary. Thus, the variance-covariance matrix was *not* invertible in line 5 of algorithm 1 for many of the early iterations, so we used the Moore-Penrose generalized inverse instead.

We now match on four selected covariates that come out most significant in an OLS regression on the full dataset: age, level of earnings, level of multitasking when performing tasks and one question from the survey gauging the subject's "need for cognition." The results are found in table 6. Note that the efficiencies are higher and do not drop off as quickly when n increases. Thus, matching on *relevant* covariates yields a performance enhancement in our procedure.

14

purported n	actual n (average)	average efficiency	approximate sample size reduction
50	37.8	1.84	45.7%
100	71.9	1.23	16.9%
168 (all)	116.1	1.06	5.4%

Table 5: Results for 200 simulations over many values of n and $\lambda = 0.10$ in the case where most of the covariates are matched on.

purported n	actual n (average)	average efficiency	approximate sample size reduction
50	34.9	2.01	50.1%
100	67.8	1.60	37.3%
168 (all)	112.1	1.57	36.3%

Table 6: Results for 200 simulations over many values of n and $\lambda = 0.10$ in the case where four cherry-picked covariates are matched on. OLS has an $R^2 = 20.8\%$, BART, 33.1% and RF, 26.5%.

4.2 Clinical Trial

We use data from Foster et al. (2010), a twelve-week, multicenter, double-blind, placebo-controlled clinical trial studying whether amitriptyline, an anti-depressant drug, can effectively treat painful bladder syndrome. The study measured many outcomes, including change in pain after 12 weeks (difference in Likert scale scores). The confidence interval for the ATE between pill and placebo for this outcome measure was $[-1.00, 0.30]$ with a significance level of $p_{\mathrm{val}} = 0.29$ (table 2, row 1, page 1856). Most likely the effect is real but there wasn't enough power to detect it either via a low sample size or a high error variance.

Once again, for matching, we first used most of the covariates found in the original dataset: age, gender, race (white / hispanic), level of education, level of employment, living with a partner, presence of sexually transmitted diseases and urinary tract infection, as well as baseline measures of pain, urination frequency and urgency, quality of life, anxiety and depression as well as syndrome symptom levels. We note that R^2 was about 25.2% under OLS, 42.6% under BART, and 82.4% under RF which is once again compelling evidence that the covariates do not combine strictly linearly to inform the subject's week 12 pain measure. By figure 1a and column 4 of table 3, our procedure should be more powerful and more efficient at estimating the ATE. The results for 200 simulations are shown in table 7. The covariates again include many binaries, so we used the Moore-Penrose generalized inverse for the inverse variance-covariance matrix calculation.

Once again, we now match on the top four covariates which are the most significant in an OLS of the full dataset: living with a partner and baseline pain, frequency and syndrome symptom levels. The results are found in table 8. Note again that the efficiencies are higher when comparing to matching on all covariates and they do not suffer the steep drop off as n increases.

purported n	actual n (average)	average efficiency	approximate sample size reduction
50	38.9	1.30	23.0%
100	75.2	1.10	9.2%
150	111.3	1.05	4.9%
224 (all)	165.5	1.07	6.7%

Table 7: Results for 200 simulations of the sequential matching procedure over many values of n and $\lambda = 0.10$ in the case where most of the covariates are matched.

purported n	actual n (average)	average efficiency	approximate sample size reduction
50	38.0	1.27	21.2%
100	72.9	1.23	18.8%
150	108.6	1.15	13.2%
224 (all)	160.3	1.13	11.3%

Table 8: Results for 200 simulations of the sequential matching procedure over many values of n and $\lambda = 0.10$ in the case where four cherry-picked covariates are matched on. OLS has an $R^2 = 18.9\%$, BART, 35.1% and RF, 45.0%.

5   Discussion

Estimation in sequential experiments can have higher power and efficiency if the covariate information is leveraged. We present a dynamic allocation of treatment and control that matches subjects on-the-fly via a novel algorithm and present modified estimators of classic approaches: average difference, linear regression, and permutation testing. We simulate under different scenarios and illustrate higher power in scenarios where competing methods cannot make proper use of covariate information. We underperform only in the case of low sample size when the covariate model is linear or non-existent. In simulations with real data, we find the efficiency of our method increases as the covariate function becomes more important. In the two real data sets from a clinical trial and an online behavioral study, our method is more efficient over complete randomization. This is most likely due to the fact that real-world response functions combine covariates non-linearly, and this is when our procedure is most advantageous.

We note that "analysis assumptions may be compromised due to the 'pseudo'-random allocation" (Scott et al., 2002) and would like to address this criticism which can be made about our procedure. Note that in table 4, the size of the tests under stratification and minimization are less than 5%. One should not use the classic estimator in these cases because one implicitly tried to balance on the covariates, but then did not include the covariates in the model (as seen by the terrible sizes using the classic model in scenario NL using the classic model in scenario LI in table 4). Simon (1979) and Senn (2000) have very good discussions about this issue and recommend using regression adjustment (as seen in the competitors in scenario LI, model linear in the table 4 for $n > 50$). As for exact testing under dynamic allocation, Kalish and Begg (1987) and many others warn that permutation distributions in stratification and minimization are incorrect unless the investigator permuted the treatment allocation *according to its structure*, that is, according to how the allocation was determined

by the covariates. This is not straightforward in practice for stratification and even more complicated for minimization.

In contrast, the sizes of the test for our apporach seem to be correct in table 4 (barring the convergence of our estimator which we talked about at the end of section 3). Usage of the classic estimator even for our method seems ill-advised for the same reasons that it is not recommended under stratification and minimization. However, using linear regression on the covariates is also ill-advised when the model is non-linear or otherwise does not satisfy the OLS model assumptions (Freedman, 2008). However, Rubin (1973, 1979) finds that covariance adjustment of matched pair differences is very robust to model misspecification.

Thus, since our permutation test performs well in the non-linear case and can avoid many of the above issues, we recommend the permutation test in practice. We permuted according to the structure of our dynamic allocation via matching, thus our permutation tests are valid (this is also a reason not to employ anything but complete randomization in our reservoir).

5.1 Further Developments

We view this contribution as a step forward in covariate-adaptive randomization in sequential experiments but it is far from complete. We begin with extensions that can be immediately implemented.

Although we assume fixed n in our construction, it is relatively straighforward to adapt to a fully or group sequential design whose methods can be found in Jennison and Turnbull (2000). It would be hard to tabulate values when our estimator has unknown convergence properties, thus it would probably have to be done by waiting until the estimator most likely converged, and then using standard sequential analysis software.

Fruitful areas of further work would be extensions to $k > 2$ treatments which can be done by making matches of size k. We believe our methods can apply beyond continuous responses to binary, ordinal, or count responses but this will involve some adjustment of the estimators. Fixed design is a simplifying assumption, but unrealistic with real-world covariates, begging a random-X robust implementation. The estimators found in Pitkin et al. (2013) can also be plugged into equations 4 and 6. We also believe there may be more clever estimators that can be constructed using the data in the matched pairs and reservoirs. We previously tried random effects models without success, but there may be others. Additionally, it would be straightforward to introduce a biased coin design to the matching algorithm to avoid possible tampering due to the partially deterministic allocation.

We feel that the most significant improvement would be better matching. Mahalanobis distance is logical, but prone to strange behavior with departures from the normality assumption. Another natural extension is to boostrap the distribution of the nominal metric in equation 2 as to not rely on probabilities from the scaled F distribution. Also, practitioners may want to weight the variables in the matching as well as force some variables to always match (these ideas and others are discussed in Rosenbaum 2010, chapter 8). Additionally, in the observational study literature, matching is elaborately engineered to improve balance across

entire groups of observations (Zubizarreta, 2012). It is possible some of these methods can be applied to better matching for single pairs.

Of course, our procedure also suffers from the central issue of matching: selecting the variables to match on. A poor choice makes a big difference as evidenced by the simulations on historic RCT data (table 5 vs. 6 and table 7 vs. 8). There may be a way to *iteratively* match on covariates that are found to be important, so the set of perceived important covariates is updated during the sequential experiment.

We have begun to consider how large n_R grows asymptotically as a function of λ (section 2.4). There is a lot of theory to be done to figure out the optimal λ to maximize estimator efficiency as functions of n, p, the variance-covariance matrix of the covariates, and how strong the signal of f is to the noise. Also, perhaps a variable rule for λ would be effective: if the sample size is large, the algorithm can afford to be conservative about the matches during the beginning of the experiment (such as waiting until n_0 to begin matching), but then become less conservative as time passes.

Acknowledgements

References

AC Atkinson. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69(1):61–67, 1982.

AC Atkinson. Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in medicine*, 18(14):1741–52, 1999.

CB Begg and B Iglewicz. A treatment allocation procedure for sequential clinical trials. *Biometrics*, 36(1):81–90, 1980.

L Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

JT Cacioppo and RE Petty. The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131, 1982.

D Chandler and A Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013.

HA Chipman, EI George, and RE McCulloch. BART: Bayesian Additive Regressive Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

SC Chow and M Chang. Adaptive design methods in clinical trials - a review. *Orphanet journal of rare diseases*, 3:11, 2008.

TD Cook and DT Campbell. *Quasi-Experimentation Design & Analysis Issues for Field Settings*. Houghton Mifflin Company, 1979.

B Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

HE Foster, PM Hanno, JC Nickel, CK Payne, RD Mayer, DA Burks, CC Yang, TC Chai, KJ Kreder, KM Peters, ES Lukacz, MP FitzGerald, L Cen, JR Landis, KJ Propert, W Yang, JW Kusek, and LM Nyberg. Effect of amitriptyline on symptoms in treatment naïve patients with interstitial cystitis/painful bladder syndrome. *The Journal of urology*, 183(5):1853–8, 2010.

DA Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

R Greevy, B Lu, JH Silber, and PR Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics (Oxford, England)*, 5(2):263–75, 2004.

B Han, NH Enas, and D McEntegart. Randomization by minimization for unbalanced treatment allocation. *Statistics in medicine*, 28:3329–3346, 2009.

JJ Horton, DG Rand, and RJ Zeckhauser. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011.

F Hu and WF Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons, Inc., 2006.

C Jennison and BW Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall / CRC, 2000.

LA Kalish and CB Begg. The impact of treatment allocation procedures on nominal significance levels and bias. *Controlled clinical trials*, 8(2):121–35, 1987.

A Kapelner and D Chandler. Preventing satisficing in online surveys: A "kapcha" to ensure higher quality data. In *Proceedings of CrowdConf*, 2010.

H Lei, I Nahum-Shani, K Lynch, D Oslin, and S a Murphy. A "SMART" design for building individualized treatment sequences. *Annual review of clinical psychology*, 8:21–48, 2012.

DJ McEntegart. The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Information Journal*, 37:293–308, 2003.

KL Morgan and DB Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

DM Oppenheimer, R Meyvis, and N Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4): 867–872, July 2009.

E Pitkin, L Brown, R Berk, A Buja, L Zhao, and E George. Improved precision in estimating average treatment effects. University of Pennsylvania mimeo, 2013.

SJ Pocock and R Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1):103–115, 1975.

D Raghavarao. Use of distance function in sequential treatment assignment for prognostic factors in the controlled clinical trial. *Calcutta Statist. Assoc. Bulletin*, 29:99–102, 1980.

SW Raudenbush, A Martinez, and J Spybrook. Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, 29(1):5–29, 2007.

PR Rosenbaum. *Design of observational studies*. Springer, New York, 2010.

WF Rosenberger and JM Lachin. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons, Inc., 2002.

WF Rosenberger and O Sverdlov. Handling Covariates in the Design of Clinical Trials. *Statistical Science*, 23(3):404–419, 2008.

DB Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, 1973.

DB Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.

NW Scott, GC McPherson, CR Ramsay, and MK Campbell. The method of minimization for allocation to clinical trials. a review. *Controlled clinical trials*, 23(6):662–74, 2002.

S Senn. Consensus and controversy in pharmaceutical statistics. *Journal of the Royal Statistical Society: Series D*, 49(2):135–176, 2000.

R Simon. Restricted randomization designs in clinical trials. *Biometrics*, 35(2):503–512, 1979.

Student. The Lanarkshire milk experiment. *Biometrika*, 23(3):398–406, 1931.

DR Taves. Minimization: a new method of assigning patients to treatment and control groups. *Clinical pharmacology and therapeutics*, 15(5):443, 1974.

R Thaler. Mental Accounting and Consumer Choice. *Marketing Science*, 4(3):199 – 214, 1985.

LJ Wei. A Class of Designs for Sequential Trials Clinical. *Journal of the American Statistical Association*, 72(358):382–386, 1977.

LJ Wei. An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, 73(363):559–563, 1978.

J Whitehead. *The Design and Analysis of Sequential Clinical Trials*. John Wiley & Sons, Inc., 2nd edition, 1997.

JR Zubizarreta. Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery. *Journal of the American Statistical Association*, 107(500): 1360–1371, 2012.