Discussion Papers in Economics

No. 14/2017

Treatment Allocation for Linear Models with Covariate Information

Tobias Aufenanger
University of Erlangen-Nürnberg

ISSN 1867-6707

Friedrich-Alexander-Universität Erlangen-Nürnberg
Institute for Economics
https://www.iwf.rw.fau.de/research/iwf-discussion-paper-series/

Treatment Allocation for Linear Models with Covariate Information

Tobias Aufenanger *

Friedrich-Alexander University Erlangen-Nürnberg (FAU)

This version: July 2017

Abstract

This paper analyses optimal treatment allocation of experimental units to treatment and control group. 'Optimal' means that the allocation of treatments should balance covariates across treatment and control group in a way that minimizes the variance of the treatment estimator in a given linear model. This paper shows the benefits as well as the limits of this approach. In particular, it presents a sample size formula as well as several simulations to give some intuition on the minimum as well as the maximum benefits of this approach compared to random allocation as well as to alternative methods of treatment allocation.

JEL Classification: C90, C61

Keywords: experiment design, treatment allocation

*Friedrich-Alexander University Erlangen-Nürnberg (FAU), School of Business and Economics, PO Box 3931, 90020 Nürnberg, Germany, e-mail: tobias.aufenanger@fau.de

1 Introduction

Economic experiments are a major part of economic research. The typical question analyzed within such experiments is whether a certain treatment causally influences a particular dependent variable of interest. The main difference to observational studies is that within an experiment, the researcher can control parts of the data generating process. In particular, given a sample of experimental units,[1] the researcher conducting the experiment can decide which of the units to allocate to the treatment group and which to the control group.

In practice, most experimenters decide to use random treatment allocation: Randomly allocate half of the sample to the treatment group and the other half to the control group (e.g. Manning et al., 1987; Bennmarker et al., 2013; Lucifora and Tonello, 2015; Potters and Stoop, 2016). This type of treatment allocation has a strong justification whenever the treatment effect is estimated by the differences in means of the dependent variable between treatment and control group. The major theorem underlying randomized experiments states that whenever treatments are allocated randomly, the difference in means estimator is an unbiased estimator for the average treatment effect in the sample (Deaton and Cartwright, 2016). This theorem is very appealing, since it is almost assumption free. No model regarding the effects of covariates on the dependent variable is required.

However, when sample sizes are small and the dependent variable strongly depends on covariates of the experimental units, the difference in means estimator will have a very high variance, which makes this approach impractical. Why is this the case? Take a step back a think about what the purpose of an experiment is. The main reason for economic experiments, or experiments in general, is to find causal effects. This means, for a given unit, one seeks to know the effect of applying a certain treatment *all other things being equal*. Consequently, considering two groups of experimental units, one of which receives a treatment, the other one not, one wishes for these two groups to be as comparable as possible (Saint-Mont, 2015). Whereas it is true that random treatment allocations on average creates comparable or balanced groups, for a single experiment random allocation can lead to a large degree of imbalance (Bruhn and McKenzie, 2009).

Whenever covariate information is available, there are two ways of reducing the variance of the treatment estimator without increasing the sample size. First, estimating a parametric model instead of difference in means estimation can help control for imbalances and therefore to reduce the variance of the treatment estimate (Duflo et al., 2008, p.3924). Second, a systematic allocation of treatment can rule out severe imbalances in the first place and thus reduce the variance of the treatment estimator (Begg and Iglewicz, 1980; Smith, 1984; Duflo et al., 2008; Senn et al., 2010; List et al., 2011; Howley and Storer, 2014; Ziliak, 2014; Sverdlov, 2015; Deaton and Cartwright, 2016; Athey and Imbens, 2017).

For the field of development economics, Bruhn and McKenzie (2009) provide a review on the usage of systematic treatment allocation algorithms. They find that

[1] For example people, groups of people, schools, hospitals or whatever unit is of interest for the experiment.

out of 18 reviewed papers, 15 use some kind of systematic treatment allocation algorithm. According to the authors, the most popular algorithms in economic research are stratification, multivariate matching and different re-randomization methods. Lately, Kasy (2016a) brought up a new method. In a Bayesian inference framework, he proposes to select treatment allocations that minimize the expected posterior mean squared error (MSE) of the treatment estimator, given a prior probability distribution on potential models. Schneider and Schlather (2017) propose a similar approach for frequentist inference. This approach, in particular finding experimental designs that minimize (functions of) MSEs[2] of estimators inside a given model, has a long tradition in the statistical field of optimal experimental design (see Pukelsheim (2006) for an overview).

This paper follows a similar approach to Schneider and Schlather (2017). In contrast to them, this paper concentrates on the analyses of power and necessary sample sizes under systematic treatment allocation on top of the MSE of the treatment estimator. I propose the following experimental procedure: First, commit to a specific model for the influence of the covariates and the treatment on the dependent variable.[3] In this paper, I will focus on a simple linear model. Generalization are discussed in appendix D. Second, sample the experimental units. Sampling can be either random or with respect to other considerations. Sampling methods are not discussed in this paper. Third, measure all important covariates of the experimental units (i.e. all covariates specified in the model). Fourth, allocate treatments to experimental units in a way that minimizes the variance of the treatment estimator inside the model. I will call this method of assignment *optimal model-based treatment allocation* or simply *optimal allocation*. Fifth, apply the treatment in the treatment group and measure the dependent variable in treatment and control group. Sixth, analyze the data with the same model used for allocating the treatments. This way, the allocation of treatments requires exactly the same assumptions as the analyses of the data.

The proposed assignments with this approach are mostly deterministic up to two possible allocations (see Kasy (2016b) for a formal proof of this result). Some words on this: There is a strong justification of allocation algorithms that involve randomness, whenever the method of inference from the experiment is based on the assumption of random allocation (see Athey and Imbens (2017) for a recent overview of randomization inference for economic experiments). Note however, that the assumptions of a linear model are *not* assumptions on random allocation of treatments but rather on random sampling from some infinitely large superpopulation (Freedman, 2008; Athey and Imbens, 2017). Therefore, in a linear model framework, there is no particular reason to allocate treatments randomly. For are more detailed discussion of this distinction, see Aickin (2001).

[2]For unbiased estimators, the MSE is equal to the variance of the estimator.

[3]One should always commit to a model *prior* to conducting the experiment (Deaton, 2010). When confronted with the experimental data, researchers have an incentive to choose exactly those models that yield the highest, or most significant treatment estimate. The model selection will thus depend on the realization of the error terms, which inhibits inference (see also Duflo et al., 2008, pp. 3908ff).

The goal of this paper is to show the benefits and the limits of optimal allocation compared to random allocation and alternative methods of systematic allocation in a frequentist inference framework with linear models. In particular, I am going to regard a situation in which the researcher commits to a particular model prior to allocating the treatments and show what could be gained by allocating subjects optimally compared to allocating them randomly or in any other way, given that the model is true. This paper contributes to the literature in three ways: First, it provides a rule of thumb for the sample size necessary to estimate a treatment effect of a given size with a given power at a given alpha level. This rule of thumb builds on a precise notion of covariate balance taken from medical research (see Atkinson, 2002). List et al. (2011) and Athey and Imbens (2017) present a similar sample size formula for difference in means estimation and random treatment allocation. This paper generalizes this formula to linear models and to different treatment allocation algorithms. Using this formula, I find that, for the same model, optimal allocation can reduce sample sizes by approximately m, the number of covariates in the model, compared to random allocation. Second, this paper gives some simulation evidence on how many covariates should be used in order to maximize the power of the experiment. This simulation shows that when using optimal allocation, one should control for more covariates than when using random allocation. Third, this paper compares different heuristics for binary optimization to find (near) optimal treatment allocations. I suggest to use a simple local search algorithm or a multiple local search algorithm with multiple random starting points.

This paper is structured as follows. Section 2 provides a short overview over the related literature. Section 3 introduces optimal treatment allocation for the case of linear models. This section discusses intuitions behind optimal allocation concerning the variance of the treatment estimator, the power of the experiment and the necessary sample size. Section 4 shows the connection between optimal allocation and the stratification and matching algorithm used in economic research. Section 5 presents numerical algorithms for finding optimal treatment allocations. Section 6 compares optimal allocation to other allocation algorithms in a simulation. Section 7 concludes.

2 Related Literature

2.1 Optimal Design

This paper is closely related to the field of optimal experimental design (see Pukelsheim (2006) for an overview). Optimal design approaches search for allocations of experimental units that minimize (functions of) the MSE in the chosen model. Traditionally, optimal design approaches target the case that covariates of the experimental units can be chosen freely (Elfving, 1952; Kiefer, 1959; Kiefer and Wolfowitz, 1959; Das et al., 2015) [4]. Applications of this theory can be found

[4]For example in an experiment to develop a law of gravity, an experimental unit would be one drop of a ball. Covariates could be the height from which the experimenter drops the ball,

in all fields of research, including engineering (Harville, 1974), biology (Khinkis et al., 2003), chemistry (Telen et al., 2016) and physics (Berger et al., 2017).

In economic experiments it is often the case that experimental units are attached to fixed covariates. The only possibility to change the covariates in the sample would be to exclude some experimental units from the sample and include others with different covariate values (i.e., through sampling). Whenever sampling is strictly exogenous, classical optimal design approaches can be applied to economic research (e.g. Aigner, 1979; Aigner and Balestra, 1988). However, in most experiments involving human subjects (not only in economic research, but also in psychology, medicine, sociology, etc.), samples cannot be drawn exogenously. One reason is that human subjects cannot be forced to participate in an experiment. Take economic laboratory experiments for an example. Even though one can invite specific individuals, it is not clear, whether these individuals will actually show up.

To deal with this issue, researchers have developed algorithms that find optimal treatment allocations inside a given sample. This literature started in the field of medical research (Atkinson, 1982; Sverdlov, 2015). The difference between economic experiments and medical trials is that participants of the latter typically enter the trial sequentially (Whitehead, 1997, preface). This means, at the time the n-th participant enters the experiment, the previous $n-1$ participants have already been allocated and the covariates of the next participants are unknown. In economic experiments it is often the case that all (or at a large part of) experimental units are known prior to allocating the treatments. This allows for different and more powerful allocation algorithms.

In economic research the issue of treatment allocation is getting increasing attention (Bruhn and McKenzie, 2009; Hahn et al., 2011; Horton et al., 2011; Deaton and Cartwright, 2016; Banerjee et al., 2016; Athey and Imbens, 2017). However, most algorithms discussed in economic research are not optimal allocation algorithms in the sense that they minimize the MSE or variance of the treatment estimator in any given model. Kasy (2016a) introduced optimal design algorithms for economic experiments. He targets optimal treatment allocation in a Bayesian inference framework, similar to most decision-theoretic models. In particular, he targets the case in which the researcher has a prior distribution on potential models and through this, a prior distribution of potential treatment effect sizes. The researcher uses the experiment to update her beliefs about the treatment effect in a Bayesian way. Kasy's paper shows how the researcher should optimally allocate subjects to treatment and control group, in order to minimze the posterior MSE of the treatment effect. Banerjee et al. (2016) extend this decision theoretic framework to cases in which the researcher not only aims at minimizing the MSE of the treatment estimator given her prior, but also at convincing an audience with presumably different priors. Finally Schneider and Schlather (2017) take the optimal design approach to frequentist inference. They provide a Stata ado-package that implements their approach.

the medium in which the ball is dropped, the size and weight of the ball, etc. For more examples see Atkinson et al. (2007).

2.2 Treatment Allocation in Economic Experiments

Since this paper targets optimal treatment allocation for economic experiments, it is also related to the literature on allocation algorithms previously applied to economic experiments. Bruhn and McKenzie (2009) provide an overview over the usage of systematic treatment allocation algorithms in the field of development economics. From a review of 18 research papers and a survey of 25 experts in this field, they conclude that stratified randomization, multivariate matching as well as re-randomization are the prevailing algorithms.

Stratified randomization, also called blocking, is, after purely random allocation, probably the most popular treatment allocation algorithm in economics as well as in other fields of research. The roots of this algorithm reach back to Fisher (1926). The idea is to identify strata of experimental units that are approximately equal concerning their covariates. Given these strata, the algorithm randomly assigns units to treatment and control group, such that within each stratum an equal number of units gets assigned to the treatment and to the control group. Whenever strata are of odd size, the last unit is assigned randomly to one group. For discrete covariates, strata are usually defined as all subjects that are equal in every covariate. For two binary covariates, this would make $2 \cdot 2 = 4$ strata. If a third variable can take on three different values, this makes $2 \cdot 2 \cdot 3 = 12$ strata. Continuous covariates or discrete covariates that can take on many different values, have to be discretized (see Bruhn and McKenzie, 2009; Ma and Hu, 2013). Take for example the income in of the experimental subjects. To balance on this variable with the stratification algorithm, one has to define categories on the income, for example small income, middle income and high income. There are two limitations to this algorithm. First, as the number of covariates increases, the number of strata quickly increases to infinity, leading to many strata of size 1.[5] Second, discretizing continuous covariates generally leads to a loss of information (Ma and Hu, 2013).

To overcome these two limitations, multivariate matching, as introduced by Greevy et al. (2004), proposes a more profound way of defining strata. Based on a sample with an even number of n experimental units, this algorithm generates $n/2$ strata with 2 subjects per block.[6] The strata are selected as to minimize the sum of Mahalanobis distances between the subjects of each block. Mahalanobis distance is a popular multivariate measure of distance between covariates (Frazer Lock, 2011; Lock Morgan and Rubin, 2012). This distance measure allows to simultaneously balance on many discrete as well as continuous covariates.

Lastly, re-randomization refers to any algorithm that draws many random treat-

[5] For example in the case of 15 binary covariates, this already makes $2^{15} = 32,768$ different strata.

[6] The matching algorithm targeted in this paper should not be confused with the many matching algorithms used in observational studies, such as for example propensity score matching (Rosenbaum and Rubin, 1983). Those matching algorithms are usually used after individuals have self selected into treatment and control group. The treatment allocation algorithms of this paper should be applied before the treatments are allocated, in cases in which the researcher is able to allocate treatments freely.

ment allocations, and selects one of the draws according to some rule. In this paper, I will regard what Bruhn and McKenzie (2009) refer to as the min-max rule: For each draw, calculate the t-statistics for the difference in covariate means between treatment and control group. This yields m t-statistics for each draw, where m is the number of covariates. Select the draw that minimizes the maximum absolute t-statistic among all draws.

It is intuitive that these algorithms create groups that are balanced concerning the covariates and intuitively, balance on cavariates will foster inference from the experiment. However, the connection between these algorithms and the quality of inference inside a particular model is not precisely clear. Even the word "balance", which is frequently used, has no precise meaning (see also Kasy, 2016a). This paper will target the connection between the treatment allocation and the distribution of the treatment estimator in a linear model. Following Atkinson (2002), I will define covariate balance in terms of the variance on the treatment estimator.

3 Treatment Allocation for Linear Models

This section repeats some well known results from linear regression theory in order to explain the impact of the treatment allocation on the distribution of the treatment estimator in a linear model. The section starts with a definition of the kinds of experiments that this paper targets, then analyzes the role of treatment allocation for the distribution of the treatment estimator, and finishes with the consequences for statistical power and sample sizes.

3.1 Setting

I regard a sample of n individuals drawn from a population of possible subjects[7]. For the data generating process, I assume a linear model:

$$Y = X\beta_x + T\beta_t + \varepsilon \qquad \text{with } \mathbb{E}[\varepsilon] = 0;\ \mathbb{V}[\varepsilon] = I\sigma^2 \tag{1}$$

where $Y \in \mathbb{R}^n$ is the dependent variable, $X = (1, X_1, ..., X_{m+1}) \in \mathbb{R}^{n \times m+1}$ is the covariate matrix, and $T \in \{0,1\}^n$ is the treatment allocation. The covariates are measured prior to allocating the treatments. Each participant can only be allocated either to the treatment group ($T_i = 1$) or to the control group ($T_i = 0$, between subjects design). After determining the treatment allocation, the dependent variable $Y = (Y_1, ..., Y_n) \in \mathbb{R}^n$ is observed. If individual i received the treatment, Y_i is given by $x_i\beta_x + \beta_t + \varepsilon_i$, if not $Y_i = x_i\beta_x + \varepsilon$, where x_i denotes the i-th row of X. The coefficients $\beta_x \in \mathbb{R}^{m+1}$ and $\beta_t \in \mathbb{R}$ are unknown to the researcher and have to be estimated on the basis of Y, X, and T.

Throughout this whole paper, I assume X to have full rank (no perfect collinearity). Further, I denote $\mathcal{T} \subset \{0,1\}^n$ as the set of all treatment allocations T for

[7]This paper regards the sample as given. The role of sampling on experimental inference will not be targeted in this paper.

which the matrix (X, T) has full rank.

I only regard one treatment and one control group and assume no interaction effects between the treatment and the covariates. I discuss some generalizations in appendix D.

3.2 Variance of the Treatment Estimator in a Linear Model

I start with the model to be estimated:

$$Y = X\beta_x + T\beta_t + \varepsilon \qquad \text{with } \mathbb{E}[\varepsilon] = 0;\ \mathbb{V}[\varepsilon] = I\sigma^2 \tag{2}$$

Let $b(T) = \begin{pmatrix} b_x(T) \\ b_t(T) \end{pmatrix} := ((X,T)'(X,T))^{-1}(X,T)'Y$ denote the OLS estimates of the coefficients. It is well known that the estimator $b_t(T)$ is unbiased, meaning $\mathbb{E}[b_t(T)] = \beta_t$ (e.g., Marquardt, 1970). The variance of $b_t(T)$ is given through the following two equivalent representations (e.g. Fox and Monette, 1992; Zuur et al., 2010):

Proposition 3.1. *Let $T \in \mathcal{T}$. Then $\mathbb{V}[b_t(T)]$ has the following two representations:*[8]
(i) $\mathbb{V}[b_t(T)] = \sigma^2(T'M_XT)^{-1}$
(ii) $\mathbb{V}[b_t(T)] = \frac{\sigma^2}{n\cdot\hat{p}_T(1-\hat{p}_T)} \cdot \frac{1}{(1-R^2_{T,X})}$
For $T \in \{0,1\}^n \setminus \mathcal{T}$, I define $\mathbb{V}[b_t(T)] = \infty$.

Here I used the following notations: $M_X := I - X(X'X)^{-1}X'$ is the projection matrix into the orthogonal space of the space spanned by the columns of X. $\hat{p}_T := \frac{\sum_{i=1}^n T_i}{n}$ is the proportion of experimental units allocated to the treatment group. $R^2_{T,X}$ is the R^2 statistics of the OLS regression $T = Y\gamma + \tilde{\varepsilon}$. $\frac{1}{R^2_{T,X}}$ is commonly known as the *variance inflation factor* (Marquardt, 1970; Kutner et al., 2004, p.408).

The first of the two representations is useful for computation, whereas the second is suited for an intuitive explanation concerning the impact of the treatment allocation on the variance of the treatment estimator. The latter representation distinguishes between the different factors that drive the variance of the treatment estimator.

The influence of the sample size n is common knowledge. The higher the sample size, the lower the variance of the treatment estimator. Also the influence of the relative group size $\hat{p}_T$ is frequently targeted (e.g. List et al., 2011). The more equal the group sizes (i.e. the closer $\hat{p}$ is to 0.5), the lower the variance of the treatment estimator.[9]

[8] Note that this proposition does not require the regression errors to be normally distributed. If the errors would be normally distributed, one could however further conclude that also $b_t(T)$ is normally distributed.

[9] Note that equal group sizes are only desirable as long as the variance of the error term is equal in treatment and control group two groups (which I assumed). For a discussion about group sizes in cases of heteroskedasticity see List et al. (2011).

The influence of the variance inflation factor is noted much less frequently in the context of experiments: The lower the linear dependence between the treatment variable and the covariates, the lower the variance of the treatment estimator. Unlike other authors claim (e.g. McClelland, 1997; List et al., 2011; Carneiro et al., 2016), the variance inflation factor is not equal to one when the random variables that induce T and X are independent (for example in case of random treatment allocation). $R^2_{T,X}$ is an empirical figure and it catches (possibly random) empirical correlations between T and X. In fact, the variance inflation factor will be the major driver for the benefits of systematic treatment allocation over random allocation.

Definition 3.2. *(Optimal treatment allocation)*
A treatment allocation $T \in \{0,1\}$ is optimal if and only if it minimizes the variance of the treatment estimator $\mathbb{V}[b_t(T)]$ over all admissible treatment allocations.

This definition leads to the a very similar result than in the Bayesian framework of Kasy (2016b). Whenever at least one covariate is continuous, the set of optimal allocations almost surely contains exactly two elements T_1 and T_2, with $T_1 = 1 - T_2$ (Kasy, 2016b, Theorem 1).

3.3 Statistical Power

Up to now, I pointed out that the treatment allocation can affect covariate balance, as measured by the loss (definition 3.4), as well as the variance of the treatment estimator (proposition 3.1). Now I will target the influence of the the treatment allocation on the probabilities of type 1 and type 2 errors while testing the null hypothesis of no treatment effect ($\beta_t = 0$). For hypothesis testing, I use a t-test on the regression estimate for the treatment effect. This is one of the most common tests within linear models. For this test to be applicable, I will assume the error ε in equation 1 to be normal distributed. If one assumes the errors to be normally distributed, $b_t(T)$ will also be normal distributed for every fixed T (Rawlings et al., 2001, p.88), and the t-statistic will actually be t-distributed for every fixed $T \in \mathcal{T}$ (Rawlings et al., 2001, p. 121):

$$\frac{b_t(T) - \beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}} \sim t_{n-m-2} \tag{3}$$

$\hat{\mathbb{V}}[b_t(T)]$ is the empirical variance of $b_t(T)$, given by $\hat{\sigma}^2(T'M_XT)^{-1}$, where $\hat{\sigma}^2$ is the estimate of σ^2, derived through the residual sum of squares. In order to test the hypothesis $\beta_t = 0$, one has to compare the statistic:

$$\left|\frac{b_t(T)}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}\right|, \tag{4}$$

to the $1 - \frac{\alpha}{2}$ quantile of the t_{n-m-2}-distribution. Under the nulle hypothesis, the probability that the test statistic exceeds the $1 - \frac{\alpha}{2}$ quantile, is obviously exactly

α. Given that $\beta_t \neq 0$, $\frac{b_t(T)}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}$ follows a noncentral t-distribution with noncentrality parameter $\frac{\beta_t}{\sqrt{\mathbb{V}[b_t(T)]}}$, where β_t is the true treatment effect and $\mathbb{V}[b_t(T)]$ is the true variance of the treatment estimator. The implications on the power, i.e. the probability that (4) exceeds the quantile $t_{n-m-2,1-\frac{\alpha}{2}}$, are summarized in the following proposition (see Ghosh (1973) for a proof of the monotonicity):

Proposition 3.3. *Let ε be normal distributed. Further, let α be given. Then the power of the experiment is given by:*

$$\mathbb{P}(|\frac{b_t(T)}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_{d,1-\frac{\alpha}{2}}) = P_\alpha(n-m-2, |\frac{\beta_t}{\sqrt{\mathbb{V}[b_t(T)]}}|), \tag{5}$$

where $P : \mathbb{N} \times \mathbb{R}_+ \to [0,1]$ is monotonously increasing in the both parameters, whenever.

Therefore, under the assumption of normal distributed errors, a treatment allocation that leads to a lower variance of the treatment estimator directly leads to a higher power, given $\beta_t > 0$.

What this section should show, is that the results for the variance and the power hold for any fixed $T \in \mathcal{T}$. No assumption on random allocation of treatments is required.

3.4 Covariate Balance and Sample Sizes

The term covariate balance is frequently used but rarely precisely defined. What most researchers will agree upon is the case of perfect balance. I will define a perfectly balanced experiment as one in which treatment and control group have the same size and the covariate means in treatment and control group are exactly equal. A simple calculation shows that any treatment allocation that would lead to perfect balance, will minimize the variance of the treatment estimator.[10] Of course, depending on the covariate matrix, there does not have to exist a treatment allocation that yields perfect balance. Therefore, the variance under perfect balance, which is given by $\mathbb{V}^* := \frac{4\sigma^2}{n}$, does not necessarily minimize the variance of the treatment estimator $\mathbb{V}[b_t(T)]$ over all $T \in \{0,1\}^n$, but serves as a lower bound. These considerations justify to define covariate balance in terms of the variance of the treatment estimator. As a measure of balance, I will use the loss due to the lack of balance, as defined by Atkinson (2002):

Definition 3.4. *Let $\mathbb{V}^* := \frac{4\sigma^2}{n}$ be the variance of the treatment estimator under perfect balance. Then for a treatment allocation $T \in \{0,1\}^n$ the loss due to the lack of balance is defined by:*

$$\mathcal{L}_n(T) := n(1 - \frac{\mathbb{V}^*}{\mathbb{V}[b_t(T)]}) = n - 4 \cdot T' M_X T \tag{6}$$

[10] Recall proposition 3.1(ii). Note that $R^2_{T,X} \geq 0$ and $R^2_{T,X} = 0$ if and only if the covariate means in treatment and control group are equal. In addition, $\hat{p_T} = 0.5$ maximizes $\hat{p_T}(1-\hat{p_T})$. Proposition 3.1 (ii) thus shows that $\mathbb{V}[b_t(T)]$ is minimized wheneve T yields perfect balance.

The loss is a multivariate measure of balance. It measures imbalances because of unequal group sizes as well as imbalance because of unequal covariate means across treatment and control group. A loss of zero relates to the case of perfect balance, whereas a higher loss indicates higher imbalances. The loss corrects for the fact that random treatment allocation leads to asymptotically balanced groups. As we will see, the average loss for random treatment allocation is more or less constant for different sample sizes. The notion of loss gives rise to two very appealing formulas:

Proposition 3.5. *(Variance)*
Consider a treatment allocation T, with $\mathcal{L}_n(T) = L$. Then:

$$\mathbb{V}[b_t(T)] = \frac{4\sigma^2}{n - L} \tag{7}$$

Proof. follows directly from the definition of the loss □

Proposition 3.6. *(Sample Size)*
Consider a treatment allocation algorithm, which results in T being drawn from a distribution η. Further assume $\mathbb{E}_\eta[\mathcal{L}_n(T)] = L(n)$. Then the sample size necessary to detect a treatment effect of a size of β_t at an alpha level α with a power P solves the following equation:

$$n = S_{\alpha,P,m}(n) + L(n) \tag{8}$$

with $S_{\alpha,P,m}(n) \approx (\frac{2\sigma(t_\alpha + t_P)}{\beta_t})^2$, and $t_\alpha := t_{n-m-2,1-\frac{\alpha}{2}}, t_P := t_{n-m-2,P}$.

Proof. See appendix B.1 □

Proposition 3.6 provides a sample size formula that takes into account the treatment allocation algorithm. The function for the loss has to be determined via simulations. As a rule of thumb, one can take $S_{\alpha,P,m}(n)$ to be constant by replacing the quantiles of the t-distribution by the corresponding quantiles of the normal distribution[11], and keep the expected loss constant by taking $L(n) = L(n^*)$. n^* should be somewhere in the region where one would suspect the necessary sample size to be. The simulation of section 6.1 helps to determine the loss of a particular algorithm.

As another rule of thumb, the loss of random allocation is equal to m (see also Atkinson, 2002), the number of covariates, and the loss for optimal allocation is approximately 0. Therefore, when inference is made with the same model, optimal treatment allocation can reduce the necessary sample size by approximately m. As section 6.2 will show, one can and should control for more covariates when using optimal allocation, then when using random allocation, leading to a further reduction in necessary sample size.

[11] A general rule of thumb is that t-quantiles a fairly close to normal quantiles, whenever the degrees of freedom are larger than 30, i.e. the sample size is larger than 32 plus the number of covariates (Meier et al., 2015, p. 191).

4 Optimal treatment allocation as a generalization of stratified and matched randomization

In this section, I will speak of optimal treatment allocation as an algorithms that randomizes among all minimizers for the variance of the treatment estimator. I will show that optimal treatment allocation is a generalization of the common stratification and matching algorithm.

When using the latter two algorithms, Bruhn and McKenzie (2009) propose to "control for the method of randomization in the analyses"(Bruhn and McKenzie, 2009). This means, if the allocation of treatments was based on k strata, they propose to analyze the data with the following linear model:

$$Y = \beta_0 + \beta_1 block_1 + ... + \beta_{k-1} block_{k-1} + \beta_T T + \varepsilon, \tag{9}$$

where $block_1, ..., block_k$ are dummy variables for the different strata.[12] Now, let us turn this around. As mentioned in the introduction, researchers should choose their model for the analyses of the data *prior* to collecting the data. Suppose the researcher commits to the model of equation 9 *prior* to conducting the experiment. Further, assume that every block contains and even number of experimental units. Then optimal treatment allocation leads to the same allocation rule as stratification:

Corollary 4.1. *Assume the model of equation 9 to be true and every block to contain an even number of experimental units. Then a treatment allocation T is optimal the sense of definition 3.2, if and only if treatment and control group contain an equal number of subjects from each block, i.e. whenever:*

$$\frac{1}{n_j} \sum_{i=1}^{n} block_j^{(i)} T_i = \frac{1}{2}, for\ all\ j = 1, ..., k$$

. where n_j is the number of units in block j.

Proof. See appendix B.2. □

Hence, whenever stratifying or matching leads to perfect balance, any optimal allocation will lead to the same result. So what if at least some strata are of unequal size? For the stratification algorithm, Bruhn and McKenzie (2009) write: "Whenever there is an odd number of units within a stratum, there will be imbalance". This is because, given an odd number of units stratification will allocate the last unit simply randomly. Optimal allocation on the contrary, will allocate the last unit in a way such that the variance of the treatment estimator and thus the imbalance is minimized.

[12] As such, they have to fulfill $\sum_{i=1}^{k} block_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, with $block_i \in \{0,1\}^n$. I removed the $k-th$ block from the model to avoid multicollinearity.

On top of that, what if one assumes a model that is not compatible with stratification or matching? For example, a simple linear model with one continuous covariate. Then, there will never be two units that are exactly equal with respect to their covariate. Hence it is not possible to build strata of units that are exactly equal. In addition, if one assumes, several covariates to be relevant: For example gender and race. Then stratification will automatically build strata for all interactions of the factor levels. There will be a stratum for white males, white females, black males, and so on. What if one assumes, that the interactions (or some of them) are not important?[13] Then the stratification algorithm will build many useless strata, which increases the risk of having many strata with an odd number of units (if the number of strata goes to infinity, strata sizes will mostly be zero or one). In all of these cases, optimal treatment allocation will arrange the allocation of treatments in a way that minimizes the variance of the treatment estimator in the chosen model.

This is best explained in a simple example. Consider the case of four subjects and one covariate on which we assume a linear effect. The subjects have covariate values of 1,2,3 and 4, respectively. In this case, reasonable strata are 1,2; 3,4.[14] Inference from a linear model will be optimized, when the difference in covariate means between treatment and control group is zero (see section 3). Therefore an allocation of $\{1, 4\}$ to one group and $\{2, 3\}$ to the other group would be optimal. When the subjects of each stratum are randomly allocated to treatment and control group, the optimal allocation will only occur in 50% of all cases.[15]

5 Numerical Optimization

In this section, I will present two algorithms to find optimal treatment allocations in practical applications. Recall the relevant optimization problem (definition 3.2):

$$\underset{T\in\{0,1\}^n}{max}\ T'M_XT \tag{10}$$

This is a binary quadratic optimization problem, which is numerically very hard to solve. Brute force solution would require to calculate $T'M_XT$ for 2^n times. Even more sophisticated methods for calculating exact solutions to this problem can usually only be applied to small problems of 100 variables or less (see Kochenberger et al. (2014) for a literature review on solvers for this problem). Much interest in the field of binary quadratic optimization is therefore on heuristics that provide near best solution very quickly. I suggest two very simple heuristics

[13] i.e. when one assumes that the effect of being a white female is just the effect of being white plus the effect of being female, without an additional effect of being a white female.

[14] Mahalanobis Matching on this one covariate will yield exactly these strata. To see this, note that for one covariate, matching with respect to Mahalanobis distance is equal to matching with respect to euclidean distance.

[15] A quick calculation shows that the loss for the allocation $\{1, 4\}$; $\{2, 3\}$ is equal to 0, for $\{1, 3\}$; $\{2, 4\}$ 0.8 and for $\{1, 2\}$; $\{3, 4\}$ 3.2. Therefore, optimal allocation would have an expected loss of 0, matching and stratification of 0.4 (since they only rule out the last allocation) and random allocation 4/3 (since it rules out non of the allocations).

for this problem. For a comparison of those two algorithms to alternative optimization algorithms, see appendix C.

The first is a *local search algorithm*. This algorithm is very simple, and provides reasonably good solutions in a short amount of time. The local search algorithm starts with some (for example random) treatment allocation T, and searches for improvements in the neighborhood of T. The neighborhood of a treatment allocation T is defined by all treatment allocations $\tilde{T}$ that differ from T in exactly one coordinate (i.e. all $\tilde{T} \in \{0,1\}^n$ with $||\tilde{T} - T|| = 1$, where $||\cdot||$ denotes the euclidean distance). The algorithm moves in every step to the neighboring allocation with the highest improvement (i.e. the highest value of $\tilde{T}'M_X\tilde{T} - T'M_XT$). It terminates when there exist no more neighboring allocations that yield any improvement over the current allocation.[16] This algorithm will terminate very quickly. However, it will terminate in every local optimum, i.e. whenever changing the treatment assignment of *one* experimental unit does not lead to any improvement. This does not rule out that there are still improvement possible once one changes the assignment for more than one experimental unit simultaneously.

If the solution of the local search algorithm is not good enough, I suggest to combine this algorithm with re-randomization, which I call *multiple local search algorithm*: Draw k treatment allocations randomly. Apply the local search algorithm to each of them. Take the treatment allocation with the lowest variance of the treatment estimator.

For both of those algorithms, I determine randomly which of the two groups receives the treatment. In particular, if T^* is the solution of one of the above algorithms, I choose $T = T^*$ or $T = (1 - T^*)$ with equal probabilities. Note that $(1 - T)$ leads to the exact same loss as T.

6 Simulations

In this section I provide some simulations, comparing optimal treatment allocation to random treatment allocation as well as to stratification, matching and re-randomization. The first part of this section compares the different algorithms for a constant model. This means, the model for estimating the data and in particular the number of covariates stays the same for all algorithms. The second part of this section compares the different algorithms for a varying number of covariates. In particular, I evaluate how the optimal number of covariates changes depending on the treatment allocation algorithm.

The simulations use the statistical software R (R Development Core Team, 2008). For implementing the matching algorithm, I use the package *nbpMatching* that implements the optimal matching approach of Greevy et al. (2004) (see Lu et al., 2011).

[16] This algorithm is also known as 1-Opt algorithm (Merz and Freisleben, 2002) or Greedy algorithm (Kasy, 2016a).

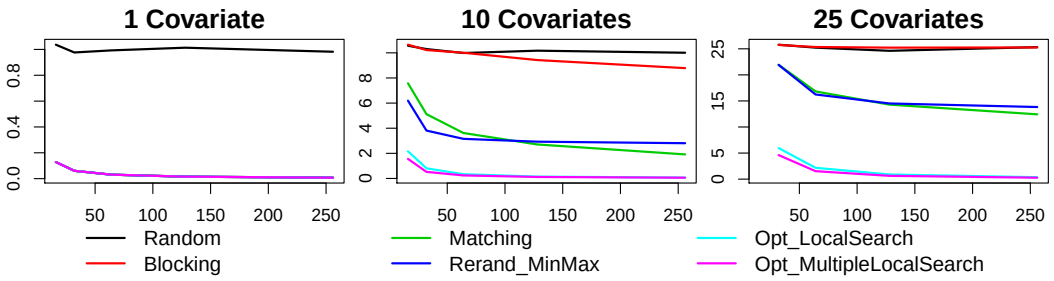
6.1 Fixed Number of Covariates

In this subsection I compare how the different algorithms perform on a given model. I focus on comparing the losses due to the lack of balance of the different algorithms. As propositions 3.5 and 3.6 show, the loss directly translates to the variance of the treatment estimator and the power or rather necessary sample size of the experiment. I simulate the average loss for the case of binary covariates.[17] The results are very robust to different covariate distributions, with the exception that stratification performs significantly worse for continuous covariates because of the discretization of the continuous variables (see appendix A). I simulate the data according to the following model:

$$Y = X\beta_x + T\beta_t + \varepsilon \qquad \text{with } \varepsilon \sim \mathcal{N}(0, I);\ \beta_x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{11}$$

and also base the treatment allocation and the estimation of the treatment effects on this model. I provide simulations for 1, 10 and 25 covariates and for 16 to 256 experimental participants. Each simulation uses $1,000$ Monte-Carlo steps. In every step, I draw a new covariate matrix and allocate treatments according to each of the algorithms based on this matrix. Given this covariate matrix, and the treatment allocation, I calculate the loss according to definition 3.4.[18] The re-randomization algorithm uses 100 redraws and the multiple local search algorithm uses 10 redraws:

Figure 1: Loss for binary covariates



For one covariate, all allocation algorithms, except for random allocation, yield the same loss. This is no surprise since for one binary covariate, treatment allocation is very simple: The experimental units with a covariate of one as well as the units with a covariate of zero have to be allocated equally across treatment and control group.

For more than one binary covariate, we see that especially stratification performs worse. In case of 10 covariates, there are already $2^{10} = 1024$ strata. Consequently, there are many strata with only one subject. Subjects of strata with size one will be allocated randomly. Therefore stratification will only yield low losses, if most

[17] In this paper, I use binary covariates that are equal to one with a probability of 0.5 and zero otherwise.

[18] Note that this procedure does not require explicit simulation of errors.

strata have sizes larger than one. For 10 covariates and 256 participants, we see that stratification works slightly better than random allocation, however for 25 covariates (and consequently 33,554,432 strata) there is no difference between random allocation and stratification anymore.

In the case of 10 and 25 covariates it also becomes apparent that the matching algorithm performs comparably poor, especially for small sample sizes. The reason for this is that matching still includes some degree of randomness. After the matches are made, one randomly selected subject of each match is allocated to the treatment group, the other to the control group. This randomness decreases the performance of the algorithm whenever the matches are not perfect. For larger sample sizes, the matches will get better and thus this problem is less severe.

The re-randomization algorithm performs worse than the local search algorithms for two reasons. First, the goal function, i.e. the maximum t-statistic does not directly relate to the variance of the treatment estimator. Second, re-randomization is not perfectly suited as a means of optimization (see appendix C.2).

Using proposition 3.6, these results on the loss directly translate to necessary sample sizes. For example, take a model with 25 covariates and assume, that the treatment effect is sufficiently strong, such that with random allocation one would need exactly 125 subjects to achieve a power of 0.8. Then with matching or re-randomization, one would only need around 115 subjects and with optimal allocation only around 100 to obtain the same power. In this case, optimal allocation can reduce necessary sample sizes by around 20% compared to random allocation, and around 13% compared to multivariate matching and re-randomization.

While these plots show, how useful systematic and especially optimal tratment allocation is for small scale experiments, they also show that there is little need for systematic allocation whenever the number of covariates is very low compared to the sample size of the experiment. Bruhn and McKenzie (2009) report that out of 18 reviewed experiments in the field of development economics 12 experiments use samples of 200 or less participants. The number of covariates to check balance on ranges from 4 to 39 among these 12 experiments. For these experiments, a systematic allocation of treatments might have been extremely useful. The authors report two other experiments with sample sizes exceeding 1,000 and 12-14 covariates to check balance on. For these experiments, a systematic allocation of treatments might not be necessary. Note however, that additional covariates to control for nonlinearities, also count as covariates. For example, if one has one continuous covariate, but assumes quadratic effects, this makes effectively two covariates.

6.2 Endogenous number of covariates

Up to now, I always assumed the model for estimation to equal the data generating process. This means, I assumed that every observable covariate that influences the dependent variable was controlled for in the regression. In practical applications this will most likely not be the case. In reality, there are often thousands of variables that might influence the dependent variable. Of those variables, only a

few are observed in the context of the experiment and even less are used in the analyses of the experimental data.

Including a variable into the regression only makes sense when the upside from including this variable exceeds the downside from including this variable. Concerning the power of the experiment, most researchers see including an additional variable as a trade-off between the degrees of freedom of the t-distribution and a lower variance of the error term (e.g. Senedecor and Cochran, 1989; Box et al., 2005; Bruhn and McKenzie, 2009; Kahan et al., 2014). However, there is another effect of an additional covariate. As (Duflo et al., 2008, p.3925) note, in a randomized experiment, a new covariate increases the loss due to the lack of balance (see also figure 1).[19] To understand this, suppose one includes a covariate X_i that has a coefficient β_i of zero. Then the estimate b_i for this covariate will not automatically be zero, but catches possible random correlations with the dependent variable. Whenever the treatment variable is not perfectly orthogonal to the covariates (perfect balance), this will lead to a more noisy estimation of the treatment effect.

Since the loss due to the lack of balance differs across treatment allocation algorithms, one might want to control for a different number of covariates, if one uses a different allocation algorithm. In this section we analyze how the optimal number of covariates changes with the allocation algorithm and what influences this has on the overall benefits of these algorithms. This analyses is fairly similar to an analysis by Therneau (1993), who compares the optimal number of covariates for stratification and minimization.[20] For simplicity of the graphic, I only compare random and optimal treatment allocation. Results for stratification, matching and re-randomization would lie somewhere in between these two extremes.

I simulate the data according to the following model:

$$Y = T\beta_t + \underbrace{X\beta_x}_{\text{observable sources of variation}} + \underbrace{\varepsilon}_{\text{unbovservable source of variation}}, \quad \text{with } \varepsilon \sim \mathcal{N}(0,1) \tag{12}$$

Further, I simulate all covariates $X_1, ..., X_m$ normal distributed with mean zero and variance one. The coefficients of the covariates linearly decrease in size:

$$\beta_i = \frac{m-i}{4m}, i = 1, ..., m \tag{13}$$

In the analyses of the data, I only control for the j strongest covariates. Therefore the model for estimation is given by:

$$Y = \beta_0 + \sum_{i=1}^{j} \beta_i X_i + \beta_t T + \tilde{\varepsilon}, \text{with } \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2), \tag{14}$$

[19] These three effects of covariates on the power of the experiment are also apparent in the sample size formula of proposition 3.6

[20] Minimization is a popular algorithms for sequential treatment allocation in medical trials, developed by Taves (1974) and Pocock and Simon (1975). This algorithm should not be confused with the optimal treatment allocation proposed in this paper.

where $\tilde{\varepsilon}$ decomposes to $\sum_{i=j+1}^{m} \beta_i + \varepsilon$

The left graphic in figure 2 shows the variance of the treatment estimator depending on the number of control variables. This figure contains the true variance of the treatment estimator, not the sample estimate thereof. Recall proposition 3.5 to see that there are only two influences of an additional covariate on the true variance of the treatment estimator: First, an additional covariate reduces the variance of the error σ^2, leading to a lower variance of the treatment estimator. Second, an additional covariate can increase the loss due to the lack of balance, leading to a higher variance of the treatment estimator. The green and the blue line are hypothetical cases. This means, there do not have to exist treatment allocations that lead to this particular loss or power. The green line in represents the case of a loss of 0 (i.e., the hypothetical case that all covariates are always perfectly balanced). In the hypothetical case of perfect balance, an additional covariate can only reduce the variance of the treatment estimator. The blue line is a lower bound on the variance of the treatment estimator obtained for a hypothetical allocation with a loss of zero in a model that controls for all 60 covariates. The variance for the local search algorithm (red line), gets very close to the lower bound. However as the number of covariates approach the sample size, there is a mild increase, since the covariate matrices do not allow for perfectly balanced allocations anymore. The variance of the treatment estimator for random allocation (black line) hardly decreases with the number of covariates. At the beginning, the reduction in the error term is slightly higher than the increase in loss. However, as the effect sizes of additional covariates get weaker, the increase in loss dominates. This figure already shows, that optimal treatment allocation is able to retrieve much more information out of the same covariates, than random allocation.

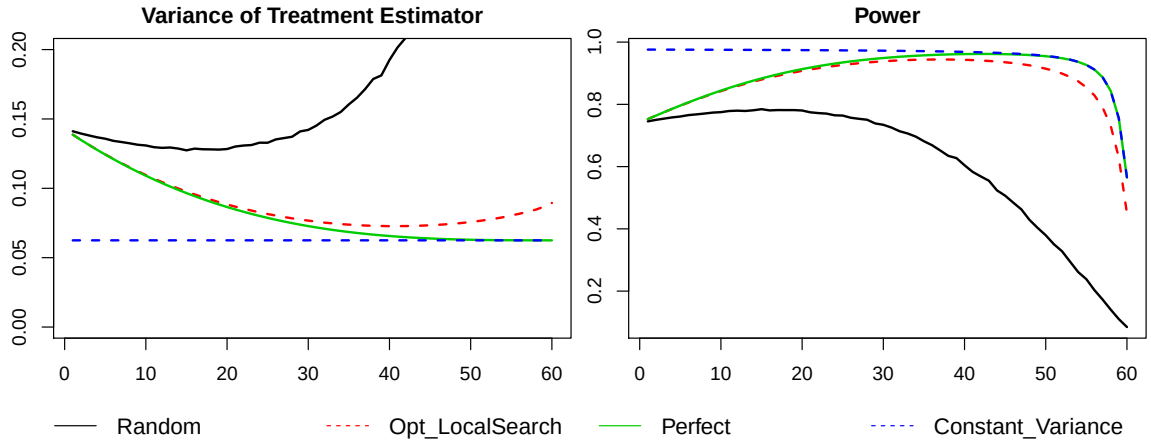


Figure 2: Variance of the treatment estimator and power depending on number of control variables

The right graphic in figure 2 presents the power, i.e. the probability of estimating a significant treatment effect. For random treatment allocation, the power would be maximized if 15 covariates are taken into account. Consequently, a researcher who uses random treatment allocation and aims at maximizing statistical power,

should control for 15 covariates. In case of 15 covariates, the power of the random allocation is 78.2% and the power of the local search algorithm is 0.88%. However, when the researcher uses the local search algorithm, it would be optimal to control for 37 covariates. In this case, the power is 94.4%. This shows, that the comparison of these two algorithms for a fixed number of covariates provides only a lower bound for the difference in statistical power in practical applications.

The blue line in Figure 2 presents the power for the hypothetical case, that the variance of the treatment estimator does not change with the number of control variables. This help to distinguish the importance of the two downsides of adding control variables on the power. The first downside of an additional control variable is an increase in the loss due to the lack of balance, the second is a decrease in the degrees of freedom of the t-distribution. Since we keep the variance of the treatment estimator constant, the only factor that leads the blue line to decrease in the right graphic, is the degrees of freedom. Up to 45 or 50 controls, the blue line decreases only slightly. For more than 50 controls, the line quickly goes to 0. This shows that as long as the number of covariates is not too close the the sample size, the degrees of freedom play only a minor role for the power of the experiment. Intuitively, one would expect a low power out of a regression with 40 covariates and 64 subjects. Figure 1 shows that this is only true for random allocation and the main factor that drives the low power is the loss due to the lack of balance.

In sum, this simulation shows that optimal treatment allocation retrieves much more information from the covariates than random treatment allocation. Even once one controls for covariates that have only weak effects on the dependent variable, the power under optimal allocation might still increase. Generally, when using optimal allocation, one should control for more covariates than when using random allocation.

7 Conclusion

This paper analyses optimal model dependent treatment allocation algorithms for linear models in the case of simultaneous allocation of treatments. Compared to the treatment allocation algorithms currently applied to economic experiments, optimal allocation results in a lower variance of the treatment estimator whenever inference is made via a linear model.

I do not claim that the algorithms of this paper should be used as a standard in every experiment. I agree with Deaton (2010) that there is no experimental design that is superior all others design in every scenario. I rather see these algorithms as part of a toolbox that should be kept in mind when conducting experiments. Especially when there are restrictions on the sample size, these algorithms can help to improve the estimation of the treatment effect and increase statistical power.

I recommend researchers to fix the model used for the analyses of the experimental data *before* the experiment is conducted for two reasons. First, whenever

the model is specified after the experiment was conducted, researchers might (consciously or unconsciously) determine the model in a way that their preferred outcome is supported. Second, when the model is specified before the experiment is conducted, treatment allocation can account for the same assumptions made in the analyses of the data.

If this model happens to be a linear model, one of the algorithms of this paper could be applied. However, the researcher has to decide whether the algorithm provides sufficient benefit over random treatment allocation. When comparing random and optimal model based treatment allocation, one should take into account, that optimal model based allocation allows to control for much more covariates which might increase the power of the experiment.

References

Aickin, M. (2001): "Randomization, balance, and the validity and efficiency of design-adaptive allocation methods," *Journal of Statistical Planning and Inference*, 94, 97 – 119.

Aigner, D. J. (1979): "A brief introduction to the methodology of optimal experimental design," *Journal of Econometrics*, 11, 7 – 26.

Aigner, D. J. and P. Balestra (1988): "Optimal Experimental Design for Error Components Models," *Econometrica*, 56, 955–971.

Athey, S. and G. Imbens (2017): "Chapter 3 - The Econometrics of Randomized Experiments," in *Handbook of Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1 of *Handbook of Economic Field Experiments*, 73 – 140.

Atkinson, A. C. (1982): "Optimum Biased Coin Designs for Sequential Clinical Trials with Prognostic factors," *Biometrika*, 69, 61–67.

——— (2002): "The comparison of designs for sequential clinical trials with covariate information," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165, 349–373.

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007): *Optimum experimental designs, with SAS*, vol. 34 of *Oxford Statistical Science Series*, Oxford University Press.

Banerjee, A., S. Chassang, and E. Snowberg (2016): "Decision Theoretic Approaches to Experiment Design and External Validity," Working Paper 22167, National Bureau of Economic Research.

Beasely, J. E. (1998): "Heuristic algorithms for the unconstrained binary quadratic programming problem," Working paper, London, UK: Management School, Imperial College.

Begg, C. B. and B. Iglewicz (1980): "A Treatment Allocation Procedure for Sequential Clinical Trials," *Biometrics*, 36, 81–90.

Bennmarker, H., E. Gr"onqvist, and B. "Ockert (2013): "Effects of contracting out employment services: Evidence from a randomized experiment," *Journal of Public Economics*, 98, 68 – 84.

Berger, J., D. Dutykh, and N. Mendes (2017): "On the optimal experiment design for heat and moisture parameter estimation," *Experimental Thermal and Fluid Science*, 81, 109 – 122.

Bernstein, D. S. (2009): *Matrix mathematics: theory, facts, and formulas*, Princeton University Press.

Box, G. E., J. S. Hunter, and W. G. Hunter (2005): *Statistics for experimenters: design, innovation, and discovery*, vol. 2, Wiley-Interscience New York.

Bruhn, M. and D. McKenzie (2009): "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal.Applied Economics*, 1, 200–232, copyright - Copyright American Economic Association Oct 2009; Zuletzt aktualisiert - 2011-07-12; SubjectsTermNotLitGenreText - United States–US.

Carneiro, P., S. Lee, and D. Wilhelm (2016): "Optimal Data Collection for Randomized Control Trials," Discussion Paper No. 9908, Institute for the Study of Labour.

Černý, V. (1985): "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm," *Journal of Optimization Theory and Applications*, 45, 41–51.

Das, P., G. Dutta, N. K. Mandal, and B. K. Sinha (2015): *Optimal Covariate Designs - Theory and Applications*, Springer.

Deaton, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48, 424–55.

Deaton, A. and N. Cartwright (2016): "Understanding and Misunderstanding Randomized Controlled Trials," Working Paper 22595, National Bureau of Economic Research.

Duflo, E., R. Glennerster, and M. Kremer (2008): *Using Randomization in Development Economics Research: A Toolkit*, Elsevier, vol. 4 of *Handbook of Development Economics*, chap. 61, 3895–3962.

Elfving, G. (1952): "Optimum Allocation in Linear Regression Theory," *Ann. Math. Statist.*, 23, 255–262.

Fisher, R. A. (1926): "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.

Fox, J. and G. Monette (1992): "Generalized Collinearity Diagnostics," *Journal of the American Statistical Association*, 87, 178–183.

Frazer Lock, K. (2011): "Rerandomization to Improve Covariate Balance in Randomized Experiments," Dissertation, Harvard University, Cambrige, Massachusetts.

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

Ghosh, B. K. (1973): "Some Monotonicity Theorems for χ^2, F and t Distributions with Applications," *Journal of the Royal Statistical Society. Series B (Methodological)*, 35, 480–492.

Glover, F. (1986): "Future paths for integer programming and links to artificial intelligence," *Computers & Operations Research*, 13, 533 – 549, applications of Integer Programming.

Greevy, R., B. Lu, J. H. Silber, and P. Rosenbaum (2004): "Optimal multivariate matching before randomization," *Biostatistics*, 5, 263–275.

Hahn, J., K. Hirano, and D. Karlan (2011): "Adaptive Experimental Design Using the Propensity Score," *Journal of Business & Economic Statistics*, 29, 96–108.

Harville, D. A. (1974): "Nearly Optimal Allocation of Experimental Units Using Observed Covariate Values," *Technometrics*, 16, 589–599.

Horton, J., D. Rand, and R. Zeckhauser (2011): "The online laboratory: conducting experiments in a real labor market," *Experimental Economics*, 14, 399–425.

Howley, R. and R. H. Storer (2014): "Balanced Assignment of Experimental Units in the Analysis of Covariance through Optimization," Have a look.

Kahan, B. C., V. Jairath, C. J. Doré, and T. P. Morris (2014): "The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies," *Trials*, 15, 139.

Kasy, M. (2016a): "Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead," *Political Analysis*.

——— (2016b): "Why Experimenters Should not Randomize and What They Shoud Do Instead," Working paper, https://scholar.harvard.edu/files/kasy/files/experimentaldesign.pdf.

Khinkis, L. A., L. Levasseur, H. Faessel, and W. R. Greco (2003): "Optimal Design for Estimating Parameters of the 4-Parameter Hill Model," *Nonlinearity in Biology, Toxicology, Medicine*, 1.

Kiefer, J. (1959): "Optimum Experimental Designs," *Journal of the Royal Statistical Society. Series B (Methodological)*, 21, 272–319.

Kiefer, J. and J. Wolfowitz (1959): "Optimum Designs in Regression Problems," *Ann. Math. Statist.*, 30, 271–294.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983): "Optimization by Simulated Annealing," *Science*, 220, 671–680.

Kochenberger, G., J.-K. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang (2014): "The unconstrained binary quadratic programming problem: a survey," *Journal of Combinatorial Optimization*, 28, 58–81.

Kutner, M. H., C. Nachtsheim, and J. Neter (2004): *Applied Linear Regression Models*, McGraw-Hill/Irwin, 5 ed.

List, J. A., S. Sadoff, and M. Wagner (2011): "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design," *Experimental Economics*, 14, 439–457.

Lock Morgan, K. and D. B. Rubin (2012): "Rerandomization to improve covariate balance in experiments," *Ann. Statist.*, 40, 1263–1282.

Lu, B., R. Greevy, X. Xu, and C. Beck (2011): "Optimal Nonbipartite Matching and Its Statistical Applications," *The American Statistician*, 65, 21–30.

Lucifora, C. and M. Tonello (2015): "Cheating and social interactions. Evidence from a randomized experiment in a national evaluation program," *Journal of Economic Behavior & Organization*, 115, 45 – 66, behavioral Economics of Education.

Ma, Z. and F. Hu (2013): "Balancing continuous covariates based on Kernel densities," *Contemporary Clinical Trials*, 34, 262 – 269.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987): "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *The American Economic Review*, 77, 251–277.

Marquardt, D. W. (1970): "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, 12, 591–612.

McClelland, G. H. (1997): "Optimal design in psychological research." *Psychological Methods*, 2, 3.

Meier, K., J. Brudney, and J. Bohte (2015): *Applied statistics for public and nonprofit administration*, Cengage Learning, 9 ed.

Merz, P. and B. Freisleben (2002): "Greedy and Local Search Heuristics for Unconstrained Binary Quadratic Programming," *Journal of Heuristics*, 8, 197–213.

Pocock, S. J. and R. Simon (1975): "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics*, 31, 103–115.

Potters, J. and J. Stoop (2016): "Do cheaters in the lab also cheat in the field?" *European Economic Review*, 87, 26 – 33.

Pukelsheim, F. (2006): *Optimal Design of Experiments*, vol. 50 of *Classics in Applied Mathematics*, SIAM.

R Development Core Team (2008): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rawlings, J. O., S. G. Pantula, and D. A. Dickey (2001): *Applied Regression Analysis: A Research Tool*, Springer.

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

Saint-Mont, U. (2015): "Randomization does not help much, comparability does," *PLOS ONE*, 10.

Schneider, S. O. and M. Schlather (2017): "A new approach to treatment assignment for one and multiple treatment groups," Tech. Rep. 228, Courant Research Centre: Poverty, Equity and Growth - Discussion Papers, Göttingen.

Senedecor, G. and W. Cochran (1989): *Statistical Methods*, Iowa State University Press, Ames, Iowa.

Senn, S., V. V. Anisimov, and V. V. Fedorov (2010): "Comparisons of minimization and Atkinson's algorithm," *Statistics in Medicine*, 29, 721–730.

Smith, R. L. (1984): "Properties of Biased Coin Designs in Sequential Clinical Trials," *The Annals of Statistics*, 12, 1018–1034.

Sverdlov, O., ed. (2015): *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*, vol. 81, CRC Press.

Taves, D. R. (1974): "Minimization: A new method of assigning patients to treatment and control groups," *Clinical Pharmacology & Therapeutics*, 15, 443–453.

Telen, D., B. Houska, F. Logist, and J. V. Impe (2016): "Multi-purpose economic optimal experiment design applied to model based optimal control," *Computers & Chemical Engineering*, 94, 212 – 220.

Therneau, T. M. (1993): "How many stratification factors are "too many" to use in a randomization plan?" *Controlled Clinical Trials*, 14, 98 – 108.

Whitehead, J. (1997): *The design and analysis of sequential clinical trials*, Statistics in Practice, John Wiley & Sons, revised second edition ed.

Ziliak, S. T. (2014): "Balanced versus Randomized Field Experiments in Economics: Why W. S. Gosset aka "Student" Matters," *Review of Behavioral Economics*, 1, 167–208.

Zuur, A. F., E. N. Ieno, and C. S. Elphick (2010): "A protocol for data exploration to avoid common statistical problems," *Methods in Ecology and Evolution*, 1, 3–14.

A Figure 1 for Alternative Covariate Distributions

In section 6.1, I simulated the losses due to the lack of balance for different treatment allocation algorithms for binary covariates. In this section, I provide the same simulation for alternative distributions of the covariates. In particular, I regard the following distributions:

- normal: A normal distribution with mean 0 and variance 1. This should be an example of a continuous distribution.
- gamma: Gamma distribution with shape parameter 2 and scale parameter 1. This should be an example of a skewed distribution.
- different: Covariates that follow this distribution are a sum a a uniformly distributed variable on [-10,10] and a second variable that is normal distributed with probability 2/3 and gamma distributed with probability 1/3. This should be an example of a slightly more complex distribution that composes of a continuous and a discrete part.

The simulations for all three covariate distributions show fairly similar results. One difference to the binary case is that stratification performs even worse. The reason is that stratification requires a discretization of continuous covariates. In the simulation, I split the continuous variable at the median. This means, for each covariate, I create a dummy variable that is equal to one, if the value of the continuous variable is above the median, and zero, if the continuous variable is below the median. Of course, this discretization leads to a loss of information.

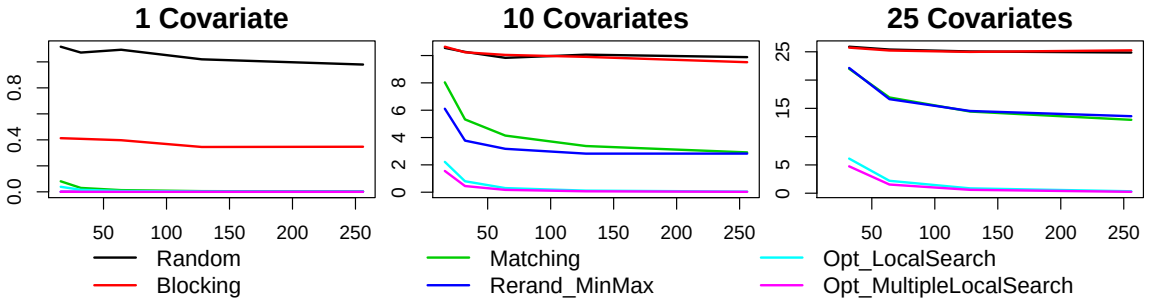


Figure 3: Loss for normal distributed covariates

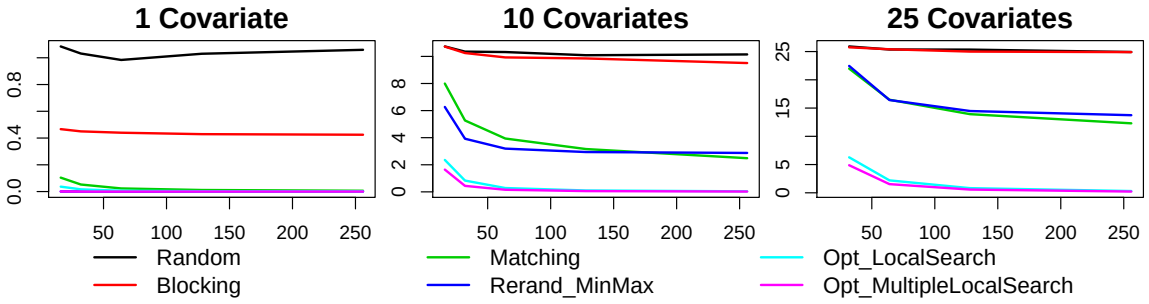


Figure 4: Loss for gamma distributed covariates

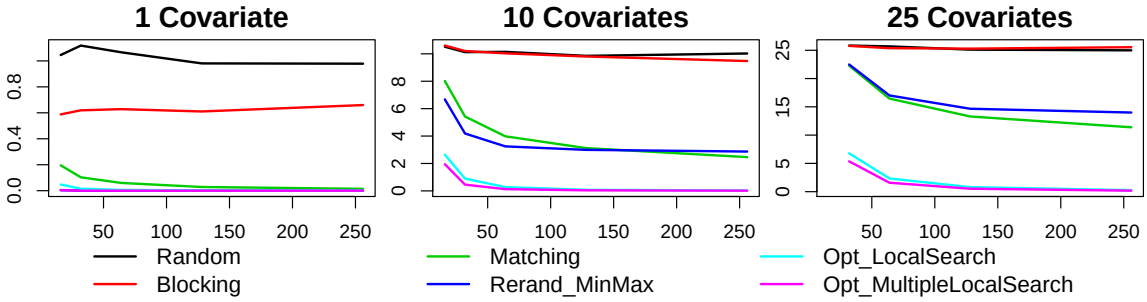


Figure 5: Loss for different distributed covariates

B Proofs for the Paper

B.1 Proof of Proposition 3.6

Proof. Let $d(n) = n - m - 2$ and $\delta(n) = \frac{\beta_t}{\sqrt{\mathbb{V}[b_t(T)]}}$. Further let $P(d, |\delta|) = \mathbb{P}(|\frac{b_t(T)}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_{d,1-\frac{\alpha}{2}})$ be the power function of proposition 3.3. For any fixed $d \in \mathbb{N}$ the range $P(d, \mathbb{R}_+)$ is equal to $[\alpha, 1)$. Thus, since P is monotonously increasing both in d and $|\delta|$, for any $d \in \mathbb{N}$ and $P \in [\alpha, 1)$ there exists a function $g_P(d)$, such that:

$$P(d, |\delta|) = P \Leftrightarrow |\delta| = g_P(d) \tag{15}$$

Writing $g_P(n) := g_P(d(n))$ and plugging the definition of δ into equation 15 yields:

$$|\frac{\beta_t}{\sqrt{\mathbb{V}[b_t(T)]}}| = g_P(n) \tag{16}$$

By proposition 3.5:

$$\Leftrightarrow \frac{\beta_t^2(n - L(n))}{4\sigma^2} = g_P(n)^2 \tag{17}$$

$$\Leftrightarrow n = \underbrace{\frac{4\sigma^2 g_P(n)^2}{\beta_t^2}}_{S_{\alpha,P,m}(n)} - L(n) \tag{18}$$

It remains to show the approximation $g_P(n) \approx t_{n-m-2,P} + t_{n-m-2,1-\frac{\alpha}{2}}$. Let $t_\alpha := t_{n-m-2,1-\frac{\alpha}{2}}$, $t_P := t_{n-m-2,P}$. I start by approximating the power function:

$$P(d, |\delta|) = \mathbb{P}(|\frac{b_t(T) - \beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(t)]}} + \frac{\beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_\alpha) = \mathbb{P}(|X + \frac{\beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_\alpha) \tag{19}$$

where X follows a central t-distribution with $n - m - 2$ degrees of freedom. I follow List et al. (2011), using two simplifications. The first approximation is to

replace $\hat{\mathbb{V}}$ by its mean $\mathbb{V}$ and thus $\frac{\beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}$ by δ:

$$P(d, |\delta|) \approx \mathbb{P}(|X + \delta| > t_\alpha) \tag{20}$$
$$= \mathbb{P}(X < -t_\alpha + \delta) + \mathbb{P}(X < -t_\alpha - \delta) \tag{21}$$

The second approximation is to neglect the smaller of the two probabilities. The error of this approximation has to be smaller than $\frac{\alpha}{2}$ and will probably be much smaller if P is large (which one would typically assume). This yields the approximation:

$$P(d, |\delta|) \approx \mathbb{P}(X < -t_\alpha + |\delta|) \tag{22}$$

Next, I invert this function to get an approximation $\tilde{g}_P(n)$:

$$\mathbb{P}(X < -t_\alpha + |\delta|) = P \Leftrightarrow -t_\alpha + |\delta| = t_P \Leftrightarrow |\delta| = t_\alpha + t_P \tag{23}$$

Consequently: $\tilde{g}_P(n) = t_\alpha + t_P$ □

B.2 Proof of corollary 4.1

Proof. Let $X := (\mathbf{1}, block_1, ..., block_{k-1})$, with $\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$.

⇐: By proposition 3.1(i): $\mathbb{V}[b_t(T)] = \sigma^2(T' M_X T)^{-1}$, with $M_X = Id - X(X'X)^{-1}X'$. Equal allocation of units from each stratum to treatment and control group implies: $X'T = \frac{1}{2}\begin{pmatrix} n \\ n_1 \\ \vdots \\ n_{k-1} \end{pmatrix} = \frac{1}{2}X'\mathbf{1} = \frac{1}{2}X'Xe_1$, where $e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ is the first unit vector in $\mathbb{R}^n$. Thus

$$T' M_X T = T' - \frac{1}{4} e_1' X'X(X'X)^{-1}X'Xe_1 \tag{24}$$
$$= \frac{n}{2} - \frac{1}{4} e_1' X'Xe_1 \tag{25}$$
$$= \frac{n}{2} - \frac{1}{4}\mathbf{1}'\mathbf{1} \tag{26}$$
$$= \frac{n}{2} - \frac{n}{4} = \frac{n}{4} \tag{27}$$

This shows that $\mathbb{V}[b_t(T)] = \frac{4\sigma^2}{n}$, which is a lower bound on the variance of the treatment estimator and thus a minimum.

⇒ Let $T^* \in \{0,1\}^n$ be a treatment allocation with $\mathbb{V}[b_t(T^*)] = \frac{4\sigma^2}{n}$. By proposition 3.1(ii): $\mathbb{V}[b_t(T)] = \frac{\sigma^2}{n \cdot \hat{p}_T(1-\hat{p}_T)} \cdot \frac{1}{(1-R^2_{T,X})}$. Note that $\hat{p}_T \cdot (1 - \hat{p}_T) \leq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ and $R^2_{T,X} \geq 0$ for all $T \in \{0,1\}$. Therefore, $\mathbb{V}[b_t(T^*)] = \frac{4\sigma^2}{n}$ implies $\hat{p}_T = \frac{1}{2}$ and

$R^2_{T,X} = 0$. Let $||\cdot||$ be the euclidean norm, then:

$$
\begin{aligned}
R^2_{T,X} = 0 & \tag{28}\\
\Leftrightarrow \frac{||X(X'X)^{-1}X'T - \hat{p}_T \cdot \mathbf{1}||^2}{||T - \hat{p}_T \cdot \mathbf{1}||^2} = 0 & \tag{29}\\
\Rightarrow ||X(X'X)^{-1}X'T - \hat{p}_T \cdot \mathbf{1}||^2 = 0 & \tag{30}\\
\Rightarrow X(X'X)^{-1}X'T = \frac{1}{2} \cdot \mathbf{1} & \tag{31}\\
\Rightarrow X'T = \frac{1}{2} \cdot X'\mathbf{1} & \tag{32}\\
\Rightarrow \frac{1}{n_j} \sum_{i=1}^{n} block_j^{(i)} T_i = \frac{1}{2}, \text{for all } j = 1, ..., k & \tag{33}
\end{aligned}
$$

□

C Comparison of Optimization Algorithms for Finding Optimal Allocations

Section 5 suggests to use a simple local search algorithm or the multiple local search algorithm with random starting points for finding optimal treatment allocations. This section should justify this suggestion. In the first part of this section (part C.1), I compare the local search to some more profound algorithms. In the second part (part C.2), I compare the local search to an optimization via re-randomization. The third part (part C.3) provides the pseudo code for all algorithms. The algorithms presented in this section aim at maximizing the goal function $T'M_XT$ over all $T \in \{0,1\}^n$. I compare the performance of the algorithms with respect to the loss due to the lack of balance, which is a monotonously decreasing transformation of the goal function (see section 3.4).

C.1 Local Search vs. Alternative Optimization Algorithms

Since exact methods for binary optimization generally work only on small problems (up to around 100 variables), I focus on heuristic methods. After all, each subjects in the experiment represents a new variable for the optimization. I regard three very popular algorithms:

1. A Randomized Greedy Algorithm (Merz and Freisleben, 2002):
   The idea of this algorithm is simple: Start with a vector $\tilde{T} = (0.5, ...0.5)'$, and sequentially set coordinates to either 0 or 1, such that in each step the improvement, i.e. the increase in the goal function, is maximized. To preserve some randomness, a random draw determines which coordinate is first and whether this coordinate should be set to 0 or 1. After that, the algorithm calculates among all coordinates that have still a value of 0.5 the coordinate with the highest improvement from changing it's value to 1 and the coordinate with the highest improvement from changing it's value to 0.

Then with a specific chance proportional the the size of the improvement, the first of the two coordinates is set to one, and otherwise the second coordinate is set to zero. This procedure continues until the final vector T consist only of zeros and ones.

2. A Tabu Search Algorithm (Glover, 1986; Beasely, 1998):
   This Algorithm works similar to the simple local search algorithm, with one difference: Whenever the algorithm is stuck in a local maximum, i.e. no neighboring allocation yields any improvement, the algorithm moves to the neighboring allocation with the lowest deterioration. In order avoid moving back right away, the algorithm blocks the coordinate along with the last move was made for a predefined number of steps. Since this algorithm will not terminate by itself, we need to specify a maximum number of iterations depending on the acceptable computing time of the algorithm. In the end, the point with the highest goal function is selected.

3. A Simulated Annealing Algorithm (Kirkpatrick et al., 1983; Černý, 1985; Beasely, 1998):
   This algorithm also works similar to the local search algorithm. However, in contrast to to simple local search algorithm, this method randomly selects exactly one neighboring allocation in each step. If this neighbor yields an improvement, the algorithm moves to this allocation. If the neighbor yields a deterioration, the move might still be made with a certain probability. This probability decreases both with the size of the deterioration and in the course of the algorithm. The algorithm terminates when a predefined number of iterations is reached.

Most modern heuristics for binary quadratic optimization are based on these three methods (see Kochenberger et al., 2014). The randomized greedy algorithm is often used to receive starting points for other algorithms. The tabu search and simulated annealing algorithm improve on the simple local search algorithm by avoiding to get stuck in local optima. In total, I compare six different algorithms: randomized greedy, tabu search, simulated annealing, basic local search, multiple local search with 10 random starting points (Opt_MLSR) and multiple local search with 10 randomized greedy starting points (Opt_MLSG). To give some bounds on the performance of these algorithms, I include random allocation as a lower bound on the performance, and a multiple local search algorithm with $1,000,000$ random starting points (Opt_MLSM) as an upper bound. These are much more redraws than in any reasonable experiment in practice, since this algorithm takes up to 2.5 hours to compute the allocation of a single covariate matrix. However, it should show how low the loss due to the lack of balance could be.

Table 1 shows a simulation for 1, 4, 10, 25 and 50 covariates and a sample size of 64. The values without parenthesis are the average losses for this algorithm, whereas the value in parenthesis are the average computing times (in s) for one allocation. In terms of computation time, the local search algorithm is much faster than any other algorithm, except for random allocation. In terms of minimizing

the loss,[21] the local search algorithm performs better than the greedy algorithm and only slightly worse than the more profound tabu search and simulated annealing algorithms. When using multiple random starting points (MLSR), the local search algorithm even leads to a lower loss than tabu search or simulated annealing, while still requiring less computation time. Randomized greedy starting points in the multiple local search algorithm (MLSG) do not improve much over random starting points and require more computation time.

Table 1: Average loss due to the lack of balance for binary optimization algorithms

	1 Covariate	4 Covariates	10 Covariates	25 Covariates	50 Covariates
Random	1.02	4.11	10.24	25.53	50.52
	(0)	(0)	(0)	(0)	(0)
Opt_Greedy	0.39	0.57	1.12	3.44	12.5
	(0.026)	(0.027)	(0.034)	(0.03)	(0.026)
Opt_LocalSearch	0.01	0.04	0.29	2.22	11.03
	(0.001)	(0.003)	(0.005)	(0.008)	(0.01)
Opt_TabuSearch	0	0.02	0.24	1.95	10.52
	(0.151)	(0.142)	(0.14)	(0.129)	(0.131)
Opt_Annealing	0	0.02	0.23	1.95	10.37
	(0.191)	(0.194)	(0.186)	(0.161)	(0.143)
Opt_MLSR	0	0.01	0.16	1.53	8.84
	(0.018)	(0.036)	(0.047)	(0.062)	(0.087)
Opt_MLSG	0	0.01	0.16	1.55	8.72
	(0.292)	(0.306)	(0.285)	(0.284)	(0.234)
Opt_MLSM	0	0	0.02	0.55	7.45
	(1372.679)	(2447.998)	(3986.627)	(6159.142)	(9622.744)

C.2 Local Search vs. Re-randomization

For a similar optimization problem, Kasy (2016a) suggests to use a re-randomization algorithm. The algorithm is very simple: Draw a predefined number of random allocations and pick the one with the highest value of the goal function. He argues that this procedure performs "reasonably well". The argument, also picked up by Banerjee et al. (2016), is the following: Suppose one re-randomizes for $k \in \mathbb{N}$ times. Then the probability the the chosen allocation is better than 99% of all allocations is $1 - 0.99^k$, which quickly converges to one as k goes to infinity. For $k = 500$ the probability is already larger than 99%.

However, what if the distribution of the loss due to the lack of balance has long but thin tails? In this case, an allocation that is better than 99% of all allocations might still be not a very good allocation. For example in the case of 64 subjects, there are $2^{64} \approx 2 \cdot 10^{19}$ possible allocations. Therefore, there are still around $2 \cdot 10^{17}$ allocations that are among the 1% of best allocations. These are $2 \cdot 10^{17}$ allocations that are potentially better than the allocation determined by

[21] Or equivalently, maximizing $T' M_X T$.

re-randomization.

To analyze the question whether re-randomization could be used as a simple alternative to the local search algorithm, I regard the density of the loss due to the lack of balance for random allocation. I simulate the density using 1,000,000 random allocations, for a sample size of 64 and for 10 as well as 50 covariates. As a benchmark, I include the average loss of the local search algorithm, as well as the minimum loss over $1,000,000$ local search algorithms (MLSM).

Figure 6 shows that for 10 covariates, the 1% quantile is already very close to 0. For this case, re-randomization might be an alternative to the local search algorithm. However, for 50 covariates, the one percent quantile is only slightly better than an average random allocation. Even though the loss could be reduced to less than 10, the one percent quantile is only slightly lower than 40. Even after 1,000,000 random allocations, the best allocation still yields a loss of 20. To calculate losses for 1,000,000 random allocations and 50 covariates, the computer needs around 30 min. The local search algorithm leads to a loss of half the size in only 10 milliseconds.

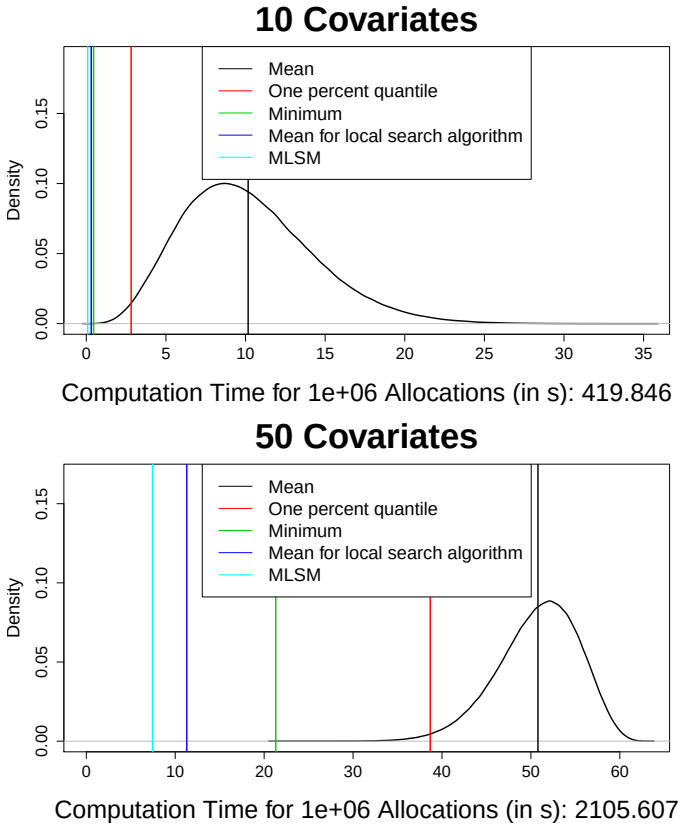


Figure 6: Distribution of loss for random allocation

C.3 Pseudo code of the optimization algorithms

Opt_Rerandomization:
Variables:
Retries % Number of Redraws

1. Draw *Retries* random treatment allocations and store them in list $\hat{T}$.
2. Calculate $T'M_XT$ for every $T \in \hat{T}$.
3. Return T^* with $T^{*\prime}M_XT^* = \underset{T\in\hat{T}}{min}\, T'M_XT$

Opt_Greedy:
Variables:
$C = \{1,...,n\}$ % indices of subjects

1. Set $T = (0.5, ..., 0.5)'$
2. For $l = 0, 1$, set $\tilde{T} = T$ and $\tilde{T}_i = l$ and compute $g_i^l = \tilde{T}'M_X\tilde{T} - T'M_XT$
3. Set $k_0 = argmax_{i\in C} g_i^0$ and $k_1 = argmax_{i\in C} g_i^1$
4. With probability $\frac{g_{k_0}^0}{g_{k_0}^0 + g_{k_1}^1}$ do
   Set $T_{k_0} = 0$ and $C = C \setminus \{k_0\}$
   else
   Set $T_{k_1} = 1$ and $C = C \setminus \{k_1\}$
5. If $C \neq \phi$
   continue with step 2.
   else
   Return T

Opt_LocalSearch:
Variables:
Start % Treatment allocation to start from (for example a random allocation)
t = 0 % iteration counter

1. Set $T = Start$ and $V = T'M_XT$.
2. Set $t = t + 1$; Store each neighbor of T in a list $\hat{T}$.[22]
3. Calculate $\tilde{T}'M_X\tilde{T}$ for every $\tilde{T} \in \hat{T}$.
4. If $\max_{\tilde{T}\in\hat{T}} \tilde{T}'M_X\tilde{T} > V$:
   Set $T = \underset{\tilde{T}\in\hat{T}}{argmax}\ \tilde{T}'M_X\tilde{T}$ and $V = \max_{\tilde{T}\in\hat{T}} \tilde{T}'M_X\tilde{T}$;
   Continue with step 2.

   Else:
   Return T; t

Note: Let $\tilde{T}$ differ from T only in the coordinate i. Then $\tilde{T}'M_X\tilde{T} = T'M_XT + (\tilde{T}_i - T_i) \cdot (M_{X\,i,i} + 2\sum_{j=1,j\neq i} M_{X\,i,j})$. I use this formula in the implementation of this algorithm to efficiently calculate $\tilde{T}'M_X\tilde{T}$ for neighbors of T.

Opt_MultipleLocalSearch:
Variables:
Retries % Number of Redraws

1. Draw *Retries* treatment allocations either randomly or with the randomized greedy algorithm and store them in list $\hat{T}$.
2. Use **Opt_LocalSearch** with *Start* parameter T for each $T \in \hat{T}$. Store resulting allocations in list $\hat{\mathcal{T}}$
3. Calculate $T'M_XT$ for every $T \in \hat{\mathcal{T}}$.
4. Return T^* with $T^{*\prime}M_XT^* = \min_{T\in\hat{\mathcal{T}}} T'M_XT$

[22] A neighbor of a treatment allocation T is defined as a treatment allocation $\tilde{T} \in \{0,1\}^n$ with $||\tilde{T} - T|| = 1$, where $||\cdot||$ denotes the euclidean distance. This means a neighbor $\tilde{T}$ differs form T in exactly one coordinate.

Opt_TabuSearch:
Variables:
Start % Treatment allocation to start from (for example a random allocation)
maxiter % Maximum Number of iterations (In the implementation, I use $max(200, 20000/n)$
T^* % best allocation found so far
$V^* = 0$ % $T^{*\prime} M_X T^*$
$L = (L_1 = 0, ..., L_n = 0)$ % The tabu value of coordinate i L^* % The tabu tenure. Determine by how much the tabu value L_i is increased if a move along coordinate i is made. (In the implementation, I use $L^* = min(10, n/8)$)
t % iteration counter

1. Set $T =$ **Start** and $V = T' M_X T$.

2. Set $t = t + 1$

3. Let $T^{(i)}$ be the neighbor that differs from T in coordinate i. Calculate ${T^{(i)}}' M_X T^{(i)}$ for every $i \in \{1, ..., n\}$ with $L_i = 0$.

4. If $\max\limits_{i \in \{1,...,n\}, L_i = 0} {T^{(i)}}' M_X T^{(i)} > V^*$:
   Set $j = \operatorname{argmax}\limits_{i \in \{1,...,n\}, L_i = 0} {T^{(i)}}' M_X T^{(i)}$
   Apply **Opt_LocalSearch** for $Start = T^{(j)}$ and $t = t$
   Set $T =$ **Opt_LocalSearch**.T and $t =$ **Opt_LocalSearch**.t
   Set $V = T' M_X T$
   Set $T^* = T$ and $V^* = V$
   Else: Set $j = \operatorname{argmax}\limits_{i \in \{1,...,n\}, L_i = 0} {T^{(i)}}' M_X T^{(i)}$
   Set $T = T^{(j)}$ and $V = T' M_X T$;

5. Reduce the tabu values: $L_i = max(L_i - 1, 0)$ for every $i = 1, ..., n$
   Set the tabu value for the most recent move: $L_j = L^*$

6. If $t <$ *maxiter*
   Continue with step 2
   Else
   Return T^*

Opt_Annealing:
Variables:
Start % Treatment allocation to start from (for example a random allocation)
maxiter % Maximum Number of iterations (In the implementation, I use $max(1000000, 10000 * n)$
T^* % best allocation found so far
$V^* = 0$ % $T^{*\prime} M_X T^*$
temperature % the value of the temperature variable determines the probability that a sub-optimal allocation will be accepted. (In the implementation, I use $temperature = n$)
α % determines how far the temperature reduces in every iteration. (In the implementation, I use $\alpha = 0.995$)
t % iteration counter

1. Set $T =$ **Start** and $V = T' M_X T$.
2. Set $t = t + 1$
3. Determine $j \in \{1, ..., n\}$ randomly. Let $T^{(j)}$ be the treatment allocation that differs from T only in coordinate j.
4. Calculate ${T^{(j)}}' M_X T^{(j)}$
5. If ${T^{(j)}}' M_X T^{(j)} > V^*$
   Set $T = T^{(j)}$ and $V = {T^{(j)}}' M_X T^{(j)}$
   Set $T^* = T^{(j)}$ and $V^* = {T^{(j)}}' M_X T^{(j)}$
   Else:
   If ${T^{(j)}}' M_X T^{(j)} > V$
   Set $T = T^{(j)}$ and $V = {T^{(j)}}' M_X T^{(j)}$
   Else:
   With a probability of $exp(-\frac{|V - {T^{(j)}}' M_X T^{(j)}|}{temperature})$:
   Set $T = T^{(j)}$ and $V = T' M_X T$ % Move to new allocation even though it is worse than the old one
6. If $t < maxiter$
   Continue with step 2
   Else
   Apply **Opt_LocalSearch** for $Start = T^*$ and $t = t$
   Set $T^* =$ **Opt_LocalSearch**.T
   Return T^*

D Multiple Treatments and/or interaction effects

In this section, I extend the analyses on linear models involving interaction effects and on experiments including multiple treatments. I consider the following model:

$$Y = X\beta_x + H(X,T)\beta_h + T\beta_t + \varepsilon, \quad \text{with } \varepsilon \sim \mathcal{N}(0, I\sigma^2), \tag{34}$$

$X = \begin{pmatrix} 1 & X_{1,1} & ... & X_{1,m} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & ... & X_{n,m} \end{pmatrix}$ denotes the covariate matrix;

The matrix $T = \begin{pmatrix} T_{1,1} & ... & T_{1,k} \\ \vdots & \vdots & \vdots \\ T_{n,1} & ... & T_{n,k} \end{pmatrix}$ describes the allocation of treatments. $T_{i,j} = 1$ means that unit i receives treatment j. Whenever $T_{i,j} = 0$ for all $j \in \{1, ..., k\}$, the unit is assigned to the control group.

$H = \begin{pmatrix} h_1(X_{1,\cdot}, T_{1,\cdot}) & ... & h_l(X_{1,\cdot}, T_{1,\cdot}) \\ \vdots & \vdots & \vdots \\ h_1(X_{n,\cdot}, T_{n,\cdot}) & ... & h_l(X_{n,\cdot}, T_{n,\cdot}) \end{pmatrix}$ is the matrix of interaction effects. $h_1, .., h_l$ are (possibly nonlinear) functions of the covariates and the treatments. One example is a simple linear interaction effect $h(X_{i,\cdot}, T_{i,\cdot}) = X_{i,j_1} \cdot T_{i,j_2}$, with $j_1 \in \{1, ..., m\}, j_2 \in \{1, ..., k\}$

Let $C = (X, H, T)$, then the OLS estimates of β_x, β_h and β_t are given by $b = \begin{pmatrix} b_x \\ b_h \\ b_t \end{pmatrix} = (C'C)^{-1}C'Y$.

In a next step, the researcher has to decide, which effects are most important. In many experiments this will be the estimators of all treatment effects β_t, but maybe the researcher is also interested in some of the interaction effects. I denote the effect that are most important to the researcher *major effects*[23], and all other effects *minor effects*. Let $\beta_z = \begin{pmatrix} \beta_{z,1} \\ \vdots \\ \beta_{z,\tilde{m}} \end{pmatrix}$ be vector of major effects. Further let Z be the columns of C that correspond to these major effects and N be the columns that correspond to the remaining minor effects. Up to a permutation of columns, $C = (N, Z)$.

The variance-covariance matrix of all estimators b is given by $\mathbb{V}[b] = \sigma^2(C'C)^{-1} = \sigma^2 \begin{pmatrix} N'N & N'Z \\ Z'N & Z'Z \end{pmatrix}^{-1}$. The variance-covariance matrix of the estimators for the major effect b_z is the lower right $\tilde{k} \times \tilde{k}$ sub matrix of $\mathbb{V}[b]$. Using an inversion

[23] In most applications the major effect will simply be all treatment effects. However, researcher might also be interested in some of the interaction effects, or they are only interested in a selection of the treatment effects

formula for block matrices,[24] this matrix is given by:

$$\mathbb{V}[b_z] = \sigma^2 (Z' M_N Z)^{-1} \tag{35}$$

Similar to the case of one treatment and no interaction effects, the goal is thus to maximize:

$$Z' M_N Z \tag{36}$$

Note that in general N as well as Z depend on the allocation of treatments T. Whenever the number of major effects $\tilde{m}$ is equal to one, this matrix reduces to a scalar, which can be maximized by the same binary optimization techniques presented in the paper. Whenever $\tilde{m} > 1$ this is however a matrix and maximization is not clearly defined. In order to define a goal function for optimization, the researcher therefore needs to specify a function g that maps the matrix $\mathbb{V}[b_z]$ to a real number.

In the field of optimal experimental design, popular functions for g are:

1. The determinant: $g(\mathbb{V}[b_z]) = det(\mathbb{V}[b_z])$. Treatment allocation that minimize $det(\mathbb{V}[b_z])$ are called *D-optimal* treatment allocations. D-optimal treatment allocations minimize the volume of the confidence region for b_z (Khinkis et al., 2003).

2. The trace: $g(\mathbb{V}[b_z]) = tr(\mathbb{V}[b_z])$. Treatment allocations that minimize $\mathbb{V}[b_z]$ are called *A-optimal* treatment allocations. A-optimal treatment allocations minimize the average variance of the estimators of the major effects. Schneider and Schlather (2017) propose to use a weighted average, i.e. $g(\mathbb{V}[b_z]) = tr(\mathbb{V}[b_z] diag(w))$, with $w = (w_1, ..., w_{\tilde{m}})$ being weights defining which effects are of most interest.

3. The maximum eigenvalue: $g(\mathbb{V}[b_z]) = \lambda_{max}(\mathbb{V}[b_z])$. Treatment allocations that minimize $\lambda_{max}(\mathbb{V}[b_z])$ are called *E-optimal* treatment allocations. E-optimal treatment allocations minimize the worst possible variance of all linear combinations of the major effects (Pukelsheim, 2006, chapter 6.4).

For more information regarding statistical properties and intuitions behind these functions and their spread in the field of experimental design, see Pukelsheim (2006, chapter 6).

Having defined the model and the major effects, the function $\gamma = g(Z' M_N Z)$ is a mapping from the set of possible covariate matrices $\mathcal{X}$ and the set of admissible treatment allocations $\mathcal{T}$ to the real numbers: $\gamma : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$. Given the covariate matrix X, $\gamma(X, T)$ solely depends on T and can be optimized according to the binary optimization techniques presented in this paper.

[24] The inversion formula yields
$\begin{pmatrix} A & B \\ B' & C \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(C - B'A^{-1}B)^{-1}B'A^{-1} & -A^{-1}B(C - B'A^{-1}B)^{-1} \\ -(C - B'A^{-1}B)^{-1}B'A^{-1} & (C - B'A^{-1}B)^{-1} \end{pmatrix}$
for a regular block matrix $\begin{pmatrix} A & B \\ B' & C \end{pmatrix}$ (Bernstein, 2009).