Discussion Papers
in Economics

No. 14/2017

Treatment Allocation for Linear Models

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

Tobias Aufenanger *

Friedrich-Alexander University Erlangen-Nürnberg (FAU)

This version: September 2017

Abstract

This paper analyzes the benefits and the limits of a systematic allocation of treatments within a linear model framework. Linear models do not necessarily require the treatment allocation to be random. Since the variance of the treatment estimator within linear models does not depend on the realization of the dependent variable, whenever the covariate information is available prior to allocating treatments it is possible to allocate treatments in a way that minimizes the variance of the treatment estimator. I show that in each experiment satisfying the linear model assumptions, there exists at least one deterministic optimal design, i.e., a deterministic way of allocating treatments that minimizes the variance of the treatment estimator over all alternative ways of allocating treatments. In finite samples, optimal design reduces the variance of the treatment estimator and increases statistical power compared to random allocation. For a given linear model and a given effect size, optimal design decreases the sample size necessary to detect a significant treatment effect on average by the number of covariates in the model. However, asymptotically, as the sample size goes to infinity, neither optimal design nor any alternative design yields any benefit over random 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. A 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. I will define the term *experimental design* as the strategy for allocating treatments.[2]

Ever since Fisher (1926) introduced formal statistical methods to experimentation, there is an ongoing debate in the statistical literature on how to ideally use the ability of allocating treatments for inference (Student, 1938; Ziliak, 2014; Bertsimas et al., 2015). In recent years, economic literature has taken on this debate (Bruhn and McKenzie, 2009; Hahn et al., 2011; Horton et al., 2011; Deaton and Cartwright, 2016; Kasy, 2016a; Banerjee et al., 2017; Athey and Imbens, 2017; Schneider and Schlather, 2017). When deciding how to allocate treatments, it is important to consider the method of inference. Randomization inference assumes the treatment allocation to be random and the dependent variable given the treatment allocation to be fixed (Kempthorne, 1955; Rubin, 1974; Holland, 1986). Consequently, this method of inference requires experimental designs that involve a certain degree of randomness. Model-based inference on the contrary, assumes the treatment allocation to be fixed and the dependent variable given the treatment to be random (Freedman, 2008; Athey and Imbens, 2017). Therefore, there is no particular need to allocate treatments randomly in this setting. Although this insight is not particularly new to the statistical literature (e.g., Aickin, 2001), there appears to be a mismatch between theory and economic experiments in practice. Whereas the econometric and related statistical literature mainly advocates randomization inference (Rosenbaum, 2002; Imai et al., 2008; Imbens and Wooldridge, 2009; Lock Morgan and Rubin, 2012; Imbens and Rubin, 2015; Schneider and Schlather, 2017; Athey and Imbens, 2017), many researchers nevertheless analyze experiments via model-based inference (Glewwe et al., 2009; Krawczyk and Smyk, 2016; Abdulkadiroğlu et al., forthcoming). Consequently, model-based inference is widely used for analyzing experiments, but there exists little statistical guidance on the role of the treatment allocation within this framework.

This paper aims at analyzing the role of treatment allocation within a linear model framework. Linear models are among the most frequently used regression models for experimental data (Bruhn and McKenzie, 2009). Inside the model framework, the variance of the treatment estimator is proportional to a function of the covariate matrix and the treatment allocation. Consequently, for a given sample with known covariates, the experimental design can minimize the vari-

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

[2] See Section 3 for a more formal definition.

ance of the treatment estimator. I refer to this type of design as *optimal design*. Optimal design aims at minimizing the linear dependency between the allocation and the covariates, i.e., experimental units with similar covariate values should be put in different groups. However, even though optimal design clearly results in a lower variance of the treatment estimator than random allocation, in practice the choice of the allocation algorithm seems to be less clear. On the one hand, most researchers neglect the benefits of a balanced allocation and use random allocation within linear model frameworks. On the other hand, many of those researchers report balance or randomization checks, so balance on observable covariates appears to matter (e.g. Blattman et al., 2017; Cristia et al., 2017).

This paper addresses the benefits as well as the limits of a systematic allocation of treatments compared to random allocation in a linear model framework. I compare random allocation to optimal design as well as more frequently used designs such as stratification or matching. The paper introduces a measure of covariate balance taken from medical research. Using this measure, I develop a sample size formula for linear models that takes into account the covariate balance. The formula shows that given a model with a fixed number of covariates m, the maximum reduction in sample size of systematic allocation compared to random allocation is approximately equal to m. Once one allows for a varying number of covariates, in case of systematic allocation one can and should control for more covariates than in case of random allocations to further reduce the necessary sample size. Lastly, I show that systematic allocation of treatments primarily fosters inference whenever the sample size is close to the number of covariates, so typically for small sample sizes. For a fixed number of covariates, as the sample size increases, the ratio of the variance of the treatment estimator under any type of systematic allocation to the variance under random allocation converges to one.

This paper is structured as follows: Section 2 provides a short overview over the related literature. Section 3 defines optimal design in the case of linear models and discusses intuitions behind this type of design. Section 4 introduces a multivariate measure of covariate balance for linear models called the *loss due to the lack of balance*. Given this measure, the section discusses the role of balance for the variance of the treatment estimator and statistical power, or more precisely necessary sample sizes to obtain a given power. Section 5 presents numerical algorithms for finding optimal treatment allocations. Section 6 compares different treatment allocations with respect to their impact on covariate balance and power. Section 7 concludes.

2 Related Literature

Since this paper studies optimal designs for linear models, it is closely related to the statistical field of optimal experimental design (see Pukelsheim (2006) for an overview). In a given model framework, optimal design approaches search for allocations of experimental units that minimize (functions of) the variance of the treatment estimator in the chosen model.[3] Applications of this theory can be found in most fields of research, including engineering (Harville, 1974), biology (Khinkis et al., 2003), chemistry (Telen et al., 2016) and physics (Berger et al., 2017).

As mentioned in the introduction, econometric and related statistical literature on experimental design primarily focuses on randomization inference. Experimental designs used in economic experiments consequently involve a certain degree of randomness. Popular allocation algorithms include stratified randomization (Athey and Imbens, 2017), non-bipartite matching (Greevy et al., 2004; Moore, 2012) and re-randomization (see Bruhn and McKenzie (2009) for an overview of the usage of these algorithms in the field of development economics).

Lately, Kasy (2016b) suggested an optimal design method for economic experiments. In a Bayesian inference framework similar to most decision theoretic models, he proposes to minimize the expected posterior mean squared error (MSE) of the treatment estimator given the experimenter's prior via the treatment allocation. Banerjee et al. (2017) extend the 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. Schneider and Schlather (2017) take the optimal design approach to frequentist inference. They propose a re-randomization approach that aims at minimizing the variance of the treatment estimator in a linear model and use this as a heuristic for treatment allocation in a randomization inference setting.

To this end, the present paper is related to a small strain of the econometric literature targeting the use of linear models in experiments (Freedman, 2008; Deaton, 2010; Schochet, 2010; Athey and Imbens, 2017). This literature takes random treatment allocation as given and shows that a random allocation of treatments does not automatically imply the linear model assumptions to be satisfied. Contrary to this literature, this paper takes the linear model assumptions as given and targets the implications for the treatment allocation.

To my best knowledge, the only economic paper that tries to capture the effect of treatment allocation from a plain linear model perspective is Bruhn and McKenzie (2009). They recommend to analyze the data with a linear model involving the same covariates as in the experimental design. However, their simulations are based on a repeated allocation of treatments on a fixed data set. This is a randomization inference setting in which the assumptions of linear models are not fulfilled (see Athey and Imbens, 2017).

[3] One typical case is a polynomial model (see Smith (1918) for an initial contribution). For example in an experiment to develop a law of gravity, the initial model for the dependence between the height of a dropped ball and the speed at which the ball hits the ground could be quadratic or cubic. Depending on that model, one question about the experimental design can be from which heights the ball should be dropped to optimally estimate the model. For more examples see Atkinson et al. (2007).

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 experimental design on the distribution of the treatment estimator in a linear model. In this paper, I assume all covariates to be observed prior to allocating the treatments. The experimental design involves the same covariates as the analysis of the data. The section starts with a definition of the experimental setting, then analyzes the role of treatment allocation for the distribution of the treatment estimator, introduces optimal design, and finishes with an analysis of stratified and matched experiments in this setting.

3.1 Setting

I consider a sample of n individuals drawn from a population of possible subjects.[4] For the data generating process, I assume a linear model:

$$Y = X\beta_x + T\beta_t + \varepsilon \qquad \text{with } \varepsilon \sim \mathcal{N}(0, I\sigma^2), \tag{1}$$

where $Y \in \mathbb{R}^n$ is the dependent variable, $X = (1, X_1, ..., X_{m+1}) \in \mathbb{R}^{n\times m+1}$ the covariate matrix, and $T \in \{0,1\}^n$ indicates the treatment allocation. The covariates are measured prior to allocating the treatments. Each participant i can only be allocated either to the treatment group ($T_i = 1$) or to the control group ($T_i = 0$). 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 $Y_i = x_i\beta_x + \varepsilon_i$, 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 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 which the matrix (X, T) has full rank. I only consider one treatment and one control group and assume no interaction effects between the treatment and the covariates. See Appendix D for a generalization on multiple treatments and interactions.

In line with Kallus (2017), I define the term *experimental design* as the distribution of T and the term *treatment allocation* as the realization of the design. I assume that the allocation of treatments takes place before observing the dependent variable but after observing the covariates. Thus any possible design has to be independent of Y given X.

Note that the definition of the linear model above does not require any assumption on the experimental design. Consequently, for the model it makes no difference whether the same allocation T was derived via a random or a deterministic design. All results of the next section will hold for any fixed $T \in \mathcal{T}$, given (1).

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

3.2 Optimal Allocation

I start with the model for estimation:

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

Let $b(T) = \begin{pmatrix} b_x(T) \\ b_t(T) \end{pmatrix} := ((X,T)'(X,T))^{-1}(X,T)'Y$ denote the ordinary least squares (OLS) estimates of the coefficients. It is well known that, under the assumptions of section 3.1, 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. *(Variance of the Treatment Estimator)*
Let $T \in \mathcal{T}$. Then $\mathbb{V}[b_t(T)]$ has the following two representations:
(i) $\mathbb{V}[b_t(T)] = \sigma^2(T'M_XT)^{-1}$
(ii) $\mathbb{V}[b_t(T)] = \frac{\sigma^2}{n\cdot\hat{p}_T(1-\hat{p}_T)} \cdot \frac{1}{(1-R^2_{T,X})}$

Here I use 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 statistic of the OLS regression $T = X\theta + \tilde{\varepsilon}$ and measures the linear dependence between the treatment variable and the covariates. $\frac{1}{1-R^2_{T,X}}$ 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. The second distinguishes between the different factors that affect the magnitude of the variance of the treatment estimator and thus provides an intuition.

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.[5]

The influence of $R^2_{T,X}$ 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. Contrary to what other authors claim (e.g. McClelland, 1997; List et al., 2011; Carneiro et al., 2016), $R^2_{T,X}$ is not equal to zero when the random variables that induce T and X are independent (for example in case of random treatment allocation). By definition, $R^2_{T,X}$ is always greater or equal to zero, with equality only if the means of all covariates are equal in treatment and control group. Random allocation will achieve this balance on average, over many experiments, but in any single experiment there will be imbalances (and thus $R^2_{T,X} > 0$). In fact, the $R^2_{T,X}$ term is the reason that

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

optimal design leads to a lower variance of the treatment estimator than random allocation.

Definition 3.2. *(Optimal Treatment Allocation and Optimal Design)*
A treatment allocation $T \in \{0,1\}^n$ is optimal if and only if it minimizes the variance of the treatment estimator $\mathbb{V}[b_t(T)]$ over all $T \in \mathcal{T}$.
An experimental design $\mu \in [0,1]^{\mathcal{T}}$ is optimal if and only if it minimizes the variance of the treatment estimator over all admissible distributions on $\mathcal{T}$.

Proposition 3.1 provides some intuition on how optimal allocations for linear models look like: They aim at minimizing the linear dependence between the treatment variable and the covariates, $R^2_{T,X}$. Whenever the covariate matrix allows for this, optimal allocations result in the covariate means in the treatment group being identical to those in the control group (in this case, $R^2_{T,X} = 0$). Thus, optimal design aims at balancing covariates across treatment and control group (see also section 4). Definition 3.2 leads to a very similar theorem as in the Bayesian framework of Kasy (2016b).

Theorem 3.3. *(Deterministic vs. Random Designs)*
In a linear model framework, there exists at least one deterministic optimal design. Any design is optimal if and only if it randomizes exclusively among optimal treatment allocations.

Proof. See Appendix B.1. □

This theorem shows that there is no benefit of any random component in the experimental design concerning the variance of the treatment estimator. Contrary to what intuitions, for example from the field of portfolio allocation, might suggest, once the design allows for more than one possible allocation, the variance does not get lower than the variance for the best deterministic treatment allocation. On top of the deterministic design, there exists, by symmetry, also at least one optimal design that involves a certain degree of randomness. For an optimal allocation T^*, this design randomizes among T^* and $1 - T^*$.[6]

What are the implications of balanced designs, such as optimal design, on reverse causality and ommitted variables?[7] As Bruhn and McKenzie (2009) argue, balancing on observable covariates can only increase the balance on unobservables. Their argument is as follows: Consider an unobserved covariate Z. Then Z can be written as $Z_1 + Z_2$, where $Z_1 = (I - M_X)Z$ is perfectly collinear with the observable covariates X and $Z_2 = M_X Z$ is uncorrelated with X. Balancing on X will thus also balance Z_1 and does not increase imbalance on Z_2 compared to random allocation. The exclusion of reverse causality and hence the estimation of causal effects does not require the treatment allocation to be random,

[6] Note that $\mathbb{V}[b_t(T)] = \mathbb{V}[b_t(1-T)]$ for all $T \in \mathcal{T}$.

[7] One argument in favor of random allocation in experiments is that this design protects against all kinds of model misspecifications. Freedman (2008) shows that this is not true. Random allocation does not ensure the linear model assumptions to be fulfilled. If the linear model assumptions do not hold, the estimator of the treatment effect and the estimator of the standard deviation can be biased. In this paper, I assume the model specification to be correct and regard the implications on experimental design.

but exogenous. In this paper, treatment allocation is a function of the covariates X and possibly a random component. Thus, the treatment allocation has to be independent of the error term ε conditioned on X. For a more thorough discussion on model assumptions under balanced designs see Aickin (2001). It is crucial, that the analysis of the data controls for the same covariates as the allocation of treatments. Omitting variables in the analysis generally results in overly conservative standard errors (Kernan et al., 1999).

As a last proposition in this section, I highlight the connection between treatment allocation and power of a two-sided t-test on the treatment estimator. The proposition shows that power monotonously decreases as the variances of the treatment estimator increases:

Proposition 3.4. *Let the assumptions in (1) hold. Further, let the probability of a type I error, $\alpha \in (0, \frac{1}{2})$, be given. Then the power of the experiment is given by:*

$$\mathbb{P}(|\frac{b_t(T)}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_{d,1-\frac{\alpha}{2}}) = P_\alpha(n - m - 2, |\frac{\beta_t}{\sqrt{\mathbb{V}[b_t(T)]}}|), \tag{3}$$

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

Proof. See Ghosh (1973). □

3.3 Stratification and Matching

To provide further intuition on optimal designs, this section shows that stratification and matching are special cases of optimal designs in a linear model framework.

In this section, I will refer to optimal design as an algorithm that randomizes among all optimal treatment allocations. Stratification divides the sample into k blocks and randomly allocates treatments within each block such that half the units out of each block are allocated to the treatment group, and the other half to the control group. For simplicity, I assume that the sample size n is divisible by k and that each block contains an even number of units. Matching is a special case of stratification for $k = \frac{n}{2}$ blocks.

When using stratification or matching, Bruhn and McKenzie (2009) propose to "control for the method of randomization in [the] analysis". This means, if the allocation of treatments was based on k strata, they propose to analyze the data with the following linear model:

$$Y = \beta_0 + \beta_1 block_1 + ... + \beta_{k-1} block_{k-1} + \beta_T T + \varepsilon, \quad \text{with } \varepsilon \sim \mathcal{N}(0, I\sigma^2) \tag{4}$$

where $block_1, ..., block_k$ are dummy variables for the different strata.[8] Now, let us turn this around. Suppose the researcher commits to the model of Equation 4

[8] 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.

prior to conducting the experiment. Then optimal treatment allocation leads to the same allocation rule as stratification:

Corollary 3.5. *Assume the model of Equation 4 to be true and every block to contain an even number of experimental units. Then a treatment allocation T is optimal in 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 all blocks are of even size and the dependent variable for units of the same block is identically distributed, stratification and matching randomize among all optimal treatment allocations. Consequently, in this case, matching and stratification are optimal designs.

4 Balance and Sample Sizes

This section, together with the simulations of Section 6, quantifies the benefits of a systematic allocation of treatments over random allocation. I provide a notion of balance within linear models taken from medical research. Based on this balance measure, this section analyzes the connection between balance and the variance of the treatment estimator as well as the necessary sample size to detect a treatment effect with a given power.

The term covariate balance is frequently used in the economic literature, 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. Recall from Section 3 that any treatment allocation that yields perfect balance is an optimal treatment allocation. Of course, depending on the covariate matrix, there does not need 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 $T \in \mathcal{T}$, but serves as a lower bound. These considerations justify defining 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 4.1. *Let $\mathbb{V}^* := \frac{4\sigma^2}{n}$ be the variance of the treatment estimator under perfect balance. Then for a treatment allocation $T \in \mathcal{T}$, the loss due to the lack of balance is defined by:*

$$\mathcal{L}_n(T) := n(1 - \frac{\mathbb{V}^*}{\mathbb{V}[b_t(T)]}) = n - 4 \cdot T' M_X T. \tag{5}$$

The loss is a multivariate measure of balance. It measures imbalance resulting from unequal group sizes as well as imbalance because of unequal covariate means across treatment and control group. A loss of zero corresponds to the case of perfect balance, whereas a higher loss indicates larger imbalances. Section 6 provides simulations of the loss due to the lack of balance for different experimental designs. The notion of loss gives rise to a third characterization of the variance of the treatment estimator:

Corollary 4.2. *(Variance)*
Consider a treatment allocation T, with $\mathcal{L}_n(T) = L$. Then:

$$\mathbb{V}[b_t(T)] = \frac{4\sigma^2}{n-L}. \tag{6}$$

Proof. Follows directly from definition 4.1. □

This representation allows for an easy comparison of the variance of the treatment estimators for two treatment allocations. Suppose the first allocation leads to a loss of L_1 and the second to a loss of L_2. Then, the ratio of the variances of the two treatment estimators is given by $\frac{\mathbb{V}[b_t(T_1)]}{\mathbb{V}[b_t(T_2)]} = \frac{n-L_2}{n-L_1}$. I will use this representation to show that asymptotically the variance ratio for random allocation to any other experimental design converges to one.

Theorem 4.3. *(Asymptotic Benefit)*
Let μ_1 be an experimental design that allocates treatments completely randomly, and μ_2 an arbitrary design that yields a lower asymptotic loss, i.e., $\lim_{n\to\infty} \mathcal{L}_n(T_1) - \mathcal{L}_n(T_2) > 0$ a.s., then:

$$\lim_{n\to\infty} \frac{\mathbb{V}[b_t(T_2)]}{\mathbb{V}[b_t(T_1)]} = 1. \tag{7}$$

Proof. See Appendix B.3. □

This theorem shows that asymptotically, designs that balance covariates across treatment and control group do not provide any benefit over a completely random allocation of treatments. Note that this only holds in case of the OLS estimator. Once experiments are analyzed by the difference in means of the dependent variable in treatment and control group, the benefits of systematically balanced designs compared to random allocation do not decrease with sample size (Kasy, 2016b; Aufenanger, 2017). The reason for this is that the OLS estimator already controls for the covariates. The larger the sample size, the lower the variance of β_x and the better the OLS estimator accounts for the imbalances of random allocation. Consequently, the benefits of optimal design are only present in finite samples.

To better quantify the benefits of optimal or alternative designs compared to random allocation in finite samples, I derive an expression for the sample sizes necessary to detect a given effect size with a given power:

Proposition 4.4. *(Sample Size)*
Consider an experimental design μ. Further assume $\mathbb{E}_\mu[\mathcal{L}_n(T)] = L(n)$. Then the

sample size necessary to detect a treatment effect of a size of β_t at level α with a power P solves the following equation:

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

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

Proof. See Appendix B.4. □

Proposition 4.4 provides a sample size formula that takes into account the covariate balance $L(n)$. The function for the loss has to be determined via simulations. As a rule of thumb, one can take $S_{\alpha,P}(n)$ to be constant by replacing the quantiles of the t-distribution by the corresponding quantiles of the normal distribution[9], 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, the number of covariates (see Section 6.1 as well as Atkinson (2002)), and the loss for optimal allocation is approximately zero. Therefore, optimal treatment allocation can reduce the necessary sample size for a given model by approximately m. As Section 6.2 will show, one can and should control for more covariates when using optimal allocation than when using random allocation, leading to a further reduction in necessary sample size.

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):

$$\underset{T\in\{0,1\}^n}{max} T'M_XT. \tag{9}$$

This is a binary quadratic optimization problem, which is numerically very hard to solve.[10] Brute force solution would require calculating $T'M_XT$ 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 experimental units 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 solutions very quickly. I suggest two very simple heuristics for this problem. For a comparison of those two algorithms to alternative optimization algorithms, see Appendix C.

[9] A general rule of thumb is that t-quantiles are 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).

[10] More precisely, the binary quardratic optimization problem belongs to the class of NP-hard problems (Wang and Kleinberg, 2009). Up to now, the does not exists any algorithm that precisely solves NP-hard problems in polynomial time (Milan et al., 2017).

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.[11] 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 exist possible improvements once one changes the assignment for more than one experimental unit simultaneously.

The second algorithm is a simple extension of the local search algorithm, which I call the *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. The larger k, the better the solution, but also the longer the computing time.

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 value of the goal function as T.

6 Simulations

In this section, I present some simulations, comparing optimal treatment allocation to random treatment allocation as well as to stratification, matching and re-randomization. Bruhn and McKenzie (2009) identify the latter three algorithms as the most popular experimental designs in economic research. Matching and stratification aim at defining blocks on the experimental units and to randomize within those blocks (see Section 3.3). For the stratification design, I start by defining dummy variables for continuous covariates that are equal to one whenever the continuous variable is greater than the median and zero otherwise. Given the discrete as well as the discretized continuous covariates, I define blocks such that units within each block are identical with respect to every covariate. The matching algorithm minimizes the sum of Mahalanobis distances between the two units of each block (see Greevy et al., 2004). Re-randomization draws a finite number of k random treatment allocations and selects the one that minimizes the largest t-statistic for the mean difference of the covariates between treatment and control group (Bruhn and McKenzie (2009) refer to this design as min-max re-randomization).

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

The first part of this section compares the different algorithms for a constant model. This means, the model for estimating the data and specifically 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.

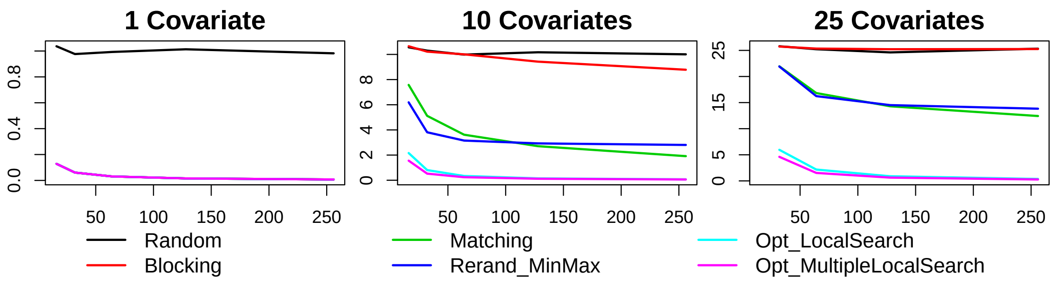
The simulations use the statistical software R (R Development Core Team, 2008). For the matching algorithm, I apply the package *nbpMatching* that implements the optimal matching approach of Greevy et al. (2004) (see Lu et al., 2011).

6.1 Fixed Number of Covariates

In this subsection I compare how the different algorithms perform in a given model. I focus on comparing the losses due to the lack of balance of the different algorithms. As Propositions 4.2 and 4.4 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.[12] 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:

$$Y = X\beta_x + T\beta_t + \varepsilon \qquad \text{with } \varepsilon \sim \mathcal{N}(0, I);\ \beta_x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{10}$$

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 4.1.[13] The re-randomization algorithm uses 100 redraws and the multiple local search algorithm uses 10 redraws:



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

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

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 ten 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 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 algorithm 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 4.4, 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 treatment allocation is for small scale experiments, they also show that there is little need for systematic allocation whenever the sample size is very large compared to the number of covariates. Bruhn and McKenzie (2009) report that out of 18 reviewed experiments in the field of development economics, 12 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 for effectively two covariates.

6.2 Endogenous Number of Covariates

Up to now, I 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 analysis 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).[14] 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 analysis is fairly similar to an analysis by Therneau (1993), who compares the optimal number of covariates for stratification and minimization.[15] 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:

$$Y = T\beta_t + X\beta_x + \varepsilon, \quad \text{with } \varepsilon \sim \mathcal{N}(0,1). \tag{11}$$

[14] These three effects of covariates on the power of the experiment are also apparent in the sample size formula of Proposition 4.4.

[15] Minimization is a popular algorithm 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.

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

$$\beta_i = \frac{m-i}{4m}, i = 1, ..., m. \tag{12}$$

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

$$Y = \beta_0 + \sum_{i=1}^{j} \beta_i X_i + \beta_t T + \tilde{\varepsilon}, \text{with } \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2), \tag{13}$$

where $\tilde{\varepsilon}$ decomposes to $\sum_{i=j+1}^{m} \beta_i X_i + \varepsilon$. Optimal design is based on the same model. In this simulation, I chose a sample size of 64 and a maximum number of $m = 60$ covariates.

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 4.2 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 in Figure 2 are hypothetical cases. This means there do not have to exist treatment allocations that lead to this particular loss or power. The green line represents the case of a loss of zero (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 approaches 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. The figure for the variance already shows that optimal treatment allocation is able to retrieve much more information out of the same covariates than random allocation.

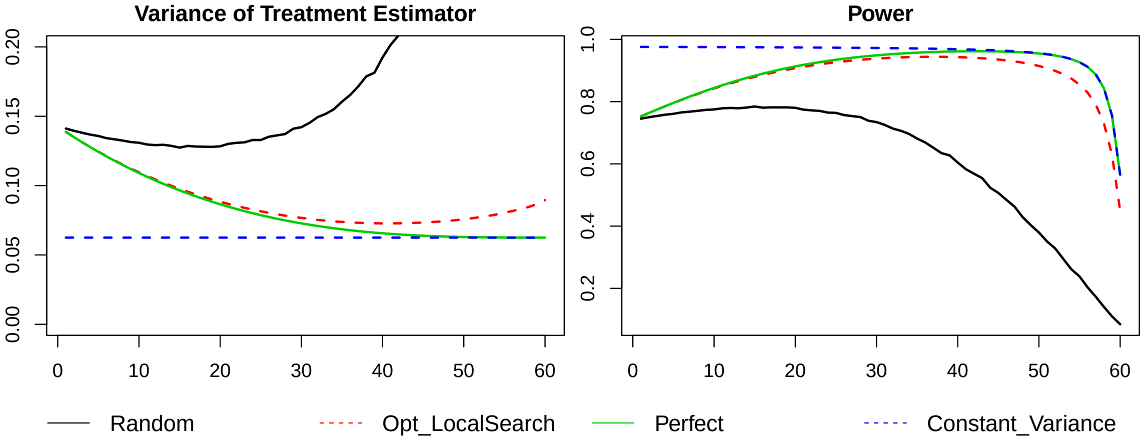


Figure 2: Variance of the Treatment Estimator and Power Depending on the Number of Control Variables for Experiments with 64 Units.

The right graphic in Figure 2 presents statistical power, i.e., the probability of estimating a significant treatment effect. For random treatment allocation, 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 power for the hypothetical case that the variance of the treatment estimator does not change with the number of control variables. This helps to distinguish the importance of the two downsides of adding control variables with respect to 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 makes the blue line 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 zero. This shows that as long as the number of covariates is not too close to 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 analyzes experimental designs from a linear model perspective. I show the benefits as well as the limits of a systematic allocation of treatments compared to random allocation.

A first result is that asymptotically, no systematic allocation of treatments yields any benefit over random allocation (Theorem 4.3).

In finite samples however, even though the OLS estimator already controls for imbalances in the covariates, a systematic allocation of treatments can reduce the variance of the treatment estimator and increase statistical power (Section 6.1). In terms of the sample size necessary to detect an effect of a given strength with a given power, optimal designs can reduce necessary sample sizes by approximately m, the number of covariates in the model (Proposition 4.4). If possible, one can and should control for more covariates in case of optimal allocation than in case of random allocation to maximize statistical power (Section 6.2).

From a linear model perspective, it makes no difference whether treatments are allocated randomly or deterministically. Experimental designs are only optimal if they randomize exclusively among optimal treatment allocations (Theorem 3.3). The popular experimental designs matching and stratification are optimal if all blocks are of even size and equally distributed dependent variables in each block (Section 3.3). For blocks of unequal size, or whenever the linear model does not imply identically distributed dependent variables in each block, those algorithms perform worse than optimal allocation (Section 6.1).

References

Abdulkadiroğlu, A., P. A. Pathak, and C. R. Walters (forthcoming): "Free to Choose: Can School Choice Reduce Student Achievement?" *American Economic Journal: Applied Economics.*

Aickin, M. (2001): "Randomization, Balance, and the Validity and Efficiency of Design-Adaptive Allocation Methods," *Journal of Statistical Planning and Inference*, 94, 97 – 119.

Athey, S. and G. Imbens (2017): "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. (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.

Aufenanger, T. (2017): "Machine Learning to Improve Experimental Design," Unpublished draft.

Banerjee, A., S. Chassang, and E. Snowberg (2017): "Decision Theoretic Approaches to Experiment Design and External Validity," in *Handbook of Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1 of *Handbook of Economic Field Experiments*, 141 – 174.

Beasely, J. E. (1998): "Heuristic Algorithms for the Unconstrained Binary Quadratic Programming Problem," Working paper, London, UK: Management School, Imperial College.

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.

Bertsimas, D., M. Johnson, and N. Kallus (2015): "The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples," *Operations Research*, 63, 868–876.

Blattman, C., J. C. Jamison, and M. Sheridan (2017): "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia," *American Economic Review*, 107, 1165–1206.

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.

Cristia, J., P. Ibarrarán, S. Cueto, A. Santiago, and E. Severín (2017): "Technology and Child Development: Evidence from the One Laptop per Child Program," *American Economic Journal: Applied Economics*, 9, 295–320.

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.

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.

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.

Glewwe, P., M. Kremer, and S. Moulin (2009): "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, 1, 112–35.

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.

Holland, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Horton, J., D. Rand, and R. Zeckhauser (2011): "The Online Laboratory: Conducting Experiments in a Real Labor Market," *Experimental Economics*, 14, 399–425.

Imai, K., G. King, and E. Stuart (2008): "Misunderstandings Among Experimentalists and Observationalists about Causal Inference," *Journal of the Royal Statistical Society, Series A*, 171, 481 – 502.

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

Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

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.

Kallus, N. (2017): "Optimal a Priori Balance in the Design of Controlled Experiments," *Journal of the Royal Statistical Society: Series B (Statistial Methodology)*.

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.

Kempthorne, O. (1955): "The Randomization Theory of Experimental Inference," *Journal of the American Statistical Association*, 50, 946–967.

Kernan, W. N., C. M. Viscoli, R. W. Makuch, L. M. Brass, and R. I. Horwitz (1999): "Stratified Randomization for Clinical Trials," *Journal of Clinical Epidemiology*, 52, 19 – 26.

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.

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.

Krawczyk, M. and M. Smyk (2016): "Author's Gender Affects Rating of Academic Articles: Evidence from an Incentivized, Deception-Free Laboratory Experiment," *European Economic Review*, 90, 326–335.

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," *The Annals of Statistics*, 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.

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.

Milan, A., S. H. Rezatofighi, R. Garg, A. R. Dick, and I. D. Reid (2017): "Data-Driven Approximations to NP-Hard Problems." in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 1453–1459.

Moore, R. T. (2012): "Multivariate Continuous Blocking to Improve Political Science Experiments," *Political Analysis*, 20, 460–479.

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

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.

Rosenbaum, P. R. (2002): "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, 17, 286–327.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Experimantal and Observational Studies," *Journal of Educational Psychology*, 66, 688 – 701.

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.

Schochet, P. Z. (2010): "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and Inference*, 140, 246 – 259.

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

Smith, K. (1918): "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they give Towards a Proper Choice of the Distribution of Observations," *Biometrika*, 12, 1–85.

Student (1938): "Comparison Between Balanced and Random Arrangements of Field Plots," *Biometrika*, 29, 363–378.

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.

Wang, D. and R. Kleinberg (2009): “Analyzing Quadratic Unconstrained Binary Optimization Problems via Multicommodity Flows,” *Discrete Applied Mathematics*, 157, 3746–3753.

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 Alternative Covariate Distributions

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

- normal: A normal distribution with mean zero and variance one. This serves as an example of a continuous distribution.
- gamma: A gamma distribution with shape parameter two and scale parameter one. This serves as an example of a skewed distribution.
- different: Covariates that follow this distribution are a sum of a uniformly distributed variable on [-10,10] and a second variable that is normally distributed with probability 2/3 and gamma distributed with probability 1/3. This serves as an example of a slightly more complex distribution that consists 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.

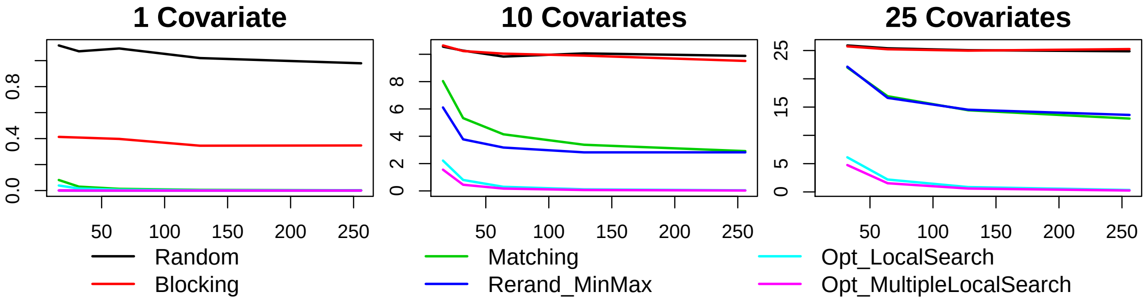


Figure 3: Loss for Normal Distributed Covariates

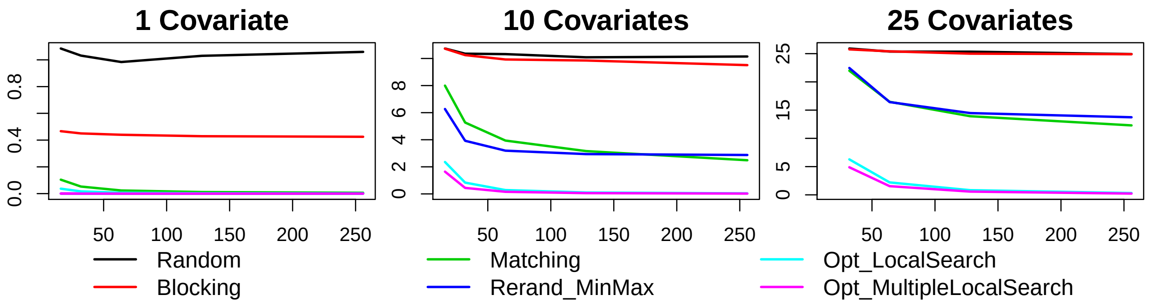


Figure 4: Loss for Gamma Distributed Covariates

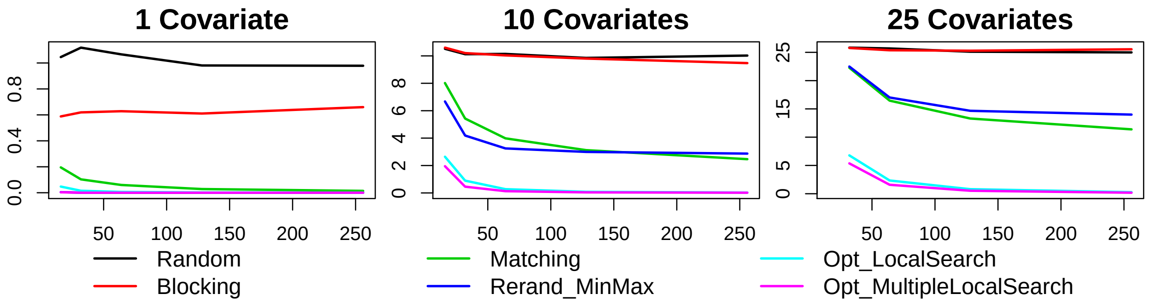


Figure 5: Loss for Different Distributed Covariates

B Proofs for the Paper

B.1 Proof of Theorem 3.3

Proof. First, consider only deterministic allocations. Since the set $\mathcal{T}$ of admissible allocations is finite, there exists at least one allocation $\hat{T}^*$ that minimizes the variance of the treatment estimator. Now, consider designs that involve some randomness. Let $\mathcal{M}$ be the set of all probability distributions on $\mathcal{T}$. For an arbitrary $\mu \in \mathcal{M}$, the variance of the treatment estimator is given by:

$$\mathbb{V}_\mu[b_t(T)] = \mathbb{V}_\mu[E[b_t(T)|T]] + \mathbb{E}_\mu[\mathbb{V}[b_t(T)|T]]$$

$$= \mathbb{V}_\mu[\beta_t] + \mathbb{E}_\mu[\mathbb{V}[b_t(T)|T]] = \mathbb{E}_\mu[\mathbb{V}[b_t(T)|T]]$$

Consequently, the variance of the treatment estimator under any random allocation μ is simply the (probability weighted) average of all deterministic deterministic allocations in the support of μ. Since $\mathbb{V}[b_t(T)|T] \geq \mathbb{V}[b_t(T^*)]$ for every $T \in \mathcal{T}$, the variance under μ is either equal to $\mathbb{V}[b_t(T^*)]$ if μ randomizes only among optimal deterministic allocations, or larger if the support of μ contains at least one sub-optimal allocation. □

B.2 Proof of Corollary 3.5

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

$\Leftarrow$: By Proposition 3.1(i): $\mathbb{V}[b_t(T)] = \sigma^2(T'M_XT)^{-1}$, 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 $\mathbb{R}^n$ and n_j is the number of subjects in stratum j. Thus

$$\begin{aligned} T'M_XT &= T'T - \frac{1}{4}e_1'X'X(X'X)^{-1}X'Xe_1 && (14)\\ &= \frac{n}{2} - \frac{1}{4}e_1'X'Xe_1 && (15)\\ &= \frac{n}{2} - \frac{1}{4}\mathbf{1}'\mathbf{1} && (16)\\ &= \frac{n}{2} - \frac{n}{4} = \frac{n}{4} && (17) \end{aligned}$$

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.

$\Rightarrow$: 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:

$$\begin{aligned} R^2_{T,X} &= 0 && (18)\\ \Leftrightarrow &\frac{||X(X'X)^{-1}X'T - \hat{p}_T \cdot \mathbf{1}||^2}{||T - \hat{p}_T \cdot \mathbf{1}||^2} = 0 && (19)\\ \Rightarrow &||X(X'X)^{-1}X'T - \hat{p}_T \cdot \mathbf{1}||^2 = 0 && (20)\\ \Rightarrow &X(X'X)^{-1}X'T = \frac{1}{2} \cdot \mathbf{1} && (21)\\ \Rightarrow &X'T = \frac{1}{2} \cdot X'\mathbf{1} && (22)\\ \Rightarrow &\frac{1}{n_j}\sum_{i=1}^{n} block_j^{(i)}T_i = \frac{1}{2}, \text{for all } j = 1, ..., k && (23) \end{aligned}$$

□

B.3 Proof of Theorem 4.3

Since the loss for random allocation $\mathcal{L}_n(T_1)$ converges almost surely to m, the number of covariates (see Atkinson, 2002), there almost surely exists a constant

$K_1 \in \mathbb{R}$ and a $N_1 \in \mathbb{N}$ such that $\mathcal{L}_n(T_1) < K_1$ for all $n > N_1$. Similarly, since the asymptotic loss of μ_2 is almost surely lower than m, there almost surely exists a $K_2 \in \mathbb{R}$ and a $N_2 \in \mathbb{N}$ such that $\mathcal{L}_n(T_2) < K_2$ for all $n > N_2$. Let $n > \max\{N_1, N_2\}$. Since the loss is always greater or equal to zero, this yields:

$$\frac{\mathbb{V}[b_t(T_2)]}{\mathbb{V}[b_t(T_1)]} = \frac{n - \mathcal{L}_n(T_2)}{n - \mathcal{L}_n(T_1)} \leq \frac{n}{n - K_1} \underset{n\to\infty}{\to} 1, \tag{24}$$

and

$$\frac{\mathbb{V}[b_t(T_2)]}{\mathbb{V}[b_t(T_1)]} = \frac{n - \mathcal{L}_n(T_2)}{n - \mathcal{L}_n(T_1)} \geq \frac{n - K_2}{n} \underset{n\to\infty}{\to} 1. \tag{25}$$

B.4 Proof of Proposition 4.4

Proof. Let μ be an experimental design. Then the loss $\mathcal{L}_n(T)$ is typically a random variable that depends on the realization of the covariate matrix and the realization of the treatment allocation. For this proof, start by assuming $\mathcal{L}_n(T) = L(n)$ to be deterministic.

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 3.4. 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:

$$P(d, |\delta|) = P \Leftrightarrow |\delta| = g_P(d). \tag{26}$$

Writing $g_P(n) := g_P(d(n))$ and plugging the definition of δ into Equation 26 yields:

$$|\frac{\beta_t}{\sqrt{\mathbb{V}[b_t(T)]}}| = g_P(n) \tag{27}$$

By Proposition 4.2:

$$\Leftrightarrow \frac{\beta_t^2(n - L(n))}{4\sigma^2} = g_P(n)^2 \tag{28}$$

$$\Leftrightarrow n = \underbrace{\frac{4\sigma^2 g_P(n)^2}{\beta_t^2}}_{S_{\alpha,P}(n)} - L(n) \tag{29}$$

Now, consider the case that $\mathcal{L}_n(T)$ is stochastic with $\mathbb{E}[\mathcal{L}_n(T)] = L(n)$, where the expectation is over the joint distribution of the covariate matrix and the treatment allocation. Note that $S_{\alpha,P}(n)$ depends neither on the covariate matrix nor on the treatment allocation. Therefore:

$$\mathbb{E}[n] = S_{\alpha,P}(n) + L(n). \tag{30}$$

Consequently, in expectation, the necessary sample size is equal to $S_{\alpha,P}(n)$ plus the expected loss $L(n)$.

It remains to show that $g_P(n)$ is approximately equal to $t_{n-m-2,P} + t_{n-m-2,1-\frac{\alpha}{2}}$.

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:

$$P(d,|\delta|) = \mathbb{P}(|\frac{b_t(T)-\beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(t)]}} + \frac{\beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_\alpha) = \mathbb{P}(|X + \frac{\beta_t}{\sqrt{\hat{\mathbb{V}}[b_t(T)]}}| > t_\alpha). \quad (31)$$

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:

$$P(d,|\delta|) \approx \mathbb{P}(|X+\delta| > t_\alpha) \quad (32)$$

$$= \mathbb{P}(X < -t_\alpha + \delta) + \mathbb{P}(X < -t_\alpha - \delta) \quad (33)$$

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 be much smaller than this whenever P is large (which one would typically assume). This yields the approximation:

$$P(d,|\delta|) \approx \mathbb{P}(X < -t_\alpha + |\delta|). \quad (34)$$

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

$$\mathbb{P}(X < -t_\alpha + |\delta|) = P \Leftrightarrow -t_\alpha + |\delta| = t_P \Leftrightarrow |\delta| = t_\alpha + t_P. \quad (35)$$

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

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 is meant to justify this suggestion. In the first part of this section (Part C.1), I compare the local search to some more sophisticated 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 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 subject in the experiment represents a new variable for the optimization. I consider 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 still have a value of 0.5 the coordinate with the highest improvement from changing its value to 1 and the coordinate with the highest improvement from changing its value to 0. Then, with a specific chance proportional to 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 consists 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 to avoid moving back right away, the algorithm blocks the coordinate along which 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 allocation with the highest value of the 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 the 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 ten random starting points (Opt_MLSR) and multiple local search with ten 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. It does, however, 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 parentheses are the average losses for this algorithm, whereas the values in parentheses are the average computing times (in seconds) 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,[16] the local search algorithm performs better than the greedy algorithm and only slightly worse than the more computationally intensive tabu search and simulated annealing algorithms. When using multiple random starting points (Opt_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 (Opt_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

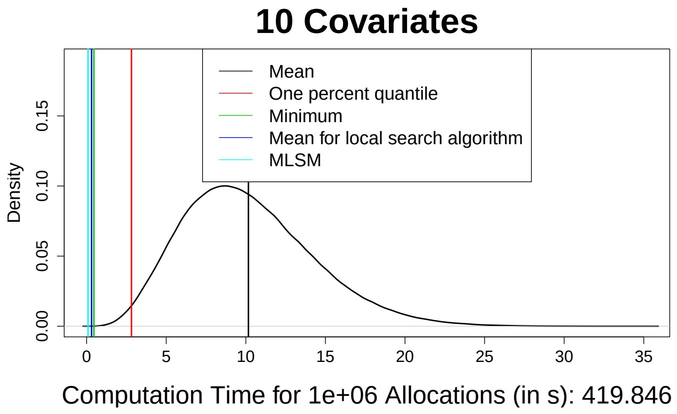
[16] Or equivalently, maximizing $T'M_XT$.

that this procedure performs "reasonably well". The argument, also picked up by Banerjee et al. (2017), is the following: Suppose one re-randomizes for $k \in \mathbb{N}$ times. Then the probability that 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 re-randomization.

To analyze the question whether re-randomization could be used as a simple alternative to the local search algorithm, I analyze 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 ten 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 ten, 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. On the same computer, the local search algorithm leads to a loss of half the size in only ten milliseconds.



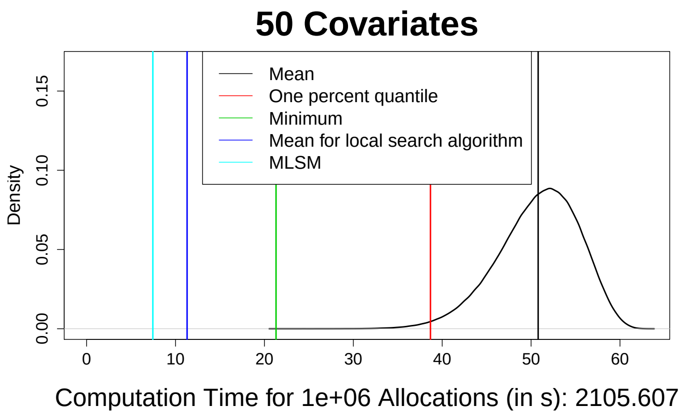


Figure 6: Distribution of Loss for Random Allocation

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 $\hat{T}$.[17]
3. Calculate $\tilde{T}'M_X\tilde{T}$ for every $\tilde{T} \in \hat{T}$.
4. If $\underset{\tilde{T}\in\hat{T}}{max}\ \tilde{T}'M_X\tilde{T} > V$:
     Set $T = \underset{\tilde{T}\in\hat{T}}{argmax}\ \tilde{T}'M_X\tilde{T}$ and $V = \underset{\tilde{T}\in\hat{T}}{max}\ \tilde{T}'M_X\tilde{T}$;
     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 $T^{*\prime}M_XT^* = \underset{T\in\hat{\mathcal{T}}}{min}\ T'M_XT$
```

[17] 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 $T^{(i)\prime} M_X T^{(i)}$ for every $i \in \{1, ..., n\}$ with $L_i = 0$.

4. If $\max_{i \in \{1,...,n\}, L_i = 0} T^{(i)\prime} M_X T^{(i)} > V^*$:
     Set $j = \operatorname{argmax}_{i \in \{1,...,n\}, L_i = 0} T^{(i)\prime} M_X T^{(i)}$
     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 $j = \operatorname{argmax}_{i \in \{1,...,n\}, L_i = 0} T^{(i)\prime} M_X T^{(i)}$
   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 $T^{(j)\prime} M_X T^{(j)}$

5. If $T^{(j)\prime} M_X T^{(j)} > V^*$
     Set $T = T^{(j)}$ and $V = T^{(j)\prime} M_X T^{(j)}$
     Set $T^* = T^{(j)}$ and $V^* = T^{(j)\prime} M_X T^{(j)}$
   Else:
     If $T^{(j)\prime} M_X T^{(j)} > V$
       Set $T = T^{(j)}$ and $V = T^{(j)\prime} M_X T^{(j)}$
     Else:
       With a probability of $exp(-\frac{|V - T^{(j)\prime} M_X T^{(j)}|}{temperature})$:
         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 $t < maxiter$
     Set $temperature = \alpha \cdot temperature$
     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 analysis on linear models involving interaction effects and on experiments including multiple treatments. I consider the following model:

$$Y = X\beta_x + H(X,T)\beta_h + T\beta_t + \varepsilon, \quad \text{with } \varepsilon \sim \mathcal{N}(0, I\sigma^2), \tag{36}$$

$X = \begin{pmatrix} 1 & X_{1,1} & ... & X_{1,m} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & ... & X_{n,m} \end{pmatrix}$ 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 ordinary least squares estimates of β_x, β_h and β_t are:

$$b = \begin{pmatrix} b_x \\ b_h \\ b_t \end{pmatrix} = (C'C)^{-1}C'Y.$$

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 or only in a selection of the treatment effects. I denote the effects that are most important to the researcher *major effects* 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 estimator of all coefficients b has the following variance-covariance matrix:

$$\mathbb{V}[b] = \sigma^2(C'C)^{-1} = \sigma^2 \begin{pmatrix} N'N & N'Z \\ Z'N & Z'Z \end{pmatrix}^{-1}.$$

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 formula for block matrices,[18] this matrix is given by:

$$\mathbb{V}[b_z] = \sigma^2 (Z' M_N Z)^{-1}. \tag{37}$$

Similar to the case of one treatment and no interaction effects, the goal is to minimize the variance of the treatment estimators. 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, the matrix $\mathbb{V}[b_z]$ reduces to a scalar, which can be minimized by the same binary optimization techniques presented in the paper. Whenever $\tilde{m} > 1$, however, this is a matrix and minimization 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 allocations 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 $tr(\mathbb{V}[b_z])$ 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 usage 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.

[18] 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).