WHY EXPERIMENTERS SHOULD NOT RANDOMIZE, AND WHAT THEY SHOULD DO INSTEAD

Maximilian Kasy[1]

This paper discusses experimental design for the case that (i) we are given a distribution of covariates from a pre-selected random sample, and (ii) we are interested in the average treatment effect (ATE) of some binary treatment. We show that in general there is a unique optimal non-random treatment assignment if there are continuous covariates. We argue that experimenters should choose this assignment. The optimal assignment minimizes the risk (e.g., expected squared error) of treatment effects estimators. We provide explicit expressions for the risk, and discuss algorithms which minimize it.

The objective of controlled trials is to have treatment groups which are similar a priori (balanced), so we can "compare apples with apples." The expressions for risk derived in this paper provide an operationalization of the notion of balance. The intuition for our non-randomization result is similar to the reasons for not using randomized estimators - adding noise can never decrease risk.

The formal setup we consider is decision-theoretic and nonparametric. In simulations and an application to project STAR we find that optimal designs have mean squared errors of up to 20% less than randomized designs.

JEL Codes: C9

1. INTRODUCTION

Economists conducting field experiments are often confronted with variants of the following situation (cf. Duflo et al., 2007; List and Rasul, 2011). They have selected a random sample from some population and have conducted a baseline survey for the individuals in this sample. Then a discrete treatment is assigned to the individuals in this sample, usually based on some randomization scheme. Finally, outcomes are realized, and the data are used to perform inference on some average treatment effect.

A key question that experimenters have to decide on is how to use covariates from the baseline survey in the assignment of treatments. Intuition and the lit-

Assistant professor, Department of Economics, Harvard University, and junior associate faculty, IHS Vienna. Address: Littauer Center 200, 1805 Cambridge Street, Cambridge, MA 02138. E-Mail: maximiliankasy@fas.harvard.edu.

[1] I thank Guido Imbens for suggesting the topic of stratification in randomized experiments to me, and Gary Chamberlain, Raj Chetty, Nathaniel Hendren, Guido Imbens, Larry Katz, and Michael Kremer, as well as seminar participants at the Harvard development retreat, the Harvard Labor economics workshop, the Harvard Quantitative Issues in Cancer Research seminar, and the UT Austin, Princeton, and Columbia Econometrics seminars for helpful discussions.

erature suggest to us stratified randomization conditional on covariates.[1] We analyze this situation as a decision problem. The experimenting economist has to choose a treatment assignment vector and an estimator, given knowledge of the covariate distribution in the sample. Her objective is to minimize risk based on a loss function such as the mean squared error of a point estimator. The decision criteria considered are Bayesian average risk and conditional minimax risk.

We show, first, that experimenters should not randomize in general. While surprising at first, the basic intuition for this result is simple. The conditional expected loss of an estimator is a function of the matrix of covariates and of the treatment assignment vector. The treatment assignment vector that minimizes conditional expected loss is generically unique if there are continuous covariates, so that a deterministic assignment strictly dominates all randomized assignments.[2]

The recommendation not to randomize raises the question of identification. We show that conditional independence of treatment and potential outcomes given covariates still holds for the deterministic assignments considered, under the usual assumptions of independent sampling and stable unit treatment values. Conditional independence only requires a controlled trial (CT), not a randomized controlled trial (RCT).

We next propose a general class of nonparametric priors for the conditional expectation of potential outcomes given covariates. If our objective is to minimize the expected squared error between our estimator and the average treatment effect (ATE), and if we restrict attention to linear estimators, then the optimal estimator is given by the posterior best linear predictor for the ATE. This estimator only uses the first two moments of the prior, and expected loss is given by the posterior variance. We also consider the case where the experimenter is committed to using simple comparison-of-means estimators when analyzing the data, yet is willing to use prior information for assigning treatment to minimize the mean squared error of such estimators.

We proceed by discussing how to pick the prior moments, and how to make the prior non-informative about treatment effects, while imposing the smoothness required to extrapolate to counterfactual outcomes. Based on this discussion, we suggest a particular prior that leaves few free parameters to be chosen by the experimenter, and which allows for an automated procedure to choose a treat-

[1] Duflo et al. (2007, section 4.5) state, for instance, "if several binary variables are available for stratification, it is a good idea to use all of them. [...] When one or several of the possible stratification variables are continuous [...] it will be necessary to make choices about which variables to use for stratification [...] taking into consideration the extent to which the candidate stratification variables are likely to explain the outcome variable."

[2] If experimenters have a preference for randomization for reasons outside the decision problem considered in the present paper, a reasonable variant of the procedure suggested here would be to randomize among a set of assignments which are "near-minimizers" of risk. If we are worried about manipulation of covariates, in particular, a final coin-flip which possibly switches treatment and control groups might be helpful. I thank Michael Kremer for this suggestion.

ment assignment minimizing expected loss. MATLAB code which implements the derived risk functions, as well as discrete optimization algorithms to find optimal designs, is available from the author's homepage. Given specification of some prior parameters, this code takes a matrix of covariates as its input and provides a treatment assignment as output.

To gain some intuition for our non-randomization result, note that in the absence of covariates the purpose of randomization is to pick treatment- and control-groups which are similar before they are exposed to different treatments. Formally, we would like to pick groups which have the same (sample) distribution of potential outcomes. Even with covariates observed prior to treatment assignment, it is not possible to make these groups identical in terms of potential outcomes. We can, however, make them as similar as possible in terms of covariates. Allowing for randomness in the treatment assignment to generate imbalanced distributions of covariates can only hurt the balance of the distribution of potential outcomes. The analogy to estimation might also be useful to understand our non-randomization result. Adding random (mean 0) noise to an estimator does not introduce any bias. But it is never going to reduce the mean squared error of the estimator.

The purpose of discussing tractable non-parametric priors - and one of the main contributions of this paper - is to operationalize the notion of "balance." In general, it will not be possible to obtain exactly identical distributions of covariates in the treatment- and control-group. When picking an assignment, we have to trade off balance across various dimensions of the joint distribution of covariates. Picking a prior distribution for the conditional expectation of potential outcomes, as well as a loss function, allows to calculate an objective function (Bayesian risk) which performs this trade-off in a coherent and principled way.

There is a large and old literature on experimental design in statistics, going back at least to Smith (1918), and receiving broader attention since Kiefer and Wolfowitz (1959) and related contributions. A good general introduction to the theory of experimental design can be found in Cox and Reid (2000); a formal treatment of the theory of optimal design is given by Shah and Sinha (1989). Bruhn and McKenzie (2009) have studied the relative variance of estimators under various designs using simulations. Some general discussions on the role of randomization in experiments took place a few decades ago, see in particular Rubin (1978). We will discuss the relationship of their arguments to our results in section 4.1.

In contrast to most of the literature on optimal design, the perspective taken in this paper is nonparametric, while allowing for continuous covariates. Here we draw on the extensive literature on inference on average treatment effects under unconfoundedness, as reviewed in Imbens (2004).

Part of this paper takes a nonparametric Bayesian perspective, considering (Gaussian) process priors for conditional expectations of potential outcomes. This follows a long tradition in the literatures on spline estimation (cf. Wahba, 1990), on “Kriging” in Geostatistics (cf. Matheron, 1973; Yakowitz and Szidarovszky, 1985), and in the more recent machine learning literature (cf. Williams and Rasmussen, 2006). For a general introduction to Bayesian methods with a focus on their decision theoretic motivation, see Robert (2007). O’Hagan and Kingman (1978) considered Gaussian process priors in the context of experimental design, taking an approach similar to ours but without allowing for covariates. A forceful argument for a Bayesian perspective on experimental design has been made by Berry (2006).

The rest of this paper is structured as follows. Section 2 discusses some motivating examples. Section 3 formally introduces the general setup we consider. Section 4 proves that generically optimal designs are non-random given the matrix of covariates, and that the non-random designs considered still satisfy unconfoundedness. Section 5 introduces a general class of nonparametric Bayesian priors and derives the corresponding estimators and expected loss functions. Section 6 discusses how to choose the moments of these priors, maintaining non-informativeness on key features of the data generating process. Section 7 briefly discusses frequentist inference based on the nonparametric Bayesian estimator. Section 8 presents simulation results, including an examination of the robustness of optimal designs to the choice of prior, and an application to project STAR. Appendix A discusses discrete optimization algorithms. Appendix B reviews very briefly the literature on optimal experimental design.

We will use the following notation throughout. We denote covariates by X, the binary treatment by D, potential outcomes by Y^d for $d \in 0, 1$, and realized outcomes by Y. Individual draws of these variables have a subscript i, whereas capital letters without subscripts denote the matrices and vectors of these objects for observations $i = 1, \ldots, n$. Lowercase letters (d and x) denote values of the corresponding variables (D_i, X_i); bold face lower case $\mathbf{d}$ denotes a vector of values d_i. We use θ to denote the family of conditional distributions $\theta := \{P(Y_i^1, Y_i^0 | X_i = x) : \ x \in \mathrm{supp}(X)\}$. We condition on θ in “frequentist” probabilities and expectations, while “Bayesian” probabilities and expectations are unconditional and average over a prior distribution for θ.

2. MOTIVATING EXAMPLES

Before we present the general setup analysed in this paper which allows for continuous covariates, let us consider the variance of an estimator for the average treatment effect in two simple cases, (i) if no covariates are available, and (ii) if a discrete covariate with finite support is available.

If no covariates are available, and $n_d := \sum \mathbf{1}(D_i = d)$ units are assigned (randomly) to treatment d, then the natural estimator for the average treatment effect is given by the difference in means,

$$\widehat{\beta} := \sum_i \left[\frac{D_i}{n_1} Y_i - \frac{1 - D_i}{n - n_1} Y_i \right].$$

Consider the following two procedures for assigning treatments: (1) randomization conditional on the number of units assigned to treatment 1, n_1, and (2) independent assignment of treatment 1 to each unit with probability p. Denote the variance of potential outcome Y^d by $\sigma_d^2 = \mathrm{Var}(Y_i^d | \theta)$. Then the variance of $\widehat{\beta}$ under the first procedure is given by

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n - n_1}.$$

The variance under the second procedure is given by

$$E_{n_1} \left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n - n_1} \right],$$

where the expectation is taken over $n_1 \sim Bin(n, p)$. In this setting it is obvious that choosing $n_1 \approx n \cdot \frac{\sigma_1}{\sigma_0 + \sigma_1}$ yields a smaller variance than any other deterministic or random assignment procedure - and indeed nobody would propose procedure (2). Randomness that puts a positive probability on values of n_1 other than the deterministic optimum strictly increases the variance. This suggests that experimenters *should not randomize* n_1. In this case, the variance does not depend on which observationally equivalent unit is assigned to treatment $d = 1$. Formally, the risk (variance) is flat over permutations of $\mathbf{d}$ which leave n_1 invariant, so randomization given n_1 does not hurt.

Consider next the setting in which a discrete covariate $X_i \in \{0, \ldots, k\}$ is available when assigning treatment, and denote $n_x := \sum_i \mathbf{1}(X_i = x)$, as well as $n_{d,x} := \sum_i \mathbf{1}(X_i = x, D_i = d)$. In this case the natural unbiased estimator for the average treatment effect is given by

$$\widehat{\beta} := \sum_x \frac{n_x}{n} \sum_i \mathbf{1}(X_i = x) \left[\frac{D_i}{n_{1,x}} Y_i - \frac{1 - D_i}{n_x - n_{1,x}} Y_i \right].$$

In this setting we can again consider (1) randomization conditional on $n_d = \sum \mathbf{1}(D_i = d)$ and (2) complete randomization as before. We can additionally consider a third possibility, (3) stratified randomization conditional on $n_{d,x}$. Denoting $\sigma_{d,x}^2 = \mathrm{Var}(Y_i^d | X_i = x, \theta)$, we get that the variance of $\widehat{\beta}$ under stratified randomization (where $n_{d,x}$ is non-random) is given by

$$V(\{n_{d,x}\}) := \sum_x \frac{n_x}{n} \left[\frac{\sigma_{1,x}^2}{n_{1,x}} + \frac{\sigma_{1,x}^2}{n_x - n_{1,x}} \right].$$

Assignment procedures (1) and (2) yield a variance which averages this expression over some distribution of $n_{d,x}$. Again it is immediate that choosing $n_{d,x} \approx n_x \cdot \frac{\sigma_{1,x}}{\sigma_{0,x}+\sigma_{1,x}}$ yields a smaller variance than any other deterministic or random choice of $n_{1,x}$. This suggests that experimenters *should not randomize* $n_{1,x}$ *for any* x. In this case, permutations of $\mathbf{d}$ which leave $n_{1,x}$ invariant for all x leave the variance constant, so it does not hurt to randomize over these permutations.

Consider now finally the case where a continuously distributed covariate $X_i \in \mathbb{R}$ is available. In this case with probability 1 no two observations will have the same X_i. A series of alternative designs can be considered in this case, including complete randomization, randomization conditional on n_d, and various forms of discretizing X_i into bins $[x_j, x_{j+1}]$ and stratifying based on these bins. A special case of this is pairwise randomization. "Full stratification" clearly is not possible in this case, since each stratum only contains one observation. What to do in this case is the topic of the present paper.

3. THE SETUP

Throughout we will consider the following setup.

Assumption 1 (Setup)

The steps of an experiment take place in the following order.

1. Sampling from a population: *We randomly sample n units $i = 1, \dots, n$ from some population. Units of observation are characterized by a vector of covariates X_i as well as potential outcomes (Y_i^0, Y_i^1). Only the covariate vector X_i is observed.*
2. Treatment assignment: *We assign treatment D_i to unit i as a function of the matrix of covariates $X = (X_1', \dots, X_n')'$, as well as (possibly) some randomization device U, so that $D_i = d_i(X, U)$.*
3. Realization of outcomes: *For every unit of observation i, we observe the outcome $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$.*
4. Estimation: *We calculate an estimator $\widehat{\beta}$ of the (conditional) average treatment effect $\beta = \frac{1}{n}\sum_i E[Y_i^1 - Y_i^0 | X_i, \theta]$ as a function of the observed data, $(X_i, D_i, Y_i)_{i=1}^n$.*

Definition 1 (Risk function, Bayesian and minimax risk)

1. *The risk function of a treatment assignment function $\mathbf{d}(X, U)$ and an estimator $\widehat{\beta}$ is given by*

$$R(\mathbf{d}, \widehat{\beta} | X, U, \theta) := E[L(\widehat{\beta}, \beta) | X, U, \theta], \tag{1}$$

   where L is a loss function. The expectation in equation (1) *averages over the conditional distributions of Y_i^1, Y_i^0 given X_i.*
2. *The conditional Bayesian risk is given by the average of the risk function*

over a prior distribution $P(\theta)$,[3]

$$R^B(\mathbf{d}, \widehat{\beta}|X, U) := \int R(\mathbf{d}, \widehat{\beta}|X, U, \theta) dP(\theta), \tag{2}$$

where we assume that the prior distribution for θ does not depend on X. Averaging additionally over the distribution of U yields the Bayesian conditional average risk, and averaging over both X and U yields the Bayesian average risk;

$$\begin{aligned} R^B(\mathbf{d}, \widehat{\beta}|X) &:= \int R^B(\mathbf{d}, \widehat{\beta}|X, U) dP(U) \\ R^B(\mathbf{d}, \widehat{\beta}) &:= \int R^B(\mathbf{d}, \widehat{\beta}|X, U) dP(X) dP(U). \end{aligned}$$

3. *The conditional minimax risk is given by the supremum of the risk function over θ,*

$$R^{mm}(\mathbf{d}, \widehat{\beta}|X, U) := \sup_{\theta} R(\mathbf{d}, \widehat{\beta}|X, U, \theta). \tag{3}$$

Assumption 2 (Decision problem)
The experimenter's objective is to minimize the risk $R(\mathbf{d}, \widehat{\beta}|X, U)$, where R is either equal to Bayesian or to minimax risk, through choice of the estimator $\widehat{\beta}(Y, X, D)$ and through choice of the treatment assignment function $\mathbf{d}(X, U)$.

Discussion

In this subsection we discuss some of the more subtle features of assumptions 1 and 2. The reader may safely skip to section 4 upon first reading.

Average treatment effect, conditional average treatment effect, and sample average treatment effect

In the setting of assumption 1, there are four sources of randomness, coming from (i) treatment assignment, (ii) sampling of covariates, (iii) sampling of potential outcomes given covariates, and possibly (if we take a Bayesian perspective) (iv) prior uncertainty.

Corresponding to these sources of randomness are different possible estimands of interest, as discussed in Imbens (2004). First, there is the (population) average treatment effect

$$ATE = E[Y_i^1 - Y_i^0|\theta]. \tag{4}$$

[3] We could equivalently write $P(\theta|X)$, since we condition on X throughout.

This effect is defined by averaging over (ii) and (iii). It is a function of the population distribution of X_i and of θ. Second there is the conditional average treatment effect,

$$CATE = \beta = \frac{1}{n}\sum_i E[Y_i^1 - Y_i^0|X_i, \theta]. \tag{5}$$

This effect is defined by averaging over (iii) while conditioning on (ii). It is a function of X and of θ. And third there is the sample average treatment effect,

$$SATE = \frac{1}{n}\sum_i (Y_i^1 - Y_i^0). \tag{6}$$

This effect is defined by conditioning on both (ii) and (iii). We have that $E[SATE|X] = CATE$ and $E[CATE] = ATE$. This implies that any estimator that is unbiased for the $SATE$ is also unbiased for the $CATE$ and the ATE, and any estimator that is unbiased for the $CATE$ is also unbiased for the ATE. We focus on the $CATE$ as an estimand because we are interested in experimental design, which only affects what we can learn about the distribution of potential outcomes (iii), conditional on covariates. This stands in contrast to questions of sampling which also affects what we can learn about the population distribution of covariates (ii). Any estimator for the $CATE$ can be rationalized as an estimator of the ATE if we take the sample distribution of X_i as an estimator of its population distribution. This is justified from a Bayesian perspective if we adopt a "non-informative" Dirichlet prior ($\alpha \equiv 0$), as we will briefly discuss in section 5. The randomness in any estimator $\widehat{\beta}$, conditional on population parameters, is driven by (i), (ii), and (iii). Given the conditioning arguments of the three estimands just defined, this implies that any estimation error for the $SATE$ is driven purely by (i), any estimation error for the $CATE$ is driven by (i) and (iii), and any estimation error for the ATE is driven by (i), (ii), and (iii). If we take a Bayesian perspective in setting up the decision problem we are going to discuss, then expected loss averages over (i), (iii), and (iv).

The standard argument for identification in randomized experiments relies on the randomness (i) of treatment assignment, which ensures independence of treatment assignment and potential outcomes on average. Under certain conditions, identification via conditional independence can also be justified using only randomness (iii) in sampling of potential outcomes. We will discuss this in more detail in the next section.

Bayesian and frequentist perspective, and the role of conditioning

We have defined our objective as minimizing Bayesian or minimax risk *conditional* on covariates X and the randomization device U. It turns out that in the Bayesian paradigm this conditioning does not matter, as demonstrated by theorem 1 below. In the frequentist paradigm it is crucial, however, since the frequentist risk function is defined as average loss over replications of the same decision problem, where "same" here is understood as the same values of X, U and θ. Understanding minimax risk as the outcome of a "game" against an adversarial nature which moves second, this means that we allow nature to choose θ as a function of U (in addition to X, $\widehat{\beta}$, and $\mathbf{d}$). In this case, there is no advantage in conditioning actions (treatment assignment) on this randomization device, as shown in theorem 1. If, in contrast, θ could depend on $\mathbf{d}$ but not on U, then it might be optimal to play a mixed strategy involving randomization.

Infinite minimax risk

Minimax risk will in general only be finite if model restrictions are imposed on the set of possible θ. To see this, consider squared loss $L(\widehat{\beta}, \beta) = (\widehat{\beta} - \beta)^2$. Consider a sequence of θ_j such that $P(Y_i|X_i, D_i = d_i, \theta_j)$ does not change along this sequence, but the counterfactual mean $E[Y_i|X_i, D_i = (1-d_i), \theta_j]$ goes to $+\infty$ if $d_i = 0$ and to $-\infty$ if $d_i = 1$. For such a sequence $P(\widehat{\beta}|\theta_j)$ does not depend on j, while $\beta \to \infty$, and thus $R(\mathbf{d}, \widehat{\beta}|X, U, \theta) \to \infty$, which implies $R^{mm}(\mathbf{d}, \widehat{\beta}|X, U) = \infty$. In contrast, Bayesian average risk $R^B(\mathbf{d}, \widehat{\beta})$ is finite for squared loss and linear estimators $\widehat{\beta}$ as long as the prior expectation $E[\beta]$, as well as the prior variance $\mathrm{Var}(Y)$ are finite.

4. OPTIMALITY OF NON-RANDOMIZED DESIGNS

We will now formally state our first result, which implies that optimal experimental designs, *given* knowledge of covariates, generically do not involve randomization. The following argument applies to any estimator $\widehat{\beta}$ and to any loss function, including (but not limited to) estimators of the average treatment effect and squared loss.

The intuition for this result is simple. The risk for a randomized design is equal to the average risk of the treatment assignments that the randomized design averages over. Randomized designs can therefore not improve upon the risk attainable by deterministic assignments, and if the optimal deterministic design is unique, then randomized designs perform strictly worse.

Let $\mathscr{D}$ be the set of all treatment assignment procedures that we consider,

$$\mathscr{D} := \{\mathbf{d} : \ (X, U) \to D\}.$$

Let $\mathscr{D}^{det}$ be the subset of deterministic treatment assignment procedures, procedures which do not depend on U,

$$\mathscr{D}^{det} := \{\mathbf{d} : \ X \to D\}.$$

Theorem 1 (Optimality of deterministic designs)
Consider the setup of assumption 1 and the decision problem of assumption 2. Take the estimator $\widehat{\beta}(Y, X, D)$ as given. Then:

1. *Bayesian average risk is minimized by an element of the set of deterministic treatment assignment procedures,*

$$\min_{\mathbf{d}\in\mathscr{D}^{det}} R^B(\mathbf{d},\widehat{\beta}) = \min_{\mathbf{d}\in\mathscr{D}} R^B(\mathbf{d},\widehat{\beta}). \tag{7}$$

2. *Assume additionally that $R^B(\mathbf{d}^1,\widehat{\beta}|X) - R^B(\mathbf{d}^2,\widehat{\beta}|X)$ is continuously distributed*[4] *for any pair of treatment assignment vectors $\mathbf{d}^1 \neq \mathbf{d}^2$.*
   Then with probability 1

$$\mathbf{d}^* = \underset{\mathbf{d}\in\mathscr{D}}{\text{argmin}}\ R^B(\mathbf{d},\widehat{\beta}) \tag{8}$$

   is unique, and $\mathbf{d}^* \in \mathscr{D}^{det}$.
3. *Similarly, there exists a deterministic treatment assignment that minimizes conditional minimax risk $R^{mm}(\mathbf{d},\widehat{\beta}|X,U)$. Furthermore, if $R^{mm}(\mathbf{d}^1,\widehat{\beta}|X,U) - R^{mm}(\mathbf{d}^2,\widehat{\beta}|X,U)$ is continuously distributed for any pair of treatment assignment vectors $\mathbf{d}^1 \neq \mathbf{d}^2$, then with probability 1 there is a unique optimal assignment $\mathbf{d}^* \in \mathscr{D}^{det}$ which dominates all other $\mathbf{d} \in \mathscr{D}$.*

Proof: Let $\mathbf{d}^*(X) \in \text{argmin}_{\,\mathbf{d}(X)\in\{0,1\}^n} R^B(\mathbf{d},\widehat{\beta}|X,U)$. $\mathbf{d}^*(X)$ is well defined since $R^B(\mathbf{d},\widehat{\beta}|X,U)$ does not depend on U for deterministic assignments.
For any randomized treatment assignment function $\mathbf{d}(X,U)$, we have that

$$R^B(\mathbf{d},\widehat{\beta}|X,U) \geq R(\mathbf{d}^*,\widehat{\beta}|X,U) \tag{9}$$

where the inequality holds by definition of $\mathbf{d}^*$. Integration over $P(X)$ and $P(U)$ yields

$$R^B(\mathbf{d},\widehat{\beta}) \geq R^B(\mathbf{d}^*,\widehat{\beta}).$$

This shows that $\mathbf{d}^*$ minimizes the Bayes risk among all treatment assignment rules.

To show the second claim, note that there are no more than 2^{2n} possible pairs of treatment assignment vectors $\mathbf{d}^1, \mathbf{d}^2$. Therefore

$$P(\exists \mathbf{d}^1,\mathbf{d}^2 :\ R^B(\mathbf{d}^1,\widehat{\beta}|X) = R^B(\mathbf{d}^2,\widehat{\beta}|X)) \leq$$
$$2^{2n} \max_{\mathbf{d}^1,\mathbf{d}^2} P(R^B(\mathbf{d}^1,\widehat{\beta}|X) = R^B(\mathbf{d}^2,\widehat{\beta}|X)) = 2^{2n}\cdot 0.$$

This implies immediately that $\text{argmin}_{\,\mathbf{d}} R^B(\mathbf{d},\widehat{\beta}|X)$ is unique, and that the inequality in (9) is strict for any treatment assignment function $\mathbf{d} \neq \mathbf{d}^*$.

The third claim is immediate once we note that $R^{mm}(\mathbf{d},\widehat{\beta}|X,U)$ does not depend on U, given $\mathbf{d}(X,U)$. This holds because neither does the risk function

[4] This requires that X has a continuously distributed component.

$R(\mathbf{d}, \widehat{\beta}|X, U, \theta)$, nor the risk maximizing θ. □

Remark: The first claim of theorem 1 is analogous to the fact that estimators that do not involve randomization are optimal with respect to Bayes risk, (cf. Robert, 2007, p66).

Remark: If the decision problem is exactly symmetric in the two values of treatment (flipping $d = 1$ and $d = 0$ leaves loss and prior unchanged) then risk is symmetric, $R^B(\mathbf{d}, \widehat{\beta}|X, U) = R^B(1 - \mathbf{d}, \widehat{\beta}|X, U)$. In this case the generic uniqueness of the optimal treatment assignment will only hold up to symmetry in $\mathbf{d}, 1 - \mathbf{d}$. As a consequence one can randomly assign which group gets the treatment and which the control without increasing risk. This might in particular be desirable if there is concern that X could be manipulable by agents aware of the treatment assignment procedure.

Theorem 1 shows that a deterministic treatment assignment is optimal in minimizing either Bayes risk or conditional minimax risk. But what about identification without random assignment? The main appeal of randomization lies in ensuring that the independence condition

$$P(Y_i|D_i = d, \theta) = P(Y_i^d|\theta)$$

holds, and more generally, under stratified randomization, that conditional independence holds,

$$P(Y_i|X_i, D_i = d, \theta) = P(Y_i^d|X_i, \theta). \tag{10}$$

These conditions enable nonparametric identification of average structural functions, quantile structural functions, average treatment effects, etc. The following argument shows that conditional independence (10) holds even without randomization in the treatment assignment. Conditional independence only requires that treatment assignment takes place in a controlled trial (CT), not necessarily in a randomized controlled trial (RCT).[5]

To make the foundations of this claim explicit, we restate the following two assumptions which are implicit in the setup of assumption 1.

Assumption 3 (i.i.d. sampling)

$$P(Y^0, Y^1, X|\theta) = \prod_i P(Y_i^0, Y_i^1, X_i|\theta)$$

The assumption of i.i.d. sampling implies in particular that $P(Y_i^d|X, \theta) = P(Y_i^d|X_i, \theta)$ for $d = 0, 1$. In words, the potential outcomes of unit i are independent of the

[5] Note that at no point we allow for self-selection into treatment!

covariates of other units, conditional on her own covariates.[6]

Assumption 4 (SUTVA) *The potential outcome $Y_i(D)$ depends only on D^i, so that $Y_i = Y_i^d$ if $D_i = d$.*

This "stable unit treatment value assumption" (following the terminology of Angrist et al. 1996) states that there is no causal effect of the treatments of other units on the outcome of a given unit of observation. This assumption excludes social interaction effects.

Theorem 2 (Conditional independence)

If assumptions 1, 3, and 4 hold, and if $D = \mathbf{d}(X, U)$ for some $U \perp (Y^0, Y^1, X)|\theta$, then conditional independence holds, $P(Y_i|X_i, D_i = d, \theta) = P(Y_i^d|X_i, \theta)$. This is true in particular for deterministic treatment assignment rules $D \in \mathscr{D}^{det}$.

Proof: By randomness of U and i.i.d. sampling,

$$P(Y_i^d|X, U, \theta) = P(Y_i^d|X, \theta) = P(Y_i^d|X_i, \theta)$$

for $d = 0, 1$. By the law of iterated expectations, this last equation, and the randomness of U again, and recalling that $D_i = d(X, U)$,

$$P(Y_i^d|X_i, D_i, \theta) = E[P(Y_i^d|X, U, \theta)|X_i, D_i, \theta] = \\ E[P(Y_i^d|X_i, \theta)|X_i, D_i, \theta] = P(Y_i^d|X_i, \theta),$$

where the expectations are taken over the distribution of X and U given X_i and D_i. This concludes the proof. □

Theorem 2 immediately implies the following corollary.

Corollary 1 (Average partial effect and average treatment effect)

If assumptions 3 and 4 hold, and if $D = \mathbf{d}(X, U)$ for some $U \perp (Y^0, Y^1, X)$, then the average partial effect is equal to the average treatment effect,

$$E[E[Y_i|X_i, D_i = 1, \theta] - E[Y_i|X_i, D_i = 0, \theta]] = E[Y_i^1 - Y_i^0|\theta].$$

Similarly, the conditional average partial effect is equal to the conditional average treatment effect,

$$\frac{1}{n}\sum_i [E[Y_i|X_i, D_i = 1, \theta] - E[Y_i|X_i, D_i = 0, \theta]] = \frac{1}{n}\sum_i E[Y_i^1 - Y_i^0|X_i, \theta] = \beta.$$

These statements are true in particular for deterministic treatment assignment rules $D = \mathbf{d}(X)$.

[6] This assumption is weaker than it might seem, as discussed in (Rubin, 1978, section 3.1). In a nutshell: If there is concern about its validity, the data can be randomly permuted to ensure exchangeability; a version of de Finetti's theorem implies then that they are i.i.d. as members of an appropriately defined population. If the position in the dataset is informative, it should furthermore be recorded as an additional covariate. Our optimal design is permutation-equivariant, so that we can actually skip the step of shuffling lines for the purpose of finding the optimal design.

Proof: This is immediate by the conditional independence shown in theorem 2, which implies $E[Y_i|X_i, D_i = d, \theta] = E[Y_i^d|X_i, \theta]$, and the law of iterated expectations. □

Remark: These results show that there is no specific role of randomization to ensure conditional independence and its implications. The important point is that we are not giving up control over treatment assignment by abandoning randomization in favor of a deterministic design; in particular we do not allow for any self-selection.

4.1. *Discussion in the context of the literature*

Discussions about the purpose of randomization have taken place repeatedly in the history of experimental design. In this section we discuss a number of defenses of randomization which have been put forward, and argue that they do not contradict our results.

Stone (1969) defended randomization by arguing that randomization provides a safeguard against *manipulation* of experimental results through experimental design. Stone's argument is interesting, but it requires a few qualifications. (i) This is an argument against *any* form of stratification. (ii) Manipulation is only possible if the researcher analyzing the data ignores the covariates used for stratification. (iii) Randomization is only manipulation-proof if it is verified by a third party - otherwise the experimenter can simply re-randomize until the outcome seems suitable for the desired manipulation. (iv) The optimal designs which we propose can be made manipulation-proof by switching treatment and control group according to a final coin-flip if the decision problem is symmetric.

Rubin (1978) provided a defense of randomization in a Bayesian framework similar to ours, based on the role of randomization for *valid causal inference*, and *practical simplicity.* Rubin's argument is essentially threefold: Randomization (i) ensures ignorability (conditional independence), (ii) leads to balance of covariates, and (iii) leads to designs and estimators which are easy to calculate. If (i) is violated, Bayesian estimators are sensitive to the prior for the treatment assignment mechanism. If (ii) is violated, Bayesian estimators are sensitive to the prior for counterfactual outcomes because of lack of overlapping support. The argument underlying claim (iii) is that under simple randomization we do not need to model the relationship between covariates and potential outcomes, and we do not need to bother searching for balanced assignments. How does this relate to our argument? As regards (i), theorem 2 implies that the assignments we consider are ignorable. As regards (ii), by the very construction of our optimal designs they lead to more balanced assignments, so we are doing better than randomization on that count. As regards (iii), this paper disagrees regarding the practical difficulties of implementing optimal design. Part of the reason is the availability of vastly superior computers than at the time of writing of Rubin (1978), so that the code provided by Kasy (2013) finds optimal design in between a few minutes and a couple of hours. Furthermore, we argue below that researchers *at the estimation stage* do not need to explicitly model the relationship of potential outcomes to covariates. We provide optimal designs for the case that researchers are committed to a simple difference-in-means estimator of the average treatment effect.

Kempthorne (1977), finally, argued in favor of randomization against optimal experimental design, claiming that optimal design requires a strong belief in the correctness of a *parametric model.* Our discussion in the next section shows that this is not true in general, as we develop optimal designs in a nonparametric Bayesian setting.

5. NONPARAMETRIC BAYESIAN OPTIMAL DESIGN AND ESTIMATION

The previous section has discussed the experimental design problem in fairly abstract terms. In this section we will derive explicit and simple expressions for risk R. These expressions provide objective functions for experimental design which can be computationally minimized in order to pick a treatment assignment. These expressions for risk also serve as an operationalization of the notion of balance of treatment groups, as will become evident in the subsection where we consider symmetric priors.

We do not discuss discrete optimization (over $\mathbf{d}$ to minimize R) in detail here, appendix A provides a brief review of some algorithms. Even simple algorithms, such as repeated draws of random assignments and picking the one which has the lowest risk, lead to good results and take only a few minutes to run. Matlab code which implements calculation of risk and discrete optimization is available from the author's homepage at Kasy (2013).

For the rest of the paper we will take a Bayesian perspective, while remaining fully nonparametric. We specialize our setup by assuming that the experimenter's objective is to minimize squared loss, by restricting attention to linear estimators, and by imposing a prior. Under these conditions, only the first two moments of the prior enter the decision problem. Given the treatment assignment D, the best estimator for β is then given by the best linear predictor with respect to the posterior distribution.

Since we are considering Bayesian probabilities now, many of the following expressions will *not* condition on θ, but rather average over the prior distribution for θ, and in particular over the conditional expectation function f defined in the following assumption. Recall also that we assumed θ to be independent of X and U. This implies in particular that observing X does not lead to updated beliefs on $P(Y_i^d|X_i)$.

Assumption 5 (Prior moments) *The experimenter has a prior over the conditional expectation f, where $f(X_i, D_i) = E[Y_i|X_i, D_i, \theta]$, which satisfies*

$$E[f(x,d)] = \mu(x,d)$$

and

$$\operatorname{Cov}(f(x_1,d_1),f(x_2,d_2)) = C((x_1,d_1),(x_2,d_2))$$

for a covariance kernel C.
The experimenter has a prior over the conditional variance of Y_i which satisfies

$$E[\operatorname{Var}(Y_i|X_i,D_i,\theta)|X_i,D_i] = \sigma^2(X_i,D_i).$$

Assumption 6 (Mean squared error objective) *Loss is given by $L(\widehat{\beta},\beta) = (\widehat{\beta}-\beta)^2$, and the experimenter has the objective to minimize expected loss. The experimenter's objective function is thus given by the mean squared difference*

$$R^B(\mathbf{d},\widehat{\beta}|X) = E[(\widehat{\beta}-\beta)^2|X]$$

between an estimator $\widehat{\beta}$ and the conditional average treatment effect $\beta = \frac{1}{n}\sum_i E[Y_i^1 - Y_i^0|X_i,\theta]$.
Assumption 7 (Linear estimators) *$\widehat{\beta}$ is restricted to lie in the class of linear estimators, so that*

$$\widehat{\beta} = w_0 + \sum_i w_i Y_i,$$

where w_i might depend on X and on D, but not on Y.
Remark:
Assumption 5 is fairly unrestrictive, since it only assumes existence of the first two moments of the prior distribution. Assumption 6 specifies that we are interested in minimizing the mean squared error of an estimator of the ATE. Assumption 7 is restrictive in terms of the class of estimators considered. This class contains most common estimators of the average partial effect, however, including estimators based on linear regression, inverse probability weighting, kernel regression, splines, and various forms of matching.
Assumption 5 can be made more specific by assuming a Gaussian process prior for f. This is not necessary in our context, however, since only the first two moments enter the decision problem under assumptions 6 and 7.

We use the following notation for the prior moments of Y and β given X and D.

$$\begin{aligned}
\mu_i &= E[Y_i|X,D] = \mu(X_i,D_i),\\
\mu_\beta &= E[\beta|X,D] = \frac{1}{n}\sum_i[\mu(X_i,1)-\mu(X_i,0)],\\
\Sigma &= E[\operatorname{Var}(Y|X,D,\theta)|X,D] = \operatorname{diag}(\sigma^2(X_i,D_i)),\\
C_{i,j} &= \operatorname{Cov}(f(X_i,D_i),f(X_j,D_j)|X,D) = C((X_i,D_i),(X_j,D_j)), \text{ and}\\
\overline{C}_i &= \operatorname{Cov}(Y_i,\beta|X,D) = \frac{1}{n}\sum_j\left[C((X_i,D_i),(X_j,1)) - C((X_i,D_i),(X_j,0))\right]
\end{aligned}$$

With this notation, $\mathrm{Var}(Y|X,D) = C + \Sigma$.

Now we can formally state the form of the best estimator and Bayesian risk.
Theorem 3 (Best linear predictor, expected loss, and unique optimal design) *Under assumptions 1 through 7, the best estimator for the conditional average treatment effect is given by*

$$\begin{aligned}\widehat{\beta} &= E[\beta|X,D] + \mathrm{Cov}(\beta, Y|X,D) \cdot \mathrm{Var}(Y|X,D)^{-1} \cdot (Y - E[Y|X,D]) \\ &= \mu_\beta + \overline{C}' \cdot (C+\Sigma)^{-1} \cdot (Y - \mu), \qquad (11)\end{aligned}$$

and the corresponding expected loss (risk) equals

$$\begin{aligned}R^B(\mathbf{d}, \widehat{\beta}|X) = E[(\widehat{\beta} - \beta)^2|X,D] &= \mathrm{Var}(\beta|X,D) - \mathrm{Var}(\widehat{\beta}|X,D) \qquad (12)\\ &= \mathrm{Var}(\beta|X) - \overline{C}' \cdot (C+\Sigma)^{-1} \cdot \overline{C}.\end{aligned}$$

Furthermore, if in addition one of the components of X_i is continuously distributed and C is not constant in this component with probability 1*, the risk function of equation* (12) *has a unique minimizer with probability* 1 *conditional on D_1.*
Proof: Under assumptions 6 and 7, the weights of the best estimator solve

$$w^* = \underset{w}{\mathrm{argmin}}\; E[(w_0 + \sum_i w_i Y_i - \beta)^2|X,D].$$

The first order conditions for this problem are

$$E(w_0 + \sum_i w_i Y_i - \beta)|X,D] = E[\widehat{\beta}|X,D] - E[\beta|X,D] = 0$$

and

$$E[Y_i \cdot (w_0 + \sum_i w_i Y_i - \beta)|X,D] = \mathrm{Cov}(\mathrm{Y}_i, \widehat{\beta}|X,D) - \mathrm{Cov}(Y_i, \beta|X,D) = 0$$

for $i = 1 \dots n$. This implies

$$w_0 + \sum_i w_i \mu_i = \beta$$

and

$$\sum w_j \,\mathrm{Cov}(Y_i, Y_j|X,D) = \mathrm{Cov}(Y_i, \beta|X,D),$$

from which equation (11) is immediate.
These first order conditions also immediately imply $\mathrm{Cov}(\widehat{\beta}, \beta - \widehat{\beta}|X,D) = 0$, so that

$$\mathrm{Var}(\beta|X,D) = \mathrm{Var}(\widehat{\beta}|X,D) + \mathrm{Var}(\beta - \widehat{\beta}|X,D),$$

which yields equation (12).
To show uniqueness of the minimizer with probability 1, by theorem 1 it is sufficient to show that

$$\Delta R := R^B(\mathbf{d}^1, \widehat{\beta}|X) - R^B(\mathbf{d}^2, \widehat{\beta}|X) = \overline{C}^{2\prime} \cdot (C^2 + \Sigma)^{-1} \cdot \overline{C}^2 - \overline{C}^{1\prime} \cdot (C^1 + \Sigma)^{-1} \cdot \overline{C}^1$$

is continuously distributed for all $\mathbf{d}^1 \neq \mathbf{d}^2$. Under the given assumptions, continuous distribution holds by continuity of the distribution of C^1, C^2, as long as no symmetry in the setup implies $C^1 = C^2$ and $\overline{C}^1 = \overline{C}^2$. The two sources of symmetry we have to worry about are (i) symmetry between $\mathbf{d}$ and $1 - \mathbf{d}$, and (ii) symmetry by permutation of $\mathbf{d}$ between observations with identical covariate values X_i. Symmetry (i) is taken care of by conditioning on D_1, symmetry (ii) plays no role here since $X_i \neq X_j$ for $i \neq j$ with probability 1 if there is a continuous covariate. □

Remark: The estimator $\widehat{\beta}$ of theorem 3 minimizes the mean squared error of all linear estimators of β. If we were to impose joint normality of the Y_i^d and $f(X_i, d)$ for all i and d, we could drop the a priori requirement of linear estimation and obtain the same estimator. In that case $\widehat{\beta}$ would be equal to the posterior expectation of β.

Remark: If we were to impose a Dirichlet prior with $\alpha \equiv 0$ for the population distribution of X_i, we would obtain the sample distribution of X_i as expected posterior probability distribution. In this case, all the estimators and risk functions derived here for estimation of the conditional average treatment effect $CATE$ would be estimators and risk functions for estimation of the ATE. To see this, note that the only components of the best linear predictor in theorem 3 that would be affected by the ATE as our object of interest are the covariances $\overline{C}_i$. These covariances have to be replaced by

$$\mathrm{Cov}(Y_i, ATE|X, D) = \int \mathrm{Cov}(Y_i, f(x, 1) - f(x, 0)) dP(x|X, D),$$

where P is the posterior distribution of x. For a Dirichlet prior with $\alpha \equiv 0$, $P(x|X, D)$ is equal to the empirical distribution of X_i.

Imposing symmetry

If we impose some additional structure on the problem, more explicit formulas for the estimator and for the expected loss can be derived.
Assumption 8 (Prior homoskedasticity) *The expected variance of Y_i^d given θ and X_i is constant in X_i and d, $E[\mathrm{Var}(Y_i^d|X_i, \theta)|X_i] = \sigma^2$.*
Assumption 9 (Restricting prior moments)

1. *The prior expectation of f is $E[f] = 0$.*
2. *The functions $f(.,0)$ and $f(.,1)$ are uncorrelated.*
3. *The prior distributions of $f(.,0)$ and $f(.,1)$ are the same.*

Under these additional assumptions we can denote $\text{Cov}(f(x_1,d), f(x_2,d)) = C(x_1,x_2)$ independent of d, and get

$$\begin{aligned}
\mu_i &= E[Y_i|X,D] = 0,\\
\mu_\beta &= E[\beta|X,D] = 0,\\
\Sigma &= E[\text{Var}(Y|X,D,\theta)|X,D] = \sigma^2 \cdot I,\\
C_{i,j} &= \text{Cov}(f(X_i,D_i), f(X_j,D_j)|X,D) = 0 \text{ for } D_i \neq D_j,\\
C_{i,j} &= \text{Cov}(f(X_i,D_i), f(X_j,D_j)|X,D) = C(X_i,X_j) \text{ for } D_i = D_j, \text{ and}\\
\overline{C}_i^d &= \text{Cov}(Y_i^d, \beta|X,D) = (-1)^{1-d} \cdot \frac{1}{n}\sum_j C(X_i,X_j).
\end{aligned}$$

We use the following superscript notation for subvectors and submatrices defined by treatment values d.

$$\begin{aligned}
Y^d &= (Y_i : D_i = d)\\
V^d &= \text{Var}(Y^d|X,D) = (C_{i,j} : D_i = d, D_j = d) + \text{diag}(\sigma^2 : D_i = d)\\
\overline{C}^d &= \text{Cov}(Y^d, \beta|X,D) = (\overline{C}_i^d : D_i = d)
\end{aligned}$$

Remark: Note that the covariance $C_{i,j}$ does not depend on treatment assignment D. The variance matrix V^d does, but only through selection of a submatrix by the treatment assignment D. Similar statements hold for the vectors Y^d, and $\overline{C}^d$.

Theorem 4 (Explicit estimator and risk function) *Under assumptions 1 through 9, the best estimator for the conditional average treatment effect is given by*

$$\widehat{\beta} = \overline{C}^{1\prime} \cdot (V^1)^{-1} \cdot Y^1 + \overline{C}^{0\prime} \cdot (V^0)^{-1} \cdot Y^0. \tag{13}$$

and the corresponding expected loss (risk) equals

$$\begin{aligned}
R^B(\mathbf{d}, \widehat{\beta}|X) &= E[(\widehat{\beta} - \beta)^2|X,D] = \text{Var}(\beta|X,D) - \text{Var}(\widehat{\beta}|X,D)\\
&= \text{Var}(\beta|X) - \overline{C}^{1\prime} \cdot (V^1)^{-1} \cdot \overline{C}^1 - \overline{C}^{0\prime} \cdot (V^0)^{-1} \cdot \overline{C}^0.
\end{aligned} \tag{14}$$

Proof: This is immediate from theorem 3, once we note that V is block diagonal, after sorting observations by their value of D_i. □

Insisting on the difference-in-means estimator

One of the main appeals of randomized experiments is their simplicity. Suppose we want to preserve this simplicity of estimation, for reasons outside the formal decision problem. Theorem 5 provides the risk function for treatment assignments under this constraint on estimation.

Assumption 10 (Simple estimator) *The estimator $\widehat{\beta}$ is given by*

$$\widehat{\beta} = \frac{1}{n_1}\sum_i D_i Y_i - \frac{1}{n_0}\sum_i (1-D_i)Y_i,$$

where $n_d = \sum_i \mathbf{1}(D_i = d)$.

This is the estimator we would use when ignoring any information about covariates at the estimation stage.

Theorem 5 (Risk function for designs using the simple estimator) *Under assumptions 1 through 10, the expected loss (risk) of treatment assignment* **d** *equals*

$$R^B(\mathbf{d}, \widehat{\beta}|X) = \sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0}\right] + \left[1 + \left(\frac{n_1}{n_0}\right)^2\right] \cdot v' \cdot \tilde{C} \cdot v,$$

where

$$v_i = \frac{1}{n} \cdot \left(-\frac{n_0}{n_1}\right)^{D_i}$$

and $\tilde{C}_{ij} = C(X_i, X_j)$. In particular, if $n_1 = n_0 = n/2$, then

$$R^B(\mathbf{d}, \widehat{\beta}|X) = \sigma^2 \cdot \frac{4}{n} + \frac{1}{n^2} \cdot (2D-1)' \cdot \tilde{C} \cdot (2D-1).$$

Proof: Let $w^1 = \frac{1}{n_1}D - \frac{1}{n}e$, $w^0 = \frac{1}{n_0}(e-D) - \frac{1}{n}e$, and $w = w^1 - w^0$, so that $\widehat{\beta} = w' \cdot Y$. Under the given assumptions

$$\widehat{\beta} - \beta = w' \cdot (Y - f) + w^{1\prime} \cdot f_1 - w^{0\prime} \cdot f_0,$$

where we use f to denote the n vector with entries $f(X_i, D_i)$, and f_d to denote the n vector with entries $f(X_i, d)$. The three terms of this decomposition have prior mean 0 and are uncorrelated. This implies

$$R^B(\mathbf{d}, \widehat{\beta}|X) = w' \cdot \Sigma \cdot w + w^{1\prime} \cdot \tilde{C} \cdot w^{1\prime} + w^{0\prime} \cdot \tilde{C} \cdot w^{0\prime}.$$

The claims of the theorem then follow, once we note that the first term equals $\sigma^2 \cdot w' \cdot w = \sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0}\right]$, that $w_i^1 = -\frac{1}{n} \cdot \left(-\frac{n_0}{n_1}\right)^{D_i}$, and that $w_i^0 = \left(\frac{n_1}{n_0}\right) \cdot w_i^1$. □

Remark: The expression for risk in theorem 5 has an easily interpretable structure. It decomposes the mean squared error into a variance term $\sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0}\right]$ and an expected squared bias term. The optimal design for this risk minimizes the variance by choosing $n_1 \approx n_0$, and then balancing the distribution of covariates to minimize the expected squared bias of the difference in means estimator.

6. HOW TO CHOOSE A PRIOR

In this section we will discuss some common approaches to choosing prior moments for f. Further background can be found in Williams and Rasmussen (2006, chapters 2 and 4), as well as Wahba (1990, chapter 1). Throughout, we will maintain assumptions 1 through 9. We discuss in particular how to make priors non-informative about key features of the data generating process, and give a general characterization in theorem 6 below.

The next subsection briefly discusses the relationship between Gaussian process regression and penalized regression. We then discuss three classes of priors: (i) Linear models, (ii) priors with a squared exponential covariance kernel, and (iii) priors which combine a general covariance kernel with a linear model where the prior is non-informative about the coefficients of the linear model. Our recommended prior will be a combination of these three, for reasons discussed below. We finally discuss how to choose $E[\sigma^2]$ given a prior for f, based on the expected share of variation in potential outcomes explained by the observed covariates.

Gaussian process regression and penalization

It is useful to note that the best linear predictor $\widehat{f}^d$ for f^d given X, D, Y can be recast as the solution to a penalized regression. Let[7]

$$\begin{aligned} f^d(X_i) &= E[Y_i^d|X_i,\theta] = f(X_i,d) \\ f^d &= (f^d(X_i) : i = 1,\dots,n). \end{aligned}$$

Under assumptions 1 through 9,

$$\widehat{f}^d = \underset{f^d}{\operatorname{argmin}} \; \frac{1}{\sigma^2} \sum_{i:D_i=d} (Y^i - f^d(X_i))^2 + \|f^d\|_d^2, \tag{15}$$

with

$$\|f^d\|_d^2 := f^{d\prime} \cdot \operatorname{Var}^{-1}(f^d) \cdot f^d = f^{d\prime} \cdot C^{-1} \cdot f^d$$

As far as estimation of f is concerned, the choice of a prior thus amounts to the choice of a penalty functional (seminorm) $\|f^d\|_d^2$. The corresponding estimator of the conditional average treatment effect is then given by

$$\widehat{\beta} = \frac{1}{n} \sum_i [\widehat{f}^1(X_i) - \widehat{f}^0(X_i)].$$

[7] Note that f^d is *not* a subvector defined by d, in contrast to Y^d.

Linear models

For an appropriate definition of X_i which might include interactions, powers, transformations etc., assume that

$$\begin{aligned}
Y_i^d &= X_i\beta^d + \epsilon_i^d \\
E[\beta^d|X] &= 0 \\
\mathrm{Var}(\beta^d|X) &= \Sigma_\beta \\
\mathrm{Var}(\epsilon^d|X, \beta^d) &= \sigma^2 I.
\end{aligned}$$

In our previous notation, we have $f(X_i, d) = X_i\beta^d$. This implies $C(x_1, x_2) = x_1' \cdot \Sigma_\beta \cdot x_2$, and thus

$$C = X\Sigma_\beta X'.$$

In this setup, the posterior expectation of β^d is given by the solution to the penalized regression

$$\widehat{\beta}^d = \underset{\beta^d}{\operatorname{argmin}} \; \frac{1}{\sigma^2} \sum_{i:D_i=d} (Y^i - X_i'\beta^d)^2 + \|\beta^d\|_d^2,$$

where $\|\beta^d\|_d^2 = \beta^{d\prime} \cdot \Sigma_\beta^{-1} \cdot \beta^d$. The solution to this penalized regression[8] is given by

$$\widehat{\beta}^d = \left(X^{d\prime}X^d + \sigma^2\Sigma_\beta^{-1}\right)^{-1} X^{d\prime}Y^d,$$

where as before X^d and Y^d denote the appropriate submatrix and vector. This implies

$$\mathrm{Var}(\widehat{\beta}^d - \beta^d|X, D) = \left(\frac{1}{\sigma^2}X^{d\prime}X^d + \Sigma_\beta^{-1}\right)^{-1}.$$

We get, finally, that the posterior expectation of the conditional average treatment effect is given by

$$\widehat{\beta} = \overline{X}\left(\widehat{\beta}^1 - \widehat{\beta}^0\right),$$

where $\overline{X} = \frac{1}{n}\sum_i X_i$, implying

$$\begin{aligned}
(16) \qquad R(\mathbf{d}, \widehat{\beta}|X) &= E\left[(\widehat{\beta} - \beta)^2|X, D\right] \\
&= \overline{X} \cdot \left(\mathrm{Var}(\widehat{\beta}^1 - \beta^1|X, D) + \mathrm{Var}(\widehat{\beta}^0 - \beta^d|X, D)\right) \cdot \overline{X}' \\
&= \sigma^2 \cdot \overline{X} \cdot \left(\left(X^{1\prime}X^1 + \sigma^2\Sigma_\beta^{-1}\right)^{-1} + \left(X^{0\prime}X^0 + \sigma^2\Sigma_\beta^{-1}\right)^{-1}\right) \cdot \overline{X}'.
\end{aligned}$$

[8] This type of regression is also known as "ridge regression," and the method of penalization is called "Tikhonov regularization" in some contexts (cf. Carrasco et al., 2007).

This is the objective function we want to minimize through choice of the design D, which enters this expression through the matrices

$$X^{d\prime}X^d = \sum_i \mathbf{1}(D_i = d)X_iX_i'.$$

Note also that the "non-informative" limit $\Sigma_\beta^{-1} \to 0$ has a particularly nice interpretation here: it implies that the $\widehat{\beta}^d$ and thus $\widehat{\beta}$ are given by simple OLS regression. The risk in this case is equal to the standard OLS variance of $\widehat{\beta}$.

Squared exponential covariance function

A common choice of prior in the machine learning literature (cf. Williams and Rasmussen, 2006) is defined by the covariance kernel

$$C(x_1, x_2) = \exp\left(-\frac{1}{2l^2}\|x_1 - x_2\|^2\right), \tag{17}$$

where $\|.\|$ is some appropriately defined norm measuring the distance between covariate vectors. The parameter l determines the length scale of the process. Figure 1 shows draws from such a process for $x \in [-1, 1]$ and various length scales l.

This prior does not restrict functional form and can accommodate any shape of f^d. In this sense it is a nonparametric prior. One attractive feature of the squared exponential covariance kernel is that is puts all its mass on smooth functions, in the sense that f^d is infinitely mean-square differentiable. A function is mean-square differentiable if the normalized differences of f converge in L^2 to some function $\partial f(x)/\partial x$,

$$\frac{f(x+\epsilon) - f(x)}{\|\epsilon\|} \to^{L^2} \frac{\partial f(x)}{\partial x}$$

as $\|\epsilon\| \to 0$, cf. Williams and Rasmussen (2006, p81). Infinite mean square differentiability holds for all processes that have a covariance kernel C which is infinitely differentiable around points where $x_1 = x_2$.

The length scale l, and more generally the norm $\|x_1 - x_2\|$, determines the smoothness of the process, where larger length scales correspond to smoother processes. One measure of smoothness are the expected number of "upcrossings" at 0, i.e., the expected number of times the process crosses 0 from below in the interval $[0, 1]$. For a one-dimensional process with squared exponential kernel, this number equals $1/(2\pi l)$, cf. again Williams and Rasmussen (2006, p81).

Noninformativeness

Researchers might rightly be concerned if experimental estimates for parameters such as average treatment effects are driven by prior information. This suggests to consider priors which are "non-informative" about the parameters of interest, while at the same time using our prior assumptions about smoothness of the underlying functions f^d.[9] One way to formalize such non-informativeness is to consider limit cases where the prior variance for the parameter of interest goes to infinity, and to use the corresponding limit estimators and implied objective functions for experimental design.

In particular, given a covariance kernel K^d for a stochastic process g^d as well as a subset of regressors x_1, consider the process

$$\begin{aligned}
Y_i^d &= g^d(X_i) + X_{1,i}\beta^d + \epsilon_i^d \\
E[g] &= 0 \\
E[\beta^d|X] &= 0 \\
E[\epsilon] &= 0 \\
\text{Cov}(g^d(x_1), g^d(x_2)) &= K(x_1, x_2) \\
\text{Var}(\beta^d|X) &= \lambda\Sigma_\beta \\
\text{Var}(\epsilon^d|X, \beta^d) &= \sigma^2 I \\
\beta^d &\perp g^d \perp \epsilon.
\end{aligned}$$

For this process we get

$$C^d = K^d + \lambda X_1^d \Sigma_\beta X_1^{d\prime},$$

where the superscript d again denotes the appropriate submatrices. We will be interested in particular in the case $\lambda \to \infty$, where the prior over β^d becomes non-informative. Let $\overline{g}^d = \frac{1}{n}\sum_i g(X_i)$, $\overline{f}^d = \overline{g}^d + \overline{X}\beta^d$, $K_y^d = K^d + \sigma^2 I$, and $\overline{K}^d = \text{Cov}(Y^d, \overline{g}^d|X, D)$.[10]

Theorem 6 (BLP and MSE for partially non-informative priors)

For this model, the best linear predictor $\widehat{\beta}$ is equal to $\widehat{\beta}_\infty = \widehat{\overline{f}}^1_\infty - \widehat{\overline{f}}^0_\infty$ up to a remainder of order $O(1/\lambda)$ as $\lambda \to \infty$, given X, D and Y, where

$$\widehat{\overline{f}}^d_\infty = \overline{X}_1\widehat{\beta}^d_\infty + \overline{K}^d K_y^{d,-1}(Y^d - X_1^d\widehat{\beta}^d_\infty) \tag{18}$$

and

$$\widehat{\beta}^d_\infty = \left(X^{d\prime}K_y^{d,-1}X^d\right)^{-1} X^{d\prime}K_y^{d,-1}Y^d. \tag{19}$$

[9] And note that *any* nonparametric estimation method has to use assumptions about smoothness!

[10] Results somewhat similar to the following theorem have been shown by O'Hagan and Kingman (1978), as well as by Wahba (1990, p19).

For any λ, we have

$$\begin{aligned}\overline{f}^d - \widehat{\overline{f}}^d_\infty &= \overline{g}^d - \overline{K}^d K_y^{d,-1}(g+\epsilon)\\ &\quad - (\overline{X} - \overline{K}^d K_y^{d,-1}X)\cdot\left(X^{d\prime}K_y^{d,-1}X^d\right)^{-1} X^{d\prime}K_y^{d,-1}(g+\epsilon)\end{aligned}$$

and

$$R(\mathbf{d},\widehat{\beta}_\infty|X) = \mathrm{Var}(\overline{f}^1 - \widehat{\overline{f}}^1_\infty|X) + \mathrm{Var}(\overline{f}^0 - \widehat{\overline{f}}^0_\infty|X) \tag{20}$$

where

$$\begin{aligned}\mathrm{Var}(\overline{f}^d - \widehat{\overline{f}}^d_\infty|X) &= \mathrm{Var}(\overline{g}^d) - \overline{K}^d K_y^{d,-1}\overline{K}^d\\ &\quad + (\overline{X} - \overline{K}^d K_y^{d,-1}X)\cdot\left(X^{d\prime}K_y^{d,-1}X^d\right)^{-1}(\overline{X} - \overline{K}^d K_y^{d,-1}X)'.\end{aligned} \tag{21}$$

Proof: All moments in this proof implicitly condition on X and D. To show the first claim, let $h^d = X^d\beta^d$, so that $Y^d = g^d + h^d + \epsilon^d$ and $\mathrm{Var}(Y^d) = \mathrm{Var}(g^d) + \mathrm{Var}(h^d) + \mathrm{Var}(\epsilon^d)$. The best linear predictor for $\overline{f}^d$ is given by

$$\begin{aligned}\widehat{\overline{f}}^d &= \mathrm{Cov}(\overline{f}^d, Y^d)\,\mathrm{Var}(Y^d)^{-1}Y^d\\ &= \mathrm{Cov}(\overline{h}^d, Y^d)\,\mathrm{Var}(Y^d)^{-1}Y + \mathrm{Cov}(\overline{g}^d, Y^d)\,\mathrm{Var}(Y^d)^{-1}Y^d\end{aligned}$$

Straightforward algebra shows that

$$\mathrm{Var}(Y^d)^{-1} = \left(\mathrm{Var}(g^d) + \mathrm{Var}(\epsilon^d)\right)^{-1}\left(I - \mathrm{Var}(h^d)\,\mathrm{Var}(Y^d)^{-1}\right)$$

so that

$$\begin{aligned}&\mathrm{Cov}(\overline{g}^d, Y^d)\,\mathrm{Var}(Y^d)^{-1}Y =\\ &Cov(\overline{g}^d, Y^d)\left(\mathrm{Var}(g^d) + \mathrm{Var}(\epsilon^d)\right)^{-1}\left(Y^d - \mathrm{Cov}(h^d, Y^d)\,\mathrm{Var}(Y^d)^{-1}Y^d\right).\end{aligned}$$

This proves the decomposition

$$\widehat{\overline{f}}^d = \overline{X}\widehat{\beta}^d + \overline{K}^d K_y^{d,-1}(Y^d - X_1^d\widehat{\beta}^d),$$

where $\widehat{h}^d = X_1^d\widehat{\beta}^d$ is given by

$$\widehat{\beta}^d = \left(X^{d\prime}K_y^{d,-1}X^d + \frac{1}{\lambda}\Sigma_\beta^{d-1}\right)^{-1} X^{d\prime}K_y^{d,-1}Y.$$

This is the penalized GLS estimator. To see this latter equality, note that after pre-multiplying X and Y by $K_y^{d,-1/2}$, this model satisfies the assumptions of the

linear model considered above. The limiting estimators $\widehat{\overline{f}}^d_\infty$ and $\widehat{\beta}^d_\infty$, as well as the form of $\overline{f}^d - \widehat{\overline{f}}^d_\infty$ now follow immediately.
It remains to derive $R(\mathbf{d}, \widehat{\beta}_\infty|X)$. From the model where we had $Y^d = g^d + \epsilon^d$ we know that

$$\mathrm{Var}(\overline{g}^d - \overline{K}^d K_y^{d,-1}(g+\epsilon)) = \mathrm{Var}(\overline{g}^d) - \overline{K}^d K_y^{d,-1}\overline{K}^d.$$

We furthermore know, by the properties of best linear predictors, that

$$\mathrm{Cov}(\overline{g}^d - \overline{K}^d K_y^{d,-1}(g+\epsilon), (g+\epsilon)) = 0.$$

These considerations and some algebra immediately yield $\mathrm{Var}(\overline{f}^1 - \widehat{\overline{f}}^d_\infty)$. □

Remark: Note that the limiting estimator of theorem 6 can be understood as penalized regression, where the penalization corresponds to the seminorm

$$\|f^d\|^2 = \min_{\widehat{\beta}}(f^d - X^d\widehat{\beta})' \cdot K_y^{d,-1} \cdot (f^d - X^d\widehat{\beta}). \tag{22}$$

This is the squared $K_y^{d,-1}$ norm of the projection of f^d onto the orthocomplement of X^d with respect to the $K_y^{d,-1}$ inner product.

Remark: Note also that the risk function $R(\mathbf{d}, \widehat{\beta}_\infty|X)$ is given by the risk function for the model without the term $X\beta^d$, plus a "correction term" of the form

$$(\overline{X} - \overline{K}^d K_y^{d,-1}X) \cdot \left(X^{d\prime} K_y^{d,-1} X^d\right)^{-1} (\overline{X} - \overline{K}^d K_y^{d,-1}X)'$$

for $d = 1, 2$.

Choice of σ^2

For all models considered above, we have to choose σ^2. A tractable way of doing so is through picking the expected share of variation in the outcome data which is explained by the covariates given θ, for a given treatment level. Under assumptions 1 through 9, this share is given by

$$R^2 = \frac{1}{1 + \sigma^2/\,\mathrm{Var}(f^d(X_i)|X,\theta)},$$

so that

$$\sigma^2 = \frac{1-R^2}{R^2}\,\mathrm{Var}(f^d(X_i)|X,\theta).$$

Here $\mathrm{Var}(f^d(X_i)|X,\theta) = f^{d\prime} M f^d/n$ is the sample variance of f^d, with M defined as the projection matrix $M = I - ee'/n$ and $e = (1,\ldots,1)'$. This implies

$$\begin{aligned} E[\mathrm{Var}(f^d(X_i)|X,\theta)|X] &= E[\mathrm{tr}(f^{d\prime} M f^d/n)|X] = \mathrm{tr}(M\cdot E[f^d f^{d\prime}|X])/n \\ &= \mathrm{tr}(M\cdot C)/n = (\mathrm{tr}\, C - e'\overline{C}/n)/n. \end{aligned}$$

This suggests picking σ^2 corresponding to the prior beliefs regarding R^2, i.e.,

$$\sigma^2 = E\left[\frac{1-R^2}{R^2}\right]\cdot(\mathrm{tr}\, C - e'\overline{C}/n)/n.$$

For the case of stationary covariance functions this simplifies further, since in that case $\mathrm{tr}(C)/n = C_{ii}$ for all i. Note also that this formula remains unchanged if we make the prior non-informative about $\overline{f}^d$.

We conclude this section by summarizing our suggested prior.

Suggested prior

1. Normalize the variance of all covariates to 1.
2. Let $K(x_1,x_2) = \exp\left(-\frac{1}{2}\|x_1 - x_2\|^2\right)$ where $\|.\|$ is the Euclidian norm.
3. Take $\sigma^2 = \frac{1-R^2}{R^2}\cdot(\mathrm{tr}\, K - e'\overline{K}/n)/n$, based on your best guess for R^2.
4. Consider the non-informative limit, w.r.t. $\mathrm{Var}(\beta^d)$, of the model

$$Y_i^d = \beta^d + g^d(X_i) + \epsilon_i^d,$$

where g^d is distributed according to the covariance kernel K.

According to theorem 6, this prior implies a best linear predictor for β of

$$\widehat{\beta}^1_\infty - \widehat{\beta}^0_\infty + \overline{K}^1 K_y^{1,-1}(Y^1 - e\widehat{\beta}^1_\infty) - \overline{K}^0 K_y^{0,-1}(Y^0 - e\widehat{\beta}^0_\infty) \tag{23}$$

where

$$\widehat{\beta}^d_\infty = \left(e' K_y^{d,-1} e\right)^{-1} e' K_y^{d,-1} Y^d. \tag{24}$$

is a weighted average of the observations for treatment d. The expected mean squared error equals

$$\begin{aligned} \mathrm{Var}(\beta|X,D,Y) = {} & \mathrm{Var}(\overline{g}^1|X) + \mathrm{Var}(\overline{g}^0|X) - \overline{K}^1 K_y^{1,-1}\overline{K}^1 - \overline{K}^0 K_y^{0,-1}\overline{K}^0 \\ & + (1 - \overline{K}^1 K_y^{1,-1} e)\cdot\left(e' K_y^{d,-1} e\right)^{-1}(1 - \overline{K}^1 K_y^{1,-1} e)' \\ & + (1 - \overline{K}^0 K_y^{0,-1} e)\cdot\left(e' K_y^{0,-1} e\right)^{-1}(1 - \overline{K}^0 K_y^{0,-1} e)'. \end{aligned} \tag{25}$$

Possible modifications:

1. Change the length scale for variables that are expected to have a more nonlinear impact by multiplying these variables by 2.
2. Make the prior non-informative about the slopes of some or all covariates; cf. theorem 6.

7. FREQUENTIST INFERENCE

This paper focuses on a decision theoretic, Bayesian foundation for experimental design. We are not bound to analyze the resulting experimental data within the Bayesian paradigm, however. In order to perform conventional, frequentist inference on average treatment effects, we need a (consistent) estimator for the variance of $\widehat{\beta}$ given X, D and the data generating process θ. For linear estimators as in assumption 7 ($\widehat{\beta} = w_0 + \sum_i w_i Y_i$), this variance is given by

$$V := \mathrm{Var}(\widehat{\beta}|X, D, \theta) = \sum w_i^2 \sigma_i^2, \tag{26}$$

where $\sigma_i^2 = \mathrm{Var}(Y_i|X_i, D_i)$. As in any method for inference on average partial effects we thus have to find an estimator of the conditional variance σ_i^2; see for instance the discussion in Imbens (2004).

A natural proposal in our setting would be to first estimate $\epsilon_i = Y_i - f_i$ by the regression residual $\widehat{\epsilon}_i = Y_i - \widehat{f}_i$, where $\widehat{f} = C \cdot (C + \Sigma)^{-1} \cdot Y$ is the nonparametric Bayes estimator of f. We can then estimate V by

$$\widehat{V} := \sum w_i^2 \widehat{\epsilon}_i^2. \tag{27}$$

This estimator of the variance V has a form similar to the standard heteroskedasticity robust estimators of the variance of linear OLS coefficients.

The following proposition gives an asymptotic justification of equation (27) by providing sufficient conditions for consistency of this variance estimator, i.e., for $\widehat{V}/V \to^p 1$.

Proposition 1 *Assume that*

1. $1/M < E[\epsilon_i^2] < M < \infty$ *and* $E[\epsilon_i^4] \leq M < \infty$ *for all* i *and for some constant* M*.*
2. $Y_i \perp Y_j | X, D, \theta$ *(this follows from i.i.d. sampling)*
3. $W_n := \frac{\max_i w_i^2}{\sum_i w_i^2} = o_p(n^{-1/2})$
4. $\|\widehat{f} - f\|_n = o_p\left(n^{-1} \cdot W_n^{-1}\right)$*, where* $\|.\|_n$ *is the* L^2 *norm w.r.t.* $P_n(x)$*,* $\|g\|_n^2 = \int g^2(x) dP_n(x)$*.*

Then $\widehat{V}/V \to^p 1$*.*

Proof: We can decompose

$$\frac{\widehat{V}}{V} = 1 + \frac{\sum w_i^2[\epsilon_i^2 - \sigma_i^2]}{\sum w_i^2 \sigma_i^2} + \frac{\sum w_i^2[\widehat{\epsilon}_i^2 - \epsilon_i^2]}{\sum w_i^2 \sigma_i^2} = 1 + R_1 + R_2.$$

We have to show $R_1 \to^p 0$ and $R_2 \to^p 0$.

Note that $E[R_1|X, D] = 0$ and, under the assumption of bounded 2nd and 4th moments,

$$\mathrm{Var}(R_1|X, D) = \frac{\sum w_i^4 \,\mathrm{Var}(\epsilon_i^2)}{\left(\sum w_i^2 \sigma_i^2\right)^2} \leq M^2 \cdot \sum \left(\frac{w_i^2}{\sum w_i^2}\right)^2 \leq M^2 \cdot n \cdot W_n^2$$

and thus $\mathrm{Var}(R_1|X, D) \to 0$ under the assumption on the large n behavior of W_n, which implies $R_1 \to^p 0$.

For the second term, note that $\widehat{\epsilon}_i = \epsilon_i + f_i - \widehat{f}_i$, therefore $\widehat{\epsilon}_i^2 - \epsilon_i^2 = (f_i - \widehat{f}_i)^2 + 2\epsilon_i(f_i - \widehat{f}_i)$, and thus

$$\begin{aligned} R_2 &= \frac{\sum w_i^2 (f_i - \widehat{f}_i)^2}{\sum w_i^2 \sigma_i^2} + \frac{\sum w_i^2 2\epsilon_i (f_i - \widehat{f}_i)}{\sum w_i^2 \sigma_i^2} \\ &\leq M \cdot n \cdot W_n \cdot \left[\|\widehat{f} - f\|_n^2 + 2\|\epsilon\|_n \cdot \|\widehat{f} - f\|_n \right]. \end{aligned}$$

It is easy to see that $\|\epsilon\|_n = O_p(1)$. We get that $R_2 = O_p(n \cdot W_n \cdot \|\widehat{f} - f\|_n)$, and thus $R_2 \to 0$ under our assumption on the rate of convergence of $\widehat{f}$. □

Remark: Conditions 1 and 2 of proposition 1 are standard regularity conditions. Condition 3 regards the large sample behavior of the maximum of the squared weights w_i^2, relative to their average. This ratio should not grow at a rate faster than $\sqrt{n}$. To show this is the case under various conditions, we can refer to the approximation of these weights by the so called “equivalent kernel;” (cf. Silverman, 1984; Sollich and Williams, 2005; Williams and Rasmussen, 2006). Condition 4 regards the consistency and rate of convergence of $\widehat{f}$. Primitive conditions for condition 4 to hold follow from theorem 3.3 of van der Vaart and van Zanten (2008a). In general, the rate of convergence of $\widehat{f}$ to the true f depends on (i) the position of f relative to the reproducing kernel Hilbert space of the Gaussian process corresponding to the covariance kernel C, and (ii) the so called “small ball probabilities” of this same process. For details, the interested reader is referred to van der Vaart and van Zanten (2008b) and van der Vaart and van Zanten (2008a).

8. MONTE CARLO SIMULATIONS AND APPLICATION

In this section we provide evidence on the potential gains achievable by using the optimal experimental design, rather than a randomized design.

Monte Carlo simulations

Let us first discuss a series of simulation results. We consider covariate vectors X drawn from a (multivariate) standard normal distribution, where X is of dimension 1 or 5, and sample sizes range from 50 to 800. The prior expected residual variance σ^2 equals 4 or 1, which corresponds to an expected R^2 of around .2 or .5 for the squared exponential kernel prior. The priors considered are those of section 6; the linear model prior (with a prior variance of slope and intercept of 1000), the squared exponential prior, and the squared exponential prior made uninformative about intercepts, both with a length scale $l = 1$.

Table I shows the average risk (expected mean squared error) $R^B(\mathbf{d}, \widehat{\beta}|X)$ of randomized designs, relative to the risk of optimal designs which are chosen to minimize R^B. As can be seen from this table, the gains of optimal designs decrease in sample size, increase in the dimension of covariates, and increase in the expected R^2 (where the latter is decreasing in σ^2 for a given prior variance of f). Furthermore, the gains from optimization are rather modest for some parameter configurations, while going up to 20% for some of the parameter values considered, equivalent to a 20% increase in sample size.

A possible concern in applied work might be the sensitivity of the optimal design to the choice of prior. To provide some evidence on sensitivity to the prior, table II compares the mean squared error under pairs of priors A and B as follows. The entries of table II shows the average of (i) the ratio of the posterior mean squared error under prior A for the design which is optimal under prior B to the mean squared error under prior A for the design which is optimal under prior A, and (ii) the same ratio, with the role of A and B reversed. The data generating processes considered are the same as in table I. The pairs of priors considered are A the same as the priors in table I and B the corresponding priors which result from a rescaling of the variable X_1 by a factor of 2. Such a rescaling increases the importance of balancing the distribution of X_1 relative to the importance of balancing other components of X. As can be seen from this table, the choice of prior does matter for the choice of an optimal treatment assignment – it makes a difference whether we aim to balance variable X_1 or variable X_2, say. That said, the designs which are optimal under prior B still perform significantly better than random assignments when evaluated under prior A, as can be seen when comparing the relative mean squared errors in table II to those in table I.

Figure 3 illustrates an optimal design for a two-dimensional covariate vector X_i, where the X_i are drawn from a sample of size 50 of i.i.d. standard normals. The picture illustrates how the optimal design aims for covariate balance, where balance is defined in terms of minimizing $R^B(\mathbf{d}, \widehat{\beta}|X)$. In comparison, random designs, like the one shown on the right hand side, can be locally imbalanced, leading to imprecise estimates of conditional average treatment effects in the imbalanced regions. This matters in particular in regions of higher covariate density.

Project STAR

To illustrate the quantitative relevance of optimal design in real experiments, we evaluate the efficiency gains that would have been feasible in the well-known "Project STAR" experiment. Data from this experiment have been used by many contributions to the economics of education literature, such as Krueger (1999) and Graham (2008).

The project STAR experiment ("Student/Teacher Achievement Ratio") was conducted in Tennessee beginning in 1985. In this experiment kindergarten students and teachers were randomly assigned to one of three groups beginning in the

1985-1986 school year: small classes (13-17 students per teacher), regular-size classes (22-25 students), and regular-size classes with a full-time teacher aide. Students were supposed to remain in the same class type for 4 years. The dataset covers students from 80 schools. Each participating school was required to have at least one of each class type, and random assignment took place within schools. Kindergarten attendance was not mandatory in Tennessee during this time; Students entering a participating school in grade one or later were randomly assigned to a class upon entry.

We use all data on students for whom demographic covariates as well as class type are observed for kindergarten or first grade in a project STAR school, which leaves 8590 students in the sample. Table III shows the means of some covariates for this sample. Here "free lunch" equals 1 for students receiving a free lunch in the first year they are observed in the sample (which is a proxy for coming from a poor family), and "birth date" equals year plus 0.25 times quarter of birth. We pool classes with and without aide, and only consider the assignment of class-size.[11] Assigned treatment D equals 1 for students assigned to a small class upon first entering a project STAR school. For the purposes of this paper we ignore issues of compliance (i.e., students switching class type over time), and consider designs which minimize the mean squared error of estimated intention to treat (ITT) effects.

The risk of a treatment assignment (expected mean squared error), and the corresponding optimal treatment assignment, are calculated as follows. We use the covariates sex, race, year and quarter of birth, free lunch received, and school ID. This stands in contrast to the original design, which stratified only based on school ID. We use a squared exponential prior, made non-informative about the level of treatment effects, as discussed in section 6. More specifically, we use the covariance kernel $K(x_1, x_2) = \exp\left(-\frac{1}{2}\|x_1 - x_2\|^2\right)$ where $\|.\|$ is the Euclidian norm, $x = 4 \cdot (\text{sex}, \text{race}, \text{birth date}, 2 \cdot \text{poor}, 2 \cdot \text{school})$, and school is a vector of school dummies. The optimal treatment assignment is chosen to (approximately) minimize the risk R^B corresponding to this prior and squared loss. We compare R^B for the actual assignment to R^B for the optimal assignment. To allow for a fair comparison of optimal and actual treatment assignment, we have to make sure that our counterfactual assignment respects the same budget constraint as the actual assignment, i.e., assigning at most 2274 out of the 8590 students to small classes. Without this constraint, we would find that improvements of mean squared error of more than 35% would have been possible, assigning around 50% of students to treatment in a balanced way.

Respecting the budget constraint, we still find that improvements of around 19% would have been feasible. This would be equivalent to an increase of sample size of about 19%, or 1632 students.

Table IV shows covariate means for some selected schools. For these schools, ran-

[11] This is done for ease of exposition. All of our arguments would go through if we allow for more than 2 treatment values, and consider minimizing a weighted average of the MSE of several treatment effects.

domization did not lead to a balanced distribution of covariates across large and small classes. In “school 16,” students in small classes were 12% more female and about four months older than students in large classes. In “school 38,” students in small classes were 15% more female, 2 months older, and much less likely to be poor (53% less students receiving free lunch) than students in large classes. In contrast to such imbalances, which are a consequence of randomization, the optimal design aims to equalize the entire distribution of covariates, and thus leads to similar means.

9. CONCLUSION

In this paper we discuss the question how information from baseline covariates should be used when assigning treatment in an experimental trial. In order to give a well grounded answer to this question we adopt a decision theoretic and non-parametric framework. The non-parametric perspective and the consideration of continuous covariates distinguish this paper from much of the previous experimental design literature. For solving the decision problem, we suggest adopting non-parametric Bayesian priors which are non-informative about the key feature of interest, the (conditional) average treatment effect.

A number of conclusions emerge from our analysis. First, randomization is in general not optimal. Rather, we should pick a risk-minimizing treatment assignment, which is generically unique in the presence of continuous covariates. Second, conditional independence between potential outcomes and treatments given covariates does not require random assignment. It is ensured by conducting *controlled* trials, and does not rely on *randomized* controlled trials. Third, there is a class of non-parametric priors - Gaussian Process priors for the conditional expectation of potential outcomes given covariates - which lead to very tractable estimators and Bayesian risk (mean squared error). The general form of risk for such priors, and squared error loss, is $\mathrm{Var}(\beta|X) - \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}$, where $\overline{C}$ and C are the appropriate covariance vector and matrix from the prior distribution, cf. section 5. We suggest to pick the treatment assignment which minimizes this prior risk. In applying this method in Monte Carlo simulations and to the project STAR data, we find gains in mean squared error, relative to random assignment, which range from moderate to sizable, where the gains are decreasing in sample size, increasing in the dimension of covariates, and increasing in the explanatory power of covariates. Gains of 20% seem realistic in many cases, corresponding to the expected gains if we were to increase sample size by 20%.

A possible objection to the practical feasibility of optimal designs might be a political perception that randomized assignments are fair, while assignment based on covariates such as sex or race is not fair, in a similar way that “tagging” in taxation is considered unfair.[12] Note however, that optimal designs seek to

[12] I thank Larry Katz for pointing this out.

balance the covariate distribution across treatments, leading to a *more* equitable distribution of treatment across demographic or other groups, relative to random assignments.

Department of Economics, Harvard University, and IHS Vienna.

APPENDIX A: DISCRETE OPTIMIZATION ALGORITHMS

The main computational challenge in implementing the approach proposed in this paper lies in solving the discrete optimization problem

$$\mathbf{d}^* \in \underset{\mathbf{d}\in\{0,1\}^n}{\operatorname{argmin}} \; R^B(\mathbf{d}, \widehat{\beta}|X).$$

For the setting considered in section 5, the risk $R^B(\mathbf{d}, \widehat{\beta}|X)$ involves inverting matrices of size of order $n \times n$ since $R^B(\mathbf{d}, \widehat{\beta}|X) = \operatorname{Var}(\beta|X) - \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}$. This is in itself not too hard, but it takes long enough to render infeasible brute-force enumeration and evaluation at all 2^n possible designs $\mathbf{d} \in \{0,1\}^n$.

There is a growing literature in applied mathematics and computer science which studies algorithms for discrete optimization problems such as this one. An overview of this literature is beyond the scope of the present paper. In our Matlab implementation (available at Kasy 2013) we use a combination of the following three algorithms.

(1) *Random search:* Draw m vectors $\mathbf{d}$ uniformly at random from the set $\{0,1\}^n$ (or an appropriately budget-constrained set), and take $\widehat{\mathbf{d}}^*$ to be the risk minimizing among these m vectors. If we choose $m = \frac{\log(1-p)}{\log(1-q)}$, then $\widehat{\mathbf{d}}^*$ will be among the best $2^n \cdot q$ of possible treatment assignments with probability p.

(2) *Greedy algorithm:* Pick a starting point $\mathbf{d}_1$. Find the vector $\mathbf{d}$ in a neighborhood of $\mathbf{d}_1$ which minimizes $R^B(\mathbf{d}, \widehat{\beta}|X)$. Here we understand "neighborhood" to mean the set of vectors which differ in at most k components. Take the minimizer from this neighborhood as your new starting point, and iterate. Keep iterating until either a local minimum is found, or a prespecified number of iterations is reached. In our applications, the greedy algorithm performed very well in terms of finding actual minima, but it also was quite slow.

(3) *Simulated annealing:* This is one of the most popular algorithms for discrete optimization and was introduced by Kirkpatrick et al. (1983). The intuition and name for this method came from a method in metallurgy where a metal is made to crystalize in a controlled manner by heating and controlled cooling. The algorithm uses noisy perturbations to a greedy search, to avoid getting stuck in a local minimum. The noise is reduced in later iterations so the algorithm converges. The algorithm, as we implemented it, is based on the following steps:

1. Pick a starting point $\mathbf{d}_1$.
2. Pick a random $\mathbf{d}$ in a neighborhood of $\mathbf{d}_1$, and evaluate $R^B(\mathbf{d}, \widehat{\beta}|X)$ at $\mathbf{d}$.
3. Take $\mathbf{d}$ as your new starting point with a probability which is decreasing in $R^B(\mathbf{d}, \widehat{\beta}|X)$, increasing in $R^B(\mathbf{d}_1, \widehat{\beta}|X)$, and decreasing in the number of iterations already made relative to the total number of iterations which are going to be made.[13] If $\mathbf{d}$ is not taken as new starting point, stay at $\mathbf{d}_1$.
4. Iterate for a prespecified total number of iterations.

APPENDIX B: A BRIEF REVIEW OF THE OPTIMAL EXPERIMENTAL DESIGN LITERATURE

To put the results of this paper into context, we briefly review the approach taken in the literature on optimal experimental design. A general introduction can be found in Cox and Reid (2000, chapter 7); a concise overview is given in Dawid and Sebastiani (1999).[14]

[13] The probability of switching which we use is given by $\min(\exp(-(R^B(\mathbf{d}) - R^B(\mathbf{d}_1))/T), 1)$, where $T = .5 \cdot R^B(\mathbf{d}_0) \cdot (m - i)/m$, with m the total number of iterations and i the current iteration.

[14] It should be emphasized that most of this literature focuses on a different type of setting than the present paper, considering parametric models, without covariates or at most discrete

Consider the linear regression model $Y = X\beta + \epsilon$, where the residuals are uncorrelated and homoskedastic, $\text{Var}(\epsilon) = \sigma^2 I$. The variance of the standard OLS estimator of β is given by

$$\text{Var}(\widehat{\beta}) = \sigma^2 (X'X)^{-1} = \frac{\sigma^2}{n} M(\xi)^{-1}$$

where

$$M(\xi) = \int xx' d\xi(x)$$

is the "design second moment," which is proportional to the Fisher information, and

$$\xi(x) = \frac{1}{n} \sum_i \delta_{x_i}(x)$$

is the design measure. Optimal experimental design is concerned with minimizing various functionals of $M(\xi)^{-1}$ through choice of the design measure ξ, subject possibly to constraints on the support $\mathcal{X}$ and other features of feasible design measures. Optimal design theory usually neglects the constraint that, in finite samples, ξ has to be a measure consisting of n point masses of size $1/n$. The optimal design measures obtained can thus be considered as an asymptotic approximation.

Two important optimality criteria are "D-optimality," which is concerned with maximizing $\det(M(\xi))$, and "G-optimality," which is concerned with minimizing the worst case variance of prediction,

$$\overline{d}(\xi) := \sup_{x \in \mathcal{X}} x' M(\xi)^{-1} x.$$

The General Equivalence Theorem of experimental design states that those two criteria are in fact optimized by the same design measure ξ^*, and that $\overline{d}(\xi^*) = \dim(\beta)$. The General Equivalence Theorem was first shown by Kiefer and Wolfowitz (1959). Other optimality criteria discussed in the literature include "D_A optimality," which is concerned with minimizing the determinant of the variance of a linear combination of parameters $A\beta$, and "A-optimality," which is concerned with minimizing $tr(M(\xi)^{-1})$.

The literature discusses many generalizations of the basic setting reviewed here. In particular, there is a literature on optimal nonlinear design, which considers optimization of functionals of the information matrix in more general, nonlinear, parametric models.

covariates, taking a frequentist approach, and minimizing functionals of the variance matrix of some parameter vector of interest. In contrast the present paper considers nonparametric models, allows for continuous covariates, takes an explicitly decision theoretic perspective, and minimizes the risk of estimating the average treatment effect.

REFERENCES

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444–455.

Berry, D. (2006): "Bayesian clinical trials," *Nature Reviews Drug Discovery*, 5, 27–36.

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.

Carrasco, M., J. Florens, and E. Renault (2007): "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization," *Handbook of econometrics*, 6, 5633–5751.

Cox, D. and N. Reid (2000): *The theory of the design of experiments.*

Dawid, A. and P. Sebastiani (1999): "Coherent dispersion criteria for optimal experimental design," *Annals of Statistics*, 65–81.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using randomization in development economics research: A toolkit," *Handbook of development economics*, 4, 3895–3962.

Graham, B. (2008): "Identifying social interactions through conditional variance restrictions," *Econometrica*, 76, 643–660.

Imbens, G. (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and statistics*, 86, 4–29.

Kasy, M. (2013): "Matlab code to find optimal experimental designs," \url{http://scholar.harvard.edu/files/kasy/files/MATLABexperimentaldesign.zip}.

Kempthorne, O. (1977): "Why randomize?" *Journal of Statistical Planning and Inference*, 1, 1–25.

Kiefer, J. and J. Wolfowitz (1959): "Optimum designs in regression problems." *Ann. Math. Stat.*, 30, 271–294.

Kirkpatrick, S., M. Vecchi, et al. (1983): "Optimization by simmulated annealing," *science*, 220, 671–680.

Krueger, A. (1999): "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 114, 497–532.

List, J. and I. Rasul (2011): "Field experiments in labor economics," *Handbook of Labor Economics*, 4, 103–228.

Matheron, G. (1973): "The intrinsic random functions and their applications," *Advances in applied probability*, 439–468.

O'Hagan, A. and J. Kingman (1978): "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–42.

Robert, C. (2007): *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Verlag.

Rubin, D. B. (1978): "Bayesian inference for causal effects: The role of randomization," *The Annals of Statistics*, 34–58.

Shah, K. and B. Sinha (1989): *Theory of Optimal Designs - Lecture Notes in Statistics*, Springer-Verlag.

Silverman, B. (1984): "Spline smoothing: the equivalent variable kernel method," *The Annals of Statistics*, 898–916.

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.

Sollich, P. and C. Williams (2005): "Using the equivalent kernel to understand gaussian process regression," *working paper.*

Stone, M. (1969): "The role of experimental randomization in bayesian statistics: Finite sampling and two bayesians," *Biometrika*, 681–683.

van der Vaart, A. and J. van Zanten (2008a): "Rates of contraction of posterior distributions based on gaussian process priors," *The Annals of Statistics*, 36, 1435–1463.

——— (2008b): "Reproducing kernel Hilbert spaces of Gaussian priors," *IMS Collections*, 3, 200–222.

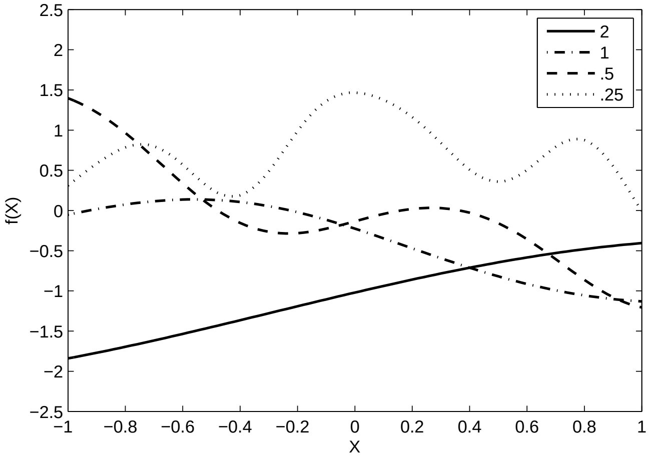
Wahba, G. (1990): *Spline models for observational data*, vol. 59, Society for Industrial Mathematics.

Williams, C. and C. Rasmussen (2006): *Gaussian processes for machine learning*, MIT Press.

Yakowitz, S. and F. Szidarovszky (1985): "A comparison of kriging with nonparametric regression methods," *Journal of Multivariate Analysis*, 16, 21–53.

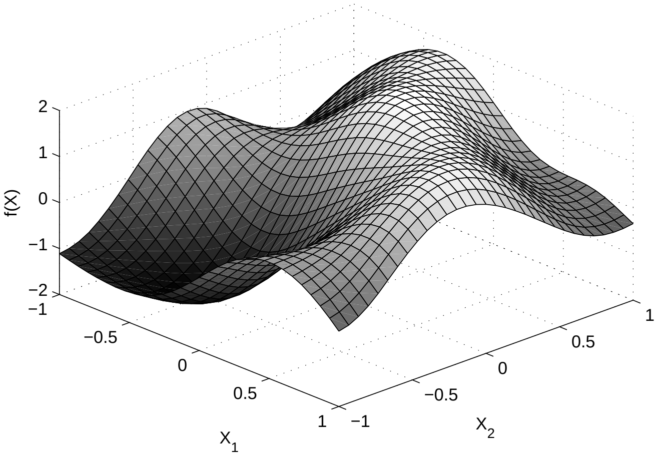
APPENDIX B: FIGURES AND TABLES

FIGURE 1.— Draws from Gaussian processes with squared exponential covariance function



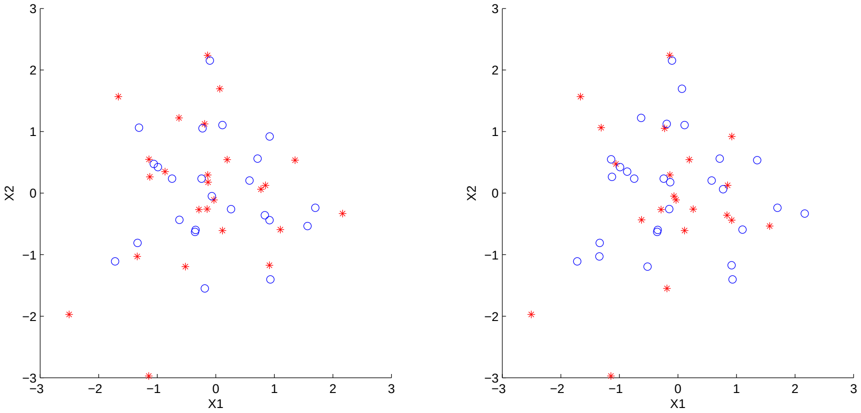
Notes: This figure shows draws from Gaussian processes with covariance kernel $C(x_1, x_2) = \exp\left(-\frac{1}{2l^2}|x_1 - x_2|^2\right)$, with the length scale l ranging from 0.25 to 2.

FIGURE 2.— Draw from a Gaussian process with squared exponential covariance function



Notes: This figure shows a draw from a Gaussian process with covariance kernel $C(x_1, x_2) = \exp\left(-\frac{1}{2l^2}\|x_1 - x_2\|^2\right)$, where $l = 0.5$ and $X \in \mathbf{R}^2$.

FIGURE 3.— Example of optimal design and random design



Optimal design Random design

Notes: This figure shows designs for a sample of size 50 of two standard normal covariates. Dots which are $*$-shaped correspond to units receiving treatment ($D_i = 1$), o-shaped dots are controls ($D_i = 0$). The left hand figure shows an optimal design for $\sigma^2 = 4$ and the squared exponential kernel prior discussed in section 6. The right hand figure shows a random design for the same data.

TABLE I

The mean squared error of randomized designs relative to optimal designs

data parameters			prior		
n	σ^2	$dim(X)$	linear model	squared exponential	non-informative
50	4.0	1	1.05	1.03	1.05
50	4.0	5	1.19	1.02	1.07
50	1.0	1	1.05	1.07	1.09
50	1.0	5	1.18	1.13	1.20
200	4.0	1	1.01	1.01	1.02
200	4.0	5	1.03	1.04	1.07
200	1.0	1	1.01	1.02	1.03
200	1.0	5	1.03	1.15	1.20
800	4.0	1	1.00	1.01	1.01
800	4.0	5	1.01	1.05	1.06
800	1.0	1	1.00	1.01	1.01
800	1.0	5	1.01	1.13	1.16

Notes: This table shows the average ratio of the posterior mean squared error, as given by equation (14), of random designs relative to optimal designs, for covariate distributions with i.i.d. standard normal components of X. The average is taken over random designs. The values 4 and 1 for σ^2 correspond to an expected R^2 of around .2 and .5 for the squared exponential kernel prior. The priors considered are those discussed and recommended in section 6.

TABLE II

Robustness to the choice of prior

data parameters			prior		
n	σ^2	$dim(X)$	linear model	squared exponential	non-informative
50	4.0	1	1.00	1.00	1.01
50	4.0	5	1.00	1.01	1.02
50	1.0	1	1.00	1.01	1.00
50	1.0	5	1.00	1.04	1.05
200	4.0	1	1.00	1.00	1.00
200	4.0	5	1.00	1.02	1.03
200	1.0	1	1.00	1.01	1.00
200	1.0	5	1.00	1.09	1.11
800	4.0	1	1.00	1.00	1.00
800	4.0	5	1.00	1.03	1.04
800	1.0	1	1.00	1.00	1.00
800	1.0	5	1.00	1.09	1.12

Notes: This table shows ratios of the posterior mean squared error, as given by equation (14), for the same data generating processes as table I. The ratios are taken between the posterior mean squared error under prior A for the design which is optimal under prior B, relative to the design which is optimal under prior A, where the priors A and B are distinguished by the length scale for X_1, see section 8 for details.

TABLE III

Covariate means for students assigned to small and regular classes

	$D=0$	$D=1$
girl	0.47	0.49
black	0.37	0.35
birth date	1980.23	1980.30
free lunch	0.46	0.45
n	6316	2274

Notes: This table shows the means of various characteristics for students assigned to small and regular classes. Note that the means are not expected to be equal, since randomization took place on the school level, not in the entire sample. The variable "free lunch" is a proxy for coming from a poor family, "birth date" equals year plus 0.25 × quarter of birth.

TABLE IV

Covariate means for selected schools - actual versus optimal treatment assignment

School 16				
	$D = 0$	$D = 1$	$D^* = 0$	$D^* = 1$
girl	0.42	0.54	0.46	0.41
black	1.00	1.00	1.00	1.00
birth date	1980.18	1980.48	1980.24	1980.27
free lunch	0.98	1.00	0.98	1.00
n	123	37	123	37

School 38				
	$D = 0$	$D = 1$	$D^* = 0$	$D^* = 1$
girl	0.45	0.60	0.49	0.47
black	0.00	0.00	0.00	0.00
birth date	1980.15	1980.30	1980.19	1980.17
free lunch	0.86	0.33	0.73	0.73
n	49	15	49	15

Notes: This table shows the means of some student characteristics for the actual treatment assignment D, as well as the (close to) optimal treatment assignment D^*. Note that, as a consequence of randomization, some covariate means are not balanced for the actual assignment, while the optimal assignment aims for a similar distribution of covariates between different treatment arms.