Why experimenters might not always want to randomize, and what they could do instead
Online Appendix

Maximilian Kasy
Department of Economics, Harvard University

March 7, 2016

Contents

1 Literature

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). The fact that deterministic designs might be optimal in the presence of covariates was noted by Atkinson (1982) in the context of a parametric model and sequential experimental design. Some general discussions on the role of randomization in experiments took place a few decades ago, see in particular Stone (1969), Kempthorne (1977), and Rubin (1978). The role of covariates in the analysis of experiments is also discussed by Imai et al. (2008).

Morgan and Rubin (2012) argue that experimenters should re-randomize until satisfactory covariate balance is achieved. Their argument is similar in spirit to ours. Moore (2012) makes an argument for blocking making use of the rich set of baseline covariates often available in field experiments. I very much agree with this argument; one way to think of the present paper is that it provides a formal foundation for this argument and takes it to its logical conclusion. Simultaneous treatment assignment to all units is not always feasible, particularly in settings where participants arrive sequentially. Such settings are discussed by Moore and Moore (2013); sequential design is not considered in the present paper. Bruhn and McKenzie (2009) have studied the relative variance of estimators under various designs using simulations.

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).

A few examples of experiments from the recent political science literature were mentioned in the introduction. Additional examples abound. Bolsen et al. (2014) sent messages on water conservation to a random set of individuals in Florida to study differential effects by voting behavior. Michelitch (2015) study whether taxi fare bargaining is affected by differences in ethnicity and/or political affiliation between driver and customer in Ghana. Further examples can be found in Gerber et al. (2014); Hanmer et al. (2014); McClendon (2014); Nyhan and Reifler (2014); Paler (2013).

2 Alternative optimization methods

In addition to the the re-randomization approach described in the main paper, there exist alternative, more sophisticated methods of optimization, of course, and there is an extensive literature discussing algorithms for discrete optimization problems such as ours. One such set of procedures, called greedy algorithms, is based on local search. The idea is to start from some assignment d^0, and search over a set of "neighboring" assignments that only differ in a few components to find the best assignment among those. This best assignment is labeled d^1 and is used as the starting point for a new local search. The procedure is left to run until a local optimum is found or timeout is reached.

A variation on greedy algorithms is so-called simulated annealing. This is one of the most popular algorithms for discrete optimization and was introduced by Kirkpatrick et al. (1983). 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.

We have implemented versions of each of these, and Matlab code is available as part of this online supplement. In practice it appears that re-randomization performs quite well, and might be preferred by practitioners on account of its simplicity.

3 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).

We 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. 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.

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[1] is given by

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

[1] 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).

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

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

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

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

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

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

where $\|.\|$ is some appropriately defined norm measuring the distance between covariate vectors. The parameter l determines the length scale of the process.

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.[2] 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

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

of regressors x_1, consider the process

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

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 $\overline{K}^d = \mathrm{Cov}(Y^d, \overline{g}^d|X, D)$.[3]

Theorem 1 (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

$$\widehat{\overline{f}}^d_\infty = \overline{X}_1 \widehat{\beta}^d_\infty + \overline{K}^d K_y^{d,-1}(Y^d - X_1^d \widehat{\beta}^d_\infty) \tag{3}$$

and

$$\widehat{\beta}^d_\infty = \left(X^{d\prime} K_y^{d,-1} X^d\right)^{-1} X^{d\prime} K_y^{d,-1} Y^d. \tag{4}$$

For any λ, we have

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

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

and

$$R(\mathbf{d}, \widehat{\beta}_\infty|X) = \mathrm{Var}(\overline{f}^1 - \widehat{\overline{f}}^1_\infty|X) + \mathrm{Var}(\overline{f}^0 - \widehat{\overline{f}}^0_\infty|X) \tag{5}$$

where

$$\begin{aligned} \mathrm{Var}(\overline{f}^d - \widehat{\overline{f}}^d_\infty|X) &= \mathrm{Var}(\overline{g}^d) - \overline{K}^d K_y^{d,-1}\overline{K}^d \\ &\quad + (\overline{X} - \overline{K}^d K_y^{d,-1}X) \cdot \left(X^{d\prime} K_y^{d,-1} X^d\right)^{-1} (\overline{X} - \overline{K}^d K_y^{d,-1}X)'. \end{aligned} \tag{6}$$

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

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

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 premultiplying 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

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

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

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

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

Remark: Note that the limiting estimator of theorem 1 can be understood as penalized regression, where the penalization corresponds to the seminorm

$$\|f^d\|^2 = \min_{\widehat{\beta}} (f^d - X^d\widehat{\beta})' \cdot K_y^{d,-1} \cdot (f^d - X^d\widehat{\beta}). \tag{7}$$

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. This share is given by

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

so that

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

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

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

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 1, this prior implies a best linear predictor for β of

$$\widehat{\beta}^1_\infty - \widehat{\beta}^0_\infty + \overline{K}^1 K_y^{1,-1}(Y^1 - e\widehat{\beta}^1_\infty) - \overline{K}^0 K_y^{0,-1}(Y^0 - e\widehat{\beta}^0_\infty) \tag{8}$$

where

$$\widehat{\beta}^d_\infty = \left(e' K_y^{d,-1} e\right)^{-1} e' K_y^{d,-1} Y^d. \tag{9}$$

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

$$
\begin{aligned}
\mathrm{Var}(\beta|X,D,Y) = {} & \mathrm{Var}(\overline{g}^1|X) + \mathrm{Var}(\overline{g}^0|X) - \overline{K}^1 K_y^{1,-1} \overline{K}^1 - \overline{K}^0 K_y^{0,-1} \overline{K}^0 \\
& + (1 - \overline{K}^1 K_y^{1,-1} e) \cdot \left(e' K_y^{d,-1} e\right)^{-1} (1 - \overline{K}^1 K_y^{1,-1} e)' \\
& + (1 - \overline{K}^0 K_y^{0,-1} e) \cdot \left(e' K_y^{0,-1} e\right)^{-1} (1 - \overline{K}^0 K_y^{0,-1} e)'. \qquad (10)
\end{aligned}
$$

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 1.

References

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

Berry, D. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36.

Bolsen, T., Ferraro, P. J., and Miranda, J. J. (2014). Are voters more likely to contribute to other public goods? evidence from a large-scale randomized policy experiment. *American Journal of Political Science*, 58(1):17–30.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

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

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

Gerber, A. S., Huber, G. A., Meredith, M., Biggers, D. R., and Hendry, D. J. (2014). Can incarcerated felons be (re)integrated into the political system? results from a field experiment. *American Journal of Political Science*, pages n/a–n/a.

Hanmer, M. J., Banks, A. J., and White, I. K. (2014). Experiments to reduce the over-reporting of voting: A pipeline to the truth. *Political Analysis*, 22(1):130–141.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2):481–502.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

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

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

Kirkpatrick, S., Vecchi, M., et al. (1983). Optimization by simmulated annealing. *science*, 220(4598):671–680.

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

McClendon, G. H. (2014). Social esteem and participation in contentious politics: A field experiment at an lgbt pride rally. *American Journal of Political Science*, 58(2):279–290.

Michelitch, K. (2015). Does electoral competition exacerbate interethnic or interpartisan economic discrimination? evidence from a field experiment in market price bargaining. *American Political Science Review*, 109:43–61.

Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4):460–479.

Moore, R. T. and Moore, S. A. (2013). Blocking for sequential political experiments. *Political Analysis*, 21(4):507–523.

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

Nyhan, B. and Reifler, J. (2014). The effect of fact-checking on elites: A field experiment on u.s. state legislators. *American Journal of Political Science*, pages n/a–n/a.

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

Paler, L. (2013). Keeping the public purse: An experiment in windfalls, taxes, and the incentives to restrain government. *American Political Science Review*, 107:706–725.

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*, pages 34–58.

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

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(1/2):1–85.

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

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

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

Yakowitz, S. and Szidarovszky, F. (1985). A comparison of kriging with nonparametric regression methods. *Journal of Multivariate Analysis*, 16(1):21–53.