Why experimenters should not randomize, and what they should do instead

Maximilian Kasy

Department of Economics, Harvard University

project STAR

Covariate means within school for the actual (D) and for the optimal (D^*) 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

Some intuitions

- "compare apples with apples"
  $\Rightarrow$ balance covariate distribution
- not just balance of means!
- don't add random noise to estimators
  – why add random noise to experimental designs?
- optimal design for STAR:
  19% reduction in mean squared error
  relative to actual assignment
- equivalent to 9% sample size, or 773 students

Some context - a very brief history of experiments

How to ensure we compare apples with apples?

1. physics - Galileo,...
   controlled experiment, not much heterogeneity, no self-selection
   $\Rightarrow$ no randomization necessary
2. modern RCTs - Fisher, Neyman,...
   observationally homogenous units with unobserved heterogeneity
   $\Rightarrow$ randomized controlled trials
   (setup for most of the experimental design literature)
3. medicine, economics:
   lots of unobserved **and observed** heterogeneity
   $\Rightarrow$ topic of this talk

The setup

1. *Sampling:*
   random sample of n units
   baseline survey $\Rightarrow$ vector of covariates X_i
2. *Treatment assignment:*
   binary treatment assigned by $D_i = d_i(X, U)$
   X matrix of covariates; U randomization device
3. *Realization of outcomes:*
   $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$
4. *Estimation:*
   estimator $\widehat{\beta}$ of the (conditional) average treatment effect,
   $\beta = \frac{1}{n} \sum_i E[Y_i^1 - Y_i^0 | X_i, \theta]$

Questions

- How should we assign treatment?
- In particular, if X has continuous or many discrete components?
- How should we estimate β?
- What is the role of prior information?

Framework proposed in this talk

1. *Decision theoretic:*
   $\mathbf{d}$ and $\widehat{\beta}$ minimize risk $R(\mathbf{d}, \widehat{\beta}|X)$
   (e.g., expected squared error)
2. *Nonparametric:*
   no functional form assumptions
3. *Bayesian:*
   $R(\mathbf{d}, \widehat{\beta}|X)$ averages expected loss over a prior.
   prior: distribution over the functions $x \to E[Y_i^d | X_i = x, \theta]$
4. *Non-informative:*
   limit of risk functions under priors such that
   $\text{Var}(\beta) \to \infty$

Main results

1. The unique optimal treatment assignment does not involve randomization.
2. Identification using conditional independence is still guaranteed without randomization.
3. Tractable nonparametric priors
4. Explicit expressions for risk as a function of treatment assignment
$\Rightarrow$ choose $\mathbf{d}$ to minimize these
5. MATLAB code to find optimal treatment assignment
6. Magnitude of gains:
    - between 5 and 20% reduction in MSE
relative to randomization, for realistic parameter values in simulations
    - For project STAR: 19% gain relative to actual assignment

Roadmap

1. Motivating examples
2. Formal decision problem and the optimality of non-randomized designs
3. Nonparametric Bayesian estimators and risk
4. Choice of prior parameters
5. Discrete optimization, and how to use my MATLAB code
6. Simulation results and application to project STAR
7. Outlook: Optimal policy and statistical decisions

Notation

- random variables: X_i, D_i, Y_i
- values of the corresponding variables: x, d, y
- matrices/vectors for observations $i = 1, \ldots, n$: X, D, Y
- vector of values: $\mathbf{d}$

- shorthand for data generating process: θ
- "frequentist" probabilities and expectations: conditional on θ
- "Bayesian" probabilities and expectations: unconditional

Example 1 - No covariates

- $n_d := \sum \mathbf{1}(D_i = d)$, $\sigma_d^2 = \mathrm{Var}(Y_i^d|\theta)$
- $$\widehat{\beta} := \sum_i \left[\frac{D_i}{n_1} Y_i - \frac{1 - D_i}{n - n_1} Y_i \right]$$
- Two alternative designs:
  1. Randomization conditional on n_1
  2. Complete randomization: D_i i.i.d., $P(D_i = 1) = p$
- Corresponding estimator variances
  1. n_1 fixed $\Rightarrow$
     $$\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n - n_1}$$
  2. n_1 random $\Rightarrow$
     $$E_{n_1}\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n - n_1} \right]$$
- Choosing (unique) minimizing n_1 is optimal.
- Indifferent which of observationally equivalent units get treatment.

Example 2 - discrete covariate

- $X_i \in \{0, \ldots, k\}$,
  $n_x := \sum_i \mathbf{1}(X_i = x)$
- $n_{d,x} := \sum_i \mathbf{1}(X_i = x, D_i = d)$,
  $\sigma^2_{d,x} = \text{Var}(Y_i^d | X_i = x, \theta)$
- $$\widehat{\beta} := \sum_x \frac{n_x}{n} \sum_i \mathbf{1}(X_i = x) \left[\frac{D_i}{n_{1,x}} Y_i - \frac{1 - D_i}{n_x - n_{1,x}} Y_i \right]$$
- Three alternative designs:
  1. Stratified randomization, conditional on $n_{d,x}$
  2. Randomization conditional on $n_d = \sum \mathbf{1}(D_i = d)$
  3. Complete randomization

Corresponding estimator variances

1. Stratified; $n_{d,x}$ fixed $\Rightarrow$

$$V(\{n_{d,x}\}) := \sum_x \frac{n_x}{n} \left[\frac{\sigma^2_{1,x}}{n_{1,x}} + \frac{\sigma^2_{0,x}}{n_x - n_{1,x}} \right]$$

2. $n_{d,x}$ random but $n_d = \sum_x n_{d,x}$ fixed $\Rightarrow$

$$E\left[V(\{n_{d,x}\}) \middle| \sum_x n_{1,x} = n_1 \right]$$

3. $n_{d,x}$ and n_d random $\Rightarrow$

$$E\left[V(\{n_{d,x}\}) \right]$$

$\Rightarrow$ Choosing unique minimizing $\{n_{d,x}\}$ is optimal.

Example 3 - Continuous covariate

- $X_i \in \mathbb{R}$ continuously distributed
- $\Rightarrow$ **no two observations have the same** X_i!
- Alternative designs:
  1. Complete randomization
  2. Randomization conditional on n_d
  3. Discretize and stratify:
     - Choose bins $[x_j, x_{j+1}]$
     - $\tilde{X}_i = \sum j \cdot \mathbf{1}(X_i \in [x_j, x_{j+1}])$
     - stratify based on $\tilde{X}_i$
  4. Special case: pairwise randomization
  5. "Fully stratify"
- But what does that mean???

Some references

- *Optimal design of experiments*:
  Smith (1918), Kiefer and Wolfowitz (1959), Cox and Reid (2000), Shah and Sinha (1989)
- *Nonparametric estimation of treatment effects*:
  Imbens (2004)
- *Gaussian process priors*:
  Wahba (1990) (Splines), Matheron (1973); Yakowitz and Szidarovszky (1985) ("Kriging" in Geostatistics), Williams and Rasmussen (2006) (machine learning)
- *Bayesian statistics, and design*:
  Robert (2007), O'Hagan and Kingman (1978), Berry (2006)
- *Simulated annealing*:
  Kirkpatrick et al. (1983)

A formal decision problem

- risk function of treatment assignment $\mathbf{d}(X,U)$, estimator $\widehat{\beta}$, under loss L, data generating process θ:

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

  ($\mathbf{d}$ affects the distribution of $\widehat{\beta}$)
- (conditional) Bayesian risk:

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

$$R^B(\mathbf{d},\widehat{\beta}|X) := \int R^B(\mathbf{d},\widehat{\beta}|X,U)dP(U) \quad (3)$$

$$R^B(\mathbf{d},\widehat{\beta}) := \int R^B(\mathbf{d},\widehat{\beta}|X,U)dP(X)dP(U) \quad (4)$$

- conditional minimax risk:

$$R^{mm}(\mathbf{d},\widehat{\beta}|X,U) := \max_\theta R(\mathbf{d},\widehat{\beta}|X,U,\theta) \quad (5)$$

- objective: $\min R^B$ or $\min R^{mm}$

Optimality of deterministic designs

Theorem

Given $\widehat{\beta}(Y, X, D)$

1. $$\mathbf{d}^*(X) \in \underset{\mathbf{d}(X) \in \{0,1\}^n}{\operatorname{argmin}} R^B(\mathbf{d}, \widehat{\beta}|X) \tag{6}$$
   minimizes $R^B(\mathbf{d}, \widehat{\beta})$ *among all* $\mathbf{d}(X, U)$ *(random or not).*
2. *Suppose* $R^B(\mathbf{d}^1, \widehat{\beta}|X) - R^B(\mathbf{d}^2, \widehat{\beta}|X)$ *is continuously distributed* $\forall \mathbf{d}^1 \neq \mathbf{d}^2 \Rightarrow \mathbf{d}^*(X)$ *is the unique minimizer of* (6).
3. *Similar claims hold for* $R^{mm}(\mathbf{d}, \widehat{\beta}|X, U)$, *if the latter is finite.*

Intuition:

- similar to why estimators should not randomize
- $R^B(\mathbf{d}, \widehat{\beta}|X, U)$ does not depend on U
  $\Rightarrow$ neither do its minimizers $\mathbf{d}^*, \widehat{\beta}^*$

Conditional independence

Theorem

Assume

- *i.i.d. sampling*
- *stable unit treatment values,*
- *and* $D = \mathbf{d}(X, U)$ *for* $U \perp (Y^0, Y^1, X)|\theta$*.*

Then conditional independence holds;

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

This is true in particular for deterministic treatment assignment rules $D = \mathbf{d}(X)$*.*

Intuition: under i.i.d. sampling

$$P(Y_i^{d_i}|X, \theta) = P(Y_i^{d_i}|X_i, \theta).$$

Nonparametric Bayes

Let $f(X_i, D_i) = E[Y_i | X_i, D_i, \theta]$.

Assumption (Prior moments)

$E[f(x,d)] = \mu(x,d)$
$\mathrm{Cov}(f(x_1,d_1), f(x_2,d_2)) = C((x_1,d_1),(x_2,d_2))$

Assumption (Mean squared error objective)

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

Assumption (Linear estimators)

$\widehat{\beta} = w_0 + \sum_i w_i Y_i$,
where w_i *might depend on* X *and on* D, *but not on* Y.

Best linear predictor, posterior variance

Notation for (prior) moments

$$\mu_i = E[Y_i|X,D], \quad \mu_\beta = E[\beta|X,D]$$
$$\Sigma = \mathrm{Var}(Y|X,D,\theta),$$
$$C_{i,j} = C((X_i,D_i),(X_j,D_j)), \text{ and } \overline{C}_i = \mathrm{Cov}(Y_i,\beta|X,D)$$

Theorem

Under these assumptions, the optimal estimator equals

$$\widehat{\beta} = \mu_\beta + \overline{C}' \cdot (C+\Sigma)^{-1} \cdot (Y-\mu),$$

and the corresponding expected loss (risk) equals

$$R^B(\mathbf{d},\widehat{\beta}|X) = \mathrm{Var}(\beta|X) - \overline{C}' \cdot (C+\Sigma)^{-1} \cdot \overline{C}.$$

More explicit formulas

Assumption (Homoskedasticity)

$$\text{Var}(Y_i^d | X_i, \theta) = \sigma^2$$

Assumption (Restricting prior moments)

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

Submatrix notation

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

Theorem (Explicit estimator and risk function)

Under these additional assumptions,

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

and

$$R^B(\mathbf{d}, \widehat{\beta} | X) = \mathrm{Var}(\beta | X) - \overline{C}^{1\prime} \cdot (V^1)^{-1} \cdot \overline{C}^1 - \overline{C}^{0\prime} \cdot (V^0)^{-1} \cdot \overline{C}^0.$$

Insisting on the comparison-of-means estimator

Assumption (Simple estimator)

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

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

Theorem (Risk function for designs using the simple estimator)

Under this additional assumption,

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

where $\tilde{C}_{ij} = C(X_i, X_j)$ *and* $v_i = \frac{1}{n} \cdot \left(-\frac{n_0}{n_1}\right)^{D_i}$.

Possible priors 1 - linear model

For X_i possibly including powers, interactions, etc.,

$$Y_i^d = X_i \beta^d + \epsilon_i^d$$
$$E[\beta^d|X] = 0, \qquad \text{Var}(\beta^d|X) = \Sigma_\beta$$

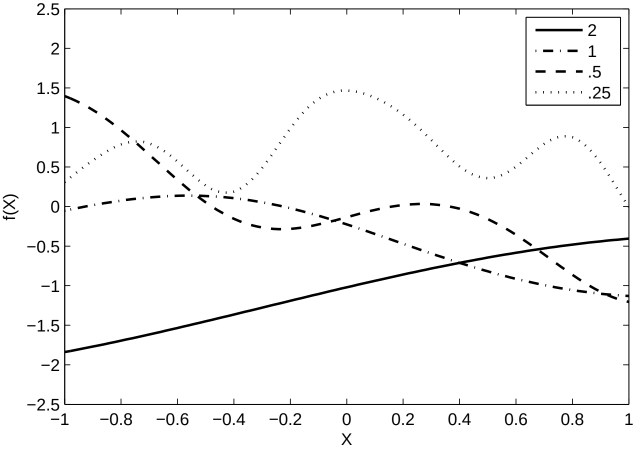
This implies

$$C = X \Sigma_\beta X'$$
$$\widehat{\beta}^d = \left(X^{d\prime} X^d + \sigma^2 \Sigma_\beta^{-1}\right)^{-1} X^{d\prime} Y^d$$
$$\widehat{\beta} = \overline{X}\left(\widehat{\beta}^1 - \widehat{\beta}^0\right)$$
$$R(\mathbf{d}, \widehat{\beta}|X) = \sigma^2 \cdot \overline{X} \cdot \left(\left(X^{1\prime} X^1 + \sigma^2 \Sigma_\beta^{-1}\right)^{-1} + \left(X^{0\prime} X^0 + \sigma^2 \Sigma_\beta^{-1}\right)^{-1}\right) \cdot \overline{X}'$$

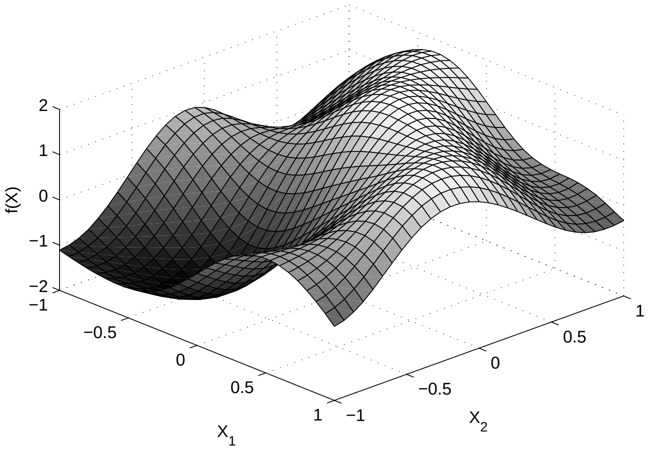
Possible priors 2 - squared exponential kernel

$$C(x_1, x_2) = \exp\left(-\frac{1}{2l^2}\|x_1 - x_2\|^2\right) \tag{7}$$

- popular in machine learning, (cf. Williams and Rasmussen, 2006)
- *nonparametric:* does not restrict functional form; can accommodate any shape of f^d
- *smooth:* f^d is infinitely differentiable (in mean square)
- length scale l / norm $\|x_1 - x_2\|$ determine smoothness



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.



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 $X \in \mathbf{R}^2$.

Possible priors 3 - noninformativeness

- "non-subjectivity" of experiments
  $\Rightarrow$ would like prior non-informative about object of interest (ATE), while maintaining prior assumptions on smoothness
- possible formalization:

$$Y_i^d = g^d(X_i) + X_{1,i}\beta^d + \epsilon_i^d$$
$$\text{Cov}(g^d(x_1), g^d(x_2)) = K(x_1, x_2)$$
$$\text{Var}(\beta^d|X) = \lambda\Sigma_\beta,$$

  and thus $C^d = K^d + \lambda X_1^d \Sigma_\beta X_1^{d\prime}$.
- paper provides explicit form of

$$\lim_{\lambda\to\infty} \min_{\widehat{\beta}} R^B(\mathbf{d}, \widehat{\beta}|X).$$

Frequentist Inference

Variance of $\widehat{\beta}$: $V := \mathrm{Var}(\widehat{\beta}|X, D, \theta)$
$\widehat{\beta} = w_0 + \sum_i w_i Y_i \Rightarrow$

$$V = \sum w_i^2 \sigma_i^2, \tag{8}$$

where $\sigma_i^2 = \mathrm{Var}(Y_i|X_i, D_i)$.

Estimator of the variance:

$$\widehat{V} := \sum w_i^2 \widehat{\epsilon}_i^2. \tag{9}$$

where $\widehat{\epsilon}_i = Y_i - \widehat{f}_i$, $\widehat{f} = C \cdot (C + \Sigma)^{-1} \cdot Y$.

Proposition

$\widehat{V}/V \to^p 1$ *under regularity conditions stated in the paper.*

Discrete optimization

optimal design solves

$$\max_{\mathbf{d}} \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}$$

- discrete support
- 2^n possible values for $\mathbf{d}$
- $\Rightarrow$ brute force enumeration infeasible
- Possible algorithms (active literature!):
  1. Search over random $\mathbf{d}$
  2. Simulated annealing (c.f. Kirkpatrick et al., 1983)
  3. Greedy algorithm: search for local improvements by changing one (or k) components of $\mathbf{d}$ at a time
- My code: combination of these

How to use my MATLAB code

```
global X n dimx Vstar
%%%input X
[n, dimx]=size(X);
vbeta = @VarBetaNI
weights = @weightsNI
setparameters
Dstar=argminVar(vbeta);
w=weights(Dstar)
csvwrite('optimaldesign.csv',[Dstar(:), w(:), X])
```

1. make sure to appropriately normalize X
2. alternative objective and weight function handles: `@VarBetaCK, @VarBetaLinear, @weightsCK, @weightsLinear`
3. modifying prior parameters: setparameters.m
4. modifying parameters of optimization algorithm: argminVar.m
5. details: readme.txt

Simulation results

Next slide:

- average risk (expected mean squared error) $R^B(\mathbf{d}, \widehat{\beta}|X)$
- average of randomized designs relative to optimal designs
- various sample sizes, residual variances, dimensions of the covariate vector, priors
- covariates: multivariate standard normal
- We find: The gains of optimal designs
  1. decrease in sample size
  2. increase in the dimension of covariates
  3. decrease in σ^2

Table : 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

Project STAR

Krueger (1999a), Graham (2008)

- 80 schools in Tennessee 1985-1986:
- Kindergarten students **randomly** assigned to small (13-17 students) / regular (22-25 students) classes **within schools**
- Sample: students observed in grades 1 - 3
- Treatment $D = 1$ for students assigned to a small class (upon first entering a project STAR school)
- Controls: sex, race, year and quarter of birth, poor (free lunch), school ID
- Prior: squared exponential, noninformative
- Respecting budget constraint (same number of small classes)
- How much could MSE be improved relative to actual design?
- Answer: 19 % (equivalent to 9% sample size, or 773 students)

Table : COVARIATE MEANS WITHIN SCHOOL

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

Summary

- What is the optimal treatment assignment given baseline covariates?
- framework: decision theoretic, nonparametric, Bayesian, non-informative
- Generically there is a unique optimal design which does not involve randomization.
- tractable formulas for Bayesian risk (e.g., $R^B(\mathbf{d}, \widehat{\beta}|X) = \text{Var}(\beta|X) - \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}$)
- suggestions how to pick a prior
- MATLAB code to find the optimal treatment assignment
- Easy frequentist inference

Outlook: Using data to inform policy

Motivation 1 (theoretical)

1. **Statistical decision theory**
   evaluates estimators, tests, experimental designs
   based on expected loss
2. **Optimal policy theory**
   evaluates policy choices
   based on social welfare
3. **this paper**
   policy choice as a statistical decision
   statistical loss $\sim$ social welfare.

objectives:

1. anchoring econometrics in economic policy problems.
2. anchoring policy choices in a principled use of data.

Motivation 2 (applied)

empirical research to inform policy choices:

1. **development economics**: (cf. Dhaliwal et al., 2011)
   (cost) effectiveness of alternative policies / treatments
2. **public finance**:(cf. Saez, 2001; Chetty, 2009)
   elasticity of the tax base
   $\Rightarrow$ optimal taxes, unemployment benefits, etc.
3. **economics of education**: (cf. Krueger, 1999b; Fryer, 2011)
   impact of inputs on educational outcomes

objectives:

1. general econometric framework for such research
2. principled way to choose policy parameters based on data
   (and based on normative choices)
3. guidelines for experimental design

The setup

1. policy maker: expected utility maximizer
2. $u(t)$: utility for policy choice $t \in \mathscr{T} \subset \mathbb{R}^{d_t}$
   u is unknown
3. $u = L \cdot m + u_0$, L and u_0 are known
   L: linear operator $\mathscr{C}^1(\mathscr{X}) \to \mathscr{C}^1(\mathscr{T})$
4. $m(x) = E[g(x, \epsilon)]$ (average structural function)
   $X, Y = g(X, \epsilon)$ observables, ϵ unobserved
   expectation over distribution of ϵ in target population
5. experimental setting: $X \perp \epsilon$
   $\Rightarrow m(x) = E[Y|X = x]$
6. Gaussian process prior: $m \sim GP(\mu, C)$

Questions

1. How to choose t optimally
   given observations X_i, Y_i, $i = 1, \ldots, n$?
2. How to choose design points X_i
   given sample size n?
   How to choose sample size n?

$\Rightarrow$ mathematical characterization:

1. How does the optimal choice of t (in a Bayesian sense)
   behave asymptotically (in a frequentist sense)?
2. How can we characterize the optimal design?

Main mathematical results

1. Explicit expression for optimal policy $\widehat{t}^*$
2. Frequentist asymptotics of $\widehat{t}^*$
   1. asymptotically normal, slower than $\sqrt{n}$
   2. distribution driven by distribution of $\widehat{u}'(t^*)$
   3. confidence sets
3. Optimal experimental design
   design density $f(x)$ increasing in (but less than proportionately)
   1. density of t^*
   2. the expected inverse of the curvature $u''(t)$

   $\Rightarrow$ algorithm to choose design based on prior, objective function

Thanks for your time!