What is to be done?
Two attempts using Gaussian process priors

Maximilian Kasy

Department of Economics, Harvard University

Oct 14 2017

What questions should econometricians work on?

- Incentives of the publication process:
  - Appeal to referees from the same subfield.
  - Danger of self-referentiality,
    untethering from external relevance.
- Versus broader usefulness:
  - Tools useful for empirical researchers, policy makers.
  - Anchored in substantive applications,
    broader methodological considerations.
- One way to get there:
  Well defined decision problems.

Decision problems

- Objects to carefully choose:
  - Objective function.
  - Space of possible decisions / policy alternatives.
  - Identifying assumptions.
  - Prior information.
  - Features the priors should be uninformative about.
- Once these are specified, coherent and well-behaved solutions can be derived.
- Useful tool for tractable solutions without functional form restrictions: Gaussian process priors.

Outline of this talk

- Brief introduction to Gaussian process regression
- Application 1: Optimal treatment assignment in experiments.
  - Setting: Treatment assignment given baseline covariates
  - General decision theory result:
    Non-random rules dominate random rules
  - Prior for expectation of potential outcomes given covariates
  - Expression for MSE of estimator for ATE
    to minimize by treatment assignment
- Application 2: Optimal insurance and taxation.
  - Economic setting: Co-insurance rate for health insurance
  - Statistical setting: prior for behavioral average response function
  - Expression for posterior expected social welfare
    to maximize by choice of co-insurance rate

References

Williams, C. and Rasmussen, C. (2006). Gaussian processes for machine learning. *MIT Press, chapter 2.*

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. Political Analysis, *24(3):324–338.*

Kasy, M. (2017). Optimal taxation and insurance using machine learning. Working Paper, Harvard University.

Brief introduction to Gaussian process regression

- Suppose we observe n i.i.d. draws of (Y_i, X_i), where Y_i is real valued and X_i is a k vector.
- $Y_i = f(X_i) + \varepsilon_i$
- $\varepsilon_i | \boldsymbol{X}, f(\cdot) \sim N(0, \sigma^2)$
- Prior: f is distributed according to a Gaussian process,

$$f | \boldsymbol{X} \sim GP(0, C),$$

where C is a covariance kernel,

$$\text{Cov}(f(x), f(x') | \boldsymbol{X}) = C(x, x').$$

- We will leave conditioning on $\boldsymbol{X}$ implicit.

Posterior mean

- The joint distribution of $(f(x), \boldsymbol{Y})$ is given by

$$\begin{pmatrix} f(x) \\ \boldsymbol{Y} \end{pmatrix} \sim N\left(0, \begin{pmatrix} C(x,x) & c(x) \\ c(x)' & C + \sigma^2 I_n \end{pmatrix}\right),$$

  where
  - $c(x)$ is the n vector with entries $C(x, X_i)$,
  - and C is the $n \times n$ matrix with entries $C_{i,j} = C(X_i, X_j)$.
- Therefore

$$E[f(x)|\boldsymbol{Y}] = c(x) \cdot \left(C + \sigma^2 I_n\right)^{-1} \cdot \boldsymbol{Y}.$$

- Read: $\widehat{f}(\cdot) = E[f(\cdot)|\boldsymbol{Y}]$
  - is a linear combination of the functions $C(\cdot, X_i)$
  - with weights $\left(C + \sigma^2 I_n\right)^{-1} \cdot \boldsymbol{Y}$.

Both applications use Gaussian process priors

1. Optimal experimental design
   - How to assign treatment to minimize mean squared error for treatment effect estimators?
   - Gaussian process prior for the conditional expectation of potential outcomes given covariates.
2. Optimal insurance and taxation
   - How to choose a co-insurance rate or tax rate to maximize social welfare, given (quasi-)experimental data?
   - Gaussian process prior for the behavioral response function mapping the co-insurance rate into the tax base.

Application 1
"Why experimenters might not always want to randomize"
Setup

1. *Sampling:*
   random sample of n units
   baseline survey $\Rightarrow$ vector of covariates X_i
2. *Treatment assignment:*
   binary treatment assigned by $D_i = d_i(\boldsymbol{X}, U)$
   $\boldsymbol{X}$ 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_i has continuous or many discrete components?
- How should we estimate β?
- What is the role of prior information?

Some intuition

- "Compare apples with apples"
  $\Rightarrow$ balance covariate distribution.
- Not just balance of means!
- We don't add random noise to estimators
  – why add random noise to experimental designs?
- Identification requires controlled trials (CTs),
  but not randomized controlled trials (RCTs).

General decision problem allowing for randomization

- General decision problem:
  - State of the world θ, observed data X, randomization device $U \perp X$,
  - decision procedure $\delta(X, U)$, loss $L(\delta(X, U), \theta)$.
- Conditional expected loss of decision procedure $\delta(X, U)$:

$$R(\delta, \theta | U = u) = E[L(\delta(X, u), \theta) | \theta]$$

- Bayes risk:

$$R^B(\delta, \pi) = \int \int R(\delta, \theta | U = u) d\pi(\theta) dP(u)$$

- Minimax risk:

$$R^{mm}(\delta) = \int \max_{\theta} R(\delta, \theta | U = u) dP(u)$$

Theorem (Optimality of deterministic decisions)

Consider a general decision problem.
Let R^* equal R^B or R^{mm}. Then:

1. *The optimal risk $R^*(\delta^*)$, when considering only deterministic procedures $\delta(X)$, is no larger than the optimal risk when allowing for randomized procedures $\delta(X, U)$.*
2. *If the optimal deterministic procedure δ^* is unique, then it has strictly lower risk than any non-trivial randomized procedure.*

Proof

- Any probability distribution $P(u)$ satisfies
  - $\sum_u P(u) = 1$, $P(u) \geq 0$ for all u.
  - Thus $\sum_u R_u \cdot P(u) \geq \min_u R_u$ for any set of values R_u.
- Let $\delta^u(x) = \delta(x, u)$.
- Then

$$R^B(\delta, \pi) = \sum_u \int R(\delta^u, \theta) d\pi(\theta) P(u)$$

$$\geq \min_u \int R(\delta^u, \theta) d\pi(\theta) = \min_u R^B(\delta^u, \pi).$$

- Similarly

$$R^{mm}(\delta) = \sum_u \max_\theta R(\delta^u, \theta) P(u)$$

$$\geq \min_u \max_\theta R(\delta^u, \theta) = \min_u R^{mm}(\delta^u).$$

Bayesian setup

- Back to experimental design setting.
- Conditional distribution of potential outcomes: for $d = 0, 1$

$$Y_i^d | X_i = x \sim N(f(x, d), \sigma^2).$$

- Gaussian process prior:

$$\begin{aligned} f &\sim GP(\mu, C), \\ E[f(x, d)] &= \mu(x, d) \\ \text{Cov}(f(x_1, d_1), f(x_2, d_2)) &= C((x_1, d_1), (x_2, d_2)) \end{aligned}$$

- Conditional average treatment effect (CATE):

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

Notation

- Covariance matrix C, where $C_{i,j} = C((X_i, D_i), (X_j, D_j))$
- Mean vector μ, components $\mu_i = \mu(X_i, D_i)$
- Covariance of observations with CATE,

$$\begin{aligned}\overline{C}_i &= \mathrm{Cov}(Y_i, \beta \,|\, \boldsymbol{X}, \boldsymbol{D}) \\ &= \frac{1}{n} \sum_j \left(C((X_i, D_i), (X_j, 1)) - C((X_i, D_i), (X_j, 0)) \right).\end{aligned}$$

Posterior expectation and risk

- The posterior expectation $\widehat{\beta}$ of β equals

$$\widehat{\beta} = \mu_\beta + \overline{C}' \cdot (C + \sigma^2 I)^{-1} \cdot (\boldsymbol{Y} - \mu).$$

- The corresponding risk equals

$$\begin{aligned} R^B(\mathbf{d}, \widehat{\beta} | \boldsymbol{X}) &= \text{Var}(\beta | \boldsymbol{X}, \boldsymbol{Y}) \\ &= \text{Var}(\beta | \boldsymbol{X}) - Var(E[\beta | \boldsymbol{X}, \boldsymbol{Y}] | \boldsymbol{X}) \\ &= \text{Var}(\beta | \boldsymbol{X}) - \overline{C}' \cdot (C + \sigma^2 I)^{-1} \cdot \overline{C}. \end{aligned}$$

Discrete optimization

- The optimal design solves

$$\max_{\mathbf{d}} \overline{C}' \cdot (C + \sigma^2 I)^{-1} \cdot \overline{C}.$$

- Possible optimization algorithms:
  1. Search over random **d**
  2. greedy algorithm
  3. simulated annealing

Special case linear separable model

- Suppose

$$f(x,d) = x' \cdot \gamma + d \cdot \beta,$$
$$\gamma \sim N(0,\Sigma),$$

  and we estimate β using comparison of means.
- Bias of $\widehat{\beta}$ equals $(\overline{X}^1 - \overline{X}^0)' \cdot \gamma$, prior expected squared bias

$$(\overline{X}^1 - \overline{X}^0)' \cdot \Sigma \cdot (\overline{X}^1 - \overline{X}^0).$$

- Mean squared error

$$MSE(d_1,\ldots,d_n) = \sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_0}\right] + (\overline{X}^1 - \overline{X}^0)' \cdot \Sigma \cdot (\overline{X}^1 - \overline{X}^0).$$

- $\Rightarrow$Risk is minimized by
  1. choosing treatment and control arms of equal size,
  2. and optimizing balance as measured by the difference in covariate means $(\overline{X}^1 - \overline{X}^0)$.

Application 2
"Optimal insurance and taxation using machine learning"
Economic setting

- Population of insured individuals i.
- Y_i: health care expenditures of individual i.
- T_i: share of health care expenditures covered by the insurance
  $1 - T_i$: coinsurance rate; $Y_i \cdot (1 - T_i)$: out-of-pocket expenditures
- Behavioral response to share covered: structural function

$$Y_i = g(T_i, \varepsilon_i).$$

- Per capita expenditures under policy t: average structural function

$$m(t) = E[g(t, \varepsilon_i)].$$

Policy objective

- Insurance provider's expenditures per person: $t \cdot m(t)$.
  - Mechanical effect of increase in t (accounting):

$$m(t)dt.$$

  - Behavioral effect of increase in t (key empirical challenge):

$$t \cdot m'(t)dt.$$

- Utility of the insured:
  - Mechanical effect of increase in t (accounting):

$$m(t)dt.$$

  - Behavioral effect: None, by envelope theorem.
  - $\Rightarrow$ effect on utility = equivalent variation = mechanical effect
- Assign relative value $\lambda > 1$ to a marginal dollar for the sick vs. the insurer.

- Marginal effect of a change in t on social welfare:

$$u'(t) = (\lambda - 1) \cdot m(t) - t \cdot m'(t) = \lambda m(t) - \frac{\partial}{\partial t}(t \cdot m(t)). \quad (1)$$

- Integrating and imposing the normalization $u(0) = 0$:

$$u(t) = \lambda \int_0^t m(x)dx - t \cdot m(t). \quad (2)$$

- Special case $\lambda = 1$: “Harberger triangle” (not the relevant case)

Observed data and prior

- n i.i.d. draws of (Y_i, T_i)
- T_i was randomly assigned in an experiment, so that $T_i \perp \varepsilon_i$, and

$$E[Y_i|T_i = t] = E[g(t, \varepsilon_i)|T_i = t] = E[g(t, \varepsilon_i)] = m(t).$$

- Y_i is normally distributed given T_i,

$$Y_i|T_i = t \sim N(m(t), \sigma^2).$$

- Gaussian process prior for $m(\cdot)$,

$$m(\cdot) \sim GP(\mu(\cdot), C(\cdot, \cdot)).$$

Prior moments

- Linear functions of normal vectors are normal.
- Linear operators of Gaussian processes are Gaussian processes.
- Prior moments:

$$\nu(t) = E[u(t)] = \lambda \int_0^t \mu(x)dx - t \cdot \mu(t),$$

$$D(t,t') = \mathrm{Cov}(u(t), m(t'))) = \lambda \cdot \int_0^t C(x,t')dx - t \cdot C(t,t'),$$

$$\mathrm{Var}(u(t)) = \lambda^2 \cdot \int_0^t \int_0^t C(x,x')dx'dx$$

$$- 2\lambda t \cdot \int_0^t C(x,t)dx + t^2 \cdot C(t,t).$$

Posterior expectation of $u(\cdot)$

- Covariance with data:

$$\begin{aligned}\boldsymbol{D}(t) = \text{Cov}(u(t), \boldsymbol{Y}|\boldsymbol{T}) &= \text{Cov}(u(t), (m(T_1), \ldots, m(T_n))|\boldsymbol{T}) \\ &= (D(t, T_1), \ldots, D(t, T_n)).\end{aligned}$$

- Posterior expectation of $u(t)$:

$$\begin{aligned}\widehat{u}(t) &= E[u(t)|\boldsymbol{Y}, \boldsymbol{T}] \\ &= E[u(t)|\boldsymbol{T}] + \text{Cov}(u(t), \boldsymbol{Y}|\boldsymbol{T}) \cdot \text{Var}(\boldsymbol{Y}|\boldsymbol{T})^{-1} \cdot (\boldsymbol{Y} - E[\boldsymbol{Y}|\boldsymbol{T}]) \\ &= \nu(t) + \boldsymbol{D}(t) \cdot \left[\boldsymbol{C} + \sigma^2 \boldsymbol{I}\right]^{-1} \cdot (\boldsymbol{Y} - \mu).\end{aligned}$$

Optimal policy choice

- Bayesian policy maker aims to maximize expected social welfare (note: different from expectation of maximizer of social welfare!)
- Thus

$$\widehat{t}^* = \widehat{t}^*(\boldsymbol{Y}, \boldsymbol{T}) \in \operatorname*{argmax}_t \widehat{u}(t).$$

- First order condition

$$\begin{aligned}\frac{\partial}{\partial t}\widehat{u}(\widehat{t}^*) &= E[u'(\widehat{t}^*)|\boldsymbol{Y}, \boldsymbol{T}] \\ &= v'(\widehat{t}^*) + \boldsymbol{B}(\widehat{t}^*) \cdot \left[\boldsymbol{C} + \sigma^2 \boldsymbol{I}\right]^{-1} \cdot (\boldsymbol{Y} - \mu) = 0,\end{aligned}$$

where $\boldsymbol{B}(t) = (B(t, T_1), \ldots, B(t, T_n))$ and

$$\begin{aligned}B(t, t') &= \operatorname{Cov}\left(\frac{\partial}{\partial t}u(t), m(t')\right) = \frac{\partial}{\partial t}D(t, t') \\ &= (\lambda - 1) \cdot C(t, t') - t \cdot \frac{\partial}{\partial t}C(t, t').\end{aligned}$$

The RAND health insurance experiment

- (cf. Aron-Dine et al., 2013)
- Between 1974 and 1981
  representative sample of 2000 households
  in six locations across the US
- families randomly assigned to
  plans with one of six consumer coinsurance rates
- 95, 50, 25, or 0 percent
  2 more complicated plans (we drop those)
- Additionally: randomized Maximum Dollar Expenditure limits
  5, 10, or 15 percent of family income,
  up to a maximum of $750 or $1,000
  (we pool across those)

Table: Expected spending for different coinsurance rates

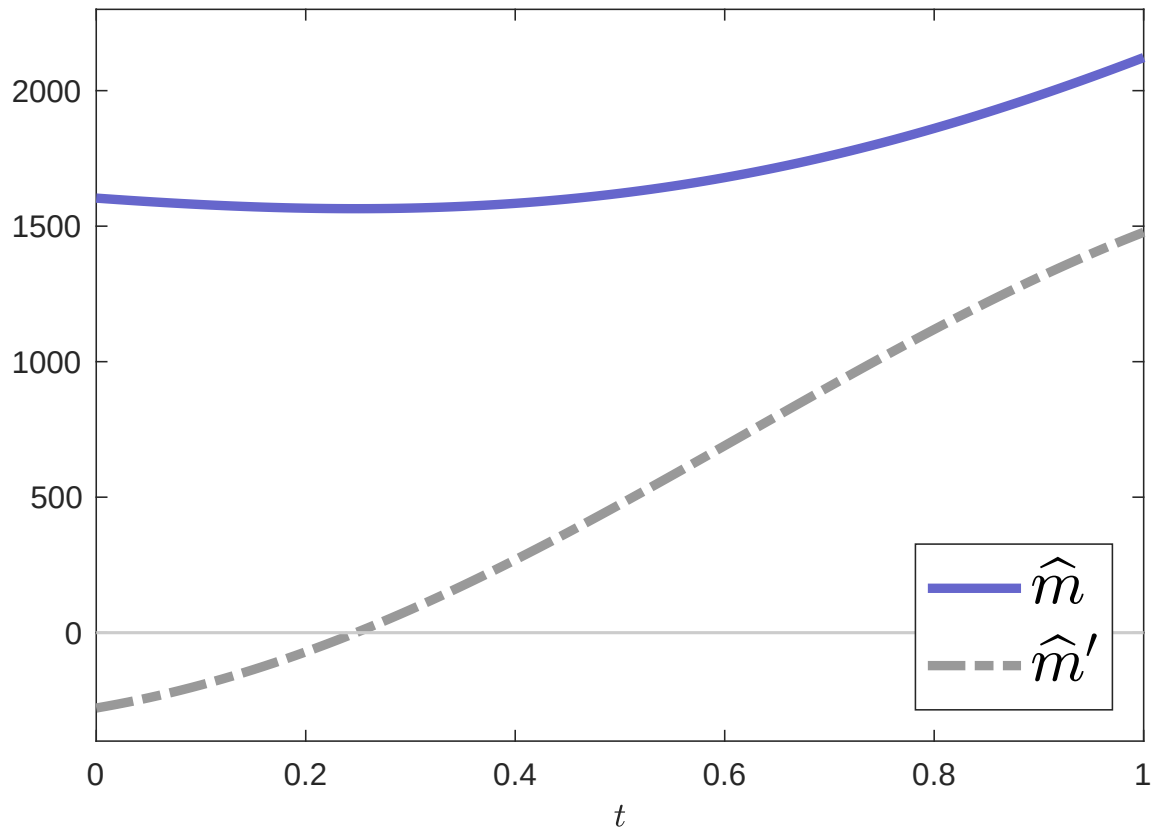
	(1) Share with any	(2) Spending in $	(3) Share with any	**(4) Spending in $**
Free Care	0.931	2166.1	0.932	2173.9
	(0.006)	(78.76)	(0.006)	(72.06)
25% Coinsurance	0.853	1535.9	0.852	1580.1
	(0.013)	(130.5)	(0.012)	(115.2)
50% Coinsurance	0.832	1590.7	0.826	1634.1
	(0.018)	(273.7)	(0.016)	(279.6)
95% Coinsurance	0.808	1691.6	0.810	1639.2
	(0.011)	(95.40)	(0.009)	(88.48)
family x month x site fixed effects	X	X	X	X
covariates			X	X
N	14777	14777	14777	14777

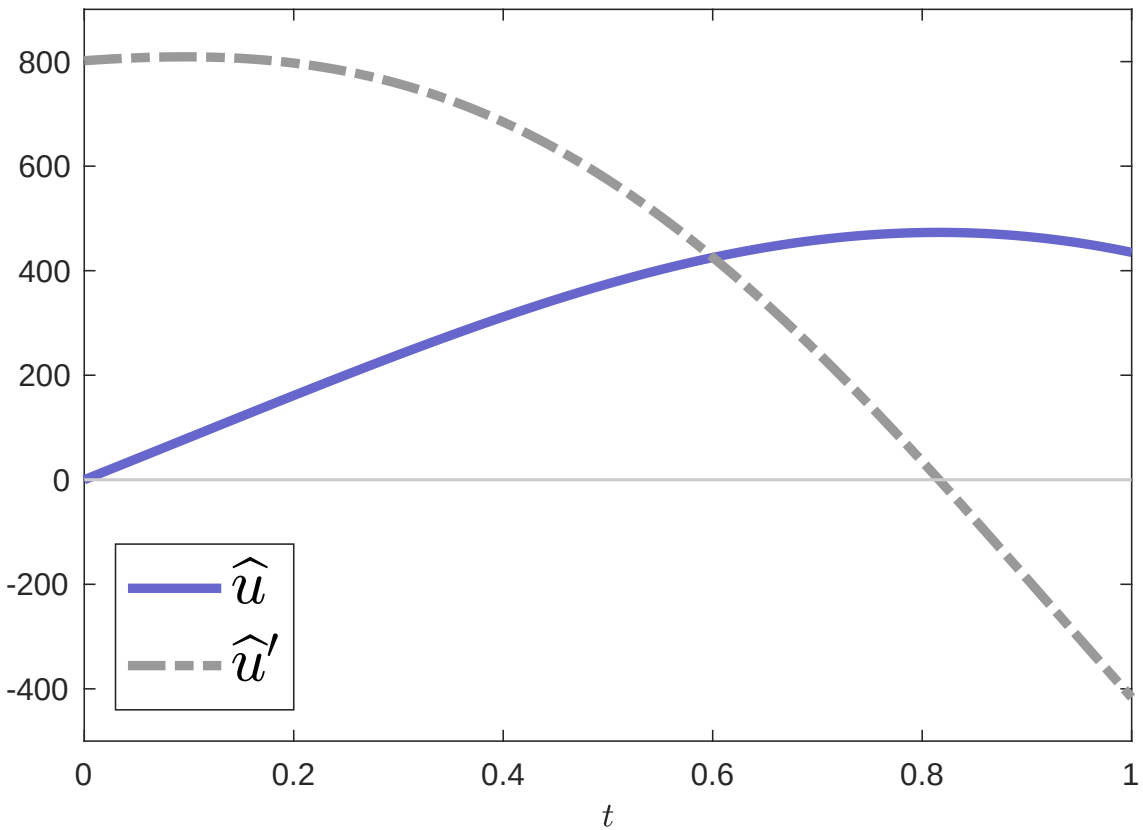
Assumptions

1. **Model**: The optimal insurance model as presented before
2. **Prior**: Gaussian process prior for m, squared exponential in distance, uninformative about level and slope
3. **Relative value** of funds for sick people vs contributors: $\lambda = 1.5$
4. Pooling data: across levels of maximum dollar expenditure

Under these assumptions we find:

Optimal copay equals 18%
(But free care is almost as good)


2000
1500
1000
500
0
0
0.2
0.4
0.6
0.8
1
t
$\widehat{m}$
$\widehat{m}'$


800
600
400
200
0
-200
-400
0
0.2
0.4
0.6
0.8
1
t
$\widehat{u}$
$\widehat{u}'$

Conclusion

- Explicit decision problems are useful to focus econometric research.
- Carefully choose:
  - Objective function.
  - Space of possible decisions / policy alternatives.
  - Identifying assumptions.
  - Prior information.
  - Features the priors should be uninformative about.
- Gaussian process priors allow for tractable solutions.
- Two examples:
  1. Optimal experimental design.
  2. Optimal insurance and taxation.

Thank you!