

Foundations of machine learning

Applications of Gaussian process priors

Maximilian Kasy

Department of Economics, University of Oxford

Hilary term 2027

Applications from my own work

- 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
- Optimal insurance and taxation.
 - Review: Envelope theorem.
 - 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

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.

Experimental design

Optimal insurance and taxation

Experimental design: 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 $\hat{\beta}$ 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”
⇒ 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.

Practice problem

Proof this.

Hints:

- Assume for simplicity that U has finite support.
- Note that a (weighted) average of numbers is always at least as large as their minimum.
- Write the risk (Bayes or minimax) of any randomized assignment rule as (weighted) average of the risk of deterministic rules.

Solution

- 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

$$\begin{aligned} R^B(\delta, \pi) &= \sum_u \int R(\delta^u, \theta) d\pi(\theta) P(u) \\ &\geq \min_u \int R(\delta^u, \theta) d\pi(\theta) = \min_u R^B(\delta^u, \pi). \end{aligned}$$

- Similarly

$$\begin{aligned} R^{mm}(\delta) &= \sum_u \max_{\theta} R(\delta^u, \theta) P(u) \\ &\geq \min_u \max_{\theta} R(\delta^u, \theta) = \min_u R^{mm}(\delta^u). \end{aligned}$$

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:

$$f \sim GP(\mu, C),$$

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

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

- 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,

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

Practice problem

- Derive the posterior expectation $\hat{\beta}$ of β .
- Derive the risk of any deterministic treatment assignment vector d , assuming
 1. The estimator $\hat{\beta}$ is used.
 2. The loss function $(\hat{\beta} - \beta)^2$ is considered.

Solution

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

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

- The corresponding risk equals

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

Discrete optimization

- The optimal design solves

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

- Possible optimization algorithms:
 1. Search over random $\mathbf{d}$
 2. greedy algorithm
 3. simulated annealing

Variation of the problem

Practice problem

- Suppose that the researcher insists on estimating β using a simple comparison of means,

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

- Derive again the risk of any deterministic treatment assignment vector d , assuming
 1. The estimator $\hat{\beta}$ is used.
 2. The loss function $(\hat{\beta} - \beta)^2$ is considered.

Solution

- Notation:

- Let $\mu_i^d = \mu(X_i, d)$ and $C_{i,j}^{d^1, d^2} = C((X_i, d^1), (X_j, d^2))$.
- Collect these terms in the vectors μ^d and matrices C^{d^1, d^2} , and let $\tilde{\mu} = (\mu^1, \mu^2)$,
$$\tilde{C} = \begin{pmatrix} C^{00} & C^{01} \\ C^{10} & C^{11} \end{pmatrix}.$$
- Weights

$$\begin{aligned} w &= (w^0, w^1), \\ w_i^1 &= \frac{d_i}{n_1} - \frac{1}{n}, \\ w_i^0 &= -\frac{1-d_i}{n_0} + \frac{1}{n}. \end{aligned}$$

- Risk: Sum of variance and squared bias,

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

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 $\hat{\beta}$ equals $(\bar{X}^1 - \bar{X}^0)' \cdot \gamma$, prior expected squared bias

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

- Mean squared error

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

- $\Rightarrow$ 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 $(\bar{X}^1 - \bar{X}^0)$.

Review: The envelope theorem

- Policy parameter t
- Vector of individual choices x
- Choice set $\mathcal{X}$
- Individual utility $v(x, t)$
- Realized choices

$$x(t) \in \operatorname{argmax}_{x \in \mathcal{X}} v(x, t).$$

- Realized utility

$$V(t) = \max_{x \in \mathcal{X}} v(x, t) = v(x(t), t)$$

- Let $x^* = x(t^*)$ for some fixed t^*
- Define

$$\tilde{V}(t) = V(t) - v(x^*, t) \quad (1)$$

$$= v(x(t), t) - v(x(t^*), t)$$

$$= \max_{x \in \mathcal{X}} v(x, t) - v(x^*, t). \quad (2)$$

- Definition of $\tilde{V}$ immediately implies:
 - $\tilde{V}(t) \geq 0$ for all t and $\tilde{V}(t^*) = 0$.
 - Thus: t^* is a global minimizer of $\tilde{V}$.
- If $\tilde{V}$ is differentiable at t^* : $\tilde{V}'(t^*) = 0$
- Thus

$$V'(t^*) = \frac{\partial}{\partial t} v(x^*, t)|_{t=t^*},$$

- Behavioral responses don't matter for effect of policy change on individual utility!

Experimental design

Optimal insurance and taxation

Optimal insurance and taxation: Setup

- 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.

Practice problem

- Write the effect $u'(t)$ on social welfare u of an increase in t as a sum of mechanical and behavioral effects on individual welfare and insurer revenues.
- Set $u(0) = 0$ and integrate to obtain an expression for social welfare.

Solution

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

$$u'(t) = (\lambda - 1) \cdot m(t) - t \cdot m'(t) = \lambda m(t) - \frac{\partial}{\partial t}(t \cdot m(t)). \quad (3)$$

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

$$u(t) = \lambda \int_0^t m(x) dx - t \cdot m(t). \quad (4)$$

- 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 $m(\cdot)$,

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

Practice problem

- What is the prior distribution of $u(t) = \lambda \int_0^t m(x)dx - t \cdot m(t)$?
- What is the prior covariance of $u(t)$ and Y given T ?
- What is the posterior expectation of $u(t)$ given Y and T ?

Solution

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

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

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

$$\begin{aligned} \text{Var}(u(t)) &= \lambda^2 \cdot \int_0^t \int_0^t C(x, x') dx' dx \\ &\quad - 2\lambda t \cdot \int_0^t C(x, t) dx + t^2 \cdot C(t, t). \end{aligned}$$

- Covariance with data:

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

- Posterior expectation of $u(t)$:

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

Optimal policy choice

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

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

- First order condition

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

where $B(t) = (B(t, T_1), \dots, B(t, T_n))$ and

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

Production objective

- Another important class of policy problems:
- Observable outcome Y_i (e.g. student test scores)
- Input vector $T_i \in \mathbb{R}^{d_t}$ (e.g., teachers per student, ...)
- (educational) production function

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

- Policy maker's objective is to maximize average (expected) outcomes $E[Y_i]$ across schools, net of the cost of inputs.
- Unit-price of input j : p_j .
- Willingness to pay for a unit-increase in Y : λ

- Yields the objective function

$$u(t) = \lambda \cdot m(t) - p \cdot t.$$

- Same type of data and prior as before.
- Posterior expectation:

$$\hat{u}(t) = v(t) + D(t) \cdot [C + \sigma^2 I]^{-1} \cdot (Y - \mu),$$

$$v(t) = \lambda \cdot \mu(t) - p \cdot t,$$

$$D(t, t') = \lambda \cdot C(t, t').$$

- First order condition:

$$\hat{u}'(\hat{t}^*) = v'(\hat{t}^*) + B(\hat{t}^*) \cdot [C + \sigma^2 I]^{-1} \cdot (Y - \mu) = 0.$$

where now $B(t, t') = \lambda \cdot \frac{\partial}{\partial t} C(t, t')$.

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

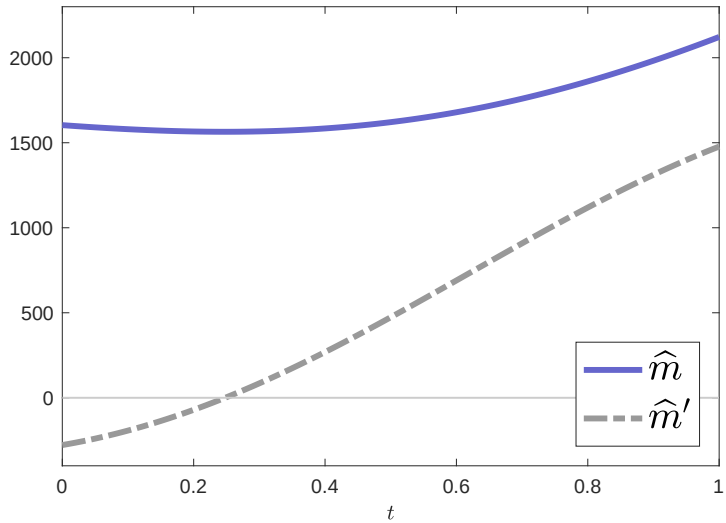
	(1)	(2)	(3)	(4)
	Share with any	Spending in \$	Share with any	Spending in \$
Free Care	0.931 (0.006)	2166.1 (78.76)	0.932 (0.006)	2173.9 (72.06)
25% Coinsurance	0.853 (0.013)	1535.9 (130.5)	0.852 (0.012)	1580.1 (115.2)
50% Coinsurance	0.832 (0.018)	1590.7 (273.7)	0.826 (0.016)	1634.1 (279.6)
95% Coinsurance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

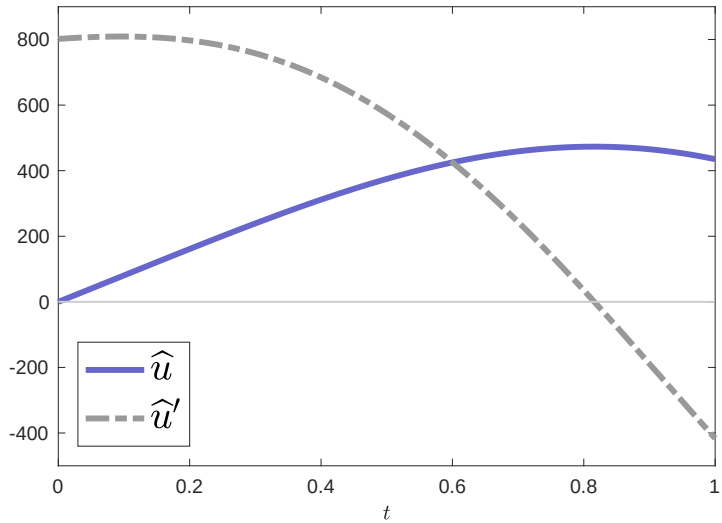
Assumptions

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)





References

- Application to experimental design:

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

- Envelope theorem:

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. Econometrica, 70(2):583–601.

- Application to optimal insurance and taxation:

Kasy, M. (2018). Optimal taxation and insurance using machine learning – sufficient statistics and beyond. Journal of Public Economics, 167.