Approximate Cross-Validation and Dynamic Experiments for Policy Choice

Maximilian Kasy

Department of Economics, Harvard University

April 23, 2018

Introduction

- Two separate, early stage projects:
  1. Approximate cross-validation
     - First order approximation to leave-one-out estimator.
     - Relationship to Stein's unbiased risk estimator.
     - Accelerated tuning.
     - Joint with Lester Mackey, MSR.
  2. Dynamic experiments for policy choice
     - Experimental design problem for choosing discrete treatment.
     - Goal: maximize average outcome.
     - Multiple waves.
     - Joint with Anja Sautman, J-PAL.
- Feedback appreciated!

Project 1: Approximate cross-validation

- Different ways of estimating risk (mean squared error):
  - Covariance penalties,
  - Stein's Unbiased Risk Estimate (SURE),
  - Cross-validation (CV).
- Result 1:
  - Consider repeated draws of some vector.
  - Then CV for estimating mean is approximately equal to SURE.
  - Without normality, unknown variance!
- Result 2:
  - Consider penalized M-estimation problem.
  - Then CV for prediction loss is approximately equal to in-sample risk plus penalty,
  - with a simple penalty based on gradient, Hessian.
- $\Rightarrow$ algorithm for accelerated tuning!

The normal means model

- $\theta, X \in \mathbb{R}^k$
- $X \sim N(\theta, \Sigma)$
- Estimator $\widehat{\theta}(X)$ of θ ("almost differentiable")
- Mean squared error:

$$\begin{aligned} MSE(\widehat{\theta}, \theta) &= \tfrac{1}{k} E_\theta \left[\|\widehat{\theta} - \theta\|^2 \right] \\ &= \tfrac{1}{k} \sum_j E_\theta \left[(\widehat{\theta}_j - \theta_j)^2 \right]. \end{aligned}$$

- Would like to estimate $MSE(\widehat{\theta}, \theta)$.
  - Choose tuning parameters to minimize estimated MSE.
  - Choose between estimators to minimize estimated MSE.
  - Theoretical tool for proving dominance results.

Covariance penalty

- Efron (2004): Adding and subtracting θ_j gives

$$(\widehat{\theta}_j - X_j)^2 = (\widehat{\theta}_j - \theta_j)^2 + 2 \cdot (\widehat{\theta}_j - \theta_j)(\theta_j - X_j) + (\theta_j - X_j)^2.$$

- Thus $MSE(\widehat{\theta}, \theta) = \frac{1}{k} \sum_j MSE_j$, where

$$\begin{aligned} MSE_j &= E_\theta \left[(\widehat{\theta}_j - \theta_j)^2 \right] \\ &= E_\theta[(\widehat{\theta}_j - X_j)^2] + 2E_\theta[(\widehat{\theta}_j - \theta_j) \cdot (X_j - \theta_j)] - E_\theta \left[(X_j - \theta_j)^2 \right] \\ &= E_\theta[(\widehat{\theta}_j - X_j)^2] + 2\,\mathrm{Cov}_\theta(\widehat{\theta}_j, X_j) - \mathrm{Var}_\theta(X_j). \end{aligned}$$

- First term: In-sample prediction error (observed).
- Second term: Covariance penalty (depends on unobserved θ).
- Third term: Doesn't depend on $\widehat{\theta}$.

Stein's Unbiased Risk Estimate

- Using partial integration and fact that $\varphi'(x) = -x \cdot \varphi(x)$, can show

$$MSE = \frac{1}{k} E_\theta \left[\|\widehat{\theta} - X\|^2 + 2\,\mathrm{trace}\left(\widehat{\theta}' \cdot \Sigma\right) - \mathrm{trace}(\Sigma) \right].$$

- All terms on the right hand side are observed! Sample version:

$$SURE = \frac{1}{k} \left(\|\widehat{\theta} - X\|^2 + 2\,\mathrm{trace}\left(\widehat{\theta}' \cdot \Sigma\right) - \mathrm{trace}(\Sigma) \right).$$

- Key assumptions that we used:
  - X is normally distributed.
  - Σ is known.
  - $\widehat{\theta}$ is almost differentiable.

Cross-validation

- Assume panel structure: X is a sample average, $i = 1, \ldots, n$ and $j = 1, \ldots, k$,

$$X = \frac{1}{n}\sum_i Y_i, \qquad Y_i \sim^{i.i.d.} (\theta, n \cdot \Sigma).$$

- Leave-one-out mean and estimator:

$$X_{-i} = \frac{1}{n-1}\sum_{i' \neq i} Y_{i'}, \qquad \widehat{\theta}_{-i} = \widehat{\theta}(X_{-i}).$$

- n-fold cross-validation:

$$CV = \frac{1}{n}\sum_i CV_i, \qquad CV_i = \| Y_i - \widehat{\theta}_{-i} \|^2.$$

Large n: $SURE \approx CV$

Proposition

Suppose $\widehat{\theta}(\cdot)$ is continuously differentiable in a neighborhood of θ, and suppose $X^n = \frac{1}{n}\sum_i Y_i^n$ with $(Y_i^n - \theta)/\sqrt{n}$ i.i.d. with expectation 0 and variance Σ. Let $\widehat{\Sigma} = \frac{1}{n^2}\sum_i (Y_i^n - X^n)(Y_i^n - X^n)'$. Then

$$CV^n = \|X^n - \widehat{\theta}^n\|^2 + 2\,\text{trace}\left(\widehat{\theta}' \cdot \widehat{\Sigma}^n\right) + (n-1)\,\text{trace}(\widehat{\Sigma}^n) + o_p(1)$$

as $n \to \infty$.

- New result, I believe.
- "For large n, CV is the same as SURE, plus the irreducible forecasting error" $n \cdot \text{trace}(\Sigma) = E_\theta[\|Y_i - \theta\|^2]$.
- Does **not** require
  - normality,
  - known Σ!

Sketch of proof

- Let $s = \sqrt{n-1}$, omit superscript n,

$$
\begin{aligned}
U_i &= \tfrac{1}{s}(Y_i - X) & U_i &\sim (0, \Sigma),\\
X_{-i} &= X - \tfrac{1}{s}U_i & Y_i &= X + sU_i\\
\widehat{\theta}(X_{-i}) &= \widehat{\theta}(X) - \tfrac{1}{s}\widehat{\theta}'(X)\cdot U_i + \Delta_i & \Delta_i &= o(\tfrac{1}{s}U_i)\\
\widehat{\Sigma} &= \tfrac{1}{n}\sum_i U_i U_i'.
\end{aligned}
$$

- Then

$$
\begin{aligned}
CV_i = \|Y_i - \widehat{\theta}_{-i}\|^2 &= \|X + sU_i - (\widehat{\theta} - \tfrac{1}{s}\widehat{\theta}'(X)\cdot U_i + \Delta_i)\|^2\\
&= \|X - \widehat{\theta}\|^2 + 2\left\langle U_i, \widehat{\theta}'(X)\cdot U_i\right\rangle + s^2\|U_i\|^2\\
&\quad + 2\left\langle X - \widehat{\theta}, (s + \tfrac{1}{s}\widehat{\theta}')U_i\right\rangle + \left(\tfrac{1}{s^2}\|\widehat{\theta}'(X)\cdot U_i\|^2 + 2\left\langle \Delta_i, Y_i - \widehat{\theta}_{-i}\right\rangle\right).
\end{aligned}
$$

$$
CV = \tfrac{1}{n}\sum_i CV_i = \|X - \widehat{\theta}\|^2 + 2\,\mathrm{trace}\left(\widehat{\theta}' \cdot \widehat{\Sigma}\right) + (n-1)\,\mathrm{trace}(\widehat{\Sigma})
$$

$$
+ 0 + o_p(\tfrac{1}{n}).
$$

More general setting: Penalized M-estimation

- Suppose $\beta = \text{argmin}_b E[m(X,\beta)]$.
- Estimate β using penalized M-estimation,

$$\widehat{\beta}(\lambda) = \underset{b}{\text{argmin}} \sum_i m(X_i, b) + \pi(b, \lambda).$$

- Would like to choose λ to minimize the out-of-sample prediction error

$$R(\lambda) = E[m(X, \widehat{\beta}(\lambda))].$$

- Leave-one-out estimator, n-fold cross-validation

$$\widehat{\beta}_{-i}(\lambda) = \underset{b}{\text{argmin}} \sum_{j \neq i} m(X_j, b) + \pi(b, \lambda).$$

$$CV(\lambda) = \frac{1}{n} \sum_i m(X_i, \widehat{\beta}_{-i}(\lambda)).$$

- Computationally costly to re-estimate β for every choice of i and λ!
- Notation for Hessian, gradients:

$$H = \left(\sum_j m_{bb}(X_j, \widehat{\beta}(\lambda)) + \pi_{bb}(\widehat{\beta}(\lambda), \lambda) \right)$$
$$g_i = m_b(X_i, \widehat{\beta}(\lambda)).$$

- First-order approximation to leave-one-out estimator (assuming 2nd derivatives):

$$\widehat{\beta}_{-i}(\lambda) - \widehat{\beta}(\lambda) \approx H^{-1} \cdot g_i.$$

- In-sample prediction error:

$$\bar{R}(\lambda) = \frac{1}{n} \sum_i m(X_i, \widehat{\beta}(\lambda)).$$

- Another first-order approximation:

$$CV(\lambda) \approx \bar{R}(\lambda) + \frac{1}{n}\sum_i g_i \cdot \left(\widehat{\beta}_{-i}(\lambda) - \widehat{\beta}(\lambda)\right).$$

- Combining the two approximations:

$$CV(\lambda) \approx \bar{R}(\lambda) + \frac{1}{n}\sum_i g_i^t \cdot H^{-1} \cdot g_i.$$

- $\bar{R}$, g_i and H are automatically available if Newton-Raphson was used for finding $\widehat{\beta}(\lambda)$!
- If not, could approximate then without bias using random subsample.

Open questions

- Implementation!
- Regularity conditions for validity of approximations?
- Gains of speed in tuning, e.g., neural nets?
- Gains of efficiency relative to wasteful sample-partition methods?

Project 2: Dynamic experiments for policy choice

- Setup:
  - Optimal treatment assignment (multiple treatments)
  - in multi-wave experiments.
  - Goal: After experiment, choose a policy
  - to maximize welfare (average outcome net of costs).
- Dynamic stochastic optimization problem,
- used normatively (for experimenter) rather than descriptively (as in structural models).
- Solution via exact backward induction.
- Outline:
  1. Setup: $\bar{d}$ treatments, binary outcomes, T waves
  2. Objective function: social welfare, max over treatment
  3. Independent Beta priors for mean potential outcomes
  4. Value functions, backward induction

Setup

- Waves $t = 1, \dots, T$, sample sizes N_t.
- Treatment $D \in \{1, \dots, \bar{d}\}$, outcomes $Y \in \{0, 1\}$, potential outcomes Y^d,

$$Y_{it} = \sum_{d=1}^{\bar{d}} \mathbf{1}(D_{it} = d) Y_{it}^d.$$

- $(Y_{it}^0, \dots, Y_{it}^{\bar{d}})$ are i.i.d. across both i and t.
- Denote

$$\begin{aligned} \theta^d &= E[Y_t^d] \\ n_t^d &= \sum_i \mathbf{1}(D_{it} = d) \\ s_t^d &= \sum_i \mathbf{1}(D_{it} = d, Y_{it} = Y_{it}^d = 1). \end{aligned}$$

Treatment assignment, outcomes, state space

- Treatment assignment in wave t: $\boldsymbol{n}_t = (n_t^1, \ldots, n_t^{\bar{d}})$.
- Outcomes of wave t: $\boldsymbol{s}_t = (s_t^1, \ldots, s_t^{\bar{d}})$.
- Cumulative versions: $M_t = \sum_{t' \leq t} N_{t'}$,

$$\boldsymbol{m}_t = (m_t^1, \ldots, m_t^{\bar{d}}) = \sum_{t' \leq t} \boldsymbol{n}_t$$

$$\boldsymbol{r}_t = (s_t^1, \ldots, s_t^{\bar{d}}) = \sum_{t' \leq t} \boldsymbol{s}_t.$$

- Relevant information for the experimenter in period $t+1$ is summarized by $\boldsymbol{m}_t$ and $\boldsymbol{r}_t$.

Design objective

- Policy objective $SW(d)$:
  Average outcome Y, net of the cost of treatment.
- Choose treatment d after experiment is completed.
- Posterior expected social welfare:

$$SW(d) = E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] - c^d,$$

where c^d is the unit cost of implementing policy d.

Bayesian prior and posterior

- By definition, $Y^d|\theta \sim Ber(\theta^d)$.
- Prior: $\theta^d \sim Beta(\alpha_0^d, \beta_0^d)$, independent across d.
- Posterior after period t:

$$\begin{aligned}\theta^d|\boldsymbol{m}_t, \boldsymbol{r}_t &\sim Beta(\alpha_t^d, \beta_t^d)\\ \alpha_t^d &= \alpha_0^d + r_t^d\\ \beta_t^d &= \beta_0^d + m_t^d - r_t^d.\end{aligned}$$

- In particular,

$$SW(d) = \frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d.$$

Dynamic optimization problem

- Dynamic optimization problem:
  - States $(\boldsymbol{m}_t, \boldsymbol{r}_t) \in \{0, \ldots, M_{t-1}\}^{2\bar{d}}$,
  - actions $\boldsymbol{n}_t \in \{0, \ldots, N_t\}^{\bar{d}}$,
  - transitions

$$\boldsymbol{m}_t = \boldsymbol{m}_{t-1} + \boldsymbol{n}_t$$
$$\boldsymbol{r}_t = \boldsymbol{r}_{t-1} + \boldsymbol{s}_t$$
$$P(s_t^d = s | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, n_t^d) = \binom{n_t^d}{s} \frac{B(\alpha_{t-1}^d + s, \beta_{t-1}^d + n_t^d - s)}{B(\alpha_{t-1}^d, \beta_{t-1}^d)}.$$

(Beta-binomial distribution)

Value functions

- Solve for the optimal experimental design using backward induction.
- Finite state space, finite time horizon: Exact solution can be computed for moderate dimensions.
- Denote by V_t the value function after completion of wave t.
- Starting at the end, we have

$$\begin{aligned} V_T(\boldsymbol{m}_T, \boldsymbol{r}_T) &= \max_d \left(E[\theta^d | \boldsymbol{m}_T, \boldsymbol{s}_T] - c^d \right) \\ &= \max_d \left(\frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d \right). \end{aligned}$$

Backward induction

- Value function before completion of wave t:

$$U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = E\left[V_t\left(\boldsymbol{m}_{t-1} + \boldsymbol{n}_t, \boldsymbol{r}_{t-1} + \boldsymbol{s}_t\right) \middle| \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t\right],$$

- Expectation is taken over the Beta-binomial distribution.
- Period t value function and the optimal experimental design satisfy

$$\begin{aligned} V_{t-1}(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \max_{\boldsymbol{n}_t :\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) \\ \boldsymbol{n}_t^*(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \operatorname*{argmax}_{\boldsymbol{n}_t :\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t). \end{aligned}$$

Open questions

- Numerical implementation when exact solution is not computationally feasible?
- State space explodes for larger N_t, $\bar{d}$, T!
  Possibly via interpolation of value functions?
- Characterization of solutions: Non-concavity of the value of information! (E-max and option value)
- Generalizations: Allowing for covariates, continuous outcomes, dependency structures in prior.
- Implementation in actual experiments.

Thank you!