Estimating risk

Maximilian Kasy

Department of Economics, Harvard University

May 4, 2018

Introduction

- Some of the topics about which I learned from Gary:
  - The normal means model.
  - Finite sample risk and point estimation.
  - Shrinkage and tuning.
  - Random coefficients and empirical Bayes.
- This talk:
  - Brief review of these topics.
  - Building on that, some new results from my own work.

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:

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

- Would like to estimate $MSE(\widehat{\theta}, \theta)$, to
  1. choose tuning parameters to minimize estimated MSE,
  2. choose between estimators to minimize estimated MSE,
  3. as a theoretical tool for proving dominance results.
- Key ingredient for machine learning!

Roadmap

- Review:
  - Covariance penalties,
  - Stein's Unbiased Risk Estimate (SURE),
  - Cross-Validation (CV).
- Panel version of (normal) means model:
  - $X \in \mathbb{R}^k$ as sample mean of n i.i.d. draws Y_i.
  - $\Rightarrow$ n-fold Cross-Validation.
- Two results that are new (I think):
  - Large $n \Rightarrow$ CV approximates SURE.
  - Large $k \Rightarrow$ CV and SURE converge to MSE,
    yield oracle optimal tuning ("uniform loss consistency").

References

- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151
- Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632
- Abadie, A. and Kasy, M. (2018). Choosing among regularized estimators in empirical economics. *Working Paper.*
- Fessler, P. and Kasy, M. (2018). How to use economic theory to improve estimators: Shrinking toward theoretical restrictions. *Working Paper*
- Kasy, M. and Mackey, L. (2018). Approximate cross-validation. *Work in progress*

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: Irreducible prediction error, doesn't depend on $\widehat{\theta}$.

Stein's Unbiased Risk Estimate

- Stein (1981): For normal pdf with variance σ^2,

$$\varphi'_\sigma(x-\theta) = -\frac{x-\theta}{\sigma} \cdot \varphi_\sigma(x-\theta).$$

- Suppose for a moment that $\Sigma = \sigma^2 I$.
- Then, by partial integration,

$$\begin{aligned}
\mathrm{Cov}_\theta(\widehat{\theta}_j, X_j) &= \int E_\theta[\widehat{\theta}_j | X_j = x_j](x_j - \theta_j)\varphi_\sigma(x_j - \theta_j)dx_j \\
&= \sigma \cdot \int -E_\theta[\widehat{\theta}_j | X_j = x_j]\varphi'_\sigma(x_j - \theta_j)dx_j \\
&= \sigma \cdot \int \partial_{x_j} E_\theta[\widehat{\theta}_j | X_j = x_j]\varphi_\sigma(x_j - \theta_j)dx_j \\
&= \sigma \cdot E_\theta[\partial_{X_j}\widehat{\theta}_j].
\end{aligned}$$

- Thus

$$MSE = \frac{1}{k}\sum_j MSE_j = \frac{1}{k}\sum_j E_\theta\left[(\widehat{\theta}_j - X_j)^2 + 2\sigma^2 \cdot \partial_{X_j}\widehat{\theta}_j - \sigma^2\right].$$

- For non-diagonal Σ, by change of coordinates we get more generally

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

- 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.

Panel setting and 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*

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

as $n \to \infty$.

- New result, I believe.
- "For large n, CV is the same as SURE, plus the irreducible forecasting error" $n \cdot \mathrm{trace}(\Sigma) = E_\theta[\|Y_i - \theta\|^2]$.
- 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

$$\begin{aligned} CV_i = \|Y_i - \widehat{\theta}_{-i}\|^2 &= \|X + sU_i - (\widehat{\theta} - \tfrac{1}{s}\widehat{\theta}'(X)\cdot U_i + \Delta_i)\|^2 \\ &= {\color{blue}\|X - \widehat{\theta}\|^2} + {\color{red}2\left\langle U_i, \widehat{\theta}'(X)\cdot U_i\right\rangle} + s^2\|U_i\|^2 \\ &\quad +2\left\langle X-\widehat{\theta},(s+\tfrac{1}{s}\widehat{\theta}')U_i\right\rangle+\left(\tfrac{1}{s^2}\|\widehat{\theta}'(X)\cdot U_i\|^2+2\left\langle\Delta_i,Y_i-\widehat{\theta}_{-i}\right\rangle\right). \\ CV = \tfrac{1}{n}\sum_i CV_i &= {\color{blue}\|X - \widehat{\theta}\|^2} + {\color{red}2\,\mathrm{trace}\left(\widehat{\theta}'\cdot\widehat{\Sigma}\right)} + (n-1)\,\mathrm{trace}(\widehat{\Sigma}) \\ &\quad +0+o_p(\tfrac{1}{n}). \end{aligned}$$

Large k: $SURE, CV \approx MSE$

- Abadie and Kasy (2018): Random effects (empirical Bayes) perspective:

$$(X_j, \theta_j) \sim^{i.i.d.} \pi, \qquad E_\pi[X_j|\theta_j] = \theta_j.$$

- Unbiasedness of SURE, CV:

$$E_\theta[SURE] = MSE, \qquad E_\theta[CV] = E_\theta[CV_i] = MSE^{n-1}.$$

- Law of large numbers: For fixed π, n,

$$\text{plim}_{k\to\infty} SURE - MSE = 0 \qquad \text{plim}_{k\to\infty} CV - MSE^{n-1} = 0.$$

- Questions:
  - Does this hold uniformly over π?
  - If so, does this yield oracle-optimal tuning parameters?

Componentwise estimators

- Answer requires more structure on estimators. Assume

$$\widehat{\theta}_j = m(X_j, \lambda).$$

  Examples:
  - Ridge: $m_R(x,\lambda) = \frac{1}{1+\lambda} x$.
  - Lasso: $m_L(x,\lambda) = \mathbf{1}(x < -\lambda)(x+\lambda) + \mathbf{1}(x > \lambda)(x-\lambda)$.
- Denote

$$SE(\lambda) = \frac{1}{k}\sum_{j=1}^{k}(m(X_j,\lambda)-\theta_j)^2, \qquad \text{(squared error loss)}$$

$$MSE(\lambda) = E_\theta[SE(\lambda)], \qquad \text{(compound risk)}$$

$$\overline{MSE}(\lambda) = E_\pi[MSE(\lambda)] = E_\pi[SE(\lambda)], \qquad \text{(empirical Bayes risk)}$$

- and $\widehat{MSE}(\lambda)$ an estimator of MSE, e.g. SURE or CV.

Theorem (Uniform loss consistency)

Assume that, as $k \to \infty$,

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0,\infty]} \left| SE(\lambda) - \overline{MSE}(\lambda) \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0,$$

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0,\infty]} \left| \widehat{MSE}(\lambda) - \overline{MSE}(\lambda) - v_\pi \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0.$$

Then

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\left| SE(\widehat{\lambda}) - \inf_{\lambda \in [0,\infty]} SE(\lambda) \right| > \varepsilon \right) \to 0, \quad \forall \varepsilon > 0,$$

where $\widehat{\lambda} \in \operatorname{argmin}_{\lambda \in [0,\infty]} \widehat{MSE}(\lambda)$.

Theorem (Uniform convergence)

Suppose that $\sup_{\pi \in \mathcal{Q}} E_\pi[X^4] < \infty$*. Under some conditions on m (satisfied for Ridge and Lasso), the assumptions of the previous theorem are satisfied.*

Remarks:

- Extension of Glivenko-Cantelli theorem.
- Need conditions on m to get uniformity over λ.
- Only need (and get) uniform convergence of $\widehat{MSE} - \overline{MSE} - v_\pi$ to 0 for some constant v_π.
- For CV, get uniform loss consistency to the estimator using λ optimal for SE^{n-1} (thus shrinking a bit too much for small n). $n \approx$ sample size / # of parameters

Outlook and work in progress

1. Approximate CV using first-order approx to leave-1-out estimator, in penalized M-estimator settings:

$$\widehat{\beta}_{-i}(\lambda) - \widehat{\beta}(\lambda) \approx \left(\sum_j m_{bb}(X_j, \widehat{\beta}(\lambda)) + \pi_{bb}(\widehat{\beta}(\lambda), \lambda) \right)^{-1} \cdot m_b(X_i, \widehat{\beta}(\lambda)).$$

   - Fast alternative to CV for tuning of neural nets, etc.
   - Additional acceleration by only calculating this for subset of i, j.
2. Risk reductions for shrinkage toward *inequality* restrictions.
   - Relevant for many restrictions implied by economic theory.
   - Proving uniform dominance using SURE, extending James-Stein.
   - Open question: Smooth choice of "degrees of freedom" that is not too conservative.

Thank you!