Econ 2140, spring 2018, Part IIc
Applications of Statistical Decision Theory

Maximilian Kasy

Department of Economics, Harvard University

Takeaways for this part of class

- In a Normal means model with Normal prior, there are a number of equivalent ways to think about regularization:
  - posterior mean,
  - penalized least squares (penalty corresponds to prior),
  - shrinkage, etc.
- Applied to linear regression:
  Ridge regression as penalized OLS with quadratic penalty.
- Alternatively, using an absolute value penalty:
  Lasso regression. (Popular in machine learning.)
- Hierarchical normal models yield the estimators used in the "value added" literature.

Roadmap

Normal posterior means – equivalent representations

Ridge regression

Lasso regression

Value added models

Normal posterior means – equivalent representations
Setup

- $\theta \in \mathbb{R}^k$
- $\boldsymbol{X}|\theta \sim N(\theta, I_k)$
- Loss

$$L(\widehat{\theta}, \theta) = \sum_i (\widehat{\theta}_i - \theta_i)^2$$

- Prior

$$\theta \sim N(0, C)$$

6 equivalent representations of the posterior mean

1. Minimizer of weighted average risk
2. Minimizer of posterior expected loss
3. Posterior expectation
4. Posterior best linear predictor
5. Penalized least squares estimator
6. Shrinkage estimator

1) Minimizer of weighted average risk

- Minimize weighted average risk (= Bayes risk),
- averaging loss $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ over both
  1. the sampling distribution $f_{\boldsymbol{X}|\theta}$, and
  2. weighting values of θ using the decision weights (prior) π_θ.
- Formally,

$$\widehat{\theta}(\cdot) = \underset{t(\cdot)}{\operatorname{argmin}} \int E_\theta[L(t(\boldsymbol{X}), \theta)] d\pi(\theta).$$

2) Minimizer of posterior expected loss

- Minimize posterior expected loss,
- averaging loss $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ over
  1. just the posterior distribution $\pi_{\theta|\boldsymbol{X}}$.
- Formally,

$$\widehat{\theta}(x) = \underset{t}{\text{argmin}} \int L(t, \theta) d\pi_{\theta|\boldsymbol{X}}(\theta|x).$$

3 and 4) Posterior expectation and posterior best linear predictor

- Note that

$$\begin{pmatrix} X \\ \theta \end{pmatrix} \sim N\left(0, \begin{pmatrix} C+I & C \\ C & C \end{pmatrix}\right).$$

- Posterior expectation:

$$\widehat{\theta} = E[\theta|\boldsymbol{X}].$$

- Posterior best linear predictor:

$$\widehat{\theta} = E^*[\theta|\boldsymbol{X}] = C \cdot (C+I)^{-1} \cdot \boldsymbol{X}.$$

5) Penalization

- Minimize
  1. the sum of squared residuals,
  2. plus a quadratic penalty term.
- Formally,

$$\widehat{\theta} = \underset{t}{\operatorname{argmin}} \sum_{i=1}^{n} (X_i - t_i)^2 + \|t\|^2,$$

- where

$$\|t\|^2 = t' C^{-1} t.$$

6) Shrinkage

- Diagonalize C: Find
  1. orthonormal matrix U of eigenvectors, and
  2. diagonal matrix D of eigenvalues, so that

$$C = UDU'.$$

- Change of coordinates, using U:

$$\tilde{\boldsymbol{X}} = U'\boldsymbol{X}$$
$$\tilde{\theta} = U'\theta.$$

- Componentwise shrinkage in the new coordinates:

$$\widehat{\tilde{\theta}}_i = \frac{d_i}{d_i + 1}\tilde{X}_i. \tag{1}$$

Practice problem

Show that these 6 objects are all equivalent to each other.

Solution (sketch)

1. Minimizer of weighted average risk = minimizer of posterior expected loss: See decision slides.
2. Minimizer of posterior expected loss = posterior expectation:
   - First order condition for quadratic loss function,
   - pull derivative inside,
   - and switch order of integration.
3. Posterior expectation = posterior best linear predictor:
   - $\boldsymbol{X}$ and θ are jointly Normal,
   - conditional expectations for multivariate Normals are linear.
4. Posterior expectation $\Rightarrow$ penalized least squares:
   - Posterior is symmetric unimodal $\Rightarrow$ posterior mean is posterior mode.
   - Posterior mode = maximizer of posterior log-likelihood = maximizer of joint log likelihood,
   - since denominator $f_{\boldsymbol{X}}$ does not depend on θ.

Solution (sketch) continued

5. Penalized least squares $\Rightarrow$ posterior expectation:
   - Any penalty of the form
   $$t'At$$
   for A symmetric positive definite
   - corresponds to the log of a Normal prior
   $$\theta \sim N\left(0, A^{-1}\right).$$
6. Componentwise shrinkage $=$ posterior best linear predictor:
   - Change of coordinates turns $\widehat{\theta} = C \cdot (C+I)^{-1} \cdot \boldsymbol{X}$ into
   $$\widehat{\tilde{\theta}} = D \cdot (D+I)^{-1} \cdot \boldsymbol{X}.$$
   - Diagonality implies
   $$D \cdot (D+I)^{-1} = \text{diag}\left(\frac{d_i}{d_i+1}\right).$$

Normal prior for linear regression

- Normal linear regression model:
- 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.
- $Y_i = X_i \cdot \beta + \varepsilon_i$
- $\varepsilon_i | \boldsymbol{X}, \beta \sim N(0, \sigma^2)$
- $\beta | \boldsymbol{X} \sim N(0, \Omega)$ (prior)
- Note: will leave conditioning on $\boldsymbol{X}$ implicit in following slides.

Practice problem ("weight space view")

- Find the posterior expectation of β
- Hints:
  1. The posterior expectation is the maximum a posteriori.
  2. The log likelihood takes a penalized least squares form.
- Find the posterior expectation of $x \cdot \beta$ for some (non-random) point x.

Solution

- Joint log likelihood of Y, β:

$$\log(f_{\boldsymbol{Y}\beta}) = \log(f_{\boldsymbol{Y}Y|\beta}) + \log(f_\beta)$$
$$= const. - \frac{1}{2\sigma^2}\sum_i (Y_i - X_i\beta)^2 - \frac{1}{2}\beta'\Omega^{-1}\beta.$$

- First order condition for maximum a posteriori:

$$0 = \frac{\partial f_{\boldsymbol{Y}\beta}}{\partial \beta} = \frac{1}{\sigma^2}\sum_i (Y_i - X_i\beta)\cdot X_i - \beta'\Omega^{-1}.$$

$$\Rightarrow \quad \widehat{\beta} = \left(\sum_i X_i' X_i + \sigma^2\Omega^{-1}\right)^{-1} \cdot \sum X_i' Y_i.$$

- Thus

$$E[x\cdot\beta|\boldsymbol{Y}] = x\cdot\widehat{\beta} = x\cdot\left(\boldsymbol{X}'\boldsymbol{X} + \sigma^2\Omega^{-1}\right)^{-1}\cdot\boldsymbol{X}'\boldsymbol{Y}.$$

- Previous derivation required inverting $k \times k$ matrix.
- Can instead do prediction inverting an $n \times n$ matrix.
- n might be smaller than k if there are many "features."
- This will lead to a "function space view" of prediction.

Practice problem ("kernel trick")

- Find the posterior expectation of

$$f(x) = E[Y|X = x] = x \cdot \beta.$$

- Wait, didn't we just do that?
- Hints:
  1. Start by figuring out the variance / covariance matrix of $(x \cdot \beta, \boldsymbol{Y})$.
  2. Then deduce the best linear predictor of $x \cdot \beta$ given $\boldsymbol{Y}$.

Solution

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

$$\begin{pmatrix} x \cdot \beta \\ \boldsymbol{Y} \end{pmatrix} \sim N \left(0, \begin{pmatrix} x\Omega x' & x\Omega \boldsymbol{X}' \\ \boldsymbol{X}\Omega x' & \boldsymbol{X}\Omega \boldsymbol{X}' + \sigma^2 I_n \end{pmatrix} \right)$$

- Denote $C = X\Omega X'$ and $c(x) = x\Omega X'$.
- Then

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

- Contrast with previous representation:

$$E[x \cdot \beta | \boldsymbol{Y}] = x \cdot \left(\boldsymbol{X}' \boldsymbol{X} + \sigma^2 \Omega^{-1} \right)^{-1} \cdot \boldsymbol{X}' \boldsymbol{Y}.$$

Lasso regression

- Ridge regression as penalization:
  Assume $\text{Var}(\beta) = \boldsymbol{I}$, denote $\sigma^2 = \lambda$ and $\|\beta\|_2^2 = \beta' \cdot \beta$. Then

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \sum_i (Y_i - X_i\beta)^2 + \lambda \|\beta\|_2^2,$$

- Could consider alternative penalties.
- For instance: $\|\beta\|_1 = \sum_j |\beta_j|$.
- This yields Lasso regression:

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \sum_i (Y_i - X_i\beta)^2 + \lambda \|\beta\|_1.$$

Lasso, simplified setting

- Consider again the normal means setting, as before, where $\boldsymbol{X}|\theta \sim N(\theta, I_k)$.
- Let

$$\widehat{\theta} = \underset{t}{\operatorname{argmin}} \sum_{i=1}^{n} (X_i - t_i)^2 + 2\lambda \|t\|_1.$$

Practice problem

Derive an explicit formula for $\widehat{\theta}$

Solution

- We can treat each component separately:

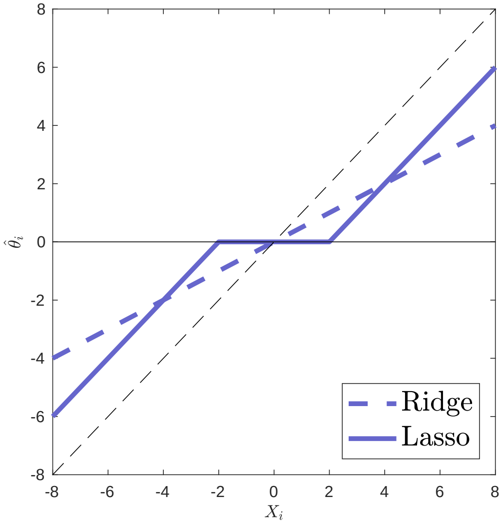
$$\widehat{\theta}_i = \underset{t_i}{\operatorname{argmin}} \ \tfrac{1}{2}(X_i - t_i)^2 + \lambda |t_i|.$$

- Sub-derivative of objective function:

$$\frac{\partial}{\partial t_i} = -(X_i - t_i) + \begin{cases} -1 & t_i < 0 \\ 0 & t_i = 0 \\ 1 & t_i < 0. \end{cases}$$

- Solution to optimization problem (first order condition):

$$\widehat{\theta}_i = \begin{cases} X_i + \lambda & X_i < -\lambda \\ 0 & -\lambda < X_i < \lambda \\ X_i - \lambda & \lambda < X_i. \end{cases}$$


8
6
4
2
0
-2
-4
-6
-8
$\hat{\theta}_i$
-8
-6
-4
-2
0
2
4
6
8
X_i
Ridge
Lasso

Comparing methods of regularized regression

- Abadie and Kasy (2017):
  compare the risk of alternative regularization methods.
- No one method that's always optimal!
- **Neighborhood effects:**
  The effect of location during childhood on adult income (Chetty and Hendren, 2015)
- **Arms trading event study:**
  Changes in the stock prices of arms manufacturers following changes in the intensity of conflicts in countries under arms trade embargoes (DellaVigna and La Ferrara, 2010)
- **Nonparametric Mincer equation:**
  A nonparametric regression equation of log wages on education and potential experience (Belloni and Chernozhukov, 2011)

Estimated Risk

- Estimated risk $\widehat{R}$ at the optimized tuning parameter $\widehat{\lambda}^*$
- for each application and estimator considered.

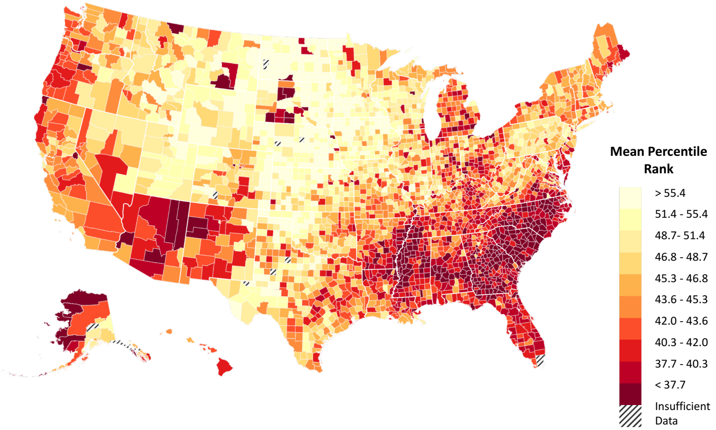
	n		Ridge	Lasso	Pre-test
location effects	595	$\widehat{R}$	**0.29**	0.32	0.41
		$\widehat{\lambda}^*$	2.44	1.34	5.00
arms trade	214	$\widehat{R}$	0.50	0.06	**-0.02**
		$\widehat{\lambda}^*$	0.98	1.50	2.38
returns to education	65	$\widehat{R}$	1.00	**0.84**	0.93
		$\widehat{\lambda}^*$	0.01	0.59	1.14

Value added models

- **Chetty, R. and N. Hendren (2015).** The impacts of neighborhoods on intergenerational mobility: Childhood exposure effects and county-level estimates. *Working Paper.*
- We are interested in the causal impact θ_i of cities i on intergenerational mobility.
- Suppose that for each city we have
  - a noisy but unbiased estimate Y_i of θ_i, with known standard error σ_i,
  - and a biased but less noisy estimate X_i.
- Suppose that estimates and parameters are jointly normally distributed,

$$Y_i|\theta_i, \sigma_i, X_i \sim N(\theta_i, \sigma_i^2)$$
$$(\theta_i, X_i)|\sigma_i \sim N((\bar{\theta}, \bar{X}), \Sigma).$$

The Geography of Intergenerational Mobility in the United States
Predicted Income Rank at Age 30 for Children with Parents at 25th Percentile



What is the Average Causal Impact of Growing Up in place with Better Outcomes?

Practice problem

- Suppose you know Σ and observe a draw of (Y_i, X_i, σ_i).
- What is the joint distribution of Y_i, X_i and θ_i?
- What is the posterior expectation of θ_i?

Solution

- Joint distribution of Y_i, X_i and θ_i:

$$(\theta_i, Y_i, X_i)|\sigma_i \sim N\left(\begin{pmatrix} \bar{\theta} \\ \bar{\theta} \\ \bar{X} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{11} & \Sigma_{11} + \sigma_i^2 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{12} & \Sigma_{22} \end{pmatrix} \right)$$

- Posterior expectation of θ_i = best linear predictor,

$$\begin{aligned} \widehat{\theta}_i &= E[\theta_i | Y_i, X_i, \sigma_i] \\ &= E[\theta_i] + \mathrm{Cov}(\theta_i, (Y_i, X_i)) \cdot \mathrm{Var}(Y_i, X_i)^{-1} \cdot ((Y_i, X_i) - E[(Y_i, X_i)]) \\ &= \bar{\theta} + (\Sigma_{11}, \Sigma_{12}) \cdot \begin{pmatrix} \Sigma_{11} + \sigma_i^2 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}^{-1} \cdot ((Y_i, X_i) - (\bar{\theta}, \bar{X})) \end{aligned}$$

Practice problem

Suppose you observe i.i.d. draws of (Y_i, X_i, σ_i). What is your estimate of $(\bar{\theta}, \bar{X})$ and of Σ?

Solution

- As shown before,

$$E[(Y_i, X_i)] = (\bar{\theta}, \bar{X}),$$
$$\text{Var}(Y_i, X_i)|\sigma_i = \begin{pmatrix} \Sigma_{11} + \sigma_i^2 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}.$$

- Replacing the expectation by the sample mean, we immediately get estimators of $(\bar{\theta}, \bar{X})$. Similarly,

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_i \begin{pmatrix} (Y_i - \bar{Y})^2 - \sigma_i^2 & (Y_i - \bar{Y})(X_i - \bar{X}) \\ (Y_i - \bar{Y})(X_i - \bar{X}) & (X_i - \bar{X})^2 \end{pmatrix}.$$