Bayesian Estimation and Frequentist Risk

STA 721: Lecture 9

Merlise Clyde (clyde@duke.edu)
Duke University

Outline

- Frequentist Risk of Bayes estimators
- Bayes and Penalized Loss Functions
- Generalized Ridge Regression
- Hierarchical Bayes and Other Penalties

Readings:

- Christensen Chapter 2.9 and Chapter 15
- Seber & Lee Chapter 10.7.3 and Chapter 12

Frequentist Risk of Bayes Estimators

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an ``action” $\mathbf{a}$ (i.e. reporting $\mathbf{a}$ as the estimate of $\boldsymbol{\beta}$)

$$\mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[L(\boldsymbol{\beta}, \mathbf{a})] = \mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[(\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})]$$

where the expectation is over the data $\mathbf{Y}$ given the true value of $\boldsymbol{\beta}$.

Expectation of Quadratic Forms

Theorem: Christensen Thm 1.3.2

If $\mathbf{W}$ is a random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then

$$\mathsf{E}[\mathbf{W}^T\mathbf{A}\mathbf{W}] = \mathsf{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$

Proof

$$(\mathbf{W}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{W}-\boldsymbol{\mu}) = \mathbf{W}^T\mathbf{A}\mathbf{W} - 2\boldsymbol{\mu}^T\mathbf{A}\mathbf{W} + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$
$$\mathsf{E}[(\mathbf{W}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{W}-\boldsymbol{\mu})] = \mathsf{E}[\mathbf{W}^T\mathbf{A}\mathbf{W}] - 2\boldsymbol{\mu}^T\mathbf{A}\mathsf{E}[\mathbf{W}] + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$

Rearranging we have

$$\mathsf{E}[\mathbf{W}^T\mathbf{A}\mathbf{W}] = \mathsf{E}[(\mathbf{W}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{W}-\boldsymbol{\mu})] + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$

▼ **Proof:** continued

Recall

$$\begin{aligned}\mathsf{E}[(\mathbf{W}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{W}-\boldsymbol{\mu})] &= \mathsf{E}[\mathsf{tr}((\mathbf{W}-\boldsymbol{\mu})\mathbf{A}(\mathbf{W}-\boldsymbol{\mu})^T)] \\ &= \mathsf{tr}(\mathsf{E}[\mathbf{A}(\mathbf{W}-\boldsymbol{\mu})(\mathbf{W}-\boldsymbol{\mu})^T]) \\ &= \mathsf{tr}(\mathbf{A}\mathsf{E}[(\mathbf{W}-\boldsymbol{\mu})(\mathbf{W}-\boldsymbol{\mu})^T]) \\ &= \mathsf{tr}(\mathbf{A}\boldsymbol{\Sigma})\end{aligned}$$

Therefore the expectation is

$$\mathsf{E}[\mathbf{W}^T\mathbf{A}\mathbf{W}] = \mathsf{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$

- Use Theorem to Explore Frequentist Risk of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta}-\mathbf{a})^T(\boldsymbol{\beta}-\mathbf{a})$$

compared to the OLS estimator $\hat{\boldsymbol{\beta}}$.

Steps to Evaluate Frequentist Risk

- MSE: $\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a}) = \mathsf{tr}(\boldsymbol{\Sigma}_{\mathbf{a}}) + (\boldsymbol{\beta} - \mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\mathbf{a}])^T(\boldsymbol{\beta} - \mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\mathbf{a}])$
- Bias of $\mathbf{a}$: $\mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\mathbf{a} - \boldsymbol{\beta}] = \mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\mathbf{a}] - \boldsymbol{\beta}$
- Covariance of $\mathbf{a}$: $\mathsf{Cov}_{\mathbf{Y}|\boldsymbol{\beta}}[\mathbf{a} - \mathsf{E}[\mathbf{a}]$
- Multivariate analog of MSE = Bias^2 + Variance in the univariate case

Mean Square Error of OLS Estimator

- MSE of OLS $\textsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$
- OLS is unbiased os mean of $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$ is $\mathbf{0}_p$
- covariance is $\textsf{Cov}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

$$\begin{aligned} \textsf{MSE}(\boldsymbol{\beta}) \equiv \textsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) & = \sigma^2 \textsf{tr}[(\mathbf{X}^T\mathbf{X})^{-1}] \\ & = \sigma^2 \textsf{tr} \mathbf{U} \boldsymbol{\Lambda}^{-1} \mathbf{U}^T \\ & = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} \end{aligned}$$

where λ_j are eigenvalues of $\mathbf{X}^T\mathbf{X}$.

- If smallest $\lambda_j \to 0$ then MSE $\to \infty$

Mean Square Error using the g-prior

- posterior mean is $\hat{\boldsymbol{\beta}}_g = \frac{g}{1+g}\hat{\boldsymbol{\beta}}$ (minimizes Bayes risk under squared error loss)
- bias of $\hat{\boldsymbol{\beta}}_g$:

$$\mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g] = \boldsymbol{\beta}\left(1 - \frac{g}{1+g}\right) = \frac{1}{1+g}\boldsymbol{\beta}$$

- covariance of $\hat{\boldsymbol{\beta}}_g$: $\mathsf{Cov}(\hat{\boldsymbol{\beta}}_g) = \frac{g^2}{(1+g)^2}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- MSE of $\hat{\boldsymbol{\beta}}_g$:

$$\begin{aligned}\mathsf{MSE}(\boldsymbol{\beta}) &= \frac{g^2}{(1+g)^2}\sigma^2 \mathsf{tr}(\mathbf{X}^T\mathbf{X})^{-1} + \frac{1}{(1+g)^2}\|\boldsymbol{\beta}\|^2 \\ &= \frac{1}{(1+g)^2}\left(g^2\sigma^2\sum_{j=1}^{p}\lambda_j^{-1} + \|\boldsymbol{\beta}\|^2\right)\end{aligned}$$

Can Bayes Estimators have smaller MSE than OLS?

Mean Square Error under Ridge Priors

Penalized Regression

Scaling and Centering

Note: usually use Ridge regression after centering and scaling the columns of $\mathbf{X}$ so that the penalty is the same for all variables. $\mathbf{Y}_c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and X_c the centered and standardized $\mathbf{X}$ matrix

- alternatively as a prior, we are assuming that the $\boldsymbol{\beta}_j$ are iid $\mathsf{N}(0, \kappa^*)$ so that the prior for $\boldsymbol{\beta}$ is $\mathsf{N}(\mathbf{0}_p, \kappa^* \mathbf{I}_p)$
- if the units/scales of the variables are different, then the variance or penality should be different for each variable.
- standardizing the $\mathbf{X}$ so that $\mathbf{X}_c^T \mathbf{X}_c$ is a constant times the correlation matrix of $\mathbf{X}$ ensures that all $\boldsymbol{\beta}$'s have the same scale
- centering the data forces the intercept to be 0 (so no shrinkage or penality)

Alternative Motivation

- If $\hat{\boldsymbol{\beta}}$ is unconstrained expect high variance with nearly singular $\mathbf{X}_c$
- Control how large coefficients may grow

$$\arg\min_{\boldsymbol{\beta}} (\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta})^T(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta})$$

subject to

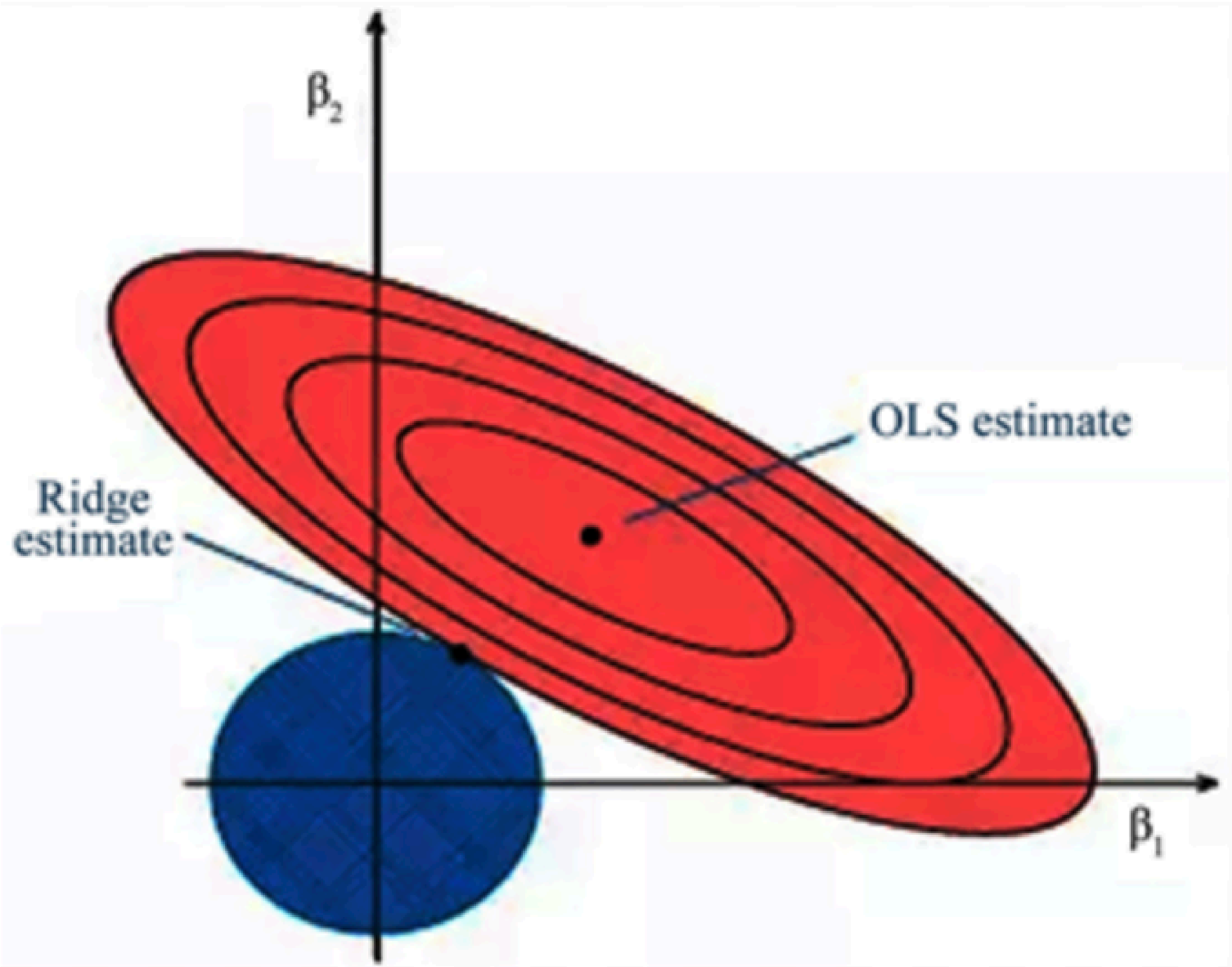
$$\sum \beta_j^2 \le t$$

- Equivalent Quadratic Programming Problem

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta}\|^2 + \kappa^*\|\boldsymbol{\beta}\|^2$$

- different approaches to selecting κ^* from frequentist ane Bayesian perspectives

Plot of Constrained Problem



Generalized Ridge Regression

- rather than a common penalty for all variables, consider a different penalty for each variable
- as a prior, we are assuming that the $\boldsymbol{\beta}_j$ are iid $\mathsf{N}(0, \frac{\kappa_j}{\phi})$ so that the prior for $\boldsymbol{\beta}$ is $\mathsf{N}(\mathbf{0}_p, \phi^{-1}\mathbf{K})$ where $\mathbf{K} = \textsf{diag}(\kappa_1, \ldots, \kappa_p)$
- hard enough to choose a single penalty, how to choose p penalties?
- place independent priors on each of the κ_j's
- a hierarchical Bayes model
- if we can integrate out the κ_j's we have a new prior for β_j
- this leads to a new penalty!
- examples include the Lasso (Double Exponential Prior) and Double Pareto Priors