MATH/STAT 355: Problem Set 6

Prof. Taylor Okonek

Due: May 2, 2025

Ridge Regression

Recall all the way back from PS2 that we derived the MLE for $\hat{\beta}$ in a linear regression framework using matrix notation as

$$\hat{\beta}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$$

where $\mathbf{Y}$ be a vector of dimension n, $\mathbf{X}$ be a matrix of dimension $n \times p$, and our regression coefficients $\boldsymbol{\beta}$ are a vector of unknown parameters of dimension p. We had the set-up that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with multivariate normally distributed errors, or equivalently,

$$\mathbf{Y} \mid \boldsymbol{\beta} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

where $\sigma^2 > 0$ is known, and the identity matrix $\mathbf{I}_n$ is of dimension $n \times n$.

Now, suppose that we approach regression from a Bayesian framework. In particular, suppose we put a Multivariate normal prior on $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} \sim MVN(0, \lambda\mathbf{I}_p)$. Show that the *posterior mean* for β is the Ridge regression estimator,

$$\hat{\beta}_R = \left(\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\lambda}\mathbf{I}\right)^{-1}\mathbf{X}^\top \mathbf{Y}$$

Hint: The matrix "version" of completing the square can be written as

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}^\top \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{X} + \mathbf{C} \\ &= (\mathbf{X} + \frac{1}{2}\mathbf{A}^{-1}\mathbf{B}^\top)^\top \mathbf{A}(\mathbf{X} + \frac{1}{2}\mathbf{A}^{-1}\mathbf{B}^\top) + \mathbf{C} - \frac{1}{4}\mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B}\end{aligned}$$

Fun Fact: If you instead set the prior on $\boldsymbol{\beta}$ to be such that $\boldsymbol{\beta} \sim Laplace(0, \lambda\mathbf{I})$, you end up with the Lasso regression estimator! Wow! Connections to Stat 253!

Jeffreys prior

Jeffreys prior is a less informative prior that has the form, for a single parameter θ,

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta)$ denotes the information matrix. One of the nice properties of Jeffreys prior is that the posteriors derived using it are *invariant* to reparameterizations.

Suppose we have $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$.

1. Derive Jeffreys prior $\pi(p)$ for this data. What common distribution does the prior follow?
2. Derive the posterior distribution for p under Jeffreys prior.
3. Use the posterior distribution you derived to construct an *exact*, 95% confidence interval for a binomial proportion p.*

*Note that this is sometimes called the *Jeffreys interval* for a binomial proportion, which has some nice properties in terms of **coverage**. We'll cover this in class when we do simulation studies!

Decision Theory

1. Show that the posterior mean is the decision rule that minimizes risk with respect to MSE loss, $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$.
2. Suppose we observe $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is known.
   (a) Find the posterior distribution of $\mu \mid X_1, \ldots, X_n$ where $\mu \sim N(\theta, \tau^2)$. (Note: You *cannot* cite Wikipedia for this question. I need to see actual math!)
   (b) Determine what value for τ would give you a posterior mean equal to $\overline{X}$, and use this to come up with a heuristic* argument that the sample mean is admissible for μ given a specific prior.

   *You can show that the sample mean is admissible without a heuristic argument, but it is much more complex mathematically. Additionally, note that we only have a *univariate* normal distribution here, not multivariate. As stated in the course notes, the sample mean is *not* admissible for the mean of a multivariate normal distribution with mean of dimension 3 or more!