STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 5 — September 15

Lecturer: Purnamrita Sarkar *Scribe: Ryan O'Donnell*

5.1 Quick Note on MLE Existence

Two examples were provided to highlight the possibility that an MLE may not exist. Example 1: Let $X \sim N(\mu, \sigma^2)$. Let $\theta =< \mu, \sigma >, theta\ \epsilon\ RxR^+$. The usual pdf for the normal is used here, which is equal to

$$f(x, \theta) = \frac{1}{\sqrt{2\pi} * \sigma} * exp(\frac{-(x-\mu)^2}{2\sigma^2})$$
$$log(f(x, \theta)) = c - \frac{(x-u)^2}{2\sigma^2} - log(\sigma)$$

So the MLE would be

$$0 = \frac{-(x-\mu)^2}{2\hat{\sigma}^3} - \frac{1}{\hat{\sigma}}$$
$$\hat{\sigma}^2 = (x-\mu)^2$$

But if $x = \mu$, then $\sigma = 0$, which is not allowed. So the MLE does not exist in this case.

Example 2: Let $X_i \sim uniform(0, \theta)$. If the interval included its boundary, then clearly the MLE would be $\theta = max[X_i]$. But since this interval does not include its boundary, the MLE cannot be the maximum, and therefore an MLE does not exist.

5.2 Expounding on the Admissibility of Shrinkage Estimators

As was previously mentioned, it is somewhat difficult to intuitively understand why these particular shrinkage estimators are admissible over the MLE. The below begins with the Bayesian approach to the problem. Beginning with

$$\boldsymbol{X}|\theta \sim N(\boldsymbol{\theta}, I), \theta \sim N(0, \tau^2 * I)$$

The posterior mean from empirical bayes is just

$$X * (1 - \frac{1}{1+\tau^2})$$

With that, we can aim to show that the MSE of this posterior estimator is preferable to the MSE of the MLE. For the moment, we will assume we knew Tau exactly. This allows for an easier proof of the MSE decreasing. In reality, we could perhaps approximate it from the data, though the classical Bayesian approach would not allow for this, as it violates the idea of a prior distribution.

$$E[(\boldsymbol{\theta}_{post} - \boldsymbol{\theta})^T(\boldsymbol{\theta}_{post} - \boldsymbol{\theta})] = E[(\frac{\tau^2}{1+\tau^2}X - \boldsymbol{\theta})^T * (\frac{\tau^2}{1+\tau^2}X - \boldsymbol{\theta})] \tag{5.1}$$

For ease of notation, let

$$c = \frac{1}{1+\tau^2} \tag{5.2}$$

Therefore, the above becomes

$$\begin{aligned}
&= E[(\boldsymbol{X} - \boldsymbol{\theta} - c\boldsymbol{X})^T(\boldsymbol{X} - \boldsymbol{\theta} - c\boldsymbol{X}) \\
&= E[(\boldsymbol{X} - \boldsymbol{\theta})^T(\boldsymbol{X} - \boldsymbol{\theta})] + c^2E[\boldsymbol{X}^T\boldsymbol{X}] - 2cE[\boldsymbol{X}^T(\boldsymbol{X} - \boldsymbol{\theta})] \\
&= MSE(\boldsymbol{X}) - (2c - c^2)E[\boldsymbol{X}^T\boldsymbol{X}] + 2cE[\boldsymbol{X}^T\boldsymbol{\theta}]
\end{aligned}$$

To show that the above does indeed equal something smaller than the MSE, it is easiest to break it up into pieces. First, recall the law of iterated expectations. Using this law,

$$\begin{aligned}
E[\boldsymbol{X}] &= E[E[\boldsymbol{X}|\boldsymbol{\theta}]] = E[\boldsymbol{\theta}] = 0 \\
\text{var}[\boldsymbol{X}] &= E[\text{var}(\boldsymbol{X}|\boldsymbol{\theta})] + \text{var}(E[\boldsymbol{X}|\boldsymbol{\theta}]) = E[I] + \text{var}(\boldsymbol{\theta}) = I(1+\tau^2)
\end{aligned}$$

As it turns out, $\boldsymbol{X}$ also has a normal distribution whose parameters using the above derivation is:

$$\boldsymbol{X} \sim N(0, (1+\tau^2)I)$$

This is useful because it implies that

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{1+\tau^2}$$

is a chi-squared distribution with degrees of freedom p. So, by properties of the chi squared distribution,

$$E[\boldsymbol{X}^T\boldsymbol{X}] = (1+\tau^2)p \tag{5.3}$$

Combining this with the original definition for c shows that:

$$\begin{aligned}
&= (2c - c^2)E[\boldsymbol{X}^T\boldsymbol{X}] \\
&= c(2-c) * E[\boldsymbol{X}^T\boldsymbol{X}] \\
&= \frac{1+2\tau^2}{(1+\tau^2)^2} * p
\end{aligned}$$

For the next part, the law of iterated expectation and the chi squared distributions are again very useful. The bulk of the work comes from simply implementing the law.

$$E[\boldsymbol{\theta}^T\boldsymbol{X}] = E[E[\boldsymbol{\theta}^T\underline{\boldsymbol{X}}|\theta]] = E[\theta^T E[X|\theta]] = E[\theta^T\theta] = \sum_i E[\theta_i^2] = \tau^2 p$$

Combining this with the original definition for c shows that:

$$\begin{aligned} &= 2c * E[\boldsymbol{\theta}^T\boldsymbol{X}] \\ &= \frac{2\tau^2}{1+\tau^2} * p \end{aligned}$$

Now, if we combine these two facts with the original definition of c, we can simplify our original expression for the MSE.

$$\begin{aligned} MSE(\boldsymbol{\theta}_{post}) &= MSE(\boldsymbol{X}) - (2c - c^2)E[\boldsymbol{X}^T\boldsymbol{X}] + 2cE[\boldsymbol{X}^T\boldsymbol{\theta}] \\ &= MSE(\boldsymbol{X}) - \frac{1+2\tau^2}{1+\tau^2} * p + \frac{2\tau^2}{1+\tau^2} * p \\ &= MSE(\boldsymbol{X}) - \frac{1}{1+\tau^2} * p \end{aligned}$$

So, as long as we know τ, we have found a way to create a shrinkage estimator that is uniformaly better than MLE in terms of its MSE. Also, this posterior mean approach creates something that is similar to the James-Stein Estimator. However, this example was not entirely realistic. What if we did not know Tau? Would we still do better than the MLE? It turns out that if we use some y to estimate tau, we arrive at the James Stein Estimator.

Recall the following: $\hat{\theta}_{post} = (1 - \frac{1}{1+\tau^2}) * \boldsymbol{X}$.
If we don't know τ, we must estimate it. Consider a random variable Y s.t.

$$E[Y] = \frac{1}{1+\tau^2}$$

Now, let $V = \frac{\boldsymbol{X}^T\boldsymbol{X}}{1+\tau^2}$. By definition, V is a chi-squared distribution, as it is equal to $\Sigma(\frac{X}{\sqrt{1+\tau^2}})^2$. Now, take$\frac{1}{V}$. This has the inverse chi squared distribution. By properties of the inverse chi squared, $E[\frac{1}{V}] = \frac{1}{p-2} = E[\frac{1+\tau^2}{\boldsymbol{X}^T\boldsymbol{X}}]$. Now, notice the following:

$$E[\frac{p-2}{\boldsymbol{X}^T\boldsymbol{X}}] = (p-2)E[\frac{1}{\boldsymbol{X}^T\boldsymbol{X}}] = (p-2)(\frac{1}{(p-2)(1+\tau^2)}) = \frac{1}{1+\tau^2}$$

Therefore, since this yields the desired expectation, $Y = \frac{p-2}{\boldsymbol{X}^T\boldsymbol{X}}$. Now, using this value of y as an estimator for $1 - \frac{1}{1+\tau^2}$ yields the following, which is equivalent to the James Stein Estimator:

$$\hat{\theta}_{\text{empirical bayes}} = (1 - \frac{p-2}{\boldsymbol{X}^T\boldsymbol{X}}) * \boldsymbol{X}$$

We call this "empirical bayes" since here we used a Bayesian model and then played frequentist by estimating the hyperparameter using the data.

5.3 Linear Regression

5.3.1 Model and MLE

Here is a linear model for linear regression. Lets first do it for one pair of data points (x, y).

$$y = \beta_0 + \beta_1 \cdot x_1 + ... + \beta_p \cdot x_p + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Now, for n data-points (x_i, y_i), where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ we can write it in matrix notation as follows:

We can write this in matrix form by stacking the datapoints as the rows of a matrix $\boldsymbol{X}$ so that x_{ij} is the j-th feature of the i-th datapoint. Then writing Y, β and ϵ as column vectors, we can write the matrix form of the linear regression model as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \text{and } \boldsymbol{X} = \begin{bmatrix} 1 & x_{12} & \ldots & x_{1p} \\ 1 & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

Assume that ϵ_i is normally distributed with variance σ^2. And so ϵ. We will now calculate the MLE $\hat{\beta}$ of β.

We are using the notation where smaller case bold letters denote vectors, capital bold dentotes matrices.

$$f(\boldsymbol{y}, \boldsymbol{\beta}) \propto \exp(\frac{-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2})$$

Take Log, we can get:

$$\frac{-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} \tag{5.4}$$

Same drill– differentiate and set it to zero.

$$-\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0 \rightarrow \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \rightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

5.3.2 Relation to least squares

Lets say I wanted to calculate an estimate that minimized the residual sum of squares (RSS).

$$\boldsymbol{\beta}_{LS} = \min_{\boldsymbol{\beta'}} RSS(\boldsymbol{\beta'}) := \min_{\beta'} \sum_i (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta'})^2$$

As it turns out, $RSS(\boldsymbol{\beta'})$ is none other than $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta'})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta'})$. But remember, because the noises are all independently drawn from the same mean zero normal distribution, maximizing log likelihood boils down to minimizing the RSS. And in this special case, the least squares estimate is identical to the MLE.

5.3.3 Expectation and Variance of $\hat{\boldsymbol{\beta}}$

Now, we want to find the $E[\hat{\beta}], Var[\hat{\beta}]$. Lets put down some ground rules for taking expectations of vector valued random variables. Say $\boldsymbol{z} = A\boldsymbol{y}$ where A is a fixed matrix. $E[\boldsymbol{z}] = AE[\boldsymbol{y}]$ and $\text{var}(\boldsymbol{z}) = A\text{var}(\boldsymbol{z})A^T$. Recall that $E[\boldsymbol{y}] = \boldsymbol{X\beta}$ and $\text{var}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$

$$E[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$
$$\text{var}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{var}[\boldsymbol{y}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Conclusion: $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. Note: this is not approximate, but exact!