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

Lecture 4 — September 13

Lecturer: Purnamrita Sarkar *Scribe: Elisa Ferracane, Carlos Zanini*

Note: These scribe notes have been slightly proofread and may have typos etc.

4.1 Review of last lecture

In the previous class, we started talking about where the MLE doesn't work. Here we review the Neyman-Scott example, where the estimator does not converge to the true parameter, because the number of parameters grows with the number of data points.

4.1.1 Neyman-Scott Example

Let's consider n groups of gaussian samples that differ only in their population mean μ:

$$X_{11}, X_{12}, \ldots, X_{1k} \sim \mathcal{N}(\mu_1, \sigma^2)$$
$$\ldots$$
$$X_{n1}, X_{n2}, \ldots, X_{nk} \sim \mathcal{N}(\mu_n, \sigma^2).$$

If our goal is to estimate the variance of the whole population σ^2, it seems obvious to consider as an estimator the mean of the sample variances of the single populations. In fact, this is the MLE for σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n S_i^2}{n}, \text{ where } S_i^2 = \text{ sample variance } = \frac{1}{k}\sum_{j=1}^k (X_{ij} - \bar{X}_i)^2.$$

It is a well known fact that

$$\frac{(k-1)S_i^2}{\sigma^2} \sim \chi^2_{k-1}, \text{ for } S_i^2 = \sum_{j=1}^k \frac{(X_{ij} - \hat{X}_i)^2}{k-1}$$

and so its expected value is $(k-1)$. Using this, as well as the law of large numbers, we get to

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \frac{(k-1)S_i^2}{k} = \frac{\sigma^2}{kn}\sum_{i=1}^n \frac{(k-1)S_i^2}{\sigma^2} \to \frac{\sigma^2}{k} E\left(\frac{(k-1)S_i^2}{\sigma^2}\right) = \frac{\sigma^2}{k} \times (k-1) \neq \sigma^2.$$

Since $\hat{\sigma}^2$ is the MLE, this result may seems to contradict the fact that the MLE converges to the real value of the parameter. It happens in this case as a consequence of a high dimensionality problem: the number of parameters to be estimated in this problem grows linearly with n.

4.2 Shrinkage estimators

4.2.1 Admissibility

Admissibility tells us when one estimator is better than another.

Definition 4.1. ***(Dominance in Mean Squared Error).*** *Consider the estimation of a parameter $\theta \in \Theta$. An estimator $\hat{\theta}$ is said to be dominated in mean squared error by $\tilde{\theta}$, if we have*

$$MSE(\tilde{\theta}, \theta) \leq MSE(\hat{\theta}, \theta), \forall \theta \in \Theta$$

and if there exists at least one $\theta_0 \in \Theta$ for which the inequality above is strict:

$$MSE(\tilde{\theta}, \theta_0) < MSE(\hat{\theta}, \theta_0).$$

Under the same conditions, we could also say that $\tilde{\theta}$ dominates $\hat{\theta}$ in mean squared error.

In other words, if we have an estimator $\tilde{\theta}$ that dominates $\hat{\theta}$ in mean squared error, then it would not be reasonable (according to this criteria) to use $\hat{\theta}$ as an estimator of θ, since for any possible value of the parameter $\theta \in \Theta$, the second one will always have a greater mean squared error. This concept is formalized in the following definition:

Definition 4.2. ***(Admissibility).*** *An estimator $\hat{\theta}$ is said to be admissible if no other estimator $\tilde{\theta}$ dominates it in Mean Square Error.*

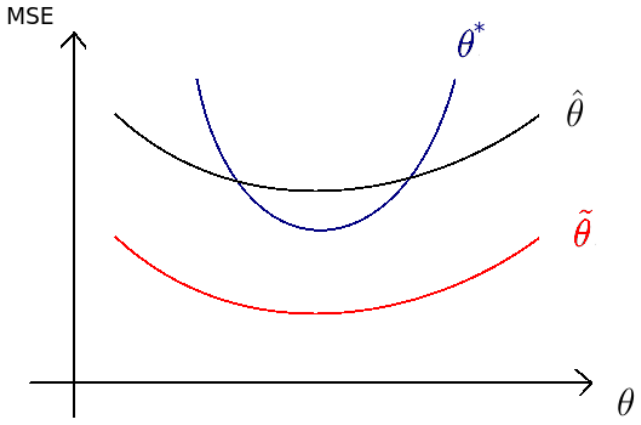

Figure 4.1. Mean Squared Error of three different estimators as a function of the real value of the parameter θ.

Figure **??** illustrates the concept. For the estimators $\hat{\theta}$ and θ^*, neither one dominates the other through the entire parameter space Θ. However, both of them are dominated in mean squared error by $\tilde{\theta}$, so they are not admissible.

Note this implies you know the true parameter (or its distribution). We use this approach not for learning θ but for evaluating different estimators. Now a concrete case is presented to illustrate the concept of admissibility.

4.2.2 The James-Stein estimator

Consider an i.i.d sample $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ from $N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ with dimension p, with known σ. If $n = 1$, the natural estimator for μ would be simply that single data point, X. The James-Stein estimator for μ is defined as

$$\hat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{(p-2)}{||\boldsymbol{X}_1||^2}\right) \boldsymbol{X}_1$$

Note how the term on the right shrinks the value of the natural estimator towards 0 (assuming $(p-2) < ||X||$). Thus, the James Stein estimator can be seen as a shrinkage estimator. This introduces some bias, but leads to lower overall variance. In fact, the $\hat{\boldsymbol{\mu}}_{JS}$ dominates $\hat{\boldsymbol{\mu}}_{MLE}$. There is another variant of the James Stein estimator that zeros out its negative entries, which in turn dominates the above James Stein estimator:

$$\hat{\theta}_+^{JS} > \hat{\theta}_{JS} > \hat{\theta}_{MLE}$$

So why hasn't the James Stein estimator caught on? Paraphrasing from the paper by Efron and Morris [1], here are a few reasons:

1. "Mistrust of the statistical interpretation of the mathematical formulation leading to Stein's result, in particular the sum of squared errors loss function.

2. Difficulties in adapting the James-Stein estimator to the many special cases that invariably arise in practice;

3. Long familiarity with the general good performance of the MLE in applied problems;

4. A feeling that any gains possible from a *complicated* procedure like Stein's could not be worth the extra trouble."

While the James-Stein estimator is not intuitive and at a first peek, looks rather suspicious, we will now show that this sort of shrinkage effect can be seen in many other naturally arising Bayesian estimators, as the next example will show.

4.2.3 Connection to Empirical Bayes:

Consider $X_j = \theta_j + \epsilon_j$, where:

$$\epsilon_1 \dots \epsilon_p \sim \mathcal{N}(0,1)$$
$$X_j|\theta_j \sim \mathcal{N}(\theta_j, 1)$$
$$\theta_1 \dots \theta_p \sim \mathcal{N}(0, \tau^2)$$

and $\boldsymbol{X}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{I})$ with a prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, \tau^2\boldsymbol{I})$. Let us calculate the posterior mean of the gaussian.

Posterior mean of Gaussian

First note the conjugate of the normal is the normal and recall the exponential in the gaussian density function:

$$exp(-\frac{(\theta-\mu)^2}{2\sigma^2})$$

We use Bayes' Rule to get the posterior distribution:

$$\begin{aligned}
f(\theta|X) &\propto f(X|\theta)f(\theta) \\
&= exp(-\frac{(X-\theta)^2)}{2})exp(\frac{-\theta^2}{2\tau^2}) \\
&= exp(-\frac{X^2}{2} - \frac{\theta^2}{2} + \theta X - \frac{\theta^2}{2\tau^2}) \\
&= exp(-\frac{\theta^2}{2}(1 + \frac{1}{\tau^2} + \theta X)) \\
&= exp(-\frac{1}{2}(1+\frac{1}{\tau^2})(\theta^2 - \frac{2\theta X}{1+\frac{1}{\tau^2}})) \\
&= exp(-\frac{(\theta - \frac{X}{1+\frac{1}{\tau^2}})^2}{2\frac{1}{1+\frac{1}{\tau^2}}})
\end{aligned}$$

Note this is the exact same format as the normal, so that we can easily derive the mean and variance of the posterior distribution:

$$\mu_{post} = X * \frac{\tau^2}{1+\tau^2}$$

$$\sigma_{post} = \frac{1}{1+\frac{1}{\tau^2}}$$

and rearranging terms, we get:

$$\mu_{post} = X(1 - \frac{1}{1+\tau^2})$$

We can see that it is a shrinkage estimator, since it multiplies the MLE by a quantity between 0 and 1. Also note that you actually don't know τ, so you estimate it from the data– hence the name *empirical*. Bayesians may argue that such a procedure violates the Bayesian principle of not using the data to express subjective prior information about parameters of the model.

As it turns out, plugging in the frequentist estimate of τ^2 gives us the James-Stein estimator. How is out of the scope of this class, but if you are interested read [1].

References

[1] Efron, Brad and Morris, Carl. Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach. *Journal of the American Statistical Association* , 1973.