LECTURE NOTES 11

1 Comparing estimators

In the last lecture we saw that most often there was no uniformly dominant estimator, i.e. most often there was not an estimator that had smaller risk than every other estimator, everywhere.

Example: Recall the Bernoulli estimation problem: two natural estimators are the MLE:

$$\widehat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and the Bayes estimator we defined previously:

$$\widehat{p}_2 = \frac{\sum_{i=1}^{n} X_i + \alpha}{n+\alpha+\beta},$$

for some values α and β that we will specify soon. Again, suppose we consider the squared loss:

$$R(p,\widehat{p}_1) = \frac{p(1-p)}{n}.$$

$$R(p,\widehat{p}_2) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i + \alpha}{n+\alpha+\beta}\right) + \left(\mathbb{E}\frac{\sum_{i=1}^{n} X_i + \alpha}{n+\alpha+\beta} - p\right)^2.$$

In the second estimator if we choose $\alpha = \beta = \sqrt{n/4}$ we obtain that the risk is constant as a function of p, i.e.

$$R(p,\widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

We can compare these two estimators' risk functions but once again we see that neither estimator dominates the other. In such cases, we need other ways to compare estimators and to find "best" estimators.

There are two main ways to choose an estimator:

Minimax estimators: The minimax estimator $\widehat{\theta}$ is one that minimizes the worst-case risk,i.e., it is one that satisfies:

$$\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}) = \inf_{\theta'}\sup_{\theta\in\Theta} \mathbb{E}_\theta L(\theta,\theta').$$

Bayes Estimator: Recall that the Bayes rsik with respect to a prior $\pi$is for some prior $\pi(\theta)$:

$$R_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta.$$

The Bayes estimator minimizes the Bayes risk. This requires choosing a prior π. Recall that for squared error loss, the Bayes estimator is the posterior mean.

Example: Let us revisit the two Bernoulli estimators from the standpoint of maximum risk and Bayes risk. Suppose we take the uniform prior, then:

$$R_\pi(\widehat{p}_1) = \int_p \frac{p(1-p)}{n} dp = \frac{1}{6n},$$

$$R_\pi(\widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2},$$

so for large n the MLE has smaller Bayes risk. On the other hand the estimator $\widehat{p}_2$ always has lower maximum risk.

2 Minimax Estimators

In general, finding exactly minimax estimators is not easy but often we can find approximately minimax (or rate minimax) estimators. There are however, some important connections between Bayes rules, minimax estimators and admissible estimators that are worth knowing.

1. For any estimator the Bayes risk with respect to any prior π lower bounds its maximum risk.

2. Suppose that for some prior π, we have that the corresponding Bayes rule $\widehat{\theta}_\pi$ has the property that:

   $$R(\theta, \widehat{\theta}_\pi) \leq R_\pi(\widehat{\theta}_\pi),$$

   for every θ. Then π is called a **least favorable prior** and $\widehat{\theta}_\pi$ is minimax.

3. A simple consequence of the above is that if for some π we have that $R(\theta, \widehat{\theta}_\pi)$ is a constant (as a function of θ) then $\widehat{\theta}_\pi$ is a minimax estimator. In words, a Bayes rule with constant risk is minimax.

   Example: Recall the Binomial Bayes estimator with $\alpha = \beta = \sqrt{n}/2$. We saw that its MSE is:

   $$R(\widehat{p}, p) = \frac{n}{4(n+\sqrt{n})^2}.$$

Since the Bayes estimator has constant risk it is minimax.

Example: Suppose we consider Bernoulli estimation with the loss function:

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

This loss function is sometimes called the Fisher loss. We will see why in a future lecture.

4. If $X_1, \ldots, X_n \sim N(\theta, 1)$ then $\widehat{\theta} = \overline{X}$ is minimax for a large variety of loss functions.

5. Under some conditions, it can be shown the MLE is approximately minimax.

3 Admissibility

The concept of **admissibility** helps to weed out bad estimators. We say that an estimator $\widehat{\theta}_1$ is inadmissible if there is another estimator $\widehat{\theta}_2$ such that,

$$R(\theta, \widehat{\theta}_2) \leq R(\theta, \widehat{\theta}_1),$$

for every $\theta \in \Theta$, and

$$R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1),$$

for some $\theta \in \Theta$. Otherwise, we say that $\widehat{\theta}_1$ is admissible. We say that an estimator $\widehat{\theta}_1$ is strongly inadmissible if there exists there is another estimator $\widehat{\theta}_2$ and an $\epsilon > 0$ such that

$$R(\theta, \widehat{\theta}_2) \leq R(\theta, \widehat{\theta}_1) - \epsilon$$

for all θ. Being admissible is a weak property. It does not guarantee that the estimator is good.

1. Bayes estimators with respect to certain "nice" priors are admissible. Formally, if the prior π is non-zero everywhere over the parameter space and if the risk of the Bayes rule is finite then the Bayes rule is admissible.
2. A minimax estimator cannot be strongly inadmissible.
3. Perhaps the most stunning example of this is due to Charles Stein. This is sometimes called Stein's phenomenon or Stein's paradox. We will not go over this in detail but its worth knowing the paradox. Let

$$Y_i \sim N(\theta_i, 1),$$

for $i \in \{1, \ldots, d\}$. We would like to estimate the vector θ. The loss function is squared error. The natural estimator of course is just the observed data, i.e.:

$$\widehat{\theta} = Y.$$

It turns out that this estimator is minimax optimal. Stein showed that this estimator is inadmissible if $d \geq 3$. An estimator that beats Y is a "shrinkage estimator" that pulls Y towards the origin.