36-705: Intermediate Statistics **Fall 2019**

Lecture 16: October 4

Lecturer: Siva Balakrishnan

In the last lecture we discussed the MSE and the bias-variance decomposition. We discussed briefly that finding uniformly optimal estimators (i.e. estimators with lowest possible MSE for every value of the unknown parameter) is hopeless, and then briefly discussed finding optimal unbiased estimators.

We then introduced some quantities: the score and the Fisher information, and used these notions to discuss the Cramér-Rao bound, which provides a lower bound on the variance of any unbiased estimator.

You might come across this terminology in your research: estimators that achieve the Cramér-Rao bound, i.e. are unbiased and achieve the Cramér-Rao lower bound on the variance, are often called *efficient* estimators. In many problems efficient estimators do not exist, and we often settle for a weaker notion of *asymptotic efficiency*, i.e. efficient but only as $n \to \infty$. Next week, we will show that in many cases the MLE is asymptotically efficient.

The Cramér-Rao bound suggests that the MSE in a parametric model typically scales as $1/(nI_1(\theta))$. In essence $1/I_1(\theta)$ behaves like the variance (we will see this more clearly in the future), so that the MSE behaves as σ^2/n, which is exactly analogous to what we have seen before with averages (think back to Chebyshev/sub-Gaussian tail bounds).

Supposing that the Fisher information is non-degenerate, i.e. that $I_1(\theta) > 0$, and treating the model and θ as fixed: $I_1(\theta)$ is a constant as $n \to \infty$. In such "nice" cases, the MSE converges to 0 at the rate of $1/n$. This is often referred to as the *parametric rate*. We will in later lectures discuss non-parametric models where the typical rate of convergence is much slower (and depends more drastically on the dimension of the model – this is called the *curse of dimensionality*).

16.1 Decision Theory

Suppose we want to estimate a parameter θ using data $X^n = (X_1, \ldots, X_n)$. What is the best possible estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of θ? Decision theory provides a framework for answering this question.

16.1.1 The Risk Function

Let $\widehat{\theta} = \widehat{\theta}(X^n)$ be an estimator for the parameter $\theta \in \Theta$. We start with a **loss function** $L(\theta, \widehat{\theta})$ that measures how good the estimator is. For example:

$$\begin{array}{ll}
L(\theta,\widehat{\theta}) = (\theta - \widehat{\theta})^2 & \text{squared error loss,} \\
L(\theta,\widehat{\theta}) = |\theta - \widehat{\theta}| & \text{absolute error loss,} \\
L(\theta,\widehat{\theta}) = |\theta - \widehat{\theta}|^p & L_p \text{ loss,} \\
L(\theta,\widehat{\theta}) = 0 \text{ if } \theta = \widehat{\theta} \text{ or } 1 \text{ if } \theta \neq \widehat{\theta} & \text{zero–one loss,} \\
L(\theta,\widehat{\theta}) = I(|\widehat{\theta} - \theta| > c) & \text{large deviation loss,} \\
L(\theta,\widehat{\theta}) = \int \log\left(\frac{p(x;\,\theta)}{p(x;\,\widehat{\theta})}\right) p(x;\,\theta)dx & \text{Kullback–Leibler loss.}
\end{array}$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector then some common loss functions are

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|^2 = \sum_{j=1}^{k} (\widehat{\theta}_j - \theta_j)^2,$$

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_p = \left(\sum_{j=1}^{k} |\widehat{\theta}_j - \theta_j|^p \right)^{1/p}.$$

When the problem is to predict a $Y \in \{0, 1\}$ based on some classifier $h(x)$ a commonly used loss is

$$L(Y, h(X)) = I(Y \neq h(X)).$$

For real valued prediction a common loss function is

$$L(Y, \widehat{Y}) = (Y - \widehat{Y})^2.$$

The **risk** of an estimator $\widehat{\theta}$ is

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\Big(L(\theta, \widehat{\theta}) \Big) = \int L(\theta, \widehat{\theta}(x_1, \ldots, x_n)) p(x_1, \ldots, x_n; \theta) dx. \qquad (16.1)$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2. \qquad (16.2)$$

If we do not state what loss function we are using, assume the loss function is squared error.

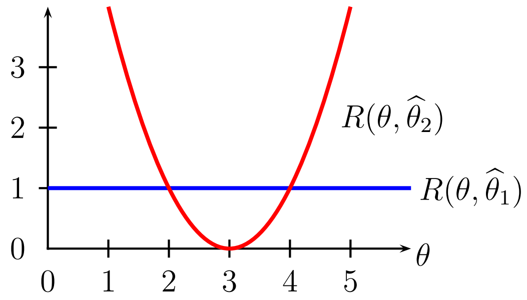

Figure 16.1: Comparing two risk functions. Neither risk function dominates the other at all values of θ.

16.1.2 Comparing Risk Functions

To compare two estimators, we compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

Example 16.1 *Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta,\widehat{\theta}_1) = \mathbb{E}_\theta(X-\theta)^2 = 1$ and $R(\theta,\widehat{\theta}_2) = \mathbb{E}_\theta(3-\theta)^2 = (3-\theta)^2$. If $2 < \theta < 4$ then $R(\theta,\widehat{\theta}_2) < R(\theta,\widehat{\theta}_1)$, otherwise, $R(\theta,\widehat{\theta}_1) < R(\theta,\widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 16.1.*

Example 16.2 *Let $X_1,\ldots,X_n \sim \text{Bernoulli}(p)$. Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has zero bias, we have that*

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

Another estimator is

$$\widehat{p}_2 = \frac{Y+\alpha}{\alpha+\beta+n}$$

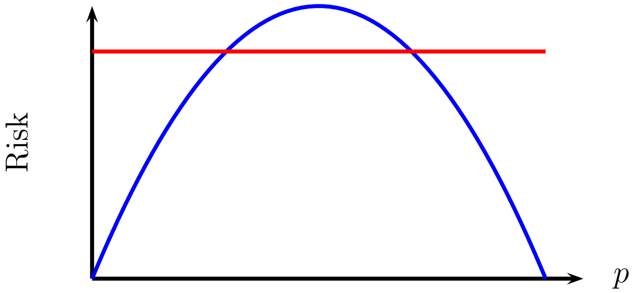

Figure 16.2: Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 16.2. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

where $Y = \sum_{i=1}^n X_i$ and α and β are positive constants.[1] *Now,*

$$\begin{aligned} R(p, \widehat{p}_2) &= \mathsf{Var}_p(\widehat{p}_2) + (\mathrm{bias}_p(\widehat{p}_2))^2 \\ &= \mathsf{Var}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2. \end{aligned}$$

Let $\alpha = \beta = \sqrt{n/4}$. The resulting estimator is

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

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

The risk functions are plotted in Figure 16.2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

[1] This is the posterior mean using a Beta (α, β) prior.

The **maximum risk** is

$$\overline{R}(\widehat{\theta}) = \sup_{\theta\in\Theta} R(\theta, \widehat{\theta}) \tag{16.3}$$

and the **Bayes risk** under prior π is

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta. \tag{16.4}$$

Example 16.3 *Consider again the two estimators in Example 16.2. We have*

$$\overline{R}(\widehat{p}_1) = \max_{0\leq p\leq 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

and

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

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when n is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like the maximum risk are imperfect.

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

An estimator $\widehat{\theta}$ that minimizes the Bayes risk is called a **Bayes estimator**. That is,

$$B_\pi(\widehat{\theta}) = \inf_{\tilde{\theta}} B_\pi(\tilde{\theta}) \tag{16.5}$$

where the infimum is over all estimators $\tilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax estimator**. That is,

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta}) \tag{16.6}$$

where the infimum is over all estimators $\tilde{\theta}$. We call the right hand side of (16.6), namely,

$$R_n \equiv R_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R(\theta, \widehat{\theta}), \tag{16.7}$$

the **minimax risk**. Statistical decision theory has two goals: determine the minimax risk R_n and find an estimator that achieves this risk.

Once we have found the minimax risk R_n we want to find the minimax estimator that achieves this risk:

$$\sup_{\theta\in\Theta} R(\theta, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R(\theta, \widehat{\theta}). \tag{16.8}$$

16.1.3 Bayes Estimators

Let π be a prior distribution. After observing $X^n = (X_1, \dots, X_n)$, the posterior distribution is, according to Bayes' theorem,

$$\mathbb{P}(\theta \in A|X^n) = \frac{\int_A p(X_1, \dots, X_n|\theta)\pi(\theta)d\theta}{\int_\Theta p(X_1, \dots, X_n|\theta)\pi(\theta)d\theta} = \frac{\int_A \mathcal{L}(\theta)\pi(\theta)d\theta}{\int_\Theta \mathcal{L}(\theta)\pi(\theta)d\theta} \tag{16.9}$$

where $\mathcal{L}(\theta) = p(x^n; \theta)$ is the likelihood function. The posterior has density

$$\pi(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{m(x^n)} \tag{16.10}$$

where $m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta$ is the **marginal distribution** of X^n. Define the **posterior risk** of an estimator $\widehat{\theta}(x^n)$ by

$$r(\widehat{\theta}|x^n) = \int L(\theta, \widehat{\theta}(x^n))\pi(\theta|x^n)d\theta. \tag{16.11}$$

Theorem 16.4 *The Bayes risk $B_\pi(\widehat{\theta})$ satisfies*

$$B_\pi(\widehat{\theta}) = \int r(\widehat{\theta}|x^n)m(x^n)\, dx^n. \tag{16.12}$$

Let $\widehat{\theta}(x^n)$ be the value of θ that minimizes $r(\widehat{\theta}|x^n)$. Then $\widehat{\theta}$ is the Bayes estimator.

Proof:

Let $p(x, \theta) = p(x|\theta)\pi(\theta)$ denote the joint density of X and θ. We can rewrite the Bayes risk as follows:

$$\begin{aligned}
B_\pi(\widehat{\theta}) &= \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = \int \left(\int L(\theta, \widehat{\theta}(x^n))p(x|\theta)dx^n \right) \pi(\theta)d\theta \\
&= \int \int L(\theta, \widehat{\theta}(x^n))p(x, \theta)dx^n d\theta = \int \int L(\theta, \widehat{\theta}(x^n))\pi(\theta|x^n)m(x^n)dx^n d\theta \\
&= \int \left(\int L(\theta, \widehat{\theta}(x^n))\pi(\theta|x^n)d\theta \right) m(x^n)\, dx^n = \int r(\widehat{\theta}|x^n)m(x^n)\, dx^n.
\end{aligned}$$

If we choose $\widehat{\theta}(x^n)$ to be the value of θ that minimizes $r(\widehat{\theta}|x^n)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x^n)m(x^n)dx^n$.

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

Theorem 16.5 *If $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x^n) = \int \theta \pi(\theta | x^n) d\theta = \mathbb{E}(\theta | X = x^n). \tag{16.13}$$

If $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$ then the Bayes estimator is the median of the posterior $\pi(\theta|x^n)$. If $L(\theta, \widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $\pi(\theta|x^n)$.

Proof:

We will prove the theorem for squared error loss. The Bayes estimator $\widehat{\theta}(x^n)$ minimizes $r(\widehat{\theta}|x^n) = \int (\theta - \widehat{\theta}(x^n))^2 \pi(\theta|x^n) d\theta$. Taking the derivative of $r(\widehat{\theta}|x^n)$ with respect to $\widehat{\theta}(x^n)$ and setting it equal to zero yields the equation $2 \int (\theta - \widehat{\theta}(x^n)) \pi(\theta|x^n) d\theta = 0$. Solving for $\widehat{\theta}(x^n)$ we get 16.13.

Example 16.6 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known. Suppose we use a $N(a, b^2)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is*

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{b^2}{b^2 + \frac{\sigma^2}{n}} \overline{X} + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}} a. \tag{16.14}$$