36-705: Intermediate Statistics **Fall 2017**

Lecture 15: October 4

Lecturer: Siva Balakrishnan

We continue our discussion of point estimation. The section on MSE is from Larry's notes. However, after discussing the MSE we will briefly depart from Larry's version to talk about unbiased estimators. We will subsequently return to Larry's notes on decision theory.

15.1 MSE

Now we discuss the evaluation of estimators. The mean squared error (MSE) is

$$\mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \int \cdots \int (\widehat{\theta}(x_1, \ldots, x_n) - \theta)^2 p(x_1; \theta) \cdots p(x_n; \theta) dx_1 \ldots dx_n.$$

The bias is

$$B = \mathbb{E}_\theta(\widehat{\theta}) - \theta$$

and the variance is

$$V = \mathrm{Var}_\theta(\widehat{\theta}).$$

Theorem 15.1 *We have*

$$MSE = B^2 + V.$$

Proof: Let $m = \mathbb{E}_\theta(\widehat{\theta})$. Then

$$\begin{aligned} MSE &= \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathbb{E}_\theta(\widehat{\theta} - m + m - \theta)^2 \\ &= \mathbb{E}_\theta(\widehat{\theta} - m)^2 + (m - \theta)^2 + 2\mathbb{E}_\theta(\widehat{\theta} - m)(m - \theta) \\ &= \mathbb{E}_\theta(\widehat{\theta} - m)^2 + (m - \theta)^2 = V + B^2. \end{aligned}$$

■

An estimator is *unbiased* if the bias is 0. In that case, the MSE = Variance. There is often a tradeoff between bias and variance. So low bias can imply high variance and vice versa.

Example 15.2 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then*

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathbb{E}(S^2) = \sigma^2.$$

The MSE's are

$$\mathbb{E}(\overline{X} - \mu)^2 = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1}.$$

It is worth thinking about how one defines the MSE when θ is multivariate (as in the example above), and what the analogous bias-variance decomposition is.

We would like to choose an estimator with small MSE. However, the MSE is a function of θ. Later, we shall discuss minimax estimators, that use the maximum of the MSE over θ as a way to compare estimators.

15.2 Unbiased estimators - Fisher Information, Cramér-Rao

Classically, an initial focus in the search for optimal estimators focused on unbiased estimators. Roughly, from a technical standpoint finding estimators which have lower possible MSE is quite difficult so one way to narrow the search space is to restrict our attention to unbiased estimators. In this case, the goal is to find minimum variance unbiased estimators and classical textbooks discuss the question of existence and finding minimum variance unbiased estimators in much detail.

More modern treatments do not often emphasize this point of view for two reasons: (1) there are many known examples where a small amount of bias can result in large reductions in variance so in general restricting attention to unbiased estimators can be bad. (2) for most problems finding minimum variance unbiased estimators is challenging or impossible (there are lots of interesting statistical quantities for which there are no unbiased estimators).

With that said there are still pieces of this classical theory that I personally find very useful. One of the important pieces is the Cramér-Rao bound which provides a lower bound on the variance of an unbiased estimator. In many problems, this bound will provide some at least heuristic guidelines into the difficulty of an estimation problem. Later on in the course we will talk about other ways of proving lower bounds that do not restrict attention to unbiased estimators (i.e. we will discuss what are called minimax lower bounds).

15.2.1 Fisher Information

We are in the setting where we observe $X_1, \ldots, X_n \sim p(X; \theta)$. We will generally suppose that $\theta \in \mathbb{R}^d$. We can compute the log-likelihood function:

$$\mathcal{LL}(\theta) = \sum_{i=1}^{n} \log p(X_i; \theta).$$

We can also define the gradient of this function, which is called the score function:

$$s(\theta) = \nabla_\theta \mathcal{LL}(\theta) = \sum_{i=1}^{n} \nabla_\theta \log p(X_i; \theta).$$

This gradient is a d-dimensional vector. The Fisher Information matrix is the expected outer product of the score, i.e.:

$$I(\theta) = \mathbb{E}[s(\theta)s(\theta)^T].$$

The Fisher information matrix is a $d \times d$ matrix. Lets take a quick look at a couple of examples:

Example 1: Suppose that $X \sim \text{Ber}(p)$, then the log-likelihood is given by:

$$\mathcal{LL}(p) = X \log(p) + (1 - X) \log(1 - p),$$

and accordingly the score is:

$$s(p) = \frac{X}{p} - \frac{1 - X}{1 - p} = \frac{X - p}{p(1 - p)}.$$

We can then compute the Fisher information:

$$I(p) = \frac{1}{p^2(1 - p)^2} \mathbb{E}\left[(X - p)^2\right] = \frac{1}{p(1 - p)}.$$

Example 2: Suppose that $X \sim N(\mu, \sigma^2)$ where σ is known, then the log-likelihood is given by:

$$\mathcal{LL}(\mu) = -\frac{1}{2\sigma^2}(X - \mu)^2,$$

so that the score is:

$$s(\mu) = \frac{X - \mu}{\sigma^2},$$

and the Fisher information is:

$$I(\mu) = \mathbb{E}\left[\frac{(X - \mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

Notice the connection between Fisher information and the variance.

An important property that we will use in the sequel is that the score function has mean zero, i.e.

$$\mathbb{E}_{p(X_1,\ldots,X_n;\theta)}[s(\theta)] = 0.$$

Proof: Notice that,

$$\begin{aligned}\mathbb{E}_{p(X_1,\ldots,X_n;\theta)}[s(\theta)] &= \sum_{i=1}^{n} \int \nabla_\theta \log p(x_i;\theta) p(x_1,\ldots,x_n;\theta) dx_1 dx_2 \ldots dx_n \\ &= \sum_{i=1}^{n} \int \nabla_\theta \log p(x_i;\theta) p(x_i;\theta) dx_i \\ &= n \int \nabla_\theta \log p(x_1;\theta) p(x_1;\theta) dx_1,\end{aligned}$$

using the i.i.d. assumption several times. Under some regularity conditions we can switch the derivative and integral (essentially the dominated convergence theorem again but see the Lehmann and Casella book for details) so we obtain,

$$\begin{aligned}\int \nabla_\theta \log p(x_1;\theta) p(x_1;\theta) dx_1 &= \int \frac{\nabla_\theta p(x_1;\theta)}{p(x_1;\theta)} p(x_1;\theta) dx_1 \\ &= \nabla_\theta \int p(x_1;\theta) dx_1 = \nabla_\theta 1 = 0.\end{aligned}$$

One simple consequence of this property is that we can interpret the Fisher information matrix as the covariance matrix of the score, i.e.

$$I(\theta) = \mathbb{E}[(s(\theta) - \mathbb{E}(s(\theta)))(s(\theta) - \mathbb{E}(s(\theta)))^T].$$

So at a very rough level, we might expect that if the Fisher information is large then parameter estimation is easy since small changes in the parameter result in a very noticeable change in the score function. We will make this more precise soon.

Suppose first that we only have 1 sample from the model. In this case, the Fisher information can alternatively be defined as:

$$I_1(\theta) = -\mathbb{E}\left[\nabla_\theta^2 \log p(X;\theta)\right].$$

To see this observe that,

$$\begin{aligned}\nabla_\theta^2 \log p(X;\theta) &= \nabla_\theta \frac{\nabla_\theta p(X;\theta)}{p(X;\theta)} \\ &= \frac{\nabla_\theta^2 p(X;\theta)}{p(X;\theta)} - \frac{(\nabla_\theta p(X;\theta) \nabla_\theta p(X;\theta)^T)}{p(X;\theta)^2} \\ &= \frac{\nabla_\theta^2 p(X;\theta)}{p(X;\theta)} - s(\theta) s(\theta)^T.\end{aligned}$$

Now, notice that,

$$\mathbb{E}\left[\frac{\nabla_\theta^2 p(X;\theta)}{p(X;\theta)}\right] = \int \nabla_\theta^2 p(X;\theta) = \nabla_\theta^2 \int p(X;\theta) = \nabla_\theta^2 1 = 0,$$

which in turn yields the desired alternate definition of the Fisher information. In essence the Fisher information is measuring the expected curvature of the log-likelihood function around the point θ. As we will see in future lectures if the log-likelihood is more curved (i.e. $I(\theta)$ is appropriately "large") then θ is easier to estimate.

Another observation from the above representation is that if we instead had n i.i.d. samples, then

$$\begin{aligned} I(\theta) &= -\mathbb{E}\left[\nabla^2_\theta \log p(X_1, \ldots, X_n; \theta)\right] \\ &= -\sum_{i=1}^{n} \mathbb{E}\left[\nabla^2_\theta \log p(X_i; \theta)\right] \\ &= nI_1(\theta). \end{aligned}$$

Example 3: The above in turn yields that if we observed $X_1, \ldots, X_n \sim \text{Ber}(p)$ the Fisher information would be,

$$I(p) = \frac{n}{p(1-p)}.$$

As a more abstract example we can consider the case of general exponential families.

Example 4: For exponential families we have seen that the log-likelihood is given as:

$$\mathcal{LL}(\theta; X_1, \ldots, X_n) = \sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(X_j) - nA(\theta),$$

so the Hessian is simply,

$$I(\theta) = n\nabla^2_\theta A(\theta) = n\mathbb{E}[(T(X) - \mathbb{E}[T(X)])(T(X) - \mathbb{E}[T(X)])^T],$$

i.e. the Fisher information matrix is simply given by (n times) the Hessian of the log-partition function or alternatively it is the covariance matrix of the vector of sufficient statistics.

15.2.2 Cramér-Rao Bound

Let us briefly consider again the Bernoulli example. We observe $X_1, \ldots, X_n \sim \text{Ber}(p)$ and estimate $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The estimator is unbiased and has variance $p(1-p)/n$ which is precisely the inverse of the Fisher information.

This turns out to be a fairly general phenomenon. Indeed, the Cramér-Rao bound assures us that this estimator is unimprovable in a certain sense. We focus first on the univariate case (when $\theta \in \mathbb{R}$) and then consider the multivariate extension.

Cramér-Rao Bound: Suppose that we observe $X_1, \ldots, X_n \sim p(X;\theta)$ and that $\widehat{\theta}$ is an unbiased estimator of θ, then:

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{nI_1(\theta)}.$$

Proof: Consider that,

$$\begin{aligned}\mathrm{cov}(\widehat{\theta}, s(\theta)) &= \mathbb{E}((\widehat{\theta}-\theta)s(\theta)) \\ &= \mathbb{E}(\widehat{\theta}s(\theta)),\end{aligned}$$

since $\mathbb{E}[s(\theta)] = 0$. Furthermore,

$$\begin{aligned}\mathbb{E}(\widehat{\theta}s(\theta)) &= \int \widehat{\theta}(x_1,\ldots,x_n)\nabla_\theta \log p(x_1,\ldots,x_n;\theta)p(x_1,\ldots,x_n;\theta)dx_1\ldots dx_n \\ &= \int \widehat{\theta}(x_1,\ldots,x_n)\frac{\nabla_\theta p(x_1,\ldots,x_n;\theta)}{p(x_1,\ldots,x_n;\theta)}p(x_1,\ldots,x_n;\theta)dx_1\ldots dx_n \\ &= \nabla_\theta\theta = 1.\end{aligned}$$

Notice that for any fixed ζ we can write:

$$\mathrm{Var}(\widehat{\theta}-\zeta s(\theta)) = \mathrm{Var}(\widehat{\theta}) + \zeta^2\mathrm{Var}(s(\theta)) - 2\zeta\mathrm{cov}(\widehat{\theta}, s(\theta)) = \mathrm{Var}(\widehat{\theta}) + \zeta^2 nI_1(\theta) - 2\zeta.$$

Using the fact that variances are positive we can write:

$$\mathrm{Var}(\widehat{\theta}) \geq 2\zeta - \zeta^2 nI_1(\theta)$$

Take $\zeta = 1/(nI_1(\theta))$ to obtain the Cramér-Rao bound.

Multivariate Generalization: The Cramér-Rao bound can be derived for a multivariate parameter θ in a very similar fashion and in this case leads to the conclusion:

$$\mathrm{Var}(\widehat{\theta}) \succeq I(\theta)^{-1} = \frac{1}{n}I_1(\theta)^{-1},$$

where we are comparing two positive semi-definite matrices so the ordering is the Loewner ordering, i.e. for any vector v we have that,

$$v^T\mathrm{Var}(\widehat{\theta})v \geq v^T I(\theta)^{-1}v.$$

Examples: In both the Gaussian and Bernoulli models, as a consequence of the Cramér-Rao bound we can conclude that the MLE is the best unbiased estimator.

Important Note: The Cramér-Rao bound holds, in an asymptotic sense, for substantially more general settings (without the unbiasedness requirement). For example, see the book of Van der Vaart (on Asymptotic Statistics) which shows that under appropriate conditions (known as quadratic mean differentiability or local asymptotic normality) that no estimator can have smaller mean squared error than Fisher information in any uniform sense. This is a deep result and central to the optimality of the MLE in more general settings and is one of Lucien Le Cam's main contributions to the theory of statistics.

15.3 Beyond unbiased estimators - decision theory

The central idea in decision theory is that we want to minimize our *expected* loss.

Let us first try to understand the decision theoretic setup. We observe data $X_1, \ldots, X_n \sim p(X;\theta)$, with $\theta \in \Theta$, and we make a decision, i.e. we select an action a.

In point estimation, the decision is just our guess of the parameter. In hypothesis testing situations our decision will instead be which of the hypotheses we believe to be true. Once we take an action we suffer a loss. The loss function in point estimation is roughly something that is large if a is far from θ and small if our guess is good, i.e., if a is close to θ.

Some very common loss functions are:

1. **Squared loss:** $L(a,\theta) = (a-\theta)^2$.
2. **Absolute loss:** $L(a,\theta) = |a-\theta|$.

There are however many other loss functions. For instance, we sometimes consider losses like:

$$L(a,\theta) = \frac{(a-\theta)^2}{|\theta|+1},$$

which penalizes errors in estimation more for small values of θ than for large values. We can similarly design a loss function that penalizes errors more strongly for large values of θ.

Another important point is that there are cases when we do not really care about estimating the parameter well but rather just the distribution $p(X;\theta)$. This is true when we care about prediction in regression or in density estimation. In this case we could define the loss between θ and a in terms of the distributions $p(X;\theta)$ and $p(X;a)$. One canonical example:

Kullback-Leibler loss:

$$L(a,\theta) = \mathrm{KL}(p(X;\theta), p(X;a)) = \mathbb{E}_{X\sim p(X;\theta)} \log\left(\frac{p(X;\theta)}{p(X;a)}\right).$$

Once we have a loss function, and an estimator, we can assess the estimator via its expected loss. This expected loss is called the *risk* of the estimator. Suppose we consider an estimator $\widehat{\theta}(X)$. Then we define:

$$R(\theta, \widehat{\theta}(X)) = \mathbb{E}_\theta L(\widehat{\theta}(X), \theta).$$

In general, we do not a-priori know anything about the value of θ so we would like estimators with low risk for all parameters $\theta \in \Theta$. So ideally, we would like to find an estimator $\widehat{\theta}$ such that for any other estimator θ' we have that:

$$R(\theta, \widehat{\theta}(X)) \leq R(\theta, \theta')$$

for all values θ. Such estimators will most often not exist – why not?

Example: Suppose $X \sim N(\theta, 1)$, and we care about estimating θ in MSE. Consider two estimators: $\widehat{\theta} = X$ and $\widehat{\theta} = 0$. The risk of X is: $\mathbb{E}(X - \theta)^2 = 1$, while the risk of 0 is $\mathbb{E}\theta^2 = \theta^2$. So when $\theta < 1$, 0 is a better estimator than the estimator X. Neither estimator dominates the other.

Example: Let us consider 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.

A lot of statistical theory was developed from this decision theoretic starting point. At a high-level there are several different paradigms and ideas:

1. The notion of admissibility: With our decision theoretic mechanism in place we could attempt to weed out the really useless estimators. Particularly, it seems natural to disregard an estimator $\widehat{\theta}_1$ 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 $\theta \in \Theta$. Estimators like θ_1 are called *inadmissible* estimators. As one might expect there are often many admissible estimators so we need other ways to narrow our search further.

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

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

More generally however the minimax risk idea suggests comparing two estimators on the basis of their worst-case risk. For various reasons this is one of the dominant paradigms for evaluating estimators. This is because for many problems of interest we can actually find the minimax estimator (or at least one that achieves the minimax risk upto constants).

3. Bayes risk: The Bayes risk of an estimator is its risk averaged with respect to a prior, i.e., for some prior $\pi(\theta)$:

$$R_\pi(\widehat{\theta}) = \mathbb{E}_{\theta \sim \pi} R(\theta, \widehat{\theta}(X)).$$

The Bayes risk (similar to the worst-case risk) of an estimator is just a number and thus it is easy to compare two estimators. A natural estimator is one that minimizes the Bayes risk, and this is sometimes called the Bayes estimator. This might seem as intractable as finding an optimal estimator by any other metric but it turns out that we can simplify the problem a bit.

Bayes Estimator: Suppose that $\theta \sim \pi$ and $X \sim p(X;\theta)$. Observe that the above is defining the conditional distribution of X: we denote the marginal distribution as $m(X)$. Then the Bayes risk of an estimator is:

$$\begin{aligned}\int_\theta R(\theta, \widehat{\theta}(X))\pi(\theta)d\theta &= \int_\theta \left[\int_X L(\theta, \widehat{\theta}(X))p(X;\theta)(X)dX\right]\pi(\theta)d\theta \\ &= \int_X \left[\int_\theta L(\theta, \widehat{\theta})\pi(\theta|X)d\theta\right] m(X)dX.\end{aligned}$$

Now, the Bayes estimator just minimizes this expression, so we can see that for every X our estimator is given by:

$$\widehat{\theta}(X) = \arg\min_{\theta'} \int_\theta L(\theta, \theta'(X))\pi(\theta|X)d\theta.$$

The term on the RHS is called the posterior expected loss. So the Bayes estimator is one that minimizes the posterior expected loss. A simple but important special case of this is that if L is the squared loss then the Bayes estimator is a conditional expectation, i.e.,

$$\widehat{\theta}(X) = \mathbb{E}[\theta|X].$$

As a point of comparison, the max-risk does not involve the choice of an arbitrary prior so in that sense has some advantages over the Bayes risk.

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. In the next lecture we will show that this estimator is actually minimax optimal.