Stats 300A session 7

TA: Song Mei

November 8, 2018

1 Overview

The main task of estimation theory is to compare the quality of estimators using risk function. Bayes risk and minimax risk are summarization of the risk function, and they are standard ways to compare the quality of estimators.

However, sometimes the statistical model is too complex (especially in high dimensional statistics) so that it is hard to calculate the minimax risk for a specific statistical model, or it is hard to establish optimality exhausting all the estimators, or in practice we are restricted to use all the estimators. There are in general two general approaches to overcome this difficulty:

- Constrain the class of estimators. Examples include: unbiased, equi-variant, linear, robust, computationally tractable, differential private, etc.
- Discuss about asymptotic minimax or approximate minimax.

Developing these two general approaches for specific tasks is still an active research direction.

Chapter 4 of the lecture notes studies methods to establish approximate minimax. Chapter 5 of the lecture notes establishes optimality within a constrained class of estimators: the unbiased estimators.

2 Theory of unbiased estimation (Chapter 5 of lecture notes)

The theory of unbiased estimation has two important ingredients:

- Establish UMVU estimator (using complete sufficient statistics).
- Establish approximate unbiased optimality (fisher information lower bound).

2.1 Uniformly minimal variance unbiased (UMVU) estimator

Consider statistical model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, and $\boldsymbol{X} \sim \mathsf{P}_{\boldsymbol{\theta}}$. Suppose we would like to estimate $g(\boldsymbol{\theta})$, and we consider the loss function $L(a, \boldsymbol{\theta}) = (a - g(\boldsymbol{\theta}))^2$.

Definition 1 (Unbiasedness)**.** *An estimator A is unbiased for $g(\boldsymbol{\theta})$ if $\mathsf{E}_\theta[A(\boldsymbol{X})] = g(\boldsymbol{\theta})$.*

Definition 2 (Uniformly minimal variance unbiased (UMVU) estimator)**.** *An unbiased estimator A for $g(\boldsymbol{\theta})$ is UMVU w.r.t. statistical model $\mathcal{P}$ for $g(\boldsymbol{\theta})$, if for any unbiased estimator $\tilde{A}$, we have*

$$Var_{\boldsymbol{\theta}}(A) = R(A, \boldsymbol{\theta}) \le R(\tilde{A}, \boldsymbol{\theta}) = Var_{\boldsymbol{\theta}}(\tilde{A}), \qquad \forall \boldsymbol{\theta} \in \Theta.$$

Remark 1. *UMVU is a stronger criteria than unbiased minimax. An UMVU estimator A is also minimax among unbiased estimators, but a minimax unbiased estimator may not have uniformly best risk function among unbiased estimators.*

There is a principled approach to find the UMVU, if we have an unbiased estimator A for $g(\boldsymbol{\theta})$, and a "complete" sufficient statistics $\boldsymbol{T}(\boldsymbol{x})$.

Theorem 1. *Let A be a "sufficiently nice" unbiased estimator for $g(\boldsymbol{\theta})$. Let $\boldsymbol{T}(\boldsymbol{x})$ be a "complete" sufficient statistics. Define*

$$A_\star(\boldsymbol{t}) = \mathsf{E}_{\boldsymbol{\theta}}[A(\boldsymbol{X})|\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t}].$$

Then $A_\star(\boldsymbol{T}(\boldsymbol{x}))$ is UMVU for $g(\boldsymbol{\theta})$.

Definition 3 (Complete statistics)**.** *A sufficient statistics $\boldsymbol{T}(\boldsymbol{X})$ is complete if: for any function f such that $\mathsf{E}_{\boldsymbol{\theta}}[f(\boldsymbol{T}(\boldsymbol{X}))] = 0$ for any $\boldsymbol{\theta}$, we have $\mathsf{P}_{\boldsymbol{\theta}}(f(\boldsymbol{T}(\boldsymbol{X})) = 0) = 1$ for any $\boldsymbol{\theta}$.*

Remark 2. *To check a sufficient statistics is complete is not an easy task. Here are some methods*

- *Exponential family (with some technical conditions).*
- *A model contains an exponential family (with some technical conditions).*
- *Check by definition.*

2.1.1 Examples: proving completeness and deriving UMVU

Example 1 (Bernoulli random experiment)**.** Consider $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ for $\theta \in [0, 1]$. The joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ gives

$$\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Since this is an exponential family, we know $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$ is a complete sufficient statistics.

Now we consider the UMVU estimator for θ. Since X_1 is an unbiased estimator for θ, we have

$$A_\star(t) = \mathsf{E}_\theta[X_1|T(\boldsymbol{X}) = t] = \frac{1}{n}\sum_{i=1}^n \mathsf{E}_\theta[X_i|T(\boldsymbol{X}) = t] = (1/n)\mathsf{E}_\theta[T(\boldsymbol{X})|T(\boldsymbol{X}) = t] = t/n.$$

Hence the UMVU estimator for θ gives $A_\star(T(\boldsymbol{x})) = T(\boldsymbol{x})/n = \sum_{i=1}^n x_i/n$.

Then we consider the UMVU estimator for θ^2. Naively we want to use the estimator $A(T(\boldsymbol{x})) = (T(\boldsymbol{x})/n)^2$. But this is not an unbiased estimator. To compute UMVU, first we need to come up with an unbiased estimator for θ^2. It can be easily checked that $A(\boldsymbol{x}) = \mathbf{1}\{x_1 = x_2 = 1\}$ is an unbiased estimator for θ^2. To compute UMVU, we have

$$\begin{aligned} A_\star(t) =& \mathsf{E}_\theta[\mathbf{1}\{X_1 = X_2 = 1\}|T(\boldsymbol{X}) = t] = \mathsf{P}_\theta(X_1 = X_2 = 1|T(\boldsymbol{X}) = t) \\ =& \mathsf{P}_\theta\Big(X_1 = X_2 = 1, \sum_{i=3}^n X_i = t-2\Big)/\mathsf{P}_\theta\Big(\sum_{i=1}^n X_i = t\Big) \\ =& \theta^2\binom{n-2}{t-2}\theta^{t-2}(1-\theta)^{n-t}\mathbf{1}\{t \geq 2\}/\Big[\binom{n}{t}\theta^t(1-\theta)^{n-t}\Big] \\ =& t(t-1)/[n(n-1)]. \end{aligned}$$

Hence the UMVU estimator for θ^2 gives $A_\star(T(\boldsymbol{x})) = T(\boldsymbol{x})(T(\boldsymbol{x})-1)/[n(n-1)]$. This estimator intuitively makes sense, since it is very similar to the naive guess $A(T(\boldsymbol{x})) = (T(\boldsymbol{x})/n)^2$ (so that we are more confident that our calculus is correct).

Example 2 (Uniform distribution)**.** Consider $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$ for some $\theta \in (0, \infty)$. The joint distribution of $\boldsymbol{X}$ gives

$$\mathsf{p}_\theta(\boldsymbol{x}) = (1/\theta^n)\mathbf{1}\{\max_i x_i \leq \theta\}.$$

It is easy to see that $T(\boldsymbol{x}) = \max_i x_i$ is a sufficient statistics. But the uniform distribution family is not an exponential family, we cannot say $T(\boldsymbol{x})$ is the complete sufficient statistics directly.

To check $T(\boldsymbol{x})$ is complete, we check the definition of completeness directly. The probability of event $\{T(\boldsymbol{x}) \leq u\}$ under measure P_θ gives (for $u \leq \theta$)

$$\mathsf{P}_\theta(\max_i X_i \leq u) = (u/\theta)^n.$$

Taking derivative with respect to u, the density of $T(\boldsymbol{x})$ under measure P_θ gives

$$q_\theta(u) = n(u^{n-1}/\theta^n)\mathbf{1}\{u \le \theta\}.$$

For any function f such that $\mathsf{E}_\theta[f(T(\boldsymbol{X}))] = 0$ for any $\theta \in (0, \infty)$, we have

$$\mathsf{E}_\theta[f(T(\boldsymbol{X}))] = \int_0^\theta q_\theta(u)g(u)\mathrm{d}u = \int_0^\theta nu^{n-1}/\theta^n g(u)\mathrm{d}u = 0,$$

which gives

$$\int_0^\theta u^{n-1}g(u)\mathrm{d}u = 0$$

for any $\theta \in (0, \infty)$. Taking derivative with respect to θ, we have

$$\theta^{n-1}g(\theta) = 0,$$

which gives $g(\theta) = 0$ for $\theta \in (0, \infty)$. Hence $T(\boldsymbol{x}) = \max_i x_i$ is complete.

To find the UMVU for parameter θ, note that $2X_1$ gives an unbiased estimator for θ. We would like to compute

$$A_\star(t) = 2\mathsf{E}_\theta[X_1 | \max_i X_i = t].$$

The conditional distribution of $[X_1 | \max_i X_i = t]$ gives

$$[X_1 | \max_i X_i = t] \sim \frac{1}{n}\delta_t + \frac{n-1}{n}\mathrm{Unif}(0, t),$$

where δ_t is the Dirac delta measure at location t. To see this, we compute $\mathsf{P}_\theta(X_1 \le x, \max_i X_i \ge t)$:

$$\begin{aligned}
&\mathsf{P}_\theta\Big(X_1 \le x, \max_i X_i \ge t\Big) = \int_0^x \mathsf{p}_\theta(x_1)\mathsf{P}_\theta\Big(\max_i X_i \ge t \Big| X_1 = x_1\Big)\mathrm{d}x_1 \\
=& \int_0^x \mathsf{p}_\theta(x_1)\Big[\mathsf{P}_\theta\Big(\max_i X_i = X_1 \Big| X_1 = x_1\Big)\mathbf{1}\{t \le x_1\} + \mathsf{P}_\theta\Big(\max_i X_i \ge t \Big| X_1 = x_1\Big)\mathbf{1}\{t \ge x_1\}\Big]\mathrm{d}x_1 \\
=& \int_0^x \mathsf{p}_\theta(x_1)\Big[\mathsf{P}_\theta\Big(\max_{i\ge 2} X_i \le x_1\Big)\mathbf{1}\{t \le x_1\} + \mathsf{P}_\theta\Big(\max_{i \ge 2} X_i \ge t\Big)\mathbf{1}\{t \ge x_1\}\Big]\mathrm{d}x_1 \\
=& \int_0^x \frac{1}{\theta}\Big[\Big(\frac{x_1}{\theta}\Big)^{n-1}\mathbf{1}\{t \le x_1\} + \Big(1 - \Big(\frac{t}{\theta}\Big)^{n-1}\Big)\mathbf{1}\{t \ge x_1\}\Big]\mathrm{d}x_1 \\
=& \int_0^x \frac{1}{\theta}\Big[\Big(\frac{x_1}{\theta}\Big)^{n-1}\mathbf{1}\{t \le x_1\}\Big]\mathrm{d}x_1 + \frac{\min\{x, t\}}{\theta}\Big(1 - \Big(\frac{t}{\theta}\Big)^{n-1}\Big).
\end{aligned}$$

Taking derivative with respect to x, t gives the density for $(X_1, T(\boldsymbol{X}))$ at (x, t), which gives (for $0 \le x, t \le \theta$)

$$q_{X_1, T(\boldsymbol{X})}(x, t) = -\partial_x\partial_t \mathsf{P}_\theta\Big(X_1 \le x, \max_i X_i \ge t\Big) = \frac{t^{n-1}}{\theta^n}\delta(t - x) + (n-1)\frac{t^{n-2}}{\theta^n}\mathbf{1}\{0 \le x \le t\}.$$

Note the density of $T(\boldsymbol{X})$ gives $q_\theta(t) = (nt^{n-1}/\theta^n)\mathbf{1}\{t \le \theta\}$. Hence the conditional density $[X_1 | \max_i X_i = t]$ gives

$$q_{X_1 | T(\boldsymbol{X})}(x | t) = \frac{1}{n}\delta(t - x) + \frac{(n-1)}{n}\frac{1}{t}\mathbf{1}\{0 \le x \le t\}.$$

Given this result, we have

$$\begin{aligned}
A_\star(t) =& 2\mathsf{E}_\theta[X_1 | \max_i X_i = t] \\
=& 2\Big[\frac{1}{n}t + \frac{n-1}{n}\int_0^t \frac{x_1}{t}\mathrm{d}x_1\Big] \\
=& \frac{n+1}{n}t.
\end{aligned}$$

The UMVU is $A_\star(T(\boldsymbol{x})) = [(n+1)/n]\max_i x_i$, an augmented max of x_i's.

2.1.2 Example: minimal sufficient but not complete statistics

Definition 4. *A sufficient statistics $\boldsymbol{T}(\boldsymbol{x})$ is minimal for the model $(\mathsf{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ if, for any other sufficient statistics $\boldsymbol{T}'(\boldsymbol{x})$, there exists a measurable function f such that $T(\boldsymbol{x}) = f(T'(\boldsymbol{x}))$ for all $\boldsymbol{x}$.*

Lemma 1. *$T(\boldsymbol{x})$ a minimal sufficient statistics if and only if the following holds for all $\boldsymbol{x}, \boldsymbol{y}$*

$$T(\boldsymbol{x}) = T(\boldsymbol{y}) \Leftrightarrow \exists C_{\boldsymbol{x},\boldsymbol{y}}, s.t., \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) = C_{\boldsymbol{x},\boldsymbol{y}}\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{y}), \forall \boldsymbol{\theta} \in \Theta.$$

Example 3 (Minimal sufficient but not complete statistics)**.** Consider the example $X_1, \ldots, X_n \sim \mathrm{Unif}([\theta, \theta+1])$ for $\theta \in \mathbb{R}$. Then $\boldsymbol{T}(\boldsymbol{x}) = (\max_i x_i, \min_i x_i)$ is a minimal sufficient statistics, but not complete.

It is easy to see that $\boldsymbol{T}$ is sufficient. To show it is minimal, for any $\boldsymbol{x}, \boldsymbol{y}$ such that $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$, we have

$$\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{1}\{\theta \le \min_i x_i \le \max_i x_i \le \theta + 1\} = \mathbf{1}\{\theta \le \min_i y_i \le \max_i y_i \le \theta + 1\} = \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{y}).$$

For any $\boldsymbol{x}, \boldsymbol{y}$ such that

$$\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{1}\{\theta \le \min_i x_i \le \max_i x_i \le \theta + 1\} = C_{\boldsymbol{x},\boldsymbol{y}}\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{y}) = C_{\boldsymbol{x},\boldsymbol{y}}\mathbf{1}\{\theta \le \min_i y_i \le \max_i y_i \le \theta + 1\},$$

for any θ, we must have $C_{\boldsymbol{x},\boldsymbol{y}} = 1$ and $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. Hence $\boldsymbol{T}$ is minimal.

To show $\boldsymbol{T}$ is not complete, let $g(\boldsymbol{T}(\boldsymbol{x})) = \max_i x_i - \min_i x_i$, it is easy to see that $\mathbb{E}_\theta[g(\boldsymbol{T}(\boldsymbol{x}))]$ is independent of θ. But g is not identically a constant. Hence $\boldsymbol{T}$ is not complete.

2.2 Cramer-Rao lower bound

We consider sufficiently smooth, positive, and fast decaying density $\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})$.

Definition 5 (Fisher information)**.** *Define $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{\theta}} \log \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{\theta}}\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})/\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})$. The fisher information matrix is defined as*

$$\mathrm{I}_{\mathrm{F}}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}[\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{X})\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{X})^{\mathsf{T}}].$$

which can be also written as

$$\mathrm{I}_{\mathrm{F}} = \int \nabla_{\boldsymbol{\theta}}\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathsf{T}}/\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = -\int \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla^2_{\boldsymbol{\theta}} \log \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Theorem 2 (Cramer-Rao lower bound)**.** *For any sufficiently good unbiased estimator $\hat{\boldsymbol{\theta}}$, define*

$$\boldsymbol{M} = \int (\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta})^{\mathsf{T}}\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Then we have

$$\boldsymbol{M} \succeq \mathrm{I}_{\mathrm{F}}(\boldsymbol{\theta})^{-1}.$$

Proof. For any $\boldsymbol{u}, \boldsymbol{v}$, we have

$$\begin{aligned}\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \int \langle \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{v}\rangle\langle \hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}, \boldsymbol{u}\rangle \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \le& \Big[\int \langle \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{v}\rangle^2 \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big] \cdot \Big[\int \langle \hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}, \boldsymbol{u}\rangle^2 \mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big] \\ =& \langle \boldsymbol{v}, \mathrm{I}_{\mathrm{F}}(\boldsymbol{\theta})\boldsymbol{v}\rangle\langle \boldsymbol{u}, \boldsymbol{M}\boldsymbol{u}\rangle,\end{aligned}$$

A special choice of $\boldsymbol{u}$ and $\boldsymbol{v}$ gives the desired inequality.

□

2.2.1 Example: computing fisher information matrix

Example 4 (Bernoulli random experiments)**.** Let $X_1, \ldots, X_n \sim \text{Ber}(\theta)$. The fisher information for their joint distribution is n times the fisher information of a single random variable.

The density $\mathsf{p}_\theta(x) = \theta^x(1-\theta)^{1-x}$, hence

$$\dot{\ell}_\theta(x) = \mathrm{d}/(\mathrm{d}\theta)[x \log \theta + (1-x)\log(1-\theta)] = x/\theta - (1-x)/(1-\theta) = x/[\theta(1-\theta)] - 1/(1-\theta).$$

The fisher information for a single Bernoulli distribution gives

$$\begin{aligned}
\mathrm{I}_{\mathrm{F},1}(\theta) =& \mathsf{E}_\theta[\dot{\ell}_\theta(X)^2] = \mathsf{E}_\theta[[X/[\theta(1-\theta)] - 1/(1-\theta)]^2] \\
=& \mathsf{E}_\theta[X^2/[\theta(1-\theta)]^2 + 1/(1-\theta)^2 - 2X/[\theta(1-\theta)^2]] \\
=& \theta/[\theta(1-\theta)]^2 + 1/(1-\theta)^2 - 2\theta/[\theta(1-\theta)^2] \\
=& 1/[\theta(1-\theta)^2] + 1/(1-\theta)^2 - 2/(1-\theta)^2 \\
=& 1/[\theta(1-\theta)].
\end{aligned}$$

The fisher information for n independent Bernoulli random variable gives $\mathrm{I}_{\mathrm{F},n}(\theta) = n/[\theta(1-\theta)]$. Therefore, for any unbiased estimator $\hat{\theta}$, the risk lower bound gives

$$R(\hat{\theta}, \theta) \geq 1/\mathrm{I}_{\mathrm{F},n}(\theta) = \theta(1-\theta)/n.$$

The sample mean estimator $\hat{\theta}_U(\boldsymbol{x}) = \sum_{i=1}^n x_i/n$ achieves this lower bound at any θ, hence the sample mean estimator is UMVU.

Remember that the minimax estimator on $\theta \in [0,1]$ is $\hat{\theta}_M(\boldsymbol{x}) = (\sum_{i=1}^n x_i + \sqrt{n}/2)/(n+\sqrt{n})$, and it has better worst risk than the mean estimator $\hat{\theta}_U(\boldsymbol{x})$. But the minimax estimator $\hat{\theta}_M(\boldsymbol{x})$ is not unbiased. In the class of unbiased estimators, $\hat{\theta}_U(\boldsymbol{x})$ is minimax optimal.

2.2.2 An important property of Fisher information (Stats 300B)

Theorem 3 (Asymptotic normality of maximum likelihood estimator)**.** *Let $X_1 \ldots, X_n \sim \mathsf{p}_{\boldsymbol{\theta}}(x)$ and let $\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \log \mathsf{p}_{\boldsymbol{\theta}}(x_i)$. Then we have*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}) \to \mathcal{N}(0, \mathrm{I}_{\mathrm{F}}(\boldsymbol{\theta})^{-1})$$

weakly as $n \to \infty$.

This theorem says that, maximum likelihood estimator asymptotically achieves Cramer-Rao lower bound.