Stats 300A session 4

TA: Song Mei

October 19, 2018

1 A short review

Within the decision theory framework, two most important tasks are estimation and testing. In the first half quarter, we are studying estimation. The core of theory of estimation 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 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. In this session, I will discuss location equi-variant estimation.

2 Location equi-variant estimation

Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ be a statistical model, $\mathbb{P}_\theta$ is a probability measure on $\mathbb{R}^n$ and has density $f_\theta(x_1, \ldots, x_n)$.

Definition 1 (Location invariant statistical model, loss function)**.** *We say the above statistical model is location invariant if* $f_{\theta+c}(x_1 + c, \ldots, x_n + c) = f_\theta(x_1, \ldots, x_n)$ *for* $c \in \mathbb{R}$*. A loss function* $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is invariant of* $L(a, \theta) = L(a + c, \theta + c) = \rho(a - \theta)$ *for* $c \in \mathbb{R}$*.*

Definition 2 (Location equi-variant statistical estimator)**.** *A statistical estimator* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}$ *is location equi-variant if*

$$\hat{\theta}(x_1, \ldots, x_n) + c = \hat{\theta}(x_1 + c, \ldots, x_n + c).$$

Example 1 (Exponential model)**.** Suppose $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta, b)$ with an unknown location parameter θ and a known scale parameter $b > 0$. The density of each of the X_i has the form

$$f_\theta(x_i) = \frac{1}{b} \exp\Big(- \frac{(x_i - \theta)}{b} \Big) \mathbf{1}\{x_i \geq \theta\}.$$

The loss function can be taken to be $L_1(a, \theta) = |a - \theta|$, $L_2(a, \theta) = (a - \theta)^2$. Then this statistical model is location equivariant, and the loss function is also equivariant.

It is very easy to find an equivariant estimator. For example, x_1, $\mathrm{median}\{x_1, \ldots, x_n\}$, $\bar{x}$, $\min\{x_1, \ldots, x_n\}$ are location equivariant estimator. The other estimators, like $n/[\sum_{i=1}^n 1/x_i]$, are not equi-variant estimators.

We would like to compare the quality of the class of equi-variant estimators.

Lemma 1. *The risk function of any equi-variant estimator is a constant function.*

Proof. We just write down the risk function

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta[L(\hat{\theta}(\boldsymbol{X}), \theta)] = \mathbb{E}_0[L(\hat{\theta}(\boldsymbol{X} + \theta), \theta)] = \mathbb{E}_0[L(\hat{\theta}(\boldsymbol{X}) + \theta, \theta)] = \mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}))].$$

The right hand side doesn't depend on θ. □

2.1 Minimum risk equivariant estimator

Since the risk function of equi-variant estimators is constant, we can easily compare their quality using risk functions.

Definition 3. *A equi-variant estimator is minimum risk equivariant estimator if it has the smallest risk.*

Theorem 1. *For location invariant statistical model, $\hat{\theta}(\boldsymbol{x})$ is a location equi-variant with finite risk. If for each $\boldsymbol{y} \in \mathbb{R}^{n-1}$, $v_\star(y_1, \ldots, y_{n-1})$ minimizes $U(v, \boldsymbol{y}) = \mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}) - v)|\boldsymbol{Y} = \boldsymbol{y}]$ as a function of v, then an minimum equi-variant estimator gives $\hat{\theta}(\boldsymbol{x}) - v_\star(x_1 - x_n, \ldots, x_{n-1} - x_n)$.*

Example 2 (Exponential model, MRE under absolute loss)**.** Here we find the MRE estimator of the exponential model under the absolute loss.

We start at one of the location equi-variant estimator $\hat{\theta}(\boldsymbol{x}) = \min\{x_1, \ldots, x_n\}$ (we can choose other MRE, but this is simpler). Note $\boldsymbol{Y} = (X_1 - X_n, \ldots, X_{n-1} - X_n)$ is independent of $\hat{\theta}(\boldsymbol{X})$ (write down their joint distribution and check it, or apply Basu's theorem here), we have $v_\star(\boldsymbol{y})$ is a constant function:

$$v_\star(\boldsymbol{y}) = \arg\min_v \mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}) - v(\boldsymbol{y}))|\boldsymbol{Y} = \boldsymbol{y}] = \arg\min_v \mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}) - v)].$$

Note under $\mathbb{P}_0$, $\hat{\theta}(\boldsymbol{X}) \sim \mathrm{EXP}(0, b/n)$. The minimizer of the above when ρ is absolute loss gives the median of $\mathrm{EXP}(0, b/n)$, which gives $v = (b/n) \log 2$.

Hence, the MRE estimator of exponential model under the square loss gives $\hat{\theta}_{M,1}(\boldsymbol{x}) = \min\{x_1, \ldots, x_n\} - (b/n) \log 2$.

Proof of Theorem 1. Proof strategy

(1) Show that all equi-variant estimator must be of form $\hat{\theta}(x_1, \ldots, x_n) - v(x_1 - x_n, \ldots, x_{n-1} - x_n)$, where v can be any function.

(2) The correct minimizing v should be calculated using the posterior.

To show point (1), it is easy to check that a function of the form $\hat{\theta}(\boldsymbol{x}) - v(x_1 - x_n, \ldots, x_{n-1} - x_n)$ is an equi-variant estimator. For any equi-variant estimator $\hat{\theta}_1(\boldsymbol{x})$, we define $v(y_1, \ldots, y_{n-1}) = \hat{\theta}(y_1, \ldots, y_{n-1}, 0) - \hat{\theta}_1(y_1, \ldots, y_{n-1}, 0)$. Then it is easy to see that $\hat{\theta}_1(\boldsymbol{x}) = \hat{\theta}(\boldsymbol{x}) - v(x_1 - x_n, \ldots, x_{n-1} - x_n)$.

To show point (2), note the risk function is constant, we would like to minimize

$$\mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}) - v(\boldsymbol{Y}))] = \int \mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}) - v(\boldsymbol{y}))|\boldsymbol{Y} = \boldsymbol{y}]\mathbb{P}(\mathrm{d}\boldsymbol{y})$$

over the function v. Then $v(\boldsymbol{y})$ is the minimizer of $U(v, \boldsymbol{y}) = \mathbb{E}_0[\rho(\hat{\theta}(\boldsymbol{X}) - v(\boldsymbol{y}))|\boldsymbol{Y} = \boldsymbol{y}]$. $\square$

2.2 Analytical expression of MRE under the square loss

Proposition 1 (Pitman's estimator)**.** *Under the square loss, the MRE estimator $\hat{\theta}_M(\boldsymbol{x})$ gives the form*

$$\hat{\theta}_M(\boldsymbol{x}) = \frac{\int_{-\infty}^{\infty} u f_0(x_1 - u, \ldots, x_n - u)\mathrm{d}u}{\int_{-\infty}^{\infty} f_0(x_1 - u, \ldots, x_n - u)\mathrm{d}u}.$$

Proof. First note $\hat{\theta}(\boldsymbol{x}) = x_n$ is an equivariant estimator. Under the square loss, the theorem above implies that the MRE gives

$$\hat{\theta}_M(\boldsymbol{x}) = x_n - \mathbb{E}_0[X_n|\boldsymbol{Y} = \boldsymbol{y}],$$

where $\boldsymbol{Y} = (X_1 - X_n, \ldots, X_{n-1} - X_n)$, $\boldsymbol{y} = (x_1 - x_n, \ldots, x_{n-1} - x_n)$, and $\mathbb{E}_0$ means $(X_1, \ldots, X_n) \sim \mathbb{P}_0$.

Now we would like to compute $\mathbb{E}_0[X_n|\boldsymbol{Y} = \boldsymbol{y}]$. Define $\boldsymbol{Z} = (X_1 - X_n, \ldots, X_{n-1} - X_n, X_n)$, and denote the density of $\boldsymbol{Z}$ to be $p_{z,0}(y_1, \ldots, y_{n-1}, x_n)$. We have

$$\mathbb{E}_0[X_n|\boldsymbol{Y} = \boldsymbol{y}] = \frac{\int t p_{z,0}(y_1, \ldots, y_{n-1}, t)\mathrm{d}t}{\int p_{z,0}(y_1, \ldots, y_{n-1}, t)\mathrm{d}t}.$$

Note $\boldsymbol{X} = (X_1, \ldots, X_n) = H\boldsymbol{Z}$, where

$$H = \begin{bmatrix} 1 & 0 & \ldots & 1 \\ 0 & 1 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{bmatrix}.$$

and $\det(H) = 1$. Hence we have

$$p_{z,0}(x_1 - x_n, \ldots, x_{n-1} - x_n, t) = p_{x,0}(x_1 - x_n + t, x_2 - x_n + t, \ldots, t) = f_0(x_1 - x_n + t, x_2 - x_n + t, \ldots, t).$$

As a result, we have

$$\mathbb{E}_0[X_n|\boldsymbol{Y} = \boldsymbol{y}] = \frac{\int t f_0(y_1 + t, \ldots, y_{n-1} + t, t)\mathrm{d}t}{\int f_0(y_1 + t, \ldots, y_{n-1} + t, t)\mathrm{d}t}.$$

Now change $t = x_n - u$, we have

$$\begin{aligned}\mathbb{E}_0[X_n|\boldsymbol{Y} = \boldsymbol{y}] =& \frac{\int (x_n - u) f_0(y_1 + x_n - u, \ldots, y_{n-1} + x_n - u, x_n - u)\mathrm{d}u}{\int f_0(y_1 + x_n - u, \ldots, y_{n-1} + x_n - u, x_n - u)\mathrm{d}u} \\ =& x_n - \frac{\int u f_0(x_1 - u, \ldots, x_{n-1} - u, x_n - u)\mathrm{d}u}{\int f_0(x_1 - u, \ldots, x_{n-1} - u, x_n - u)\mathrm{d}u}.\end{aligned}$$

This gives the desired result. □

Example 3 (Exponential model, MRE under square loss)**.** Here we find the MRE estimator of the exponential model under the square loss.

We have

$$\begin{aligned}\hat{\theta}_{M,2}(\boldsymbol{x}) =& \frac{\int_{-\infty}^{\infty} u f_0(x_1 - u, \ldots, x_n - u)\mathrm{d}u}{\int_{-\infty}^{\infty} f_0(x_1 - u, \ldots, x_n - u)\mathrm{d}u} \\ =& \frac{\int_{-\infty}^{\infty} u \prod_i \exp\{-(x_i - u)/b\}\mathbf{1}\{u \le \min\{x_i\}\}\mathrm{d}u}{\int_{-\infty}^{\infty} \prod_i \exp\{-(x_i - u)/b\}\mathbf{1}\{u \le \min\{x_i\}\}\mathrm{d}u} \\ =& \frac{\int_{-\infty}^{\infty} u \exp\{nu/b\}\mathbf{1}\{u \le \min\{x_i\}\}\mathrm{d}u}{\int_{-\infty}^{\infty} \exp\{nu/b\}\mathbf{1}\{u \le \min\{x_i\}\}\mathrm{d}u} = \min\{x_1, \ldots, x_n\} - b/n.\end{aligned}$$

We recall here the MRE under absolute loss is $\hat{\theta}_{M,1}(\boldsymbol{x}) = \min\{x_1, \ldots, x_n\} - (b/n)\log 2$, and the MRE under square loss is $\hat{\theta}_{M,2}(\boldsymbol{x}) = \min\{x_1, \ldots, x_n\} - b/n$.

2.3 The big picture of data science

Here we discuss a little bit about some big pictures upon theories of data science. Statistical decision theory is not the only way to model the data, here we list several useful frameworks and theories:

- Statistical decision theory: statistical model, risk function, Bayes/minimax risk, hypothesis testing.
- Statistical learning theory: statistical model, empirical risk minimization, excessive risk.
- Information theory: statistical model, entropy, channel capacity, mutual information, coding.
- Causal inference theory: potential outcome framework, randomized experiment.
- Complexity theory, approximation theory, numerical linear algebra, theory of optimization...
- Your own framework.

Statistical decision theory focuses on parameter estimation and hypothesis testing. Statistical learning theory focuses on prediction. Information theory focuses on compression and communication. Causal inference theory focuses on finding causation through experiments.