Econ 508B: Lecture 8

Model Estimation

Hongyi Liu

Washington University in St. Louis

August 6, 2017

Outline

Outline

Motivation

Bayes estimators, are one of *Average Risk Optimality* in decision problem. In Bayes, even for a parameter θ, it also has a probability distribution. Hence, we define a measure ν on the parameter space Θ, which has a measurable space. In a sense, ν is regarded as an assignment of different importance weights to the parameter θ, named a *priori*. It represents the subjective beliefs for the parameter based on previous experiments.

Bayes Risk

our optimality goal is to find an estimator δ_ν such that minimizing the average risk below:

$$r(\nu) = \int R(\theta, \delta) d\nu(\theta) = \int \int L(\theta, \delta(x)) dP_\theta(x) d\nu(\theta)$$

- If ν is a probability distribution on Θ, it is referred to as a **prior** distribution.
- The estimator δ_ν is called the **Bayes estimator** with respect to ν if it exists.
- Thus the minimized average risk $r(\nu, \delta_\nu)$ is referred to as the **Bayes risk**.

Bayesian setup

In the Bayesian setup, θ is also treated as a random variable with distribution ν, and the average risk $r(\nu)$ can be represented by $E[L(\Theta, \delta(X))]$.

- In this setup, (X, Θ) are jointly distributed random variables such that a Bayesian statistical model is composed of a *data generation model*, $p(x|\theta)$, and a *prior* distribution on the parameters, $p(\theta)$.
- Using the law of iterated expectations, the average risk could be rewrote as:

$$\begin{aligned} r(\nu) &= E[L(\Theta, \delta(X))] \\ &= E[E[L(\Theta, \delta(X))|\Theta]] \\ &= E[E[L(\Theta, \delta(X))|X]] \end{aligned}$$

Bayesian Paradigm

Theorem 1.1

Suppose that $\Theta \sim \nu$ and $X|\Theta = \theta \sim P_\theta$. If

1. $\exists \delta_0$, an estimator of $g(\theta)$ with finite risk, and
2. a value $\delta_\nu(x)$ which minimizes

$$E[L(\Theta, \delta_\nu(X))|X = x] \text{ for almost every } x,$$

Then δ_ν is a *Bayes estimator* with respect to ν.

Comments

The main idea of the above theorem is that it suffices to consider the conditional risk $E[L(\Theta, \delta(X))|X = x]$ at almost every x.

- To minimize the conditional risk, we should know the marginal distribution of X, where the marginal distribution is given by

$$P(X \in A) = \int P_\theta(X \in A) d\nu(\theta).$$

- The conditional risk is taken with respect to the posterior distribution of Θ given X.

Example 1.1

Consider the loss function $L(\theta, d) = (\theta - d)^2$, the optimality goal is to minimize $E[(g(\Theta) - \delta(X))^2 | X = x]$. Under this case, the Bayes estimator turns out to be $\delta_\nu(X) = E[g(\Theta)|X]$, where the expectation is taken w.r.t. the posterior distribution of Θ given X.

The Posterior Density

the posterior density is given by the formula,

$$\pi(\theta|x) = \underbrace{\frac{\pi(\theta)f(x|\theta)}{f(x)}}_{\frac{joint}{marginal}} = \underbrace{\frac{\pi(\theta)f(x|\theta)}{\int \pi(\theta')f(x|\theta')d\theta'}}_{\frac{prior \times likelihood}{marginal}}$$

- Note that the marginal component is a normalizing term and it does not depend on θ. Hence, we can obtain the following useful property:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Example

Suppose that $X \sim Bin(n, \theta)$ given $\Theta = \theta$ and that Θ has prior distribution $Beta(a, b)$.

- The prior density is given by

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{0<\theta<1}.$$

- The likelihood function is given by

$$f(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{(n-x)}.$$

- By the direct proportion property of posterior density, we have

$$\begin{aligned}\pi(\theta|x) &\propto \binom{n}{x}\theta^{x}(1-\theta)^{(n-x)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{x+a-1}(1-\theta)^{(n-x+b-1)} \sim Beta(x+a, n-x+b).\end{aligned}$$

Continued

Hence, the Bayes estimator of θ under the squared error loss is

$$E[\Theta|X=x] = \frac{x+a}{n+a+b} = \underbrace{\frac{n}{n+a+b}}_{\to 1}\left(\frac{x}{n}\right) + \underbrace{\frac{a+b}{n+a+b}}_{\to 0}\left(\frac{a}{a+b}\right)$$

- It turns out that the Bayes estimate is a **convex combination** of the **sample proportion** X/n (which is the UMVUE) and the **prior mean** $a/a+b$.
- Thus, although the Bayes estimate is **biased** under the *finite* sample because of the modification in light of prior information, it is **asymptotically unbiased** as $n \to \infty$.

Outline

Motivation: Repeated Sampling

Since estimates of population quantities can vary from sample to sample, we regard estimates (and other statistical quantities) as values **sampled** from random variables.

Example 2.1

the estimate $\hat{\theta} = y/n$ depends on y, which is a value sampled from a random Y. Hence the estimator $\hat{\theta}$ is a function of y.

This viewpoint gives rise to the **repeated sampling principle** which means regarding $\hat{\theta}$ as a sampled value of the random variable.

- The repeated sampling principle can be applied to *likelihood methods* and non-likelihood methods. We focus on the likelihood methods in this slide.

The Parametric Statistical Model

Proposition 2.1

Let $\mathcal{F} = \{\mathcal{X}, p_X(x;\theta), \Theta\}$ be a statistical model.

- $\mathcal{X}$ is the sample space; $\Theta \subset \mathbb{R}^p$
- $p_X(x;\theta)$ is a density with respect to a dominating measure μ^*.

- For n independent observations, the sample space is the Cartesian product $\mathcal{X}_n$, and the joint density of the sample is

$$p_X(x;\theta) = \prod_{i=1}^{n} p_{X_i}(x_i, \theta)$$

  with $p_{X_i}(x_i, \theta)$ the density of the ith observation.
- When the n observations are **dependent**, it is often convenient to write their joint density in product form,

$$p_X(x_1, ..., x_n; \theta) = p_{X_1}(x_1; \theta) p_{X_2|X_1=x_1}(x_2; x_1, \theta) \cdots$$
$$p_{X_n|X_1=x_1,...,X_{n-1}=x_{n-1}}(x_n; x_1, ..., x_{n-1}; \theta)$$

The Likelihood Function & Log-likelihood function

Definition 2.1 (likelihood function)

Let $\mathcal{F}$ be a parametric statistical model for a data x specified as above. The **likelihood function** is

$$L = L(\theta) = L(\theta; x) = c(x)p_X(x; \theta),$$

where $\theta \in \Theta$ and $c(x) > 0$ is an arbitrary constant of proportionality.

Definition 2.2 (Log-likelihood function)

For the convenience of calculation, **log-likelihood function** is

$$l = l(\theta) = l(\theta; x) = \log L,$$

For independent observations,

$$l(\theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta).$$

Basic Regularity Conditions

In what follows, assume the log-likelihood:

(1) : is sufficiently smooth function of θ, i.e., it has partial derivatives w.r.t. components of θ up to the required order.

(2) : all the null moments of these derivatives are finite, and whenever we refer to non-null moments, we assume they **exist**.

Partial derivatives of the log-likelihood function are indicated by

$$l_r = l_r(\theta; x) = \frac{\partial}{\partial \theta^r} l(\theta)$$

$$l_{rs} = l_{rs}(\theta; x) = \frac{\partial^2}{\partial \theta^r \partial \theta^s} l(\theta) \qquad \text{(Hessian matrix)}$$

and so forth.

The Score Function

Definition 2.3

The **score function** l_* is the vector of partial derivatives of $l(\theta)$ w.r.t. θ, i.e.,$l_* = l_*(\theta; x) = (l_1, ..., l_p)$.

- Assume that conditions for differentiation and integration are fulfilled, then

$$E_\theta(l_*) = E_\theta[l_*(\theta; X)] = 0$$

i.e. the null first moment of the score is zero.

Information Matrix

Definition 2.4 (Observed information matrix)

The **observed information matrix**, $j(\theta)$, is

$$j = j(\theta) = \begin{pmatrix} -l_{11} & \cdots & -l_{1p} \\ \vdots & \ddots & \vdots \\ -l_{p1} & \cdots & -l_{pp} \end{pmatrix}$$

or in simplified notation, $j = [-l_{rs}]$.

Definition 2.5 (Fisher Information)

The **expected information** or **Fisher information matrix** is

$$i = i(\theta) = E_\theta\{j(\theta)\} = [E_\theta\{-l_{rs}(\theta; Y)\}]$$

Information Identity

We assume conditions that ensure the validity of the **information identity**:

$$E_\theta\{-l_{rs}(\theta)\} = E_\theta\{l_r(\theta)l_s(\theta)\}$$

It says that expected information matrix is the null second moment of the score and is a non-negative definite matrix.

Maximum Likelihood Estimate

A value of θ that maximizes $L(\theta; x)$ over Θ, i.e., a value $\hat{\theta}$ such that $L(\theta) = \sup_{\theta \in \Theta} L(\theta)$ is called **maximum likelihood estimate** of θ.
To find the MLE,

- It is worth trying to the find the MLE through the solutions of the likelihood equation

$$l_*(\theta) = 0. \tag{1}$$

- Intuitively, the true parameter value θ_0 is located close to the value $\hat{\theta}$ which has maximum empirical support in terms of likelihood.
- If (1) fails, we could make use of the monotonicity of $l(\theta)$.

Outline

Motivation: Overidentified Model and Moment Condition

Let $g(x;\theta)$ be an $L \times 1$ function of a $K \times 1$ parameter θ with $L > K$ such that

$$\mathbb{E}g(x;\theta_0) = 0 \tag{M}$$

where θ_0 is the true value of θ.

- We **cannot** in general choose an K-dimensional estimator $\hat{\delta}$ to satisfy the L equations in (M).

GMM Estimator

Define the sample analog of $\mathbb{E}g(x;\theta_0)$

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$$

- If we cannot set $g_n(\hat{\theta})$ exactly equal to 0, we can **at least choose** $\hat{\delta}$ so that $g_n(\hat{\theta})$ is as "close" to 0 as possible.
- To make precise what we mean by **close**, we define the distance between any two L-dimensional vectors ξ and η by the quadratic form $(\xi-\eta)'W_n(\xi-\eta)$, where W_n, referred to as the **weighting matrix**, is a symmetric and positive definite matrix defining the distance.

GMM Estimator

For some $L \times L$ weight matrix $W_n > 0$ such that $W_n \xrightarrow{p} W$ with W symmetric and positive definite, let

$$J_n(\theta, W_n) = n \cdot \bar{g}_n(\theta)' W_n \bar{g}_n(\theta)$$

For example, if $W_n = I$, then $J_n(\theta, I)$ is reduced into $n \cdot ||\bar{g}_n(\theta)||^2$, n multiplying the square of the Euclidean length.

Definition 3.1 (GMM estimator)

The **GMM estimator minimizes** $J_n(\theta, W_n)$,

$$\hat{\theta}_{GMM} = \arg\min_{\theta} J_n(\theta, W_n)$$

GMM Estimator

For some $L \times L$ weight matrix $W_n > 0$ such that $W_n \xrightarrow{p} W$ with W symmetric and positive definite, let

$$J_n(\theta, W_n) = n \cdot \bar{g}_n(\theta)' W_n \bar{g}_n(\theta)$$

For example, if $W_n = I$, then $J_n(\theta, I)$ is reduced into $n \cdot ||\bar{g}_n(\theta)||^2$, n multiplying the square of the Euclidean length.

Definition 3.1 (GMM estimator)

The **GMM estimator minimizes** $J_n(\theta, W_n)$,

$$\hat{\theta}_{GMM} = \arg\min_{\theta} J_n(\theta, W_n)$$

Question: It turns out that $\hat{\theta}_{GMM}$ relies on the choice of W_n, so what is the strategy of choosing W_n?

Estimation of the Efficient Weight Matrix

In Econometrics, a common example to illustrate the **efficient weight matrix** is to refer to the overidentification problem for IV estimation. Please refer to *Fumio Hayashi (2000) & Bruce Hansen (2017)*.

The strategy of choosing efficient weight matrix is as follows

(1) : Set a common weight matrix firstly, which could be wrong, then obtain $\hat{\theta}_1$ by estimation and plug-in the $\hat{\theta}_1$ back to J_n, then we will have the $\hat{g}_i$.

(2) : Construct $W_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i \hat{g}_i'\right)$.

The difference among Bayesian, MLE and GMM

Thoughts?