Chapter 2: Likelihood-based approach (Part 1)

Basic Setup (No missing data)

- $\mathbf{y} = (y_1, y_2, \cdots y_n)$ is a realization of the random sample from a distribution P with density $f(y)$ with dominating measure μ. Thus, $f = dP/d\mu$.
- Assume that the true density $f(y)$ belongs to a parametric family of densities $\mathcal{P} = \{f(y;\theta);\theta \in \Omega\}$ indexed by $\theta \in \Omega \subset \mathbb{R}^p$. That is, there exist $\theta_0 \in \Omega$ such that $f(y;\theta_0) = f(y)$ for all y.

Definitions for likelihood theory

- The likelihood function of θ is defined as

$$L(\theta) = f(\mathbf{y}; \theta)$$

where $f(\mathbf{y}; \theta)$ is the joint density of $\mathbf{y}$.

- Let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of θ_0 if it satisfies

$$L(\hat{\theta}) = \max_{\theta \in \Omega} L(\theta).$$

Identifiability

Definition

Let $\mathcal{P} = \{P_\theta; \theta \in \Omega\}$ be a statistical model with parameter space Ω. Model $\mathcal{P}$ is identifiable if the mapping $\theta \mapsto P_\theta$ is one-to-one:

$$P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2.$$

If the distributions are defined in terms of the probability density functions (pdfs), then two pdfs should be considered distinct only if they differ on a set of non-zero measure (for example two functions $f_1(x) = I(0 \le x < 1)$ and $f_2(x) = I(0 \le x \le 1)$ differ only at a single point $x = 1$, a set of measure zero, and thus cannot be considered as distinct pdfs).

Lemma 2.1 Properties of identifiable distribution

Lemma 2.1.

If $\mathcal{P} = \{f(y;\theta);\theta \in \Omega\}$ is identifiable and $E\{|\ln f(Y;\theta)|\} < \infty$ for all θ, then

$$Q(\theta) = E_{\theta_0}[\ln f(Y;\theta)] = \int \{\log f(y;\theta)\} f(y;\theta_0) d\mu(y)$$

has a unique maximum at $\theta = \theta_0$.

Kullback-Leibler divergence measure

Shannon-Kolmogorov information inequality

Let $f_0(y)$ and $f_1(y)$ be two density functions (wrt the denomating measure μ). The Kullback-Leibler divergence measure defined by

$$D_{\mathrm{KL}}(f_0 \parallel f_1) \equiv \int \left\{ \ln \frac{f_0(y)}{f_1(y)} \right\} f_0(y) d\mu(y) = E_0 \left[\ln \frac{f_0(Y)}{f_1(Y)} \right]$$

satisfies

$$D_{\mathrm{KL}}(f_0 \parallel f_1) \geq 0,$$

with equality if and only if $P_0 \{ f_0(Y) = f_1(Y) \} = 1$.

Proof

Jensen's inequality

If $g(t)$ is a convex function, then for any random variable X, $g\{E(X)\} \leq E\{g(X)\}$. Furthermore, if $g(t)$ is strictly convex, then $E\{g(X)\} = g\{E(X)\}$ if and only if $P(X = c) = 1$ for some constant c.

Since $\phi(x) = -\ln(x)$ is a strictly convex function of x, we have, using Jensen's inequality,

$$D_{\mathrm{KL}}(f_0 \parallel f_1) = E_0\left\{-\ln \frac{f_1(Y)}{f_0(Y)}\right\} \geq -\ln E_0\left\{\frac{f_1(Y)}{f_0(Y)}\right\} = 0$$

with equality if and only if $P_0\left\{f_0(Y) = f_1(Y)\right\} = 1$.

Proof of Lemma 2.1

1. By Shannon-Kolmogorov information inequality, we can obtain that

$$Q(\theta) \leq Q(\theta_0)$$

for all $\theta \in \Omega$, with equality iff $P_{\theta_0}\{f(Y;\theta) = f(Y;\theta_0)\} = 1$, which means that $Q(\theta)$ has a maximum at $\theta = \theta_0$.

2. To show uniqueness, note that if θ_1 satisfies $Q(\theta_1) = Q(\theta_0)$, it should satisfy

$$P_{\theta_0}\{f(Y;\theta_1) = f(Y;\theta_0)\} = 1,$$

which implies $\theta_1 = \theta_0$, by the identifiability assumption.

Remark

1. Lemma 2.1 simply says that, under identifiability, $Q(\theta) = E_{\theta_0}\{\log f(Y;\theta)\}$ takes the (unique) maximum value at $\theta = \theta_0$.
2. Define
$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\theta)$$
then the MLE $\hat{\theta}$ is the maximizer of $Q_n(\theta)$. Since $Q_n(\theta)$ converges in probability to $Q(\theta) = E_{\theta_0}\{\log f(Y;\theta)\}$ for each θ, can we say that the maximizer of $Q_n(\theta)$ converges to the maximizer of $Q(\theta)$?

Theorem 2.1: (Weak) consistency of MLE

Theorem 2.1

Let

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i, \theta)$$

and $\hat{\theta}_n = \arg\max_{\theta\in\Omega} Q_n(\theta)$. That is

$$Q_n(\hat{\theta}) = \max_{\theta\in\Omega} Q_n(\theta).$$

Assume the following two conditions:

1. Uniform weak convergence:

$$\sup_{\theta\in\Omega} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$$

   for some non-stochastic function $Q(\theta)$

2. Identification: $Q(\theta)$ is uniquely maximized at $\theta = \theta_0$.

Then, $\hat{\theta} \xrightarrow{p} \theta_0$.

Remark

- Convergence in probability:

$$\hat{\theta} \xrightarrow{p} \theta_0 \iff P\left\{|\hat{\theta} - \theta_0| > \epsilon\right\} \to 0 \text{ as } n \to \infty,$$

  for any $\epsilon > 0$.
- If $\mathcal{P} = \{f(y; \theta); \theta \in \Omega\}$ is not identifiable, then $Q(\theta)$ may not have a unique maximum and $\hat{\theta}$ may not converge (in probability) to a single point.
- If the true distribution $f(y)$ does not belong to the class $\mathcal{P} = \{f(y; \theta); \theta \in \Omega\}$, which point does $\hat{\theta}$ converge to?

Other properties of MLE

- Asymptotic normality (Theorem 2.2)
- Asymptotic optimality: MLE achieves Cramer-Rao lower bound
- Wilks' theorem:

$$2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} \xrightarrow{d} \chi_p^2$$

where $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i; \theta)$.

Definition

1. Score function = the gradient function (wrt θ) of the log-density:

$$S(\theta; y) = \frac{\partial \log f(y; \theta)}{\partial \theta}$$

2. (Fisher) information function = negative Hessian matrix of the log-density:

$$\mathrm{I}(\theta; y) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(y; \theta) = -\frac{\partial}{\partial\theta^T} S(\theta; y)$$

3. Observed (Fisher) information: $\mathrm{I}(\hat{\theta}; y)$ where $\hat{\theta}$ is the MLE.
4. Expected (Fisher) information: $\mathcal{I}(\theta) = \mathrm{E}_\theta \left\{ \mathrm{I}(\theta; Y) \right\}$

- The Fisher information $\mathrm{I}(\theta)$ is only meaningful in the neighborhood of $\hat{\theta}$.
- The observed information is always positive. The observed information applies to a single dataset.
- The expected information is meaningful as a function of θ across the admissible values of θ. The expected information is an average quantity over all possible datasets.
- $\mathcal{I}(\hat{\theta}) = \mathrm{I}(\hat{\theta})$ for exponential family.

Theorem 2.3. Bartlett identities

Under the regularity conditions allowing for the exchange of the order of integration and differentiation,

$$\mathrm{E}_{\theta}\left\{S(\theta; Y)\right\} = 0 \quad \text{and} \quad \mathrm{V}_{\theta}\left\{S(\theta; Y)\right\} = \mathcal{I}(\theta).$$

Proof

Remark

- Under some regularity conditions, the MLE $\hat{\theta}$ converges in probability to the true parameter θ_0. (Theorem 2.1)
- Since $\hat{\theta} \xrightarrow{p} \theta_0$, we can apply a Taylor linearization on $n^{-1}\sum_{i=1}^{n} S(\hat{\theta}; y_i) = 0$ to get

$$\hat{\theta} - \theta_0 \cong \{\mathcal{I}(\theta_0)\}^{-1} n^{-1} \sum_{i=1}^{n} S(\theta_0; y_i).$$

  Here, we use the fact that $n^{-1}\sum_{i=1}^{n} I(\theta; y_i)$ converges in probability to $\mathcal{I}(\theta)$.
- Thus, the (asymptotic) variance of MLE is

$$\begin{aligned} V(\hat{\theta}) &\doteq n^{-1} \{\mathcal{I}(\theta_0)\}^{-1} V\{S(\theta_0; Y)\} \{\mathcal{I}(\theta_0)\}^{-1} \\ &= n^{-1} \{\mathcal{I}(\theta_0)\}^{-1}, \end{aligned}$$

  where the last equality follows from Theorem 2.3.

Advanced topic: Information projection

- The *entropy* $H(P)$ of a probability distribution P is defined as

$$H(P) = -\int P(x) \log P(x) d\mu(x)$$

- The KL divergence of P with respect to Q is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int P(x) \log \left\{ \frac{P(x)}{Q(x)} \right\} d\mu(x)$$

- Let $\hat{P}$ be the empirical distribution of the sample. Assume that P belongs to a family $\mathcal{P}$ of distributions (closed, convex). The maximum likelihood estimator of P can be defined as the minimizer of $D_{\mathrm{KL}}(\hat{P} \parallel P)$ for $P \in \mathcal{P}$.

- Let Π be a (non-empty) closed, convex set of distributions
- The information projection of Q onto Π is $P^* \in \Pi$ such that

$$D_{\mathrm{KL}}(P^* \parallel Q) = \min_{P \in \Pi} D_{\mathrm{KL}}(P \parallel Q).$$

- One important family of distributions is a linear family:

$$\mathcal{L}(\alpha) = \left\{ P; \int T_i(x) P(x) d\mu(x) = \alpha_i, i = 1, \cdots, k \right\} \subset \Pi.$$

Note that the linear family is orthogonal to $T_i - \alpha_i$ for $i = 1, \cdots, k$. Moreover, $\mathcal{L}$ is closed, convex (indeed, linear). To show the linearity of $\mathcal{L}$, it suffices to show that $\mathcal{L}$ is closed under the scalar multiplication. Scalar addition is not under consideration since

$$\int P(x) d\mu(x) = 1 \text{ for all } \ P \in \Pi.$$

Information projection (Csiszár and Shields, 2004)

- Since the function $D_{\mathrm{KL}}(P \parallel Q)$ is continuous and strictly convex in P, so that P^* satisfying
$$D_{\mathrm{KL}}(P^* \parallel Q) = \min_{P \in \mathcal{L}(\alpha)} D_{\mathrm{KL}}(P \parallel Q).$$
exists and is unique.
- Moreover, P^*, the information projection of Q onto $\mathcal{L}(\alpha)$ is of the form
$$P^*(x) = Q(x) \frac{\exp\left\{\sum_{i=1}^K \theta_i T_i(x)\right\}}{E_Q\left[\exp\left\{\sum_{i=1}^K \theta_i T_i(x)\right\}\right]}. \tag{1}$$
- Thus, the exponential family of distributions can be derived as the information projection onto the space $\mathcal{L}$ using $Q(\cdot)$ as the baseline distribution.
- Note that there is an one-to-one correspondence between $\theta_1, \ldots, \theta_k$ (canonical parameter) and $\alpha_1, \ldots, \alpha_k$ (natural parameter).

REFERENCES

Csiszár, Imre and P. C. Shields (2004), *Information theory and Statistics: A tutorial*, Now Publishers Inc.