Maximum Likelihood Estimates: Asymptotic Properties

Ching-Kang Ing

Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan

Outline

Consistency of MLEs

Let $L(\boldsymbol{\theta})$ be a likelihood function and $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ denote the true parameter. If

(i) $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$,

(ii) There exists a sufficiently small number $a > 0$ and a nonrandom matrix $\boldsymbol{G}(\boldsymbol{\theta})$ such that for any $\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq a\}$,

$$\sup_{\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0)} \left\| -\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} - \boldsymbol{G}(\boldsymbol{\theta}) \right\| \to 0 \quad \text{in probability,}$$

$$\inf_{\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0)} \lambda_{\min}(\boldsymbol{G}(\boldsymbol{\theta})) > 0,$$

where $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$,

(iii) $\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0) = O_p(\sqrt{n})$,

Norm

- For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|$ denotes the Euclidean norm of $\boldsymbol{v}$.
- For a matrix $\boldsymbol{V}$, $\|\boldsymbol{V}\|$ denotes the spectral norm of $\boldsymbol{V}$.

then the likelihood equation

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{1}$$

has a root $\tilde{\boldsymbol{\theta}}_n$ satisfying

$$\tilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \quad \text{in probability.} \tag{2}$$

Moreover, if

(iv) $P\left(\ell(\boldsymbol{\theta}_0) > \sup_{\boldsymbol{\theta} \in \overline{B}_a^c(\boldsymbol{\theta}_0)} \ell(\boldsymbol{\theta})\right) \to 1,$

then

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) \to \boldsymbol{\theta}_0 \quad \text{in probability.} \tag{3}$$

Proof

- Let $\tilde{\boldsymbol{\theta}}_n$ be the solution of (1) closest to $\boldsymbol{\theta}_0$. We will first show that for any $0 < \epsilon \leq a$ (note that a is defined in (ii)),

$$P\left(\sup_{\boldsymbol{\theta} \in B_\epsilon^*(\boldsymbol{\theta}_0)} \ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}_0)\right) \to 1 \text{ as } n \to \infty, \tag{4}$$

  where $B_\epsilon^*(\boldsymbol{\theta}_0) = \overline{B}_a(\boldsymbol{\theta}_0) - B_\epsilon(\boldsymbol{\theta}_0)$.

- Relation (4) implies

$$P\left(\tilde{\boldsymbol{\theta}}_n \in B_\epsilon(\boldsymbol{\theta}_0)\right) = P(\text{there exists a solution of (1) lying in } B_\epsilon(\boldsymbol{\theta}_0)) \to 1 \text{ as } n \to \infty,$$

  and hence the desired conclusion (2) follows.

- Define

$$E_n(\delta) = \left\{\inf_{\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0)} \lambda_{\min}\left(-\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right) < \delta\right\},$$

  where $\delta > 0$ is arbitrarily small.

Then, on the set $E_n^c(\delta)$ (the complement of $E_n(\delta)$), one has by Taylor's theorem,

$$\begin{aligned}
& \sup_{\boldsymbol{\theta}\in B_\epsilon^*(\boldsymbol{\theta}_0)} \ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) \\
\overset{\text{why?}}{\leq} & \sup_{\boldsymbol{\theta}\in B_\epsilon^*(\boldsymbol{\theta}_0)} \left\| \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\| \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \\
& - \inf_{\boldsymbol{\theta}\in B_\epsilon^*(\boldsymbol{\theta}_0)} \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \left(-\frac{\partial^2 \ell(\boldsymbol{\theta}^*)}{\partial \theta_i \partial \theta_j} \right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \qquad (\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\| \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) \\
\leq & \left\| \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\| a - \frac{\epsilon^2}{2} \inf_{\boldsymbol{\theta}\in \overline{B}_a(\boldsymbol{\theta}_0)} \lambda_{\min} \left(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \\
\leq & \left\| \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\| a - \frac{n\epsilon^2}{2} \delta.
\end{aligned}$$

This and condition (iii) yield

$$P\left(\sup_{\boldsymbol{\theta}\in B_\epsilon^*(\boldsymbol{\theta}_0)} \ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) \geq 0, E_n^c(\delta) \right) \leq P\left(\left\| \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\| a \geq \frac{\epsilon^2}{2} n\delta \right) \to 0 \text{ as } n \to \infty. \tag{5}$$

- In view of (5), (4) follows if we can show that for any $\epsilon_1 > 0$, there exists a sufficiently large N and sufficiently small $\delta > 0$ such that for all $n \geq N$,

$$P(E_n(\delta)) < \epsilon_1 . \tag{6}$$

- To show (6), denote $-\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$ by $\boldsymbol{A}_n(\boldsymbol{\theta})$. Then,

$$\begin{aligned} \lambda_{\min}(\boldsymbol{A}_n(\boldsymbol{\theta})) &\geq \lambda_{\min}(\boldsymbol{A}_n(\boldsymbol{\theta}) - \boldsymbol{G}(\boldsymbol{\theta})) + \lambda_{\min}(\boldsymbol{G}(\boldsymbol{\theta})) \\ &\geq \lambda_{\min}(\boldsymbol{G}(\boldsymbol{\theta})) - \|\boldsymbol{A}_n(\boldsymbol{\theta}) - \boldsymbol{G}(\boldsymbol{\theta})\|, \end{aligned}$$

  yielding

$$\inf_{\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0)} \lambda_{\min}(\boldsymbol{A}_n(\boldsymbol{\theta})) \geq \inf_{\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0)} \lambda_{\min}(\boldsymbol{G}(\boldsymbol{\theta})) - \sup_{\boldsymbol{\theta} \in \overline{B}_a(\boldsymbol{\theta}_0)} \|\boldsymbol{A}_n(\boldsymbol{\theta}) - \boldsymbol{G}(\boldsymbol{\theta})\|.$$

  This and condition (ii) give (6). Thus, (2) is proved.

- Finally, it follows from (4) and (iv) that

$$P\left(\sup_{\boldsymbol{\theta} \in B_\epsilon^c(\boldsymbol{\theta}_0)} \ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}_0) \right) \to 1 \text{ as } n \to \infty,$$

  which implies $\hat{\boldsymbol{\theta}}_n$ converges $\boldsymbol{\theta}_0$ in probability.

Logistic Regression Model

- Recall logistic regression. For $i = 1, \ldots, n$, assume

$$P(Y_i = 1) = 1 - P(Y_i = 0) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_0}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_0}} \equiv p_{i,0}, \qquad \left(\log\left(\frac{p_{i,0}}{1-p_{i,0}}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}_0\right)$$

  where $\boldsymbol{x}_i$ is the explanatory vector and $\boldsymbol{\beta}_0$ is an unknown regression coefficient vector.
- Therefore, the likelihood function is (assuming Y_i are independent)

$$L(\boldsymbol{\beta}) = P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

  where $p_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}/(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}})$ and $\boldsymbol{\beta} \in \boldsymbol{B}$, the parameter space, and the log-likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \\ &= \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}\log(1-p_i) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \sum_{i=1}^{n}\log(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}).\end{aligned}$$

- Moreover,
  $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\boldsymbol{x}_i = \sum_{i=1}^{n} \boldsymbol{x}_i\left(y_i - \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right) \underset{\text{if } \boldsymbol{\beta}=\boldsymbol{\beta}_0}{=} \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - p_{i,0})$.
- Note that $y_i = E(y_i) + (y_i - E(y_i)) = p_{i,0} + \epsilon_i$ where $\epsilon_i \sim (0, p_{i,0}(1-p_{i,0}))$.

Define $\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}\in\boldsymbol{B}}\ell(\boldsymbol{\beta})$. Does $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_0$ in probability?

(0) $\boldsymbol{\beta}_0$ is an interior point of $\boldsymbol{B}$, ← This can be assumed.

(1) $\frac{1}{\sqrt{n}}\frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} = O_p(1)$,
$\frac{1}{\sqrt{n}}\frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i\epsilon_i$ in linear regression model.

(2) $P(\inf_{\boldsymbol{\beta}\in\overline{B}_a(\boldsymbol{\beta}_0)}\lambda_{\min}(-\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j}) > \delta) \to 1$, where δ and a are small constants and $\overline{B}_a(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le a\}$,
$-\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ in linear regression model.

(3) $P(\ell(\boldsymbol{\beta}_0) > \sup_{\boldsymbol{\beta}\in\overline{B}_a^c(\boldsymbol{\beta}_0)}\ell(\boldsymbol{\beta})) \to 1$.
In linear regression model,

$$-\frac{1}{2\sigma^2}\sum_{i=1}^n\epsilon_i^2 = -\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_0)^2 > -\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2,\quad \boldsymbol{\beta}\in\overline{B}_a^c(\boldsymbol{\beta}_0),$$

because

$$\begin{aligned}
-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2 &= -\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_0+\boldsymbol{x}_i'(\boldsymbol{\beta}_0-\boldsymbol{\beta}))^2\\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^n(\epsilon_i+\boldsymbol{x}_i'(\boldsymbol{\beta}_0-\boldsymbol{\beta}))^2\\
&\dot{\sim} -\frac{1}{2\sigma^2}\left\{\sum_{i=1}^n\epsilon_i^2+(\boldsymbol{\beta}_0-\boldsymbol{\beta})'\left(\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i'\right)(\boldsymbol{\beta}_0-\boldsymbol{\beta})\right\}.
\end{aligned}$$

Remark

- For (1), does $\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i(y_i - p_{i,0}) = O_p(1)$? Note that

$$\begin{aligned}
\frac{1}{n}E\left(\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\left(\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)'\right) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\mathrm{Var}(y_i) &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' p_{i,0}(1-p_{i,0}) \\
&= \frac{1}{n}\boldsymbol{X}'\boldsymbol{D}_0\boldsymbol{X} \\
&\le \lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\lambda_{\max}(\boldsymbol{D}_0) \\
&\le \lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\min_{1\le i\le n} p_{i,0}(1-p_{i,0}) \\
&\le \frac{1}{4}\lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right).
\end{aligned}$$

  where $\boldsymbol{D}_0 = \mathrm{diag}(p_{1,0}(1-p_{1,0}), \ldots, p_{n,0}(1-p_{n,0}))$. If $\lambda_{\max}(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})$ is bounded, then (1) holds.

- For (2), since $-\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \dot{\sim} \frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}$ where $\boldsymbol{D} = \mathrm{diag}(p_1(1-p_1), \ldots, p_n(1-p_n))$, we have

$$\inf_{\boldsymbol{\beta}\in\overline{B}_a(\boldsymbol{\beta}_0)} \lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right) \ge \lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) \inf_{\boldsymbol{\beta}\in\overline{B}_a(\boldsymbol{\beta}_0)} \min_{1\le i\le n} p_i(1-p_i).$$

  Therefore, (2) holds if $\lambda_{\min}(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}) > \underline{c}_1 > 0$ and $\inf_{\boldsymbol{\beta}\in\boldsymbol{B}} \min_{1\le i\le n} p_i(1-p_i) > \underline{c}_2 > 0$ for all large n, where $\underline{c}_1$ and $\underline{c}_2$ are some positive constants.

Remark (cont.)

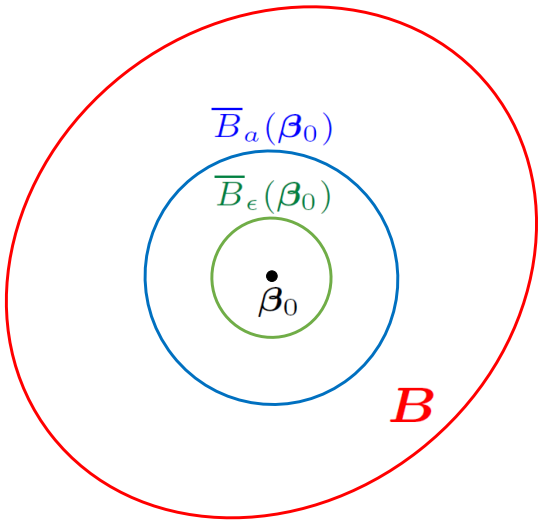
- For (3), it can be shown through (0)–(2) that

$$P\left(\ell(\boldsymbol{\beta}_0) > \sup_{\boldsymbol{\beta}\in(\overline{B}_a(\boldsymbol{\beta}_0)-B_\epsilon(\boldsymbol{\beta}_0))} \ell(\boldsymbol{\beta})\right) \to 1,$$

for any $\epsilon > 0$. Moreover, since

$$\frac{\partial \ell^2(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j} \dot{\sim} -\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X},$$

which is negative definite (assuming $\boldsymbol{X}'\boldsymbol{X}$ is p.d. and $\min_{1\le i\le n} p_i(1-p_i)$ is bounded away from 0 in $\boldsymbol{B}$) and hence $\ell(\boldsymbol{\beta})$ is a convex function, (3) follows.



A central limit theorem for MLEs

- Define

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta} \in \boldsymbol{\Lambda}}{\operatorname{argmin}}\, \ell(\boldsymbol{\eta}),$$

  where $\ell(\boldsymbol{\eta})$ is the log-likelihood function and the true parameter $\boldsymbol{\eta}_0$ is an interior point of the parameter space $\boldsymbol{\Lambda}$.

- I have shown in my previous note that

$$\hat{\boldsymbol{\eta}} \xrightarrow{pr.} \boldsymbol{\eta}_0 \,.$$

- In the following, I will present a CLT for $\hat{\boldsymbol{\eta}}$ using a way far from being rigorous.

- Note that

$$\mathbf{0} = \frac{\partial \ell(\hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} = \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} + \frac{\partial^2 \ell(\boldsymbol{\eta}^*)}{\partial \eta_i \partial \eta_j}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0), \tag{7}$$

  where $\|\boldsymbol{\eta}^* - \boldsymbol{\eta}_0\| \leq \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|$.

- Since it can be shown that $\boldsymbol{\eta}^* \to \boldsymbol{\eta}_0$ in probability (because $\hat{\boldsymbol{\eta}} \to \boldsymbol{\eta}_0$ in probability), we have by (7) and some tedious arguments (which we skip in this note),

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) - \left[-\left(\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\eta}_0)}{\partial \eta_i \partial \eta_j} \right)^{-1} \left(\frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} \right) \right] \xrightarrow{pr.} \mathbf{0}. \tag{8}$$

  ($\boldsymbol{\eta}^*$ in (7) has been replaced here.)

Assume

$$E\left(-\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\eta}_0)}{\partial\eta_i\partial\eta_j}\right) = E\left[\frac{1}{n}\left(\frac{\partial\ell(\boldsymbol{\eta}_0)}{\partial\boldsymbol{\eta}}\right)\left(\frac{\partial\ell(\boldsymbol{\eta}_0)}{\partial\boldsymbol{\eta}}\right)'\right] \underset{n\to\infty}{\longrightarrow} \boldsymbol{I}(\boldsymbol{\eta}_0),$$

and

$$\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\eta}_0)}{\partial\eta_i\partial\eta_j} - E\left(\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\eta}_0)}{\partial\eta_i\partial\eta_j}\right) \xrightarrow{pr.} \mathbf{0}. \text{ (zero matrix)}$$

Then

$$\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\eta}_0)}{\partial\eta_i\partial\eta_j} \xrightarrow{pr.} -\boldsymbol{I}(\boldsymbol{\eta}_0),$$

and hence

$$-\left(\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\eta}_0)}{\partial\eta_i\partial\eta_j}\right)^{-1} \xrightarrow{pr.} \boldsymbol{I}^{-1}(\boldsymbol{\eta}_0). \tag{9}$$

- Under certain regularity conditions, it can be shown that

$$\frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\eta}_0)). \tag{10}$$

- By (8)–(10) and Slutsky's Theorem,

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\eta}_0)). \tag{11}$$

Example I

Consider logistic regression model. Does $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{A})$, for some $\boldsymbol{A}$?

- By $\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \dot{\sim} \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} + \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \dot{\sim} \left(-\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} \right)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}.$$

- Assume $\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} - E\left(\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} \right) \xrightarrow{pr.} 0$ and

$$E\left(-\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} \right) = E\left(\frac{1}{n} \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \left(\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right)' \right) \underset{n \to \infty}{\longrightarrow} \boldsymbol{G} \text{ (p.d.)}.$$

  Then

$$-\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} \dot{\sim} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' p_{i,0}(1 - p_{i,0}) \underset{n \to \infty}{\longrightarrow} \boldsymbol{G}.$$

- Since $\frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i \epsilon_i$, it follows from Lindeberg's CLT that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i \epsilon_i \xrightarrow{d} N(\mathbf{0}, \boldsymbol{G}),$$

  and hence $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{G}^{-1})$.

Example II

- Consider $y_i \overset{\text{independent}}{\sim} N(\boldsymbol{x}_i'\boldsymbol{\beta}_0, e^{\boldsymbol{x}_i'\boldsymbol{\alpha}_0})$, or equivalently,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + \sigma_{i,0}\epsilon_i,$$

  with $\sigma_{i,0} = e^{\frac{1}{2}\boldsymbol{x}_i'\boldsymbol{\alpha}_0}$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$.

- log-likelihood function

$$\ell(\boldsymbol{\eta}) = \ell(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{\alpha} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 e^{-\boldsymbol{x}_i'\boldsymbol{\alpha}}.$$

- To illustrate (11), we start by considering the special case where $\sigma_{i,0} = 1$ or equivalently, $\boldsymbol{\alpha}_0 = \mathbf{0}$. Then,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1).$$

- In addition,

$$\ell(\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2,$$

yielding

$$\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0) = \sum_{i=1}^{n}\boldsymbol{x}_i\epsilon_i.$$

- Note that we've shown before

$$\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i\epsilon_i \xrightarrow{d} N(\mathbf{0}, \boldsymbol{R}),$$

where $\boldsymbol{R} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$ (assuming the limit exists).

- Moreover, we have

$$\frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} = -\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i',$$

hence

$$\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} \to -\boldsymbol{R}.$$

- As a result, we have by (11),

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{R}^{-1}),$$

which coincides with the result that we've obtained for LSE. (Noting that in this example, we assume $\sigma^2 = 1$.)

We now get back to the general case $\boldsymbol{\alpha}_0 \neq \mathbf{0}$.

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\boldsymbol{x}_i}{\sigma_{i,0}} \epsilon_i, \\
\frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\alpha}} &= \frac{1}{2} \sum_{i=1}^{n} \left(\frac{(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_0)^2}{\sigma_{i,0}^2} - 1 \right) \boldsymbol{x}_i, \qquad \left(\frac{(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_0)^2}{\sigma_{i,0}^2} = \epsilon_i^2 \right) \\
\frac{\partial^2 \ell(\boldsymbol{\eta}_0)}{\partial \beta_i \partial \beta_j} &= -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_{i,0}^2}, \\
\frac{\partial^2 \ell(\boldsymbol{\eta}_0)}{\partial \alpha_i \partial \alpha_j} &= -\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \epsilon_i^2, \\
\frac{\partial^2 \ell(\boldsymbol{\eta}_0)}{\partial \alpha_i \partial \beta_j} &= -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_{i,0}^2} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_0).
\end{aligned}
$$

- Assume

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i'}{\sigma_{i,0}^2}\underset{n\to\infty}{\longrightarrow}\boldsymbol{B}\qquad\text{and}\qquad\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\underset{n\to\infty}{\longrightarrow}\boldsymbol{R}.$$

  Then,

$$-\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\eta}_0)}{\partial\eta_i\partial\eta_j}\xrightarrow{pr.}\begin{pmatrix}\boldsymbol{B} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{1}{2}\boldsymbol{R}\end{pmatrix}.$$

- Moreover, it can be shown that

$$\frac{1}{\sqrt{n}}\frac{\partial\ell(\boldsymbol{\eta}_0)}{\partial\boldsymbol{\eta}}=\begin{pmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i}{\sigma_{i,0}}\epsilon_i\\ \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i(\epsilon_i^2-1)\end{pmatrix}\xrightarrow{d}N\left(\boldsymbol{0},\begin{pmatrix}\boldsymbol{B} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{1}{2}\boldsymbol{R}\end{pmatrix}\right).$$

  Consequently,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\xrightarrow{d}N\left(\boldsymbol{0},\begin{pmatrix}\boldsymbol{B}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & 2\boldsymbol{R}^{-1}\end{pmatrix}\right).$$

Question

If $\epsilon_i \overset{\text{i.i.d.}}{\sim} (0,1)$ (without assuming normality; $E(\epsilon_i)=0$ and $Var(\epsilon_i)=1$), $\sqrt{n}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\xrightarrow{d}$?

- In practice, $\boldsymbol{B}$ and $\boldsymbol{R}$ are unknown and we can use

$$\hat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{e^{\boldsymbol{x}_i'\hat{\boldsymbol{\alpha}}}} \qquad \text{and} \qquad \hat{\boldsymbol{R}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$$

  in place of $\boldsymbol{B}$ and $\boldsymbol{R}$.

- Moreover, it can be shown that

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{B}}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{\sqrt{2}}\hat{\boldsymbol{R}}^{\frac{1}{2}} \end{pmatrix} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}),$$

  which allow us to construct confidence regions for $\boldsymbol{\eta}_0$.

Question

If $\epsilon_i \overset{\text{i.i.d.}}{\sim} (0,1)$ (without assuming normality; $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = 1$), please find

a "data-driven matrix" $\hat{\boldsymbol{D}}$

such that

$$\sqrt{n}\hat{\boldsymbol{D}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}).$$

Asymptotics for likelihood ratio tests

- Let $x_1, \ldots, x_n$ be a random sample from the pdf $f_{\boldsymbol{\theta}_0}(\cdot)$, where

$$\boldsymbol{\theta}_0 = \begin{pmatrix} \boldsymbol{\theta}_{0,K(r)} \\ \boldsymbol{\theta}_{0,r} \end{pmatrix} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^K,$$

  $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$, $\boldsymbol{\theta}_{0,r}$ and $\boldsymbol{\theta}_{0,K(r)}$, respectively, are r-dimensional and $(K-r)$-dimensional vectors, and $K > r$.

- For the hypothesis

$$\mathrm{H}_0 : \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta} \quad \text{versus} \quad \mathrm{H}_A : \boldsymbol{\theta}_0 \notin \boldsymbol{\Theta}_0,$$

  the likelihood ratio test rejects H_0 if $\Lambda_n \leq C$, where

$$\Lambda_n = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})} = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \prod_{i=1}^n f_{\boldsymbol{\theta}}(x_i)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \prod_{i=1}^n f_{\boldsymbol{\theta}}(x_i)},$$

  and $0 \leq C \leq 1$ is determined by the distribution of Λ_n and the level of the test.

- In the following, I'll show that

$$-2\log\Lambda_n \xrightarrow{d} \chi^2(\dim(\boldsymbol{\Theta}) - \dim(\boldsymbol{\Theta}_0)).$$

- Without loss of generality, we may assume

$$\mathrm{H}_0 : \boldsymbol{\theta}_{0,r} = \boldsymbol{\theta}^*_{0,r} \quad \text{and} \quad \mathrm{H}_A : \boldsymbol{\theta}_{0,r} \neq \boldsymbol{\theta}^*_{0,r},$$

and hence

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} : (\boldsymbol{\theta}'_{K(r)}, \boldsymbol{\theta}^{*\prime}_{0,r}) \in \boldsymbol{\Theta}\},$$

where $\boldsymbol{\theta}_{K(r)}$ is $(K-r)$-dimensional.
($\boldsymbol{\theta}_{K(r)}$: free parameters; $\boldsymbol{\Theta}$: non-degenerate subset in $\mathbb{R}^K$)

Assume

$$\hat{\boldsymbol{\theta}}^* = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{K(r)} \\ \boldsymbol{\theta}^*_{0,r} \end{pmatrix} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0}{\operatorname{argmax}} L(\boldsymbol{\theta}) \qquad \text{and} \qquad \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} L(\boldsymbol{\theta})$$

are unique. Then, under some regularity conditions mentioned in my previous note,

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_K \end{pmatrix} \xrightarrow{pr.} \boldsymbol{\theta}_0 \equiv \begin{pmatrix} \theta_{0,1} \\ \vdots \\ \theta_{0,K-r} \\ \theta_{0,K-r+1} \\ \vdots \\ \theta_{0,K} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}_{0,K(r)} \\ \boldsymbol{\theta}_{0,r} \end{pmatrix}, \quad \hat{\boldsymbol{\theta}} \xrightarrow[\mathsf{H}_0]{pr.} \boldsymbol{\theta}^*_0 \equiv \begin{pmatrix} \boldsymbol{\theta}_{0,K(r)} \\ \boldsymbol{\theta}^*_{0,r} \end{pmatrix} = \begin{pmatrix} \theta_{0,1} \\ \vdots \\ \theta_{0,K-r} \\ \theta^*_{0,K-r+1} \\ \vdots \\ \theta^*_{0,K} \end{pmatrix},$$

and

$$\hat{\boldsymbol{\theta}}^* = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{0,K(r)} \\ \boldsymbol{\theta}^*_{0,r} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_{K(r),1} \\ \vdots \\ \hat{\theta}_{K(r),K-r} \\ \theta^*_{0,K-r+1} \\ \vdots \\ \theta^*_{0,K} \end{pmatrix} \xrightarrow[\mathsf{H}_0]{pr.} \boldsymbol{\theta}^*_0.$$

Now, by Taylor's theorem, we have under H_0,

$$\ell(\hat{\boldsymbol{\theta}}^*) = \ell(\boldsymbol{\theta}_0^*) + \left(\frac{\partial \ell(\boldsymbol{\theta}_0^*)}{\partial \boldsymbol{\theta}_{K(r)}}\right)' (\hat{\boldsymbol{\theta}}_{K(r)} - \boldsymbol{\theta}_{0,K(r)})$$
$$+ \frac{1}{2}(\hat{\boldsymbol{\theta}}_{K(r)} - \boldsymbol{\theta}_{0,K(r)})' \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0^*)}{\partial \theta_i \partial \theta_j}\right)_{1 \le i,j \le K-r} (\hat{\boldsymbol{\theta}}_{K(r)} - \boldsymbol{\theta}_{0,K(r)})$$
$$+ \frac{1}{3!} \sum_{1 \le i,j,l \le K-r} \left(\frac{\partial^3 \ell(\boldsymbol{\theta}^\triangle)}{\partial \theta_i \partial \theta_j \partial \theta_l}\right) (\hat{\theta}_{K(r),i} - \theta_{0,i})(\hat{\theta}_{K(r),j} - \theta_{0,j})(\hat{\theta}_{K(r),l} - \theta_{0,l}), \quad (12)$$

where $\|\boldsymbol{\theta}^\triangle - \boldsymbol{\theta}_0^*\| \le \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*\|$ and $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{K(r)}} = \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{K-r}}\right)'$, and in general

$$\ell(\hat{\boldsymbol{\theta}}) = \ell(\boldsymbol{\theta}_0) + \left(\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)' (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$
$$+ \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \theta_i \partial \theta_j}\right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$
$$+ \frac{1}{3!} \sum_{1 \le i,j,l \le K} \left(\frac{\partial^3 \ell(\boldsymbol{\theta}^\nabla)}{\partial \theta_i \partial \theta_j \partial \theta_l}\right) (\hat{\theta}_i - \theta_{0,i})(\hat{\theta}_j - \theta_{0,j})(\hat{\theta}_l - \theta_{0,l}), \quad (13)$$

where $\|\boldsymbol{\theta}^\nabla - \boldsymbol{\theta}_0\| \le \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$.

- By the CLT for $\hat{\boldsymbol{\theta}}_{K(r)}$ and $\hat{\boldsymbol{\theta}}$ and assuming there exists small $\delta > 0$ such that

$$\sup_{\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}_0^*)} \left| \frac{1}{n} \frac{\partial^3 \ell(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_l} \right| = O_p(1) \text{ for } 1 \le i, j, l \le K,$$

  it can be shown that under H_0, the third terms on the RHS of (12) and (13) are of order $o_p(1)$.

- Therefore, it remains to show that

$$(\mathrm{I}) - (\mathrm{II}) \xrightarrow[H_0]{d} \chi^2(r), \tag{14}$$

  where

$$\begin{aligned}
(\mathrm{I}) &= 2\left(\frac{\partial \ell(\boldsymbol{\theta}_0^*)}{\partial \boldsymbol{\theta}}\right)' (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*)' \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0^*)}{\partial\theta_i \partial\theta_j}\right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*), \\
(\mathrm{II}) &= 2\left(\frac{\partial \ell(\boldsymbol{\theta}_0^*)}{\partial \boldsymbol{\theta}_{K(r)}}\right)' (\hat{\boldsymbol{\theta}}_{K(r)} - \boldsymbol{\theta}_{0,K(r)}) \\
&\quad + (\hat{\boldsymbol{\theta}}_{K(r)} - \boldsymbol{\theta}_{0,K(r)})' \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0^*)}{\partial\theta_i \partial\theta_j}\right)_{1 \le i,j \le K-r} (\hat{\boldsymbol{\theta}}_{K(r)} - \boldsymbol{\theta}_{0,K(r)}).
\end{aligned}$$

Define

$$\boldsymbol{x}_t = \frac{\partial \log f_{\boldsymbol{\theta}_0^*}(x_t)}{\partial \boldsymbol{\theta}} \qquad \text{and} \qquad \boldsymbol{x}_{t,K(r)} = \frac{\partial \log f_{\boldsymbol{\theta}_0^*}(x_t)}{\partial \boldsymbol{\theta}_{K(r)}}.$$

Then, under the regularity conditions similar to those used to obtain the CLTs for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{K(r)}$, we have

$$\begin{aligned} \text{(I)} &\underset{\mathsf{H}_0}{=} \mathbf{1}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{1} + o_p(1), \\ \text{(II)} &\underset{\mathsf{H}_0}{=} \mathbf{1}'\boldsymbol{X}_{K(r)}(\boldsymbol{X}'_{K(r)}\boldsymbol{X}_{K(r)})^{-1}\boldsymbol{X}'_{K(r)}\mathbf{1} + o_p(1), \end{aligned} \tag{15}$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}'_1 \\ \vdots \\ \boldsymbol{x}'_n \end{pmatrix}, \boldsymbol{X}_{K(r)} = \begin{pmatrix} \boldsymbol{x}'_{1,K(r)} \\ \vdots \\ \boldsymbol{x}'_{n,K(r)} \end{pmatrix}, \quad \text{and} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Let $\boldsymbol{X}_{-1}$ satisfy $(\boldsymbol{X}_{K(r)}, \boldsymbol{X}_{-1}) = \boldsymbol{X}$ and define

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I}_{(K-r)\times(K-r)} & -(\boldsymbol{X}'_{K(r)}\boldsymbol{X}_{K(r)})^{-1}\boldsymbol{X}'_{K(r)}\boldsymbol{X}_{-1} \\ \boldsymbol{0}_{r\times(K-r)} & \boldsymbol{I}_{r\times r} \end{pmatrix}.$$

Then

$$\begin{aligned} & \mathbf{1}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{1} \\ = \quad & \mathbf{1}'\boldsymbol{X}\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{X}'\mathbf{1} \\ \overset{\text{why?}}{=} \quad & \mathbf{1}'\boldsymbol{X}_{K(r)}(\boldsymbol{X}'_{K(r)}\boldsymbol{X}_{K(r)})^{-1}\boldsymbol{X}'_{K(r)}\mathbf{1} \\ & +\mathbf{1}'(\boldsymbol{I}-\boldsymbol{M}_{K(r)})\boldsymbol{X}_{-1}(\boldsymbol{X}'_{-1}(\boldsymbol{I}-\boldsymbol{M}_{K(r)})\boldsymbol{X}_{-1})^{-1}\boldsymbol{X}'_{-1}(\boldsymbol{I}-\boldsymbol{M}_{K(r)})\mathbf{1}, \end{aligned} \tag{16}$$

where

$$\boldsymbol{M}_{K(r)} = \boldsymbol{X}_{K(r)}(\boldsymbol{X}'_{K(r)}\boldsymbol{X}_{K(r)})^{-1}\boldsymbol{X}'_{K(r)}$$

is the orthogonal projection matrix onto $C(\boldsymbol{X}_{K(r)})$.

Since

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow[\mathsf{H}_0]{pr.} \boldsymbol{I}(\boldsymbol{\theta}_0^*) = \begin{pmatrix} \boldsymbol{I}_{11}(\boldsymbol{\theta}_0^*) & \boldsymbol{I}_{12}(\boldsymbol{\theta}_0^*) \\ \boldsymbol{I}_{21}(\boldsymbol{\theta}_0^*) & \boldsymbol{I}_{22}(\boldsymbol{\theta}_0^*) \end{pmatrix} \quad \text{(p.d.)}$$

and

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{x}_t \xrightarrow[\mathsf{H}_0]{d} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0^*))$$

under some regularity conditions, it holds that

$$\begin{aligned}
\frac{1}{n}\boldsymbol{X}'_{-1}(\boldsymbol{I} - \boldsymbol{M}_{K(r)})\boldsymbol{X}_{-1} &\xrightarrow[\mathsf{H}_0]{pr.} \boldsymbol{I}_{22}(\boldsymbol{\theta}_0^*) - \boldsymbol{I}_{21}(\boldsymbol{\theta}_0^*)\boldsymbol{I}_{11}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{I}_{12}(\boldsymbol{\theta}_0^*), \\
\frac{1}{\sqrt{n}}\boldsymbol{X}'_{-1}(\boldsymbol{I} - \boldsymbol{M}_{K(r)})\mathbf{1} &\overset{\text{why?}}{=} (-\boldsymbol{X}'_{-1}\boldsymbol{X}_{K(r)}(\boldsymbol{X}'_{K(r)}\boldsymbol{X}_{K(r)})^{-1}, \boldsymbol{I}_{r\times r})\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{x}_t \\
&\overset{\mathsf{H}_0}{=} (-\boldsymbol{I}_{21}(\boldsymbol{\theta}_0^*)\boldsymbol{I}_{11}^{-1}(\boldsymbol{\theta}_0^*), \boldsymbol{I}_{r\times r})\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{x}_t + o_p(1),
\end{aligned}$$

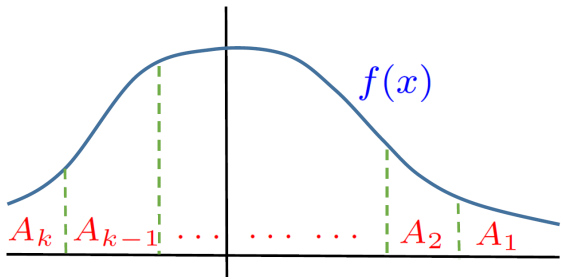
and

$$\mathbf{1}'(\boldsymbol{I} - \boldsymbol{M}_{K(r)})\boldsymbol{X}_{-1}(\boldsymbol{X}'_{-1}(\boldsymbol{I} - \boldsymbol{M}_{K(r)})\boldsymbol{X}_{-1})^{-1}\boldsymbol{X}'_{-1}(\boldsymbol{I} - \boldsymbol{M}_{K(r)})\mathbf{1} \xrightarrow[\mathsf{H}_0]{d} \chi^2(r). \tag{17}$$

Now, the desired conclusion (14) follows from (15)–(17).

Pearson's Chi-Squared Test with unknown parameters

- Assume $x_1, \ldots, x_n$ is a random sample generated from $f(x)$. Consider



- Let $p_i = P(x_1 \in A_i) = \int_{z \in A_i} f(z)\,\mathrm{d}z$, where $\bigcup_{i=1}^{k} A_i = \mathbb{R}$ and $A_i \cap A_j = \emptyset$ if $i \neq j$.
- Define

$$\boldsymbol{y}_i = \begin{pmatrix} I_{\{x_i \in A_1\}} \\ \vdots \\ I_{\{x_i \in A_{k-1}\}} \end{pmatrix} \equiv \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,k-1} \end{pmatrix}.$$

Then, the likelihood function of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is

$$L(p_1, \ldots, p_{k-1}) = \prod_{i=1}^{n} p_1^{y_{i,1}} \cdots p_k^{y_{i,k}} = p_1^{\sum_{i=1}^{n} y_{i,1}} \cdots p_k^{\sum_{i=1}^{n} y_{i,k}} \equiv p_1^{O_1} \cdots p_k^{O_k},$$

where $y_{i,k} = 1 - \sum_{j=1}^{k-1} y_{i,j}$ and $p_k = 1 - \sum_{j=1}^{k-1} p_j$.

- Now Consider the null hypothesis,

$$\mathsf{H}_0 : p_i = \int_{z \in A_i} f_{\boldsymbol{\theta}}(z)\, \mathrm{d}z = p_i(\boldsymbol{\theta}), \quad i = 1, \ldots, k-1,$$

versus the alternative,

$$\mathsf{H}_\mathsf{A} : \mathsf{H}_0 \text{ is wrong},$$

where $f_{\boldsymbol{\theta}}(\cdot)$ is a probability density function indexed by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

- Then, one can use the likelihood ratio statistics

$$-2 \log \Lambda_n \xrightarrow{d} \chi^2(k - 1 - \dim(\boldsymbol{\Theta})),$$

where

$$\Lambda_n = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(p_1(\boldsymbol{\theta}), \cdots, p_{k-1}(\boldsymbol{\theta}))}{\sup_{\sum_{i=1}^k p_i = 1} L(p_1, \ldots, p_{k-1})} = \frac{\prod_{j=1}^k (p_j(\hat{\boldsymbol{\theta}}))^{O_j}}{\prod_{j=1}^k \hat{p}_j^{O_j}},$$

where $\hat{p}_j = O_j / n$, and reject H_0 if $-2 \log \Lambda_n$ is large.

- In the following, I'll show that

$$\sum_{j=1}^k \frac{(O_j - n p_j(\hat{\boldsymbol{\theta}}))^2}{n p_j(\hat{\boldsymbol{\theta}})} \xrightarrow[\mathsf{H}_0]{d} \chi^2(k - 1 - \dim(\boldsymbol{\Theta})).$$

Note that under H_0,

$$
\begin{aligned}
-2\log\Lambda_n &= 2\sum_{j=1}^{k} O_j \log\left(\frac{\hat{p}_j}{p_j(\hat{\boldsymbol{\theta}})} - 1 + 1\right) \\
&\dot{\sim} 2\sum_{j=1}^{k} O_j \left\{\frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{p_j(\hat{\boldsymbol{\theta}})} - \frac{1}{2}\left[\frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{p_j(\hat{\boldsymbol{\theta}})}\right]^2\right\} \\
&= 2\sum_{j=1}^{k} O_j \left\{\frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{\hat{p}_j}\left(\frac{\hat{p}_j}{p_j(\hat{\boldsymbol{\theta}})} - 1 + 1\right) - \frac{1}{2}\left[\frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})}\left(\frac{\hat{p}_j}{p_j(\hat{\boldsymbol{\theta}})} - 1 + 1\right)\right]\right\} \\
&= 2\sum_{j=1}^{k} O_j \frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{\hat{p}_j} + \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})} - \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})}\left(\frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{p_j(\hat{\boldsymbol{\theta}})}\right) \\
&= 2n\sum_{j=1}^{k}(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})) + \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})} - \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})}\left(\frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{p_j(\hat{\boldsymbol{\theta}})}\right) \\
&= \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})} - \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})}\left(\frac{\hat{p}_j - p_j(\hat{\boldsymbol{\theta}})}{p_j(\hat{\boldsymbol{\theta}})}\right) \\
&= \sum_{j=1}^{k} O_j \frac{(\hat{p}_j - p_j(\hat{\boldsymbol{\theta}}))^2}{\hat{p}_j p_j(\hat{\boldsymbol{\theta}})}(1 + o_p(1)) \qquad \left(\text{under } H_0,\ \max_{1\le i\le k}\left|\frac{\hat{p}_i - p_i(\hat{\boldsymbol{\theta}})}{p_i(\hat{\boldsymbol{\theta}})}\right| = o_p(1)\right) \\
&= \sum_{j=1}^{k} \frac{(O_j - np_j(\hat{\boldsymbol{\theta}}))^2}{np_j(\hat{\boldsymbol{\theta}})}(1 + o_p(1)).
\end{aligned}
$$

- Hence, by the following Lamma,

$$-2\log\Lambda_n - \sum_{j=1}^{k}\frac{(O_j - np_j(\hat{\boldsymbol{\theta}}))^2}{np_j(\hat{\boldsymbol{\theta}})} = o_p(1).$$

Lemma: If $A_n = B_n(1+o_p(1))$ and $A_n = O_p(1)$, then $A_n - B_n \xrightarrow{pr.} 0$.

Proof. Note that $A_n - B_n = (B_n - A_n + A_n)o_p(1) = (B_n - A_n)o_p(1) + o_p(1)$, which implies $(A_n - B_n)(1+o_p(1)) = o_p(1)$, and hence $A_n - B_n = \frac{o_p(1)}{1+o_p(1)} = o_p(1)$.

- By Slutsky's Theorem,

$$\sum_{j=1}^{k}\frac{(O_j - np_j(\hat{\boldsymbol{\theta}}))^2}{np_j(\hat{\boldsymbol{\theta}})} \xrightarrow[\mathsf{H}_0]{d} \chi^2(k-1-\dim(\boldsymbol{\Theta})).$$

Therefore, $-2\log\Lambda_n$ and $\sum_{j=1}^{k}\frac{(O_j - np_j(\hat{\boldsymbol{\theta}}))^2}{np_j(\hat{\boldsymbol{\theta}})}$ have the same limiting distribution.

Application

Assume that there are r rows and s columns in the contingency table, and the total number in the table is n. Let O_{ij} be the observed frequency for the i^{th} row and j^{th} column, and p_{ij} be the probability that an observation falls into the i^{th} row and j^{th} column. Define $p_{i\cdot} = \sum_{j=1}^{s} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$. For the hypothesis

$$\mathsf{H}_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \text{ for } i = 1, \ldots, r \text{ and } j = 1, \ldots, s, \qquad \text{versus} \qquad \mathsf{H}_\mathsf{A} : \mathsf{H}_0 \text{ is wrong},$$

the test statistic is

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}} \xrightarrow[\mathsf{H}_0]{d} \chi^2((r-1)(s-1)),$$

where $\hat{p}_{i\cdot} = \sum_{j=1}^{s} O_{ij}/n$ and $\hat{p}_{\cdot j} = \sum_{i=1}^{r} O_{ij}/n$.

Remark

$rs - 1 - (r-1) - (s-1) = rs - 1 - r + 1 - s + 1 = rs - r - s + 1 = (r-1)(s-1)$.

Akaike's information criterion (AIC)

- Let

  $g(\boldsymbol{y}^*) = g(y_1^*, \ldots, y_n^*)$ denote the likelihood function of $\boldsymbol{y}^*$ (true pdf of $\boldsymbol{y}^*$)

  and

  $f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)$ is a family of approximation models indexed by $\boldsymbol{\theta}$ (a family of pdfs indexed by $\boldsymbol{\theta}$).
- The "distance" between $g(\boldsymbol{y}^*)$ and $f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)$ is measured by Kullback-Leibler distance

$$\mathrm{KL}(f_{\boldsymbol{\theta}}, g) = -E_g\left(\log \frac{f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{g(\boldsymbol{y}^*)}\right) = -E_g(\log f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)) + E_g(\log g(\boldsymbol{y}^*)). \tag{18}$$

Remark

$\mathrm{KL}(g, g) = 0$ and $\mathrm{KL}(f_{\boldsymbol{\theta}}, g) \geq 0$, since $-\log t$ is a convex function, leading to

$$E(-\log X) \geq -\log E(X)$$

by Jensen's inequality.

- More specifically, letting

$$X = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{g(\boldsymbol{y}^*)},$$

we have by Jensen's inequality and the convexity of $-\log t$,

$$E_g\left(-\log \frac{f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{g(\boldsymbol{y}^*)}\right) \geq -\log E_g\left(\frac{f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{g(\boldsymbol{y}^*)}\right).$$

- Moreover,

$$E_g\left(\frac{f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{g(\boldsymbol{y}^*)}\right) = \int \frac{f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{g(\boldsymbol{y}^*)} g(\boldsymbol{y}^*)\, \mathrm{d}\boldsymbol{y}^* = \int f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)\, \mathrm{d}\boldsymbol{y}^* = 1$$

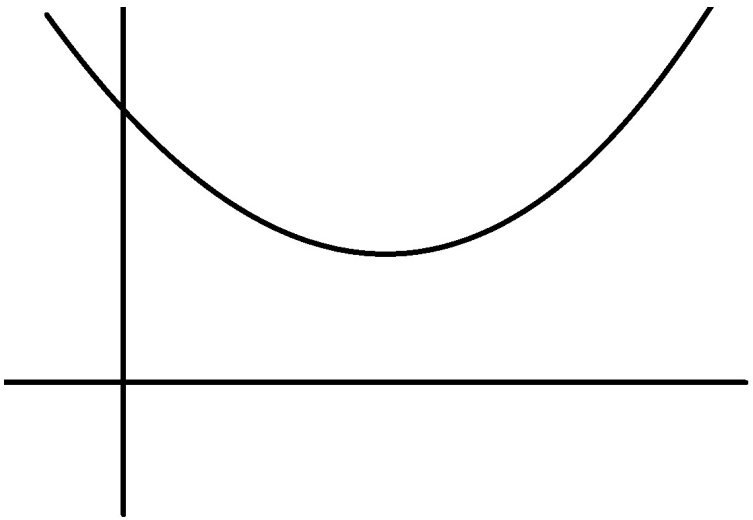
because $f_{\boldsymbol{\theta}}(\boldsymbol{y}^*)$ is a pdf and hence integrates to "1".

Note that Jensen's inequality says that for any convex function $h(\cdot)$, and rnadom variable X,

$$E(h(X)) \geq h(E(X)),$$

whenever the expectations exist. For example,

$$\begin{aligned} E(X^2) &\geq (EX)^2, \\ -E(\log X) &\geq -\log EX, \\ E(e^X) &\geq e^{E(X)}, \\ &\vdots \end{aligned}$$



- In the derivation of AIC, $\boldsymbol{y}^*$ is regarded as the "future", whereas $\boldsymbol{y}$, an independent copy of $\boldsymbol{y}^*$ (namely $\boldsymbol{y}$ and $\boldsymbol{y}^*$ are independent, but have the same distribution.), denotes the observations at hand.
- Since the factor $E_g(\log g(\boldsymbol{y}^*))$ in (18) will be cancelled out when comparing across different approximation models (families), the goal here is to construct an asymptotically unbiased estimate of

$$-E_g(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}^*)), \quad (g\text{: true model; } \hat{\boldsymbol{\theta}}(\boldsymbol{y})\text{: present; } \boldsymbol{y}^*\text{: future})$$

  in which $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ is the MLE of $\boldsymbol{\theta}$ using observations $\boldsymbol{y}$.
- A natural estimate of $-E_g(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}^*))$ is given by

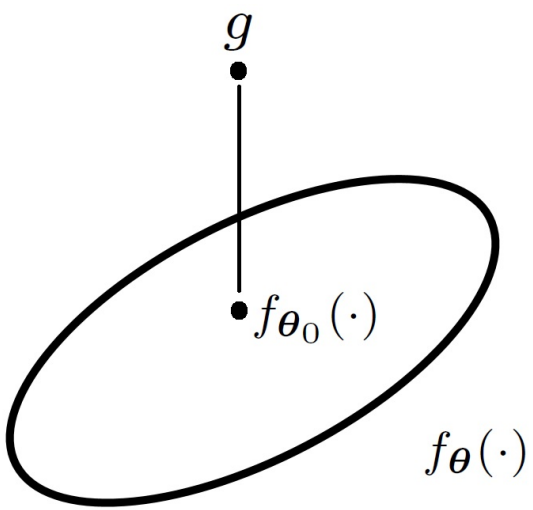
$$-\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}), \quad (g \to \text{unknown} \to \text{dropped here; } \boldsymbol{y}^* \text{ is replaced by } \boldsymbol{y})$$

  the log likelihood function of $\boldsymbol{y}$ based on $f_{\boldsymbol{\theta}}(\cdot)$.
- However, a "bias correction" term is needed to achieve asymptotic unbiasedness.

To find the bias correction term, note first that by Taylor's theorem,

$$\begin{aligned} E_g(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}^*)) &= E_g(\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)) \\ &\quad + E_g\left[\left(\frac{\partial \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial \boldsymbol{\theta}}\right)' (\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)\right] \\ &\quad + \frac{1}{2} E_g\left[(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \left(\frac{\partial^2 \log f_{\boldsymbol{\theta}^*}(\boldsymbol{y}^*)}{\partial \theta_i \partial \theta_j}\right) (\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)\right], \end{aligned} \tag{19}$$

where $\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \operatorname{KL}(f_{\boldsymbol{\theta}}, g)$ and $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0\|$.



Assume $f_{\boldsymbol{\theta}_0} = g$ (the true model is included among the approximation family). (19) becomes

$$\begin{aligned}
& E_g(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}^*)) \\
= \quad & E_{f_{\boldsymbol{\theta}_0}}(\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)) + E_{f_{\boldsymbol{\theta}_0}}\left[\left(\frac{\partial \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial \boldsymbol{\theta}}\right)'(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)\right] \\
& + \frac{1}{2} E_{f_{\boldsymbol{\theta}_0}}\left[(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \frac{\partial^2 \log f_{\boldsymbol{\theta}^*}(\boldsymbol{y}^*)}{\partial \theta_i \partial \theta_j}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)\right] \\
\dot{\sim} \quad & E_{f_{\boldsymbol{\theta}_0}}(\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)) + \frac{1}{2} E_{f_{\boldsymbol{\theta}_0}}\left[(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \frac{\partial^2 \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial \theta_i \partial \theta_j}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)\right] \qquad (20)
\end{aligned}$$

because $\boldsymbol{\theta}^* \dot{\sim} \boldsymbol{\theta}_0$ and

$$\begin{aligned}
& E_{f_{\boldsymbol{\theta}_0}}\left[\left(\frac{\partial \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial \boldsymbol{\theta}}\right)'(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)\right] \\
= \quad & E_{f_{\boldsymbol{\theta}_0}}\left(\frac{\partial \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial \boldsymbol{\theta}}\right)' E(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0) \quad (\boldsymbol{y}^* \text{ and } \boldsymbol{y} \text{ are independent}) \\
= \quad & 0. \quad \left(E_{f_{\boldsymbol{\theta}_0}}\left(\frac{\partial \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial \boldsymbol{\theta}}\right) = \mathbf{0}\right)
\end{aligned}$$

- Moreover, assume

$$\lim_{n\to\infty} \frac{1}{n} E_{f_{\boldsymbol{\theta}_0}} \left(-\frac{\partial^2 \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial\theta_i \partial\theta_j} \right) = \boldsymbol{I}(\boldsymbol{\theta}_0) \text{ exist.} \quad (\boldsymbol{y}^* \text{ can also be replaced by } \boldsymbol{y})$$

- Then, we have shown (heuristically) that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)),$$

and

$$-\frac{1}{n} \frac{\partial^2 \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial\theta_i \partial\theta_j} \xrightarrow{pr.} \boldsymbol{I}(\boldsymbol{\theta}_0),$$

leading to

$$-(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \frac{\partial^2 \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)}{\partial\theta_i \partial\theta_j} (\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2(p),$$

assuming $\boldsymbol{\theta}$ is a p-dimensional vector.

- As a result, (20) becomes

$$\begin{aligned} & E_{f_{\boldsymbol{\theta}_0}}(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}^*)) \\ \dot{\sim} \quad & E_{f_{\boldsymbol{\theta}_0}}(\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}^*)) - \frac{p}{2} \\ = \quad & E_{f_{\boldsymbol{\theta}_0}}(\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y})) - \frac{p}{2} \quad \text{(this replacement won't change anything)} \end{aligned} \tag{21}$$

- By Taylor's expansion again,

$$\begin{aligned}\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y}) &= \log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}) + \left(\frac{\partial \log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right)' (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(\boldsymbol{y})) \\ &\quad + \frac{1}{2}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \left(\frac{\partial^2 \log f_{\check{\boldsymbol{\theta}}}(\boldsymbol{y})}{\partial \theta_i \partial \theta_j}\right) (\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0),\end{aligned}$$

  where $\|\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0\|$.

- Since

$$\frac{\partial \log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad \text{(why?)}$$

  and

$$\begin{aligned}&-(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \frac{\partial^2 \log f_{\check{\boldsymbol{\theta}}}(\boldsymbol{y})}{\partial \theta_i \partial \theta_j} (\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0) \\ \dot{\sim} \quad &-(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0)' \frac{\partial^2 \log f_{\boldsymbol{\theta}_0}(\boldsymbol{y})}{\partial \theta_i \partial \theta_j} (\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2(p),\end{aligned}$$

  it follows that

$$E_{f_{\boldsymbol{\theta}_0}}(\log f_{\boldsymbol{\theta}_0}(\boldsymbol{y})) \dot{\sim} E_{f_{\boldsymbol{\theta}_0}}(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y})) - \frac{p}{2}. \tag{22}$$

- Combining (21) and (22) yields

$$-E_{f_{\boldsymbol{\theta}_0}}(\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}^*)) \dot{\sim} E_{f_{\boldsymbol{\theta}_0}}(-\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})}(\boldsymbol{y}) + p). \tag{23}$$

  This is what we call "asymptotic unbiasedness". (p: This is our bias correction term!!)

- Now, the definition of AIC is given by

$$-2\log f_{\hat{\boldsymbol{\theta}}(\boldsymbol{y})} + 2p,$$

  which is the quantity inside the expectation on the RHS of (23) multiplied by "2".

More on information criteria

Let

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right\}, \tag{24}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$. Then, the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}^2)'$, where

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2, \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix},$$

yielding

$$\begin{aligned} \log f_{\hat{\boldsymbol{\theta}}}(\cdot) &= -\frac{n}{2}(\log(2\pi) + \log\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\hat{\sigma}^2. \end{aligned}$$

($-\frac{n}{2}\log(2\pi) - \frac{n}{2}$: this part is independent of the model)

- Therefore, the essential part of AIC is

$$n \log \hat{\sigma}^2 + 2\text{dim}(\boldsymbol{\beta}), \tag{25}$$

  noting that the number of parameters in model (24) is $\text{dim}(\boldsymbol{\beta}) + 1$ in which "1" will be cancelled out when model comparisons are performed.

- In most practical situations, the distribution of error in model (24) is unknown. Therefore, AIC is in general not obtainable. But we can still "borrow" the AIC obtained in the Gaussian case, namely (25), to do model selection.

Bayesian information criterion (BIC)

- Bayesian information criterion (BIC):

$$-2 \log f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}) + \log n \, \sharp(\text{estimated parameters})$$

  - $\log n$: in contrast to "2" for AIC
  - $\sharp$(estimated parameters): the number of meaning estimated parameters
- BIC in linear regression model with Gaussian error:

$$n \log \hat{\sigma}^2 + \log n \, \dim(\boldsymbol{\beta}).$$

Consistency of BIC in regression models

- Consider a linear model

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \epsilon_t, \quad \epsilon_t \overset{\text{i.i.d.}}{\sim} (0, \sigma^2)$$

where $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tK})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ in which some of β_i may equal 0.

- Define $J_0 = \{\beta_i : 1 \le i \le K, \beta_i \neq 0\}$ and

$$\text{BIC}(J) = n \log \hat{\sigma}_J^2 + \log n \times \sharp(J),$$

where J is a subset of $\mathcal{K} = \{1, \ldots, K\}$, $\sharp(J)$ is the number of elements in J,

$$\hat{\sigma}_J^2 = \frac{1}{n} \sum_{t=1}^{n} (y_t - \boldsymbol{x}_t'(J)\hat{\boldsymbol{\beta}}(J))^2$$

is the residual mean squared error of model J, with $\boldsymbol{x}_t(J) = (x_{ti}, i \in J)$ and

$$\hat{\boldsymbol{\beta}}(J) = \left(\sum_{t=1}^{n} \boldsymbol{x}_t(J)\boldsymbol{x}_t'(J) \right)^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t(J) y_t$$

denoting the LSE of model J.

- Define

$$\hat{J} = \underset{J \subseteq 2^{\mathcal{K}}}{\operatorname{argmin}} \, \mathrm{BIC}(J),$$

  where $2^{\mathcal{K}}$ denotes all subsets of $\mathcal{K}$.

- In the following, I shall show that

$$\lim_{n \to \infty} P(\hat{J} = J_0) = 1, \quad (J_0 : \text{ true model}) \tag{26}$$

  "without" assuming that ϵ_t are Gaussian.

Proof

- If we can show that for $J_0 - J = \emptyset$ and $\sharp(J) > \sharp(J_0)$ (J 嚴格包含 J_0),

$$P(\text{BIC}(J) \leq \text{BIC}(J_0)) \xrightarrow[n\to\infty]{} 0, \tag{27}$$

  and for $J_0 - J \neq \emptyset$ (J 不包含 J_0),

$$P(\text{BIC}(J) \leq \text{BIC}(J_0)) \xrightarrow[n\to\infty]{} 0, \tag{28}$$

  then combining (27) and (28) leads to the desired conclusion (26).
- To show (27), note first that

$$\{\text{BIC}(J) \leq \text{BIC}(J_0)\} = \{n(\log \hat{\sigma}^2_{J_0} - \log \hat{\sigma}^2_J) \geq \log n(\sharp(J) - \sharp(J_0))\}. \tag{29}$$

Moreover, since

$$\begin{aligned} n(\log\hat{\sigma}^2_{J_0} - \log\hat{\sigma}^2_J) = n\left(\log\left(1 + \frac{\hat{\sigma}^2_{J_0} - \hat{\sigma}^2_J}{\hat{\sigma}^2_J}\right)\right) &\dot{\sim} \frac{n}{\sigma^2}(\hat{\sigma}^2_{J_0} - \hat{\sigma}^2_J) \\ &= \frac{1}{\sigma^2}\boldsymbol{\epsilon}'(\boldsymbol{M}_J - \boldsymbol{M}_{J_0})\boldsymbol{\epsilon} \\ &\leq \frac{1}{\sigma^2}\boldsymbol{\epsilon}'\boldsymbol{M}_J\boldsymbol{\epsilon}, \end{aligned}$$

and since

$$\frac{1}{\sigma^2}E(\boldsymbol{\epsilon}'\boldsymbol{M}_J\boldsymbol{\epsilon}) = \sharp(J) < \infty,$$

we have

$$n(\log\hat{\sigma}^2_{J_0} - \log\hat{\sigma}^2_J) = O_p(1). \quad \text{(why?)} \tag{30}$$

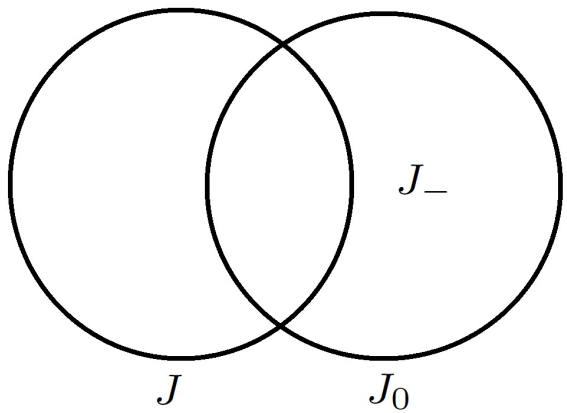
- $\dot{\sim}$ holds due to "correctness" of those two models ($\hat{\sigma}^2_{J_0} \xrightarrow{pr.} \sigma^2$ and $\hat{\sigma}^2_J \xrightarrow{pr.} \sigma^2$) and by Taylor's expansion.
- $\boldsymbol{M}_J$ and $\boldsymbol{M}_{J_0}$ are orthogonal projection matrices for the column spaces of

$$\boldsymbol{X}_J = \begin{pmatrix} \boldsymbol{x}'_1(J) \\ \vdots \\ \boldsymbol{x}'_n(J) \end{pmatrix} \quad \text{and} \quad \boldsymbol{X}_{J_0} = \begin{pmatrix} \boldsymbol{x}'_1(J_0) \\ \vdots \\ \boldsymbol{x}'_n(J_0) \end{pmatrix}.$$

- Now, (27) follows directly from (29), (30) and $\log n \to \infty$.
- To show (28), define

$$J^* = J \cup J_0, \quad \text{and} \quad J_- = J_0 - J,$$

noting that $J^* = J \cup J_-$ and $J \cap J_- = \emptyset$.



- In view of (27), (28) is guaranteed by

$$P(\text{BIC}(J) \leq \text{BIC}(J^*)) \xrightarrow[n\to\infty]{} 0. \tag{31}$$

- It is clear that

$$\begin{aligned} &\{\text{BIC}(J) \leq \text{BIC}(J^*)\} = \{n(\log \hat{\sigma}_J^2 - \log \hat{\sigma}_{J^*}^2) \leq \log n(\sharp(J^*) - \sharp(J))\}, \\ &\text{and} \quad \log \hat{\sigma}_{J^*}^2 \xrightarrow{pr.} \log \sigma^2. \end{aligned} \tag{32}$$

(Since J^* is a correct model, namely, a model including J_0 as a subset model)

In addition, we have

$$\boldsymbol{y} = \boldsymbol{X}_{J^*}\boldsymbol{\beta}_{J^*} + \boldsymbol{\epsilon}, \quad \text{(why?)}$$

where $\boldsymbol{\beta}_{J^*} = (\beta_j, j \in J^*)$, and hence

$$\begin{aligned}
\hat{\sigma}_J^2 &= \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{M}_J)\boldsymbol{y} \\
&= \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{M}_J)(\boldsymbol{I} - \boldsymbol{M}_J)(\boldsymbol{X}_J\boldsymbol{\beta}_J + \boldsymbol{X}_{J_}\boldsymbol{\beta}_{J_} + \boldsymbol{\epsilon}) \\
&= \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{M}_J)(\boldsymbol{I} - \boldsymbol{M}_J)(\boldsymbol{X}_{J_}\boldsymbol{\beta}_{J_} + \boldsymbol{\epsilon}) \\
&= \frac{1}{n}(\boldsymbol{X}_{J_}\boldsymbol{\beta}_{J_} + \boldsymbol{\epsilon})'(\boldsymbol{I} - \boldsymbol{M}_J)(\boldsymbol{X}_{J_}\boldsymbol{\beta}_{J_} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta}_{J_}'\left(\frac{1}{n}\boldsymbol{X}_{J_}'(\boldsymbol{I} - \boldsymbol{M}_J)\boldsymbol{X}_{J_}\right)\boldsymbol{\beta}_{J_} + \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{M}_J\boldsymbol{\epsilon} + \frac{2}{n}\boldsymbol{\beta}_{J_}'\boldsymbol{X}_{J_}'(\boldsymbol{I} - \boldsymbol{M}_J)\boldsymbol{\epsilon} \\
&\equiv \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)}. \qquad (33)
\end{aligned}$$

Define $\hat{\boldsymbol{B}} = (\boldsymbol{X}'_J \boldsymbol{X}_J)^{-1} \boldsymbol{X}'_J \boldsymbol{X}_{J_-}$. Then

$$\begin{aligned} & \boldsymbol{X}'_{J_-}(\boldsymbol{I} - \boldsymbol{M}_J)\boldsymbol{X}_{J_-} \quad (\boldsymbol{I}\text{: this is the identity matrix of dimension } \sharp(J_-)) \\ = \quad & (\boldsymbol{I}, -\hat{\boldsymbol{B}}') \begin{pmatrix} \boldsymbol{X}'_{J_-} \\ \boldsymbol{X}_J \end{pmatrix} (\boldsymbol{X}_{J_-}, \boldsymbol{X}_J) \begin{pmatrix} \boldsymbol{I} \\ -\hat{\boldsymbol{B}} \end{pmatrix} \\ = \quad & (\boldsymbol{I}, -\hat{\boldsymbol{B}}') \boldsymbol{X}'_{J^*} \boldsymbol{X}_{J^*} \begin{pmatrix} \boldsymbol{I} \\ -\hat{\boldsymbol{B}} \end{pmatrix}, \end{aligned}$$

and hence

$$\begin{aligned} \lambda_{\min}(\boldsymbol{X}'_{J_-}(\boldsymbol{I} - \boldsymbol{M}_J)\boldsymbol{X}_{J_-}) \quad & \geq \quad \lambda_{\min}(\boldsymbol{X}'_{J^*}\boldsymbol{X}_{J^*}) \lambda_{\min}\left((\boldsymbol{I}, -\hat{\boldsymbol{B}}') \begin{pmatrix} \boldsymbol{I} \\ -\hat{\boldsymbol{B}} \end{pmatrix}\right) \\ & = \quad \lambda_{\min}(\boldsymbol{X}'_{J^*}\boldsymbol{X}_{J^*}) \lambda_{\min}(\boldsymbol{I} + \hat{\boldsymbol{B}}'\hat{\boldsymbol{B}}) \\ & \geq \quad \lambda_{\min}(\boldsymbol{X}'_{J^*}\boldsymbol{X}_{J^*}). \end{aligned}$$

Assume

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow[n\to\infty]{} \boldsymbol{R} \text{ with } \lambda_{\min}(\boldsymbol{R}) > 0. \tag{34}$$

Then,

$$\begin{aligned}
\lambda_{\min}\left(\frac{\boldsymbol{X}'_{J_-}(\boldsymbol{I}-\boldsymbol{M}_J)\boldsymbol{X}_{J_-}}{n}\right) &\geq \lambda_{\min}\left(\frac{\boldsymbol{X}'_{J^*}\boldsymbol{X}_{J^*}}{n}\right) \\
&\overset{\text{this is obvious}}{\geq} \lambda_{\min}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right) \\
&\xrightarrow{\text{by assumption}} \lambda_{\min}(\boldsymbol{R}) > 0. \qquad (35)
\end{aligned}$$

- Therefore,

$$\text{(I)} \xrightarrow{\text{by (34)}} V(J_-) \geq \|\boldsymbol{\beta}_{J_-}\|^2 \lim_{n\to\infty} \lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'_{J_-}(\boldsymbol{I}-\boldsymbol{M}_J)\boldsymbol{X}_{J_-}\right) > 0, \tag{36}$$

where

$$V(J_-) = \boldsymbol{\beta}'_{J_-}\left(\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}'_{J_-}(\boldsymbol{I}-\boldsymbol{M}_J)\boldsymbol{X}_{J_-}\right)\boldsymbol{\beta}_{J_-}.$$

  - $\|\boldsymbol{\beta}_{J_-}\|^2 > 0$ because $\boldsymbol{\beta}_{J_-}$ contains non-zero coefficients
  - by (35), $\lim_{n\to\infty}\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'_{J_-}(\boldsymbol{I}-\boldsymbol{M}_J)\boldsymbol{X}_{J_-}\right) > 0$

- Moreover, it is easy to see that

$$\frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \to \sigma^2 \text{ in probability,} \tag{37}$$

$$\frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{M}_J\boldsymbol{\epsilon} \to 0 \text{ in probability,} \tag{38}$$

and

$$E(\text{IV})^2 \leq \frac{4\sigma^2}{n}\|\boldsymbol{\beta}_{J_-}\|^2\lambda_{\max}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right) \xrightarrow{\text{by (34)}} 0. \tag{39}$$

- Now, it follows from (33) and (36)–(39) that

$$\log \hat{\sigma}_J^2 \xrightarrow{pr.} \log(\sigma^2 + V(J_-)),$$

which, together with the second equation of (32), yields

$$\log \hat{\sigma}_J^2 - \log \hat{\sigma}_{J^*}^2 \xrightarrow{pr.} \log(\sigma^2 + V(J_-)) - \log \sigma^2 > 0. \tag{40}$$

- Equation (40) and the first equation of (32) imply

$$\begin{aligned} & P(\mathsf{BIC}(J) \leq \mathsf{BIC}(J^*)) \\ = \quad & P\left(\log \hat{\sigma}_J^2 - \log \hat{\sigma}_{J^*}^2 \leq \frac{\log n}{n}(\sharp(J^*) - \sharp(J))\right) \xrightarrow[n\to\infty]{} 0, \quad \left(\frac{\log n}{n} \to 0\right) \end{aligned}$$

which is (28). Thus, the proof is complete.

A Newton-Raphson method and its asymptotics

- Define

$$\hat{\boldsymbol{\eta}}^{\text{New}} = \tilde{\boldsymbol{\eta}} - \left(\frac{\partial^2 \ell(\tilde{\boldsymbol{\eta}})}{\partial \eta_i \partial \eta_j} \right)^{-1} \frac{\partial \ell(\tilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}},$$

  where $\ell(\boldsymbol{\eta})$ is the log-likelihood function and $\tilde{\boldsymbol{\eta}}$ is an initial estimate of the true parameter $\boldsymbol{\eta}_0$.

- In the following, I'll show that if

$$\|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = o_p(n^{-\frac{1}{4}}), \tag{41}$$

  then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}^{\text{New}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\eta}_0)),$$

  where $\boldsymbol{I}(\boldsymbol{\eta}_0)$ is the Fisher information matrix.

Sketch

$$\hat{\boldsymbol{\eta}}^{\text{New}} = \tilde{\boldsymbol{\eta}} - \left(\frac{\partial^2 \ell(\tilde{\boldsymbol{\eta}})}{\partial \eta_i \partial \eta_j}\right)^{-1} \left(\frac{\partial \ell(\tilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} - \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}\right) - \left(\frac{\partial^2 \ell(\tilde{\boldsymbol{\eta}})}{\partial \eta_i \partial \eta_j}\right)^{-1} \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}$$

By Taylor's theorem,

$$\frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} = \frac{\partial \ell(\tilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} + \frac{\partial^2 \ell(\tilde{\boldsymbol{\eta}})}{\partial \eta_i \partial \eta_j}(\boldsymbol{\eta}_0 - \tilde{\boldsymbol{\eta}}) + \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{pmatrix}, \tag{42}$$

where

$$\gamma_k = \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^3 \ell(\boldsymbol{\eta}_k^*)}{\partial \eta_k \partial \eta_i \partial \eta_j} (\eta_{0,i} - \tilde{\eta}_i)(\eta_{0,j} - \tilde{\eta}_j)$$

with

$$\begin{pmatrix} \tilde{\eta}_1 \\ \vdots \\ \tilde{\eta}_p \end{pmatrix} = \tilde{\boldsymbol{\eta}}, \begin{pmatrix} \eta_{0,1} \\ \vdots \\ \eta_{0,p} \end{pmatrix} = \boldsymbol{\eta}_0 \quad \text{and} \quad \|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| \geq \|\boldsymbol{\eta}_k^* - \boldsymbol{\eta}_0\|, k = 1, \ldots, p.$$

(assuming there are p parameters)

- Assume Condition (ii) in the proof of the consistency of the MLE holds, and there exists $\delta > 0$ such that

$$\sup_{\boldsymbol{\eta} \in \mathsf{B}_\delta(\boldsymbol{\eta}_0)} \frac{1}{n} \frac{\partial^3 \ell(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_i \partial \eta_j} = O_p(1).$$

- Then, by (41) and (42), one gets

$$\hat{\boldsymbol{\eta}}^{\mathsf{New}} = \tilde{\boldsymbol{\eta}} - (\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + o_p(n^{-\frac{1}{2}}) - \left(\frac{\partial^2 \ell(\tilde{\boldsymbol{\eta}})}{\partial \eta_i \partial \eta_j} \right)^{-1} \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}},$$

and hence

$$\sqrt{n}(\hat{\boldsymbol{\eta}}^{\mathsf{New}} - \boldsymbol{\eta}_0) = o_p(1) - \left(\frac{1}{n} \frac{\partial^2 \ell(\tilde{\boldsymbol{\eta}})}{\partial \eta_i \partial \eta_j} \right)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}. \tag{43}$$

By (41) and an argument similar to that used to prove (42), the RHS of (43) becomes

$$\left(-\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\eta}_0)}{\partial \eta_i \partial \eta_j}\right)^{-1} \frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} + o_p(1),$$

and hence the desired conclusion follows from

$$\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\eta}_0)),$$

and

$$-\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\eta}_0)}{\partial \eta_i \partial \eta_j} \xrightarrow{pr.} \boldsymbol{I}(\boldsymbol{\eta}_0).$$

Question

Please show that if (41) is replaced by

$$\|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = O_p(n^{-\frac{1}{q}})$$

for some $q \geq 4$, then

$$\|\hat{\boldsymbol{\eta}}^{\text{New}} - \boldsymbol{\eta}_0\| = O_p(n^{-\frac{2}{q}}).$$

Example

- Assume

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + a_i, \quad i = 1, \ldots, n,$$

where $a_i = \sigma_i \epsilon_i$, $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$, $\sigma_i^2 = \exp\{\boldsymbol{x}_i'\boldsymbol{\alpha}_0\}$, and $\boldsymbol{x}_i$ is a $(p+1)$-dimensional explanatory vector with 1 as its first component (namely the intercept term is included).

- Assume also $1 \le \|\boldsymbol{x}_i\| \le M$ for some $1 < M < \infty$ and

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \to \boldsymbol{R} \text{ (p.d.)}, \text{ where } \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}.$$

Then, it is not difficult to see that the LSE of $\boldsymbol{\beta}_0$,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \text{ where } \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

satisfies

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(n^{-\frac{1}{2}}). \tag{44}$$

- In the following, I'll construct an initial estimate of

$$\boldsymbol{\theta}_0 = \begin{pmatrix} u + \alpha_{0,0} \\ \alpha_{0,1} \\ \vdots \\ \alpha_{0,p} \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} \alpha_{0,0} \\ \vdots \\ \alpha_{0,p} \end{pmatrix} = \boldsymbol{\alpha}_0 \quad \text{and} \quad u = E(\log \epsilon_i^2).$$

- Define

$$\tilde{a}_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} = a_i - \boldsymbol{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

and

$$\hat{a}_i = \begin{cases} \tilde{a}_i, & \text{if } |\tilde{a}_i| > n^{-\xi}, \\ n^{-\xi}, & \text{if } |\tilde{a}_i| \leq n^{-\xi}, \end{cases}$$

where $0 < \xi < \frac{1}{2} - \theta_1$. Let $A_n = \{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq n^{-\frac{1}{2}+\theta_1}\}$, $\theta_1 > 0$ is a small positive number.

- By (44), it holds that

$$\lim_{n \to \infty} P(A_n) = 1. \tag{45}$$

- Let

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}, \quad \text{where} \quad \boldsymbol{z} = \begin{pmatrix} \log \hat{a}_1^2 \\ \vdots \\ \log \hat{a}_n^2 \end{pmatrix}.$$

- Since

$$\log \hat{a}_i^2 = \boldsymbol{x}_i'\boldsymbol{\theta}_0 + (\log \hat{a}_i^2 - \log a_i^2) + (\log \epsilon_i^2 - u),$$

we obtain

$$\begin{aligned} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(\log \epsilon_i^2 - u) \\ &\quad + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(\log \hat{a}_i^2 - \log a_i^2) I_{|a_i| < n^{-\theta}} \\ &\quad + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(\log \hat{a}_i^2 - \log a_i^2) I_{|a_i| > n^{-\theta}} \\ &\equiv (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}), \end{aligned}$$

where $\xi < \theta < \frac{1}{2} - \theta_1$.

- It is easy to see that

$$(\mathrm{I}) = O_p(n^{-\frac{1}{2}}). \tag{46}$$

- Denote $\boldsymbol{x}_i$ by $\begin{pmatrix} x_{i1} \\ \vdots \\ x_{i,p+1} \end{pmatrix}$. Then, we have for $1 \le j \le p+1$,

$$E\left\{ \frac{1}{n} \sum_{i=1}^{n} \left| x_{ij} (\log \hat{a}_i^2 - \log a_i^2) \right| I_{|a_i| < n^{-\theta}} I_{A_n} \right\}$$

$$\overset{\text{why?}}{\le} \quad C \log n \, P(a_i^2 < n^{-2\theta}) + C P^{\frac{1}{2}}(a_i^2 < n^{-2\theta}) \quad (C : \text{ some positive constant})$$

$$\overset{\text{why?}}{\le} \quad C^* n^{-\frac{\theta}{2}}. \quad (C^* : \text{ some positive constant})$$

This, the positive definiteness of $\boldsymbol{R}$, and (45) yield

$$(\mathrm{II}) = O_p(n^{-\frac{\theta}{2}}). \tag{47}$$

Moreover,

$$E\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}(\log \hat{a}_i^2 - \log a_i^2) I_{|a_i|>n^{-\theta}} I_{A_n}\right|$$

$$\overset{\text{why?}}{\leq} \quad E\left|\frac{C}{n}\sum_{i=1}^{n} |x_{ij}| \frac{|\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|}{|a_i|} I_{|a_i|>n^{-\theta}} I_{A_n}\right| \quad (C: \text{ some positive constant})$$

$$\overset{\text{why?}}{\leq} \quad C^* n^{\theta - \frac{1}{2} + \theta_1}. \quad (C^*: \text{ some positive constant})$$

This, the positive definiteness of $\boldsymbol{R}$, and (45) imply

$$(\text{III}) = O_p(n^{-\frac{1}{2}+\theta_1+\theta}),$$

which, together with (46) and (47), gives

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(n^{-\zeta}) \tag{48}$$

with $\zeta = \min\{\frac{1}{2} - \theta_1 - \theta, \frac{\theta}{2}\}$.

Remark

- Recall that the location-dispersion model is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \exp\left\{\frac{1}{2}\boldsymbol{x}_i'\boldsymbol{\alpha}\right\}\epsilon_i, \quad i = 1, \ldots, n,$$

  with $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and the log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{\alpha} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 e^{-\boldsymbol{x}_i'\boldsymbol{\alpha}}.$$

- The function $-\ell(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is not jointly convex in coefficients $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. This also reveals the importance of finding a good initial estimate of the true parameters.

Model

$$\begin{aligned} y_t &= \beta x_t + \eta_t, \quad t = 1, 2, \ldots, n, \\ \eta_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \exp\{\alpha x_t\} \end{aligned}$$

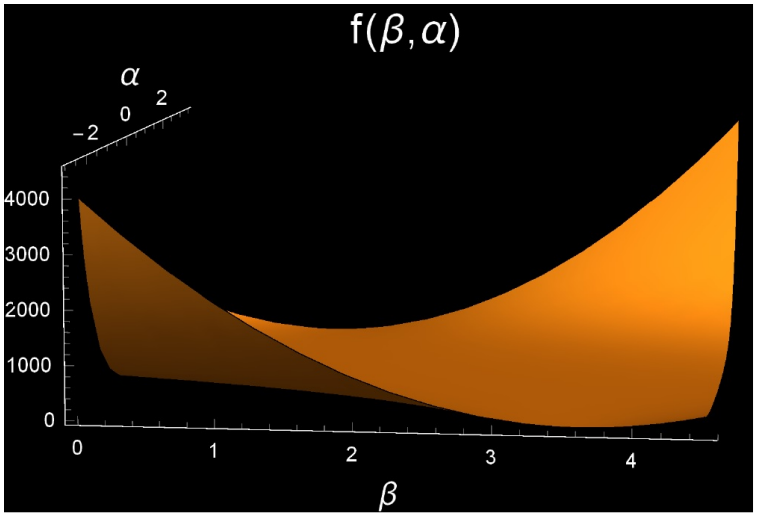
Function

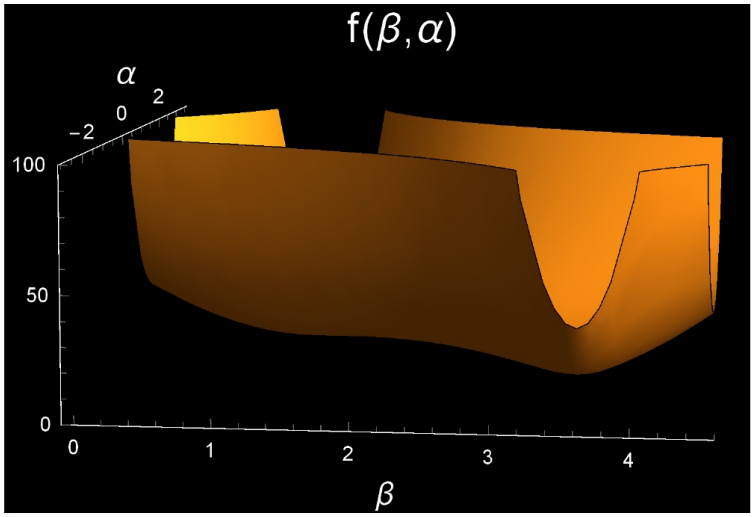
$$f(\beta, \alpha) = \sum_{i=1}^{n} (\alpha x_i) + \sum_{j=1}^{n} \frac{(y_j - \beta x_j)^2}{\exp\{\alpha x_j\}}$$

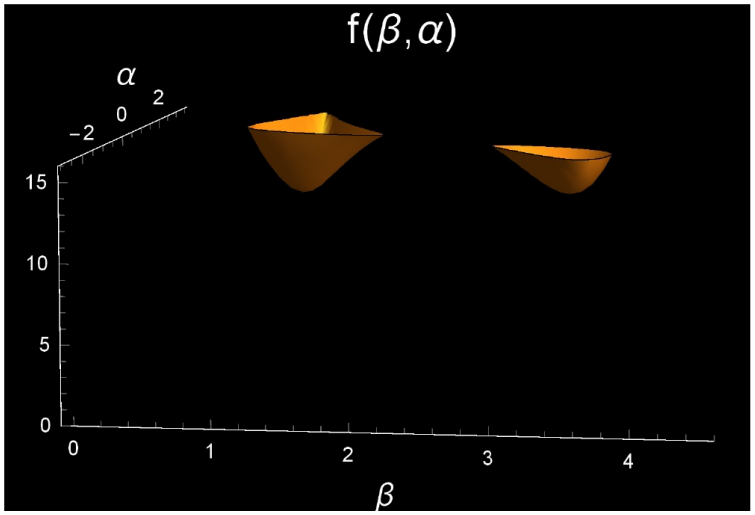
Setting

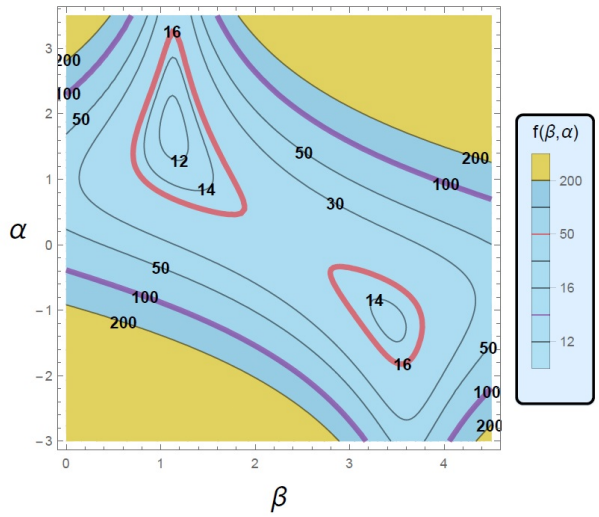
- $x_t \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1)$
- $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1)$
- $\beta = 1$ and $\alpha = 2$
- $n = 10$

3D plot and contour plot









Question

Define

$$\tilde{\alpha}_0 = \log \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{a}_i^2}{e^{x_{i1}\hat{\theta}_1 + \cdots + x_{ip}\hat{\theta}_p}},$$

where

$$\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \end{pmatrix} = \hat{\boldsymbol{\theta}}.$$

a. Please show that $\tilde{\alpha}_0 - \alpha_{0,0} = O_p(n^{-\delta})$ for some $\delta > 0$.

b. Can you obtain similar results when ϵ_i's are non-Gaussian?

Asymptotics for the MLEs of Weibull distribution parameters

- Weibull distribution

$$f(x) = \frac{\alpha_0}{\lambda_0^{\alpha_0}} x^{\alpha_0 - 1} \exp\left\{ -\left(\frac{x}{\lambda_0} \right)^{\alpha_0} \right\}, \tag{49}$$

  where $x > 0$, $\lambda_0 > 0$ is called the scale parameter, and $\alpha_0 > 0$ is called the shape parameter.

- The density function (49) can be rewritten as

$$f(x) = \frac{\alpha_0}{\eta_0} x^{\alpha_0 - 1} \exp\left\{ -\frac{x^{\alpha_0}}{\eta_0} \right\}, \quad x > 0, \alpha_0 > 0, \eta_0 > 0,$$

  in which $\eta_0 = \lambda_0^{\alpha_0}$.

Define the average log-likelihood function

$$\frac{1}{n}\ell(\alpha,\eta) = \log\frac{\alpha}{\eta} + (\alpha-1)\frac{1}{n}\sum_{i=1}^{n}\log x_i - \frac{1}{\eta}\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}.$$

Then

$$\begin{aligned}
&\frac{1}{n}\{\ell(\alpha_0,\eta_0) - \ell(\alpha,\eta)\} \\
= \quad &\log\frac{\alpha_0}{\eta_0} - \log\frac{\alpha}{\eta} + (\alpha_0-\alpha)\frac{1}{n}\sum_{i=1}^{n}(\log x_i - E(\log x_i)) \\
&+(\alpha_0-\alpha)E(\log x_i) - \frac{1}{n}\sum_{i=1}^{n}\frac{x_i^{\alpha_0} - Ex_i^{\alpha_0}}{\eta_0} \\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{x_i^{\alpha} - Ex_i^{\alpha}}{\eta} + \frac{Ex_i^{\alpha}}{\eta} - \frac{Ex_i^{\alpha_0}}{\eta_0}. \qquad (50)
\end{aligned}$$

- Note that

$$\begin{aligned} E(\log x_i) &= \frac{1}{\alpha_0}\int_0^\infty \left(\log \frac{x_i^{\alpha_0}}{\eta_0} + \log \eta_0\right) \frac{\alpha_0}{\eta_0} x_i^{\alpha_0 - 1} e^{-\frac{x_i^{\alpha_0}}{\eta_0}} \, \mathrm{d}x_i \\ &= \frac{1}{\alpha_0}\int_0^\infty \log u \, e^{-u} \, \mathrm{d}u + \frac{1}{\alpha_0}\log \eta_0 \\ &= \frac{1}{\alpha_0}(-\gamma + \log \eta_0), \end{aligned}$$

  where $\gamma = \lim_{n\to\infty}\left(\log n - \sum_{i=1}^n \frac{1}{i}\right) \doteq 0.5772$. ($\gamma$: Euler-Mascheroni constant)

- In addition, it holds that for $m > -\alpha_0$,

$$E(x_i^m) = \eta_0^{\frac{m}{\alpha_0}} \Gamma\left(\frac{m}{\alpha_0} + 1\right),$$

  and by the Weierstrass product for the Γ function,

$$\frac{\Gamma'(z+1)}{\Gamma(z)} = -\gamma + \sum_{i\geq 1}\frac{1}{i} - \frac{1}{i+z}, \; z > 0. \tag{51}$$

- (51) implies

$$\frac{\mathrm{d}}{\mathrm{d}z}\frac{\Gamma'(z+1)}{\Gamma(z)} = \frac{\Gamma(z+1)\Gamma''(z+1) - (\Gamma'(z+1))^2}{\Gamma^2(z)} = \sum_{i\geq 1}\frac{1}{(i+z)^2} > 0,\ z > 0,$$

and hence

$$\frac{\Gamma''(z+1)}{\Gamma(z)} \geq \left(\frac{\Gamma'(z+1)}{\Gamma(z)}\right)^2,\ z > 0. \tag{52}$$

- Now, for $(\alpha, \eta) \in [\delta_1, M_1] \times [\delta_2, M_2]$, where $0 < \delta_1, \delta_2 < \infty$ are small constants and $0 < M_1, M_2 < \infty$ are large constants,

$$\begin{aligned}
& \frac{1}{n}\{\ell(\alpha_0, \eta_0) - \ell(\alpha, \eta)\} \\
= \quad & (\alpha_0 - \alpha)\frac{1}{n}\sum_{i=1}^{n}\{\log x_i - E(\log x_i)\} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^{\alpha_0}}{\eta_0} - 1\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{x_i^{\alpha}}{\eta} - \frac{Ex_i^{\alpha}}{\eta}\right\} + g(\alpha_0, \eta_0) - g(\alpha, \eta) \\
\equiv \quad & (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}) + (\mathrm{IV}),
\end{aligned} \tag{53}$$

where $(\mathrm{IV}) = g(\alpha_0, \eta_0) - g(\alpha, \eta)$ and

$$g(\alpha, \eta) = \log\frac{\alpha}{\eta} + \alpha E(\log x_i) - \frac{\eta_0^{\frac{\alpha}{\alpha_0}}\Gamma(\frac{\alpha}{\alpha_0} + 1)}{\eta}.$$

In the following, I shall show that

(i) (α_0, η_0) is the only critical point satisfying

$$\frac{\partial g(\alpha,\eta)}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial g(\alpha,\eta)}{\partial \eta} = 0,$$

(ii) there exist small positive constants s_1 and s_2 such that

$$g(\alpha_0, \eta_0) - g(\alpha, \eta) \geq s_2 \|\boldsymbol{v} - \boldsymbol{v}_0\|^2$$

for all $\boldsymbol{v} = (\alpha, \eta)' \in \mathsf{B}_{s_1}(\boldsymbol{v}_0)$, where $\boldsymbol{v}_0 = (\alpha_0, \eta_0)'$,

(iii) (α_0, η_0) is the unique maximizer of $g(\alpha, \eta)$.

To show (i), note that

$$\frac{\partial g(\alpha,\eta)}{\partial \eta} = 0 \iff \eta = \eta_0^{\frac{\alpha}{\alpha_0}} \Gamma\left(\frac{\alpha}{\alpha_0} + 1\right),$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} g\left(\alpha, \eta_0^{\frac{\alpha}{\alpha_0}} \Gamma\left(\frac{\alpha}{\alpha_0} + 1\right)\right) = 0 \overset{\text{why?}}{\iff} \sum_{i \geq 1} \left(\frac{1}{i} - \frac{1}{i + \frac{\alpha}{\alpha_0}}\right) = \frac{1}{\frac{\alpha}{\alpha_0}}. \tag{54}$$

- Since the RHS (LHS) of (54) is decreasing (increasing) in $\frac{\alpha}{\alpha_0}$, and $\frac{\alpha}{\alpha_0} = 1$ is a solution of the equation, the desired conclusion (i) follows (why?).
- To show (ii), note first that by (i) and Taylor's theorem, one has for $\|\boldsymbol{v} - \boldsymbol{v}_0\| \leq \delta$ with δ being arbitrarily small,

$$g(\alpha, \eta) = g(\alpha_0, \eta_0) - \frac{1}{2}(\alpha - \alpha_0, \eta - \eta_0)' \begin{pmatrix} \frac{\partial^2 g(\alpha^*, \eta^*)}{\partial \alpha^2} & \frac{\partial^2 g(\alpha^*, \eta^*)}{\partial \alpha \partial \eta} \\ \frac{\partial^2 g(\alpha^*, \eta^*)}{\partial \eta \partial \alpha} & \frac{\partial^2 g(\alpha^*, \eta^*)}{\partial \eta^2} \end{pmatrix} \begin{pmatrix} \alpha - \alpha_0 \\ \eta - \eta_0 \end{pmatrix}, \quad (55)$$

where

$$g(\alpha, \eta) = \frac{\eta_0^{\frac{\alpha}{\alpha_0}} \Gamma\left(\frac{\alpha}{\alpha_0} + 1\right)}{\eta} - \log \frac{\alpha}{\eta} \quad \text{and} \quad \left\| \begin{pmatrix} \alpha^* - \alpha_0 \\ \eta^* - \eta_0 \end{pmatrix} \right\| \leq \left\| \begin{pmatrix} \alpha - \alpha_0 \\ \eta - \eta_0 \end{pmatrix} \right\| \leq \delta.$$

It is straightforward to see that

$$\frac{\partial^2 g(\alpha,\eta)}{\partial \alpha^2} = \frac{\eta_0^{\frac{\alpha}{\alpha_0}}}{\eta}\left[\left(\frac{\log\eta_0}{\alpha_0}\right)^2\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)+\frac{2\log\eta_0}{\alpha_0^2}\Gamma'\left(\frac{\alpha}{\alpha_0}+1\right)+\frac{\Gamma''(\frac{\alpha}{\alpha_0}+1)}{\alpha_0^2}\right]+\frac{1}{\alpha^2},$$

$$\frac{\partial^2 g(\alpha,\eta)}{\partial \eta^2} = \frac{\eta_0^{\frac{\alpha}{\alpha_0}}}{\eta}\left(\frac{2\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)}{\eta^2}\right)-\frac{1}{\eta^2},$$

$$\frac{\partial^2 g(\alpha,\eta)}{\partial \alpha\partial \eta} = \frac{\eta_0^{\frac{\alpha}{\alpha_0}}}{\eta}\left\{-\frac{1}{\alpha_0\eta}\left[\log\eta_0\,\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)+\Gamma'\left(\frac{\alpha}{\alpha_0}+1\right)\right]\right\}.$$

Since $\|\boldsymbol{v}-\boldsymbol{v}_0\| \le \delta$ with δ being arbitrarily small, there exists an arbitrarily small $\theta_\delta > 0$ such that for all $\|\boldsymbol{v}-\boldsymbol{v}_0\| \le \delta$,

$$\frac{\eta_0^{\frac{\alpha}{\alpha_0}}}{\eta}\left(\frac{2\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)}{\eta^2}\right) - \frac{1}{\eta^2} \ge (1-\theta_\delta)\frac{\eta_0^{\frac{\alpha}{\alpha_0}}}{\eta}\frac{\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)}{\eta^2},$$

which, together with (52), yields

$$\begin{aligned}
&\det\begin{pmatrix} \frac{\partial^2 g(\alpha,\eta)}{\partial\alpha^2} & \frac{\partial^2 g(\alpha,\eta)}{\partial\alpha\partial\eta} \\ \frac{\partial^2 g(\alpha,\eta)}{\partial\eta\partial\alpha} & \frac{\partial^2 g(\alpha,\eta)}{\partial\eta^2} \end{pmatrix} \\
\ge\ & \left(\frac{\eta_0^{\frac{\alpha}{\alpha_0}}}{\eta}\right)^2\left(\frac{1}{\alpha_0\eta}\right)^2\left\{(1-\theta_\delta)\left[(\log\eta_0)^2\Gamma^2\left(\frac{\alpha}{\alpha_0}+1\right)\right.\right. \\
& \left. +2\log\eta_0\,\Gamma'\left(\frac{\alpha}{\alpha_0}+1\right)\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)+\Gamma''\left(\frac{\alpha}{\alpha_0}+1\right)\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)\right] \\
& \left. -\left[\left(\log\eta_0\,\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)\right)^2+2\log\eta_0\,\Gamma\left(\frac{\alpha}{\alpha_0}+1\right)\Gamma'\left(\frac{\alpha}{\alpha_0}+1\right)+\left(\Gamma'\left(\frac{\alpha}{\alpha_0}+1\right)\right)^2\right]\right\} \\
>\ & \underline{c},
\end{aligned}$$

for some small $\underline{c} > 0$ depending only on δ and θ_δ.

- Therefore,

$$\inf_{\boldsymbol{v}\in \mathsf{B}_\delta(\boldsymbol{v}_0)} \lambda_{\min}\begin{pmatrix} \frac{\partial^2 g(\alpha,\eta)}{\partial \alpha^2} & \frac{\partial^2 g(\alpha,\eta)}{\partial \alpha \partial \eta} \\ \frac{\partial^2 g(\alpha,\eta)}{\partial \eta \partial \alpha} & \frac{\partial^2 g(\alpha,\eta)}{\partial \eta^2} \end{pmatrix} \overset{\text{why?}}{>} \underline{c}^*, \tag{56}$$

  for some $\underline{c}^* > 0$.

- By (55) and (56), it holds that for $\|\boldsymbol{v} - \boldsymbol{v}_0\| \le \delta$,

$$g(\alpha_0, \eta_0) - g(\alpha, \eta) \ge \frac{\underline{c}^*}{2}\|\boldsymbol{v} - \boldsymbol{v}_0\|^2,$$

  and hence (ii) follows.

- Now, (iii) follows directly from (i), (ii) and the differentiability of $g(\alpha, \beta)$.

- In view of (i)–(iii) and (53), the consistency of

$$(\hat{\alpha}, \hat{\eta}) = \underset{(\alpha,\eta)\in[\delta_1, M_1]\times[\delta_2, M_2]}{\operatorname{argmax}} \frac{1}{n}\ell(\alpha, \eta)$$

  is ensured by

$$\sup_{\alpha\in[\delta_1, M_1]} |(\mathrm{I})| = o_p(1),\ |(\mathrm{II})| = o_{p(1)}, \tag{57}$$

$$\sup_{(\alpha,\eta)\in[\delta_1, M_1]\times[\delta_2, M_2]} |(\mathrm{III})| = o_p(1). \tag{58}$$

- (57) is an immediate consequence of the (classical) law of large numbers, whereas (58) relies on the so-called uniform law of large numbers.
- In the following, I'll provide a proof of (58).

- Note first that (58) is guaranteed by (why?)

$$\sup_{\alpha\in[\delta_1,M_1]}\left|\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}-E(x_i^{\alpha})\right|=o_p(1). \tag{59}$$

- Define $g_t(\alpha)=x_t^{\alpha}-E(x_t^{\alpha})$. Then,

$$\begin{aligned}
& E\sup_{\alpha\in[\delta_1,M_1]}\left(\frac{1}{n}\sum_{t=1}^{n}g_t(\alpha)-g_t(\alpha_0)\right)^2 \\
\underset{\text{why?}}{\leq} \quad & (M_1-\delta_1)E\left\{\int_{\delta_1}^{M_1}\left(\frac{1}{n}\sum_{t=1}^{n}g_t'(x)\right)^2 \mathrm{d}x\right\} \\
\underset{\text{why?}}{\leq} \quad & (M_1-\delta_1)^2\sup_{\alpha\in[\delta_1,M_1]}E\left(\frac{1}{n}\sum_{t=1}^{n}g_t'(\alpha)\right)^2 \\
\underset{\text{why?}}{\leq} \quad & \frac{(M_1-\delta_1)^2}{\eta}\sup_{\alpha\in[\delta_1,M_1]}Var(x_1^{\alpha}\log x_1)=O(n^{-1}),
\end{aligned}$$

which implies

$$\sup_{\alpha\in[\delta_1,M_1]}\frac{1}{n}\sum_{t=1}^{n}g_t(\alpha)-g_t(\alpha_0)\overset{\text{why?}}{=}O_p(n^{-\frac{1}{2}}). \tag{60}$$

- Moreover, it is easy to see that

$$\frac{1}{n}\sum_{t=1}^{n} g_t(\alpha_0) = O_p(n^{-\frac{1}{2}}). \tag{61}$$

- Combining (60) and (61) gives (59). Thus, the proof is complete.

Question

Please find the limiting distribution of $(\hat{\alpha}, \hat{\eta})$.