Lecture 5. Profile Likelihood and Generalized Profile Likelihood Approach

[1]*School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

1. Profile Likelihood

Consider the case where we have a sample $X = \{X_1, \cdots, X_n\}$ with joint distribution function $f(x|\theta_0) \in \mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$. Now, assume the unknown parameter θ can be partitioned into two parts, $\theta^T = (\phi^T, \eta^T)$, where $\phi \in \mathcal{M}_{p\times 1}$ refers to the parameters of interests, and $\eta \in \mathcal{M}_{q\times 1}$ refers to the nuisance parameter. Denote Θ_ϕ as the parameter space of η with a fixed ϕ (or one can think it as the parameter space of η conditional on ϕ), e.g., $\Theta_{\phi'} = \left\{\eta : (\phi'^T, \eta) \in \Theta\right\}$, and Θ_η can be similarly defined. We again use $\theta_0^T = (\phi_0^T, \eta_0^T)$ to denote the unknown underlying true parameter value of the data generating process. When we are seeking the MLE, we will have to estimate both ϕ and η despite our interests only lays on ϕ. To achieve this one often profiles out the nuisance parameters, known as the profile likelihood approach. Formally,

Definition 1.1 (♣ **Profile Likelihood**). Denote

$$L(\theta|X) = L(\phi, \eta|X) = f(X|\theta), \text{ and } \ell(\theta) = \ell(\theta|X) = \ell(\phi, \eta|X) = \log L(\theta|X)$$

the likelihood and the log-likelihood of the sample. Then the profile likelihood and the profile log-likelihood of ϕ is defined as

$$\mathcal{PL}(\phi|X) = \sup_{\eta\in\Theta_\phi} L(\phi, \eta|X), \quad \text{and} \quad \mathcal{P}\ell(\phi|X) = \sup_{\eta\in\Theta_\phi} \ell(\phi, \eta|X).$$

Apparently, if we pretend that ϕ is known for a while, we may rewrite the log-likelihood as $\ell(\phi, \eta|X) = \ell_\phi(\eta|X)$ to indicate that ϕ is fixed but η varies within Θ_ϕ. This $\ell_\phi(\eta|X)$ can be think as a new curve over η. Therefore, the "MLE" of η with the ϕ we pretend to know is

$$\hat{\eta}_\phi = \arg\max_{\eta\in\Theta_\phi} \ell_\phi(\eta|X) = \arg\max_{\eta\in\Theta_\phi} \ell(\phi, \eta|X).$$

Accordingly, for ϕ varying within the parameter space, we have the profile log-likelihood

$$\mathcal{P}\ell(\phi|X) = \sup_{\eta\in\Theta_\phi} \ell(\phi,\eta|X) = \ell(\phi,\hat{\eta}_\phi|X).$$

Meaning our profile likelihood is nothing but just a new "curve" on the original "manifold" where for each ϕ, this "curve" takes the largest value of the original "manifold" when η varies.

If we further denote the MLE of this profile log-likelihood as

$$\begin{aligned}\hat{\phi} &= \arg \max_{\phi\in\{(\phi,\hat{\eta}_\phi)\in\Theta\}} \mathcal{P}\ell(\phi|X) = \arg\max_{\phi\in\Theta_{\hat{\eta}}} \mathcal{P}\ell(\phi|X)\\ &= \arg\max_{\phi} \ell(\phi,\hat{\eta}_\phi|X) = \arg\max_{\phi}\left[\arg\max_{\eta\in\Theta_\phi}\ell(\phi,\eta|X)\right],\end{aligned}$$

then for arbitrary $(\phi_1,\eta_1)\in\Theta$, it's clear that

$$\begin{aligned}\ell(\phi_1,\eta_1|X) &\le \sup_{\eta\in\Theta_{\phi_1}} \ell(\phi_1,\eta|X) = \ell(\phi_1,\hat{\eta}_{\phi_1}|X)\\ &\le \sup_{\phi}\ell(\phi,\hat{\eta}_\phi|X) = \ell(\hat{\phi},\hat{\eta}_{\hat{\phi}}|X).\end{aligned}$$

Hence $(\hat{\phi},\hat{\eta}_{\hat{\phi}})$ is the MLE of the original log-likelihood. The calculation is devided into two steps simply cause we were profiling out the nuisance parameter η (to write η into a function of ϕ) in the first step. Later in this notes, we will use $\hat{\eta} = \hat{\eta}_{\hat{\phi}}$ for notation simplicity.

• *Example* 1.2 (Two Sample Exponential Distribution). Suppose we have two mutually independent samples $X = \{X_1,\cdots,X_m\}$ and $Y = \{Y_1,\cdots,Y_n\}$, where each $X_i \sim_{i.i.d}$ Exponential(λ) and $Y_i \sim_{i.i.d}$ Exponential($\lambda\alpha$), please give the MLE of $\theta = (\lambda,\alpha)$.

2. Nested Likelihood Ratio and Wilks' Theorem

In this section, we consider that $X = \{X_1, \cdots, X_n\}$ is a random sample, we abuse the notation a little bit by denote each $X_i's$ distribution function again as $f(x|\theta_0) \in \mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$ (instead of using $f(x|\theta)$ to denote joint distribution of X. Commonly, it can be distinguished since we will write $f(X_i|\theta)$ when we refer to each observations distribution). Therefore, we have

$$\ell(\theta) = \ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i) = \sum_{i=1}^{n} \log f(X_i|\theta).$$

Now, one quantity that are closely related to profile likelihood is the nested likelihood ratio statistic, which admits the form

$$\delta_{\ell R}(X) = 2\left\{ \sup_{\phi,\eta} \ell(\phi,\eta|X) - \sup_{\eta} \ell(\phi_0,\eta|X) \right\} = 2\left\{ \sup_{\phi,\eta} \ell(\phi,\eta|X) - \mathcal{P}\ell(\phi_0|X) \right\}$$

As a test statistic, we are by natural interested in its limiting distribution. However, to derive the limiting distribution for this statistic is a little more complicated than derive the CAN property of MLE which does not involve nuisance parameter and therefore we can directly apply Taylor expansion with respect to the true parameter values. To adjust for this "$\sup_\eta$" appeared in both terms of $\delta_{\ell R}(X)$, notice that

$$\begin{aligned}\delta_{\ell R}(X) &= 2\left\{ \sup_{\phi,\eta} \ell(\phi,\eta|X) - \ell(\phi_0,\eta_0|X) \right\} - 2\left\{ \sup_{\eta} \ell(\phi_0,\eta|X) - \ell(\phi_0,\eta_0|X) \right\} \\ &= 2\left\{ \ell(\hat{\phi},\hat{\eta}|X) - \ell(\phi_0,\eta_0|X) \right\} - 2\left\{ \ell(\phi_0,\hat{\eta}_{\phi_0}|X) - \ell(\phi_0,\eta_0|X) \right\}\end{aligned}$$

Accordingly, we may think about using Taylor expansions several times to analyze the limit behavior of this nested likelihood ratio statistic $\delta_{\ell R}$. Below, we first present the ultimate result for its limiting distribution.

Theorem 2.1 (♣ **Wilks' Theorem**). *Under Condition.5.1, we have*

$$\frac{1}{\sqrt{n}} \left(\left.\frac{\partial \ell(\phi,\eta|X)}{\partial \phi}\right|_{\phi_0,\hat{\eta}_{\phi_0}} - \left.\frac{\partial \ell(\phi,\eta|X)}{\partial \phi}\right|_{\phi_0,\eta_0} + \left.\frac{\partial \ell(\phi,\eta|X)}{\partial \eta}\right|_{\phi_0,\eta_0} \cdot I^{-1}_{\eta_0\eta_0} I_{\eta_0\phi_0} \right) \xrightarrow{p} 0. \tag{2.1}$$

and

$$\frac{1}{\sqrt{n}} \left.\frac{\partial \ell(\phi,\eta|X)}{\partial \phi}\right|_{\phi_0,\hat{\eta}_{\phi_0}} \xrightarrow{d} N\Big(0, \big(I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I^{-1}_{\eta_0\eta_0} I_{\eta_0\phi_0}\big)\Big). \tag{2.2}$$

Further more, when the second Barlett Indentity holds, we have

$$\delta_{\ell R}(X) \xrightarrow{d} \chi_p^2. \tag{2.3}$$

where p denotes the dimension of θ. This result is often called Wilks' Theorem.

Proof. First, notice that $\hat{\eta}_{\phi_0}$ is the MLE of η when we "pretend" ϕ_0 is known, therefore, under (a) and (b) of Condition.5.1, we have $\hat{\eta}_{\phi_0}$ is consistent to η_0, i.e., $\hat{\eta}_{\phi_0} \xrightarrow{p} \eta_0$. And according to (c) of Condition.5.1, for large sample size n, we will have $\hat{\eta}_{\phi_0}$ also lays in the interior of Θ_{ϕ_0}, and thus $\hat{\eta}_{\phi_0}$ satisfy

$$0 = \frac{1}{\sqrt{n}} \frac{\partial \ell(\phi, \eta|X)}{\partial \eta}\bigg|_{\phi_0, \hat{\eta}_{\phi_0}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell(\phi, \eta|X_i)}{\partial \eta}\bigg|_{\phi_0, \hat{\eta}_{\phi_0}}. \tag{2.4}$$

By conduct Taylor expansion on the right-hand-side of (2.4), we have

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell(\phi, \eta|X_i)}{\partial \eta}\bigg|_{\phi_0, \eta_0} + \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\phi, \eta|X_i)}{\partial \eta \partial \eta^T}\bigg|_{\phi_0, \eta_0} + R_1 \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell(\phi, \eta|X_i)}{\partial \eta}\bigg|_{\phi_0, \eta_0} + \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \cdot I_{\eta_0\eta_0} + R_1 + R_2, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} R_1 &= \frac{1}{2\sqrt{n}} \cdot n\big(\hat{\eta}_{\phi_0} - \eta_0\big)^T \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3 \ell(\phi, \eta|X_i)}{\partial \eta^3}\bigg|_{\phi_0, \tilde{\eta}_1} \cdot \big(\hat{\eta}_{\phi_0} - \eta_0\big), \\ \text{and} \quad R_2 &= \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\phi, \eta|X_i)}{\partial \eta \partial \eta^T}\bigg|_{\phi_0, \eta_0} - I_{\eta_0\eta_0}\right]. \end{aligned}$$

Recall that $\sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \to N(0, I_{\eta_0\eta_0}^{-1})$, and by (h) of Condition.5.1, we conclude

$$|R_1| \leq \frac{1}{2\sqrt{n}} \cdot n\big\|\hat{\eta}_{\phi_0} - \eta_0\big\|^2 \cdot M \xrightarrow{d} 0, \quad \text{hence} \;\; R_1 \xrightarrow{p} 0.$$

While by law of large numbers and (g) of Condition.5.1, we have $R_2 \xrightarrow{d} 0$, hence $R_2 \xrightarrow{p} 0$. Overall, we obtain from (2.5) that

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell(\phi, \eta|X_i)}{\partial \eta}\bigg|_{\phi_0, \eta_0} \cdot I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0} + \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \cdot I_{\eta_0\phi_0}\right) \xrightarrow{p} 0. \tag{2.6}$$

By substituting (2.6) into (2.1), we see it's sufficient to prove

$$\frac{1}{\sqrt{n}} \left(\frac{\partial \ell(\phi, \eta|X)}{\partial \phi}\bigg|_{\phi_0, \hat{\eta}_{\phi_0}} - \frac{\partial \ell(\phi, \eta|X)}{\partial \phi}\bigg|_{\phi_0, \eta_0} - n\big(\hat{\eta}_{\phi_0} - \eta_0\big) \cdot I_{\eta_0\phi_0}\right) \xrightarrow{p} 0. \tag{2.7}$$

Again by Taylor expansion, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\phi,\eta|X)}{\partial\phi}\bigg|_{\phi_0,\hat\eta_{\phi_0}} - \frac{\partial \ell(\phi,\eta|X)}{\partial\phi}\bigg|_{\phi_0,\eta_0}\right) &= \frac{1}{\sqrt{n}}\big(\hat\eta_{\phi_0}-\eta_0\big)\cdot\frac{\partial^2\ell(\phi,\eta|X)}{\partial\phi\partial\eta^T}\\ =\sqrt{n}\big(\hat\eta_{\phi_0}-\eta_0\big)\cdot\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\phi,\eta|X_i)}{\partial\phi\partial\eta^T} &= \sqrt{n}\big(\hat\eta_{\phi_0}-\eta_0\big)\cdot I_{\eta_0\phi_0}+R_3,\end{aligned} \tag{2.8}$$

where by law of large numbers and (g) of Condition.5.1, we again have

$$R_3=\sqrt{n}\big(\hat\eta_{\phi_0}-\eta_0\big)\cdot\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\phi,\eta|X_i)}{\partial\phi\partial\eta^T}\bigg|_{\phi_0,\eta_0} - I_{\eta_0\phi_0}\right]\xrightarrow{d} 0, \tag{2.9}$$

hence $R_3\xrightarrow{p}0$, and by combining (2.6)-(2.9), we conclude (2.1) holds. Now, to prove (2.2), notice that according to central limit theorem and (g) of Condition.5.1, we have

$$\frac{1}{\sqrt{n}}\begin{pmatrix}\frac{\partial\ell(\phi,\eta|X)}{\partial\phi}\\ \frac{\partial\ell(\phi,\eta|X)}{\partial\eta}\end{pmatrix}\Bigg|_{\phi_0,\eta_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\begin{pmatrix}\frac{\partial\ell(\phi,\eta|X_i)}{\partial\phi}\\ \frac{\partial\ell(\phi,\eta|X_i)}{\partial\eta}\end{pmatrix}\Bigg|_{\phi_0,\eta_0}\xrightarrow{d} N\left(0,\begin{pmatrix}I_{\phi_0\phi_0} & I_{\phi_0\eta_0}\\ I_{\eta_0\phi_0} & I_{\eta_0\eta_0}\end{pmatrix}\right).$$

therefore, by (2.1), we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial\ell(\phi,\eta|X)}{\partial\phi}\bigg|_{\phi_0,\hat\eta_{\phi_0}} &= \begin{pmatrix}1\\ -I^{-1}_{\eta_0\eta_0}I_{\eta_0\phi_0}\end{pmatrix}^T\cdot\frac{1}{\sqrt{n}}\begin{pmatrix}\frac{\partial\ell(\phi,\eta|X)}{\partial\phi}\\ \frac{\partial\ell(\phi,\eta|X)}{\partial\eta}\end{pmatrix}\Bigg|_{\phi_0,\eta_0}\\ \xrightarrow{d}N&\left(0,\begin{pmatrix}1\\ -I^{-1}_{\eta_0\eta_0}I_{\eta_0\phi_0}\end{pmatrix}^T\cdot\begin{pmatrix}I_{\phi_0\phi_0} & I_{\phi_0\eta_0}\\ I_{\eta_0\phi_0} & I_{\eta_0\eta_0}\end{pmatrix}\cdot\begin{pmatrix}1\\ -I^{-1}_{\eta_0\eta_0}I_{\eta_0\phi_0}\end{pmatrix}\right)\\ =N&\Big(0,\big(I_{\phi_0\phi_0}-I_{\phi_0\eta_0}I^{-1}_{\eta_0\eta_0}I_{\eta_0\phi_0}\big)\Big),\end{aligned}$$

which finishes the prove of (2.2). As for (2.3), by (a) and (b) of Condition.5.1, we have $\hat\theta=(\hat\phi,\hat\eta)\xrightarrow{p}\theta_0=(\phi_0,\eta_0)$ and $\hat\eta_{\phi_0}\xrightarrow{p}\eta_0$. While by (c) of Condition.5.1, we have $\hat\theta$ lays in the interior of Θ and $\hat\eta_{\phi_0}$ lays in the interior of Θ_{ϕ_0} for large sample size. Therefore,

$$\frac{\partial\ell(\phi,\eta|X)}{\partial\theta}\bigg|_{\hat\phi,\hat\eta}=0,\quad\text{and}\quad\frac{\partial\ell(\phi_0,\eta|X)}{\partial\eta}\bigg|_{\phi_0,\hat\eta_{\phi_0}}=0.$$

Thus, by Taylor expansion with respect to point $\hat\theta=(\hat\phi,\hat\eta)$ and the second Bartlett identity, we have

$$\begin{aligned}2\Big\{\ell(\hat\phi,\hat\eta|X)-\ell(\phi_0,\eta_0|X)\Big\} &= -\sqrt{n}\big(\hat\theta-\theta_0\big)^T\cdot\frac{1}{n}\sum_{i=1}^n\frac{\partial^2\ell(\theta|X_i)}{\partial\theta\partial\theta^T}\bigg|_{\hat\theta}\cdot\sqrt{n}\big(\hat\theta-\theta_0\big)-R_4\\ &= \sqrt{n}\big(\hat\theta-\theta_0\big)^T\cdot I(\theta_0)\cdot\sqrt{n}\big(\hat\theta-\theta_0\big)-R_4-R_5,\end{aligned}\tag{2.10}$$

where by (h) of Condition.(5.1),

$$|R_4| \leq \frac{1}{3\sqrt{n}} \cdot \left\| \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \right\|^3 \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial^3 \ell(\phi, \eta | X_i)}{\partial \theta^3} \bigg|_{\tilde{\theta}_2} \leq \frac{M}{3\sqrt{n}} \cdot \left\| \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \right\|^3 \xrightarrow{d} 0, \tag{2.11}$$

hence $R_4 \xrightarrow{p} 0$, and by CAN property of $\hat{\theta}$, law of large numbers, (g) and (f) of Condition.5.1 (notice that (f) also infers the continuity of $I(\theta)$), we have

$$\begin{aligned} R_5 = \sqrt{n}\big(\hat{\theta} - \theta_0\big)^T \cdot \Bigg[& -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ell(\theta | X_i)}{\partial \theta \partial \theta^T} \bigg|_{\hat{\theta}} + \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ell(\theta | X_i)}{\partial \theta \partial \theta^T} \bigg|_{\theta_0} \\ & -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ell(\theta | X_i)}{\partial \theta \partial \theta^T} \bigg|_{\theta_0} - I(\theta_0) \Bigg] \cdot \sqrt{n}\big(\hat{\theta} - \theta_0\big) \xrightarrow{d} 0, \end{aligned} \tag{2.12}$$

hence $R_5 \xrightarrow{p} 0$. Similarly, we have

$$2\Big\{ \ell(\phi_0, \hat{\eta}_{\phi_0} | X) - \ell(\phi_0, \eta_0 | X) \Big\} = \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big)^T I_{\eta_0 \eta_0} \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) - R_6 - R_7, \tag{2.13}$$

with $R_6 \xrightarrow{p} 0$, $R_7 \xrightarrow{p} 0$. Combining (2.10)-(2.13), we obtain

$$\delta_{\ell R}(X) = n\big(\hat{\theta} - \theta_0\big)^T I(\theta_0) \big(\hat{\theta} - \theta_0\big) - n\big(\hat{\eta}_{\phi_0} - \eta_0\big)^T I_{\eta_0 \eta_0} \big(\hat{\eta}_{\phi_0} - \eta_0\big) + R_8, \tag{2.14}$$

where by Slutsky's theorem, $R_8 = -R_4 - R_5 + R_6 + R_7 \xrightarrow{p} 0$. Further, similar to (2.6), we have

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \ell(\theta | X_i)}{\partial \theta} \bigg|_{\phi_0, \eta_0} \cdot I(\theta_0)^{-1} + \sqrt{n}\big(\hat{\theta} - \theta_0\big) \right) \xrightarrow{p} 0. \tag{2.15}$$

Combining (2.6) and (2.15), we get

$$\begin{aligned} R_9 &= -\left(I_{\eta_0 \eta_0}^{-1} I_{\eta_0 \phi_0},\ 1 \right) \cdot \sqrt{n}\big(\hat{\theta} - \theta_0\big) + \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \\ &= -\sqrt{n}\big(\hat{\theta} - \theta_0\big)^T \cdot I(\theta_0) \cdot \begin{pmatrix} 0_{p \times 1} \\ 1_{q \times 1} \end{pmatrix} \cdot I_{\eta_0 \eta_0}^{-1} + \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \\ &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \ell(\theta | X_i)}{\partial \theta^T} \bigg|_{\phi_0, \eta_0} \cdot I(\theta_0)^{-1} \cdot I(\theta_0) \cdot \begin{pmatrix} 0_{p \times 1} \\ 1_{q \times 1} \end{pmatrix} \cdot I_{\eta_0 \eta_0}^{-1} + \sqrt{n}\big(\hat{\eta}_{\phi_0} - \eta_0\big) \right) \xrightarrow{p} 0. \end{aligned} \tag{2.16}$$

together with the fact that $\sqrt{n}\big(\hat{\theta} - \theta_0\big) \xrightarrow{d} N(0, I(\theta_0)^{-1})$, (2.16) implies

$$\begin{aligned} & n\big(\hat{\eta}_{\phi_0} - \eta_0\big)^T I_{\eta_0 \eta_0} \big(\hat{\eta}_{\phi_0} - \eta_0\big) - n\big(\hat{\theta} - \theta_0\big) \begin{pmatrix} I_{\eta_0 \eta_0}^{-1} I_{\eta_0 \phi_0} \\ 1 \end{pmatrix} I_{\eta_0 \eta_0} \left(I_{\eta_0 \eta_0}^{-1} I_{\eta_0 \phi_0},\ 1 \right) \big(\hat{\theta} - \theta_0\big) \\ = & n\big(\hat{\eta}_{\phi_0} - \eta_0\big)^T I_{\eta_0 \eta_0} \big(\hat{\eta}_{\phi_0} - \eta_0\big) - n\big(\hat{\theta} - \theta_0\big) \begin{pmatrix} I_{\phi_0 \eta_0} I_{\eta_0 \eta_0}^{-1} I_{\eta_0 \phi_0} & I_{\phi_0 \eta_0} \\ I_{\eta_0 \phi_0} & I_{\eta_0 \eta_0} \end{pmatrix} \big(\hat{\theta} - \theta_0\big) = R_{10} \xrightarrow{p} 0. \end{aligned} \tag{2.17}$$

If we denote

$$
\begin{aligned}
Z &= \begin{pmatrix} I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0} & 0 \\ 0 & 0 \end{pmatrix}^{1/2} \cdot \sqrt{n}(\hat{\theta} - \theta_0) \\
&\xrightarrow{d} N\left(0, \begin{pmatrix} I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0} & 0 \\ 0 & 0 \end{pmatrix}^{1/2} I(\theta_0)^{-1} \begin{pmatrix} I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0} & 0 \\ 0 & 0 \end{pmatrix}^{1/2}\right) \\
&= N\left(0, \begin{pmatrix} I_{p\times p} & 0 \\ 0 & 0 \end{pmatrix}\right),
\end{aligned}
$$

where we used the fact that

$$
I(\theta_0)^{-1} = \begin{pmatrix} I_{\phi_0\phi_0} & I_{\phi_0\eta_0} \\ I_{\eta_0\phi_0} & I_{\eta_0\eta_0} \end{pmatrix}^{-1} = \begin{pmatrix} \left(I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0}\right)^{-1} & -I_{\phi_0\phi_0}^{-1} I_{\phi_0\eta_0} \left(I_{\eta_0\eta_0} - I_{\eta_0\phi_0} I_{\phi_0\phi_0}^{-1} I_{\phi_0\eta_0}\right)^{-1} \\ -I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0} \left(I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0}\right)^{-1} & \left(I_{\eta_0\eta_0} - I_{\eta_0\phi_0} I_{\phi_0\phi_0}^{-1} I_{\phi_0\eta_0}\right)^{-1} \end{pmatrix}
$$

So by (2.14), (2.17) and Slutsky's theorem,

$$
\begin{aligned}
\delta_{\ell R}(X) &= n\big(\hat{\theta} - \theta_0\big) \begin{pmatrix} I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0} & 0 \\ 0 & 0 \end{pmatrix} \big(\hat{\theta} - \theta_0\big) + R_{10} + R_8 \\
&= Z^T \cdot Z + R_{10} + R_8 \xrightarrow{d} \chi_p^2
\end{aligned}
$$

which finishes the prove. □

Remark 2.2. (i) As a matter of fact, for $\hat{\phi}$ in MLE, we have

$$
\sqrt{n}\big(\hat{\phi} - \phi_0\big) \xrightarrow{d} N\Big(0, \big(I_{\phi_0\phi_0} - I_{\phi_0\eta_0} I_{\eta_0\eta_0}^{-1} I_{\eta_0\phi_0}\big)^{-1}\Big),
$$

which is MLE's marginal limit distribution. Similar results holds for $\hat{\eta}$.

(ii) Roughly speaking, we have

$$
\delta_{\ell R}(X) = \underbrace{2\Big\{\sup_{\phi,\eta} \ell(\phi, \eta|X) - \ell(\phi_0, \eta_0|X)\Big\}}_{\approx \chi_{p+q}^2} - \underbrace{2\Big\{\sup_{\eta} \ell(\phi_0, \eta|X) - \ell(\phi_0, \eta_0|X)\Big\}}_{\approx \chi_q^2}.
$$

More generally, the Wilks' theorem works beyond the case of nested likelihood ratio. Specifically, consider a hypothesis testing problem,

$$
H_0 : \theta \in \Theta_0 \quad v.s. \quad H_1 : \theta \in \Theta/\Theta_0,
$$

then for a random sample $X = \{X_1, \cdots, X_n\}$ such that each X_i has distribution function $f(x|\theta_0) \in \mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$. The log-likelihood ratio statistic is

$$
\delta_{\ell R}(X) = -2 \log \left[\frac{\max_{\theta\in\Theta_0} \prod_{i=1}^n f(X_i|\theta)}{\max_{\theta\in\Theta} \prod_{i=1}^n f(X_i|\theta)}\right] = 2\Big\{\sup_{\theta\in\Theta} \ell(\theta|X) - \sup_{\theta\in\Theta_0} \ell(\theta|X)\Big\}
$$

Theorem 2.3 (♣ **Wilks' Theorem** (General Case))**.** *Under "Regularity Condition* $\mho$*", assume that* $\dim(\Theta) - \dim(\Theta_0) = p > 0$*, then* $\delta_{\ell R}(X) \xrightarrow{d} \chi_p^2$*.*

A proof of Theorem.2.3 has originally arisen in Wilks (1938), but later been pointed out that was not rigorous. We refer to Chernoff (1954), Wilks (1962) and Van der Vaart (2000) for more details of "Regularity Condition $\mho$" and the proof. As one would expect, the regularity condition $\mho$ would not differ from Condition.5.1 too much but only slightly stronger.

3. Applications of Profile Likelihood and Wilks' Theorem

Definition 3.1 (♣ **Pivotal Quantity and Asymptotic Pivotal Quantity**). A pivotal quantity or pivot is a function of observations and unobservable parameters, i.e., $\phi(X, \theta)$, such that its probability distribution does not depend on the unknown parameter θ. And $\phi(X, \theta)$ is call an asymptotic pivotal quantity if its asymptotic distribution does not depend on the unknown parameter θ.

- *Example* 3.2 (♣ **Confidence Interval for Normal Distribution**). For a random sample $X = \{X_1, \cdots, X_n\}$ been drawn from $N(\mu, \sigma^2)$ with μ and σ^2 both unknown, please give a 95% confidence interval for μ.

Remark 3.3. For this specific case, $t(\mu)$ is a pivotal quantity and we actually have the exact distribution of $t(\mu)$ being the t-distribution with degree of freedom $n-1$. So we can construct a more "accurate" confidence interval of μ.

- *Example* 3.4 (♣ **Contingency Tables**). A total number of n subjects are drawn independently such that each subject is classified according to two characteristics: A, with possible outcomes $A_1, \ldots, A_k$, and B, with possible outcomes $B_1, \ldots, B_k$. The probability that a subject has properties (A_i, B_j) will be denoted by p_{ij} and the number of such subjects in the sample by n_{ij}. The joint distribution of the k^2 variables n_{ij} are from an unrestricted multinomial distribution, and the results of the sample can be represented in an k^2 contingency table, i.e., Table.1. Please give a joint confidence region for $\{p_{ii}, i = 1, \cdots, m\}$ for some fixed $1 \leq m \leq k$.

TABLE 1

$k \times k$ Contingency Table

	B_1	$\ldots$	B_k	Total
A_1	n_{11}	$\ldots$	n_{1k}	n_{1+}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	n_{k1}	$\ldots$	n_{kk}	n_{k+}
Total	n_{+1}	$\ldots$	n_{+k}	n

$\Leftrightarrow$

	B_1	$\ldots$	B_k	Total
A_1	p_{11}	$\ldots$	p_{1k}	p_{1+}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	p_{k1}	$\ldots$	p_{kk}	p_{k+}
Total	p_{+1}	$\ldots$	p_{+k}	1

4. Generalized Profile Likelihood Approach

As useful as the profile likelihood in profiling the nuisance parameters, it can still bring complication when seeking $\hat{\eta}_\phi$ and the profile log-likelihood $\mathcal{P}\ell(\phi|X)$. A coming up reasonable question is, whether can we use some other estimator $\hat{\eta}$ of η (hopefully which is easier to obtain or more trackable or has some nice property compare to $\hat{\eta}_\phi$) instead of using $\hat{\eta}_\phi$. Luckily, under Condition.5.1, we have the following Lemma to ensure that the log-likelihood with the nuisance parameter η substituted by some consistent estimator $\hat{\eta}$ would behave similarly to the original log-likelihood.

Lemma 4.1 (Likelihood Approximation)**.** *For a random sample $X = \{X_1, \cdots, X_n\}$ such that each X_i has distribution function $f(x|\theta_0) \in \mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$, if $\hat{\eta} = \hat{\eta}(X)$ is an arbitrary consistent estimator of the nuisance parameter η, i.e., $\hat{\eta} \xrightarrow{p} \eta$, then under Condition.5.1, we have*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \ell(\phi, \hat{\eta}|X_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(\phi, \eta|X_i) \right| \xrightarrow{p} 0.$$

A proof of Lemma.4.1 is postponed to Supplement.

Following the thought, in practice, we define

Definition 4.2 (♣ **Generalized Profile Likelihood**). For every given ϕ, denote $\hat{\eta} = \hat{\eta}(X, \phi)$ an abitrary consistent estimator of η, then we call $\ell_p(\phi, \hat{\eta}|X)$ a generalized profile log-likelihood, and

$$\hat{\phi}_p = \arg\max_{\phi} \ell_p(\phi, \hat{\eta}|X) = \arg\max_{\phi} \ell_p(\phi, \hat{\eta}(X, \phi)|X)$$

a maximum generalized profile log-likelihood estimator.

4.1. Applications of Generalized Profile Likelihood Approach

- *Example* 4.3 (♣ **Random Design Regression Model**). Consider the regression model

$$Y_{n\times 1} = \alpha \cdot \mathbb{1}_{n\times 1} + X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1},$$

where $X^T = (X_1^T, \cdots, X_n^T)$ are i.i.d with each $\mathbb{E}X_i = 0_{p\times 1}$ and its distribution does not depend on $\gamma^T = (\alpha, \beta^T)$, and $\epsilon \sim N(0_{n\times 1}, I_{n\times n})$ is the white noise that is independent of X. Please estimate the parameter of interests γ using MLE and the generalized profile likelihood approach.

- *Example* 4.4 (♣ **Fixed Effect One-Way-Layout**). If this gain of simplicity in Example.4.3 does not arouse your interests, consider the following simple variation been known as fixed effect one-way-layout regression model,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} \mathbb{1}_{k_1\times 1} & 0_{k_1\times 1} & \cdots & 0_{k_1\times 1} & X_1 \\ 0_{k_2\times 1} & \mathbb{1}_{k_2\times 1} & \cdots & 0_{k_2\times 1} & X_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{k_p\times 1} & 0_{k_p\times 1} & \cdots & \mathbb{1}_{k_p\times 1} & X_p \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}, \quad \text{or } Y = Z\gamma + \epsilon,$$

where X is independent of ϵ, $\sum_{i=1}^p k_i = n$, $k_i(n) \to \infty$ for each $1 \le i \le p$, and

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{ik_i} \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ik_i} \end{pmatrix}, \quad \epsilon_i = \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ik_i} \end{pmatrix}, \quad i = 1, \cdots, p.$$

Assume $\{X_{ij}, 1 \leq j \leq k_i, 1 \leq i \leq p\}$ and $\{\epsilon_{ij}, 1 \leq j \leq k_i, 1 \leq i \leq p\}$ are two random samples, where for $1 \leq j \leq k_i$, $1 \leq i \leq p$, each

$$\mathbb{E}X_{ij} = 0, \quad \text{its distribution does not depend on } \gamma, \quad \text{and } \ \epsilon_{ij} \sim N(0,1).$$

Or we can formulate the model as

$$Y_{ij} = \alpha_i + X_{ij}\beta + \epsilon_{ij}, \ \ 1 \leq j \leq k_i, \ \ 1 \leq i \leq p.$$

Please estimate the parameter of interests γ using MLE and the generalized profile likelihood approach.

• *Example* 4.5 (♣ **Two-parameter Exponential Distribution**). Let $X_1, \cdots, X_n$ be a random sample from the Exponential(a, b), each with density

$$f(x|a,b) = \frac{1}{b} e^{-(x-a)/b} \cdot \mathbb{1}(x \geq a), \quad a \in \mathbb{R}, b > 0.$$

Now, for each b, if we pretend b is fixed for a moment, we may then seek the UMVUE of a and obtain a corresponding generalized profile likelihood by plugin the UMVUE and let b vary again. Please estimate $\theta = (a, b)$ using this generalized profile likelihood.

• *Example* 4.6 (♣ **Gamma Distribution**). Let $X_1, \cdots, X_n$ be a random sample from the Gamma(α, β), each with density

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \cdot \mathbb{1}(x > 0), \quad \alpha, \beta > 0.$$

First, please give the estimator of $\theta = (\alpha, \beta)$ using profile likelihood. Second, please find a method of moments estimator of α and obtain a corresponding generalized profile likelihood by plugin this method of moments estimator, and estimate $\theta = (\alpha, \beta)$ using this generalized profile likelihood.

*4.2. *Asymptotic Behavior of the Estimator in Generalized Profile Likelihood Approach*

In practice, we don't really have to requiring an estimator $\hat{\eta}$ such that

$$\hat{\eta} \xrightarrow{p} \eta \ \text{ for all } \ \eta \in \Theta_{\cdot,\eta} \triangleq \Big\{\eta : \exists \phi, s.t., (\phi, \eta) \in \Theta\Big\} = \bigcup_{\phi} \Theta_\phi,$$

i.e., $\Theta_{\cdot,\eta}$ is the set of all possible values of η. Instead, we only have to find an consistent estimator $\hat{\eta}$ such that

$$\hat{\eta} \xrightarrow{p} \eta \ \text{ for all } \ \eta \in \Theta_{L,\eta} \triangleq \Big\{\eta : \eta = \tilde{\eta}(\phi), (\phi, \tilde{\eta}(\phi)) \in \Theta\Big\} \subset \Theta_{\cdot\eta}, \tag{4.1}$$

where $\tilde{\eta}(\phi)$ is a specific curve call the least-favorable curve defined as follows.

Definition 4.7 (Least-favorable Curves). For every given ϕ within its domain, $\tilde{\eta}(\phi)$ is the point such that the KL-divergence $KL(f(x|\phi, \tilde{\eta}(\phi))\|f(x|\phi_0, \eta_0))$ is minimized among all $\eta \in \Theta_{\cdot,\eta}$, i.e.,

$$\begin{aligned}\tilde{\eta}(\phi) &= \arg \min_{\eta:(\phi,\eta)\in\Theta} KL\big(f(x|\phi,\eta)\|f(x|\phi_0,\eta_0)\big)\\ &= \arg \min_{\eta\in\Theta_\phi} -\int \left(\log \frac{f(x|\phi,\eta)}{f(x|\phi_0,\eta_0)}\right) \cdot f(x|\phi_0,\eta_0)dx.\end{aligned}$$

Specially, we have $\tilde{\eta}(\phi_0) = \eta_0$.

Theorem 4.8 (♣ **Asymptotic Behavior of the Generalized Profile Likelihood Estimator**). *For any least-favorable curve $\tilde{\eta}(\phi)$ and assume $\hat{\eta}(X, \phi)$ is a consistent estimator of $\tilde{\eta}(\phi)$, i.e., (4.1) holds, then under "Regularity Condition $\mho'$", the maximum generalized profile log-likelihood estimator $\hat{\phi}_p$ is asymptotically normal and efficient, i.e.,*

$$\sqrt{n}\Big(\hat{\phi}_p - \phi_0\Big) \xrightarrow{d} N(0, I(\phi_0)^{-1}).$$

*A "Regularity Condition $\mho'$" and the proof of Theorem.4.8 is provided in Severini and Wong (1992).

Remark 4.9.

- Notice that we use the phrase ”for any least-favorable curve” because the least-favorable curve is not necessarily unique.
- Intuitively, for each given ϕ within its domain, when the KL-divergence is minimized, meaning there aren't too much "differences" between these two curves and foreseeably this is the most difficult case to estimate parameter values correctly, and that's why it's called the "least-favorable curve".
- For a more comprehensive introduction of the generalized profile likelihood approach, we refer to Fan and Wong (2000).

5. *Supplement

Here we list some of the conditions and proofs of theorems listed before.

Condition 5.1 (♣ **Condition for Wilks' Theorem**). The condition of Wilks' Theorem are tedious but can be separated into several parts.

(i) The first part is the consistency condition of MLE,

(a) The overall parameter $\theta = (\phi, \eta)$ satisfy the separation condition, i.e., for $\forall \epsilon > 0$,

$$\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \mathbb{E}\ell(\theta) < \mathbb{E}\ell(\theta_0). \tag{5.1}$$

(b) $\frac{1}{n}\ell(\theta)$ converges uniformly to its mean $\frac{1}{n}\mathbb{E}\ell(\theta)$ in probability, i.e.,

$$\sup_{\theta \in \Theta} \frac{1}{n} |\ell(\theta) - \mathbb{E}\ell(\theta)| \xrightarrow{p} 0.$$

Therefore, we will have $\hat{\theta} = (\hat{\phi}, \hat{\eta}) \xrightarrow{p} \theta_0 = (\phi_0, \eta_0)$, and $\hat{\eta}_{\phi_0} \xrightarrow{p} \eta_0$ according to the consistency of MLE (Theorem 1.3 of Notes 4).

(ii) The second part is the CAN condition of MLE, for $\forall \theta \in \Theta$, we have

(c) θ_0 is in the interior of Θ.

(d) $f(x|\theta)$ has a common support $\mathcal{X}$ who does not depend on θ.

(e) The partial derivative operator can exchange with the integral operator, i.e., the second Bartlett's Identity holds (for each term of the vector/matrix),

$$\frac{\partial^i}{\partial \theta^i} \mathbb{E}\Big[\log f(x|\theta)\Big] = \mathbb{E}\Big[\frac{\partial^i \log f(x|\theta)}{\partial \theta^i}\Big], \quad \text{for } i = 1, 2.$$

(f) the derivative of score function $s'(\theta|x)$ (i.e., $\ell''(\theta)$) is continuous in θ uniformly for all $x \in \mathcal{X}$.

(g) $0 < I(\theta) < \infty$ is well defined.

(iii) The third part is the uniform boundness of the derivative in the neighborhood of θ_0,

(h) There exists some $\epsilon_0 > 0$, such that within the neighborhood $\{\theta : |\theta - \theta_0| \leq \epsilon_0\}$, the derivative

$$\left\| \frac{\partial^i \log f(x|\theta)}{\partial \theta^i} \right\|_\infty \leq M, \quad \text{for } i = 1, 2, 3, \quad \text{and for } \forall x \in \mathcal{X}.$$

□

Proof of Lemma.4.1. By Taylor expansion, we have

$$\frac{1}{n} \sum_{i=1}^n \Big[\ell(\phi, \hat{\eta}|X_i) - \ell(\phi, \eta|X_i)\Big] = \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \ell(\phi, \eta|X_i)}{\partial \eta}\right] (\hat{\eta} - \eta)$$

$$+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\phi,\tilde{\eta}|X_i)}{\partial\eta\partial\eta^T}\right]\left(\hat{\eta}-\eta\right)^2. \tag{5.2}$$

for some $\tilde{\eta}$ lays in between of $\hat{\eta}$ and η. Now, since $\ell''(\theta)$ is continuous in θ uniformly for all $x \in \mathcal{X}$ according to (f) of Condition.5.1, so for arbitrary $\theta = (\phi,\eta) \in \Theta$ and arbitrary $\delta > 0$, there exists some $\epsilon_1 > 0$, s.t.,

$$\|\hat{\eta}-\eta\|_2 \le \epsilon_1 \quad \Rightarrow \quad \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\phi,\tilde{\eta}|X_i)}{\partial\eta\partial\eta^T} - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\phi,\eta|X_i)}{\partial\eta\partial\eta^T}\right\|_2 \le \frac{\delta}{2}. \tag{5.3}$$

Meanwhile, according to law of large numbers, for arbitrary $\epsilon > 0$, there exists an integer N_1, such that for $\forall n \ge N_1$,

$$\mathbb{P}\left(\left\|-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\phi,\eta|X_i)}{\partial\eta\partial\eta^T} - I(\phi,\eta)\right\|_2 \le \frac{\delta}{2}\right) \ge 1-\frac{\epsilon}{2}. \tag{5.4}$$

Besides, $\hat{\eta} \xrightarrow{p} \eta$ implies that there exists an integer N_2, such that for $\forall n \ge N_2$,

$$\mathbb{P}\Big(\|\hat{\eta}-\eta\|_2 \le \epsilon_1\Big) \ge 1-\frac{\epsilon}{2}. \tag{5.5}$$

Combining (5.3) (5.5) we conclude

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\phi,\tilde{\eta}|X_i)}{\partial\eta\partial\eta^T} \xrightarrow{p} -I(\phi,\eta). \tag{5.6}$$

Therefore, by applying Slutsky's theorem to (5.2), using (5.6), the fact that $(\hat{\eta}-\eta) \xrightarrow{p} 0$ and the weak law of large numbers, we conclude the result of Lemma.4.1.

□

References

Chernoff, H. (1954). On the distribution of the likelihood ratio. The Annals of Mathematical Statistics, 573-578.

Fan, J., & Wong, W. H. (2000). On profile likelihood: Comment. Journal of the American Statistical Association, 95(450), 468-471.

Severini, T. A., & Wong, W. H. (1992). Profile likelihood and conditionally parametric models. The Annals of statistics, 1768-1802.

Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. The annals of mathematical statistics, 9(1), 60-62.

Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York; 2d printing, corrected, 1963.