

Lecture 4. Likelihood, Quasi-Likelihood & Their Variants

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

1. Likelihood

$\frac{1}{n} \sum_{i=1}^n \log f(x_i | \theta)$
 $= \log L(\theta | X)$
 $= \log L(\theta)$

Definition 1.1 (♣ Likelihood, Log-Likelihood and the Maximum Likelihood Estimator). Let $X = \{X_1, \dots, X_n\}$ with distribution function $X \sim f(x|\theta)$, $\theta \in \Theta$. The corresponding likelihood function is defined as

$$L(\theta|x) = f(x|\theta)$$

and the corresponding log-likelihood function is $\ell(\theta) = \ell(\theta|x) = \log L(\theta) = \log L(\theta|x)$. The maximum likelihood estimator (MLE) of θ is

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta|X) = \arg \max_{\theta \in \Theta} \ell(\theta|X).$$

Now, assume $X = \{X_1, \dots, X_n\}$ is a random sample draw from the distribution function $f(x|\theta_0) \in \mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$, where θ_0 is the unknown underlying true parameter value of the data generating process. Denote $\hat{\theta}_{MLE}$ as the MLE of θ_0 , then

Condition 1.2 (♣ Consistency Condition of MLE). (i) Θ is compact or we have the separation condition, i.e., for $\forall \epsilon > 0$,

$$\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \mathbb{E} \ell(\theta) < \mathbb{E} \ell(\theta_0). \quad (1.1)$$

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

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

Theorem 1.3 (♣ Consistency of MLE). Under condition.1.2, $\hat{\theta}_{MLE}$ is a consistent estimator of θ_0 , i.e., $\hat{\theta}_{MLE} \xrightarrow{P} \theta_0$.

A proof of Theorem.1.3 is postponed to Supplement.

Further, if we assume the following

Condition 1.4 (♣ CAN Condition of MLE). For $\forall \theta \in \Theta$, we have

- (i) θ_0 is in the interior of Θ .
- (ii) $f(x|\theta)$ has a common support \mathcal{X} who does not depend on θ .
- (iii) The partial derivative operator can exchange with the integral operator,

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

- (iv) the derivative of score function $s'(\theta|x)$ (i.e., $\ell''(\theta)$) is continuous in θ .
- (v) $0 < I(\theta) < \infty$ is well defined.

Theorem 1.5 (♣ Consistency and Asymptotic Normality Property of MLE). Under condition.1.2 and condition.1.4, $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$.

Proof. Since θ_0 is in the interior of Θ , $\hat{\theta}_{MLE}$ is a consistent estimator of θ_0 (Thm.1.3), and thus it is also in the interior of Θ so for large enough n , which means $\ell'(\hat{\theta}_{MLE}) = 0$. By Taylor expansion, we have

$$0 = \ell'(\hat{\theta}_{MLE}) = \ell'(\theta_0) + \ell''(\tilde{\theta})(\hat{\theta}_{MLE} - \theta_0),$$

where $\tilde{\theta}$ lays in between θ_0 and $\hat{\theta}_{MLE}$. Again, since $\ell''(\theta)$ is continuous in θ and $\hat{\theta}_{MLE}$ is a consistent estimator of θ_0 , therefore, $|\ell''(\tilde{\theta}) - \ell''(\theta_0)|/n \xrightarrow{p} 0$. By (iii) of Condition.1.4 and central limit theorem, we have

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(X_i|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \xrightarrow{d} N(\mathbb{E}[s(\theta_0|X_1)], I(\theta_0)) \triangleq N(0, I(\theta_0)),$$

and by (iii) of Condition.1.4 and the weak law of large number,

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(X_i|\theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \xrightarrow{p} \mathbb{E}\left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \Big|_{\theta=\theta_0}\right] = -I(\theta_0).$$

Overall, by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{\frac{1}{n}\ell''(\tilde{\theta})} \xrightarrow{d} N(0, I(\theta_0)^{-1}).$$

□

Definition 1.6. Assume $\ell(\theta)$ is the log-likelihood function of θ , then for any function $g(\theta)$, the log-likelihood function of $g(\theta)$ is defined as

$$\ell_g(\phi) \triangleq \max_{\theta \in \Theta, g(\theta) = \phi} \ell(\theta).$$

Theorem 1.7 (♣ Invariance Principle of MLE). If $\hat{\theta}_{MLE}$ is the MLE of θ , then for any function $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta}_{MLE})$.

A proof of Theorem.1.7 is postponed to Supplement.

Remark 1.8. MLE does not have to exist. For example, suppose $X = \{X_1, \dots, X_n\}$ is a random sample drawn from $\text{Uniform}(0, \theta)$, then the MLE of θ does not exist. This is commonly known as the boundary problems. A simple way to get around this barrier is to define $\hat{\theta}_{MLE}$ slightly different as

$$\hat{\theta}_{MLE} = \max_{\theta \in \bar{\Theta}} \ell(\theta), \quad \text{where } \bar{\Theta} \text{ is the closure of } \Theta.$$

Here in notes, for illustration purpose, we simplify the situation and always assume MLE exists.

• *Example 1.9 (♣ MLE in Multinomial Distribution).* Suppose there are k scenarios each with probability p_i to occur respectively, $1 \leq i \leq k$, i.e.

$$A_1, \dots, A_k \text{ are disjoint with } \Omega = \cup_{i=1}^n A_i, \text{ and } \mathbb{P}(A_i) = p_i, \text{ for } 1 \leq i \leq k,$$

Apparently, $\sum_{i=1}^k p_i = 1$. We repeat the trial independently for n times. Each time j , $1 \leq j \leq n$, we observe one of the scenarios happens and record $Y_j = (Y_{j1}, \dots, Y_{jk}) = (0, \dots, 0, 1, 0, \dots, 0)_{k \times 1}$ with s -th element being 1 if s -th element occurs. We denote X_i as the number of times A_i occurs, so we have the sample $X = (X_1, \dots, X_k) = (x_1, \dots, x_k)$ and knowing that $X = (X_1, \dots, X_k)$ follows a multinomial distribution. Please find the MLE for p_1, \dots, p_k and please derive the CAN property of the MLE.

• *Example 1.10 (MLE does not have to be unique).* Let $X = \{X_1, \dots, X_n\}$ be a random sample drawn from $\text{Uniform}[\theta, \theta + 1]$. Please find the MLE of θ .

• *Example 1.11 (♣ MLE in Generalized Linear Models)*. Assume the sample $X = \{X_1, \dots, X_n\}$ has mutually independent observations X_1, \dots, X_n . Each X_i , has probability distribution function

$$f_{X_i}(x_i|\eta_i, \phi_i) = \exp\left(\frac{\eta_i x_i - \zeta(\eta_i)}{\phi_i}\right) \cdot h(x_i, \phi_i),$$

where $\{\phi_i > 0, 1 \leq i \leq n\}$ is called the dispersion parameters and here we assume there are known for simplicity. Apparently, we have f_{X_i} forms an exponential family and by the differentiation of exponential family density, we have

$$0 = \frac{\partial}{\partial \eta_i} \int f_{X_i}(x_i|\eta_i) dx_i = \int \frac{\partial f_{X_i}(x_i|\eta_i)}{\partial \eta_i} dx_i = \frac{1}{\phi_i} [\mathbb{E}X_i - \zeta'(\eta_i)].$$

Hence $\mathbb{E}X_i = \zeta'(\eta_i)$ for $1 \leq i \leq n$. It's assumed that the mean can be explained by known covariates (effects) Z_i and interested effect size β as

$$g(\mathbb{E}X_i) = g(\zeta'(\eta_i)) = \beta^T Z_i, \quad i = 1, \dots, n$$

for some known link function g , here we consider $g \circ \zeta'$ to be the identical function for simplicity, i.e., $\eta_i = \beta^T Z_i$. If ζ is smooth and convex (i.e., $\zeta''(x) > 0$ for all x in its domain), please give the MLE of β .

Answer. Notice that the log-likelihood is given by

$$\ell(\beta) = \sum_{i=1}^n \left[\log h(x_i, \phi_i) + \frac{\beta^T Z_i x_i - \zeta(\beta^T Z_i)}{\phi_i} \right]$$

by differentiate with respect to β , we have

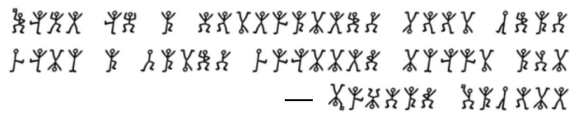
$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^n \frac{1}{\phi_i} (x_i - \zeta'(\beta^T Z_i)) Z_i, \quad \text{and} \quad \frac{\partial^2 \ell(\beta)}{\partial \beta^2} = - \sum_{i=1}^n \frac{\zeta''(\beta^T Z_i)}{\phi_i} Z_i Z_i^T.$$

Since $\frac{\partial^2 \ell(\beta)}{\partial \beta^2}$ is negative definite, so the $\hat{\beta}$ satisfy

$$\left. \frac{\partial \ell(\beta)}{\partial \beta} \right|_{\beta=\hat{\beta}} = \sum_{i=1}^n \frac{1}{\phi_i} (x_i - \zeta'(\hat{\beta}^T Z_i)) Z_i, \quad \text{and} \quad \hat{\beta}^T Z_i = \eta_i \in \Theta_i,$$

will be the MLE of this generalized linear model. □

2. Composite Likelihood



In complicated statistical models, the likelihood functions are sometimes intractable but low-dimensional distributions are readily computable. Therefore, we combine low dimensional terms to construct a substitution, which is called the composite likelihood.

Composite likelihood are sometimes refers to as pseudo-likelihood or approximate likelihood, we refer to [Varin, Reid and Firth \(2011\)](#) for a more comprehensive introduction on the history and general development of composite likelihood. Rigorously speaking,

Definition 2.1 (♣ Composite Likelihood). For a sample X with joint density $f(x|\theta)$, $\theta \in \Theta$. Denote by $\{\mathcal{A}_1, \dots, \mathcal{A}_K\}$ a set of marginal or conditional events with associated likelihoods

$$L_k(\theta | x) \propto f(x \in \mathcal{A}_k | \theta)$$

the composite log-likelihood is defined as the weighted sum

$$\mathcal{C}\ell(\theta | x) = \sum_{k=1}^K \omega_k \log L_k(\theta | x),$$

where $\{\omega_k\}_{1 \leq k \leq K}$, are non-negative weights to be further specified, and the particular used collection of set $\{\mathcal{A}_1, \dots, \mathcal{A}_K\}$ is often determined by the context.

Since the rigorous definition can be elusory, so we look at these following examples for understanding the composite likelihood.

2.1. Composite Marginal Likelihoods

• **Example 2.2 (♣ Sensitivity and Specificity Evaluation for Multiple Diagnostic Tests).** As the advancement of biomedical research, multiple tests may be available for diagnosis of a certain disease, but the accuracies and costs of tests could be different. In order to improve the accuracy of disease assessment and to minimize medical cost, it's important to quantify and compare accuracies of tests. For $i = 1, \dots, K$ diagnostic tests, $j = 1, \dots, m$ patients, we denote

$$Y_{ij} = \mathbb{1}(\text{The } j - \text{th patient in } i - \text{th test is tested positive})$$

$$Y'_j = \mathbb{1}(\text{The } j - \text{th patient in is truely positive})$$

Then conditional on the sensitivity $\alpha = (\alpha_1, \dots, \alpha_K)$, specificity $\beta = (\beta_1, \dots, \beta_K)$ of each diagnostic test, and the prevalence π , for the diagnostic test i , the distribution of the simple random effect model of Y_{ij} and Y'_j can be written as

$$\begin{aligned} \mathbb{P}(y_{ij}, y'_j | \alpha_i, \beta_i, \pi) &= \left(\alpha_i \pi \right)^{y_{ij} y'_j} \left[\beta_i (1 - \pi) \right]^{(1 - y_{ij})(1 - y'_j)} \\ &\quad \times \left[(1 - \alpha_i) \pi \right]^{(1 - y_{ij}) y'_j} \left[(1 - \beta_i) (1 - \pi) \right]^{y_{ij} (1 - y'_j)}. \end{aligned}$$

If we further assume

$$(\alpha, \beta, \pi)^T \sim N((\alpha_0, \beta_0, \pi_0)^T, \Sigma)$$

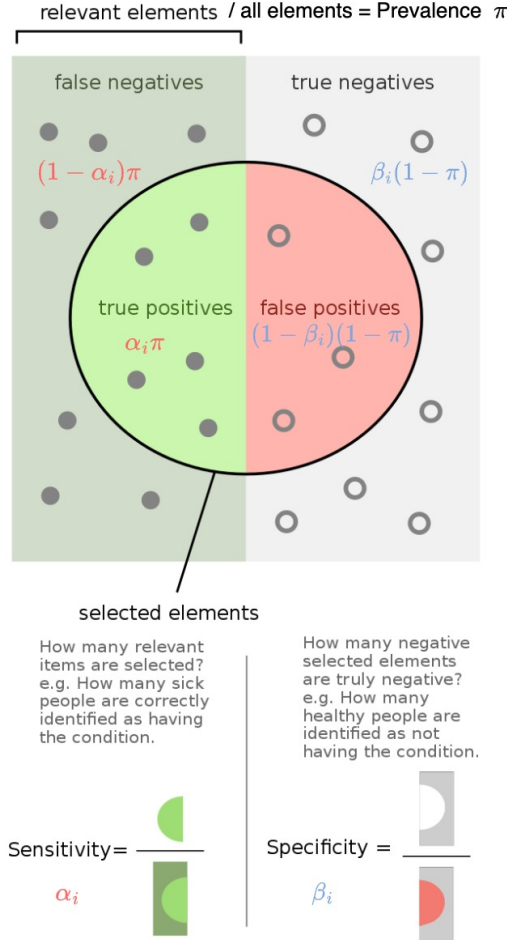


Figure 1: Sensitivity, Specificity and Prevalence. Pic from Wikipedia

where we are interested in each diagnostic tests' accuracy (the mean parameter $(\alpha_0, \beta_0, \pi_0)^T$) and their reliability (the diagonal term of Σ , i.e., $\text{diag}(\Sigma) = (\sigma_{\alpha_1}^2, \dots, \sigma_{\alpha_K}^2, \sigma_{\beta_1}^2, \dots, \sigma_{\beta_K}^2, \sigma_{\pi}^2)$). **Then, instead of write of the whole likelihood, we can see we may just use the marginal likelihood cause it's much more traceable**, i.e., one composite log-likelihood is given by

$$\mathcal{C}\ell(\alpha, \beta, \pi | Y, Y') = \sum_{j=1}^m \sum_{i=1}^K \left\{ y_{ij} y'_j \log(\alpha_i \pi) + (1 - y_{ij})(1 - y'_j) \log[\beta_i(1 - \pi)] \right\}$$

$$\begin{aligned}
& + (1 - y_{ij})y'_j \log \left[(1 - \alpha_i)\pi \right] + y_{ij}(1 - y'_j) \log \left[(1 - \beta_i)(1 - \pi) \right] \\
& + \frac{1}{m} \left[-\frac{1}{2} \log \left(\det \Sigma_i \right) - \frac{1}{2} \begin{pmatrix} \alpha_i - \alpha_{0i} \\ \beta_i - \beta_{0i} \\ \pi - \pi_0 \end{pmatrix}^T \cdot \Sigma_i^{-1} \cdot \begin{pmatrix} \alpha_i - \alpha_{0i} \\ \beta_i - \beta_{0i} \\ \pi - \pi_0 \end{pmatrix} \right] \Bigg\}
\end{aligned}$$

where

$$\Sigma_i = \begin{pmatrix} \sigma_{\alpha i}^2 & \text{Cov}(\alpha_i, \beta_i) & \text{Cov}(\alpha_i, \pi) \\ \text{Cov}(\alpha_i, \beta_i) & \sigma_{\beta i}^2 & \text{Cov}(\beta_i, \pi) \\ \text{Cov}(\alpha_i, \pi) & \text{Cov}(\beta_i, \pi) & \sigma_{\pi}^2 \end{pmatrix}$$

is the covariance matrix of (α_i, β_i, π) . Notice this is just the summation of log-arithm of the marginal distributions, and we have greatly reduced the number of parameters we have to estimate.

• *Example 2.3 (♣ Contingency Tables)*. A number n of subjects is drawn at random from a population sufficiently large that the drawings can be considered to be independent. Each subject is classified according to two characteristics: A , with possible outcomes A_1, \dots, A_I , and B , with possible outcomes B_1, \dots, B_J . For example, students might be classified as being male or female ($I = 2$) and according to their average performance (A, B, C, D , or F ; $J = 5$). 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 IJ variables n_{ij} is an unrestricted multinomial distribution with number of categories $s = IJ - 1$, and the results of the sample can be represented in an $I \times J$ table, such as Table.1. Now, assume for an observation Y_k , the distribution of $Y_k \in A_i \cap B_j$, $1 \leq k \leq n$, is

$$\mathbb{P}(Y_k \in A_i \cap B_j) = p_{ij} = \exp \left\{ \sum_{\ell=1}^I \sum_{m=1}^J \theta_{\ell m} \mathbb{1}(\ell = i, m = j) \right\}$$

where $\theta_{\ell m} = \log p_{\ell m}$ and $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1$. Say we are only interested in the $p_{i+} = \sum_{j=1}^J p_{ij}$ and $p_{+j} = \sum_{i=1}^I p_{ij}$. **Therefore, instead of seeking the whole likelihood, we can looking for the marginal distribution**

$$\mathbb{P}(Y_k \in A_i) = \sum_{j=1}^J p_{ij} = \exp \left\{ \sum_{\ell=1}^I \theta_{\ell+}^* \mathbb{1}(\ell = i) \right\}$$

where $\theta_{\ell+}^* = \log p_{\ell+}$ and $\sum_{i=1}^I p_{i+} = 1$. Notice that, by using the composite log-likelihood

$$\mathcal{C}\ell(\theta) = \sum_{k=1}^n \left[\sum_{i=1}^I \theta_{i+}^* \mathbb{1}(Y_k \in A_i) + \sum_{j=1}^J \theta_{+j}^* \mathbb{1}(Y_k \in B_j) \right]$$

we reduced the number of parameters in our composite likelihood from $I \cdot J - 1$ to $I + J - 2$.

TABLE 1
 $I \times J$ Contingency Table

	B_1	\dots	B_J	Total
A_1	n_{11}	\dots	n_{1J}	n_{1+}
\vdots	\vdots	\ddots	\vdots	\vdots
A_I	n_{I1}	\dots	n_{IJ}	n_{I+}
Total	n_{+1}	\dots	n_{+J}	n

• *Example 2.4* (**♣ Autoregressive Regressive Time Series**). Assume we have a time series $X = \{X_1, \dots, X_n\}$ with $X_{k+1} = \phi X_k + \epsilon_{k+1}$, $k \geq 1$. Where we have $X_0 = \epsilon_0/\sqrt{1-\phi^2}$, $|\phi| < 1$ and $\{\epsilon_i\}_{i \geq 0}$ is the sequence of white noise with each having the normal distribution $N(0, \sigma^2)$. Therefore, it's trivial to see that

$$X_i \sim N(0, \sigma^2/(1-\phi^2)), \quad i = 1, \dots, n$$

The joint distribution of X is given by

$$\begin{aligned} f_X(x|\sigma^2, \phi) &= f_{X_1}(x_1|\sigma^2, \phi) \prod_{k=2}^n f_{X_k|X_{k-1}}(x_k|x_{k-1}, \sigma^2, \phi) \\ &= \frac{\sqrt{1-\phi^2}}{\sqrt{2\pi\sigma^2}} e^{-\frac{(1-\phi^2)x_1^2}{2\sigma^2}} \cdot \prod_{k=2}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k - \phi x_{k-1})^2}{2\sigma^2}} \end{aligned} \quad (2.1)$$

Say we are interested in estimating the variance of X_i , i.e., $\sigma^2/(1-\phi^2)$. Apparently, if we seek estimator from the whole likelihood, solving $\partial \log f_X(x|\sigma^2, \phi)/\partial \phi$ along is complicated enough. However, if we using the composite log-likelihood

$$\mathcal{C}\ell(\sigma^2, \phi) = \sum_{i=1}^n \left[-\frac{1}{2} \log \left(\frac{\sigma^2}{1-\phi^2} \right) - \frac{x_i^2(1-\phi^2)}{2\sigma^2} \right]$$

we can easily have an estimator for $\sigma^2/(1-\phi^2)$ as $\sum_{i=1}^n X_i^2/n$.

2.2. Composite Conditional Likelihoods

The composite conditional likelihoods discussed here also works mainly in cases where MLE's are difficult to compute. The idea follows roughly like this, consider $\theta = (\theta_1, \theta_2)$, θ_1 is the parameter of interests, and θ_2 is the nuisance parameter. Suppose that there is a sufficient statistic of θ_2 given by $T_2(X)$ for each fixed θ_1 . Then, by the sufficiency, the conditional distribution of X conditional on T_2 does not depend on the nuisance parameter θ_2 . A "MCLE" of θ_1 can then be obtained by maximizing this composite conditional likelihood and therefore reduce our workloads in estimating θ_1 . This method can be applied to the case

where the dimension of θ is considerably larger than the dimension of θ_1 , so that computing the unconditional MLE of θ is much more difficult than computing the MCLE of θ_1 . Note that the MCLE's are usually different from the MLE's.

• *Example 2.5.* We again revisit Example 2.4. Consider that our sample $X = \{X_1, \dots, X_n\}$ follows AR(1) time series model but this time with location parameter μ :

$$X_{k+1} - \mu = \phi(X_k - \mu) + \epsilon_{k+1}, \quad k = 0, \dots, n-1.$$

where $X_0 = \mu + \epsilon_0/\sqrt{1-\phi^2}$, $\mu \in \mathbb{R}$, $|\phi| < 1$ and $\{\epsilon_i\}_{i \geq 0}$ is the sequence of white noise with each having the normal distribution $N(0, \sigma^2)$. Let $\theta = (\mu, \rho, \sigma^2)$, apparently, the joint likelihood function is a slight modification of (2.1), and the whole log-likelihood is given by

$$\ell(\theta|x) = -\frac{n}{2} \log \left(\frac{\sigma^2}{1-\phi^2} \right) - \frac{1}{2\sigma^2} \left[(x_1 - \mu)^2(1-\phi^2) + \sum_{t=2}^n [x_t - \mu - \phi(x_{t-1} - \mu)]^2 \right]$$

The computation of the MLE is greatly simplified if we consider the composite conditional likelihood by conditional on $X_1 = x_1$, then the composite log-likelihood is given by

$$\mathcal{C}\ell(\theta|x) = -\frac{n-1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \left[\sum_{t=2}^n [x_t - \mu - \phi(x_{t-1} - \mu)]^2 \right]$$

and we can readily get the MCLE as

$$\hat{\phi} = \frac{\sum_{t=2}^n (X_t - \bar{X}_0)(X_{t-1} - \bar{X}_{-1})}{\sum_{t=2}^n (X_{t-1} - \bar{X}_{-1})^2}$$

where $(\bar{X}_{-1}, \bar{X}_0) = \left(\frac{1}{n-1} \sum_{t=2}^n X_{t-1}, \frac{1}{n-1} \sum_{t=2}^n X_t \right)$. and the MCLE of μ and σ^2 are, respectively,

$$\hat{\mu} = \frac{\bar{X}_0 - \hat{\phi}\bar{X}_{-1}}{1 - \hat{\phi}}, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=2}^n [X_t - \bar{X}_0 - \hat{\phi}(X_{t-1} - \bar{X}_{-1})]^2.$$

• *Example 2.6 (Graph Data).* Now let's consider the graph data. Say we have observations $Y = \{Y_1, \dots, Y_n\}$, $Z = \{Z_{ij}, 1 \leq i, j \leq n\}$, from a graph, each Y_i is the observed value of the i -th vertex, and

$$Z_{ij} = \mathbb{1}(i\text{-th vertex is connected to } j\text{-th vertex}).$$

A commonly seen undirectional graph data is given in Figure 2.

Now, consider a problem where Y_i represents the impact of i -th paper (i -th vertex) and conditional on all other vertices $Y_{-i} = Y \setminus Y_i$ (all other topic related papers), the distribution of $Y_i|Y_{-i}$ is given by

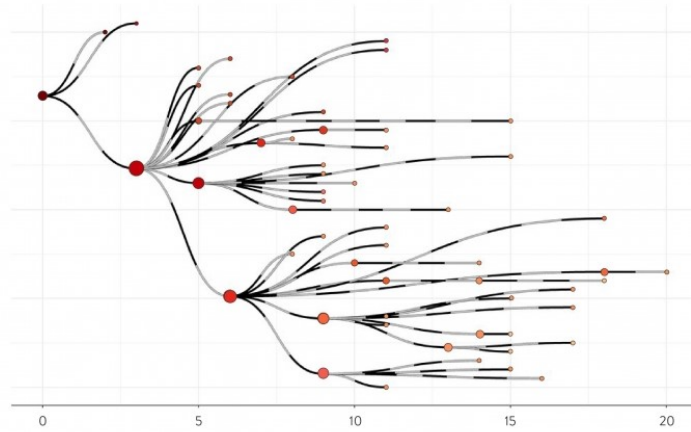
$$\mathbb{P}(Y_i \leq y_i | Y_{-i}, Z) = \frac{1}{1 + \exp \left(-\phi \cdot y_i + \sum_{j=1, j \neq i}^n \beta \cdot Y_j \cdot Z_{ij} \right)}.$$

$$\mathcal{C}\ell(\phi, \beta|Y, Z) = \sum_{i=1}^n \left[\log \phi - \phi \cdot y_i + \sum_{j=1, j \neq i}^n \beta \cdot y_j \cdot z_{ij} - 2 \log \left(1 + \exp \left(-\phi \cdot y_i + \sum_{j=1, j \neq i}^n \beta \cdot Y_j \cdot Z_{ij} \right) \right) \right].$$

This model can be used to tell other stories as well. For instance, we consider we have n listed company in the entertainment industry in NASDAQ, such as Bona Film Group, Sony, Universal Music, Warner Music, Netflix, NetEase, etc. We have Y_i represent the amount of the financing activity on the market of the i -th company, and its influenced by other companies' financing activity, who is in the same industry. Therefore, when other companies are financing more on

the market, then it will be hard for the i -th company to finance more (i.e., that is intuitively saying β would be negative). Or, if other companies are financing more on the market, causing the market to be irrational where hot money are desperate to get in, then i -th company would also be able to finance more (i.e., that is intuitively saying β would be positive).

- *Example 2.7 (Time-Since-Infection Model for Infectious Disease)*. Without confusion, we use subscript “ s ” and “ t ” to refer to the time since infection and the calendar time separately. We denote the total number of newly infected cases and hospital admission cases at time t as I_t and H_t , and define the filtration $\mathcal{F}_t = \sigma(\{I_r, 0 \leq r \leq t\})$ and $\mathcal{G}_t = \sigma(\{I_r, H_r, 0 \leq r \leq t\})$ which represents the information of past incident cases data and the past information of both incident cases and hospital admission data accordingly. Clearly, the disease transmission has a tree-type structure as shown in Figure.3.



Source: <https://www.derstandard.at/story/2000117352807/wie-sich-covid-19-in-oesterreich-ausbreitete>

Figure 3: Transmission Tree Structure

So a natural way to model the infectious disease transmission would be a point process, such as

$$I_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}\left(R_t \sum_{s=1}^t \omega_s I_{t-s}\right), \quad (2.2)$$

where $\omega = \omega_1, \dots$ are unknown parameters called the infectiousness profile.

Now, if we introduce $h_{t,s}$ to be the total number of patients who were infected at time t and are admitted at time $t + s$, i.e., admitted s time since infection, and we denote $h_{t,-1}$ to be the total number of patients infected at time t who

never go to hospitals, then a reasonable model for those “ h ” is

$$(h_{t,-1}, h_{t,0}, \dots, h_{t,\tilde{\eta}}) \mid I_t \sim \text{Multinomial}(I_t, \tilde{\omega}_{-1}, \tilde{\omega}_0, \dots, \tilde{\omega}_{\tilde{\eta}}). \quad (2.3)$$

where $\tilde{\omega} = \tilde{\omega}_{-1}, \dots$ are unknown parameters called the hospitalization profile.

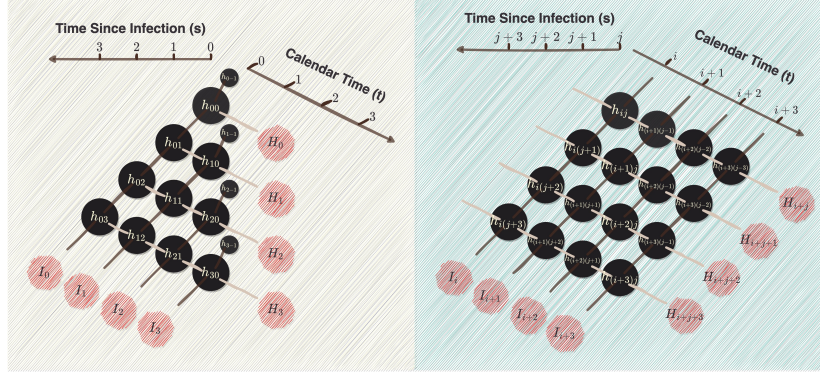


Figure 4: The relation among incident cases I_t and hospital admission data H_t .

Accordingly, we have a deeply connected and complicated model between our observation $\{(I_t, H_t)\}_{t \geq 1}$. If we look at the whole log-likelihood of the model

$$\ell = \sum_{0 \leq r \leq t} \log \mathbb{P}(H_r, I_r \mid \mathcal{G}_{r-1})$$

We found it being impossible to derive. But, if we look at the composite conditional distribution, we found

$$\begin{aligned} \mathcal{C}\ell &= \sum_{0 \leq r \leq t} \log \mathbb{P}(H_r, I_r \mid \mathcal{F}_{r-1}) \\ &= \sum_{0 \leq r \leq t} \sum_{h_{r-\tilde{\eta}}, \dots, h_{r-1,1}} \mathbb{P}\left(\text{Poisson}((1 - \tilde{\omega}_0)R_r\Lambda_r) = I_r - H_r + \sum_{s=1}^{\tilde{\eta}} h_{r-s,s}\right) \\ &\quad \cdot \mathbb{P}\left(\text{Poisson}(\tilde{\omega}_0 R_r\Lambda_r) = H_r - \sum_{s=1}^{\tilde{\eta}} h_{r-s,s}\right) \\ &\quad \cdot \prod_{s=1}^{\tilde{\eta}} \mathbb{P}\left(\text{Binomial}(I_{r-s}, \tilde{\omega}_s) = h_{r-s,s}\right) \mathbb{1}\left(H_r - I_r \leq \sum_{s=1}^{\tilde{\eta}} h_{r-s,s} \leq H_r\right). \end{aligned}$$

which is at least trackable.

2.3. Composite Marginal and Conditional Likelihoods

- *Example 2.8* (**♣ Generalized Proportional Likelihood Ratio Model**). Assume the outcome is Y , the covariate is X and the offset is Z , the proportional

likelihood ratio model assume the distribution of Y condition on X and Z is given by

$$f_{Y|X,Z}(Y = y|X = x, Z = z) = \frac{\exp(y(x^T\beta + z))g(y)}{\sum_{\tilde{y}=0}^{\infty} \exp(\tilde{y}(x^T\beta + z))g(\tilde{y})},$$

where β is the parameter of interests which can be interpreted as a “generalized log odds ratio” for a count outcome, and the function $g(\cdot)$ being a unknown nuisance paramter (unknown function). Since estimating using the whole likelihood is difficult because estimating $g(\cdot)$ would be tough task. As an alternative, for a independent sample $\{(Y_i, X_i, Z_i)\}_{1 \leq i \leq n}$, we consider a composite likelihood that is constitute of pairwise conditional likelihood, notice that

$$\begin{aligned} f(y_1, y_2|y_{(1)}, y_{(2)}, x_1, x_2, z_1, z_2) &= \frac{f(y_1|x_1, z_1)f(y_2|x_2, z_2)}{f(y_1|x_1, z_1)f(y_2|x_2, z_2) + f(y_1|x_2, z_2)f(y_2|x_1, z_1)} \\ &= \frac{\exp(y_1(x_1^T\beta + z_1)) \exp(y_2(x_2^T\beta + z_2))}{\exp(y_1(x_1^T\beta + z_1)) \exp(y_2(x_2^T\beta + z_2)) + \exp(y_1(x_2^T\beta + z_2)) \exp(y_2(x_1^T\beta + z_1))} \\ &= \frac{1}{1 + \exp\left[-(y_1 - y_2)\left((x_1 - x_2)^T\beta + (z_1 - z_2)\right)\right]}. \end{aligned}$$

Therefore, a much easier composite log-likelihood to deal with can be

$$\mathcal{C}\ell(\beta) = \sum_{i \neq j} -\log \left\{ 1 + \exp \left[-(y_i - y_j) \left((x_i - x_j)^T \beta + (z_i - z_j) \right) \right] \right\}.$$

2.4. Validity and Drawbacks of Composite Likelihood

2.4.1. Validity of Composite Likelihood

When introduced the composite likelihood, one question we have to ask is why does this likelihood works? Recall the definition of the composite likelihood

$$\mathcal{C}\ell(\theta | x) = \sum_{k=1}^K \omega_k \log L_k(\theta | x),$$

which is just a linear combination of some “valid” likelihood functions. Each of them will attained the maximum when the parameter θ is taking the underlying true value θ_0 under the expectation sense, i.e., the Shannon entropy is minimized when the parameter θ is taking the underlying true value θ_0 ,

$$H(\theta_0|X) = -\mathbb{E}_{\theta_0}[\log L_k(\theta_0|X)] \leq -\mathbb{E}_{\theta_0}[\log L_k(\theta|X)].$$

Therefore, intuitively speaking, this linear combination of some “valid” likelihood function would still (more likely to) attained its maximum

when the parameter θ is taking the underlying true value θ_0 . If we have multiple sets of independent sample (say M sets of sample with M being large), we would have

$$\sum_{i=1}^M \mathcal{C}\ell(\theta \mid X^{(i)}) = \sum_{i=1}^M \sum_{k=1}^K \omega_k \log L_k(\theta \mid X^{(i)}),$$

would lead us to a CAN MCLE under similar regularity condition.1.2 and condition.1.4. But since we don't have multiple sets of sample in reality for most of the time, so **the exact consistency and the asymptotic normality of the MCLE still needs to be derived case by case.**

2.4.2. Drawback of Composite Likelihood

- *Example 2.9 (Symmetric Normal Distribution Correlation Estimation).* Consider $Y = \{Y_1, \dots, Y_n\}$ be a random sample with each $Y_i = (Y_{i1}, \dots, Y_{iq})$ follows a normal distribution, and we have $\mathbb{E}Y_{ir} = 0$, $\text{Var } Y_{ir} = 1$ and $\text{Cov}(Y_{ir}, Y_{is}) = \rho$ for $i = 1, \dots, n$ and $1 \leq r, s \leq q$, $r \neq s$.

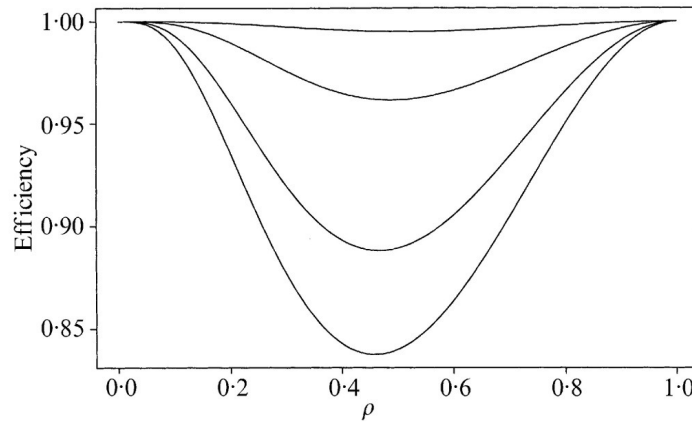


Fig. 5. Ratio of asymptotic variance of $\hat{\rho}$ to $\tilde{\rho}$, as a function of ρ , for fixed q . At $q = 2$ the ratio is identically 1. The lines shown are for $q = 3, 5, 8, 10$ (descending).

Denote the MLE of ρ based on the whole likelihood as $\hat{\rho}$, and the MCLE of ρ based on the following composite likelihood as $\tilde{\rho}$,

$$\mathcal{C}\ell(\rho|Y) = \sum_{i=1}^n \sum_{r>s} \log f_{Y_{ir}|Y_{is},\rho}(Y_{ir}|Y_{is},\rho)$$

$$= \sum_{i=1}^n \sum_{r>s} \left[C_0 - \frac{1}{2} \log(1 - \rho^2) - \frac{(Y_{ir} - \rho Y_{is})^2}{2(1 - \rho^2)} \right]$$

where C_0 is some constant. Cox and Reid (2004) considered the ratio of the asymptotic variance of $\hat{\rho}$ and $\tilde{\rho}$, turns out the ratio can get lower than 1 significantly when ρ is close to 0.5 and when the dimension q increases, which means the variation of our new estimator $\tilde{\rho}$ is clearly larger than the variation of $\hat{\rho}$, and known as the loss of efficiency. Therefore, they will be a tradeoff between using composite likelihood and estimation efficiency.

3. Quasi-Likelihood

$$\begin{aligned} & \frac{\partial}{\partial \theta} \log L(\theta) = \frac{1}{L(\theta)} \frac{\partial L(\theta)}{\partial \theta} = \frac{1}{L(\theta)} \sum_{i=1}^n \frac{\partial l_i(\theta)}{\partial \theta} \\ & = \frac{1}{L(\theta)} \sum_{i=1}^n \frac{\partial}{\partial \theta} \left(\log \pi(y_i | \theta) \right) = \frac{1}{L(\theta)} \sum_{i=1}^n \frac{\partial}{\partial \theta} \left(\log \pi(y_i) \right) \\ & = \frac{1}{L(\theta)} \sum_{i=1}^n \frac{\partial}{\partial \theta} \left(\log \pi(y_i) \right) = \frac{1}{L(\theta)} \sum_{i=1}^n \frac{\partial}{\partial \theta} \left(\log \pi(y_i) \right) \end{aligned}$$

3.1. General Definition of Quasi-Likelihood

Recall our procedure in calculating the MLE. It has always been going from

$$\text{Likelihood} \rightarrow \text{Log-Likelihood} \rightarrow \text{Score} \rightarrow \text{Estimating Equation}$$

So a natural question would be, can we skip specifying the likelihood and directly specify the score or estimating equation?

Definition 3.1 (\clubsuit General definition of the quasi-likelihood). Consider a random sample $Y = \{Y_1, \dots, Y_n\}$, an estimating equation is

$$\sum_{i=1}^n \psi(\theta | y_i) = 0,$$

with $\psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ being some known function. **Presumably this function ψ is the score function of some likelihood that we are just not bothering to find or specify it, and this unspecified likelihood associated with the above estimating equation is called the quasi-likelihood.**

The word "Quasi" refers to the fact that we could misspecify the estimating equation in which case it don't even has to be a score function that is corresponds to some distribution function.

• *Example 3.2.* Let us revisit Example 1.11, the generalized linear model. Say our observation $Y = \{Y_1, \dots, Y_n\}$, each Y_i has log-likelihood of an exponential dispersion family

$$\ell(\theta) = \frac{y\theta - \zeta(\theta)}{\phi}$$

Since this is an exponential family for arbitrary fixed dispersion parameter ϕ , so by the differentiation of exponential family, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int f_Y(y|\theta, \phi) dy = \int \frac{\partial}{\partial \theta} \exp\left(\frac{y\theta - \zeta(\theta)}{\phi}\right) \cdot h(y, \phi) dy = \frac{\mathbb{E}Y_i - \zeta'(\theta)}{\phi}, \\ 0 &= \frac{\partial^2}{\partial \theta^2} \int f_Y(y|\theta, \phi) dy = \int \frac{\partial^2}{\partial \theta^2} \exp\left(\frac{y\theta - \zeta(\theta)}{\phi}\right) \cdot h(y, \phi) dy = \frac{\text{Var } Y_i - \phi \zeta''(\theta)}{\phi^2}, \end{aligned}$$

so we have

$$\mathbb{E}Y_i = \zeta'(\theta) \triangleq \mu(\theta), \quad \text{Var } Y_i = \phi \zeta''(\theta) = \phi \cdot \nu(\mu(\theta)), \quad (3.1)$$

Notice that

$$\begin{aligned} &\mathbb{E}\left[(Y_i - \mu(\theta))^T \nu(\mu(\theta))^{-1} (Y_i - \mu(\theta))\right] \\ &= \mathbb{E}\left[\text{Tr} \left\{ (Y_i - \mu(\theta))^T \nu(\mu(\theta))^{-1} (Y_i - \mu(\theta)) \right\}\right] \\ &= \mathbb{E}\left[\text{Tr} \left\{ \nu(\mu(\theta))^{-1} (Y_i - \mu(\theta)) (Y_i - \mu(\theta))^T \right\}\right] \\ &= \text{Tr} \left[\nu(\mu(\theta))^{-1} \mathbb{E}\left\{ (Y_i - \mu(\theta)) (Y_i - \mu(\theta))^T \right\} \right] \\ &= \text{Tr} \left[\nu(\mu(\theta))^{-1} \cdot \phi \cdot \nu(\mu(\theta)) \right] = \phi, \end{aligned}$$

therefore, intuitively we have

$$\begin{aligned} 0 &\approx \sum_{i=1}^n \left\{ \frac{\partial}{\partial \theta} \left[(Y_i - \mu(\theta))^T \nu(\mu(\theta))^{-1} (Y_i - \mu(\theta)) \right] \right\} \\ &= \sum_{i=1}^n \left\{ \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu) + (Y_i - \mu)^T \frac{\partial \nu}{\partial \theta} (Y_i - \mu) \right\} \\ &= \sum_{i=1}^n \left\{ \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu) + o\left(\|Y_i - \mu\|_2^2\right) \right\}. \end{aligned}$$

Follow this intuition, quasi-likelihood directly construct our estimating equation and require that the estimator $\hat{\theta}$ satisfy the equation

$$0 = \sum_{i=1}^n \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu), \quad (3.2)$$

and the estimator is commonly called the M-estimator because we obtain it in a way we do for MLE.

Definition 3.3 (♣ Quasi-Likelihood approach (Generalized Estimating Equation)). For a random sample $Y = \{Y_1, \dots, Y_n\}$, assume

$$\mathbb{E}Y_i \triangleq \mu(\theta), \quad \text{Var } Y_i \triangleq \phi \cdot \nu(\mu(\theta)),$$

then we call

$$0 = \sum_{i=1}^n \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu)$$

the generalized estimating equation and obtaining an estimator by solving this equation refers as the quasi-likelihood approach.

Remark 3.4. Notice that we didn't make any distributional assumption in the definition of quasi-likelihood approach, even though the form of the generalized estimating equation is borrowed from the example of generalized linear model. It allows us to focusing on modeling the mean while the only real distributional assumption we make is the mean-variance relationship $\nu(\mu(\theta))$.

3.2. Miscellany on Quasi-Likelihood

Quasi-Likelihood approach is often invoked, first, when complex correlation structures arise in the model, such as in longitudinal data, spatial statistics and time series analysis. Specifying only the mean and variance will bring great simplicity; second, there might be different models that generates the same score function and therefore, focusing only on properties of the score allows us to consider a wider class of models.

The consistency and the asymptotic normality of quasi-likelihood estimator is proved in a way parallel to the MLE. The proof of the consistency requires slightly more efforts but the proof of asymptotic normality simply utilize the Taylor expansion after we conclude consistency. If we denote $\hat{\theta}$ and θ_0 as the estimator obtained based on generalized estimating equation and the underlying true value of the parameter, then

$$\begin{aligned} 0 &= \sum_{i=1}^n \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu) \Big|_{\theta=\hat{\theta}} \\ &= \sum_{i=1}^n \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu) \Big|_{\theta=\hat{\theta}_0} + \left[\frac{\partial^2 \mu}{\partial \theta^2} \nu^{-1} + \frac{\partial \mu}{\partial \theta} \frac{\partial \nu^{-1}}{\partial \theta} \right] \Big|_{\theta=\hat{\theta}} \left(\sum_{i=1}^n (Y_i - \mu) \right) (\tilde{\theta} - \theta_0) \\ &\quad - n \cdot \left(\frac{\partial \mu}{\partial \theta} \right)^2 \nu^{-1} \Big|_{\theta=\hat{\theta}} \cdot (\hat{\theta} - \theta_0) \end{aligned}$$

where $\tilde{\theta}$ lays in between $\hat{\theta}$ and θ_0 , if we admit consistency, then it leads to

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \mu}{\partial \theta} \nu^{-1}(Y_i - \mu) \Big|_{\theta=\hat{\theta}_0}}{- \left[\frac{\partial^2 \mu}{\partial \theta^2} \nu^{-1} + \frac{\partial \mu}{\partial \theta} \frac{\partial \nu^{-1}}{\partial \theta} \right] \Big|_{\theta=\hat{\theta}} \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \mu) \right) + \left(\frac{\partial \mu}{\partial \theta} \right)^2 \nu^{-1} \Big|_{\theta=\hat{\theta}}}$$

where the numerator is a partial sum of independent random variables and converge to $N(0, \left(\frac{\partial \mu}{\partial \theta} \right)^2 \phi \nu)$, while the first term in the denominator converges to

zero by weak law of large number, and the second term converges to $\left(\frac{\partial \mu}{\partial \theta}\right)^2 \nu^{-1} \Big|_{\theta=\theta_0}$, overall, hence

$$\left[\left(\frac{\partial \mu}{\partial \theta}\right)^2 / (\phi \cdot \nu) \right] \cdot \sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, 1).$$

However, Quasi-Likelihood approach can bring loss of efficiency just like the case in composite likelihood. Besides, when using Quasi-Likelihood, our hand will be tight in some well developed methods such as AIC and likelihood ratio tests.

4. *Supplement

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

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \rightarrow \mathbb{E} \left[\log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \right] = -KL(f_{\theta_0} \| f_{\theta}) < 0, \\ & \text{where the last inequality sign changes to the equal sign iff } f(x|\theta) \equiv f(x|\theta_0), a.s., \\ & \text{i.e., } \theta = \theta_0. \end{aligned}$$

Proof of Theorem 1.3. Notice that,

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} \ell(\theta) = \arg \max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \right] \triangleq \arg \max_{\theta \in \Theta} M(\theta)$$

Since we have

$$M(\theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \rightarrow \mathbb{E} \left[\log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \right] = -KL(f_{\theta_0} \| f_{\theta}) < 0,$$

where the last inequality sign changes to the equal sign iff $f(x|\theta) \equiv f(x|\theta_0)$, a.s., i.e., $\theta = \theta_0$. Therefore, if Θ is compact, then for $\forall \epsilon > 0$, we have

$$\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \mathbb{E} M(\theta) < \mathbb{E} M(\theta_0) = -KL(f_{\theta_0} \| f_{\theta_0}) = 0. \quad (4.1)$$

Similarly we can also conclude (4.1) under the separation condition (1.1). Now, if we denote

$$\delta = \mathbb{E} M(\theta_0) - \sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \mathbb{E} M(\theta) > 0,$$

Since $M(\theta)$ converges uniformly to $\mathbb{E} M(\theta)$ in probability according to (ii) of Condition 1.2. Therefore, there exists $N \in \mathbb{N}^+$ s.t. when $n \geq N$,

$$\mathbb{P} \left(|\hat{\theta}_{MLE} - \theta_0| \geq \epsilon \right) = \mathbb{P} \left(\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} M(\theta) > M(\theta_0) \right)$$

$$\begin{aligned}
&\leq \mathbb{P} \left(\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} [M(\theta) - \mathbb{E}M(\theta)] > [M(\theta_0) - \mathbb{E}M(\theta_0)] + \delta \right) \\
&\leq \mathbb{P} \left(\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} [M(\theta) - \mathbb{E}M(\theta)] > \frac{\delta}{2} \right) + \mathbb{P} \left([M(\theta_0) - \mathbb{E}M(\theta_0)] > -\frac{\delta}{2} \right) \\
&\leq 2\mathbb{P} \left(\sup_{\theta \in \Theta} |M(\theta) - \mathbb{E}M(\theta)| > \frac{\delta}{2} \right) \rightarrow 0.
\end{aligned}$$

Hence $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$. □

Proof of Theorem 1.7. We proof by contradiction. If there exist some other $\theta_0 \neq \hat{\theta}_{MLE}$ such that $\phi_0 = g(\theta_0) \neq g(\hat{\theta}_{MLE})$, and $\phi_0 = g(\theta_0)$ is the MLE of $g(\theta)$. Then

$$\ell_g(\phi_0) = \max_{\theta \in \Theta, g(\theta) = \phi_0} \ell(\theta) > \ell_g(g(\hat{\theta}_{MLE})) = \max_{\theta \in \Theta, g(\theta) = g(\hat{\theta}_{MLE})} \ell(\theta) = \ell(\hat{\theta}_{MLE}),$$

which contradict with the fact that $\hat{\theta}_{MLE}$ being the MLE of $\ell(\theta)$. Hence $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$. □

References

- Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3), 729-737.
- Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5-42.