Identifying the Most Appropriate Order for Categorical Responses

Tianmeng Wang and Jie Yang

University of Illinois at Chicago

Supplementary Material

Proof. of Theorem 2.1: The log-likelihood of the model at $\boldsymbol{\theta}_1$ with permuted responses $\mathbf{Y}_i^{\sigma_1}$ is

$$l_1(\boldsymbol{\theta}_1) = \sum_{i=1}^m \log(n_i!) - \sum_{i=1}^m \sum_{j=1}^J \log(Y_{i\sigma_1(j)}!) + \sum_{i=1}^m \sum_{j=1}^J Y_{i\sigma_1(j)} \log \pi_{ij}(\boldsymbol{\theta}_1)$$

while the log-likelihood at $\boldsymbol{\theta}_2$ with $\mathbf{Y}_i^{\sigma_2}$ is

$$l_2(\boldsymbol{\theta}_2) = \sum_{i=1}^m \log(n_i!) - \sum_{i=1}^m \sum_{j=1}^J \log(Y_{i\sigma_2(j)}!) + \sum_{i=1}^m \sum_{j=1}^J Y_{i\sigma_2(j)} \log \pi_{ij}(\boldsymbol{\theta}_2)$$

Since $\mathbf{Y}_i^{\sigma_1}$ and $\mathbf{Y}_i^{\sigma_2}$ are different only at the order of individual terms,

$$\sum_{j=1}^J \log(Y_{i\sigma_1(j)}!) = \sum_{j=1}^J \log(Y_{i\sigma_2(j)}!)$$

On the other hand, $\pi_{ij}(\boldsymbol{\theta}_2) = \pi_{i\sigma_2(\sigma_1^{-1}(j))}(\boldsymbol{\theta}_1)$ implies that

$$\begin{aligned}
\sum_{j=1}^{J} Y_{i\sigma_2(j)} \log \pi_{ij}(\boldsymbol{\theta}_2) &= \sum_{j=1}^{J} Y_{i\sigma_2(j)} \log \pi_{i\sigma_2(\sigma_1^{-1}(j))}(\boldsymbol{\theta}_1) \\
&= \sum_{j=1}^{J} Y_{ij} \log \pi_{i\sigma_1^{-1}(j)}(\boldsymbol{\theta}_1) \\
&= \sum_{j=1}^{J} Y_{i\sigma_1(j)} \log \pi_{ij}(\boldsymbol{\theta}_1)
\end{aligned}$$

Therefore, $l_1(\boldsymbol{\theta}_1) = l_2(\boldsymbol{\theta}_2)$, which implies $\max_{\boldsymbol{\theta}_1} l_1(\boldsymbol{\theta}_1) \leq \max_{\boldsymbol{\theta}_2} l_2(\boldsymbol{\theta}_2)$. Similarly, $\max_{\boldsymbol{\theta}_2} l_2(\boldsymbol{\theta}_2) \leq \max_{\boldsymbol{\theta}_1} l_1(\boldsymbol{\theta}_1)$. Thus $\max_{\boldsymbol{\theta}_1} l_1(\boldsymbol{\theta}_1) = \max_{\boldsymbol{\theta}_2} l_2(\boldsymbol{\theta}_2)$.

Given that (2.4) is true for σ_1 and σ_2, for any permutation $\sigma \in \mathcal{P}$,

$$\pi_{i\sigma_1^{-1}(\sigma^{-1}(j))}(\boldsymbol{\theta}_1) = \pi_{i\sigma_2^{-1}(\sigma^{-1}(j))}(\boldsymbol{\theta}_2)$$

for all i and j. That is, $\pi_{i(\sigma\sigma_1)^{-1}(j)}(\boldsymbol{\theta}_1) = \pi_{i(\sigma\sigma_2)^{-1}(j)}(\boldsymbol{\theta}_2)$ for all i and j. Following the same proof above, we have $\sigma\sigma_1 \sim \sigma\sigma_2$. □

Proof. of Theorem 2.2: We first show that $\sigma_1 = \mathrm{id}$, the identity permutation, and any permutation σ_2 satisfying $\sigma_2(J) = J$ are equivalent. Actually, given any $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T, \boldsymbol{\zeta}^T)^T$ for $\sigma_1 = \mathrm{id}$, we let $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{\sigma_2(1)}^T, \ldots, \boldsymbol{\beta}_{\sigma_2(J-1)}^T, \boldsymbol{\zeta}^T)^T$ for σ_2. Then $\eta_{ij}(\boldsymbol{\theta}_2) = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_{\sigma_2(j)} + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta} = \mathbf{h}_{\sigma_2(j)}^T(\mathbf{x}_i)\boldsymbol{\beta}_{\sigma_2(j)} + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta} = \eta_{i\sigma_2(j)}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J-1$. According to (2.2) and (2.3), $\pi_{ij}(\boldsymbol{\theta}_2) = \pi_{i\sigma_2(j)}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J$. Then $\mathrm{id} \sim \sigma_2$ is obtained by Theorem 2.1.

For general σ_1 and σ_2 satisfying $\sigma_1(J) = \sigma_2(J) = J$, $\mathrm{id} \sim \sigma_1^{-1}\sigma_2$ implies

$\sigma_1 \sim \sigma_2$ by Theorem 2.1. □

Proof. of Theorem 2.3: Given $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_{J-1}^T, \boldsymbol{\zeta}^T)^T$ with σ_1, we let $\boldsymbol{\theta}_2 = (-\boldsymbol{\beta}_{J-1}^T, -\boldsymbol{\beta}_{J-2}^T, \ldots, -\boldsymbol{\beta}_1^T, -\boldsymbol{\zeta}^T)^T$ for σ_2. Then $\eta_{ij}(\boldsymbol{\theta}_2) = -\mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_{J-j} - \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta} = -\mathbf{h}_{J-j}^T(\mathbf{x}_i)\boldsymbol{\beta}_{J-j} - \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta} = -\eta_{i,J-j}(\boldsymbol{\theta}_1)$ and thus $\rho_{ij}(\boldsymbol{\theta}_2) = 1 - \rho_{i,J-j}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J-1$. It can be verified that $\pi_{ij}(\boldsymbol{\theta}_2) = \pi_{i,J+1-j}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J$ according to (2.2) and (2.3). Then for all $i = 1, \ldots, m$ and $j = 1, \ldots, J$,

$$\begin{aligned}\pi_{i\sigma_2^{-1}(j)}(\boldsymbol{\theta}_2) &= \pi_{i,J+1-\sigma_2^{-1}(j)}(\boldsymbol{\theta}_1) = \pi_{i\sigma_1^{-1}(\sigma_1(J+1-\sigma_2^{-1}(j)))}(\boldsymbol{\theta}_1)\\ &= \pi_{i\sigma_1^{-1}(\sigma_2(\sigma_2^{-1}(j)))}(\boldsymbol{\theta}_1) = \pi_{i\sigma_1^{-1}(j)}(\boldsymbol{\theta}_1)\end{aligned}$$

That is, (2.4) holds given $\boldsymbol{\theta}_1$. Since it is one-to-one from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$, (2.4) holds given $\boldsymbol{\theta}_2$ as well. According to Theorem 2.1, $\sigma_1 \sim \sigma_2$. □

Proof. of Theorem 2.4: Similar as the proof of Theorem 2.3, for ppo models satisfying $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i)$, given $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_{J-1}^T, \boldsymbol{\zeta}^T)^T$ with σ_1, we let $\boldsymbol{\theta}_2 = (-\boldsymbol{\beta}_{J-1}^T, -\boldsymbol{\beta}_{J-2}^T, \ldots, -\boldsymbol{\beta}_1^T, -\boldsymbol{\zeta}^T)^T$ for σ_2 . Then $\eta_{ij}(\boldsymbol{\theta}_2) = -\mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_{J-j} - \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta} = -\mathbf{h}_{J-j}^T(\mathbf{x}_i)\boldsymbol{\beta}_{J-j} - \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta} = -\eta_{i,J-j}(\boldsymbol{\theta}_1)$ and thus $\rho_{ij}(\boldsymbol{\theta}_2) = 1 - \rho_{i,J-j}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J-1$. It can be verified that for $j = 1, \ldots, J-1$,

$$\prod_{l=j}^{J-1} \frac{\rho_{il}(\boldsymbol{\theta}_2)}{1 - \rho_{il}(\boldsymbol{\theta}_2)} = \prod_{l=j}^{J-1} \frac{1 - \rho_{i,J-l}(\boldsymbol{\theta}_1)}{\rho_{i,J-l}(\boldsymbol{\theta}_1)} = \prod_{l=1}^{J-j} \frac{1 - \rho_{il}(\boldsymbol{\theta}_1)}{\rho_{il}(\boldsymbol{\theta}_1)}$$

implies $\pi_{ij}(\boldsymbol{\theta}_2) = \pi_{i,J+1-j}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J$ due to (2.2) and (2.3). Then $\pi_{i\sigma_2^{-1}(j)}(\boldsymbol{\theta}_2) = \pi_{i\sigma_1^{-1}(j)}(\boldsymbol{\theta}_1)$ similarly as in the proof of Theorem 2.3, which leads to $\sigma_1 \sim \sigma_2$ based on Theorem 2.1. $\square$

Proof. of Theorem 2.5: According to Theorem 2.1, we only need to show that (2.4) holds for $\sigma_1 = \text{id}$ and an arbitrary permutation $\sigma_2 \in \mathcal{P}$.

Case one: $\sigma_2(J) = J$. In this case, for any $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T$, we let $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{\sigma_2(1)}^T, \ldots, \boldsymbol{\beta}_{\sigma_2(J-1)}^T)^T$. Then $\eta_{ij}(\boldsymbol{\theta}_2) = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_{\sigma_2(j)} = \mathbf{h}_{\sigma(j)}^T(\mathbf{x}_i)\boldsymbol{\beta}_{\sigma_2(j)}$ $= \eta_{i\sigma_2(j)}(\boldsymbol{\theta}_1)$, which leads to $\rho_{ij}(\boldsymbol{\theta}_2) = \rho_{i\sigma_2(j)}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J-1$. According to (2.2) in Lemma 1, $\pi_{ij}(\boldsymbol{\theta}_1) = \pi_{i\sigma_2^{-1}(j)}(\boldsymbol{\theta}_2)$, which is (2.4) in this case. Since the correspondence between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is one-to-one, then (2.4) holds for given $\boldsymbol{\theta}_2$ as well.

Case two: $\sigma_2(J) \neq J$. Given $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T$, we let $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{21}^T, \ldots, \boldsymbol{\beta}_{2,J-1}^T)^T$ such that for $j = 1, \ldots, J-1$,

$$\boldsymbol{\beta}_{2j} = \begin{cases} \boldsymbol{\beta}_{\sigma_2(j)} - \boldsymbol{\beta}_{\sigma_2(J)} & \text{if } j \neq \sigma_2^{-1}(J) \\ -\boldsymbol{\beta}_{\sigma_2(J)} & \text{if } j = \sigma_2^{-1}(J) \end{cases}$$

Then for $i = 1, \ldots, m$ and $j = 1, \ldots, J-1$,

$$\eta_{ij}(\boldsymbol{\theta}_2) = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_{2j} = \begin{cases} \eta_{i\sigma_2(j)}(\boldsymbol{\theta}_1) - \eta_{i\sigma_2(J)}(\boldsymbol{\theta}_1) & \text{if } j \neq \sigma_2^{-1}(J) \\ -\eta_{i\sigma_2(J)}(\boldsymbol{\theta}_1) & \text{if } j = \sigma_2^{-1}(J) \end{cases}$$

It can be verified that: (i) If $\sigma_2^{-1}(j) \neq J$ and $j \neq J$, then $\pi_{i\sigma_2^{-1}(j)}(\boldsymbol{\theta}_2) = \pi_{ij}(\boldsymbol{\theta}_1)$ according to (2.2); (ii) $\pi_{i\sigma_2^{-1}(J)}(\boldsymbol{\theta}_2) = \pi_{iJ}(\boldsymbol{\theta}_1)$ according to (2.2);

and (iii) if $\sigma_2^{-1}(j) = J$, then $\pi_{i\sigma_2^{-1}(j)}(\boldsymbol{\theta}_2) = \pi_{ij}(\boldsymbol{\theta}_1)$ according to (2.3). Thus (2.4) holds given $\boldsymbol{\theta}_1$. Given σ_2, the correspondence between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is one-to-one, then (2.4) holds given $\boldsymbol{\theta}_2$ as well. Thus id $\sim \sigma_2$ according to Theorem 2.1.

For general σ_1 and σ_2, id $\sim \sigma_1^{-1}\sigma_2$ implies $\sigma_1 \sim \sigma_2$ according to Theorem 2.1. □

Proof. of Theorem 2.6: Similar as the proof of Theorem 2.5, we first show that (2.4) holds for $\sigma_1 =$ id and an arbitrary permutation σ_2 .

For any $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T$ with $\sigma_1 =$ id, we let $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{21}^T, \ldots, \boldsymbol{\beta}_{2,J-1}^T)^T$ for σ_2, where

$$\boldsymbol{\beta}_{2j} = \begin{cases} \sum_{l=\sigma_2(j)}^{\sigma_2(j+1)-1} \boldsymbol{\beta}_l & \text{if } \sigma_2(j) < \sigma_2(j+1) \\ -\sum_{l=\sigma_2(j+1)}^{\sigma_2(j)-1} \boldsymbol{\beta}_l & \text{if } \sigma_2(j) > \sigma_2(j+1) \end{cases} \tag{S.1}$$

For $j = 1, \ldots, J-1$,

$$\eta_{ij}(\boldsymbol{\theta}_2) = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_{2j} = \sum_{l=\sigma_2(j)}^{\sigma_2(j+1)-1} \mathbf{h}_l^T(\mathbf{x}_i)\boldsymbol{\beta}_l = \sum_{\sigma_2(j)}^{\sigma_2(j+1)-1} \eta_{il}(\boldsymbol{\theta}_1)$$

if $\sigma_2(j) < \sigma_2(j+1)$; and $\eta_{ij}(\boldsymbol{\theta}_2) = -\sum_{l=\sigma_2(j+1)}^{\sigma_2(j)-1} \eta_{il}(\boldsymbol{\theta}_1)$ if $\sigma_2(j) > \sigma_2(j+1)$. It can be verified that (S.1) implies for any $1 \leq j < k \leq J-1$,

$$\sum_{l=j}^{k} \boldsymbol{\beta}_{2j} = \begin{cases} \sum_{l=\sigma_2(j)}^{\sigma_2(k+1)-1} \boldsymbol{\beta}_l & \text{if } \sigma_2(j) < \sigma_2(k+1) \\ -\sum_{l=\sigma_2(k+1)}^{\sigma_2(j)-1} \boldsymbol{\beta}_l & \text{if } \sigma_2(j) > \sigma_2(k+1) \end{cases} \tag{S.2}$$

Then for $j = 1, \ldots, J-1$,

$$\prod_{l=j}^{J-1} \frac{\rho_{il}(\boldsymbol{\theta}_2)}{1-\rho_{il}(\boldsymbol{\theta}_2)} = \exp\left\{\sum_{l=j}^{J-1} \eta_{il}(\boldsymbol{\theta}_2)\right\}$$

$$= \begin{cases} \prod_{l=\sigma_2(j)}^{\sigma_2(J)-1} \frac{\rho_{il}(\boldsymbol{\theta}_1)}{1-\rho_{il}(\boldsymbol{\theta}_1)} & \text{if } \sigma_2(j) < \sigma_2(J) \\ \frac{1}{\prod_{l=\sigma_2(J)}^{\sigma_2(j)-1} \frac{\rho_{il}(\boldsymbol{\theta}_1)}{1-\rho_{il}(\boldsymbol{\theta}_1)}} & \text{if } \sigma_2(j) > \sigma_2(J) \end{cases}$$

According to (2.2) and (2.3), it can be verified that $\pi_{ij}(\boldsymbol{\theta}_2) = \pi_{i\sigma_2(j)}(\boldsymbol{\theta}_1)$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, J$. That is, (2.4) holds given $\boldsymbol{\theta}_1$.

On the other hand, (S.2) implies an inverse transformation from $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_1$

$$\boldsymbol{\beta}_j = \sum_{l=\sigma_2^{-1}(j)}^{\sigma_2^{-1}(j+1)} \boldsymbol{\beta}_{2l} \tag{S.3}$$

with $j = 1, \ldots, J-1$. That is, it is one-to-one from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$. According to Theorem 2.1, id $\sim \sigma_2$.

Similar as in the proof of Theorem 2.5, we have $\sigma_1 \sim \sigma_2$ for any two permutations σ_1 and σ_2 . □

Proof. of Theorem 2.7: We frist verify condition (2.4) for $\sigma_1 = \text{id}$ and $\sigma_2 = (J-1, J)$. In this case, given $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-2}^T, \boldsymbol{\beta}_{J-1}^T)^T$ for σ_1, we let $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-2}^T, -\boldsymbol{\beta}_{J-1}^T)^T$ for σ_2. Then $\eta_{ij}(\boldsymbol{\theta}_2) = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j = \eta_{ij}(\boldsymbol{\theta}_1)$ for $j = 1, \ldots, J-2$, and $\eta_{i,J-1}(\boldsymbol{\theta}_2) = -\mathbf{h}_{J-1}^T(\mathbf{x}_i)\boldsymbol{\beta}_{J-1} = -\eta_{i,J-1}(\boldsymbol{\theta}_1)$. We further obtain $\rho_{ij}(\boldsymbol{\theta}_2) = \rho_{ij}(\boldsymbol{\theta}_1)$ for $j = 1, \ldots, J-2$ and $\rho_{i,J-1}(\boldsymbol{\theta}_2) = 1 - \rho_{i,J-1}(\boldsymbol{\theta}_1)$. According to (2.2) and (2.3), we obtain $\pi_{ij}(\boldsymbol{\theta}_2) = \pi_{ij}(\boldsymbol{\theta}_1)$ for

$j = 1, \ldots, J-2$; $\pi_{i,J-1}(\boldsymbol{\theta}_2) = \pi_{iJ}(\boldsymbol{\theta}_1)$; and $\pi_{iJ}(\boldsymbol{\theta}_2) = \pi_{i,J-1}(\boldsymbol{\theta}_1)$. That is, (2.4) holds given $\boldsymbol{\theta}_1$, which also holds given $\boldsymbol{\theta}_2$ since it is one-to-one from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$. According to Theorem 2.1, id $\sim (J-1, J)$ and thus $\sigma_1 \sim \sigma_1(J-1, J)$ for any permutation σ_1 . □

Proof. of Lemma 2: It is well known that for each $i = 1, \ldots, m$, $\left(\frac{Y_{i1}}{N_i}, \ldots, \frac{Y_{iJ}}{N_i}\right)$ maximizes $\sum_{j=1}^{J} Y_{ij} \log \pi_{ij}$ as a function of $(\pi_{i1}, \ldots, \pi_{iJ})$ under the constraints $\sum_{j=1}^{J} \pi_{ij} = 1$ and $\pi_{ij} \geq 0$, $j = 1, \ldots, J$ (see, for example, Section 35.6 of Johnson et al. (1997)). If $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ and $\hat{\sigma} \in \mathcal{P}$ satisfy $\pi_{i\hat{\sigma}^{-1}(j)}(\hat{\boldsymbol{\theta}}) = \frac{Y_{ij}}{N_i}$ for all i and j, then $\{\pi_{i\hat{\sigma}^{-1}(j)}(\hat{\boldsymbol{\theta}})\}_{ij}$ maximizes $\sum_{i=1}^{m} \sum_{j=1}^{J} Y_{ij} \log \pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta})$, which implies $(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ maximizes $l_N(\boldsymbol{\theta}, \sigma)$ and thus $l(\boldsymbol{\theta}, \sigma)$. □

Proof. of Lemma 3: According to the strong law of large numbers (see, for example, Chapter 4 in Ferguson (1996)), $\frac{N_i}{N} = N^{-1} \sum_{l=1}^{N} \mathbf{1}_{\{X_l = \mathbf{x}_i\}} \to E(\mathbf{1}_{\{X_l = \mathbf{x}_i\}}) = \frac{n_i}{n}$ almost surely, as $N \to \infty$, for each $i = 1, \ldots, m$. Since $n_0 \geq 1$, it can be verified that $\min\{N_1, \ldots, N_m\} \to \infty$ almost surely, as $N \to \infty$. Similarly, we have $\frac{Y_{ij}}{N_i} \to \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)$ almost surely, as $N_i \to \infty$, for each $i = 1, \ldots, m$ and $j = 1, \ldots, J$. Then as N goes to infinity,

$$\begin{aligned} N^{-1} l_N(\boldsymbol{\theta}_0, \sigma_0) &= \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{J} Y_{ij} \log \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{J} \frac{N_i}{N} \cdot \frac{Y_{ij}}{N_i} \log \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \\ &\xrightarrow{a.s.} \sum_{i=1}^{m} \sum_{j=1}^{J} \frac{n_i}{n} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \log \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{J} \frac{n_i}{n} \pi_{ij}(\boldsymbol{\theta}_0) \log \pi_{ij}(\boldsymbol{\theta}_0) \;=\; l_0 \;<\; 0 \end{aligned}$$

□

Proof. of Theorem 3.1: First we claim that for large enough N, all $i = 1, \ldots, m$ and $j = 1, \ldots, J$, $0 > \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) \geq \frac{2n(l_0-1)}{n_0\pi_0}$, which is a finite constant. Actually, since $(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N)$ is an MLE, we have $N^{-1} l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) \geq N^{-1} l_N(\boldsymbol{\theta}_0, \sigma_0)$ for each N. According to Lemma 3, $N^{-1} l_N(\boldsymbol{\theta}_0, \sigma_0) \to l_0$ almost surely, then $N^{-1} l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) > l_0 - 1$ for large enough N almost surely. On the other hand, since $\log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) < 0$ for all i and j, then

$$N^{-1} l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) = \sum_{i=1}^{m} \sum_{j=1}^{J} \frac{N_i}{N} \cdot \frac{Y_{ij}}{N_i} \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) < \frac{N_i}{N} \cdot \frac{Y_{ij}}{N_i} \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N)$$

for each i and j. Since $\frac{N_i}{N} \to \frac{n_i}{n}$ almost surely and $\frac{Y_{ij}}{N_i} \to \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)$ almost surely, then $\frac{N_i}{N} \cdot \frac{Y_{ij}}{N_i} \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) < \frac{1}{2} \cdot \frac{n_i}{n} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N)$ for large enough N almost surely. Then we have

$$0 > \frac{1}{2} \cdot \frac{n_i}{n} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) > l_0 - 1$$

almost surely for large enough N and each i and j. Since $l_0 - 1 < 0$, we further have almost surely for large enough N,

$$0 > \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) > \frac{l_0 - 1}{\frac{1}{2} \cdot \frac{n_i}{n} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)} \geq \frac{l_0 - 1}{\frac{1}{2} \cdot \frac{n_0}{n} \pi_0} = \frac{2n(l_0 - 1)}{n_0 \pi_0}$$

Now we are ready to check the asymptotic difference between $N^{-1} l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N)$ and $N^{-1} l_N(\boldsymbol{\theta}_0, \sigma_0)$. According to Lemma 2 and its proof, $(\boldsymbol{\theta}_0, \sigma_0)$ maximizes $\sum_{i=1}^m \sum_{j=1}^J \frac{N_i}{N} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \log \pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta})$. Then

$$\begin{aligned}
0 &\leq N^{-1}[l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) - l_N(\boldsymbol{\theta}_0, \sigma_0)] \\
&= \sum_{i=1}^m \sum_{j=1}^J \frac{N_i}{N} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) \\
&- \sum_{i=1}^m \sum_{j=1}^J \frac{N_i}{N} \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \log \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0) \\
&+ \sum_{i=1}^m \sum_{j=1}^J \frac{N_i}{N} \left[\frac{Y_{ij}}{N_i} - \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)\right] \cdot \left[\log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) - \log \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)\right] \\
&\leq \sum_{i=1}^m \sum_{j=1}^J \frac{N_i}{N} \left[\frac{Y_{ij}}{N_i} - \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)\right] \cdot \left[\log \pi_{i\hat{\sigma}_N^{-1}(j)}(\hat{\boldsymbol{\theta}}_N) - \log \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)\right]
\end{aligned}$$

Then for large enough N, we have almost surely

$$\begin{aligned}
&\frac{1}{N} |l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) - l_N(\boldsymbol{\theta}_0, \sigma_0)| \\
\leq &\sum_{i=1}^m \sum_{j=1}^J \left|\frac{Y_{ij}}{N_i} - \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)\right| \cdot \left[\frac{-2n(l_0 - 1)}{n_0 \pi_0} - \log \pi_0\right]
\end{aligned}$$

Since $\frac{Y_{ij}}{N_i} \to \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)$ almost surely for each i and j, then $N^{-1}|l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) - l_N(\boldsymbol{\theta}_0, \sigma_0)| \to 0$ almost surely as N goes to infinity. The rest parts of the theorem are straightforward. □

Proof. of Corollary 1: Since AIC $-$ AIC$(\boldsymbol{\theta}_0, \sigma_0)$ = BIC $-$ BIC$(\boldsymbol{\theta}_0, \sigma_0)$ = $-2l(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) + 2l(\boldsymbol{\theta}_0, \sigma_0) = -2l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) + 2l_N(\boldsymbol{\theta}_0, \sigma_0)$, the conclusion follows directly by Theorem 3.1. □

Bibliography

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman & Hall.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions.* John Wiley & Sons.