Online Supplementary Material for GRF Proposal 2019: *"Statistical Inference after Model Averaging"*

Alan WAN (P.I.), Xinyu ZHANG (Co-I.)

Summary

This document provides the proofs of the preliminary theoretical results of the captioned GRF Proposal.

1 Model framework and notations

Let y be generated from the density

$$f(y, \boldsymbol{\beta}, \boldsymbol{\gamma}) \tag{1.1}$$

on $\boldsymbol{\Omega}$, a measurable Euclidean space, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $q_1 \times 1$ and $q_2 \times 1$ vectors of unknown parameters. We consider inference on $\boldsymbol{\mu} = \boldsymbol{\mu}(f)$ when the unknowns are estimated by model averaging. Hjort & Claeskens (2003) considered the local misspecification framework

$$f_{true}(y) = f_n(y) = f(y, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n}),$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_0$ is fixed and known, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{q_2})$ contains parameters that signify the degrees of the model departures in directions $1, ..., q_2$. As discussed in the proposal, the local misspecification framework has the advantage of simplifying the asymptotic analysis but its realism has been subject to criticism. Here, we study the asymptotic distribution of model averaging estimators under the general fixed parameter setup (1.1) without invoking local misspecification.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ and $\boldsymbol{\theta}_0$ denote the true value of θ, with $\boldsymbol{\theta} \subset \boldsymbol{\Theta} \subset \mathbb{R}^q$. Define the likelihood function

$$L_n(\boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t, \boldsymbol{\theta}),$$

and the log likelihood function $l_n(\boldsymbol{\theta}) = \ln L_n(\boldsymbol{\theta})$. It is assumed that the first and second partial derivatives of $f(y, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist. Let

$$\Psi(y, \boldsymbol{\theta}) = \partial \ln f(y, \boldsymbol{\theta}) / \partial \boldsymbol{\theta},$$

and

$$\dot{\Psi}(y,\boldsymbol{\theta}) = \partial^2 \ln f(y,\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T.$$

Then the first and second partial derivatives of $l_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are $\dot{l}_n(\boldsymbol{\theta}) = \sum_{t=1}^n \Psi(y_t,\boldsymbol{\theta})$ and $\ddot{l}_n(\boldsymbol{\theta}) = \sum_{t=1}^n \dot{\Psi}(y_t,\boldsymbol{\theta})$ respectively, and the Fisher Information matrix is

$$\mathcal{F}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Psi(y,\boldsymbol{\theta})\Psi(y,\boldsymbol{\theta})^T.$$

We combine M sub-models within (1.1). Denote k_m as the number of parameters in the m^{th} sub-model. It is assumed that a sub-model contains all q_1 elements in $\boldsymbol{\beta}$ and some of the q_2 elements in $\boldsymbol{\gamma}$. If all possible combinations of elements in $\boldsymbol{\gamma}$ are considered, then $M = 2^{q_2}$; if only nested candidate models are considered, then $M = q_2 + 1$. Let $S = \{1,\ldots,q_2\}$, $S_m = \{i_1,\cdots,i_{k_m-q_1}\} \subset S$, $S_m^c = \{i_{k_m-q_1+1},\cdots,i_{q_2}\}$, the complement of S_m, and γ_j be the j^{th} element of $\boldsymbol{\gamma}$. Write $\boldsymbol{\gamma}_{S_m} = (\gamma_{i_1},\cdots,\gamma_{i_{k_m-q_1}})^T$, $\boldsymbol{\gamma}_{S_m^c} = (\gamma_{i_{k_m-q_1+1}},\cdots,\gamma_{i_{q_2}})^T$, and $\boldsymbol{\theta}_m = (\boldsymbol{\beta}^T,\boldsymbol{\gamma}_m^T)^T$, where $\boldsymbol{\gamma}_m = (\gamma_1,\gamma_2,\cdots,\gamma_{q_2})^T$ such that $\boldsymbol{\gamma}_{S_m^c} = \mathbf{0}$. Define the permutation matrix

$$\Pi_m = (\mathbf{e}_1^T,\cdots,\mathbf{e}_{q_1}^T,\mathbf{e}_{q_1+i_1}^T,\cdots,\mathbf{e}_{q_1+i_{k_m-q_1}}^T,\mathbf{e}_{q_1+i_{k_m-q_1+1}}^T,\cdots,\mathbf{e}_{q_1+i_{q_2}}^T)^T,$$

where $\mathbf{e}_j$ is a unit vector with the j^{th} element being 1 and all other elements zero. Then we can write $\boldsymbol{\theta}_m' = \Pi_m\boldsymbol{\theta}_m = (\boldsymbol{\beta}^T,\boldsymbol{\gamma}_{S_m}^T,\mathbf{0}^T)^T = (\boldsymbol{\theta}_{S_m}^T,\mathbf{0}^T)^T$, with $\boldsymbol{\theta}_{S_m} = (\boldsymbol{\beta}^T,\boldsymbol{\gamma}_{S_m}^T)^T$. Similarly, we have $\boldsymbol{\theta}_{0,m} = \Pi_m\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,S_m}^T,\boldsymbol{\gamma}_{0,S_m^c}^T)^T$, where $\boldsymbol{\theta}_{0,S_m}$ contains the first k_m elements of $\boldsymbol{\theta}_{0,m}$ and $\boldsymbol{\gamma}_{0,S_m^c}$ contains the remaining $q - k_m$ elements.

Denote

$$\mathcal{F}_m(\boldsymbol{\theta}_{0,m}) = E_{\boldsymbol{\theta}_{0,m}}\Psi(y,\boldsymbol{\theta}_{0,m})\Psi(y,\boldsymbol{\theta}_{0,m})^T = \Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T.$$

Assuming that partial derivatives exist, write

$$\begin{aligned}
\Psi_m(y,\boldsymbol{\theta}_{S_m}) &= \partial \ln f(y,\boldsymbol{\theta}_m)/\partial\boldsymbol{\theta}_{S_m},\\
\dot{\Psi}_m(y,\boldsymbol{\theta}_{S_m}) &= \partial^2 \ln f(y,\boldsymbol{\theta}_m)/\partial\boldsymbol{\theta}_{S_m}\boldsymbol{\theta}_{S_m}^T,\\
A_{m,n}(\boldsymbol{\theta}_{S_m}) &= n^{-1}\sum_{t=1}^n \dot{\Psi}_m(y_t,\boldsymbol{\theta}_{S_m}),\\
B_{m,n}(\boldsymbol{\theta}_{S_m}) &= n^{-1}\sum_{t=1}^n \Psi_m(y_t,\boldsymbol{\theta}_{S_m})\Psi_m(y_t,\boldsymbol{\theta}_{S_m})^T,\\
\dot{l}_{m,n}(\boldsymbol{\theta}_{S_m}) &= \sum_{t=1}^n \Psi_m(y_t,\boldsymbol{\theta}_{S_m}), \text{and}\\
\ddot{l}_{m,n}(\boldsymbol{\theta}_{S_m}) &= \sum_{t=1}^n \dot{\Psi}_m(y_t,\boldsymbol{\theta}_{S_m}).
\end{aligned}$$

Assuming that expectations exist, write

$$\begin{aligned}
A_m(\boldsymbol{\theta}_{S_m}) &= \mathrm{E}\left(\dot{\Psi}_m(y,\boldsymbol{\theta}_{S_m})\right), \text{and}\\
B_m(\boldsymbol{\theta}_{S_m}) &= \mathrm{E}\left(\Psi_m(y,\boldsymbol{\theta}_{S_m})\Psi_m(y,\boldsymbol{\theta}_{S_m})^T\right).
\end{aligned}$$

Assuming that appropriate inverses exist, write

$$\begin{aligned} C_{m,n}(\boldsymbol{\theta}_{S_m}) &= A_{m,n}(\boldsymbol{\theta}_{S_m})^{-1} B_{m,n}(\boldsymbol{\theta}_{S_m}) A_{m,n}(\boldsymbol{\theta}_{S_m})^{-1}, \text{and} \\ C_m(\boldsymbol{\theta}_{S_m}) &= A_m(\boldsymbol{\theta}_{S_m})^{-1} B_m(\boldsymbol{\theta}_{S_m}) A_m(\boldsymbol{\theta}_{S_m})^{-1}. \end{aligned}$$

The Fisher Information matrix of $\boldsymbol{\theta}_{0,S_m}$ is given by

$$\mathcal{F}_m(\boldsymbol{\theta}_{0,S_m}) = E_{\boldsymbol{\theta}_{0,S_m}} \Psi_m(y, \boldsymbol{\theta}_{0,S_m}) \Psi_m(y, \boldsymbol{\theta}_{0,S_m})^T.$$

Denote $\hat{\boldsymbol{\theta}}_{S_m}$ as the Maximum Likelihood estimator of $\boldsymbol{\theta}_{S_m}$ under the m^{th} sub-model, which is the solution of the log-likelihood equation $\sum_{t=1}^{n} \Psi_m(y, \boldsymbol{\theta}_{S_m}) = 0$. Define $\hat{\boldsymbol{\theta}}'_m = (\hat{\boldsymbol{\theta}}_{S_m}^T, \mathbf{0}^T)^T$, where $\hat{\boldsymbol{\theta}}_{S_m}$ is an estimator of $\boldsymbol{\theta}^*_{S_m}$, the parameter vector which minimizes the Kullback-Leibler Information Criterion (KLIC),

$$I\left(f(y, \boldsymbol{\theta}_0) : f(y, \boldsymbol{\theta}_m), \boldsymbol{\theta}'_m\right) = \mathrm{E}\left(\ln\left[\frac{f(y, \boldsymbol{\theta}_0)}{f(y, \boldsymbol{\theta}_m)}\right]\right). \tag{1.2}$$

By writing $\boldsymbol{\theta}^{*\prime}_m = \Pi_m \boldsymbol{\theta}^*_m = (\boldsymbol{\theta}^{*T}_{S_m}, \mathbf{0}^T)^T$, we can obtain $\boldsymbol{\theta}^*_m$. The parameters in $\boldsymbol{\theta}^*_m$ follow the same order as in $\boldsymbol{\theta}_0$. Taking expectations with respect to the true distribution, we have

$$I\left(f(y, \boldsymbol{\theta}_0) : f(y, \boldsymbol{\theta}_m), \boldsymbol{\theta}'_m\right) = \int f(y, \boldsymbol{\theta}_0) \ln f(y, \boldsymbol{\theta}_0) \mathrm{d}\nu - \int f(y, \boldsymbol{\theta}_0) \ln f(y, \boldsymbol{\theta}_m) \mathrm{d}\nu. \tag{1.3}$$

Define $\hat{\boldsymbol{\theta}}_m$ as the Maximum Likelihood estimator of $\boldsymbol{\theta}_m$ under the m^{th} sub-model, so the model averaging estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \hat{\boldsymbol{\theta}}_m$, where w_m is weight for the m^{th} sub-model and $\mathbf{w} = (w_1, \ldots, w_M)^T$, belonging to weight set $\mathcal{W} = \{\mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}$.

Now, the AIC and BIC scores under the m^{th} candidate model are:

$$\text{AIC}_m = -2\sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_m) + 2k_m \text{ and } \text{BIC}_m = -2\sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_m) + k_m \ln n$$

respectively. The model weights based on these scores are given by

$$\hat{w}_{\text{xIC},m} = \exp(-\text{xIC}_m/2) / \sum_{m=1}^{M} \exp(-\text{xIC}_m/2), \quad m = 1, \ldots, M, \tag{1.4}$$

where xIC_m is the AIC or BIC score from the m^{th} sub-model. The model average estimators resulting from these weights are commonly referred to as the S-AIC or S-BIC estimators, defined as

$$\begin{aligned} \hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{AIC}) &= \sum_{m=1}^{M} \hat{w}_{AIC,m} \hat{\boldsymbol{\theta}}_m, \text{and} \\ \hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{BIC}) &= \sum_{m=1}^{M} \hat{w}_{BIC,m} \hat{\boldsymbol{\theta}}_m \end{aligned} \tag{1.5}$$

respectively. The regularity conditions required for the asymptotic results are as follows. All limiting processes presented here are with respect to $n \to \infty$ and the notations $\xrightarrow{d}$, $\xrightarrow{a.s.}$ and $\xrightarrow{p}$ denote convergence in distribution, almost surely and in probability, respectively. We denote $\hat{\theta}_j(\hat{\mathbf{w}}_{AIC})$, $\hat{\theta}_j(\hat{\mathbf{w}}_{BIC})$ and $\theta_{j,0}$ as the j^{th} component of $\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{AIC})$, $\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{BIC})$ and $\boldsymbol{\theta}_0$ respectively.

Condition 1.1. *The density $f(y,\boldsymbol{\theta})$ is measurable in y for every $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, a compact subset of $\mathbb{R}^q$, continuous in $\boldsymbol{\theta}$ for every y in $\boldsymbol{\Omega}$, a measurable Euclidean space, and the true parameter point $\boldsymbol{\theta}_0$ is identifiable.*

Condition 1.2. *(a)* $\mathrm{E}(\ln f(y_t))$ *exists and* $|\ln f(y,\boldsymbol{\theta}_m)| \leq K_m(y)$ *for all* $\boldsymbol{\theta}$ *in* $\boldsymbol{\Theta}$, *where* $K_m(y)$ *is integrable with respect to* $\mathrm{d}\nu$;

(b) $I(f(y,\boldsymbol{\theta}_{0,m}) : f(y,\boldsymbol{\theta}_m), \boldsymbol{\theta}'_m)$ *has a unique minimum at* $\boldsymbol{\theta}^*_m$ *in* $\boldsymbol{\Theta}_m$.

Condition 1.3. *(a)* $\partial \ln f(y,\boldsymbol{\theta}_m)/\partial\theta_i, i = 1,\ldots,k_m$, *are bounded in absolute value by a function integrable with respect to* $\mathrm{d}\nu$ *uniformly in some neighborhood of* $\boldsymbol{\theta}_0$.

(b) The second partial derivative of $f(y,\boldsymbol{\theta}_m)$ with respect to $\boldsymbol{\theta}_m$ exists and is continuous for all y, and may be passed under the integral sign in $\int f(y,\boldsymbol{\theta}_m)\mathrm{d}\nu$.

Condition 1.4. $|\partial^2 \ln f(y,\boldsymbol{\theta}_m)/\partial\theta_i\partial\theta_j|$ *and* $|\partial \ln f(y,\boldsymbol{\theta}_m)/\partial\theta_i \cdot \partial \ln f(y,\boldsymbol{\theta}_m)/\partial\theta_j|$, $i,j = 1,\ldots,k_m$ *are dominated by functions integrable with respect to* $\mathrm{d}\nu$ *for all y in $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}_m$ in $\boldsymbol{\Theta}_m$.*

Condition 1.5. *(a) $\boldsymbol{\theta}^*_{S_m}$ is in the interior of $\boldsymbol{\Theta}_m$; (b) $B(\boldsymbol{\theta}^*_{S_m})$ is nonsingular; (c) $\boldsymbol{\theta}^*_{S_m}$ is a regular point of $A_m(\boldsymbol{\theta}_{S_m})$, defined as the value for $\boldsymbol{\theta}_{S_m}$ such that $A_m(\boldsymbol{\theta}_{S_m})$ has a constant rank in some open neighborhood of $\boldsymbol{\theta}_{S_m}$.*

Condition 1.6. *The Fisher Information $\mathcal{F}(\boldsymbol{\theta}_0)$ is a positive definite matrix.*

Condition 1.7. *The derivatives $|\partial[\partial f(y,\boldsymbol{\theta})/\theta_i \cdot f(y,\boldsymbol{\theta})]/\partial\theta_j|$, $i,j = 1,\cdots,q$, are dominated by functions integrable with respect to ν for all θ in Θ, and the minimal support of $f(y,\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$.*

Remark 1.1. *Condition 1.2 ensures that the KLIC is well-defined. Condition 1.3(a) allows us to apply the Uniform Law of Large Numbers. Condition 1.3(b) ensures that the first two derivatives with respect to $\boldsymbol{\theta}_{S_m}$ exist. This condition allows us to apply Taylor's Theorem and Mean Value Theorem for random functions. Condition 1.4 ensures that the derivatives are appropriately dominated by functions integrable with respect to* $\mathrm{d}\nu$. *This in turns ensures that $A_m(\boldsymbol{\theta}_{S_m})$ and $B_m(\boldsymbol{\theta}_{S_m})$ are continuous in $\boldsymbol{\theta}_{S_m}$, and that we can apply the Uniform Law of Large Numbers to $A_{m,n}(\boldsymbol{\theta}_{S_m})$ and $B_{m,n}(\boldsymbol{\theta}_{S_m})$. These assumptions are adopted from White (1982) and Ferguson (1996).*

2 Main results

Assume that the m^{th}_o model is the true model, i.e., $\boldsymbol{\theta}_{m_o} = \boldsymbol{\theta}_0$. Denote $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_{0,m}$ as the Maximum Likelihood estimator of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_{0,m}$ under the true model respectively. Any sub-model containing all regressors of the true model is referred to as an overfitted model, contained by the set $\mathcal{O}$. The remaining sub-models are referred to as underfitted models, contained by the set $\mathcal{U}$. Let $d_m = \exp\left\{\left(\kappa^T P_{m_o}\kappa - \kappa^T P_m \kappa\right)/2 + k_{m_o} - k_m\right\}$, where $\kappa^T P_{m_o}\kappa \sim \mathcal{X}^2(q-k_{m_o})$ and $\kappa^T P_m \kappa \sim \mathcal{X}^2(q-k_m)$. Let $\mathbf{w}^o_m$ be a vector with the m^{th} element taking on the value of unity and other elements zeros and $w_{\text{AIC},m} = \{\mathbf{I}(m \in \mathcal{O}/\{m_o\})d_m + \mathbf{I}(m = m_o)\} \Big/ \left\{1 + \sum_{m\in\mathcal{O}/\{m_o\}} d_m\right\}$, where $\mathbf{I}(\cdot)$ is an indicator function. Under the m^{th} model, the likelihood ratio test statistic is given by

$$\lambda_{m,n} = \frac{sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m}\prod_{t=1}^n f(y_t,\boldsymbol{\theta})}{sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\prod_{t=1}^n f(y_t,\boldsymbol{\theta})} = \frac{\prod_{t=1}^n f(y_t,\hat{\boldsymbol{\theta}}_m)}{\prod_{t=1}^n f(y_t,\hat{\boldsymbol{\theta}}_o)}.$$

Lemma 2.1. *Suppose that Conditions 1.1, 1.3 and 1.6 are satisfied and $m \in \mathcal{O}$. Then we have*

$$\begin{aligned} -2\ln\lambda_{m,n} &= (\Pi_m\xi_n)^T[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m]\Pi_m\xi_n + o(1) \\ &\xrightarrow{d} \kappa^T P_m\kappa \sim \chi^2(q-k_m), \end{aligned}$$

where $\xi_n = \frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0)$, $\kappa \sim \mathcal{N}(0, I_{q\times q})$ *and* $P_m = (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{\frac{1}{2}}[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m](\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{\frac{1}{2}}$.

Proof: See Subsection 3.

Lemma 2.2. *For any underfitted model $s \in \mathcal{U}$, under Conditions 1.1,1.2 and Conditions 1.3(b)-1.5, we have*

$$n^{-1}l_n(\hat{\boldsymbol{\theta}}_s) = \mathrm{E}(\ln f(y_t, \boldsymbol{\theta}_s^*)) + o_p(1). \tag{2.1}$$

Proof: See Subsection 3.

Lemma 2.3. *Suppose that Conditions 1.1-1.6 are satisfied. Then*

$$\hat{w}_{\text{AIC},m} \xrightarrow{d} w_{\text{AIC},m}, \;\; m = 1, 2, \cdots, M \;\; \textit{and} \;\; \hat{\mathbf{w}}_{\text{BIC}} \equiv (\hat{w}_{\text{BIC},1}, \ldots, \hat{w}_{\text{BIC},M})^T \xrightarrow{p} \mathbf{w}_{m_o}^o. \tag{2.2}$$

Proof: See Subsection 3.

Let Δ_m be a diagonal selection matrix such that

$$\Delta_m = \begin{pmatrix} \mathbf{I}_{q_1} & & & & \\ & \delta_1 & & & \\ & & \delta_2 & & \\ & & & \ddots & \\ & & & & \delta_{q_2} \end{pmatrix},$$

where $\mathbf{I}_{q_1}$ is an $q_1 \times q_1$ identity matrix and

$$\delta_j = \begin{cases} 1, & j \in S_m, \\ 0, & j \notin S_m, \end{cases}$$

The asymptotic distributions of the S-AIC and S-BIC estimators are established in the following theorem.

Theorem 2.1. *Suppose that Conditions 1.1-1.7 are satisfied. Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\hat{w}_{\text{AIC}}) - \boldsymbol{\theta}_0) \xrightarrow{d} \sum_{m\in\mathcal{O}} (G_m/G)\Pi_m^T H_{S_m}\Pi_m\Delta_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}, \tag{2.3}$$

$\hat{\theta}_j(\hat{\mathbf{w}}_{\text{BIC}}) \xrightarrow{p} 0$ *for* $j \notin S_{m_o}$, *and* $\hat{\theta}_j(\hat{\mathbf{w}}_{\text{BIC}}) \xrightarrow{d} Z_j$ *for* $j \in S_{m_o}$,

where

$$G_m = \exp\left\{(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta})^T[H_m - (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}/2 - k_m\right\},$$

$G = \sum_{m\in\mathcal{O}} G_m$, $\boldsymbol{\beta} \sim N(0, \mathcal{F}^{-1}(\boldsymbol{\theta}_0))$, $Z_j \sim \mathcal{N}(0, \sigma_j)$ *and* σ_j *is the* j^{th} *element on the diagonal of matrix* $\Sigma = \Pi_{m_o}^T H_{S_{m_o}}\Pi_{m_o}\Delta_{m_o}\mathcal{F}(\boldsymbol{\theta}_0)(\Pi_{m_o}^T H_{S_{m_o}}\Pi_{m_o}\Delta_{m_o})^T$.

Proof: See Subsection 3.

Denote $\hat{\boldsymbol{\theta}}_m = (\hat{\theta}_{1,m}, \cdots, \hat{\theta}_{q,m})^T$, where $\hat{\theta}_{j,m}$ is the jth component of $\hat{\boldsymbol{\theta}}_m$. Following Buckland et al. (1997), we also consider the scaled SAIC and scaled SBIC estimators such that $\sum_k \hat{w}_{xIC,m_k} = 1$, where xIC is AIC or BIC. This leads to the scaled SAIC and scaled SBIC estimators of θ_j, defined as

$$\hat{\theta}_j(\hat{\mathbf{w}}_{AICs}) = \sum_{k=1}^{M_j} \frac{\hat{w}_{AIC,m_k}}{\sum_{k=1}^{M_j} \hat{w}_{AIC,m_k}} \hat{\theta}_{j,k} \text{ and } \hat{\theta}_j(\hat{\mathbf{w}}_{BICs}) = \sum_{k=1}^{M_j} \frac{\hat{w}_{BIC,m_k}}{\sum_{k=1}^{M_j} \hat{w}_{BIC,m_k}} \hat{\theta}_{j,k} \tag{2.4}$$

respectively.

Denote $\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{AICs}) = (\hat{\theta}_1(\hat{\mathbf{w}}_{AICs}), \cdots, \hat{\theta}_q(\hat{\mathbf{w}}_{AICs}))$ and $\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{BICs}) = (\hat{\theta}_1(\hat{\mathbf{w}}_{BICs}), \cdots, \hat{\theta}_q(\hat{\mathbf{w}}_{BICs}))$ Next, we establish the asymptotic distributions of $\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{AICs})$ and $\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{BICs})$.

Corollary 2.1. *Suppose that Conditions 1.1-1.7 are satisfied. Then*

$$\sqrt{n}(\hat{\theta}_j(\hat{\mathbf{w}}_{AICs}) - \boldsymbol{\theta}_{0,j}) \xrightarrow{d} \sum_{k=1}^{M_j} \mathbf{1}\{m_k \in \mathcal{O}\}\left(\frac{G_{m_k}}{\sum_{k=1}^{M_j} \mathbf{1}\{m_k \in \mathcal{O}\} G_{m_k}}\right) \Delta_{m_k} \mathcal{F}^{-1}(\boldsymbol{\theta}_0) \Delta_{m_k} \mathcal{F}(\boldsymbol{\theta}_0) \eta_j$$

and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{BICs}) - \boldsymbol{\theta}_0) \xrightarrow{d} \boldsymbol{\beta},$$

where

$$G_{m_k} = \exp\left\{(\Pi_{m_k}\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta})^T [H_{m_k} - (\Pi_{m_k}\mathcal{F}(\boldsymbol{\theta}_0)\Pi_{m_k}^T)^{-1}]\Pi_{m_k}\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}/2 - k_m\right\},$$

$\boldsymbol{\beta} \sim N(0, \mathcal{F}^{-1}(\boldsymbol{\theta}_0))$, η_j *is the* j*th component of* $\boldsymbol{\beta}$ *and*

$$\mathbf{1}\{m_k \in \mathcal{O}\} = \begin{cases} 1, & m_k \in \mathcal{O}, \\ 0, & m_k \notin \mathcal{O}. \end{cases}$$

Proof: The corollary is a direct consequence of Theorem 2.1.

Furthermore, let the parameter of interest be $\mu = \mu(\boldsymbol{\theta})$, a smooth real-valued function. Denote the sub-model estimator as $\hat{\mu}_m = \mu(\hat{\boldsymbol{\theta}}_m)$). Then the model averaging estimators of μ based on SAIC weight and SBIC weight are

$$\hat{\mu}(\hat{\mathbf{w}}_{\text{AIC}}) = \sum_{m=1}^{M} \hat{w}_{\text{AIC},m}\hat{\mu}_m \text{ and } \hat{\mu}(\hat{\mathbf{w}}_{\text{BIC}}) = \sum_{m=1}^{M} \hat{w}_{\text{BIC},m}\hat{\mu}_m,$$

respectively. Then by Theorems 2.1 and Theorem 7 in Ferguson (1996), we obtain the following corollary.

Corollary 2.2. *Under the assumptions of Theorem 2.1 and assuming that* $\dot{\mu}(\boldsymbol{\theta}) = \frac{\partial \mu}{\partial \boldsymbol{\theta}}$ *is continuous in a neighborhood of* $\boldsymbol{\theta}_0$*, we have*

$$\sqrt{n}(\hat{\mu}(\hat{\mathbf{w}}_{\text{AIC}}) - \mu(\boldsymbol{\theta}_0)) \xrightarrow{d} \sum_{m \in \mathcal{O}} (G_m/G)\dot{\mu}(\boldsymbol{\theta}_0)^T \Pi_m^T H_{S_m} \Pi_m \Delta_m \mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta} \tag{2.5}$$

and

$$\sqrt{n}(\hat{\mu}(\hat{\mathbf{w}}_{\text{BIC}}) - \mu(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathcal{N}(0, \dot{\mu}(\boldsymbol{\theta}_0)^T \Sigma \dot{\mu}(\boldsymbol{\theta}_0)). \tag{2.6}$$

3 Proof of Theorem 2.1

Proof of Lemma 2.1. When $m \in \mathcal{O}$, the last $q - k_m$ components of the true value $\boldsymbol{\theta}_{0,m}$ are 0. Given Conditions 1.1, 1.3 and 1.6, and assuming that Theorems 18 and 22 of Ferguson (1996) are satisfied, then we have $-2\ln\lambda_{m,n} = 2[l_n(\hat{\boldsymbol{\theta}}_0) - l_n(\hat{\boldsymbol{\theta}}_m)]$. Now, expanding $l_n(\hat{\boldsymbol{\theta}}_m)$ about $\hat{\boldsymbol{\theta}}_{0,m}$ yields:

$$l_n(\hat{\boldsymbol{\theta}}_m) = l_n(\hat{\boldsymbol{\theta}}_{0,m}) + \dot{l}_{m,n}(\hat{\boldsymbol{\theta}}_{0,m})(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}) - n(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m})^T I_n(\hat{\boldsymbol{\theta}}'_m)(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}),$$

where $I_n(\hat{\boldsymbol{\theta}}'_m) = -\frac{1}{n}\int_0^1\int_0^1 v\ddot{l}_{m,n}(\hat{\boldsymbol{\theta}}_{0,m} + uv(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}))\mathrm{d}u\mathrm{d}v \xrightarrow{a.s.} \frac{1}{2}\mathcal{F}_m(\boldsymbol{\theta}_{0,m})$, as in the proof of Theorem 18 of Ferguson (1996). Let $o(1)$ represent a random variable matrix with each element converging almost surely to 0 as $n \to \infty$. By $\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}_{0,m}) = 0$, we have

$$\begin{aligned}
-2\ln\lambda_{m,n} &= -2(l_n(\hat{\boldsymbol{\theta}}_m) - l_n(\hat{\boldsymbol{\theta}}_0)) \\
&= 2n(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_0)^T I_n(\hat{\boldsymbol{\theta}}_m)(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_0) \\
&= 2n(\Pi_m(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_0))^T \Pi_m I_n(\hat{\boldsymbol{\theta}}_m)\Pi_m^T \Pi_m(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_0) \\
&= 2n(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m})^T I_n(\hat{\boldsymbol{\theta}}'_m)(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}) \\
&= n(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m})^T(\mathcal{F}_m(\boldsymbol{\theta}_{0,m}) + o(1))(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}).
\end{aligned}$$

To ascertain the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m})$, consider the following expansion of $\dot{l}_n(\hat{\boldsymbol{\theta}}'_m)$ about $\hat{\boldsymbol{\theta}}_{0,m}$:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m) &= \frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}_{0,m}) + \frac{1}{n}\int_0^1 \ddot{l}_{m,n}(\hat{\boldsymbol{\theta}}_{0,m} + v(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}))\mathrm{d}v\sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}) \\
&= (-\mathcal{F}_m(\boldsymbol{\theta}_{0,m}) + o(1))\sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}).
\end{aligned}$$

Hence $\sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \hat{\boldsymbol{\theta}}_{0,m}) = -(\mathcal{F}_m(\boldsymbol{\theta}_{0,m})^{-1} + o(1))\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m)$ and

$$-2\ln\lambda_{m,n} = \frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m)^T(\mathcal{F}_m(\boldsymbol{\theta}_{0,m})^{-1} + o(1))\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m). \tag{3.1}$$

To seek the asymptotic distribution of $\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m)$, consider the following expansion about $\theta_{0,m}$:

$$\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m) = \frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m}) + \frac{1}{n}\int_0^1 \ddot{l}_{m,n}(\boldsymbol{\theta}_{0,m} + v(\hat{\boldsymbol{\theta}}'_m - \boldsymbol{\theta}_{0,m}))\mathrm{d}v\sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \boldsymbol{\theta}_{0,m}). \tag{3.2}$$

For the m^{th} sub-model, write $\mathcal{F}_m(\boldsymbol{\theta}_{0,m})$ as

$$\mathcal{F}_m(\boldsymbol{\theta}_{0,m}) = \begin{pmatrix} k_m \times k_m & k_m \times (q-k_m) \\ G_{m,1} & G_{m,2} \\ (q-k_m) \times k_m & (q-k_m) \times (q-k_m) \\ G_{m,3} & G_{m,4} \end{pmatrix}$$

and let

$$H_m = \begin{pmatrix} G_{m,1}^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that the first k_m components of $\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m)$ are zero, yielding $H_m\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m) = 0$ and

$$H_m\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m}) = H_m(\mathcal{F}_m(\boldsymbol{\theta}_{0,m}) + o(1))\sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \boldsymbol{\theta}_{0,m}) = \sqrt{n}(\hat{\boldsymbol{\theta}}'_m - \boldsymbol{\theta}_{0,m}) + o(1)$$

as the last $q - k_m$ components of $\hat{\boldsymbol{\theta}}'_m$ and $\boldsymbol{\theta}_{0,m}$ are equal. Substituting this result into (3.2), we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m) &= [I - \mathcal{F}_m(\boldsymbol{\theta}_{0,m})H_m]\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m}) + o(1)\\ &= [I - \Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T H_m]\frac{1}{\sqrt{n}}\Pi_m\dot{l}_n(\boldsymbol{\theta}_0) + o(1). \qquad (3.3)\end{aligned}$$

From the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\theta}_0) = \sqrt{n}\left(\frac{1}{n}\dot{l}_n(\boldsymbol{\theta}_0)\right) = \boldsymbol{\xi}_n \xrightarrow{d} \boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathcal{F}(\boldsymbol{\theta}_0))$. Hence,

$$\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\hat{\boldsymbol{\theta}}'_m) \xrightarrow{d} [I - \Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T H_m]\Pi_m\boldsymbol{\xi},$$

so that, from (3.1) and (3.3),

$$\begin{aligned}-2\ln\lambda_{m,n} &= \frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m})^T[I - \mathcal{F}_m(\boldsymbol{\theta}_{0,m})H_m]\mathcal{F}_m(\boldsymbol{\theta}_{0,m})^{-1}[I - \mathcal{F}_m(\boldsymbol{\theta}_{0,m})H_m] \qquad (3.4)\\ &\times \frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m}) + o(1)\\ &= \frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m})^T[\mathcal{F}_m(\boldsymbol{\theta}_{0,m})^{-1} - H_m]\frac{1}{\sqrt{n}}\dot{l}_{m,n}(\boldsymbol{\theta}_{0,m}) + o(1)\ (by\ H_m\mathcal{F}_m(\boldsymbol{\theta}_{0,m})H_m = H_m)\\ &= \frac{1}{\sqrt{n}}(\Pi_m\dot{l}_n(\boldsymbol{\theta}_0))^T[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m]\frac{1}{\sqrt{n}}\Pi_m\dot{l}_n(\boldsymbol{\theta}_0) + o(1)\\ &= (\Pi_m\boldsymbol{\xi}_n)^T[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m]\Pi_m\boldsymbol{\xi}_n + o(1)\\ &\xrightarrow{d} (\Pi_m\boldsymbol{\xi})^T[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m]\Pi_m\boldsymbol{\xi}\\ &= \boldsymbol{\kappa}^T(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{\frac{1}{2}}[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m](\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{\frac{1}{2}}\boldsymbol{\kappa} \qquad (3.5)\end{aligned}$$

where $\boldsymbol{\kappa} = \mathcal{F}(\boldsymbol{\theta}_0)^{-\frac{1}{2}}\boldsymbol{\xi} \sim \mathcal{N}(0, I_{q\times q})$. Let $P_m = (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{\frac{1}{2}}[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1} - H_m](\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{\frac{1}{2}}$ be a projection and $rank(P_m) = trace(P_m) = q - k_m$. Hence,

$$-2\ln\lambda_{m,n} \xrightarrow{d} \boldsymbol{\kappa}^T P_m\boldsymbol{\kappa} \sim \chi^2(q - k_m). \qquad (3.6)$$

Proof of Lemma 2.2. Consider an underfitted model $s \in \mathcal{U}$. Let Conditions 1.1-1.2, Conditions 1.3(b)-1.5, and Theorems 2.2 and 3.2 of White (1982) hold. Then we have

$$\hat{\boldsymbol{\theta}}_{S_s} \xrightarrow{a.s.} \boldsymbol{\theta}^*_{S_s} \qquad (3.7)$$

and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{S_s} - \boldsymbol{\theta}^*_{S_s}) \xrightarrow{d} N(0, C_s(\boldsymbol{\theta}^*_{S_s})). \qquad (3.8)$$

As well, it can be shown that $C_{s,n}(\hat{\boldsymbol{\theta}}_{S_s}) \xrightarrow{a.s.} C_s(\boldsymbol{\theta}^*_{S_s})$. Then by applying the Taylor's Theorem argument of Roy (1957), we have

$$\begin{aligned} l_n(\hat{\boldsymbol{\theta}}_s) &= \sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_s) \\ &= l_n(\boldsymbol{\theta}^*_s) + \sum_{t=1}^{n} \Psi_s(y_t, \boldsymbol{\theta}^*_{S_s})(\hat{\boldsymbol{\theta}}_{S_s} - \boldsymbol{\theta}^*_{S_s}) + \frac{n}{2}(\hat{\boldsymbol{\theta}}_{S_s} - \boldsymbol{\theta}^*_{S_s})^T A_{s,n}(\hat{\boldsymbol{\theta}}_{S_s} + \alpha(\boldsymbol{\theta}^*_{S_s} - \hat{\boldsymbol{\theta}}_{S_s}))(\hat{\boldsymbol{\theta}}_{S_s} - \boldsymbol{\theta}^*_{S_s}), \end{aligned}$$

where $\alpha \in (0,1)$. Given Conditions 1.1-1.2 and 1.3(b)-1.5, and the proof of Theorem 3.2 of White (1982), we have $\mathrm{E}(\Psi_s(y_t, \boldsymbol{\theta}^*_{S_s})) = 0$. In addition, by the Laws of Large Numbers, we have

$$n^{-1} \sum_{t=1}^{n} \Psi_s(y_t, \boldsymbol{\theta}^*_{S_s}) \xrightarrow{p} \mathrm{E}(\Psi_s(y_t, \boldsymbol{\theta}^*_{S_s})), \tag{3.9}$$

and

$$n^{-1} l_n(\boldsymbol{\theta}^*_s) = n^{-1} \sum_{t=1}^{n} \ln f(y_t, \boldsymbol{\theta}^*_s) \xrightarrow{p} \mathrm{E}(\ln f(y_t, \boldsymbol{\theta}^*_s)). \tag{3.10}$$

By Theorem 2.2 of White (1982), it can be shown that

$$A_{s,n}(\hat{\boldsymbol{\theta}}_{S_s} + \alpha(\boldsymbol{\theta}^*_{S_s} - \hat{\boldsymbol{\theta}}_{S_s})) \xrightarrow{a.s.} A_s(\boldsymbol{\theta}^*_{S_s}) \tag{3.11}$$

Then, by (3.7)-(3.11), we obtain

$$n^{-1} l_n(\hat{\boldsymbol{\theta}}_s) = \mathrm{E}(\ln f(y_t, \boldsymbol{\theta}^*_s)) + o_p(1). \tag{3.12}$$

Proof of Lemma 2.3. From Lemma 2.1, when $m \in \mathcal{O}$, we have

$$\begin{aligned} & \hat{w}_{\text{AIC},m} \hat{w}^{-1}_{\text{AIC},m_o} = \exp\{\text{AIC}_{m_o}/2 - \text{AIC}_m/2\} \\ = \;& \exp\left\{ -\sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_{m_o}) + \sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_m) + k_{m_o} - k_m \right\} \\ = \;& \exp\left\{ -\ln\left(\prod_{t=1}^{n} f(y_t, \hat{\boldsymbol{\theta}}_{m_o}) \Big/ \prod_{t=1}^{n} f(y_t, \hat{\boldsymbol{\theta}}_o) \right) + \ln\left(\prod_{t=1}^{n} f(y_t, \hat{\boldsymbol{\theta}}_m) \Big/ \prod_{t=1}^{n} f(y_t, \hat{\boldsymbol{\theta}}_o) \right) + k_{m_o} - k_m \right\} \\ = \;& \exp\left\{ \left(-2\ln\lambda_{m_o,n} + 2\ln\lambda_{m,n}\right)/2 + k_{m_o} - k_m \right\} \\ = \;& \exp\Big\{ \big((\Pi_{m_o}\boldsymbol{\xi}_n)^T [(\Pi_{m_o}\mathcal{F}(\boldsymbol{\theta}_0)\Pi^T_{m_o})^{-1} - H_{m_o}]\Pi_m\boldsymbol{\xi}_n - (\Pi_m\boldsymbol{\xi}_n)^T[(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi^T_m)^{-1} - H_m]\Pi_m\boldsymbol{\xi}_n \\ & \quad + o(1)\big)/2 + k_{m_o} - k_m \Big\} \\ \xrightarrow{d} \;& \exp\left\{ \left(\boldsymbol{\kappa}^T P_{m_o} \boldsymbol{\kappa} - \boldsymbol{\kappa}^T P_m \boldsymbol{\kappa} \right)/2 + k_{m_o} - k_m \right\}, \end{aligned} \tag{3.13}$$

where $\boldsymbol{\kappa}^T P_{m_o}\boldsymbol{\kappa} \sim \mathcal{X}^2(q - k_{m_o})$ and $\boldsymbol{\kappa}^T P_{m_o}\boldsymbol{\kappa} \sim \mathcal{X}^2(q - k_m)$.

On the other hand, when $s \in \mathcal{U}$, from Lemma 2.2, we have

$$n^{-1} l_n(\hat{\boldsymbol{\theta}}_s) = \mathrm{E}(\ln f(y_t, \boldsymbol{\theta}^*_s)) + o_p(1). \tag{3.14}$$

Similarly, from the proof of Lemma 2.2, it can be proven that

$$n^{-1}l_n(\hat{\boldsymbol{\theta}}_{m_o}) = \mathrm{E}(\ln f(y_t, \boldsymbol{\theta}_0)) + o_p(1). \tag{3.15}$$

By the definition of KLIC and Theorem of Bowden (1973), we have

$$\mathrm{E}(\ln f(y_t, \boldsymbol{\theta}_0)) - \mathrm{E}(\ln f(y_t, \boldsymbol{\theta}_s^*)) = \delta > 0. \tag{3.16}$$

Hence, when $s \in \mathcal{U}$, by (3.14)-(3.16), we have

$$\begin{aligned}
& \hat{w}_{\text{AIC},s}\hat{w}^{-1}_{\text{AIC},m_o} = \exp(\text{AIC}_{m_o}/2 - \text{AIC}_s/2) \\
= \;& \exp\left(-\sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_{m_o}) + \sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_s) + k_{m_o} - k_s\right) \\
= \;& \exp\left(-n\left(n^{-1}\sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_{m_o}) - n^{-1}\sum_{t=1}^{n} \ln f(y_t, \hat{\boldsymbol{\theta}}_s)\right) + k_{m_o} - k_s\right) \\
= \;& \exp\left(-n\left(n^{-1}l_n(\hat{\boldsymbol{\theta}}_{m_o}) - n^{-1}l_n(\hat{\boldsymbol{\theta}}_s)\right) + k_{m_o} - k_s\right) \\
= \;& [\exp(-n)]^{\left(n^{-1}l_n(\hat{\boldsymbol{\theta}}_{m_o}) - n^{-1}l_n(\hat{\boldsymbol{\theta}}_s) + (k_{m_o} - k_s)/n\right)} \\
= \;& O_p(\exp(-n)) \xrightarrow{p} 0 \;\; as \; n \to \infty.
\end{aligned} \tag{3.17}$$

As $\hat{\mathbf{w}}_{\text{AIC}} \in \mathcal{W}$, (3.17) implies that $\hat{w}_{\text{AIC},m} = O_p(\exp(-n))$. Using (3.13), (3.17) and

$$\hat{w}_{\text{AIC},m} = \hat{w}_{\text{AIC},m}\hat{w}^{-1}_{\text{AIC},m_o} / \sum_{m=1}^{M} \hat{w}_{\text{AIC},m}\hat{w}^{-1}_{\text{AIC},m_o},$$

we have $\hat{w}_{\text{AIC}} \xrightarrow{d} w_{\text{AIC}}$.

Similarly, we can show that for any overfitted model m and $m \neq m_o$,

$$\hat{w}_{\text{BIC},m}\hat{w}^{-1}_{\text{BIC},m_o} = O_p(1) \cdot n^{(k_{m_o} - k_m)/2} \xrightarrow{p} 0,$$

and that for any underfitted model $s \in \mathcal{U}$,

$$\hat{w}_{\text{BIC},s}\hat{w}^{-1}_{\text{BIC},m_o} = [\exp(-n)]^{n^{-1}l_n(\hat{\boldsymbol{\theta}}_{m_o}) - n^{-1}l_n(\hat{\boldsymbol{\theta}}_s)} n^{(k_{m_o} - k_s)/2} \xrightarrow{p} 0.$$

From above two formulae, we have $\hat{w}_{\text{BIC}} \xrightarrow{p} w^o_{m_o}$.

Proof of Theorem 2.1. When $m \in \mathcal{O}$, from the proof of the Lemmas 2.1 and 2.3, we have

$$\begin{aligned}
\hat{w}_{\text{AIC},m} &= \hat{w}_{\text{AIC},m}\hat{w}^{-1}_{\text{AIC},m_o} / \sum_{m=1}^{M} \hat{w}_{\text{AIC},m}\hat{w}^{-1}_{\text{AIC},m_o} \\
&= \exp\left\{\left(-2\ln\lambda_{m_o,n} + 2\ln\lambda_{m,n}\right)/2 + k_{m_o} - k_m\right\} / \Big\{\sum_{m\in\mathcal{O}} \exp\Big\{\left(-2\ln\lambda_{m_o,n} + 2\ln\lambda_{m,n}\right)/2 \\
&\qquad + k_{m_o} - k_m\Big\} + o_p(1)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{(\Pi_m\boldsymbol{\xi}_n)^T[H_m - (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\boldsymbol{\xi}_n/2 - k_m + o(1)\right\} / \{\sum_{m\in\mathcal{O}} \exp\left\{(\Pi_m\boldsymbol{\xi}_n)^T\right.\\
&\qquad \left.[(H_m - \Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\boldsymbol{\xi}_n/2 - k_m + o(1)\right\} + o_p(1)\}\\
&\xrightarrow{d} \exp\left\{(\Pi_m\boldsymbol{\xi})^T[H_m - (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\boldsymbol{\xi}/2 - k_m\right\} / \Big\{ \sum_{m\in\mathcal{O}} \exp\left\{(\Pi_m\boldsymbol{\xi})^T\right.\\
&\qquad \left.[(H_m - \Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\boldsymbol{\xi}/2 - k_m\right\}\Big\}\\
&= \exp\left\{(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta})^T[H_m - (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}/2 - k_m\right\}\\
&\qquad / \Big\{ \sum_{m\in\mathcal{O}} \exp\left\{(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta})^T[H_m - (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}/2 - k_m\right\}\Big\}\\
&= G_m / \sum_{m\in\mathcal{O}} G_m = \ G_m/G, \qquad (3.18)
\end{aligned}
$$

where

$$
G_m = \exp\left\{(\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta})^T[H_m - (\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\Pi_m^T)^{-1}]\Pi_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}/2 - k_m\right\}
$$

$\boldsymbol{\beta} \sim \mathcal{N}(0, \mathcal{F}^{-1}(\boldsymbol{\theta}_0))$, and $G = \sum_{m\in\mathcal{O}} G_m$.

On the other hand, when $m \in \mathcal{U}$, from the proof of the Lemma 2.3, we obtain

$$
\hat{w}_{\text{AIC},m} = O_p(\exp(-n)) \ \ and \ \ \hat{w}_{\text{BIC},m} = O_p(\exp(-n)n^{(k_{m_o}-k_m)/2}). \qquad (3.19)
$$

As $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^q$, by Theorems 2.2 and 3.2 of White (1982), we can conclude, for any sub-model m, that

$$
\boldsymbol{\theta}_m^* = O(1), \ \boldsymbol{\theta}_0 = O(1), \ \hat{\boldsymbol{\theta}}_m \xrightarrow{a.s.} \boldsymbol{\theta}_m^*, \qquad (3.20)
$$

and

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_{S_m} - \boldsymbol{\theta}_{S_m}^*) = -A_m^{-1}(\boldsymbol{\theta}_{S_m}^*)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial \ln f(y_t, \boldsymbol{\theta}_m^*)}{\partial \boldsymbol{\theta}_{S_m}} + o_p(1). \qquad (3.21)
$$

When $m \in \mathcal{O}$, by the definition of KLIC and Theorem of Bowden (1973), we have $\boldsymbol{\theta}_{S_m}^* = \boldsymbol{\theta}_{0,S_m}$ and $\boldsymbol{\theta}_m^* = \boldsymbol{\theta}_0$. Then by Theorem 3.3 of White (1982), we have $-A_m^{-1}(\boldsymbol{\theta}_{S_m}^*) = G_{S_m,1}^{-1}$, and hence

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_{S_m} - \boldsymbol{\theta}_{S_m}^*) = G_{S_m,1}^{-1}\sqrt{n}(\frac{1}{n}\sum_{t=1}^{n}\frac{\partial \ln f(y_t, \boldsymbol{\theta}_m^*)}{\partial \boldsymbol{\theta}_{S_m}}) + o_p(1). \qquad (3.22)
$$

From the proof of Lemma 2.1, we have $\sqrt{n}(\frac{1}{n}\sum_{t=1}^{n}\frac{\partial \ln f(y_t,\boldsymbol{\theta}_m^*)}{\partial \boldsymbol{\theta}_{S_m}}) = \boldsymbol{\xi}_{m,n} \xrightarrow{d} \boldsymbol{\xi}_m$, where $\boldsymbol{\xi}_m \sim \mathcal{N}(0, G_{S_m,1})$. Hence

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{\text{AIC}}) - \boldsymbol{\theta}_0) = \sum_{m=1}^{M}\hat{w}_{\text{AIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)
$$

$$
\begin{aligned}
&= \sum_{m\in\mathcal{U}} \hat{w}_{\text{AIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m + \boldsymbol{\theta}^*_m - \boldsymbol{\theta}_0) + \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)\\
&= \sum_{m\in\mathcal{U}} O_p(\exp(-n))\sqrt{n}O_p(1) + \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0) + o_p(1)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m) + o_p(1)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}\Pi_m^{-1}(\hat{\boldsymbol{\theta}}'_m - \boldsymbol{\theta}'_m) + o_p(1)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}\Pi_m^{T}((\hat{\boldsymbol{\theta}}^T_{S_m}, \mathbf{0}^T)^T - (\boldsymbol{\theta}^{*T}_{S_m}, \mathbf{0}^T)^T) + o_p(1)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\sqrt{n}\Pi_m^{T}((\hat{\boldsymbol{\theta}}_{S_m} - \boldsymbol{\theta}^*_{S_m})^T, \mathbf{0}^T_{q-k_m})^T + o_p(1)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\Pi_m^{T}\begin{pmatrix} G^{-1}_{S_m,1} & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{t=1}^n \frac{\partial \ln f(y_t,\boldsymbol{\theta}^*_m)}{\partial \boldsymbol{\theta}_{S_m}} \\ 0\end{pmatrix} + o_p(1)\\
&= \sum_{m\in\mathcal{O}} \hat{w}_{\text{AIC},m}\Pi_m^{T}H_{S_m}\Pi_m\Delta_m\boldsymbol{\xi}_n\\
&\xrightarrow{d} \sum_{m\in\mathcal{O}} (G_m/G)\Pi_m^{T}H_{S_m}\Pi_m\Delta_m\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta}
\end{aligned}
\tag{3.23}
$$

Similarly, we can prove that

$$
\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\theta}}(\hat{\mathbf{w}}_{\text{BIC}}) - \boldsymbol{\theta}_0) = \sum_{m=1}^{M} \hat{w}_{\text{BIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)\\
=&\sum_{m\in\mathcal{U}} \hat{w}_{\text{BIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0) + \sum_{m\in\mathcal{O}} \hat{w}_{\text{BIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)\\
=&\sum_{m\in\mathcal{U}} O_p(\exp(-n)n^{(k_{m_o}-k_m)/2})\sqrt{n}O_p(1) + \sum_{m\in\mathcal{O}} \hat{w}_{\text{BIC},m}\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0).\\
\xrightarrow{d}&\Pi_{m_o}^{T}H_{S_{m_o}}\Pi_{m_o}\Delta_{m_o}\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta},
\end{aligned}
$$

where $\Pi_{m_o}^{T}H_{S_{m_o}}\Pi_{m_o}\Delta_{m_o}\mathcal{F}(\boldsymbol{\theta}_0)\boldsymbol{\beta} \sim \mathcal{N}(0,\Sigma)$, and $\Sigma = \Pi_{m_o}^{T}H_{S_{m_o}}\Pi_{m_o}\Delta_{m_o}\mathcal{F}(\boldsymbol{\theta}_0)(\Pi_{m_o}^{T}H_{S_{m_o}}\Pi_{m_o}\Delta_{m_o})^T$. It is known that

$$
\Delta_{m_o} = \begin{pmatrix} \mathbf{I}_{q_1} & & & & \\ & \delta_1 & & & \\ & & \delta_2 & & \\ & & & \ddots & \\ & & & & \delta_{q_2}\end{pmatrix},
$$

where $\mathbf{I}_{q_1}$ is an $q_1 \times q_1$ identity matrix and

$$
\delta_j = \begin{cases} 1, & j \in S_{m_o},\\ 0, & j \notin S_{m_o}.\end{cases}
$$

As the m_oth candidate model is the true model, when $j \notin S_{m_o}$, $\theta_{j,0} = 0$. If σ_j is the j element on the diagonal of matrix Σ, then $\sigma_j = 0$. This leads to $\hat{\theta}_j(\hat{\mathbf{w}}_{\text{BIC}}) \xrightarrow{p} \theta_{j,0} = 0$. For $j \in S_{m_o}$,

$$\sqrt{n}(\hat{\theta}_j(\hat{\mathbf{w}}_{\text{BIC}}) - \theta_{j,0}) \xrightarrow{d} Z_j,$$

where $Z_j \sim \mathcal{N}(0, \sigma_j)$.

References

BOWDEN, ROGER (1973). The theory of parametric identification. *Econometrica* **41**, 1069–1074.

BUCKLAND, S. T., BURNHAM, K. P. & AUGUSTIN, N. H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603–618.

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

HJORT, N. L. & CLAESKENS, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

ROY, KP (1957). A note on the asymptotic distribution of likelihood ratio. *Calcutta Statistical Association Bulletin* **7**, 73–77.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.