

Generalized AIC for componentwise model selection

Youngjo Lee and Kwok Pui Choi

Seoul National University and National University of Singapore

Supplementary Material

This online supplementary material consists of the following: Proofs of Theorems 1–3, Proposition 2, detailed description of the four real data examples and Table A for the components $\boldsymbol{\psi}$ and $\boldsymbol{\varsigma}$.

S1 Proof of Theorem 1

We first consider the case $\boldsymbol{\psi}$ contains $\mathbf{v}$, so $E = E_{\mathbf{y}|\mathbf{u}}$ and $\widehat{\boldsymbol{\psi}} = \widehat{\boldsymbol{\psi}}_{\text{MHLE}}$. Recall that $\boldsymbol{\psi}_0$ denotes the true value. Let $\widehat{\boldsymbol{\psi}} = \widehat{\boldsymbol{\psi}}(\mathbf{y})$, $\mathbf{J}(\boldsymbol{\psi}; \mathbf{y}) = -\partial^2 \ell_*(\boldsymbol{\psi}; \mathbf{y}) / \partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top$, and $K(\boldsymbol{\psi}) = E_{\mathbf{y}^*|\mathbf{u}}[\ell_*(\boldsymbol{\psi}; \mathbf{y}^*)]$. By symmetry, $K(\boldsymbol{\psi}) = E_{\mathbf{y}|\mathbf{u}}[\ell_*(\boldsymbol{\psi}; \mathbf{y})]$. We have

$$\text{GAI} = -2E_{\mathbf{y}|\mathbf{u}}E_{\mathbf{y}^*|\mathbf{u}}[\ell_*(\widehat{\boldsymbol{\psi}}; \mathbf{y}^*)] = -2E_{\mathbf{y}|\mathbf{u}}[K(\widehat{\boldsymbol{\psi}})] = -2E_{\mathbf{y}|\mathbf{u}}[\ell_*(\widehat{\boldsymbol{\psi}}; \mathbf{y})] + A + B$$

where

$$A = -2E_{\mathbf{y}|\mathbf{u}}[\ell_*(\boldsymbol{\psi}_0; \mathbf{y}) - \ell_*(\widehat{\boldsymbol{\psi}}; \mathbf{y})], \quad B = -2E_{\mathbf{y}|\mathbf{u}}[K(\widehat{\boldsymbol{\psi}}) - \ell_*(\boldsymbol{\psi}_0; \mathbf{y})].$$

By Taylor expansion to the second order of $\ell_*(\hat{\boldsymbol{\psi}}; \mathbf{y}^*)$,

$$\begin{aligned}
 B &= -2\mathbb{E}_{\mathbf{y}|\mathbf{u}}\mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[\ell_*(\hat{\boldsymbol{\psi}}; \mathbf{y}^*) - \ell_*(\boldsymbol{\psi}_0; \mathbf{y})] \\
 &= -2\mathbb{E}_{\mathbf{y}|\mathbf{u}}\mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[\ell_*(\boldsymbol{\psi}_0; \mathbf{y}^*) - \ell_*(\boldsymbol{\psi}_0; \mathbf{y})] - 2\mathbb{E}_{\mathbf{y}|\mathbf{u}}\mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top \nabla \ell_*(\boldsymbol{\psi}_0; \mathbf{y}^*)] \\
 &\quad + \mathbb{E}_{\mathbf{y}|\mathbf{u}}\mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top \mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y}^*)(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)] + o(1) \\
 &= \text{tr}\{\mathbb{E}_{\mathbf{y}|\mathbf{u}}[n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})]\text{var}_{\mathbf{y}|\mathbf{u}}(\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0))\} + o(1) = \text{df} + o(1). \quad (\text{S1.1})
 \end{aligned}$$

In the second equality above, the first term $= -2\mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[\ell_*(\boldsymbol{\psi}_0; \mathbf{y}^*)] + 2\mathbb{E}_{\mathbf{y}|\mathbf{u}}[\ell_*(\boldsymbol{\psi}_0; \mathbf{y})] = 0$; the second term $= -2\mathbb{E}_{\mathbf{y}|\mathbf{u}}[(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top] \mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[\nabla \ell_*(\boldsymbol{\psi}_0; \mathbf{y}^*)] = 0$, and the third term is simplified as

$$\mathbb{E}_{\mathbf{y}|\mathbf{u}}\{(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top \mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y}^*)](\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)\} = \text{tr}\{\mathbb{E}_{\mathbf{y}^*|\mathbf{u}}[n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y}^*)]\text{var}(\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0))\} = \text{df}.$$

It remains to show that $A = \text{df} + o(1)$. Towards this, we let $\tilde{\mathbf{J}} = \mathbb{E}_{\mathbf{y}|\mathbf{u}}[n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})]$ and $\mathbf{Q} = n(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top$. By Taylor expansion to the second order of $\ell_*(\boldsymbol{\psi}_0; \mathbf{y})$,

$$\begin{aligned}
 A &= -2\mathbb{E}_{\mathbf{y}|\mathbf{u}}\{(\boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}})^\top \nabla \ell_*(\hat{\boldsymbol{\psi}}; \mathbf{y})\} + \mathbb{E}_{\mathbf{y}|\mathbf{u}}\{(\boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}})^\top \mathbf{J}(\hat{\boldsymbol{\psi}}; \mathbf{y})(\boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}})\} + o(1) \\
 &= \sum_{r,s} \mathbb{E}_{\mathbf{y}|\mathbf{u}}\{[n^{-1}\mathbf{J}(\hat{\boldsymbol{\psi}}; \mathbf{y})]_{rs} \mathbf{Q}_{sr}\} + o(1), \quad (\text{S1.2})
 \end{aligned}$$

where $\mathbf{Q}_{rs}$ denotes the rs -th entry of matrix $\mathbf{Q}$, and the others are similarly defined. Let $\hat{\boldsymbol{\psi}}_r$ and $\boldsymbol{\psi}_{0r}$ denote the r -th entry of the vectors $\hat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}_0$, respectively. Let $\tilde{J}_{rs} = \mathbb{E}_{\mathbf{y}|\mathbf{u}}\{n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})\}_{rs}$. We shall approximate the rs -th summand in (S1.2) by

Step 1: $E_{\mathbf{y}|\mathbf{u}}[\{n^{-1}\mathbf{J}(\hat{\boldsymbol{\psi}}; \mathbf{y})\}_{rs}\mathbf{Q}_{sr}] = E_{\mathbf{y}|\mathbf{u}}[\{n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})\}_{rs}\mathbf{Q}_{sr}] + o(1)$; and

Step 2: $E_{\mathbf{y}|\mathbf{u}}[\{n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})\}_{rs}\mathbf{Q}_{sr}] = \tilde{J}_{rs} \text{var}_{\mathbf{y}|\mathbf{u}}(\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0))_{sr} + o(1)$.

For Step 1, let $X_n = \mathbf{Q}_{sr} = n(\hat{\boldsymbol{\psi}}_r - \boldsymbol{\psi}_{0r})(\hat{\boldsymbol{\psi}}_s - \boldsymbol{\psi}_{0s})$, and $Y_n = n^{-1}[\mathbf{J}(\hat{\boldsymbol{\psi}}; \mathbf{y}) - \mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})]_{rs}$. Since $\hat{\boldsymbol{\psi}}$ is an MHL estimator and due to asymptotic normality, $X_n \rightarrow X$ in distribution. The dependence of X_n , X and Y_n on r and s are not indicated for simplicity. By mean value theorem, $Y_n = n^{-1} \sum_{i=1}^n \{(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top \mathbf{J}'(\bar{\boldsymbol{\psi}}; y_i)\}$, where $\bar{\boldsymbol{\psi}} = \boldsymbol{\psi}_0 + \epsilon(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)$ for some $\epsilon \in [0, 1]$. By the regularity condition in Lee and Lee (2025) that $\|\mathbf{J}'(\bar{\boldsymbol{\psi}}; y_i)\|$ is bounded by an integrable random variable, and that $(\hat{\boldsymbol{\psi}}_r - \boldsymbol{\psi}_{0r}) = O_p(n^{-1/2})$, then $Y_n \rightarrow 0$. Therefore, $X_n Y_n \rightarrow 0$ in distribution, and $E[X_n Y_n] \rightarrow 0$, proving Step 1.

For Step 2, we write $E_{\mathbf{y}|\mathbf{u}}[\{n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})\}_{rs}\mathbf{Q}_{sr}] = E_{\mathbf{y}|\mathbf{u}}[\{n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})\}_{rs}]E_{\mathbf{y}|\mathbf{u}}[\mathbf{Q}_{sr}] + c_{rs}$ where $c_{rs} = \text{cov}(n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})_{rs}, \mathbf{Q}_{sr})$. By Cauchy-Schwarz inequality,

$$\begin{aligned} |c_{rs}|^2 &\leq \text{var}_{\mathbf{y}|\mathbf{u}}(\mathbf{Q}_{sr}) \text{var}_{\mathbf{y}|\mathbf{u}}(n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})_{rs}) \\ &= \text{var}_{\mathbf{y}|\mathbf{u}}(\mathbf{Q}_{sr}) n^{-2} \sum_{i=1}^n \text{var}_{\mathbf{y}|\mathbf{u}}([\partial^2 \ell_*(\boldsymbol{\psi}_0; y_i) / \partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top]_{rs}) \\ &\leq C/n \text{var}_{\mathbf{y}|\mathbf{u}}(\mathbf{Q}_{sr}) \rightarrow 0, \end{aligned}$$

where C denotes an upper bound of $E_{\mathbf{y}|\mathbf{u}}([\partial^2 \ell_*(\boldsymbol{\psi}_0; y_i) / \partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top]_{rs}^2)$.

Using the approximations in Steps 1 and 2, (S1.2) becomes

$$\begin{aligned}
 A &= \sum_{r,s} \tilde{J}_{rs} \text{var}_{\mathbf{y}|\mathbf{u}}(\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0))_{sr} + o(1) \\
 &= \text{tr}\{\mathbf{E}_{\mathbf{y}|\mathbf{u}}[n^{-1}\mathbf{J}(\boldsymbol{\psi}_0; \mathbf{y})]\text{var}(\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0))\} + o(1) \\
 &= \text{df} + o(1).
 \end{aligned}$$

Combining this and (S1.1), (10) in the paper follows. This proves Theorem 1 when $\boldsymbol{\psi}$ contains $\mathbf{v}$.

When $\boldsymbol{\varsigma}$ contains $\mathbf{v}$ with $\mathbf{E} = \mathbf{E}_{\mathbf{y}}$, the proof is similar. This completes the proof of Theorem 1.

S2 Proof of Theorem 2

Proof of (i) Without loss of generality, we assume that the true model is M_1 , which gives the density of the true data generation,

$$g(\mathbf{y}) = g(\mathbf{y}|M_1) = \prod_{i=1}^n g_i(y_i|M_1) = \prod_{i=1}^n \int f(y_i|\boldsymbol{\theta}_1, M_1) |\partial\delta^1(\boldsymbol{\theta}_1)/\partial\boldsymbol{\theta}_1| d\boldsymbol{\theta}_1$$

for a parametrization function $\delta^1(\cdot)$. In this proof, we assume that the parametrization function $\delta^k(\cdot)$ for M_k is known for $1 \leq k \leq K$. Denote $D_k = \mathbf{E}_{\mathbf{y}}[\ell_0(M_k; \mathbf{y}, \boldsymbol{\delta}_k)] = n\mathbf{E}_y[\ell_0(M_k; y, \boldsymbol{\delta}_k)] = nd_k$ for $1 \leq k \leq K$. Let $\Delta_n(k) = \sum_{i=1}^n \ell_0(M_k; y_i, \boldsymbol{\delta}_k)$, for $1 \leq k \leq K, n \geq 1$; and

$$\widehat{M}_n = \arg \max\{\Delta_n(k) : 1 \leq k \leq K\}, \quad n \geq 1.$$

Since M_1 is the maximizer of $\{D_k : 1 \leq k \leq K\}$, it follows that $D_k < D_1$ and equivalently, $d_k < d_1$ for $2 \leq k \leq K$, and so there exists $\epsilon > 0$ such that $d_1 > \epsilon + d_k$ for $2 \leq k \leq K$. Assuming that $\ell_0(M_k; y_i, \boldsymbol{\delta}_k)$'s are of bounded variances, $\lim_{n \rightarrow \infty} n^{-1} \Delta_n(k) = d_k$ by strong law of large numbers. For this ϵ , there exist $N_1, \dots, N_K$ such that for $1 \leq k \leq K$,

$$d_k - \epsilon/2 < n^{-1} \Delta_n(k) < d_k + \epsilon/2, \quad n \geq N_k.$$

Let $N = \max\{N_1, \dots, N_K\}$, then for $n \geq N$ and $2 \leq k \leq K$,

$$n^{-1} \Delta_n(1) > d_1 - \epsilon/2 > d_k + \epsilon/2 > n^{-1} \Delta_n(k).$$

So, $\widehat{M}(n) = M_1 = \widehat{M}$ for $n \geq N$, and $E_{\mathbf{y}}[\text{MIC}(\delta)] = E_{\mathbf{y}}[\sum_{i=1}^n \ell_0(M_1; y_i, \delta^1)] = nd_1 = \max\{D_k : 1 \leq k \leq K\} = \text{MI}(\delta)$, proving part (i).

Proof of (ii) By Laplace approximation to the likelihood of the candidate model M_k , we have

$$\begin{aligned} L_0(M_k; \mathbf{y}, \boldsymbol{\delta}_k) &= \int L_{\mathbf{m}}(\boldsymbol{\theta}_k; \mathbf{y}) d\boldsymbol{\delta}_k = L_{\mathbf{m}}(\widehat{\boldsymbol{\theta}}_k; \mathbf{y}) |n\mathbf{J}_k/(2\pi)|^{-1/2} + o_p(1) \\ &= (2\pi)^{(p_k+t_k+1)/2} |\mathbf{J}_k|^{-1/2} n^{-(p_k+t_k+1)/2} L_{\mathbf{m}}(\widehat{\boldsymbol{\theta}}_k; \mathbf{y}) + o_p(1), \end{aligned}$$

where $\mathbf{J}_k = n^{-1} \{-\partial^2 \ell_m(\boldsymbol{\theta}_k; \mathbf{y}) / \partial \boldsymbol{\delta}_k \partial \boldsymbol{\delta}_k^\top\} |_{\boldsymbol{\theta}_k = \widehat{\boldsymbol{\theta}}_k} = O_p(1)$. The log-likelihood for M_k leads to

$$\text{BIC} = -2\ell_{\mathbf{m}}(\widehat{\boldsymbol{\theta}}_k; \mathbf{y}) + (p_k + t_k + 1) \log n = \text{MIC}(\boldsymbol{\delta}_k) + O_p(1).$$

This completes the proof of part (ii), and hence Theorem 2.

S3 Proof of Proposition 2

Let $a = \Gamma((n-2)/2)/(2\pi^{n/2})$. With $t = \|\mathbf{y} - \boldsymbol{\mu}\|^2/(2\phi)$, we have

$$\begin{aligned} L_i(\boldsymbol{\mu}; \mathbf{y}) &= \int L_c(\boldsymbol{\mu}, \phi; \mathbf{y}|\mathbf{v})d\phi \\ &= (2\pi)^{-n/2} \int_0^\infty \phi^{-n/2} \exp\{-\|\mathbf{y} - \boldsymbol{\mu}\|^2/(2\phi)\}d\phi \\ &= (2\pi)^{-n/2} \{\|\mathbf{y} - \boldsymbol{\mu}\|^2/2\}^{-n/2+1} \int_0^\infty t^{n/2-2} e^{-t} dt \\ &= a \{\|\mathbf{y} - \boldsymbol{\mu}\|^2\}^{-n/2+1}. \end{aligned}$$

S4 Proof of Theorem 3

We write $\hat{\boldsymbol{\mu}}$ for $\hat{\boldsymbol{\mu}}_{\text{MHLE}}$. For given ϕ and $\boldsymbol{\lambda}$, let $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}} = \mathbf{P}\mathbf{y}$, where

$$\mathbf{P} = [\mathbf{X} \ \mathbf{Z}] \begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{Z} \\ \mathbf{Z}^\top \mathbf{X} & \mathbf{Z}^\top \mathbf{Z} + \phi \boldsymbol{\Lambda}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^\top \\ \mathbf{Z}^\top \end{bmatrix}.$$

With $\mathbf{Q} = (\mathbf{I} - \mathbf{P})^2$ and $\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 = \mathbf{y}^\top \mathbf{Q} \mathbf{y}$, we shall show that

$$(n-2) \{ \mathbb{E}_{\mathbf{y}|\mathbf{u}} \mathbb{E}_{\mathbf{y}^*|\mathbf{u}} \log \|\mathbf{y}^* - \hat{\boldsymbol{\mu}}\|^2 - \mathbb{E}_{\mathbf{y}^*|\mathbf{u}} \log \|\mathbf{y}^* - \boldsymbol{\mu}\|^2 \} = \rho + o(\mathfrak{S}) \quad (4.3)$$

$$(n-2) \{ \mathbb{E}_{\mathbf{y}^*|\mathbf{u}} \log \|\mathbf{y}^* - \boldsymbol{\mu}\|^2 - \mathbb{E}_{\mathbf{y}|\mathbf{u}} \log \|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 \} = \rho + o(\mathfrak{S}) \quad (4.4)$$

Adding (S4.3) and (S4.4) gives $\text{SAI} - \mathbb{E}[\text{SAIC}] = o(1)$.

Let φ denote the digamma function, $\chi_n^2(c)$ a noncentral chi-squared distribution with n degrees of freedom and noncentral parameter c , and χ_n^2 a central chi-squared distribution with n degrees of freedom. We recall

known results about central and noncentral chi-squared distributions, and digamma function:

$$\begin{aligned} \mathbb{E}[\log \chi_n^2] &= \log 2 + \varphi(n/2), \\ \mathbb{E}[\log \chi_n^2(c)] &= \log 2 + \sum_{j=0}^{\infty} \frac{(c/2)^j e^{-c/2}}{j!} \varphi((n+2j)/2); \end{aligned} \quad (\text{S4.5})$$

$$\varphi(x+1) = \varphi(x) + \frac{1}{x}, \quad \varphi(x) = \log x - \frac{1}{2x} + O(x^{-2}); \quad (\text{S4.6})$$

$$2j/n - 4j^2/n^2 \leq \varphi(n/2 + j) - \varphi(n/2) \leq 2j/n. \quad (\text{S4.7})$$

Note that $\|\mathbf{y}^* - \hat{\boldsymbol{\mu}}\|^2 \sim \chi_n^2(\nu)$ with $\nu = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2$. Let $w_j = (\nu/2)^j e^{-\nu/2} / (j!)$, the point probability of a Poisson with mean $\nu/2$ at j . Thus, $\sum_{j=1}^{\infty} j w_j = \nu/2$ and $\sum_{j=1}^{\infty} j^2 w_j = \nu^2/4 + \nu/2$. By (S4.5) and (S4.7),

$$\begin{aligned} \text{LHS of (S4.3)} &= (n-2) \mathbb{E}_{\mathbf{y}|\mathbf{u}} \left\{ \sum_{j=1}^{\infty} w_j [\varphi(n/2 + j) - \varphi(n/2)] \right\} \\ &= (n-2) \mathbb{E}_{\mathbf{y}|\mathbf{u}} \left\{ \sum_{j=1}^{\infty} w_j \{ \varphi(n/2 + j) - \varphi(n/2) \} \right\} \\ &\leq (n-2) \mathbb{E}_{\mathbf{y}|\mathbf{u}} \left\{ \sum_{j=1}^{\infty} 2j w_j / n \right\} = \mathbb{E}_{\mathbf{y}|\mathbf{u}}(\nu) + o(1) \\ &= \rho + o(1). \end{aligned}$$

Similarly, $\text{LHS of (S4.3)} \geq (n-2) \mathbb{E}_{\mathbf{y}|\mathbf{u}}(\nu/n - (\nu^2 + 2\nu)/n^2) = \mathbb{E}_{\mathbf{y}|\mathbf{u}}(\nu) + o(1) = \rho + o(1)$. Combining the upper and lower bounds, we have proved (S4.3).

Let $U = \|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2$, $\tau = \mathbb{E}_{\mathbf{y}|\mathbf{u}}[U]$, $\sigma^2 = \text{var}_{\mathbf{y}|\mathbf{u}}(U)$ and $W \sim \chi_{n-\rho}^2$.

We have $\tau = n - \rho + o(1)$, $\sigma^2 = 2(n - \rho) + o(1)$, $\mathbb{E}[W] = n - \rho$ and $\text{var}(W) = 2(n - \rho)$.

We have $\log U = \log\{\tau[1 + (U - \tau)/\tau]\} = \log(\tau) + (U - \tau)/\tau - (U - \tau)^2/(2\tau^2) + O_P(n^{-3/2})$, and a similar expansion for W . Taking expectation of each expansion, and subtracting them, we have

$$\begin{aligned} E_{\mathbf{y}|\mathbf{u}} \log\{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2/W\} &= \log\{\tau/(n - \rho)\} - \sigma^2/(2\tau^2) + 1/(n - \rho) + O(n^{-3/2}) \\ &= O(n^{-3/2}). \end{aligned}$$

We proceed to prove (S4.4). Recall that $\|\mathbf{y}^* - \boldsymbol{\mu}\|^2 \sim \chi_n^2$, and definition of W . By (S4.5) and (S4.6),

$$\begin{aligned} \text{LHS of (S4.4)} &= (n - 2)\{E_{\mathbf{y}^*|\mathbf{u}} \log[\|\mathbf{y}^* - \boldsymbol{\mu}\|^2/W] - E_{\mathbf{y}|\mathbf{u}} \log[\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2/W]\} \\ &= (n - 2)[\varphi(n) - \varphi(n - \rho) + O(n^{-3/2})] \\ &= (n - 2)[- \log(1 - \rho/n) + O(n^{-3/2})] \\ &= \rho + o(1), \end{aligned}$$

where we applied (S4.6) in the second last equality. Hence, the proof of (S4.4) and Theorem 3.

S5 Detailed description of datasets

Various information criteria have been proposed for model selection; however, difficulties arise when different criteria select different models. These differences reflect distinct inferential objectives, and the criteria should

therefore be applied in a manner aligned with the specific goals of the analysis. For illustration, we consider four models derived from Model (2.1)

M1: GLM with $\mathbf{v} = \mathbf{0}$.

M2: GLM with fixed $\mathbf{v}$.

M3: GLMM with $\mathbf{v} \sim N(0, \mathbf{\Lambda})$, where $\mathbf{\Lambda} = \lambda_0 \mathbf{G}$.

M4: robust HGLM with $\mathbf{v}|\mathbf{u} \sim N(0, \mathbf{\Lambda}(\mathbf{u}))$, inverse-gamma($\alpha+1, \alpha$), u_i , $E(u_i) = 1$ and $\text{var}(u_i) = \delta$, where u_i is the i -th entry of $\mathbf{u}$ in Example 1 below. We apply these GAICs for component model selection in the four real datasets below.

The first three real data examples are chosen for LMM in order to illustrate the effects of model selection with increasing sample sizes, where $n = 60, 241$ and 4059 . The last example is concerned with count data. We use the second example to illustrate the log-normal and gamma random effects models.

Example 1 (Cadrilazine data). Vaida and Blanchard (2005) and Liang et al. (2008) applied CAIC for model selection of the Cadrilazine dataset in Wakefield et al. (1994) with $n = 60$, which consists of plasma drug concentrations from 10 cardiac failure patients who were given a single intravenous dose of 30 mg of cadralazine, an anti-hypertensive drug. Each subject has the plasma drug concentration (y), in mg/l, measured at 2, 4,

6, 8, 10 and 24 hours (t), for a total of 6 observations per subject (Wakefield et al., 1994). Vaida and Blanchard (2005) removed two of the ten subjects with observations at 28 and 32 hours in order to work with a simple balanced dataset. Also, three concentrations of 0 were replaced with 0.005 mg. For log-transformation of drug concentration y_{ij} in the $i(= 1, \dots, 10)$ th subject and the $j(= 1, \dots, 6)$ th time, they considered the two-component random effects model

$$\log(y_{ij}) = \beta_0 + \beta_1 t_j + v_{0i} + t_j v_{1i} + e_{ij}, \quad e_{ij} \sim N(0, \phi).$$

We consider M1 with $v_{0i} = 0$ and $v_{1i} = 0$; M2 with fixed v_{0i} and v_{1i} ; M3 with $(v_{0i}, v_{1i})^\top \sim N(\mathbf{0}, \mathbf{\Lambda}_i)$ where $\mathbf{\Lambda}_i = \lambda_1 \mathbf{G}$, $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2]$, $\mathbf{g}_1 = (1, \lambda_3 \sqrt{\lambda_2/\lambda_1})^\top$, $\mathbf{g}_2 = (\lambda_3 \sqrt{\lambda_2/\lambda_1}, \lambda_2/\lambda_1)^\top$ and $-1 \leq \lambda_3 \leq 1$. M4 with $(v_{0i}, v_{1i})^\top \sim N(0, u_i \mathbf{\Lambda}_i)$, and u_i are independent inverse-gamma($\alpha + 1, \alpha$).

GLMs with independent errors, M1 and M2 correspond to the random effects models with $\lambda_1 = 0$ and $\lambda_1 = \infty$, respectively (Lee and Nelder, 1996). Thus, random effects model is a compromise between two GLMs, M1 and M2. For robust HGLMs, Lee and Nelder (2006b) preferred this parametrization for $E(u_i) = 1$ to the conventional t -distribution because the first two marginal moments, $E(\mathbf{v}) = \mathbf{0}$ and $\text{var}(\mathbf{v}) = \mathbf{\Lambda} = E[\mathbf{\Lambda}(\mathbf{u})]$ remain the same as those of the normal distribution, while higher moments are different. The use of exponential u_i leads to LASSO estimation in Bayesian

approach (Park and Casella, 2008) and that in frequentist approach (Tibshirani, 1996). For further variable selection and estimation of elements of $\boldsymbol{\psi}$, LASSO type estimators can replace $\hat{\boldsymbol{\psi}}_{\text{MHLE}}$ in GAIC for $\boldsymbol{\psi} = \boldsymbol{\beta}$ (Lee and Oh, 2014) and $\boldsymbol{\psi} = \boldsymbol{\phi}$ (Zeng et al., 2025). However, in this paper, we restrict to inverse gamma u_i for robust HGLMs; see Lee et al. (2017) for further development. The robust HGLM, M4, being shrunk more than M3, leans more towards M1 than ordinary random effects model, M3. Models M1, M4, M3 and M2 increase in order of $\dim(\mathbf{v})$; whereas M1, M4, M3 and M2 decrease in the (deviance) residual sum of squares when effects $\mathbf{v}$ are present.

Example 2 (Crack Growth data). In the Crack Growth data of Lu and Meeker (1993), there are 21 metallic specimens, each subjected to 120,000 loading cycles, with the crack lengths recorded every 10^4 cycles. We take $t = \text{no. of cycles}/10^6$ here; hence $t_j = j/100$ for $j = 1, \dots, 12$. The crack increment sequences look rather irregular. Let l_{ij} be the crack length of the i th specimen at the j th observation and let $y_{ij} = l_{ij} - l_{i(j-1)}$ be the corresponding increment in the crack length, which is always a positive value. For response y_{ij} , we consider the model with $x_{ij} = l_{i(j-1)}$

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + v_i. \quad (\text{S5.8})$$

We consider M1 with $v_i = 0$; M2 with fixed v_i ; M3 with $v_i \sim N(0, \lambda)$; and

M4 with $v_i \sim N(0, u_i \lambda)$ and u_i are independent inverse-gamma($\alpha + 1, \alpha$).

Example 3 (Exam data). We consider a data-set for exam scores of $n = 4,059$ students from 65 schools in inner London taken from the R package ‘mlmRev’. For normalized exam score y_{ij} in the $i(= 1, \dots, 65)$ th school and the j th observation with standardized LR test score as a covariate x_{ij} . We consider model (S5.8).

Example 4 (Crack Growth data revisited). Because crack growth is positive, we can apply LMM for responses $\log y_{ij}$, which is a log-normal random effects model. We also apply gamma random effects model with the response y_{ij} and the linear predictor $\eta_{ij} = \log \mu_{ij}$ in (S5.8). In log-normal model, the log-likelihood is $\ell_*(\boldsymbol{\psi}; \log \mathbf{y})$ whereas in LMM and gamma models, it is $\ell_*(\boldsymbol{\psi}; \mathbf{y})$. In order to compare these three models, a correction term should be added in the log-normal model due to the Jacobian term $|\partial \log \mathbf{y} / \partial \mathbf{y}|$ (Lee et al., 2017).

Example 5 (Epilepsy data). Thall and Vail (1990) presented longitudinal data with $n = 236$ from a clinical trial of 59 epileptics who were randomly assigned to a new drug or a placebo with five covariates in Lee et al. (2017). A multivariate response variable consisted of the seizure counts during the two-week periods before each of the four visits to the clinic. Let y_{ij} be the corresponding response variable for the $i(= 1, \dots, 59)$ th patient and

the $j(= 1, \dots, 4)$ th visit. We consider four Poisson models with linear predictor, $\eta_{ij} = \log(\mu_{ij}) = \beta_0 + x_{ij}^\top \boldsymbol{\beta} + v_i$. We let M1 with $v_i = 0$; M2 with fixed v_i ; M3 with $v_i \sim N(0, \lambda)$; and M4 with $v_i \sim N(0, u_i \lambda)$ and u_i is inverse-gamma($\alpha + 1, \alpha$) distributed.

The Cadralazine data can be found in Wakefield et al. (1994), Crack Growth Data in Lu and Meeker (1993), and Exam Data from the R package ‘mlmRev’.

S6 Table A

Bibliography

- Lee, H. and Y. Lee (2025). Statistical inference for random unknowns via modifications of extended likelihood. *arXiv*: 2310.09955v3.
- Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(4), 619–656.
- Lee, Y. and J. A. Nelder (2006). Fitting via alternative random-effect models. *Statistics and Computing* 16, 69–75.
- Lee, Y., J. A. Nelder, and Y. Pawitan (2017). *Generalized linear models with random effects: unified analysis via H-likelihood*. Chapman and Hall/CRC.
- Lee, Y. and H.-S. Oh (2014). A new sparse variable selection via random-effect model. *Journal of Multivariate Analysis* 125, 89–99.
- Liang, H., H. Wu, and G. Zou (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* 95(3), 773–778.
- Lu, C. J. and W. O. Meeker (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics* 35(2), 161–174.

- Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 103(482), 681–686.
- Thall, P. F. and S. C. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* 46(3), 657–671.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1), 267–288.
- Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed effects models. *Biometrika* 92, 351–370.
- Wakefield, J. C., A. F. Smith, A. Racine-Poon, and A. E. Gelfand (1994). Bayesian analysis of linear and non-linear population models by using the gibbs sampler. *Journal of the Royal Statistical Society Series C: Applied Statistics* 43(1), 201–221.
- Zeng, Y., Z. Wang, Y. Lee, and N. Tang (2025). Simultaneous outlier detection and prediction for kriging with true identification. *Journal of Computational and Graphical Statistics*, 1–12.

Table A: Parameters ψ and ς in M1–M4 for each information criterion. $\mathbf{0}, \mathbf{1}$ are column vectors of zeros and ones, respectively. $\mathbf{u}, \alpha$ and λ_1 are introduced in Example 1.

		M1	M2	M3	M4
Cadrilazine, Crack growth and Exam datasets					
BIC	ψ	null	null	null	null
	ς	$\beta, \phi, \lambda_1 = 0, \mathbf{v} = \mathbf{0}$	$\beta, \phi, \lambda_1 = \infty, \mathbf{v}$	$\beta, \mathbf{v}, \phi, \boldsymbol{\lambda}, \alpha = \infty, \mathbf{u} = \mathbf{1}$	$\beta, \mathbf{v}, \phi, \boldsymbol{\lambda}, \alpha, \mathbf{u}$
MAIC	ψ	$\beta, \phi, \lambda_1 = 0, \mathbf{v} = \mathbf{0}$	$\beta, \phi, \lambda_1 = \infty, \mathbf{v}$	$\beta, \phi, \boldsymbol{\lambda}, \alpha = \infty$	$\beta, \phi, \boldsymbol{\lambda}, \alpha$
	ς	null	null	$\mathbf{v}, \mathbf{u} = \mathbf{1}$	$\mathbf{v}, \mathbf{u}$
RAIC	ψ	$\phi, \lambda_1 = 0$	$\phi, \lambda_1 = \infty$	$\phi, \boldsymbol{\lambda}, \alpha = \infty$	$\phi, \boldsymbol{\lambda}, \alpha$
	ς	$\beta, \mathbf{v} = \mathbf{0}$	$\beta, \mathbf{v}$	$\beta, \mathbf{v}, \mathbf{u} = \mathbf{1}$	$\beta, \mathbf{v}, \mathbf{u}$
CAIC	ψ	$\beta, \phi, \mathbf{v} = \mathbf{0}$	$\beta, \phi, \mathbf{v}$	$\beta, \phi, \mathbf{v}$	$\beta, \phi, \mathbf{v}$
	ς	$\lambda_1 = 0$	$\lambda_1 = \infty$	$\boldsymbol{\lambda}, \alpha = \infty, \mathbf{u} = \mathbf{1}$	$\boldsymbol{\lambda}, \alpha, \mathbf{u}$
SAIC	ψ	$\beta, \mathbf{v} = \mathbf{0}$	$\beta, \mathbf{v}$	$\beta, \mathbf{v}$	$\beta, \mathbf{v}$
	ς	$\phi, \lambda_1 = 0$	$\phi, \lambda_1 = \infty$	$\phi, \boldsymbol{\lambda}, \alpha = \infty, \mathbf{u} = \mathbf{1}$	$\phi, \boldsymbol{\lambda}, \alpha, \mathbf{u}$
Epilepsy dataset					
BIC	ψ	null	null	null	null
	ς	$\beta, \phi = 1, \lambda_1 = 0, \mathbf{v} = \mathbf{0}$	$\beta, \phi = 1, \lambda_1 = \infty, \mathbf{v}$	$\beta, \mathbf{v}, \phi = 1, \boldsymbol{\lambda}, \alpha = \infty, \mathbf{u} = \mathbf{1}$	$\beta, \mathbf{v}, \phi = 1, \boldsymbol{\lambda}, \alpha, \mathbf{u}$
MAIC	ψ	$\beta, \phi = 1, \lambda_1 = 0, \mathbf{v} = \mathbf{0}$	$\beta, \phi = 1, \lambda_1 = \infty, \mathbf{v}$	$\beta, \phi = 1, \boldsymbol{\lambda}, \alpha = \infty$	$\beta, \phi = 1, \boldsymbol{\lambda}, \alpha$
	ς	null	null	$\mathbf{v}, \mathbf{u} = \mathbf{1}$	$\mathbf{v}, \mathbf{u}$
RAIC	ψ	$\phi = 1, \lambda_1 = 0$	$\phi = 1, \lambda_1 = \infty$	$\phi = 1, \boldsymbol{\lambda}, \alpha = \infty$	$\phi = 1, \boldsymbol{\lambda}, \alpha$
	ς	$\beta, \mathbf{v} = \mathbf{0}$	$\beta, \mathbf{v}$	$\beta, \mathbf{v}, \mathbf{u} = \mathbf{1}$	$\beta, \mathbf{v}, \mathbf{u}$
CAIC	ψ	$\beta, \phi = 1, \mathbf{v} = \mathbf{0}$	$\beta, \phi = 1, \mathbf{v}$	$\beta, \phi = 1, \mathbf{v}$	$\beta, \phi = 1, \mathbf{v}$
	ς	$\lambda_1 = 0$	$\lambda_1 = \infty$	$\boldsymbol{\lambda}, \alpha = \infty, \mathbf{u} = \mathbf{1}$	$\boldsymbol{\lambda}, \alpha, \mathbf{u}$