information criterion (TIC; Takeuchi, 1976), the regularization information criterion (RIC; Shibata, 1989), the generalized information criterion (GIC; Konishi & Kitagawa, 1996, 2003), the extended information criterion (EIC; Ishiguro et al., 1997) and AIC_b (Cavanaugh and Shumway, 1997) among others.

In each of these information criteria, $D+a$ is an estimator of -2 times the predictive expected log-L, which is essentially equivalent to the Kullback-Leibler (1951) information or distance. The penalty term a is also seen as that of bias correction up to order $O(1)$ for AIC when the model holds (correctly specified model) or includes a correct model (overspecified model) and for the other information criteria given above when the model is possibly misspecified or underspecified. The stochastic a of TIC is given by observed information matrices or its expectations followed by substituting the maximum likelihood estimators (MLEs) for parameters. In EIC and AIC_b the term a is obtained using the bootstrap (see also Kishino & Hasegawa, 1989; Shibata, 1997; Shimodaira & Hasegawa, 1999; Konishi & Kitagawa, 2008).

While D is typically given by the MLEs of parameters in these criteria, RIC and GIC extended it to include robust and ridge-type estimators of parameters. Ogasawara (2015a) gives the asymptotic cumulants of $n^{-1}\mathrm{AIC}_\mathrm{W}$ and $n^{-1}\mathrm{TIC}_\mathrm{W}$, where the subscript W indicates that the weighted score estimators (WSEs) of parameters are used in place of the MLEs in $n^{-1}\mathrm{AIC}$ and $n^{-1}\mathrm{TIC}$. Note that the MLE is a special case of the WSE when the weight is null. The factor n^{-1} is not for essential transformation but for tractability so that $n^{-1}\mathrm{AIC}_\mathrm{W}$ and $n^{-1}\mathrm{TIC}_\mathrm{W}$ are $O_p(1)$ as for the usual MLEs of parameters.

A well-known exception of (1.1) is the Bayes information criterion (BIC; Schwarz, 1978), where a is replaced by $a_n = q\log n = O(\log n)$. The original definition of Mallows' (1973) $C_P(= O_p(1))$ for multiple linear regression is also an exception of (1.1). However, C_P can be transformed to the form of (1.1) using the predictive mean square errors (MSEs) of observations (Fujikoshi and Satoh, 1997). Further, it can be shown that C_P is asymptotically equivalent to AIC under the assumption of normality with a correct model or overspecified model.

Assume that we have n independent and identically distributed observations. Divide (1.1) by n when $D = -2\log\widehat{L}_\mathrm{W} \equiv -2\widehat{l}_\mathrm{W}$ and $a = -\widehat{b}$, where $\widehat{l}_\mathrm{W} = \log\widehat{L}_\mathrm{W} = \log L(\mathbf{X},\widehat{\boldsymbol{\theta}}_\mathrm{W}) = l(\mathbf{X},\widehat{\boldsymbol{\theta}}_\mathrm{W})$; $\mathbf{X}$ is the $n\times p$ matrix of n observations for p variables; $\widehat{\boldsymbol{\theta}}_\mathrm{W}$ is the $q\times 1$ vector of the WSEs of parameters defined explicitly in the next section; $n^{-1}b$ is the bias of $-n^{-1}2\widehat{l}_\mathrm{W} = O_p(1)$ up to order $O(n^{-1})$ under possible model misspecification;

and $\widehat{b} = \widehat{b}^{(1)}_{\mathrm{W}} = b(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{\mathrm{W}}), \widehat{b} = \widehat{b}^{(2)}_{\mathrm{W}} = b(\widehat{\boldsymbol{\theta}}_{\mathrm{W}}), \widehat{b} = \widehat{b}^{(1)}_{\mathrm{ML}} = b(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ or $\widehat{b} = \widehat{b}^{(2)}_{\mathrm{ML}} = b(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ is a consistent estimator of b. Then, (1.1) becomes

$$-n^{-1}2\widehat{l}_{\mathrm{W}} - n^{-1}\widehat{b} \equiv -2\widehat{\bar{l}}_{\mathrm{W}} - n^{-1}\widehat{b}, \tag{1.2}$$

where $\widehat{\bar{l}}_{\mathrm{W}} = n^{-1}\widehat{l}_{\mathrm{W}} = O_p(1)$ is the sample log-L given by $\mathbf{X}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}(= \boldsymbol{\theta}_{\mathrm{W}}(\mathbf{X}))$ averaged over observations. The population counterpart of $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ denoted by $\boldsymbol{\theta}_0$ under model misspecification is defined as the solution of $\boldsymbol{\theta}$ given by the expectation of the first order condition of $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$, where the expectation is that of the true distribution rather the fitted model.

Note that (1.2) is seen as a case of usual bias correction for e.g., a MLE of order $O_p(1)$ when we see $D = -2\widehat{\bar{l}}_{\mathrm{W}}$ as an estimator of a "parameter". Let $\widehat{\theta}_{\mathrm{W}}$ be an element of $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$. Ogasawara (2013, 2014, 2015b) dealt with the bias adjusted estimator $\widehat{\theta}_{\mathrm{W}(k)} \equiv \widehat{\theta}_{\mathrm{W}} - n^{-1}k\widehat{\alpha}_1$, where $n^{-1}\alpha_1$ is the bias of $\widehat{\theta}_{\mathrm{W}}$ up to order $O(n^{-1})$, $\widehat{\alpha}_1$ is a consistent estimator of α_1 and k is a constant. In Ogasawara (2013, 2014, 2015b), the optimal value k was derived such that k minimizes the asymptotic mean square error (AMSE) of $\widehat{\theta}_{\mathrm{W}}$ up to order $O(n^{-2})$.

Recall that $-2\widehat{\bar{l}}_{\mathrm{W}}$ is an estimator of $-2n^{-1}$ times the predictive expected log-L. In this paper, Ogasawara's method of bias adjustment is applied to information criteria. Further, another type of adjustment for $-2\widehat{\bar{l}}_{\mathrm{W}}$ by shrinkage is considered in combination with bias correction. In the following sections, it is shown that the AMSEs of $-2\widehat{\bar{l}}_{\mathrm{W}}$ by the adjustments are smaller than that of (1.2). Three examples using basic distributions in statistics are shown under model misspecification or correct model specification. A simulation and a real example for model selection in linear regression are also given.

2 Optimal Bias Adjustment of TIC

Define $f(\mathbf{X}^* = \mathbf{X} \,|\, \boldsymbol{\theta}_0) = f(\mathbf{X} \,|\, \boldsymbol{\theta}_0)$ as a probability density or mass of $\mathbf{X}^*$ at $\mathbf{X}$ when the model holds while $g(\mathbf{X}^* = \mathbf{X} \,|\, \boldsymbol{\zeta}_0) = g(\mathbf{X} \,|\, \boldsymbol{\zeta}_0) = g(\mathbf{X})$ is the corresponding true density or mass under model misspecification with $\boldsymbol{\zeta}_0$ being a vector of the population parameters of an appropriate size. Let $\boldsymbol{x}^{*\prime}_j$ be the j-th row of $\mathbf{X}^*(j = 1, \ldots, n)$. Then, the log-$L$ for $\boldsymbol{\theta}$ averaged over observations is

$$\bar{l} \equiv \bar{l}(\boldsymbol{\theta} \,|\, \mathbf{X}^*) \equiv n^{-1}l \equiv n^{-1}\sum_{j=1}^{n} l_j \equiv n^{-1}\sum_{j=1}^{n} \log f(\boldsymbol{x}^*_j \,|\, \boldsymbol{\theta}). \tag{2.1}$$

Let $\mathbf{Z}^*$ be an independent copy of $\mathbf{X}^*$ to be obtained in the future. Denote n^{-1} times the expected log-L by $\bar{l}_0^*$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ from a viewpoint of prediction:

$$\bar{l}_0^* \equiv \mathrm{E}_g\{\bar{l}(\boldsymbol{\theta}_0 \,|\, \mathbf{Z}^*)\} \equiv \int_{R(\mathbf{Z})} \bar{l}(\boldsymbol{\theta}_0 \,|\, \mathbf{Z}) g(\mathbf{Z}) d\mathbf{Z} (= \mathrm{E}_g\{\bar{l}(\boldsymbol{\theta}_0 \,|\, \mathbf{X}^*)\}), \tag{2.2}$$

where the summation is to be used when $g(\cdot)$ is a probability function. The sample counterpart of (2.2) using $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ is

$$\begin{aligned}\widehat{\bar{l}}_{\mathrm{W}}^* \equiv \mathrm{E}_g[\bar{l}\{\boldsymbol{\theta}_{\mathrm{W}}(\mathbf{X}^*) \,|\, \mathbf{Z}^*\}] &= \int_{R(\mathbf{Z})} \bar{l}\{\boldsymbol{\theta}_{\mathrm{W}}(\mathbf{X}^*) \,|\, \mathbf{Z}\} g(\mathbf{Z}) d\mathbf{Z} \\ &= \int_{R(\mathbf{Z})} \bar{l}(\widehat{\boldsymbol{\theta}}_{\mathrm{W}} \,|\, \mathbf{Z}) g(\mathbf{Z}) d\mathbf{Z}, \end{aligned} \tag{2.3}$$

where the vector $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ of WSEs addressed earlier is defined as the solution of

$$\frac{\partial \bar{l}(\boldsymbol{\theta} \,|\, \mathbf{X}^*)}{\partial \boldsymbol{\theta}} + n^{-1}\boldsymbol{q}^* = \mathbf{0}, \tag{2.4}$$

$\boldsymbol{q}^*$ is the $q \times 1$ vector of weights for the WSEs, which correspond to the log-prior derivatives in the case of Bayesian estimation but can be other weights.

The expected predictive log-L averaged over observations when $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\mathrm{W}} = \boldsymbol{\theta}_{\mathrm{W}}(\mathbf{X}^*)$ is

$$\mathrm{E}_g(\widehat{\bar{l}}_{\mathrm{W}}^*) = \int_{R(\mathbf{X})} \int_{R(\mathbf{Z})} \bar{l}\{\boldsymbol{\theta}_{\mathrm{W}}(\mathbf{X}) \,|\, \mathbf{Z}\} g(\mathbf{Z}) g(\mathbf{X}) d\mathbf{Z} d\mathbf{X}. \tag{2.5}$$

On the other hand, the usual sample log-L averaged over observations denoted by $\widehat{\bar{l}}_{\mathrm{W}}$ using $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ is

$$\widehat{\bar{l}}_{\mathrm{W}} = \bar{l}(\widehat{\boldsymbol{\theta}}_{\mathrm{W}} \,|\, \mathbf{X}^*) = \bar{l}\{\boldsymbol{\theta}_{\mathrm{W}}(\mathbf{X}^*) \,|\, \mathbf{X}^*\}. \tag{2.6}$$

When we use information criteria, $\mathrm{E}_g(\widehat{\bar{l}}_{\mathrm{W}}^*)$ of (2.5) is of primary interest although $\bar{l}_0^*$ of (2.2) is also of interest since $\bar{l}_0^*$ is the limiting value of $\widehat{\bar{l}}_{\mathrm{W}}^*$ when n goes to infinity. The bias of $-2\widehat{\bar{l}}_{\mathrm{W}}$ as an estimator of $-2\mathrm{E}_g(\widehat{\bar{l}}_{\mathrm{W}}^*)$ is assumed to be of order $O_p(n^{-1})$ i.e.,

$$-2\mathrm{E}_g(\widehat{\bar{l}}_{\mathrm{W}}) + 2\mathrm{E}_g(\widehat{\bar{l}}_{\mathrm{W}}^*) = n^{-1}b + n^{-2}c + O(n^{-3}), \tag{2.7}$$

where b, which is unchanged when $\mathrm{W} = \mathrm{ML}$, is well-known:

$$\begin{aligned} b &= 2\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}), \boldsymbol{\Lambda} = \mathrm{E}_g\left(\left.\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \equiv \mathrm{E}_g\left(\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'}\right), \\ \boldsymbol{\Gamma} &\equiv n\mathrm{E}_g\left(\frac{\partial \bar{l}}{\partial\boldsymbol{\theta}_0}\frac{\partial \bar{l}}{\partial\boldsymbol{\theta}_0'}\right) \end{aligned} \tag{2.8}$$

(see e.g., Ogasawara, 2015a) and the general formulas of c are available (see Konishi & Kitagawa, 2003 and Ogasawara, 2015a) but are not used in this paper.

Under correct model specification, $-\boldsymbol{\Lambda} = \boldsymbol{\Gamma} = \mathbf{I}$, where $\mathbf{I}$ is the Fisher information matrix per observation. For $n^{-1}\mathrm{TIC}_\mathrm{W}$ addressed in the introductory section, the estimator $\widehat{b}_\mathrm{W}^{(1)}$ or $\widehat{b}_\mathrm{W}^{(2)}$ for b in (2.8) is used as follows:

$$\begin{aligned} \widehat{b}_\mathrm{W}^{(1)} &\equiv 2\,\mathrm{tr}(\widehat{\boldsymbol{\Lambda}}_\mathrm{W}^{-1}\widehat{\boldsymbol{\Gamma}}_\mathrm{W}) \\ &\equiv 2\,\mathrm{tr}\left\{\left(\left.\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\mathrm{W}}\right)^{-1} n^{-1}\sum_{j=1}^{n}\left.\frac{\partial l_j}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\mathrm{W}}\left.\frac{\partial l_j}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\mathrm{W}}\right\} \\ &\equiv 2\,\mathrm{tr}\left\{\left(\frac{\partial^2 \bar{l}}{\partial\widehat{\boldsymbol{\theta}}_\mathrm{W}\partial\widehat{\boldsymbol{\theta}}_\mathrm{W}'}\right)^{-1} n^{-1}\sum_{j=1}^{n}\frac{\partial l_j}{\partial\widehat{\boldsymbol{\theta}}_\mathrm{W}}\frac{\partial l_j}{\partial\widehat{\boldsymbol{\theta}}_\mathrm{W}'}\right\}, \\ \widehat{b}_\mathrm{W}^{(2)} &\equiv 2\,\mathrm{tr}(-\widehat{\mathbf{I}}_\mathrm{W}^{(\boldsymbol{\Lambda})-1}\widehat{\mathbf{I}}_\mathrm{W}^{(\boldsymbol{\Gamma})}) \\ &\equiv 2\,\mathrm{tr}\left(\left[\left\{\mathrm{E}_g\left(\frac{\partial^2 \bar{l}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\right\}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\mathrm{W}}\right]^{-1}\mathrm{E}_g\left(\frac{\partial l_j}{\partial\boldsymbol{\theta}}\frac{\partial l_j}{\partial\boldsymbol{\theta}'}\right)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\mathrm{W}}\right), \end{aligned} \tag{2.9}$$

where $\widehat{b}_\mathrm{W}^{(2)}$ is generally simpler than $\widehat{b}_\mathrm{W}^{(1)}$ though $\widehat{b}_\mathrm{W}^{(2)}$ is not always available whereas $\widehat{b}_\mathrm{W}^{(1)}$ is available even when $g(\mathbf{X}\,|\,\boldsymbol{\zeta}_0)$ is unknown.

Define

$$n^{-1}\mathrm{TIC}_{\mathrm{W}(k)}^{(j)} = -2\widehat{\bar{l}}_\mathrm{W} - n^{-1}k\widehat{b}_\mathrm{W}^{(j)}\ (j = 1, 2). \tag{2.10}$$

When $k = 1$ and $\mathrm{W} = \mathrm{ML}$, (2.10) becomes

$$n^{-1}\mathrm{TIC}_{\mathrm{ML}(1)}^{(j)} = n^{-1}\mathrm{TIC}_\mathrm{ML}^{(j)} = n^{-1}\mathrm{TIC}^{(j)}\ (j = 1, 2), \tag{2.11}$$

where $\mathrm{TIC}^{(2)} = \mathrm{TIC}$ is the original definition of TIC by Takeuchi (1976). Ogasawara (2015a) showed that under regularity conditions $-2\widehat{\bar{l}}_{\mathrm{W}}$ is expanded as

$$\begin{aligned}
- \quad & 2\widehat{\bar{l}}_{\mathrm{W}} = -2(\bar{l}_0^*)_{O(1)} - 2(\bar{l}_0 - \bar{l}_0^*)_{O_p(n^{-1/2})} + \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0'}\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)_{O_p(n^{-1})} \\
+ \quad & \left[-\mathrm{vec}'(\mathbf{M})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<2>} + \frac{1}{3}\mathrm{vec}'\{\mathrm{E}_g(\mathbf{J}_0^{(3)})\}\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<3>}\right]_{O_p(n^{-3/2})} \\
+ \quad & O_p(n^{-2}) \\
\equiv \quad & -2(\bar{l}_0^*)_{O(1)} + \sum_{j=1}^{3}(\bar{l}_{\mathrm{ML}}^{(j)})_{O_p(n^{-j/2})} + O_p(n^{-2}), \qquad (2.12)
\end{aligned}$$

where $\bar{l}_0 = \bar{l}(\boldsymbol{\theta}_0 \,|\, \mathbf{X}), \mathbf{M} = \dfrac{\partial^2 \bar{l}}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} - \mathrm{E}_g\left(\dfrac{\partial^2 \bar{l}}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right), \mathbf{J}_0^{(3)} = \dfrac{\partial^3 \bar{l}}{\partial \boldsymbol{\theta}_0 (\partial \boldsymbol{\theta}_0')^{<2>}}$, $\mathrm{vec}(\cdot)$ is the vectorizing operator stacking the columns of a matrix sequentially, $\mathrm{vec}'(\cdot) = \{\mathrm{vec}(\cdot)\}'$, $\mathbf{x}^{<k>} = \mathbf{x} \otimes \cdots \otimes \mathbf{x}$ (k times of $\mathbf{x}$) is the k-fold Kronecker product of $\mathbf{x}$ and $(\cdot)_{O_p(n^{-j/2})}$ indicates that $(\cdot)$ is of order $O_p(n^{-j/2})$ for clarity. Equation (2.12) shows that $-2\widehat{\bar{l}}_{\mathrm{W}}$ is equal to $-2\widehat{\bar{l}}_{\mathrm{ML}}$ up to order $O_p(n^{-3/2})$. Similarly, it can be shown that $\widehat{b}_{\mathrm{W}}^{(j)} = \widehat{b}_{\mathrm{ML}}^{(j)} + O_p(n^{-1})$ $(j = 1, 2)$ due to $\widehat{\boldsymbol{\theta}}_{\mathrm{W}} = \widehat{\boldsymbol{\theta}}_{\mathrm{ML}} + O_p(n^{-1})$ (see Ogasawara, 2014, Equation (2.4)). The vector $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is expanded as

$$\begin{aligned}
& \widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0 = \sum_{j=1}^{3}(\boldsymbol{\Lambda}^{(j)})_{O(1)}(\mathbf{l}_0^{(j)})_{O_p(n^{-j//2})} + O_p(n^{-2}), \qquad (2.13) \\
& \boldsymbol{\Lambda}^{(1)}\mathbf{l}_0^{(1)} = -\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}, \\
& \boldsymbol{\Lambda}^{(2)}\mathbf{l}_0^{(2)} = \boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} - \frac{1}{2}\boldsymbol{\Lambda}^{-1}\mathrm{E}_g(\mathbf{J}_0^{(3)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}\right)^{<2>}, \\
& \mathbf{l}_0^{(1)} = \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}, \mathbf{l}_0^{(2)} = \left\{\mathrm{v}'(\mathbf{M}) \otimes \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0'}, \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0'}\right)^{<2>}\right\}', \boldsymbol{\Lambda}^{(1)} = -\boldsymbol{\Lambda}^{-1},
\end{aligned}$$

and $\boldsymbol{\Lambda}^{(2)}$ is implicitly defined by (2.13), where $\mathrm{v}(\cdot)$ is the vectorizing operator taking the non-duplicated elements of a symmetric matrix (Ogasawara, 2010, Equation (2.4)).

Then, using the expansion of $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0$ up to order $O_p(n^{-1/2})$ in (2.13), we have

$$\begin{aligned}\widehat{b}_{\mathrm{W}}^{(1)} &= 2\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) \\ &\quad -2\,\mathrm{tr}\Bigg[\underset{(\mathrm{A})}{\Bigg[}\left[\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\mathrm{E}_g(\mathbf{J}_0^{(3)})\left\{\boldsymbol{\Lambda}^{-1}\otimes\left(\boldsymbol{\Lambda}^{-1}\frac{\partial\bar{l}}{\partial\boldsymbol{\theta}_0}\right)\right\}\right]\boldsymbol{\Gamma} \\ &\quad -\boldsymbol{\Lambda}^{-1}\left\{\mathbf{M}_G - \sum_{j=1}^{q}\mathrm{E}_g(\mathbf{G}_{0(j)}^{(3)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial\bar{l}}{\partial\boldsymbol{\theta}_0}\right)_j\right\}\Bigg]_{(\mathrm{A})O_p(n^{-1/2})} + O_p(n^{-1}),\end{aligned} \tag{2.14}$$

where

$$\begin{aligned}&\mathbf{M}_G = n^{-1}\sum_{j=1}^{n}\frac{\partial l_j}{\partial\boldsymbol{\theta}_0}\frac{\partial l_j}{\partial\boldsymbol{\theta}_0'} - n\mathrm{E}_g\left(\frac{\partial\bar{l}}{\partial\boldsymbol{\theta}_0}\frac{\partial\bar{l}}{\partial\boldsymbol{\theta}_0'}\right) \equiv \mathbf{G}_0 - \mathrm{E}_g(\mathbf{G}_0) = O_p(n^{-1/2}), \\ &\mathbf{G}_{0(j)}^{(3)} = \frac{\partial\mathbf{G}_0}{\partial(\boldsymbol{\theta}_0)_j}\ (j=1,\ldots,q),\end{aligned} \tag{2.15}$$

$(\cdot)_j$ indicates the j-th element of a vector and $\underset{(\mathrm{A})}{[}\ \cdot\ \underset{(\mathrm{A})}{]}$ is for ease of finding correspondence. On the other hand, $\widehat{b}_{\mathrm{W}}^{(2)}$ for $n^{-1}\mathrm{TIC}_{\mathrm{W}}^{(2)}$ is given by omitting $\mathbf{M}$ and $\mathbf{M}_G$ in (2.14) i.e.,

$$\begin{aligned}\widehat{b}_{\mathrm{W}}^{(2)} &= 2\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + 2\,\mathrm{tr}\Bigg[\left[\boldsymbol{\Lambda}^{-1}\mathrm{E}_g(\mathbf{J}_0^{(3)})\left\{\boldsymbol{\Lambda}^{-1}\otimes\left(\boldsymbol{\Lambda}^{-1}\frac{\partial\bar{l}}{\partial\boldsymbol{\theta}_0}\right)\right\}\right]\boldsymbol{\Gamma} \\ &\quad -\boldsymbol{\Lambda}^{-1}\sum_{j=1}^{q}\mathrm{E}_g(\mathbf{G}_{0(j)}^{(3)})\left(\boldsymbol{\Lambda}^{-1}\frac{\partial\bar{l}}{\partial\boldsymbol{\theta}_0}\right)_j\Bigg] + O_p(n^{-1}).\end{aligned} \tag{2.16}$$

Define $\mathrm{MSE}(\cdot)_{\to O(n^{-2})}$ as the MSE of $(\cdot)$ up to order $O(n^{-2})$. Then,

$$\begin{aligned}&\mathrm{MSE}(n^{-1}\mathrm{TIC}_{\mathrm{W}(k)}^{(j)})_{\to O(n^{-2})} \\ &= \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})} \\ &\quad +n^{-2}\{(1-k)^2b^2 - 2kn\mathrm{acov}_g(\widehat{b}_{\mathrm{W}}^{(j)}, -2\widehat{\bar{l}}_{\mathrm{W}})\}(j=1,2),\end{aligned} \tag{2.17}$$

where $\mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})}$ is the higher-order asymptotic variance of $-2\widehat{\bar{l}}_{\mathrm{W}}$ up to order $O(n^{-2})$ with $\mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})} = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}$ due to (2.12), and from (2.14)

$$
\begin{aligned}
&n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{W}}, -2\widehat{\bar{l}}_{\mathrm{W}}) = n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{W}}, \bar{l}^{(1)}_{\mathrm{W}}) = n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, -2\bar{l}_0) \\
&= n\mathrm{acov}_g \Bigg[\underset{\text{(A)}}{-2\,\mathrm{tr}} \Bigg[\underset{\text{(B)}}{\Big[\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\mathrm{E}_g(\mathbf{J}^{(3)}_0) \Big\{ \boldsymbol{\Lambda}^{-1} \otimes \Big(\boldsymbol{\Lambda}^{-1} \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \Big) \Big\} \Big]} \boldsymbol{\Gamma} \\
&\quad - \boldsymbol{\Lambda}^{-1} \Bigg\{ \mathbf{M}_G - \sum_{j=1}^{q} \mathrm{E}_g(\mathbf{G}^{(3)}_{0(j)}) \Big(\boldsymbol{\Lambda}^{-1} \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \Big)_j \Bigg\} \Bigg]_{\text{(B)}} , -2\bar{l}_0 \Bigg]_{\text{(A)}} \\
&= 4 \Bigg[\underset{\text{(A)}}{\mathrm{vec}'(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}) n\mathrm{E}_g\{\mathrm{vec}(\mathbf{M}), \bar{l}_0\}} \\
&\quad - \mathrm{vec}'\{\mathrm{E}_g(\mathbf{J}^{(3)}_0)\} \Big[\mathrm{vec}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}) \otimes \Big\{ \boldsymbol{\Lambda}^{-1} n\mathrm{E}_g \Big(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \bar{l}_0 \Big) \Big\} \Big] \\
&\quad - \mathrm{tr}\{\boldsymbol{\Lambda}^{-1} n\mathrm{E}_g(\mathbf{M}_G \bar{l}_0)\} \\
&\quad + \mathrm{tr} \Bigg[\boldsymbol{\Lambda}^{-1} \sum_{j=1}^{q} \mathrm{E}_g(\mathbf{G}^{(3)}_{0(j)}) \Big\{ \boldsymbol{\Lambda}^{-1} n\mathrm{E}_g \Big(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \bar{l}_0 \Big) \Big\}_j \Bigg] \Bigg]_{\text{(A)}} .
\end{aligned} \tag{2.18}
$$

where $\mathrm{acov}_g(\cdot)$ is the asymptotic covariance of order $O(n^{-1})$ for two variables in parentheses under possible model misspecification.

The result for $\widehat{b}^{(2)}_{\mathrm{W}}$ corresponding to (2.18) is given by

$$
\begin{aligned}
&n\mathrm{acov}_g(\widehat{b}^{(2)}_{\mathrm{W}}, -2\widehat{\bar{l}}_{\mathrm{W}}) = n\mathrm{acov}_g(\widehat{b}^{(2)}_{\mathrm{W}}, \bar{l}^{(1)}_{\mathrm{W}}) = n\mathrm{acov}_g(\widehat{b}^{(2)}_{\mathrm{ML}}, -2\bar{l}_0) \\
&= 4 \Bigg[\underset{\text{(A)}}{- \mathrm{vec}'\{\mathrm{E}_g(\mathbf{J}^{(3)}_0)\} \Big[\mathrm{vec}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}) \otimes \Big\{ \boldsymbol{\Lambda}^{-1} n\mathrm{E}_g \Big(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \bar{l}_0 \Big) \Big\} \Big]} \\
&\quad + \mathrm{tr} \Bigg[\boldsymbol{\Lambda}^{-1} \sum_{j=1}^{q} \mathrm{E}_g(\mathbf{G}^{(3)}_{0(j)}) \Big\{ \boldsymbol{\Lambda}^{-1} n\mathrm{E}_g \Big(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \bar{l}_0 \Big) \Big\}_j \Bigg] \Bigg]_{\text{(A)}} .
\end{aligned} \tag{2.19}
$$

Then, from (2.17) to (2.19), we have

Theorem 1. *Under regularity conditions of* (2.12) *and* (2.3) *with possible model misspecification, the constant k minimizing the MSE up to order* $O_p(n^{-2})$ *for* $n^{-1}TIC^{(j)}_{W(k)}$ *denoted by* $k^{(Tj)}_{\min}$ $(j=1,2)$ *defined by* (2.10) *with* (2.9) *using the WSE* $\widehat{\boldsymbol{\theta}}_W$ *is*

$$k^{(Tj)}_{\min} = 1 + b^{-2} nacov_g(\widehat{b}^{(j)}_{ML}, -2\bar{l}_0), \tag{2.20}$$

which is equal to $k^{(Tj)}_{\min}$ *by* $\widehat{\boldsymbol{\theta}}_{ML}$. *The minimized value is*

$$\begin{aligned} &MSE(n^{-1}TIC^{(j)}_{W(k^{(Tj)}_{\min})})_{\to O(n^{-2})} = MSE(n^{-1}TIC^{(j)}_{ML(k^{(Tj)}_{\min})})_{\to O(n^{-2})} \\ &= avar_g(-2\widehat{\bar{l}}_{ML})_{\to O(n^{-2})} + n^{-2}b^2\{1-(k^{(Tj)}_{\min})^2\} \\ &= avar_g(-2\widehat{\bar{l}}_{ML})_{\to O(n^{-2})} \\ &\quad -n^{-2}[2nacov_g(\widehat{b}^{(j)}_{ML}, -2\bar{l}_0) + b^{-2}\{nacov_g(\widehat{b}^{(j)}_{ML}, -2\bar{l}_0)\}^2] \ (j=1,2). \end{aligned} \tag{2.21}$$

Theorem 1 shows that generally $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}(k^{(\mathrm{T}j)}_{\min})})_{\to O(n^{-2})}$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}})_{\to O(n^{-2})}$ $(=\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{ML}})_{\to O(n^{-2})} = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} - n^{-2}\{2n\mathrm{acov}_g(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0)\}_{O(1)})$ by $n^{-2}b^{-2}\{n\mathrm{acov}_g(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0)\}^2$ $(j=1,2)$. Note that for the derivation of $k^{(\mathrm{T}j)}_{\min}$ $(j=1,2)$, $\mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}$ was not used although $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}(k^{(\mathrm{T}j)}_{\min})})_{\to O(n^{-2})}$ depends on this value, which is somewhat complicated but can be obtained, if necessary, using Ogasawara (2015a). Note also that generally $k^{(\mathrm{T}j)}_{\min}$ $(j=1,2)$ depend on $\boldsymbol{\theta}_0$ and is not directly available in practice. This issue will be addressed in a discussion section.

Under correct model specification, the bias adjustment of $n^{-1}\mathrm{AIC}_{\mathrm{W}}$ as $n^{-1}\mathrm{AIC}_{\mathrm{W}(k)} = -2\widehat{\bar{l}}_{\mathrm{W}} + n^{-1}2kq$ does not decrease the original value $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})_{\to O(n^{-2})}$ since by this adjustment the variance is unchanged whereas the absolute value of the bias up to order $O(n^{-1})$ when $k \neq 1$ is larger than that of $n^{-1}\mathrm{AIC}_{\mathrm{W}}$. Under possible model misspecification, the optimal k minimizing $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}(k)})_{\to O(n^{-2})}$ is $k^{(\mathrm{A})}_{\min} = -b/(2q)$, which gives $n^{-1}\mathrm{AIC}_{\mathrm{W}(k^{(\mathrm{A})}_{\min})} = -2\widehat{\bar{l}}_{\mathrm{W}} - n^{-1}b$. This bias-adjusted $n^{-1}\mathrm{AIC}_{\mathrm{W}}$ is equal to $n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}$ $(j=1,2)$ when $\widehat{b}^{(j)}_{\mathrm{W}}$ is replaced by b though b is not directly available in practice. Note, however, that $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}(k)})_{\to O(n^{-2})}$ and $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}(k)})_{\to O(n^{-2})}$ are generally different, which is shown as follows.

Corollary 1. *Under the same conditions as in Theorem 1,*

$$\begin{aligned}
&MSE(n^{-1}TIC^{(j)}_{W(k)})_{\to O(n^{-2})} - MSE(n^{-1}AIC_{W(k)})_{\to O(n^{-2})} \\
&= avar_g(-2\widehat{\bar{l}}_{ML})_{\to O(n^{-2})} + n^{-2}\{(1-k)^2b^2 - 2knacov_g(\widehat{b}^{(j)}_{ML}, -2\bar{l}_0)\} \\
&\quad - avar_g(-2\widehat{\bar{l}}_{ML})_{\to O(n^{-2})} - n^{-2}(b+2kq)^2 \qquad (2.22) \\
&= -n^{-2}\{2k(b+2q)b + (-b^2+4q^2)k^2 + 2knacov_g(\widehat{b}^{(j)}_{ML}, -2\bar{l}_0)\} \ (j=1,2).
\end{aligned}$$

Note that the sign of (2.22) is indeterminable. When a model is true, consequently when $b=-2q$, if $n\mathrm{acov}_f(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0)$ is positive, where $n\mathrm{acov}_f(\cdot)$ indicates the asymptotic covariance under the true model, a positive k gives smaller $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}(k)})_{\to O(n^{-2})}$ than $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}(k)})_{\to O(n^{-2})}$. An example of the last result is $k=1$ which gives the usual $n^{-1}\mathrm{AIC}_{\mathrm{W}}$ and $n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}$ $(j=1,2)$. That is, in this case with $n\mathrm{acov}_f(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0) > 0$, $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}})_{\to O(n^{-2})}$ $(j=1,2)$ are smaller than $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})_{\to O(n^{-2})}$ by $n^{-2}\{2n\mathrm{acov}_f(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0)\}$. On the other hand, when $n\mathrm{acov}_f(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0) < 0$, an opposite result is obtained. When $n\mathrm{acov}_f(\widehat{b}^{(j)}_{\mathrm{ML}}, -2\bar{l}_0) = 0$, $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}})_{\to O(n^{-2})} = \mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})_{\to O(n^{-2})}$ $(j=1,2)$.

3 Optimal Information Criteria by Shrinkage

3.1. $n^{-1}OIC_{W(k)}$. In the expression $D+a$ of (1.1), consider the case of $D = n(-2\widehat{\bar{l}}_{\mathrm{W}})$ and $a = k2\widehat{\bar{l}}_{\mathrm{W}}$ with positive k. Divide $D+a$ by n and define this as a candidate optimal information criterion:

$$n^{-1}\mathrm{OIC}_{\mathrm{W}(k)} \equiv (1-n^{-1}k)(-2\widehat{\bar{l}}_{\mathrm{W}}), \qquad (3.1)$$

which is a shrinkage estimator of $-2\mathrm{E}_g(\widehat{\bar{l}}^{*}_{\mathrm{W}})$. Since

$$\begin{aligned}
&\mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k)})_{\to O(n^{-2})} \\
&= \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})} + n^{-2}\{(b+2k\bar{l}^{*}_0)^2 - 8kn\mathrm{var}_g(\bar{l}_0)\}, \qquad (3.2)
\end{aligned}$$

we have

Theorem 2. *Under the same conditions as in Theorem 1, the constant k minimizing $MSE(n^{-1}OIC_{W(k)})_{\to O(n^{-2})}$ is*

$$k_{\min} = \frac{1}{2(\bar{l}^{*}_0)^2}\{-\bar{l}^{*}_0 b + 2nvar_g(\bar{l}_0)\} \qquad (3.3)$$

and the minimized value is

$$\begin{aligned}
&MSE(n^{-1}OIC_{W(k_{\min})})_{\to O(n^{-2})}\\
&= avar_g(-2\widehat{\bar{l}}_W)_{\to O(n^{-2})} + n^{-2}\{b^2 - 4(\bar{l}_0^*)^2 k_{\min}^2\}\\
&= avar_g(-2\widehat{\bar{l}}_W)_{\to O(n^{-2})} - n^{-2}4\left[-\frac{b}{\bar{l}_0^*}nvar_g(\bar{l}_0) + \frac{1}{(\bar{l}_0^*)^2}\{nvar_g(\bar{l}_0)\}^2\right]. \quad (3.4)
\end{aligned}$$

Note that $\mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})})_{\to O(n^{-2})}$ is smaller than $\mathrm{MSE}(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})}$ by $4(\bar{l}_0^*)^2 k_{\min}^2 = (\bar{l}_0^*)^{-2}\{-\bar{l}_0^* b + 2n\mathrm{var}_g(\bar{l}_0)\}^2$. Comparison of the results of $n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})}$, $n^{-1}\mathrm{AIC}_{\mathrm{W}}$ and $n^{-1}\mathrm{TIC}_{\mathrm{W}}^{(j)}$ $(j = 1, 2)$ are given as follows.

$$\begin{aligned}
&\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})})_{\to O(n^{-2})}\\
&= n^{-2}(b + 2q)^2 - n^{-2}\{b^2 - 4(\bar{l}_0^*)^2 k_{\min}^2\}\\
&= n^{-2}\left[(b + 2q)^2 + 4\left\{-\frac{b}{\bar{l}_0^*}n\mathrm{var}_g(\bar{l}_0) + \frac{1}{(\bar{l}_0^*)^2}\{n\mathrm{var}_g(\bar{l}_0)\}^2\right\}\right]. \quad (3.5)
\end{aligned}$$

The sign of $b = \mathrm{tr}(\mathbf{\Lambda}^{-1}\mathbf{\Gamma})$ is usually negative while $\bar{l}_0^*$ can also be negative. So, when $b/\bar{l}_0^*$ is negative or the value of positive $b/\bar{l}_0^*$ is relatively small, (3.5) is positive especially under model misspecification with $(b + 2q)^2 > 0$ (note that $b = -2q$ under correct model specification). Similarly, we have

$$\begin{aligned}
&\mathrm{MSE}(n^{-1}\mathrm{TIC}_{\mathrm{W}}^{(j)})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})})_{\to O(n^{-2})}\\
&= n^{-2}4n\mathrm{acov}_g(\widehat{b}_{\mathrm{ML}}^{(j)}, \bar{l}_0) - n^{-2}\{b^2 - 4(\bar{l}_0^*)^2 k_{\min}^2\}\\
&= n^{-2}4\left[n\mathrm{acov}_g(\widehat{b}_{\mathrm{ML}}^{(j)}, \bar{l}_0) - \frac{b}{\bar{l}_0^*}n\mathrm{var}_g(\bar{l}_0) + \frac{1}{(\bar{l}_0^*)^2}\{n\mathrm{var}_g(\bar{l}_0)\}^2\right] \quad (j = 1, 2).
\end{aligned} \quad (3.6)$$

3.2. $n^{-1}OIC_{W(k)}^{(A)}$. In Section 3.1, the bias adjustment of $-2\widehat{\bar{l}}_{\mathrm{W}}$ was not explicitly considered as for $n^{-1}\mathrm{AIC}_{\mathrm{W}(k)}$ and $n^{-1}\mathrm{TIC}_{\mathrm{W}(k)}^{(j)}$ $(j = 1, 2)$ in Section 2, though $k_{\min}$ in Theorem 2 depends on b. In this section, a shrinkage estimator using $n^{-1}\mathrm{AIC}_{\mathrm{W}}$ is considered. Define

$$\begin{aligned}
n^{-1}\mathrm{OIC}_{\mathrm{W}(k)}^{(\mathrm{A})} &\equiv (1 - n^{-1}k)n^{-1}\mathrm{AIC}_{\mathrm{W}} = (1 - n^{-1}k)(-2\widehat{\bar{l}}_{\mathrm{W}} + n^{-1}2q)\\
&= n^{-1}\mathrm{AIC}_{\mathrm{W}} + n^{-1}2k\widehat{\bar{l}}_{\mathrm{W}} + O(n^{-2}). \quad (3.7)
\end{aligned}$$

Let $l_{0j} = l_j\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = f(\boldsymbol{x}_j^* \,|\, \boldsymbol{\theta}_0)$ $(j = 1, \ldots, n)$ (see (2.1)) and $c_{\mathrm{v}}(\cdot)$ as the coefficient of variation of a variable. Then, since

$$\begin{aligned}&\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k)})_{\to O(n^{-2})}\\ &\quad = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})} + n^{-2}\{(b+2q+2k\bar{l}_0^*)^2 - 8kn\mathrm{var}_g(\bar{l}_0)\},\end{aligned} \tag{3.8}$$

we have

Theorem 3. *Under the same conditions as in Theorem 1, the constant k minimizing $MSE(n^{-1}OIC^{(A)}_{W(k)})_{\to O(n^{-2})}$ is*

$$k_{\min}^{(OA)} = -\frac{b+2q}{2\bar{l}_0^*} + \frac{var_g(l_{0j})}{\{E_g(l_{0j})\}^2} = -\frac{b+2q}{2\bar{l}_0^*} + c_v^2(l_{0j}) \tag{3.9}$$

and the minimized value is

$$\begin{aligned}&MSE(n^{-1}OIC^{(A)}_{W(k_{\min}^{(OA)})})_{\to O(n^{-2})}\\ &\quad = avar_g(-2\widehat{\bar{l}}_W)_{\to O(n^{-2})} + n^{-2}\{(b+2q)^2 - 4(\bar{l}_0^*)^2(k_{\min}^{(OA)})^2\}\\ &\quad = avar_g(-2\widehat{\bar{l}}_W)_{\to O(n^{-2})} + n^{-2}4\{\bar{l}_0^*(b+2q)c_v^2(l_{0j}) - (\bar{l}_0^*)^2c_v^4(l_{0j})\}.\end{aligned} \tag{3.10}$$

Theorem 3 with (3.8) shows that $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k_{\min}^{(\mathrm{OA})})})_{\to O(n^{-2})}$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})_{\to O(n^{-2})}$ by
$n^{-2}4(\bar{l}_0^*)^2(k_{\min}^{(\mathrm{OA})})^2 = n^{-2}\{-(b+2q) + 2\bar{l}_0^* c_{\mathrm{v}}^2(l_{0j})\}^2$.

3.3. $n^{-1}OIC^{(Tj)}_{W(k)}$. In this subsection, shrinkage estimators using $n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}$ $(j = 1, 2)$ are considered. Define

$$\begin{aligned}n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k)} &= (1-n^{-1}k)n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}} = (1-n^{-1}k)(-2\widehat{\bar{l}}_{\mathrm{W}} - n^{-1}\widehat{b}^{(j)}_{\mathrm{W}})\\ &= n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}} + n^{-1}2k\widehat{\bar{l}}_{\mathrm{W}} + O_p(n^{-2})\ (j=1,2).\end{aligned} \tag{3.11}$$

Noting that

$$\begin{aligned}&\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k)})_{\to O(n^{-2})}\\ &\quad = \mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}})_{\to O(n^{-2})} + n^{-2}4\{(\bar{l}_0^*)^2k^2 - 2kn\mathrm{var}_g(\bar{l}_0)\}\ (j=1,2),\end{aligned} \tag{3.12}$$

we have

Theorem 4. *Under the same conditions as in Theorem 1, the constant k minimizing MSE$(n^{-1}OIC^{(Tj)}_{W(k)})_{\to O(n^{-2})}$ is*

$$k^{(OTj)}_{\min} = \frac{nvar_g(\bar{l}_0)}{(\bar{l}^*_0)^2} = \frac{var_g(l_{0i})}{\{E_g(l_{0i})\}^2} = c^2_v(l_{0i}) \tag{3.13}$$

and the minimized value is

$$\begin{aligned} &MSE(n^{-1}OIC^{(Tj)}_{W(k^{(OTj)}_{\min})})_{\to O(n^{-2})} \\ &\quad = MSE(n^{-1}TIC^{(j)}_W)_{\to O(n^{-2})} - n^{-2}4(\bar{l}^*_0)^2c^4_v(l_{0i})(j=1,2). \end{aligned} \tag{3.14}$$

Note that $k^{(\mathrm{OT1})}_{\min} = k^{(\mathrm{OT2})}_{\min}$ in (3.13). Theorem 4 shows that $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k^{(\mathrm{OT}j)}_{\min})})_{\to O(n^{-2})}$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}})_{\to O(n^{-2})}$ by $n^{-2}4(\bar{l}^*_0)^2c^4_{\mathrm{v}}(l_{0i})$ $(j=1,2)$.

Define $\widehat{\theta}_{\mathrm{u}}(=O_p(1))$ as an asymptotically unbiased estimator up to order $O(n^{-1})$ with the corresponding population value being θ_0. Then, it is known that the multiplicative constant c^* for $\widehat{\theta}_{\mathrm{u}}$ minimizing the MSE of $c^*\widehat{\theta}_{\mathrm{u}}$ up to order $O(n^{-2})$ is given by $c^*_{\min} = 1 - n^{-1}\theta_0^{-2}n\mathrm{var}_g(\widehat{\theta}_{\mathrm{u}})$ (Ogasawara, 2015b, the paragraph including Equation (2.6)). Since $n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}$ $(j=1,2)$ are asymptotically unbiased up to order $O(n^{-1})$ and $k^{(\mathrm{OT}j)}_{\min} = c^2_{\mathrm{v}}(l_{0i})$, $n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k^{(\mathrm{OT}j)}_{\min})} = (1-n^{-1}k^{(\mathrm{OT}j)}_{\min})n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}$ $(j=1,2)$ in (3.11) is seen as a special case of $c^*_{\min}\widehat{\theta}_{\mathrm{u}}$. Note also that

$$\begin{aligned} n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k^{(\mathrm{OT}j)}_{\min})} &= (1-n^{-1}k^{(\mathrm{OT}j)}_{\min})n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}} = \frac{n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}}{1+n^{-1}k^{(\mathrm{OT}j)}_{\min}} + O_p(n^{-2}) \\ &= \frac{n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{W}}}{1+n^{-1}c^2_{\mathrm{v}}(l_{0i})} + O_p(n^{-2}) \ (j=1,2). \end{aligned} \tag{3.15}$$

Then, we find that the factor $1/\{1+n^{-1}c^2_{\mathrm{v}}(l_{0i})\}$ in (3.15) is the multiplicative constant for $\bar{l}_0$ exactly minimizing the MSE of $\bar{l}_0$, where $\bar{l}_0$ is seen as a usual unbiased sample mean (Gruber, 1998, Section 1.6; Ogasawara, 2015b, Section 3.1).

It is of interest to see that $n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})}$ given in Section 3.1 is asymptotically equivalent to $n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k^{(\mathrm{OT}j)}_{\min})}$ $(j=1,2)$ as follows.

Corollary 2. *Under the same conditions as in Theorem 1,*

$$n^{-1}OIC_{W(k_{\min})} = n^{-1}OIC^{(Tj)}_{W(k^{(OTj)}_{\min})} + O_p(n^{-3/2}) \ (j=1,2). \tag{3.16}$$

PROOF.

$$\begin{aligned} n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})} &= (1-n^{-1}k_{\min})(-2\widehat{\bar{l}}_{\mathrm{W}}) \\ &= \left[1-\frac{n^{-1}}{2(\bar{l}_0^*)^2}\{-\bar{l}_0^* b+2n\mathrm{var}_g(\bar{l}_0)\}\right](-2\widehat{\bar{l}}_{\mathrm{W}}) \\ &= -2\widehat{\bar{l}}_{\mathrm{W}}-n^{-1}b-n^{-1}\frac{n\mathrm{var}_g(\bar{l}_0)}{(\bar{l}_0^*)^2}(-2\widehat{\bar{l}}_{\mathrm{W}})+O_p(n^{-3/2}) \\ &= \{1-n^{-1}c_{\mathrm{v}}^2(l_{0i})\}n^{-1}\mathrm{TIC}_{\mathrm{W}}^{(j)}+O_p(n^{-3/2}) \\ &= n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k_{\min}^{(\mathrm{OT}j)})}+O_p(n^{-3/2})\ (j=1,2). \end{aligned} \tag{3.17}$$

□

For comparison of the results of $n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k_{\min}^{(\mathrm{OA})})}$ and $n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k_{\min}^{(\mathrm{OT}j)})}$, we have from (3.10) and (3.14) with (2.17)

$$\begin{aligned} &\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k_{\min}^{(\mathrm{OA})})})_{\to O(n^{-2})}-\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{W}(k_{\min}^{(\mathrm{OT}j)})})_{\to O(n^{-2})} \\ &\quad = n^{-2}\{4\bar{l}_0^*(b+2q)c_{\mathrm{v}}^2(l_{0i})-4n\mathrm{acov}_g(\widehat{b}_{\mathrm{ML}}^{(j)},\bar{l}_0)\}(j=1,2). \end{aligned} \tag{3.18}$$

The sign of (3.18) is indeterminable. Under correct model specification (3.18) becomes $-n^{-2}4n\mathrm{acov}_g(\widehat{b}_{\mathrm{ML}}^{(j)},\bar{l}_0)$, which can be positive, negative or zero.

4 Examples

In this section, three simple examples are shown for illustration of computational aspects of the optimal information criteria. The first two examples are given under possible model misspecification using $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$. The third example is shown under correct model specification when $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ is used.

Example 1. The exponential distribution is used when the gamma distribution holds, where the shape parameter α is not necessarily equal to 1. That is,

$$\begin{aligned} f(x^* &= x\,|\,\lambda_0) = \lambda_0\exp(-\lambda_0 x)(x>0), \\ g(x^* &= x\,|\,\lambda_1,\alpha) = x^{\alpha-1}\lambda_1\exp(-\lambda_1 x)/\Gamma(\alpha)\ (x>0), \\ \widehat{\boldsymbol{\theta}}_{\mathrm{ML}} &= 1/\overline{x}^*, \theta_0=\lambda_0, \boldsymbol{\zeta}_0=(\lambda_1,\alpha)', \end{aligned} \tag{4.1}$$

$\overline{x}^*$is the usual sample mean,

$$\begin{aligned} \mathrm{E}_g(\overline{x}^*) &= \alpha/\lambda_1 = 1/\lambda_0, \lambda_0=\lambda_1/\alpha, l_{0j}=-\lambda_0 x_j^*+\log\lambda_0, \\ \bar{l}_0^* &= \mathrm{E}_g(l_{0j}) = -\lambda_0(\alpha/\lambda_1)+\log\lambda_0 = \log(\lambda_1/\alpha)-1, \end{aligned}$$

$$n\mathrm{var}_g(\bar{l}_0) = \mathrm{var}_g(l_{0j}) = \lambda_0^2\mathrm{var}_g(x_j^*) = \lambda_0^2(\alpha/\lambda_1^2) = (\lambda_1/\alpha)^2(\alpha/\lambda_1^2) = 1/\alpha,$$

$$\begin{aligned}
c_{\mathrm{v}}^2(l_{0j}) &= \frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}^2},\ \frac{\partial\bar{l}}{\partial\theta_0} = \frac{1}{\lambda_0} - \overline{x}^*, \\
\boldsymbol{\Lambda} &= \lambda = \frac{\partial^2\bar{l}}{\partial\theta_0^2} = -\frac{1}{\lambda_0^2} = -\frac{\alpha^2}{\lambda_1^2}, \\
\boldsymbol{\Gamma} &= \gamma = n\mathrm{E}_g\left\{\left(\frac{\partial\bar{l}}{\partial\theta_0}\right)^2\right\} = \mathrm{E}_g\left\{\left(\frac{1}{\lambda_0} - x_j^*\right)^2\right\} = \frac{\alpha}{\lambda_1^2}, \\
b = 2\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) &= 2\lambda^{-1}\gamma = -\frac{2}{\alpha}, \\
n^{-1}\mathrm{AIC}_{\mathrm{ML}} &= -2\widehat{\bar{l}}_{\mathrm{ML}} + n^{-1}2,\ n^{-1}\mathrm{TIC}_{\mathrm{ML}}^{(2)} = -2\widehat{\bar{l}}_{\mathrm{ML}} + n^{-1}2\widehat{\alpha}^{-1}.
\end{aligned}$$

Note that in this example $n^{-1}\mathrm{TIC}_{\mathrm{ML}}^{(1)}$ is not used since $n^{-1}\mathrm{TIC}_{\mathrm{ML}}^{(2)}$ is available.

(1) $n^{-1}\mathrm{AIC}_{\mathrm{ML}}$

The MSEs up to order $O(n^{-2})$ of $n^{-1}\mathrm{AIC}_{\mathrm{ML}}$ and $n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k_{\min}^{(\mathrm{OA})})}$ are compared. From (3.9) and (3.10),

$$\begin{aligned}
k_{\min}^{(\mathrm{OA})} &= -\frac{b+2q}{2\bar{l}_0^*} + c_{\mathrm{v}}^2(l_{0j}) = -\frac{-(2/\alpha)+2}{2\{\log(\lambda_1/\alpha)-1\}} + c_{\mathrm{v}}^2(l_{0j}) \\
&= \frac{\alpha^{-1}-1}{\log(\lambda_1/\alpha)-1} + \frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}^2} = \frac{(1-\alpha)\log(\lambda_1/\alpha)+\alpha}{\alpha\{\log(\lambda_1/\alpha)-1\}^2}, \\
&\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{ML}})_{\to O(n^{-2})} = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} + n^{-2}(b+2q)^2 \\
&= \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} + n^{-2}4(1-\alpha^{-1})^2, \\
&\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{ML}(k_{\min}^{(\mathrm{OA})})})_{\to O(n^{-2})} \\
&= \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} + n^{-2}4\{\bar{l}_0^*(b+2q)c_{\mathrm{v}}^2(l_{0j}) - (\bar{l}_0^*)^2c_{\mathrm{v}}^4(l_{0j})\} \\
&= \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} \\
&\quad + n^{-2}4\left[2(1-\alpha^{-1})\frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}} - \frac{1}{\alpha^2\{\log(\lambda_1/\alpha)-1\}^2}\right] \\
&= \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} + n^{-2}4\frac{2(\alpha-1)\log(\lambda_1/\alpha) - 2\alpha + 1}{\alpha^2\{\log(\lambda_1/\alpha)-1\}^2},
\end{aligned} \tag{4.2}$$

where $\mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} = n^{-1}4\alpha^{-1} + n^{-2}2\alpha^{-2}$ (see Ogasawara, 2015a). That is, $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{ML}(k^{(\mathrm{OA})}_{\mathrm{min}})})_{\to O(n^{-2})}$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{ML}})_{\to O(n^{-2})}$ by

$$n^{-2}\{-(b+2q) + 2\bar{l}_0^* c_{\mathrm{v}}^2(l_{0j})\}^2 = n^{-2}\left[-2(1-\alpha^{-1}) + \frac{2}{\alpha\{\log(\lambda_1/\alpha)-1\}}\right]^2$$
$$= n^{-2}4\left[\frac{(\alpha-1)\log(\lambda_1/\alpha)-\alpha}{\alpha\{\log(\lambda_1/\alpha)-1\}}\right]^2.$$

(2) $n^{-1}\mathrm{OIC}_{\mathrm{ML}(k)}$ and $n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}}$

The constant k minimizing $\mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k)})_{\to O(n^{-2})}$ with $n^{-1}\mathrm{OIC}_{\mathrm{W}(k)} = (1-n^{-1}k)(-2\widehat{\bar{l}}_{\mathrm{W}})$ (see (3.1)) is given from (3.3) as

$$\begin{aligned} k_{\mathrm{min}} &= \frac{1}{2(\bar{l}_0^*)^2}\{-\bar{l}_0^* b + 2n\mathrm{var}_g(\bar{l}_0)\} = -\frac{1}{2\bar{l}_0^*}(2\lambda^{-1}\gamma) + c_{\mathrm{v}}^2(l_{0i}) \\ &= \frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}} + \frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}^2} \\ &= \frac{\log(\lambda_1/\alpha)}{\alpha\{\log(\lambda_1/\alpha)-1\}^2}. \end{aligned} \tag{4.3}$$

The constant k minimizing $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{ML}(k)})_{\to O(n^{-2})}$ with $n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{ML}(k)} = (1-n^{-2}k)n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}}$ (see (3.11) and (3.13)) is

$$k^{(\mathrm{OT2})}_{\mathrm{min}} = c_{\mathrm{v}}^2(l_{0i}) = \frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}^2}. \tag{4.4}$$

From (4.3) and (4.4),

$$k_{\mathrm{min}} = -\frac{b}{2\bar{l}_0^*} + k^{(\mathrm{OT2})}_{\mathrm{min}} = \frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}} + k^{(\mathrm{OT2})}_{\mathrm{min}}. \tag{4.5}$$

From (3.14), $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{ML}(k^{(\mathrm{OT2})}_{\mathrm{min}})})_{\to O(n^{-2})}$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}})_{\to O(n^{-2})}$ by

$$n^{-2}4(\bar{l}_0^*)^2 c_{\mathrm{v}}^4(l_{0i}) = n^{-2}\frac{4\{\mathrm{var}_g(l_{0i})\}^2}{(\bar{l}_0^*)^2} = n^{-2}\frac{4}{\alpha^2\{\log(\lambda_1/\alpha)-1\}^2}. \tag{4.6}$$

From (3.18) and $4n\mathrm{acov}_g(\widehat{b}^{(2)}_{\mathrm{ML}}, \bar{l}_0) = 0$ in this example (see Ogasawara, 2015a),

$$\begin{aligned}
&\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{ML}(k^{(\mathrm{OA})}_{\mathrm{min}})})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{ML}(k^{(\mathrm{OT2})}_{\mathrm{min}})})_{\to O(n^{-2})} \\
&\quad = n^{-2}4\bar{l}^*_0(b+2q)c^2_{\mathrm{v}}(l_{0i}) = n^{-2}4(2-2\alpha^{-1})\frac{1}{\alpha\{\log(\lambda_1/\alpha)-1\}} \\
&\quad = n^{-2}\frac{8(\alpha-1)}{\alpha^2\{\log(\lambda_1/\alpha)-1\}}, \qquad (4.7)
\end{aligned}$$

whose sign is indeterminable.

From (3.5).

$$\begin{aligned}
&\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{ML}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{ML}(k_{\mathrm{min}})})_{\to O(n^{-2})} \\
&\quad = n^{-2}\left[(b+2q)^2 + 4\left\{-\frac{b}{\bar{l}^*_0}n\mathrm{var}_g(\bar{l}_0) + \frac{1}{(\bar{l}^*_0)^2}\{n\mathrm{var}_g(\bar{l}_0)\}^2\right\}\right] \\
&\quad = n^{-2}\left[4(1-\alpha^{-1})^2 + 4\left\{-\frac{1}{\log(\lambda_1/\alpha)-1}\left(-\frac{2}{\alpha}\right)\frac{1}{\alpha} + \frac{1}{\alpha^2\{\log(\lambda_1/\alpha)-1\}^2}\right\}\right] \\
&\quad = n^{-2}4\left[(1-\alpha^{-1})^2 + \frac{1}{\alpha^2}\left\{\frac{2}{\log(\lambda_1/\alpha)-1} + \frac{1}{\{\log(\lambda_1/\alpha)-1\}^2}\right\}\right]. \qquad (4.8)
\end{aligned}$$

When λ_1/α is sufficiently large, (4.8) is positive.

From (3.6) and $n\mathrm{acov}_g(\widehat{b}^{(2)}_{\mathrm{ML}}, \bar{l}_0) = 0$,

$$\begin{aligned}
&\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{ML}(k_{\mathrm{min}})})_{\to O(n^{-2})} \\
&\quad = n^{-2}4\left\{-\frac{b}{\bar{l}^*_0}n\mathrm{var}_g(\bar{l}_0) + \frac{1}{(\bar{l}^*_0)^2}\{n\mathrm{var}_g(\bar{l}_0)\}^2\right\} \\
&\quad = n^{-2}\frac{4}{\alpha^2}\left\{\frac{2}{\log(\lambda_1/\alpha)-1} + \frac{1}{\{\log(\lambda_1/\alpha)-1\}^2}\right\}. \qquad (4.9)
\end{aligned}$$

When $\log(\lambda_1/\alpha) - 1 > 0$, (4.9) is positive.

Example 2. In this example, the normal distribution with a known variance and the MLE of the mean is used although the true distribution is possibly non-normal with the known variance. That is,

$$\begin{aligned}
&f(x^* = x \,|\, \mu_0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu_0)^2}{2\sigma^2}\right\}, \\
&\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \overline{x}^*, \bar{l}^*_0 = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2},\ \mathrm{var}_g(z^2) = \kappa_4 + 2, \qquad (4.10)
\end{aligned}$$

where $z = (x^* - \mu_0)/\sigma$, $\kappa_j = \kappa_j(z)$ is the j-th cumulant of z under possible non-normality,

$$
\begin{aligned}
& c_{\mathrm{v}}^2(l_{0i}) = \frac{(\kappa_4+2)/4}{\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\right\}^2} = \frac{\kappa_4+2}{\{\log(2\pi\sigma^2)+1\}^2}, \\
& \boldsymbol{\Lambda} = \lambda = -1/\sigma^2,\ \gamma = 1/\sigma^2, b = 2\lambda^{-1}\gamma = -2,\ b+2q = 0, \\
& n^{-1}\mathrm{AIC}_{\mathrm{ML}} = -2\widehat{\bar{l}}_{\mathrm{ML}} + n^{-1}2, \qquad (4.11) \\
& n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}} = -2\widehat{\bar{l}}_{\mathrm{ML}} + n^{-1}2(-\lambda^{-1}\gamma) = -2\widehat{\bar{l}}_{\mathrm{ML}} + n^{-1}2 = n^{-1}\mathrm{AIC}_{\mathrm{ML}}.
\end{aligned}
$$

As before $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}}$ is not used. Note that in this example, $n^{-1}\mathrm{AIC}_{\mathrm{ML}} = n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}}$ even under non-normality.

From (3.9) and (3.10),

$$
\begin{aligned}
& k^{(\mathrm{OA})}_{\min} = k^{(\mathrm{OT2})}_{\min} = -\frac{b+2q}{2\bar{l}^*_0} + c_{\mathrm{v}}^2(l_{0i}) = c_{\mathrm{v}}^2(l_{0i}), \\
& \mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{ML}}) = \mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}}) \\
& \quad = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} + n^{-2}(b+2q)^2 = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}, \\
& \mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{ML}(k^{(\mathrm{OA})}_{\min})})_{\to O(n^{-2})} = \mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{ML}(k^{(\mathrm{OT2})}_{\min})})_{\to O(n^{-2})} \\
& \quad = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} - n^{-2}4(\bar{l}^*_0)^2 c_{\mathrm{v}}^4(l_{0i}) \\
& \quad = \mathrm{avar}_g(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})} - n^{-2}\frac{(\kappa_4+2)^2}{\{\log(2\pi\sigma^2)+1\}^2}. \qquad (4.12)
\end{aligned}
$$

That is, $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{ML}(k^{(\mathrm{OA})}_{\min})})_{\to O(n^{-2})}$ $(=\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{ML}(k^{(\mathrm{OT2})}_{\min})})_{\to O(n^{-2})})$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{ML}})$ $(=\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}}))$ by $n^{-2}\frac{(\kappa_4+2)^2}{\{\log(2\pi\sigma^2)+1\}^2}$.

From (3.3) to (3.6),

$$
\begin{aligned}
& k_{\min} = \frac{1}{2(\bar{l}^*_0)^2}\{-\bar{l}^*_0 b + 2n\mathrm{var}_g(\bar{l}_0)\} = \frac{1}{\bar{l}^*_0} + c_{\mathrm{v}}^2(l_{0i}) \\
& \quad = \frac{1}{\bar{l}^*_0} + k^{(\mathrm{OA})}_{\min} = \frac{1}{\bar{l}^*_0} + k^{(\mathrm{OT2})}_{\min}, \\
& \mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{ML}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{ML}(k_{\min})})_{\to O(n^{-2})} \\
& \quad = \mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{ML}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{ML}(k_{\min})})_{\to O(n^{-2})} \\
& \quad = n^{-2}\left[(b+2q)^2 + 4\left\{-\frac{b}{\bar{l}^*_0}n\mathrm{var}_g(\bar{l}_0) + \frac{1}{(\bar{l}^*_0)^2}\{n\mathrm{var}_g(\bar{l}_0)\}^2\right\}\right]
\end{aligned}
$$

$$= n^{-2}4\left[\left\{\frac{1}{2}\log(2\pi\sigma^2)+\frac{1}{2}\right\}^{-1}(-2)\frac{\kappa_4+2}{4} + \left\{\frac{1}{2}\log(2\pi\sigma^2)+\frac{1}{2}\right\}^{-2}\frac{1}{16}(\kappa_4+2)^2\right]$$
$$= n^{-2}[-4\{\log(2\pi\sigma^2)+1\}^{-1}(\kappa_4+2)+\{\log(2\pi\sigma^2)+1\}^{-2}(\kappa_4+2)^2]. \quad (4.13)$$

Since $\kappa_4+2 = 4n\mathrm{var}_g(\bar{l}_0) > 0$, when $\bar{l}_0^* = -\{\log(2\pi\sigma^2)+1\}/2 > 0$, (4.13) is positive.

Example 3. In this example, it is assumed that the Bernoulli distribution holds, where the WSE of the logit with $a^*/2$ pseudocounts for each of two categories is used:

$$\begin{aligned}
\Pr(x^* &= x\,|\,\theta) = \pi^x(1-\pi)^{1-x}\ (x=0,1), \pi = \frac{1}{1+\exp(-\theta)},\\
l_{0i} &= x_i^*\log\frac{\pi_0}{1-\pi_0}+\log(1-\pi_0) = x_i^*\theta_0+\log(1-\pi_0), \pi_0 = \frac{1}{1+\exp(-\theta_0)},\\
\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} &= \log\frac{\overline{x}^*}{1-\overline{x}^*},\ \widehat{\boldsymbol{\theta}}_{\mathrm{W}} = \log\frac{\overline{x}^*+n^{-1}0.5a^*}{1-\overline{x}^*+n^{-1}0.5a^*},\\
\boldsymbol{\Lambda} = \lambda &= -\pi_0(1-\pi_0) \equiv -\bar{i}_0, \gamma = \bar{i}_0, b = 2\lambda^{-1}\gamma = -2, b+2q = 0,\\
\bar{l}_0^* &= \pi_0\theta_0+\log(1-\pi_0),\ \mathrm{var}_f(l_{0i}) = \theta_0^2\bar{i}_0,\\
c_{\mathrm{v}}^2(l_{0i}) &= \frac{\theta_0^2\bar{i}_0}{\{\pi_0\theta_0+\log(1-\pi_0)\}^2},\ n^{-1}\mathrm{AIC}_{\mathrm{W}} = n^{-1}\mathrm{TIC}_{\mathrm{W}}^{(2)} = -2\widehat{\bar{l}}_{\mathrm{W}}+n^{-1}2. \quad (4.14)
\end{aligned}$$

From (3.9) and (3.10),

$$\begin{aligned}
k_{\min}^{(\mathrm{OA})} &= k_{\min}^{(\mathrm{OT2})} = -\frac{b+2q}{2\bar{l}_0^*}+c_{\mathrm{v}}^2(l_{0i}) = c_{\mathrm{v}}^2(l_{0i}),\\
\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}}) &= \mathrm{MSE}(n^{-1}\mathrm{TIC}_{\mathrm{W}}^{(2)})\\
&= \mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}+n^{-2}(b+2q)^2 = \mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})},\\
\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k_{\min}^{(\mathrm{OA})})})_{\to O(n^{-2})} &= \mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{W}(k_{\min}^{(\mathrm{OT2})})})_{\to O(n^{-2})}\\
&= \mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}+n^{-2}\{(b+2q)^2-4(\bar{l}_0^*)^2(k_{\min}^{(\mathrm{OA})})^2\}\\
&= \mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}-n^{-2}4(\bar{l}_0^*)^2c_{\mathrm{v}}^4(l_{0i})\\
&= \mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}-n^{-2}\frac{4\theta_0^4\bar{i}_0^2}{\{\pi_0\theta_0+\log(1-\pi_0)\}^2}, \quad (4.15)
\end{aligned}$$

which shows that $\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{W}(k^{(\mathrm{OA})}_{\min})})_{\to O(n^{-2})}$ $(=\mathrm{MSE}(n^{-1}\mathrm{OIC}^{(\mathrm{T2})}_{\mathrm{W}(k^{(\mathrm{OT2})}_{\min})})_{\to O(n^{-2})})$ is smaller than $\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})(=\mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{W}}))$ by $n^{-2}\dfrac{4\theta_0^4\bar{i}_0^2}{\{\pi_0\theta_0+\log(1-\pi_0)\}^2}$.

From (3.3) to (3.6),

$$\begin{aligned}
k_{\min} &= \frac{1}{2(\bar{l}_0^*)^2}\{-\bar{l}_0^* b + 2n\mathrm{var}_f(\bar{l}_0)\} = \frac{1}{\bar{l}_0^*} + c_{\mathrm{v}}^2(l_{0i}) \\
&= \frac{1}{\bar{l}_0^*} + k^{(\mathrm{OA})}_{\min} = \frac{1}{\bar{l}_0^*} + k^{(\mathrm{OT2})}_{\min} \\
&= \frac{1}{\pi_0\theta_0+\log(1-\pi_0)} + \frac{\theta_0^2\bar{i}_0}{\{\pi_0\theta_0+\log(1-\pi_0)\}^2}, \\
&\mathrm{MSE}(n^{-1}\mathrm{AIC}_{\mathrm{W}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})})_{\to O(n^{-2})} \\
&= \mathrm{MSE}(n^{-1}\mathrm{TIC}^{(2)}_{\mathrm{W}})_{\to O(n^{-2})} - \mathrm{MSE}(n^{-1}\mathrm{OIC}_{\mathrm{W}(k_{\min})})_{\to O(n^{-2})} \\
&= n^{-2}\left[(b+2q)^2 + 4\left\{-\frac{b}{\bar{l}_0^*}n\mathrm{var}_f(\bar{l}_0) + \frac{1}{(\bar{l}_0^*)^2}\{n\mathrm{var}_f(\bar{l}_0)\}^2\right\}\right] \\
&= n^{-2}4\left[-\frac{1}{\pi_0\theta_0+\log(1-\pi_0)}(-2)\theta_0^2\bar{i}_0 + \frac{\theta_0^4\bar{i}_0^2}{\{\pi_0\theta_0+\log(1-\pi_0)\}^2}\right] \\
&= n^{-2}\left[\frac{8\theta_0^2\bar{i}_0}{\pi_0\theta_0+\log(1-\pi_0)} + \frac{4\theta_0^4\bar{i}_0^2}{\{\pi_0\theta_0+\log(1-\pi_0)\}^2}\right]. \qquad (4.16)
\end{aligned}$$

Note that the above asymptotic results by $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ are the same as those by $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ with $\mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{W}})_{\to O(n^{-2})} = \mathrm{avar}_f(-2\widehat{\bar{l}}_{\mathrm{ML}})_{\to O(n^{-2})}$ (see (2.12)).

5 A Simulation for Model Selection in Normal Linear Regression

Since information criteria are used primarily for selecting appropriate models, a simulation for model selection is carried out in the case of usual normal-theory linear regression with non-stochastic regressors or covariates under normality and non-normality. Let $\boldsymbol{y}$ be the $n\times 1$ random vector with

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{e} \text{ and } \boldsymbol{e} \sim N(\mathbf{0}, \sigma_0^2\mathbf{I}_{(n)}), \qquad (5.1)$$

where n is the sample size, $\mathbf{X}$ is a $n\times p$ design matrix, $\boldsymbol{\beta}_0$ is a $p\times 1$ vector of population regression coefficients, $\boldsymbol{e}$ is a $n\times 1$ random vector of errors, σ_0^2 is the common variance for the errors and $\mathbf{I}_{(n)}$ is the $n\times n$ identity

matrix. Using the likelihood of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ corresponding to the population counterpart $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \sigma_0^2)'$ i.e.,

$$L(\boldsymbol{\theta} \,|\, \boldsymbol{y}, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\{-(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2)\}, \tag{5.2}$$

we have

$$\begin{aligned} \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} &= \left\{ \frac{n^{-1}}{\sigma_0^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_0)'\mathbf{X}, -\frac{1}{2\sigma_0^2} + \frac{n^{-1}}{2\sigma_0^4}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_0)'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_0) \right\}' \\ &= \left(\frac{n^{-1}}{\sigma_0^2}\boldsymbol{e}'\mathbf{X}, -\frac{1}{2\sigma_0^2} + \frac{n^{-1}}{2\sigma_0^4}\boldsymbol{e}'\boldsymbol{e} \right)', \end{aligned} \tag{5.3}$$

and

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\mathrm{ML}} &= (\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}', \widehat{\sigma}_{\mathrm{ML}}^2)' = \{\boldsymbol{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, n^{-1}\boldsymbol{y}'(\mathbf{I}_{(n)} - \mathbf{P})\boldsymbol{y}\}' \\ &= \{\boldsymbol{\beta}_0' + \boldsymbol{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, n^{-1}\boldsymbol{e}'(\mathbf{I}_{(n)} - \mathbf{P})\boldsymbol{e}\}', \end{aligned} \tag{5.4}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projection matrix.

In this section, n^{-1}AIC, n^{-1}CAIC, $n^{-1}\mathrm{TIC}^{(1)}$ and $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$ with various k's including $k^{(\mathrm{T1})}_{\min}$ are considered using $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ of (5.4), where CAIC is a corrected AIC (Sugiura, 1978) to have an exactly unbiased estimator of $\mathrm{E}_f(-2\bar{l}^*_{\mathrm{ML}})$ under correct model specification or under over specification with true and redundant regressors. Note that in the case of simulation in this section $n^{-1}\mathrm{OIC}^{(\mathrm{A})}_{\mathrm{ML}(k)}$ and $n^{-1}\mathrm{OIC}^{(\mathrm{T}j)}_{\mathrm{ML}(k)}$ $(j = 1, 2)$ give the same model selection as those by $n^{-1}\mathrm{AIC}_{\mathrm{ML}}$ and $n^{-1}\mathrm{TIC}^{(j)}_{\mathrm{ML}}$ $(j = 1, 2)$, respectively due to proportionality.

Since $-2\bar{l}_{\mathrm{ML}} = \log(2\pi\widehat{\sigma}^2_{\mathrm{ML}}) + 1$, we obtain

$$n^{-1}\mathrm{AIC} = \log(2\pi\widehat{\sigma}^2_{\mathrm{ML}}) + 1 + n^{-1}2(p+1). \tag{5.5}$$

On the other hand n^{-1}CAIC is defined by

$$\begin{aligned} n^{-1}\mathrm{CAIC} &= \log(2\pi\widehat{\sigma}^2_{\mathrm{ML}}) + \frac{n+p}{n-p-2} = \log(2\pi\widehat{\sigma}^2_{\mathrm{ML}}) + 1 + \frac{2(p+1)}{n-p-2} \\ &> \log(2\pi\widehat{\sigma}^2_{\mathrm{ML}}) + 1 + n^{-1}2(p+1) = n^{-1}\mathrm{AIC}. \end{aligned} \tag{5.6}$$

For the TIC,

$$n^{-1}\mathrm{TIC}^{(1)} = n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}} = -2\bar{l}_{\mathrm{ML}} - n^{-1}2\,\mathrm{tr}(\widehat{\boldsymbol{\Lambda}}^{-1}_{\mathrm{ML}}\widehat{\boldsymbol{\Gamma}}_{\mathrm{ML}}) \tag{5.7}$$

is used under normality and non-normality. It is assumed that even under non-normality, $\mathrm{E}_g(\boldsymbol{e}) = \mathbf{0}$ and $\mathrm{E}_g(\boldsymbol{e}\boldsymbol{e}') = \sigma_0^2\mathbf{I}_{(n)}$ hold. In (5.7),

$$\begin{aligned}\widehat{\boldsymbol{\Lambda}}_{\mathrm{ML}} &= \frac{\partial^2 \bar{l}}{\partial\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}\partial\widehat{\boldsymbol{\theta}}'_{\mathrm{ML}}} \\ &= \begin{bmatrix} -n^{-1}\mathbf{X}'\mathbf{X}/\widehat{\sigma}^2_{\mathrm{ML}} & -n^{-1}\mathbf{X}'(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ML}})/\widehat{\sigma}^4_{\mathrm{ML}} \\ -n^{-1}(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ML}})'\mathbf{X}/\widehat{\sigma}^4_{\mathrm{ML}} & \{1/(2\widehat{\sigma}^4_{\mathrm{ML}})\} \\ & -n^{-1}(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ML}})'(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ML}})/\widehat{\sigma}^6_{\mathrm{ML}} \end{bmatrix} \\ &= \begin{bmatrix} -n^{-1}\mathbf{X}'\mathbf{X}/\widehat{\sigma}^2_{\mathrm{ML}} & \mathbf{0} \\ \mathbf{0}' & -1/(2\widehat{\sigma}^4_{\mathrm{ML}}) \end{bmatrix} \end{aligned} \tag{5.8}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}_{\mathrm{ML}} &= n^{-1}\sum_{j=1}^{n}\frac{\partial l_j}{\partial\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}}\frac{\partial l_j}{\partial\widehat{\boldsymbol{\theta}}'_{\mathrm{ML}}} \\ &= n^{-1}\sum_{j=1}^{n}\begin{bmatrix} (y_j-\widehat{y}_j)\boldsymbol{x}_j/\widehat{\sigma}^2_{\mathrm{ML}} \\ -\dfrac{1}{2\widehat{\sigma}^2_{\mathrm{ML}}}+\dfrac{(y_j-\widehat{y}_j)^2}{2\widehat{\sigma}^4_{\mathrm{ML}}} \end{bmatrix}\left[\frac{y_j-\widehat{y}_j}{\widehat{\sigma}^2_{\mathrm{ML}}}\boldsymbol{x}'_j, -\frac{1}{2\widehat{\sigma}^2_{\mathrm{ML}}}+\frac{(y_j-\widehat{y}_j)^2}{2\widehat{\sigma}^4_{\mathrm{ML}}}\right] \\ &= n^{-1}\sum_{j=1}^{n}\begin{bmatrix} [\{(\mathbf{I}_{(n)}-\mathbf{P})\boldsymbol{e}\}_j]^2\boldsymbol{x}_j\boldsymbol{x}'_j/\widehat{\sigma}^4_{\mathrm{ML}} & [\{(\mathbf{I}_{(n)}-\mathbf{P})\boldsymbol{e}\}_j]^3\boldsymbol{x}_j/(2\widehat{\sigma}^6_{\mathrm{ML}}) \\ \dfrac{[\{(\mathbf{I}_{(n)}-\mathbf{P})\boldsymbol{e}\}_j]^3}{2\widehat{\sigma}^6_{\mathrm{ML}}}\boldsymbol{x}'_j & -\dfrac{1}{4\widehat{\sigma}^4_{\mathrm{ML}}}+\dfrac{[\{(\mathbf{I}_{(n)}-\mathbf{P})\boldsymbol{e}\}_j]^4}{4\widehat{\sigma}^8_{\mathrm{ML}}} \end{bmatrix}, \end{aligned} \tag{5.9}$$

where $\widehat{y}_j = (\widehat{\boldsymbol{y}})_j = (\mathbf{P}\boldsymbol{y})_j$ and $\boldsymbol{x}'_j$ is the j-th row of $\mathbf{X}$ ($j = 1, \ldots, n$). From (5.7) to (5.9),

$$\begin{aligned} n^{-1}\mathrm{TIC}^{(1)} &= -2\widehat{\bar{l}}_{\mathrm{ML}} - n^{-1}\widehat{b}^{(1)}_{\mathrm{ML}} \\ &= -2\widehat{\bar{l}}_{\mathrm{ML}} - n^{-1}\underset{(\mathrm{A})}{\Big[} -2\,\mathrm{tr}\underset{(\mathrm{B})}{\Big[}(\mathbf{X}'\mathbf{X})^{-1}\sum_{j=1}^{n}[\{(\mathbf{I}_{(n)}-\mathbf{P})\boldsymbol{e}\}_j]^2\boldsymbol{x}_j\boldsymbol{x}'_j/\widehat{\sigma}^2_{\mathrm{ML}}\underset{(\mathrm{B})}{\Big]} \\ &\quad +1 - n^{-1}\sum_{j=1}^{n}[\{(\mathbf{I}_{(n)}-\mathbf{P})\boldsymbol{e}\}_j]^4/\widehat{\sigma}^4_{\mathrm{ML}}\underset{(\mathrm{A})}{\Big]} \end{aligned} \tag{5.10}$$

follows.

For $k^{(\mathrm{T1})}_{\min}$, we derive $n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, -2\widehat{\bar{l}}_{\mathrm{ML}})$ under possible non-normality. Note that

$$n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, -2\widehat{\bar{l}}_{\mathrm{ML}}) = n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, -2\bar{l}_0) = n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, n^{-1}\boldsymbol{e}'\boldsymbol{e}/\sigma_0^2) \tag{5.11}$$

On the other hand,

$$\begin{aligned} n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, -2\widehat{\bar{l}}_{\mathrm{ML}}) &= n\mathrm{acov}_g\{\widehat{b}^{(1)}_{\mathrm{ML}}, \log(2\pi\widehat{\sigma}^2_{\mathrm{ML}}) + 1\} \\ &= n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, \log\widehat{\sigma}^2_{\mathrm{ML}}) = n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, \widehat{\sigma}^2_{\mathrm{ML}}/\sigma_0^2). \end{aligned} \tag{5.12}$$

Note also that $\{(\mathbf{I}_{(n)} - \mathbf{P})\boldsymbol{e}\}_j = e_j - (\mathbf{P}\boldsymbol{e})_j$, where $e_j = O_p(1)$ and $(\mathbf{P}\boldsymbol{e})_j = O_p(n^{-1/2})$ under some regularity conditions since $(\mathbf{P}\boldsymbol{e})_j$ is a weighted mean of $e_1, \ldots, e_n$. Define $\bar{z} = e_1/\sigma_0$ and $\kappa_j = \kappa_j(\bar{z})$ is the j-th cumulant of $\bar{z}$. Then,

$$\begin{aligned} &n\mathrm{acov}_g(\widehat{b}^{(1)}_{\mathrm{ML}}, -2\widehat{\bar{l}}_{\mathrm{ML}}) \\ &= n\mathrm{acov}_g\Big[\underset{(\mathrm{A})}{-2\,\mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\sum_{j=1}^n e_j^2\boldsymbol{x}_j\boldsymbol{x}_j'/\widehat{\sigma}^2_{\mathrm{ML}}]} - n^{-1}\sum_{j=1}^n e_j^4/\widehat{\sigma}^4_{\mathrm{ML}}, \underset{(\mathrm{A})}{n^{-1}\boldsymbol{e}'\boldsymbol{e}/\sigma_0^2} \Big] \\ &= -n2\,\mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}n^{-1}\sum_{j=1}^n \left\{ \mathrm{cov}_g\left(\frac{e_j^2}{\sigma_0^2}, \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma_0^2}\right) - \frac{\mathrm{E}_g(e_j^2)}{\sigma_0^6} n\mathrm{avar}_g(\widehat{\sigma}^2_{\mathrm{ML}}) \right\} \boldsymbol{x}_j\boldsymbol{x}_j' \right] \\ &- nn^{-2}\sum_{j=1}^n \left\{ \mathrm{cov}_g\left(\frac{e_j^4}{\sigma_0^4}, \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma_0^2}\right) - \frac{2\mathrm{E}_g(e_j^4)}{\sigma_0^8} n\mathrm{avar}_g(\widehat{\sigma}^2_{\mathrm{ML}}) \right\} \\ &= -2\,\mathrm{tr}[\mathbf{I}_{(p)}\{(\kappa_4+2)-(\kappa_4+2)\}] - \{\mathrm{E}_g(\bar{z}^6) - (\kappa_4+3) - 2(\kappa_4+3)(\kappa_4+2)\} \\ &= -(\kappa_6 + 15\kappa_4 + 10\kappa_3^2 + 15) + (2\kappa_4+5)(\kappa_4+3) \\ &= -(\kappa_6 + 4\kappa_4 + 10\kappa_3^2) + 2\kappa_4^2. \end{aligned} \tag{5.13}$$

From (2.20) and (5.13), we find that $k^{(\mathrm{T1})}_{\mathrm{min}} = 1$ and consequently, $n^{-1}\,\mathrm{TIC}^{(1)}_{\mathrm{ML}(k^{(\mathrm{T1})}_{\mathrm{min}})} = n^{-1}\mathrm{TIC}^{(1)}$ under normality.

For $b = 2\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})$, we have

$$\boldsymbol{\Lambda} = \begin{bmatrix} -n^{-1}\mathbf{X}'\mathbf{X}/\sigma_0^2 & \mathbf{0} \\ \mathbf{0}' & -1/(2\sigma_0^4) \end{bmatrix} = -\mathbf{I}_0 \tag{5.14}$$

even under non-normality and

$$\begin{aligned} \boldsymbol{\Gamma} &= \mathrm{E}_g\left(n^{-1}\sum_{j=1}^n \frac{\partial l_j}{\partial\boldsymbol{\theta}_0}\frac{\partial l_j}{\partial\boldsymbol{\theta}_0'} \right) \\ &= n^{-1}\mathrm{E}_g\left[\sum_{j=1}^n \begin{bmatrix} e_j\boldsymbol{x}_j/\sigma_0^2 \\ -\dfrac{1}{2\sigma_0^2} + \dfrac{e_j^2}{2\sigma_0^4} \end{bmatrix} \left[\frac{e_j}{\sigma_0^2}\boldsymbol{x}_j', -\frac{1}{2\sigma_0^2} + \frac{e_j^2}{2\sigma_0^4} \right] \right] \\ &= \begin{bmatrix} n^{-1}\mathbf{X}'\mathbf{X}/\sigma_0^2 & n^{-1}\sum_{j=1}^n \kappa_3\boldsymbol{x}_j/(2\sigma_0^3) \\ n^{-1}\sum_{j=1}^n \kappa_3\boldsymbol{x}_j'/(2\sigma_0^3) & (\kappa_4+2)/(4\sigma_0^4) \end{bmatrix}, \end{aligned} \tag{5.15}$$

which gives

$$b = 2\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) = -2\,\mathrm{tr}\begin{pmatrix} \mathbf{I}_{(p)} & \mathbf{0} \\ \mathbf{0}' & (\kappa_4+2)/2 \end{pmatrix} = -(2p+2+\kappa_4). \quad (5.16)$$

In the simulation, two sets of population values are used: [1] $p = 2$, $\boldsymbol{\beta}_0 = (2,1)'$, $\sigma_0^2 = 1$, $n = 20$ and 50, [2] $p = 5$, $\boldsymbol{\beta}_0 = (2,1,1,1,1)'$, $\sigma_0^2 = 1$, $n = 50$. The covariates are generated by the independent standard normal distributions except the first unit covariate corresponding to an intercept which is always included. In the first case, candidate models with $p = 1$ (only an intercept), 2, 3, 4 are used while in the second case, those with $p = 1, 2, \ldots, 8$ are used. For non-normal distributions of e_j, the chi-square distributions with 3 and 1 degrees of freedom followed by standardization to have mean zero and variance σ_0^2 are used. Their skewnesses are $2\sqrt{2}/\sqrt{3}$ and $2\sqrt{2}$, the excess kurtoses are 4 and 12, and the k_6's are 160/3 and 480 for the two non-normal distributions, respectively. For $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$, $k = k^{(\mathrm{T1})}_{\min}, 2k^{(\mathrm{T1})}_{\min}, 4k^{(\mathrm{T1})}_{\min}, 5k^{(\mathrm{T1})}_{\min}$ are used after some preliminary investigation.

Tables 1 and 2 show the proportions of correct model selection by the 7 information criteria when the 4 and 8 candidate models are available, respectively. In Table 1, with $n = 20$ n^{-1}CAIC shows the best results. The added proportions of correct model selection over n^{-1}AIC are substantial even under non-normality. On the other hand, usual or non-adjusted $n^{-1}\mathrm{TIC}^{(1)}$ gives poor results. When $k^{(\mathrm{T1})}_{\min}$ is used for $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$, the proportions of correct model selection have increased under non-normality. However, the increased proportions are still smaller than those by those by n^{-1}CAIC. When $k = 2k^{(\mathrm{T1})}_{\min}$ or $4k^{(\mathrm{T1})}_{\min}$ is used, the correct proportions have surprisingly increased especially when $n = 50$. In Table 2, under normality the correct proportion by $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(4k^{(\mathrm{T1})}_{\min})}$ is as large as 99.4 %.

It is known that AIC tends to select overspecified models or models with redundant regressor(s) (Hurvich and Tsai, 1989; Fujikoshi and Satoh, 1997). Tables 1 and 2 repeat this tendency. We find that in the tables, n^{-1}CAIC also has this tendency. On the other hand, when k is increased from $k = k^{(\mathrm{T1})}_{\min}$ to $k = 5k^{(\mathrm{T1})}_{\min}$, it is clearly shown that the opposite tendency of selecting underspecified models gradually appears. The information criteria with the largest correct proportions in the tables seem to select under and overspecified models in a balanced way. Although $k^{(\mathrm{T1})}_{\min}$ is not directly available in practice since $k^{(\mathrm{T1})}_{\min}$ depends on $\boldsymbol{\theta}_0$, the results in the tables are

encouraging in that when information about $k_{\min}^{(\mathrm{T1})}$ and optimal c for $ck_{\min}^{(\mathrm{T1})}$ accumulates using e.g., simulations, it is expected that the proportions of correct model selection can be increased as in Tables 1 and 2.

6 A Real Example for Model Selection

A real example is shown for model selection in linear regression using the house price data analyzed by Gilmour (1996, Table 1) in the context of model selection by Mallows' C_P and its modification, where the dependent variable is the house price and nine regressors are available with $n = 24$: x_1 taxes (\$/1000), x_2 number of baths, x_3 lot size (ft^2/1000), x_4 living space (ft^2/1000), x_5 number of garage stalls, x_6 number of rooms, x_7 number of bed rooms, x_8 age (years) and x_9 number of fireplaces. An intercept is always included in candidate models. That is, the number of possible models considered is $2^9 = 512$, where the model with an intercept only is also considered as a candidate.

Six information criteria n^{-1}AIC, n^{-1}CAIC, $n^{-1}\mathrm{TIC}^{(1)}$ and $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$ with $k = k_{\min}^{(\mathrm{T1})}, 2k_{\min}^{(\mathrm{T1})}, 4k_{\min}^{(\mathrm{T1})}$ are used for model selection. For $k_{\min}^{(\mathrm{T1})}$, two conditions for the distributions of e_i are illustrated: the normal and chi-square (3 degrees of freedom) distributions, where the latter distribution is employed as a substantially non-normal distribution. For σ_0^2, the unbiased estimator $n\widehat{\sigma}^2_{\mathrm{ML}}/(n-p)$ in the full model with $p = 10$ is used with the assumption that the model includes the true set of regressors as in Mallows' C_P. The value $-2\mathrm{E}_f(\widehat{\bar{l}}^*_{\mathrm{ML}})$ under normality is given by

$$-2\mathrm{E}_f(\widehat{\bar{l}}^*_{\mathrm{ML}}) = \log\frac{4\pi\sigma_0^2}{n} + \psi\left(\frac{n-p}{2}\right) + \frac{n+p}{n-p-2} \tag{6.1}$$

(see Davies et al., 2006, Equation (3.5)), where $\psi(\cdot)$ is the digamma function, and the last term $(n+p)/(n-p-2)$ is equal to the correction term plus 1 in n^{-1}CAIC (see (5.6)).

Figures 1 and 2 show the values of the six information criteria under normality and non-normality, respectively, where the horizontal axis is for the number of regressors including an intercept. A fine horizontal line indicates the value 5.168 of (6.1), when $p = 2$, which is also used under non-normality for comparison. Note that Gilmour (1996, p.54) concluded that the model of $p = 2$ with regressor x_1 and an intercept is the most appropriate model. In Fig. 1, $n^{-1}\mathrm{TIC}^{(1)} = n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k_{\min}^{(\mathrm{T1})})}$ since $k_{\min}^{(\mathrm{T1})} = 1$ under normality.

The models selected by the minimum values of the six information criteria in Fig. 1 are $(1, x_1, x_2, x_5, x_7), (1, x_1, x_2), (1, x_1, x_2, x_4, x_5, x_7)$,

Table 1: 1,000 times the proportions of model selection when the true number of regressors is 2 with 1,000 replications

	$n = 20$				$n = 50$			
	$p = 1$	2^*	3	4	$p = 1$	2^*	3	4
Normal	$\left(k_{\min}^{(\mathrm{T1})} = 1\right.$	1	1	1)				
n^{-1}AIC	11	685	177	127	0	779	134	87
n^{-1}CAIC	25	<u>824</u>	114	37	0	833	116	51
$n^{-1}\mathrm{TIC}^{(1)}$	7	618	210	165	0	736	152	112
$n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$								
$k = k_{\min}^{(\mathrm{T1})}$	7	618	210	165	0	736	152	112
$k = 2k_{\min}^{(\mathrm{T1})}$	39	806	108	47	0	928	53	19
$k = 4k_{\min}^{(\mathrm{T1})}$	169	808	19	4	5	981	14	0
$k = 5k_{\min}^{(\mathrm{T1})}$	274	716	10	0	11	<u>982</u>	7	0
Chi-square ($df = 3$)	$\left(k_{\min}^{(\mathrm{T1})} = 0\right.$	.36	.56	.67)				
n^{-1}AIC	11	720	151	118	0	761	148	91
n^{-1}CAIC	28	<u>838</u>	95	39	0	818	127	55
$n^{-1}\mathrm{TIC}^{(1)}$	28	616	184	172	5	703	176	116
$n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$								
$k = k_{\min}^{(\mathrm{T1})}$	27	727	137	109	1	820	115	64
$k = 2k_{\min}^{(\mathrm{T1})}$	94	834	49	23	9	<u>959</u>	24	8
$k = 4k_{\min}^{(\mathrm{T1})}$	312	681	7	0	64	936	0	0
$k = 5k_{\min}^{(\mathrm{T1})}$	435	561	4	0	106	894	0	0
Chi-square ($df = 1$)	$\left(k_{\min}^{(\mathrm{T1})} = -.25\right.$	.01	.20	.34)				
n^{-1}AIC	27	715	173	85	0	743	149	108
n^{-1}CAIC	39	824	109	28	0	801	126	73
$n^{-1}\mathrm{TIC}^{(1)}$	66	572	201	161	12	656	186	146
$n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$								
$k = k_{\min}^{(\mathrm{T1})}$	18	672	194	116	0	778	143	79
$k = 2k_{\min}^{(\mathrm{T1})}$	57	812	98	33	5	927	52	16
$k = 4k_{\min}^{(\mathrm{T1})}$	133	<u>841</u>	22	4	24	<u>971</u>	5	0
$k = 5k_{\min}^{(\mathrm{T1})}$	178	806	15	1	40	958	2	0

Note. 2^* = the true value. An underscore indicates the largest proportion of correct model selection under each condition

Table 2: 1,000 times the proportions of model selection when the true number of regressors is 5 with 1,000 replications

$n = 50$	$p = 1$	2	3	4	5*	6	7	8
Normal	$\left(k_{\min}^{(\mathrm{T1})} = 1\right.$	1	1	1	1	1	1	1)
n^{-1}AIC	0	0	0	0	736	120	78	66
n^{-1}CAIC	0	0	0	0	849	77	47	27
$n^{-1}\mathrm{TIC}^{(1)}$	0	0	0	0	707	128	85	80
$n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$								
$k = k_{\min}^{(\mathrm{T1})}$	0	0	0	0	707	128	85	80
$k = 2k_{\min}^{(\mathrm{T1})}$	0	0	0	0	925	41	25	9
$k = 4k_{\min}^{(\mathrm{T1})}$	0	0	0	0	<u>994</u>	5	1	0
$k = 5k_{\min}^{(\mathrm{T1})}$	0	0	1	4	993	2	0	0
Chi-square ($df = 3$)	$\left(k_{\min}^{(\mathrm{T1})} = 0\right.$	.36	.56	.67	.75	.80	.84	.87)
n^{-1}AIC	0	0	0	0	706	139	74	81
n^{-1}CAIC	0	0	0	0	815	106	46	33
$n^{-1}\mathrm{TIC}^{(1)}$	0	0	0	0	642	149	104	105
$n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$								
$k = k_{\min}^{(\mathrm{T1})}$	0	0	0	0	729	134	70	67
$k = 2k_{\min}^{(\mathrm{T1})}$	1	1	2	5	922	47	16	6
$k = 4k_{\min}^{(\mathrm{T1})}$	6	3	6	27	<u>954</u>	4	0	0
$k = 5k_{\min}^{(\mathrm{T1})}$	13	4	23	48	911	1	0	0
Chi-square ($df = 1$)	$\left(k_{\min}^{(\mathrm{T1})} = -.25\right.$	.01	.20	.34	.44	.53	.59	.64)
n^{-1}AIC	0	0	0	0	703	147	85	65
n^{-1}CAIC	0	0	0	0	819	110	43	28
$n^{-1}\mathrm{TIC}^{(1)}$	0	0	3	6	602	175	111	103
$n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k)}$								
$k = k_{\min}^{(\mathrm{T1})}$	0	0	1	0	739	134	66	60
$k = 2k_{\min}^{(\mathrm{T1})}$	1	1	6	8	928	42	6	8
$k = 4k_{\min}^{(\mathrm{T1})}$	4	4	16	39	<u>934</u>	3	0	0
$k = 5k_{\min}^{(\mathrm{T1})}$	5	5	42	48	899	1	0	0

Note. 5^* = the true value. An underscore indicates the largest proportion of correct model selection under each condition

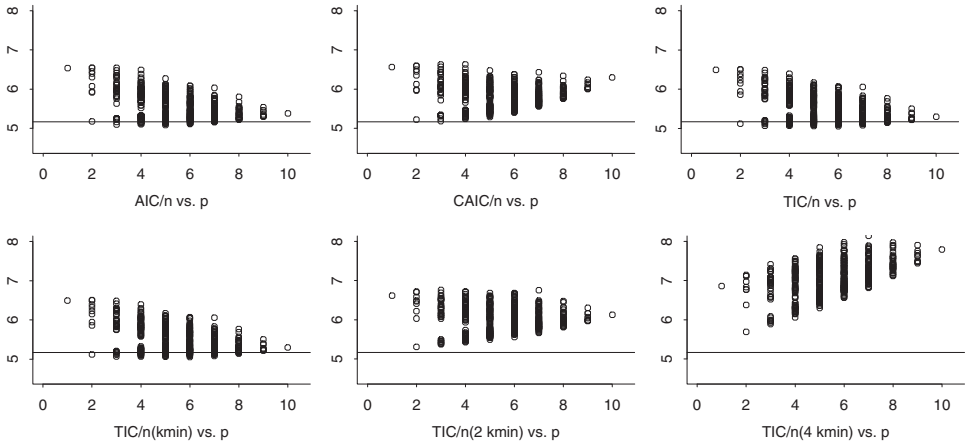

Figure 1: Information criteria for the house price data under normality

$(1, x_1, x_2, x_4, x_5, x_7), (1, x_1)$ and $(1, x_1)$ for n^{-1}AIC, n^{-1}CAIC, $n^{-1}\mathrm{TIC}^{(1)}$, $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k_{\min}^{(\mathrm{T1})})}$, $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(2k_{\min}^{(\mathrm{T1})})}$ and $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(4k_{\min}^{(\mathrm{T1})})}$, respectively, where e.g., $(1, x_1)$ indicates the model with regressors constant 1 for an intercept and x_1. The corresponding results in Fig. 2 are $(1, x_1, x_2, x_5, x_7), (1, x_1, x_2)$, $(1, x_1, x_2, x_4, x_5, x_7), (1, x_1, x_2), (1, x_1)$ and $(1, x_1)$, where the results are unchanged except that $(1, x_1, x_2)$ by $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(k_{\min}^{(\mathrm{T1})})}$ was $(1, x_1, x_2, x_4, x_5, x_7)$ in Fig. 1.

The appropriate model with $(1, x_1)$ is selected only by $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(2k_{\min}^{(\mathrm{T1})})}$ and $n^{-1}\mathrm{TIC}^{(1)}_{\mathrm{ML}(4k_{\min}^{(\mathrm{T1})})}$ in Figs. 1 and 2, which is consistent with the results in Section 5. Selection of the appropriate model was made possible by penalizing relatively complicated models. It is seen from the figures that n^{-1}CAIC also has this tendency in comparison with n^{-1}AIC though the correction is not sufficient.

7 Discussion

As mentioned earlier, optimal k values minimizing the associated MSEs generally depend on $\boldsymbol{\theta}_0$, which is unknown in practice. The problem of k depending on $\boldsymbol{\theta}_0$ is similar to that choosing optimal ridge parameters depending on the population parameters (Hoerl & Kennard, 1970). Since $\boldsymbol{\theta}_0$ can be estimated, the optimal k can also be estimated using the same data as those for $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$. However, if we use the estimated k, the optimality does not generally hold.

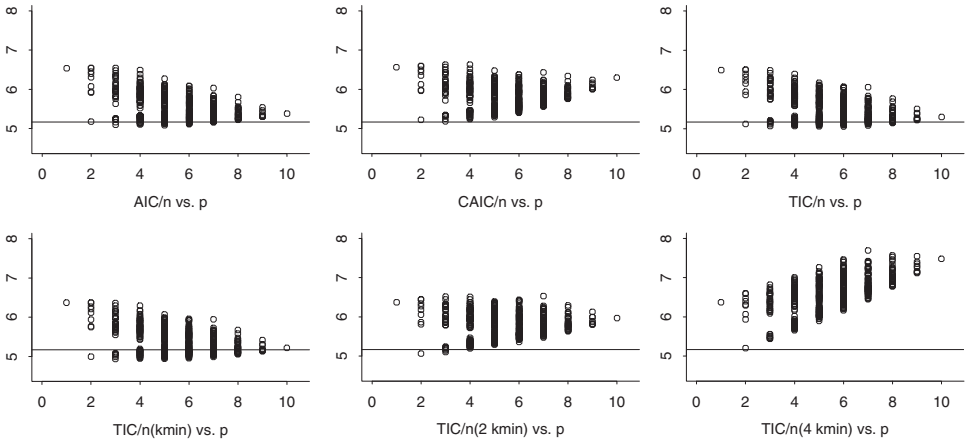

Figure 2: Information criteria for the house price data under chi-square (3 df)

Let an information criterion

$$\mathrm{IC}_{\mathrm{W}(k)} \equiv -2\widehat{\bar{l}}_{\mathrm{W}} + n^{-1}k\widehat{a} \tag{7.1}$$

and a similar one

$$\mathrm{IC}_{\mathrm{W}(\widehat{k})} \equiv -2\widehat{\bar{l}}_{\mathrm{W}} + n^{-1}\widehat{k}\widehat{a}, \tag{7.2}$$

where $\widehat{k}$ is an estimated k using the same data. Then,

$$\begin{aligned}&\mathrm{MSE}(\mathrm{IC}_{\mathrm{W}(\widehat{k})})_{\to O(n^{-2})}\\ &\quad = \mathrm{MSE}(\mathrm{IC}_{\mathrm{W}(k)})_{\to O(n^{-2})} + n^{-2}\{2an\mathrm{acov}_g(-2\widehat{\bar{l}}_{\mathrm{W}}, \widehat{k})\}_{O(1)}.\end{aligned} \tag{7.3}$$

That is, except under the condition of $n\mathrm{acov}_g(-2\widehat{\bar{l}}_{\mathrm{W}}, \widehat{k}) = 0$, the MSE up to order $O(n^{-2})$ changes although the MSE may decrease when $an\mathrm{acov}_g(-2\widehat{\bar{l}}_{\mathrm{W}},\widehat{k})$ is negative.

While it is difficult to overcome the above problem completely, several methods can be used. First, in a limited number of cases, optimal k's do not depend on (the whole of) $\boldsymbol{\theta}_0$. Note that in Example 2, the optimal k's are not functions of $\theta_0(=\mu_0)$. In Sections 5 and 6 for linear regression, $k_{\min}^{(\mathrm{T1})}$ does not depend on $\boldsymbol{\beta}_0$, but depends on the distributions of errors. Secondly, we often have some prior information about the values of θ_0, which can give e.g., a lower bound of $k_{\min}^{(\mathrm{OA})}$. Since the MSEs of the information criteria up

to order $O(n^{-2})$ shown earlier with k are quadratic functions of k, we find that using the lower/upper bound decreases the original MSE. For instance, in Example 1 when $\alpha = 1$ or under correct model specification, (4.4) gives $k_{\min}^{(\mathrm{OT2})} = 1/(\log\lambda_1 - 1)^2 = 1/(\log\lambda_0 - 1)^2$. When it is known that an upper bound of λ_0 is $\lambda_0^*(> e)$, $1/(\log\lambda_0^* - 1)^2$ can be used as a lower bound for $k_{\min}^{(\mathrm{OT2})}$. Similarly, in Example 2, $k_{\min}^{(\mathrm{OA})} = c_{\mathrm{v}}^2(l_{0i}) = (\kappa_4 + 2)/\{\log(2\pi\sigma^2) + 1\}^2$ is a linear function of the excess kurtosis of z under arbitrary distributions. When a lower bound of $\kappa_4 + 2(= \mathrm{var}_g(z^2))$ is available, this gives a lower bound of $k_{\min}^{(\mathrm{OA})}$.

Thirdly, as in cross validation, if another independent data set of size n^* is available for estimation of k, the estimated k, say $\widehat{k}^*$, can be used. Note that $\widehat{k}^*$ does not change the optimality of using k since the expectation of $n^{-1}\widehat{k}^*\widehat{a}$ is the same as that of $n^{-1}k\widehat{a}$ up to order $O(n^{-1})$ under usual conditions with the term $n\mathrm{acov}_g(-2\widehat{\bar{l}}_{\mathrm{W}}, \widehat{k}^*) = 0$ in (7.3). Note that the variance of $n^{-1}\widehat{k}^*\widehat{a}$ is of order $o(n^{-2})$, which can be asymptotically neglected when we consider the MSE up to order $O(n^{-2})$ for the associated information criterion of order $O_p(n^{-1})$.

Acknowledgments. This work was partially supported by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology (JSPS KAKENHI, Grant No.26330031).

References

AKAIKE, H. (1973). *Information theory and an extension of the maximum likelihood principle.* Académiai Kiado, Budapest. Proceedings of the 2nd international symposium on information theory, B. N. PETROV and F. CSÁKI (eds.), p. 267–281.

CAVANAUGH, J.E. and SHUMWAY, R.H. (1997). A bootstrap variant of AIC for state-space model selection. *Statistica Sinica* **7**, 473–496.

DAVIES, S.L., NEATH, A.A. and CAVANAUGH, J.E. (2006). Estimation optimality of corrected AIC and modified Cp in linear regression. *Int. Stat. Rev.* **74**, 161–168.

FUJIKOSHI, Y. and SATOH, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707–716.

GILMOUR, S.G. (1996). The interpretation of Mallows's *Cp*-statistic. *J. R. Stat. Soc. D* **45**, 45–56.

GRUBER, M.H.J. (1998). *Improving efficiency by shrinkage: The James-Stein and ridge regression estimators.* Marcel Dekker, New York.

HOERL, A.E. and KENNARD, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

HURVICH, C.M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

ISHIGURO, M., SAKAMOTO, Y. and KITAGAWA, G. (1997). Bootstrapping log-likelihood and EIC, An extension of AIC. *Ann. Inst. Stat. Math.* **49**, 411–434.

KISHINO, H. and HASEGAWA, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29**, 170–179.

KONISHI, S. and KITAGAWA, G. (1996). Generalized information criteria in model selection. *Biometrika* **83**, 875–890.

KONISHI, S. and KITAGAWA, G. (2003). Asymptotic theory for information criteria in model selection – functional approach. *Journal of Statistical Planning and Inference* **114**, 45–61.

KONISHI, S. and KITAGAWA, G. (2008). *Information criteria and statistical modeling.* Springer, New-York.

KULLBACK, S. and LEIBLER, R.A. (1951). On information and sufficiency. *Ann. Stat.* **22**, 79–86.

MALLOWS, C.L. (1973). Some comments on C_P. *Technometrics* **15**, 661–675.

OGASAWARA, H. (2010). Asymptotic expansions for the pivots using log-likelihood derivatives with an application in item response theory. *J. Multivar. Anal.* **101**, 2149–2167.

OGASAWARA, H. (2013). Asymptotic cumulants of the estimator of the canonical parameter in the exponential family. *Journal of Statistical Planning and Inference* **143**, 2142–2150.

OGASAWARA, H. (2014). Optimization of the Gaussian and Jeffreys power priors with emphasis on the canonical parameters in the exponential family. *Behaviormetrika* **41**, 195–223.

OGASAWARA, H. (2015a) Asymptotic cumulants of some information criteria (2nd version). Discussion Paper, Center for Business Creation, Otaru University of Commerce, No. 174. http://barrel.ih.otaru-uc.ac.jp/.

OGASAWARA, H. (2015b). Bias adjustment minimizing the asymptotic mean square error. *Communications in Statistics – Theory and Methods* **44**, 3503–3522.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.

SHIBATA, R. (1989). *Statistical aspects of model selection.* Springer, Berlin. From data to model, J. C. WILLEMS (ed.), p. 215–240.

SHIBATA, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Stat. Sin.* **7**, 375–394.

SHIMODAIRA, H. and HASEGAWA, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116.

SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods* **7**, 13–26.

TAKEUCHI, K. (1976). Distributions of information statistics and criteria of the goodness of models. *Math. Sci.* **153**, 12–18. (in Japanese).

Haruhiko Ogasawara
Department of Information and
Management Science
Otaru University of Commerce
3-5-21, Otaru 047-8501, Japan
E-mail: hogasa@res.otaru-uc.ac.jp

Paper received: 12 May 2015.