A Class of Model Selection Criteria Based on Cross-Validation Method

(Last Modified: January 13, 2006)

Hirokazu YANAGIHARA[1], Ke-Hai YUAN[2], Hironori FUJISAWA[3]
AND
Kentaro HAYASHI[4]

[1]*Department of Mathematics, Graduate School of Science, Hiroshima University*
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

[2]*Department of Psychology, University of Notre Dame*
Notre Dame, Indiana 46556, USA

[3]*Department of Mathematical Analysis and Statistical Inference*
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

[4]*Department of Psychology, University of Hawaii at Manoa*
Honolulu, Hawaii 96822, USA

Abstract

In this paper, we define a class of cross-validatory model selection criteria as an estimator of the predictive risk based on a discrepancy between a candidate model and the true model. For a vector of unknown parameters, n estimators are required for the definition of the class, where n is the sample size. The ith estimator ($i = 1, \ldots, n$) is obtained by minimizing a weighted discrepancy function in which the ith observation has a weight of $1 - \lambda$ and others have weight of 1. Cross-validatory model selection criteria in the class are specified by the individual λ. The sample discrepancy function and the ordinary cross-validation (CV) criterion are special cases of the class. One may choose λ to minimize the biases. The

[1]Corresponding author. E-mail: yanagi@math.sci.hiroshima-u.ac.jp

optimal λ makes the bias-corrected CV (CCV) criterion a second-order unbiased estimator for the risk, while the ordinary CV criterion is a first-order unbiased estimator of the risk.

AMS 2000 *subject classifications*. Primary 62H25; Secondary 62F07.
Key words: Asymptotic expansion, Bias correction, Cross-validation criterion, Model misspecification, Model selection, Predictive discrepancy, Sample discrepancy function, Structural equation model.

1. Introduction

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be a random sample from a p-dimensional population $\boldsymbol{y}$ whose probability density function $\varphi(\boldsymbol{y})$ is unknown. Nevertheless, the true model can be expressed as

$$M^* : \ \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ \varphi(\boldsymbol{y}). \tag{1}$$

Consider a family of parametric models $\mathcal{F} = \{f(\boldsymbol{y}|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q\}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is a q-dimensional vector of unknown parameters. This implies that a candidate model is given by

$$M : \ \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ f(\boldsymbol{y}|\boldsymbol{\theta}). \tag{2}$$

Most model selection criteria (or information criteria) for determining the best model among all candidate models are estimators of the predictive risk based on the discrepancy between the candidate model and the true model. For example, Akaike's information criterion (AIC; Akaike, 1973, 1974), Takeuchi's (1976) bias-corrected information criterion (TIC), and the extended information criterion (EIC; Ishiguro, Sakamoto, & Kitagawa, 1997) are estimators of the predictive Kullback-Leibler (K-L) discrepancy (Kullback & Leibler, 1951). On the other hand, the cross-validation (CV) criterion (Stone, 1974, 1977) serves as an estimator of the predictive risk based on an arbitrary discrepancy, e.g., K-L discrepancy, L_2 distance, and density power divergence (Basu *et al.*, 1998).

In this paper, we propose a class of model selection criteria by the cross-validatory method. For a given discrepancy, n estimators of $\boldsymbol{\theta}$ are required to define the class. The ith estimator ($i = 1, \ldots, n$) is obtained by minimizing a weighted discrepancy function in which the ith observation has a weight of $1 - \lambda$ ($0 \leq \lambda \leq 1$) and others have weights of 1. Each λ represents a cross-validatory model selection criteria. The sample discrepancy function and the ordinary CV criterion correspond to $\lambda = 0$ and 1, respectively. From the viewpoint of second order asymptotics for biases, the optimal λ can be expanded as $\lambda = 1 - 1/(2n) + O(n^{-2})$.

The optimal λ yield a bias-corrected CV (CCV) criterion that corrects the bias to $O(n^{-2})$ while the bias of the ordinary CV criterion is $O(n^{-1})$. The CCV criterion extends the result of Yanagihara, Tonda, and Matsumoto (2006), which consists of the K-L discrepancy.

In Section 2, we define the class of cross-validatory model selection criteria and study its mathematical properties. In Section 3, we describe other model selection criteria and their properties. In Section 4, the developed criteria will be applied to selecting structural equation models (SEM) under the normal distribution assumption. In Section 5, via the Monte Carlo method, we check the mathematical properties of the developed model selection criteria and compare CV and CCV criteria with other criteria such as AIC, TIC, and EIC. Conclusions and discussion are given in Section 6. Technical details are provided in an appendix.

2. A Class of Cross-Validatory Model Selection Criteria

Suppose that $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ is a discrepancy function for the candidate model M in (2), which is typically a function of $f(\boldsymbol{y}|\boldsymbol{\theta})$. Let $\Psi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{w})$ be a weighted discrepancy function defined by

$$\Psi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{w}) = \sum_{i=1}^{n} w_i \psi(\boldsymbol{y}_i|\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)'$ and $\boldsymbol{w} = (w_1,\ldots,w_n)'$. Then an estimator of $\boldsymbol{\theta}$ is obtained by minimizing the discrepancy function $\Psi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{w})$ in (3), i.e.,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{w}). \tag{4}$$

It is easy to see that $\hat{\boldsymbol{\theta}}(\boldsymbol{w})$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ when $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta})$ and $\boldsymbol{w} = \boldsymbol{1}_n = (1,\ldots,1)'$. When

$$\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{\beta} f(\boldsymbol{y}|\boldsymbol{\theta})^{\beta} + \frac{1}{1+\beta}\int \{f(\boldsymbol{x}|\boldsymbol{\theta})\}^{1+\beta} d\boldsymbol{x},$$

$\Psi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{w})$ is the density power divergence (Basu *et al.*, 1998). An application of the power density divergence can be found in Fujisawa and Eguchi (2006). For simplicity, we use $\Psi(\boldsymbol{\theta}|\boldsymbol{Y}) = \Psi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{1}_n)$ and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{1}_n)$. Furthermore, we write $\hat{\boldsymbol{\theta}}_{[-i]} = \hat{\boldsymbol{\theta}}(\boldsymbol{1}_n - \boldsymbol{e}_i)$, where $\boldsymbol{e}_i$ is a $n \times 1$ vector whose ith element is 1, and the other elements are 0. Notice that $\hat{\boldsymbol{\theta}}_{[-i]}$ becomes the jackknife estimator evaluated from the ith jackknife sample, which is obtained from $\boldsymbol{Y}$ by deleting $\boldsymbol{y}_i$. Let $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n$ be $p \times 1$ independent random vectors from $\boldsymbol{u} \sim \varphi(\boldsymbol{u})$ with $\boldsymbol{U} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)'$, which is also independent from $\boldsymbol{Y}$. Notice that $\hat{\boldsymbol{\theta}}$ is a function of

$\boldsymbol{Y}$. We define a risk based on the predictive discrepancy $\Psi(\boldsymbol{\theta}|\cdot)$ as

$$R_{\mathrm{PD}} = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}\left[\Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{U})\right] = n\mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}\left[\psi(\boldsymbol{u}|\hat{\boldsymbol{\theta}})\right], \tag{5}$$

where $\mathrm{E}^*_{\boldsymbol{y}}$ and $\mathrm{E}^*_{\boldsymbol{u}}$ are expectations under the true model M^* in (1) with respect to $\boldsymbol{y}$ and $\boldsymbol{u}$, respectively. In model selection based on $\psi(\boldsymbol{y}|\boldsymbol{\theta})$, we regard the model having the smallest R_{PD} as the best model, which is typically different from the true model. In many contexts of statistical modeling, the aim is to determine the best model. Obtaining an unbiased estimator of R_{PD} will allows us to correctly evaluate the discrepancy between data and model, which will further facilitate the selection of the best model. The simplest estimator of R_{PD} is the sample discrepancy function $\Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})$. The CV criterion proposed by Stone (1974, 1977),

$$\mathrm{CV} = \sum_{i=1}^{n}\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]}), \tag{6}$$

is also an estimator of R_{PD}. Let

$$\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\vartheta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\psi(\boldsymbol{y}|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}}, \quad \boldsymbol{H}(\boldsymbol{y}|\boldsymbol{\vartheta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\psi(\boldsymbol{y}|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}}, \tag{7}$$

and

$$\boldsymbol{r}(\boldsymbol{\theta}) = \mathrm{E}^*_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta})], \qquad \boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}^*_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta})'], \qquad \boldsymbol{J}(\boldsymbol{\theta}) = \mathrm{E}^*_{\boldsymbol{y}}[\boldsymbol{H}(\boldsymbol{y}|\boldsymbol{\theta})]. \tag{8}$$

Suppose that $\boldsymbol{\theta}_0$ is a $q\times 1$ vector such that $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $n\to\infty$. Under proper conditions, as specified in White (1982), $\boldsymbol{\theta}_0$ satisfies

$$\boldsymbol{r}(\boldsymbol{\theta}_0) = \mathbf{0}_q, \tag{9}$$

where $\mathbf{0}_q$ is a vector of q zeros. Notice that $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is called the Fisher's information matrix when $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta})$. Because $\boldsymbol{\theta}_0$ is a local minimum of $\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta})]$, equation (9) leads to a natural assumption that $\boldsymbol{J}(\boldsymbol{\theta}_0)$ is positive definite.

Let $\hat{\boldsymbol{\theta}}_i(\lambda)$ $(0\le\lambda\le 1)$ be the estimator of $\boldsymbol{\theta}$, which is obtained by minimizing the weighted discrepancy function $\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - \lambda\boldsymbol{e}_i)$, i.e., $\hat{\boldsymbol{\theta}}_i(\lambda) = \hat{\boldsymbol{\theta}}(\mathbf{1}_n - \lambda\boldsymbol{e}_i)$. Notice that, with weight $\mathbf{1}_n - \lambda\boldsymbol{e}_i$, the effect of the ith observation $\boldsymbol{y}_i$ on $\hat{\boldsymbol{\theta}}_i(\lambda)$ decreases as λ increases. The estimator $\hat{\boldsymbol{\theta}}_i(\lambda)$ includes the ordinary estimator and the ith jackknife estimator as special cases, i.e., $\hat{\boldsymbol{\theta}}_i(0) = \hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_i(1) = \hat{\boldsymbol{\theta}}_{[-i]}$. Replacing $\hat{\boldsymbol{\theta}}$ by $\hat{\boldsymbol{\theta}}_i(\lambda)$, we define the following cross-validatory model selection criterion:

$$\mathrm{CV}(\lambda) = \sum_{i=1}^{n}\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda)), \quad (0\le\lambda\le 1). \tag{10}$$

Let

$$\mathcal{G}_\lambda = \{\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(\lambda)]|\ 0 \leq \lambda \leq 1\},$$

and

$$R_1 = \sum_{i=1}^{n} \mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{r}(\hat{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})\right],\ R_2 = \sum_{i=1}^{n} \mathrm{E}^*_{\boldsymbol{y}}\left[(\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})'\boldsymbol{J}(\bar{\boldsymbol{\theta}}_i(\delta_i))(\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})\right], \tag{11}$$

where

$$\bar{\boldsymbol{\theta}}_i(\delta_i) = \hat{\boldsymbol{\theta}} + \delta_i(\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \quad (i = 1, \ldots, n), \tag{12}$$

with $\delta_i \in (0, 1)$. The following theorem characterizes the properties of $\mathrm{CV}(\lambda)$ (the proof is given in Appendix A.1).

THEOREM 1. *The model selection criterion* $\mathrm{CV}(\lambda)$ *has the following properties:*

1. $\mathrm{CV}(0) = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})$ *and* $\mathrm{CV}(1) = \mathrm{CV}$.
2. $\mathrm{CV}(\lambda)$ *is an increasing function of* λ.
3. $n\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta}_0)] \in \mathcal{G}_\lambda$ *when* $\boldsymbol{\theta}_0$ *is a global minimum of* $\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta})]$.
4. $R_{\mathrm{PD}} \in \mathcal{G}_\lambda$ *when* $R_1 + R_2/2 \geq 0$.

Appendix A.2 provides the detail leading to $R_1 = O(n^{-2})$ and $R_2 = \gamma_1 + O(n^{-2})$, where γ_1 is given by

$$\gamma_1 = \mathrm{tr}\{\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\}. \tag{13}$$

Because $\boldsymbol{J}(\boldsymbol{\theta}_0)$ is positive definite, γ_1 is positive. Thus, $R_1 + R_2/2 \geq 0$ asymptotically holds. Consequently, $R_{\mathrm{PD}} \in \mathcal{G}_\lambda$ when n is adequate. When $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ is a strictly convex function of $\boldsymbol{\theta}$, $\boldsymbol{H}(\boldsymbol{y}|\boldsymbol{\theta})$ is positive definite for any $\boldsymbol{\theta}$ and $\boldsymbol{y}$ (see e.g., Lehmann & Casella, 1998, p. 49). Then, $\boldsymbol{J}(\boldsymbol{\theta})$ will be positive definite for any $\boldsymbol{\theta}$. This directly implies that $R_2 > 0$ when $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ is a strictly convex function of $\boldsymbol{\theta}$. Thus, $R_{\mathrm{PD}} \in \mathcal{G}_\lambda$ when $R_1 \geq 0$. Although the order of R_1 is $O(n^{-2})$, $R_1 \geq 0$ holds under special cases, as in the following example.

EXAMPLE 1. Suppose that the candidate model M and the true model M^* are given by

$$\begin{aligned} M:&\quad \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ M^*:&\quad \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\mu}^* \text{ and } \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}^*. \end{aligned}$$

If the K-L discrepancy is used to define CV(λ), Appendix A.3 shows that $R_1 > 0$ always holds. Thus, $R_{\rm PD} \in \mathcal{G}_\lambda$.

An important issue is how to choose λ. It follows from Theorem 1 that, when $R_1 + R_2/2 \geq 0$, a λ_0 exists such that $\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(\lambda_0)] = R_{\rm PD}$. However, since λ_0 depends on the unknown distribution $\varphi(\boldsymbol{y})$, it is very difficult to find the exact λ_0. Even if we can obtain λ_0 somehow, it may be difficult to put it to practice. This is because the optimal λ_0 may depend on cumulants of $\varphi(\boldsymbol{y})$. It is difficult to obtain good estimates of higher-order cumulants even when n is relatively large (see Yanagihara (2007) for the case of kurtosis). Thus, an estimator of λ that does not involve higher-order cumulants is preferable. Let $\beta_{abcd} = \mathrm{E}^*_{\boldsymbol{y}}[\partial^4\psi(\boldsymbol{y}|\boldsymbol{\theta})/(\partial\theta_a\partial\theta_b\partial\theta_c\partial\theta_d)]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ $(1 \leq a, b, c, d \leq q)$. The following theorem characterizes the bias of CV(λ) (the proof is given in Appendix A.4).

THEOREM 2. *Under the condition* $|\beta_{abcd}| < \infty$ $(1 \leq a, b, c, d \leq q)$, *the bias of* CV($\lambda$) *is characterized as*

$$R_{\rm PD} - \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(\lambda)] = \begin{cases} (1-\lambda)\gamma_1 + O(n^{-1}) & (\lambda \textit{ is independent of } n) \\ \{1-\lambda-1/(2n)\}\,\gamma_1 + O(n^{-2}) & (\lambda = 1 + O(n^{-1})) \end{cases}, \tag{14}$$

where γ_1 *is given by* (13).

The moment condition in Theorem 2 (also Theorems A.1 and A.2) may be weakened as in Hall (1987). If $\lambda = 1 - 1/(2n)$, then the $O(n^{-1})$ term in the bias of CV(λ) in Theorem 2 vanishes. Thus, using second-order asymptotics, the optimal value of λ is $\lambda = 1 - 1/(2n) + O(n^{-2})$. Based on this, we propose a bias-corrected CV (CCV) criterion as in the following theorem.

THEOREM 3. *Let* $a_n \in (0, 1)$ *that can be expanded as* $a_n = 1 - 1/(2n) + O(n^{-2})$, *and*

$$\mathrm{CCV} = \mathrm{CV}(a_n) = \sum_{i=1}^{n} \psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(a_n)). \tag{15}$$

Then the bias of the CCV criterion is $O(n^{-2})$, *while the bias in the ordinary CV criterion is* $O(n^{-1})$. *Because* $a_n < 1$, *the CCV criterion is always smaller than the ordinary CV criterion.*

Notice that the CCV in (15) coincides with the CCV criterion in Yanagihara, Tonda, and Matsumoto (2006) when $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -2\log f(\boldsymbol{y}|\boldsymbol{\theta})$.

Since our assumption is that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are *i.i.d.*, it may seem that Theorem 3 does not apply to selecting explanatory variables in regression models, which are widely used in data analysis. Let $\boldsymbol{y} = (\boldsymbol{z}', \boldsymbol{x}')'$, where $\boldsymbol{z}$ is the vector of response variables and $\boldsymbol{x}$ is the vector of explanatory variables. Then our result immediately applies to the regression model. In order to calculate $\mathrm{CV}(\lambda)$, it is often necessary to obtain each $\hat{\boldsymbol{\theta}}_i(\lambda)$. However, $\mathrm{CV}(\lambda)$ in the linear regression model under the normal distribution assumption can be derived using $\hat{\boldsymbol{\theta}}$ alone, as in the following example.

EXAMPLE 2. Let $\boldsymbol{z}$ and $\boldsymbol{x}$ be $m \times 1$ and $k \times 1$ vectors and $\boldsymbol{y} = (\boldsymbol{z}', \boldsymbol{x}')'$. Suppose that the candidate model M and the true model M^* are given by

$$\begin{aligned} M: \quad & \boldsymbol{z}_i | \boldsymbol{x}_i \sim \mathrm{N}_m(\boldsymbol{\Xi}' \tilde{\boldsymbol{x}}_i, \boldsymbol{\Gamma}), \\ M^*: \quad & \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\mu}^* \text{ and } \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}^*, \end{aligned}$$

where $\tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i')'$. Notice that the MLEs of $\boldsymbol{\Xi}$ and $\boldsymbol{\Gamma}$ are $\hat{\boldsymbol{\Xi}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{Z}$ and $\hat{\boldsymbol{\Gamma}} = \boldsymbol{Z}'\{\boldsymbol{I}_n - \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\}\boldsymbol{Z}/n$, where $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ and $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)'$. Then, $\mathrm{CV}(\lambda)$ in the case of $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -2\log f(\boldsymbol{y}|\boldsymbol{\theta})$ is given by

$$\begin{aligned} \mathrm{CV}(\lambda) &= n\log|\hat{\boldsymbol{\Gamma}}| + nm\log\left(\frac{2n\pi}{n-\lambda}\right) + \sum_{i=1}^{n}\log\left\{1 - \frac{\lambda\hat{r}_i^2}{n(1-\lambda c_i)}\right\} \\ &\quad + \left(1 - \frac{\lambda}{n}\right)\sum_{i=1}^{n}\frac{\hat{r}_i^2}{(1-\lambda c_i)^2}\left\{1 - \frac{\lambda\hat{r}_i^2}{n(1-\lambda c_i)}\right\}^{-1}, \end{aligned}$$

where $c_i = \tilde{\boldsymbol{x}}_i'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{x}}_i$ and $\hat{r}_i^2 = (\boldsymbol{z}_i - \hat{\boldsymbol{\Xi}}'\tilde{\boldsymbol{x}}_i)'\hat{\boldsymbol{\Gamma}}^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\Xi}}'\tilde{\boldsymbol{x}}_i)$.

Yanagihara, Kamo, and Tonda (2006) proposed a second-order bias-corrected AIC, called $\mathrm{CAIC_J}$, in multivariate linear models. The order of the bias of $\mathrm{CAIC_J}$ is the same as that of CCV. However, $\mathrm{CAIC_J}$ was obtained under the assumption that the explanatory variables $\boldsymbol{x}$ are nonstochastic, while the condition here is that both the explanatory variables $\boldsymbol{x}$ and the response variables $\boldsymbol{z}$ are stochastic.

For linear regression models, the well-known CV criterion is defined by the predicted residual sum of squares. Our general formula also applies to this case, and the $\mathrm{CV}(\lambda)$ is given by the following example.

EXAMPLE 3. Let $\boldsymbol{x}$ be a $k \times 1$ vector and $\boldsymbol{y} = (z, \boldsymbol{x}')'$. Suppose that the candidate model M and the true model M^* are

$$M: \quad \mathrm{E}[z_i|\boldsymbol{x}_i] = \boldsymbol{\beta}'\tilde{\boldsymbol{x}}_i,$$

$$M^* : \ \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d. \ \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\mu}^* \text{ and } \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}^*,$$

where $\tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i')'$. Notice that the least square estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{z}$, where $\boldsymbol{z} = (z_1, \ldots, z_n)'$ and $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)'$. Thus, $\mathrm{CV}(\lambda)$ in the case of the predicted residual sum of squares is given by

$$\mathrm{CV}(\lambda) = \sum_{i=1}^{n} \left\{ \frac{z_i - \hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{x}}_i}{1 - \lambda\tilde{\boldsymbol{x}}_i'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{x}}_i} \right\}^2 .$$

3. Other Model Selection Criteria

In this section, we discuss other criteria for selecting the best model among all the candidate models using the general discrepancy function $\psi(\boldsymbol{y}|\boldsymbol{\theta})$. The AIC-type criterion can be defined by adding the number of parameters to the sample discrepancy function as

$$\mathrm{AIC} = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + q. \tag{16}$$

However, unless $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta})$ and $\mathcal{F}$ contains $\varphi(\boldsymbol{y})$, (16) has a constant bias in estimating R_{PD}. The TIC-type criterion corrects the bias of the AIC-type criterion, reducing the bias to $O(n^{-1})$. The TIC-type criterion is given by

$$\mathrm{TIC} = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \mathrm{tr}\{\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\}, \tag{17}$$

where

$$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})', \quad \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \tag{18}$$

with $\boldsymbol{g}(\cdot|\cdot)$ and $\boldsymbol{H}(\cdot|\cdot)$ being given by (7). Although the order of the bias in TIC is the same as that in the CV criterion, the bias of TIC tends to be larger than that of CV because $\mathrm{tr}\{\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\}$ may contain a large bias. Actually, Theorems A.1 and A.2 in Appendix A.2 show that the n^{-1} term of the bias in TIC contains more terms of higher-order moments than that of the CV criterion.

The bootstrap method can also correct the bias of the AIC-type criterion. The resulting criterion is called the EIC-type criterion. Let $\boldsymbol{y}^\star_{b,1}, ..., \boldsymbol{y}^\star_{b,n}$ be the bth bootstrap resample from $\boldsymbol{Y}$ $(b = 1, \ldots, B)$ and $\hat{\boldsymbol{\theta}}^\star_b$ be the estimator of $\boldsymbol{\theta}$, where

$$\hat{\boldsymbol{\theta}}^\star_b = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \psi(\boldsymbol{y}^\star_{b,i}|\boldsymbol{\theta}).$$

Replacing the log-likelihood function by the discrepancy function $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ in the formula of Konishi (1999), the EIC-type criterion can be defined by

$$\mathrm{EIC} = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \frac{1}{B}\sum_{b=1}^{B}\left\{\sum_{i=1}^{n}\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_b^{\star}) - \sum_{i=1}^{n}\psi(\boldsymbol{y}_{b,i}^{\star}|\hat{\boldsymbol{\theta}}_b^{\star})\right\}. \tag{19}$$

By using random vectors distributed according to the multinomial distribution, we can rewrite the definition of EIC in (19). Let $\boldsymbol{d}_b = (d_{b1}, \ldots, d_{bn})'$ $(b = 1, \ldots, B)$ be random samples of size n from the multinomial distribution $\mathrm{Multi}_n(n; 1/n, \ldots, 1/n)$. Then, the EIC in (19) is equivalent to the following formula (the derivation is given in Appendix A.5):

$$\mathrm{EIC} = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \frac{1}{B}\sum_{b=1}^{B}\Psi(\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)|\boldsymbol{Y}, \boldsymbol{1}_n - \boldsymbol{d}_b). \tag{20}$$

where $\hat{\boldsymbol{\theta}}(\cdot)$ is given by (4). Because the bias of EIC is $O(n^{-1})$, the order of bias in EIC is the same as those in TIC and the ordinary CV criterion. However, since EIC does not contain the term $\mathrm{tr}\{\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\}$, the bias of EIC tends to be smaller than that of TIC. On the other hand, EIC involves more computation than the CV criterion. Furthermore, EIC may behave poorly when the sample size is small and the number of parameters is large. Caution is needed when using EIC with small samples.

4. Application to Selecting Structural Equation Models Under the Normal Distribution Assumption

SEM is a multivariate statistical technique designed to model the covariance matrix by a structure with relatively few parameters (see e.g., Lee & Kontoghiorghes, 2007; Yuan & Bentler, 2007). The normal distribution assumption is typically used in the practice of SEM and is the default option of all statistical software (AMOS, EQS, LISREL, Mplus, SAS Calis). We will obtain the analytical expression of $\mathrm{CV}(\lambda)$ when the candidate model is from the normal family while the true model is unknown.

Let the candidate model M and the true model M^* be

$$\begin{aligned} M &: \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\xi})), \\ M^* &: \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d.\ \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\mu}^* \text{ and } \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}^*, \end{aligned} \tag{21}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)'$ are $p \times 1$ and $q \times 1$ unknown vectors of parameters, respectively, and the true distribution of $\boldsymbol{y}$ is unknown. Consider the K-L

discrepancy with

$$\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -2\log f(\boldsymbol{y}|\boldsymbol{\theta}) = p\log(2\pi) + \log|\boldsymbol{\Sigma}(\boldsymbol{\xi})| + (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{\xi})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{22}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\xi}')'$. Let

$$\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i, \quad \boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})', \tag{23}$$

and

$$F(\boldsymbol{\xi}|\boldsymbol{A}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\xi})| + \mathrm{tr}\left\{\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\xi})^{-1}\right\}. \tag{24}$$

Then, the CV(λ) defined in (10) is given by the following corollary (the proof is given in Appendix A.6).

COROLLARY 1. *The* CV(λ) *under the candidate model* M *in* (21) *is given by*

$$\begin{aligned}\mathrm{CV}(\lambda) &= np\log(2\pi) \\ &\quad + \sum_{i=1}^{n}\left\{\log\left|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_i(\lambda))\right| + \left(\frac{n}{n-\lambda}\right)^2(\boldsymbol{y}_i-\bar{\boldsymbol{y}})'\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}}_i(\lambda))^{-1}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})\right\},\end{aligned} \tag{25}$$

where $\hat{\boldsymbol{\xi}}_i(\lambda)$ *is the estimator of* $\boldsymbol{\xi}$ *defined by*

$$\hat{\boldsymbol{\xi}}_i(\lambda) = \arg\min_{\boldsymbol{\xi}} F(\boldsymbol{\xi}|\boldsymbol{S}_i(\lambda)), \tag{26}$$

with

$$\boldsymbol{S}_i(\lambda) = \frac{n}{n-\lambda}\left\{\boldsymbol{S} - \frac{\lambda}{n-\lambda}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})'\right\}. \tag{27}$$

The following corollary provides the analytical expression for other model selection criteria.

COROLLARY 2. *Let* $\hat{\boldsymbol{\xi}}$ *be the estimator of* $\boldsymbol{\xi}$ *such that*

$$\hat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} F(\boldsymbol{\xi}|\boldsymbol{S}). \tag{28}$$

Then, AIC, TIC, and EIC under the candidate model M *in* (21) *are given by*

$$\mathrm{AIC} = nF(\hat{\boldsymbol{\xi}}|\boldsymbol{S}) + np\log(2\pi) + 2(p+q);$$

$$\begin{aligned}\mathrm{TIC} &= \mathrm{AIC} - 2(p+q) + 2F(\hat{\boldsymbol{\xi}}|\boldsymbol{S}) \\ &\quad + \mathrm{tr}\left\{\hat{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1}\otimes\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1})\boldsymbol{Q}(\hat{\boldsymbol{\xi}}|\boldsymbol{S})(\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1}\otimes\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1})\right\} \\ &\quad - \mathrm{vec}(\boldsymbol{S})'(\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1}\otimes\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1})\boldsymbol{Q}(\hat{\boldsymbol{\xi}}|\boldsymbol{S})(\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1}\otimes\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}})^{-1})\mathrm{vec}(\boldsymbol{S}),\end{aligned}$$

where

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}((\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})')\mathrm{vec}((\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})')',$$

$$\boldsymbol{Q}(\hat{\boldsymbol{\xi}}|\boldsymbol{S}) = \left\{\frac{\partial}{\partial \boldsymbol{\xi}'}\mathrm{vec}(\boldsymbol{\Sigma}(\boldsymbol{\xi}))\right\}\left\{\frac{\partial^2}{\partial \boldsymbol{\xi}'\partial \boldsymbol{\xi}'}F(\boldsymbol{\xi}|\boldsymbol{S})\right\}^{-1}\left\{\frac{\partial}{\partial \boldsymbol{\xi}}\mathrm{vec}(\boldsymbol{\Sigma}(\boldsymbol{\xi}))'\right\}\Bigg|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}};$$

$$\mathrm{EIC} = \mathrm{AIC} - 2(p+q) + \frac{n}{B}\sum_{b=1}^{B}\mathrm{tr}\left\{\boldsymbol{V}(\boldsymbol{d}_b)\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}}(\boldsymbol{d}_b))^{-1}\right\}, \tag{29}$$

where

$$\hat{\boldsymbol{\xi}}(\boldsymbol{d}_b) = \arg\min_{\boldsymbol{\xi}} F(\boldsymbol{\xi}|\boldsymbol{S}(\boldsymbol{d}_b)), \tag{30}$$

$$\boldsymbol{V}(\boldsymbol{d}_b) = \frac{1}{n}\boldsymbol{Y}'\left\{\boldsymbol{I}_p - \mathrm{diag}(\boldsymbol{d}_b) + \frac{1}{n}\left(2\boldsymbol{d}_b\boldsymbol{d}_b' - \boldsymbol{1}_n\boldsymbol{d}_b' - \boldsymbol{d}_b\boldsymbol{1}_n'\right)\right\}\boldsymbol{Y}, \tag{31}$$

with

$$\boldsymbol{S}(\boldsymbol{d}_b) = \frac{1}{n}\boldsymbol{Y}'\left\{\mathrm{diag}(\boldsymbol{d}_b) - \frac{1}{n}\boldsymbol{d}_b\boldsymbol{d}_b'\right\}\boldsymbol{Y}. \tag{32}$$

The use of AIC for selecting the number of factors in the explanatory factor model was discussed by Akaike (1987). TIC for selecting SEM models under the normal distribution assumption was obtained by Yanagihara (2005). The details leading to the expression for EIC are provided in Appendix A.7.

5. Numerical Examinations

In this section, we verify the mathematical properties of model selection criteria using a Monte Carlo method. In particular, we compare CV and CCV criteria with AIC, TIC, and EIC. Bayesian information criterion (BIC; Schwarz, 1978) and the consistent Akaike's information criterion (CAIC; Bozdogan, 1987) are also frequently used for model selection, but their expectations do not convergence to R_{PD}. Thus, our study will not include BIC and CAIC.

In designing the Monte Carlo, we let the candidate distribution be multivariate normal as in the previous section, while the true distribution varies. Let $\boldsymbol{y}$ be the 6×1 vector defined

by $\boldsymbol{y} = \boldsymbol{\Sigma}^{*1/2}\boldsymbol{\varepsilon}$, where

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 6 \end{pmatrix}.$$

We use Mardia's (1970) multivariate skewnesses $\kappa_{3,3}^{(1)}$ and $\kappa_{3,3}^{(2)}$ and kurtosis $\kappa_4^{(1)}$ to measure the departure of the candidate distribution from the true distribution. These are given by

$$\kappa_{3,3}^{(1)} = \mathrm{E}[(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_2)^3], \quad \kappa_{3,3}^{(2)} = \mathrm{E}[(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_1)(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_2)(\boldsymbol{\varepsilon}_2'\boldsymbol{\varepsilon}_2)], \quad \kappa_4^{(1)} = \mathrm{E}[(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_1)^2] - 48,$$

where $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independent random vectors having the same distribution of $\boldsymbol{\varepsilon}$.

Six populations or true models are created when the elements ε_j of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_6)'$ are independently and identically distributed standardized variables from each of the following six distributions:

1. *Normal Distribution*: $\varepsilon_j \sim \mathrm{N}(0,1)$, ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = 0$).

2. *Laplace Distribution*: ε_j is generated from the Laplace distribution with mean 0 and standard deviation 1 ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = 18$).

3. *Uniform Distribution*: ε_j is generated from the uniform distribution on $(-1, 1)$, divided by the standard deviation $1/\sqrt{3}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = -7.2$).

4. *Skew-Laplace Distribution*: ε_j is generated from the skew-Laplace distribution with location parameter 0, dispersion parameter 1 and skew parameter 1, standardized by mean 3/4 and standard deviation $\sqrt{23}/4$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} \approx 7.32$ and $\kappa_4^{(1)} \approx 19.56$).

5. *Chi-Square Distribution*: ε_j is generated from the chi-square distribution with 2 degrees of freedom, standardized by mean 2 and standard deviation 2 ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 12$ and $\kappa_4^{(1)} = 36$).

6. *Log-Normal Distribution*: ε_j is generated from the lognormal distribution such that $\log \varepsilon_j \sim \mathrm{N}(0, 1/2)$, standardized by mean $e^{1/4}$ and standard deviation $\sqrt{e^{1/2}(e^{1/2} - 1)}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} \approx 17.64$ and $\kappa_4^{(1)} \approx 111.06$).

The skew-Laplace distribution was proposed by Balakrishnan and Ambagaspitiya (1994) (for the probability density function, see e.g., Yanagihara & Yuan, 2005). The distributions in 1, 2, and 3 are symmetric, and distributions in 4, 5, and 6 are skewed.

A sample of size 20 is generated from $\boldsymbol{y} = \boldsymbol{\Sigma}^{*1/2}\boldsymbol{\varepsilon}$. The three candidate models are:

$$\begin{aligned}
\text{Model 1, } M_1 &: \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{20} \sim i.i.d.\ \mathrm{N}_6(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_6), \\
\text{Model 2, } M_2 &: \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{20} \sim i.i.d.\ \mathrm{N}_6(\boldsymbol{\mu}, (\sigma^2 - \rho)\boldsymbol{I}_6 + \rho \mathbf{1}_6 \mathbf{1}_6'), \\
\text{Model 3, } M_3 &: \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{20} \sim i.i.d.\ \mathrm{N}_6(\boldsymbol{\mu}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2, \sigma_6^2)).
\end{aligned}$$

Because the sample size n $(= 20)$ is rather small compared with the dimension p $(= 6)$, the saturated model, i.e., $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{20} \sim i.i.d.\ \mathrm{N}_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is not considered here. Since $\boldsymbol{\Sigma}^* \neq \boldsymbol{\Sigma}(\boldsymbol{\xi})$ for any $\boldsymbol{\xi}$ in any of the candidate models, all the candidate models are misspecified. We use the K-L discrepancy to select the best model among the three candidates. For each of the candidate models and distributions, results of Appendix A.3 imply that $R_1 > 0$ and $R_2 > 0$. Thus, $R_{\mathrm{PD}} \in \mathcal{G}_\lambda$ in all three models.

The number of replications is chosen as $N_r = 10,000$. The following quantities are evaluated at each replication: $\mathrm{CV}(\lambda)$ with $\lambda = 0.00, 0.01, 0.02, \ldots, 0.98, 0.99, 1.00$; CCV $= \mathrm{CV}(a_n)$ with $a_n = \sqrt{n/(n+1)} = \sqrt{20/21}$; AIC; TIC; and EIC using $B = 1,000$ nested resamples. For each of the N_r $\hat{\boldsymbol{\theta}}$'s, $R = \sum_{i=1}^{20} \psi(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})$ with $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{20}$ being simulated from $\boldsymbol{u} = \boldsymbol{\Sigma}^{*1/2}\boldsymbol{\varepsilon}$ is also obtained, where $\boldsymbol{u}_i$ are independent of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{20}$. The average of R across the N_r replications, $\bar{R}$, is regarded as the risk R_{PD}. Let $\overline{\mathrm{IC}}$ be the average of any of the above criteria; the relative bias and relative root mean square error (RMSE) of the criterion are evaluated by

$$\text{Relative Bias} = \frac{\bar{R} - \overline{\mathrm{IC}}}{|\bar{R}|} \times 100, \quad \text{Relative RMSE} = \frac{\sqrt{\sum_{l=1}^{N_r} (\bar{R} - \mathrm{IC}_l)^2 / N_r}}{|\bar{R}|} \times 100.$$

The smallest IC at each replication for a given model is recorded, as are its frequencies among the 10,000 replications.

Table 1 contains the risks $(\bar{R})$ of all the candidate models at each true distribution. Model 3 has the smallest risk when the true distribution is normal and uniform; model 2 becomes the best when the true distribution is Laplace, skew-Laplace, chi-square, or log-normal.

Insert Table 1 around here

Figures 1 and 2 contain the plots of relative biases and RMSEs of $\mathrm{CV}(\lambda)$ against λ, respectively. Figure 3 contains the frequencies of the model being selected by $\mathrm{CV}(\lambda)$. The plots in Figure 1 clearly show that there is an λ_0 which makes $\mathrm{CV}(\lambda)$ an unbiased estimator of R_{PD}. In all the figures, the optimal λ_0 is close to 1.0 or approximately $1 - 1/(2n) = 39/40$.

The bias approaches 0 as λ moves towards λ_0, and departs from 0 as λ moves away from λ_0. Larger biases of $\mathrm{CV}(\lambda)$ are associated with more unknown parameters or larger multivariate kurtosis of the true distribution ($\kappa_4^{(1)}$). Comparing the plots for Laplace and skew-Laplace distributions, we may notice that the sizes of multivariate skewnesses $\kappa_{3,3}^{(1)}$ and $\kappa_{3,3}^{(2)}$ have little effect on the bias of $\mathrm{CV}(\lambda)$. Similar to Figure 1, the plots in Figure 2 clearly show that, regardless of the model and distribution, there exists an $\lambda_{\mathrm{M}} \in (0, 1)$ such that $\mathrm{CV}(\lambda_{\mathrm{M}})$ has the smallest RMSE. Furthermore, Figure 3 shows that $\mathrm{CV}(\lambda)$ tends to choose model 3 for smaller λ and model 2 for larger λ.

Insert Figures 1, 2, and 3 around here

Table 2 contains the relative biases, RMSEs, and the frequencies of each of the models being selected by AIC, TIC, EIC, CV, and CCV criteria. The table clearly shows that the CCV criterion has the smallest bias among all the criteria. Moreover, the CCV criterion not only improves the bias of the CV criterion, but also its RMSE. The biases of AIC and TIC are greater than those of EIC, CV, and CCV criteria. In particular, AIC has a very large bias when $\kappa_4^{(1)}$ is large. RMSE of EIC tends to be smaller than that of the CV criterion, although the bias of EIC tends to be greater than that of the CV criterion. Comparing Tables 1 and 2, CV and CCV select the model with the smallest risk most often. But AIC and TIC select model 3 most often while the best model changes with the true distribution. Notice that the frequency of choosing the best model by each criterion varies when the true distribution changes. Table 3 contains the average frequencies of choosing the best model by each criterion across all the true models. Among the 5 criteria, CCV chooses the best model most frequently; EIC and CV also work well.

In addition to the results reported above, several other models were also studied and similar results were obtained. While the frequency of choosing the best model by each criterion changes with the true model/distribution, the best criterion is mostly among EIC, CV and CCV.

Insert Tables 2 and 3 around here

6. Conclusion

In this paper, we defined the class of cross-validatory model selection criterion $\mathrm{CV}(\lambda)$ $(0 \leq \lambda \leq 1)$, which includes the sample discrepancy function and the ordinary CV criterion

as special cases. $\mathrm{CV}(\lambda)$ is an increasing function of λ. In particular, under proper conditions, there exists an $\lambda_0 \in [0,1]$ such that $\mathrm{CV}(\lambda_0)$ is unbiased for R_{PD}. Because $R_1 = O(n^{-2})$ and $R_2 = \gamma_1 + O(n^{-2})$ with $\gamma_1 > 0$, $|R_1|$ tends to be smaller than $|R_2|$. Thus, $R_{\mathrm{PD}} \in \mathcal{G}_\lambda$ in most cases. From the viewpoint of second-order asymptotics for the bias, $\lambda = 1 - 1/(2n) + O(n^{-2})$ is optimal. We found that $\lambda = \sqrt{n/(n+1)}$ worked well empirically. In particular, without estimating any higher-order cumulants, such a λ reduces the bias in CCV to $O(n^{-2})$. Such a result is especially valuable with small samples, where any criterion involving higher-order cumulants will inevitably perform poorly. The Monte Carlo results in the previous section verify the merit of CCV. In addition to the CCV criterion, other second-order bias-corrected criteria also exist. Those other criteria were generally obtained under specified models and distributions. The CCV criterion here is obtained under the general assumption, and can be applied broadly.

The aim of the CCV criterion is to minimize the bias in estimating R_{PD}. More important theme is to have a criterion that selects the model with the smallest risk. An unbiased estimator of R_{RD} does not necessarily lead to the model with the smallest R_{PD} being selected most frequently. Fortunately, the merits of least bias and selecting the best model both occur most frequently with CCV. Thus, we recommend the use of the CCV criterion for general model selection.

Acknowledgements

Hirokazu Yanagihara's research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #17700274, 2005–2007. Ke-Hai Yuan's research was supported by NSF grant DMS-0437167 and the James McKeen Cattell Fund.

Appendix

A.1. Proof of Theorem 1

Proof of Property 1: We omit the proof because it is easy to verify.

Proof of Property 2: Let $\Psi_i(\boldsymbol{\theta}|\boldsymbol{Y},\lambda) = \Psi(\boldsymbol{\theta}|\boldsymbol{Y},\mathbf{1}_n - \lambda\boldsymbol{e}_i)$, and $\lambda_1 < \lambda_2$. Because $\hat{\boldsymbol{\theta}}_i(\lambda)$ minimizes $\Psi_i(\boldsymbol{\theta}|\boldsymbol{Y},\lambda)$, there exist $\Psi_i(\hat{\boldsymbol{\theta}}(\lambda_1)|\boldsymbol{Y},\lambda_1) \leq \Psi_i(\hat{\boldsymbol{\theta}}(\lambda_2)|\boldsymbol{Y},\lambda_1)$ and $\Psi_i(\hat{\boldsymbol{\theta}}(\lambda_2)|\boldsymbol{Y},\lambda_2) \leq$

$\Psi_i(\hat{\boldsymbol{\theta}}(\lambda_1)|\boldsymbol{Y}, \lambda_2)$. By using these relations we obtain

$$\begin{aligned}
\Psi(\hat{\boldsymbol{\theta}}_i(\lambda_1)|\boldsymbol{Y}) - \lambda_1\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_1)) &= \Psi_i(\hat{\boldsymbol{\theta}}_i(\lambda_1)|\boldsymbol{Y}, \lambda_1) \\
&\leq \Psi_i(\hat{\boldsymbol{\theta}}_i(\lambda_2)|\boldsymbol{Y}, \lambda_1) \\
&= \Psi(\hat{\boldsymbol{\theta}}_i(\lambda_2)|\boldsymbol{Y}) - \lambda_1\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_2)) \\
&= \Psi_i(\hat{\boldsymbol{\theta}}_i(\lambda_2)|\boldsymbol{Y}, \lambda_2) + (\lambda_2 - \lambda_1)\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_2)) \\
&\leq \Psi_i(\hat{\boldsymbol{\theta}}_i(\lambda_1)|\boldsymbol{Y}, \lambda_2) + (\lambda_2 - \lambda_1)\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_2)) \\
&= \Psi(\hat{\boldsymbol{\theta}}_i(\lambda_1)|\boldsymbol{Y}) - \lambda_2\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_1)) + (\lambda_2 - \lambda_1)\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_2)).
\end{aligned}$$

Thus,

$$\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_1)) \leq \psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_2)). \tag{A1}$$

It follows from (A1) that

$$\mathrm{CV}(\lambda_1) = \sum_{i=1}^{n} \psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_1)) \leq \sum_{i=1}^{n} \psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda_2)) = \mathrm{CV}(\lambda_2).$$

Consequently, $\mathrm{CV}(\lambda)$ is an increasing function of λ.

Proof of Property 3: Because $\hat{\boldsymbol{\theta}}_{[-i]}$ minimizes $\sum_{j\neq i}^{n} \psi(\boldsymbol{y}_j|\boldsymbol{\theta})$, there exists

$$\sum_{j\neq i}^{n} \psi(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}_{[-i]}) \leq \sum_{j\neq i}^{n} \psi(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}).$$

Thus,

$$\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}_{[-i]})] \leq \mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}})], \quad (j \neq i). \tag{A2}$$

Let $\hat{\boldsymbol{\theta}}_n$ be the minimizer of the discrepancy function based on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, and $\alpha_n = \mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_1|\hat{\boldsymbol{\theta}}_n)]$. Then $\alpha_{n-1} = \mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}_{[-i]})]$ and $\alpha_n = \mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}})]$. It follows from (A2) that $\alpha_{n-1} \leq \alpha_n$ for any n. Thus, α_n monotonically increases. Let $\gamma(\boldsymbol{\theta}) = \mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta})]$. Then $\lim_{n\to\infty} \alpha_n = \gamma(\boldsymbol{\theta}_0)$ follows from $\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$. Therefore, α_n is bounded and monotonically increases. This directly implies that $\alpha_n \leq \gamma(\boldsymbol{\theta}_0)$ and

$$\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(0)] \leq n\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta}_0)]. \tag{A3}$$

On the other hand, if $\boldsymbol{\theta}_0$ is the global minimum of $\gamma(\boldsymbol{\theta})$, there must exist $\gamma(\boldsymbol{\theta}_0) \leq \gamma(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$. Thus, $\gamma(\boldsymbol{\theta}_0) \leq \mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})]$, or equivalently

$$n\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta}_0)] \leq \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(1)]. \tag{A4}$$

Equations (A3) and (A4) imply $n\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}|\boldsymbol{\theta}_0)] \in \mathcal{G}_\lambda$.

Proof of Property 4: We will first show $\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}] \leq R_{\mathrm{PD}}$, where R_{PD} is given by (5). Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{U}}$ be the minimizer of $\Psi(\boldsymbol{\theta}|\boldsymbol{U})$. Because $\boldsymbol{U}$ and $\boldsymbol{Y}$ are identically distributed, $\mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}[\Psi(\hat{\boldsymbol{\theta}}_{\boldsymbol{U}}|\boldsymbol{Y})] = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}[\Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{U})] = R_{\mathrm{PD}}$. The property $\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(0)] \leq R_{\mathrm{PD}}$ follows by noticing that $\mathrm{CV}(0) = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) \leq \Psi(\hat{\boldsymbol{\theta}}_{\boldsymbol{U}}|\boldsymbol{Y})$. We next show that $R_{\mathrm{PD}} \leq \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(1)]$ when $R_1 + R_2/2 \geq 0$, where R_1 and R_2 are given by (11). Notice that $\mathrm{CV}(1) = \mathrm{CV}$ and $\hat{\boldsymbol{\theta}}_{[-i]}$ and $\boldsymbol{y}_i$ are independent. Because the distribution of $\boldsymbol{u}_i$ is identical to that of $\boldsymbol{y}_i$, $\mathrm{E}^*_{\boldsymbol{y}}[\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})] = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}[\psi(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}}_{[-i]})]$. Applying the Taylor expansion at $\hat{\boldsymbol{\theta}}$, we obtain

$$\sum_{i=1}^n \psi(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}}_{[-i]}) = \sum_{i=1}^n \psi(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}}) \quad + \quad \sum_{i=1}^n \boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})'\boldsymbol{H}(\boldsymbol{u}_i|\bar{\boldsymbol{\theta}}_i(\delta_i))(\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}),$$

where $\bar{\boldsymbol{\theta}}_i(\delta_i)$ is given by (12). Thus,

$$\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}] = R_{\mathrm{PD}} + R_1 + \frac{1}{2}R_2. \tag{A5}$$

Consequently, $R_{\mathrm{PD}} \leq \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(1)]$ whenever $R_1 + R_2/2 \geq 0$.

A.2. Expansions of Biases of CV and TIC

Let

$$\boldsymbol{L}(\boldsymbol{y}|\boldsymbol{\vartheta}) = \left(\frac{\partial}{\partial\boldsymbol{\theta}} \otimes \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\psi(\boldsymbol{y}|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}},$$

and

$$\boldsymbol{K}(\boldsymbol{\theta}) = \mathrm{E}^*_{\boldsymbol{y}}[\boldsymbol{L}(\boldsymbol{y}|\boldsymbol{\theta})], \quad \hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{L}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}).$$

Because $\hat{\boldsymbol{\theta}}_i(\lambda)$ is the minimizer of $\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - \lambda\boldsymbol{e}_i)$, there exists

$$\sum_{j=1}^n \boldsymbol{g}(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}_i(\lambda)) = \lambda\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i(\lambda)), \tag{A6}$$

where $\boldsymbol{g}(\cdot|\cdot)$ is given by (7). The following stochastic expansion is needed

$$\hat{\boldsymbol{\theta}}_i(\lambda) = \hat{\boldsymbol{\theta}} + \frac{\lambda}{n}\boldsymbol{z}_{1,i} + \frac{\lambda^2}{n^2}\boldsymbol{z}_{2,i} + O_p(n^{-3}), \tag{A7}$$

where $\lambda = O(1)$ and $\boldsymbol{z}_{1,i}$ and $\boldsymbol{z}_{2,1}$ are to be determined. Applying the Taylor expansion to both sides of (A6) at $\hat{\boldsymbol{\theta}}$, replacing $\hat{\boldsymbol{\theta}}_i(\lambda)$ by (A7), and comparing the $O(n^{-1})$ and $O(n^{-2})$

terms in both sides of the resulting equation in sequence, we obtain

$$\boldsymbol{z}_{1,i} = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \quad \boldsymbol{z}_{2,i} = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\left\{\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\boldsymbol{z}_{1,i} - \frac{1}{2}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\mathrm{vec}(\boldsymbol{z}_{1,i}\boldsymbol{z}_{1,i}')\right\}, \tag{A8}$$

where $\boldsymbol{H}(\cdot|\cdot)$ and $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})$ are given by (7) and (18), respectively. Notice that $\hat{\boldsymbol{\theta}}_{[-i]} = \hat{\boldsymbol{\theta}}_i(1)$. Substituting $\lambda = 1$ into (A7), we obtain the stochastic expansion of $\hat{\boldsymbol{\theta}}_{[-i]}$ as

$$\hat{\boldsymbol{\theta}}_{[-i]} = \hat{\boldsymbol{\theta}} + \frac{1}{n}\boldsymbol{z}_{1,i} + \frac{1}{n^2}\boldsymbol{z}_{2,i} + O_p(n^{-3}). \tag{A9}$$

In order to calculate the asymptotic expansion of the bias of the CV criterion in (6), we first substitute the stochastic expansion of $\hat{\boldsymbol{\theta}}_{[-i]}$ in (A9) into R_1 and R_2, where R_1 and R_2 are given by (11); we then use the relation $\bar{\boldsymbol{\theta}}_i(\delta_i) \xrightarrow{a.s.} \boldsymbol{\theta}_0$, where $\bar{\boldsymbol{\theta}}_i(\delta_i)$ and $\boldsymbol{\theta}_0$ are given by (12) and (9), respectively. These two steps yield

$$\begin{aligned} R_1 &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{r}(\hat{\boldsymbol{\theta}})'\left(\boldsymbol{z}_{1,i} + \frac{1}{n}\boldsymbol{z}_{2,i}\right)\right] + O(n^{-2}), \\ R_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{z}_{1,i}'\boldsymbol{J}(\boldsymbol{\theta}_0)\boldsymbol{z}_{1,i}'\right] + O(n^{-2}), \end{aligned}$$

where $\boldsymbol{r}(\cdot)$ and $\boldsymbol{J}(\cdot)$ are given by (8). Notice that $\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \boldsymbol{0}_q$ due to $\hat{\boldsymbol{\theta}}$ being the minimizer of $\Psi(\boldsymbol{\theta}|\boldsymbol{Y})$. Thus,

$$\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{r}(\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{1,i}\right] = \mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{r}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\right] = 0.$$

Moreover, from $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ and $\boldsymbol{r}(\boldsymbol{\theta}_0) = \boldsymbol{0}_q$, the second term in the expansion of R_1 is expanded as

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{r}(\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{2,i}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{r}(\boldsymbol{\theta}_0)'\boldsymbol{z}_{i,2}\right] + O(n^{-2}) = O(n^{-2}).$$

Consequently, $R_1 = O(n^{-2})$. Using $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{J}(\boldsymbol{\theta}_0)$, R_2 is expanded as

$$R_2 = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}_i|\boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{J}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\boldsymbol{\theta}_0)\right] + O(n^{-2}) = \frac{1}{n}\gamma_1 + O(n^{-2}),$$

where γ_1 is given by (13). Substituting the above two results into (A5) yields the following theorem.

THEOREM A.1. *When $|\beta_{abcd}| < \infty$ holds $(1 \le a, b, c, d \le q)$, the bias of the CV criterion is expanded as*

$$R_{\mathrm{PD}} - \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}] = -\frac{1}{2n}\gamma_1 + O(n^{-2}). \tag{A10}$$

Using Theorem A.1, we can easily obtain an expansion of the bias of TIC in (17). Applying the Taylor expansion of CV at $\hat{\boldsymbol{\theta}}$ yields

$$\mathrm{CV} = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + C_1 + \frac{1}{n}\left(C_2 + \frac{1}{2}C_3\right) + O_p(n^{-2}),$$

where C_1, C_2, and C_3 are given by

$$C_1 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{1,i}, \quad C_2 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{2,i}, \quad C_3 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{1,i}'\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{1,i}. \quad \text{(A11)}$$

Notice that $C_1 = \mathrm{tr}\{\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\}$ and $\mathrm{TIC} = \Psi(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \mathrm{tr}\{\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\}$. Thus,

$$\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{TIC}] = \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}] - \frac{1}{n}\left\{\mathrm{E}^*_{\boldsymbol{y}}[C_2] + \frac{1}{2}\mathrm{E}^*_{\boldsymbol{y}}[C_3]\right\} + O(n^{-2}).$$

By using $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$, $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{J}(\boldsymbol{\theta}_0)$, and $\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{K}(\boldsymbol{\theta}_0)$, we obtain

$$\begin{aligned}
\mathrm{E}^*_{\boldsymbol{y}}[C_2] &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\right] \\
&\quad -\frac{1}{2}\,\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\left\{[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})] \otimes [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})]\right\}\right]\Big\} \\
&= \gamma_2 - \frac{1}{2}\gamma_3 + O(n^{-1}), \\
\mathrm{E}^*_{\boldsymbol{y}}[C_3] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\right] \\
&= \gamma_2 + O(n^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
\gamma_2 &= \mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y}|\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)\right], \\
\gamma_3 &= \mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)\left\{[\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)] \otimes [\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)]\right\}\right].
\end{aligned}$$

Thus,

$$\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{TIC}] = \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}] - \frac{1}{2n}(3\gamma_2 - \gamma_3) + O(n^{-2}). \quad \text{(A12)}$$

Equations (A10) and (A12) lead to the following theorem.

THEOREM A.2. *When $|\beta_{abcd}| < \infty$ holds $(1 \le a, b, c, d \le q)$, the bias of TIC is expanded as*

$$R_{\mathrm{PD}} - \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{TIC}] = -\frac{1}{2n}(\gamma_1 - 3\gamma_2 + \gamma_3) + O(n^{-2}).$$

A.3. Proof of Example 1

The discrepancy function corresponding to the multivariate normal distribution is

$$\psi(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{2}\left\{p\log(2\pi) + \log|\boldsymbol{\Sigma}| + (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}', \mathrm{vech}(\boldsymbol{\Sigma})')$ with $\mathrm{vech}(\boldsymbol{A})$ being the vector of staking the distinct elements of a symmetric matrix $\boldsymbol{A}$ columnwise. Let $\boldsymbol{D}_p$ be the duplication matrix such that $\mathrm{vec}(\boldsymbol{A}) = \boldsymbol{D}_p\mathrm{vech}(\boldsymbol{A})$ (see Magnus & Neudecker, 1999, p. 48). Then, the corresponding $\boldsymbol{r}(\boldsymbol{\theta})$ in (8) is given by

$$\boldsymbol{r}(\boldsymbol{\theta}) = \frac{1}{2}\begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{D}_p'\boldsymbol{D}_p \end{pmatrix}\begin{pmatrix} 2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}^*) \\ \mathrm{vech}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\{\boldsymbol{\Sigma}^* + (\boldsymbol{\mu}-\boldsymbol{\mu}^*)(\boldsymbol{\mu}-\boldsymbol{\mu}^*)'\}\boldsymbol{\Sigma}^{-1}) \end{pmatrix}.$$

It is well known that the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\bar{\boldsymbol{y}}', \mathrm{vech}(\boldsymbol{S})')'$, where $\bar{\boldsymbol{y}}$ and $\boldsymbol{S}$ are the sample mean and covariance matrix given by (23). On the other hand, the ith jackknife estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{[-i]} = (\bar{\boldsymbol{y}}_{[-i]}', \mathrm{vech}(\boldsymbol{S}_{[-i]})')'$, where $\bar{\boldsymbol{y}}_{[-i]} = (n-1)^{-1}\sum_{j\neq i}^{n}\boldsymbol{y}_j$ and $\boldsymbol{S}_{[-i]} = (n-1)^{-1}\sum_{j\neq i}^{n}(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_{[-i]})(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_{[-i]})'$. Fujikoshi *et al.* (2003) gives

$$\bar{\boldsymbol{y}}_{[-i]} = \bar{\boldsymbol{y}} - \frac{1}{n-1}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}), \quad \boldsymbol{S}_{[-i]} = \frac{n}{n-1}\left\{\boldsymbol{S} - \frac{1}{n-1}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'\right\}.$$

Therefore, $\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}$ becomes

$$\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}} = \frac{1}{n-1}\begin{pmatrix} -(\boldsymbol{y}_i - \bar{\boldsymbol{y}}) \\ \mathrm{vech}(\boldsymbol{S}) - \dfrac{n}{n-1}\mathrm{vech}((\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})') \end{pmatrix}.$$

Notice that $\mathrm{vech}(\boldsymbol{A})'\boldsymbol{D}_p'\boldsymbol{D}_p\mathrm{vech}(\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}'\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}'\boldsymbol{B})$ for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. It follows from the definition of R_1 in (11) that

$$R_1 = \frac{n}{2(n-1)^2}\left\{\mathrm{E}^*_{\boldsymbol{y}}\left[\mathrm{tr}(\boldsymbol{\Sigma}^*\boldsymbol{S}^{-1})\right] + \mathrm{E}^*_{\boldsymbol{y}}\left[(\bar{\boldsymbol{y}}-\boldsymbol{\mu}^*)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}^*)\right] - p\right\}.$$

Jensen's inequality implies $\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{tr}(\boldsymbol{\Sigma}^*\boldsymbol{S}^{-1})] \geq np/(n-1) > p$, and thus $R_1 > 0$.

A.4. Proof of Theorem 2

It follows from (A9) and (A7) that

$$\hat{\boldsymbol{\theta}}_i(\lambda) = \hat{\boldsymbol{\theta}}_{[-i]} + \frac{1}{n}(\lambda-1)\boldsymbol{z}_{1,i} + \frac{1}{n^2}(\lambda^2-1)\boldsymbol{z}_{2,i} + O_p(n^{-3}),\ (\lambda = O(1)), \tag{A13}$$

where $\boldsymbol{z}_{1,i}$ and $\boldsymbol{z}_{2,i}$ are given by (A8). Using the expansion (A13) after applying the Taylor expansion of $\mathrm{CV}(\lambda)$ at $\hat{\boldsymbol{\theta}}_{[-i]}$ yields

$$\mathrm{CV}(\lambda) \;=\; \mathrm{CV} + (\lambda-1)C_1 + \frac{1}{n}\left\{(\lambda^2-1)C_2 + \frac{1}{2}(\lambda-1)^2 C_3\right\} + O_p(n^{-3}),$$

where C_1, C_2 and C_3 are given by (A11). Notice that $\mathrm{E}^*_{\boldsymbol{y}}[C_1] = O(1)$, $\mathrm{E}^*_{\boldsymbol{y}}[C_2] = O(1)$ and $\mathrm{E}^*_{\boldsymbol{y}}[C_3] = O(1)$. Using the expansion of $\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}]$ in (A10), we obtain

$$\begin{aligned}\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(\lambda)] \;&=\; R_{\mathrm{PD}} + (\lambda-1)\mathrm{E}^*_{\boldsymbol{y}}[C_1] \\ &\quad + \frac{1}{2n}\left\{\gamma_1 + 2(\lambda^2-1)\mathrm{E}^*_{\boldsymbol{y}}[C_2] + (\lambda-1)^2\mathrm{E}^*_{\boldsymbol{y}}[C_3]\right\} + O(n^{-2}),\end{aligned}$$

where γ_1 is given by (13). The first equation in (14) follows by noticing that $\mathrm{E}^*_{\boldsymbol{y}}[C_1] = \gamma_1 + O(n^{-1})$. If $\lambda = 1 + O(n^{-1})$, then $\lambda - 1 = O(n^{-1})$ and $\lambda^2 - 1 = O(n^{-1})$. Consequently,

$$\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}(\lambda)] = R_{\mathrm{PD}} + \frac{1}{2n}\left\{\gamma_1 + 2n(\lambda-1)\mathrm{E}^*_{\boldsymbol{y}}[C_1]\right\} + O(n^{-2}),$$

and from which the second equation in (14) follows.

A.5. Derivation of Redefining EIC

Notice that the bth bootstrap resample $\boldsymbol{y}^\star_{b,i}$ $(i = 1, \ldots, n)$ is one of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Let $\boldsymbol{d}_b = (d_{b1}, d_{b2}, \ldots, d_{bn})$ with d_{bi} equal to the number of times $\boldsymbol{y}_i$ appeares in the bth bootstrap resample. Then

$$\sum_{i=1}^n \psi(\boldsymbol{y}^\star_{b,i}|\boldsymbol{\theta}) = \sum_{i=1}^n d_{bi}\psi(\boldsymbol{y}_i|\boldsymbol{\theta}) = \Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d}_b).$$

Thus, $\hat{\boldsymbol{\theta}}^\star_b = \hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)$. Consequently,

$$\begin{aligned}\sum_{i=1}^n \psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}^\star_b) - \sum_{i=1}^n \psi(\boldsymbol{y}^\star_{b,i}|\hat{\boldsymbol{\theta}}^\star_b) \;&=\; \sum_{i=1}^n \psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)) - \sum_{i=1}^n d_{bi}\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)) \\ &=\; \sum_{i=1}^n (1-d_{bi})\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)) \\ &=\; \Psi(\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)|\boldsymbol{Y}, \boldsymbol{1}_n - \boldsymbol{d}_b). \qquad \text{(A14)}\end{aligned}$$

Substituting (A14) into (19) yields equation (20). The distribution property $\boldsymbol{d}_b \sim \mathrm{Multi}_n(n; 1/n, \ldots, 1/n)$ is the definition of the bootstrap sampling.

A.6. Proof of Corollary 1

For the discrepancy function given by (22), we obtain by direct calculation

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - \lambda \boldsymbol{e}_i) &= -2\boldsymbol{\Sigma}(\boldsymbol{\xi})^{-1}\left\{(n-\lambda)\boldsymbol{\mu} - (n\bar{\boldsymbol{y}} - \lambda \boldsymbol{y}_i)\right\}, \\ \frac{\partial}{\partial \boldsymbol{\xi}}\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - \lambda \boldsymbol{e}_i) &= \frac{\partial}{\partial \boldsymbol{\xi}}(n-\lambda)F(\boldsymbol{\xi}|\boldsymbol{M}_i(\boldsymbol{\mu}, \lambda)),\end{aligned}$$

where $F(\boldsymbol{\xi}|\cdot)$ is given by (24) and

$$\boldsymbol{M}_i(\boldsymbol{\mu}, \lambda) = \frac{1}{n-\lambda}\left\{\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{\mu})(\boldsymbol{y}_j - \boldsymbol{\mu})' - \lambda(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})'\right\}.$$

Denote $\hat{\boldsymbol{\theta}}_i(\lambda) = (\hat{\boldsymbol{\mu}}_i(\lambda)', \hat{\boldsymbol{\xi}}_i(\lambda)')'$. Solving the equation $\partial\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - \lambda\boldsymbol{e}_i)/\partial\boldsymbol{\mu} = \mathbf{0}_p$ leads to

$$\hat{\boldsymbol{\mu}}_i(\lambda) = \bar{\boldsymbol{y}} - \frac{\lambda}{n-\lambda}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}).$$

Notice that

$$\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i(\lambda) = \begin{cases} \dfrac{n}{n-\lambda}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}) & (j = i) \\ \boldsymbol{y}_j - \bar{\boldsymbol{y}} + \dfrac{\lambda}{n-\lambda}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}) & (j \neq i) \end{cases}. \tag{A15}$$

It follows from (A15) that $\boldsymbol{M}_i(\hat{\boldsymbol{\mu}}_i(\lambda), \lambda) = \boldsymbol{S}_i(\lambda)$, where $\boldsymbol{S}_i(\lambda)$ is given by (27). Equation (22) implies

$$\mathrm{CV}(\lambda) = np\log(2\pi) + \sum_{i=1}^{n}\left\{\log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}}_i(\lambda))| + (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i(\lambda))'\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}}_i(\lambda))^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i(\lambda))\right\},$$

where $\hat{\boldsymbol{\xi}}_i(\lambda)$ is given by (26). Substituting (A15) into the above equation yields (25).

A.7. Derivation of EIC in Corollary 2

Notice that $\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)$ is the minimizer of $\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d}_b)$ and $\mathbf{1}_n'\boldsymbol{d}_b = n$. With the discrepancy function given by (22), by direct calculations we obtain

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d}_b) &= -2\boldsymbol{\Sigma}(\boldsymbol{\xi})^{-1}(\boldsymbol{Y}'\boldsymbol{d}_b - n\boldsymbol{\mu}), \\ \frac{\partial}{\partial \boldsymbol{\xi}}\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d}_b) &= \frac{\partial}{\partial \boldsymbol{\xi}}(n-\lambda)F(\boldsymbol{\xi}|\boldsymbol{M}(\boldsymbol{\mu}, \boldsymbol{d}_b)),\end{aligned}$$

where $F(\boldsymbol{\xi}|\boldsymbol{S})$ is given by (24) and

$$\boldsymbol{M}(\boldsymbol{\mu}, \boldsymbol{d}_b) = \frac{1}{n}(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}')'\mathrm{diag}(\boldsymbol{d}_b)(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}').$$

Denote $\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b) = (\hat{\boldsymbol{\mu}}(\boldsymbol{d}_b)', \hat{\boldsymbol{\xi}}(\boldsymbol{d}_b)')'$. Solving the equation $\partial\Psi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d}_b)/\partial\boldsymbol{\mu} = \mathbf{0}_p$ leads to

$$\hat{\boldsymbol{\mu}}(\boldsymbol{d}_b) = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{d}_b. \tag{A16}$$

Substituting (A16) into $\boldsymbol{M}(\boldsymbol{\mu}, \boldsymbol{d}_b)$ yields $\boldsymbol{M}_i(\hat{\boldsymbol{\mu}}(\boldsymbol{d}_b), \boldsymbol{d}_b) = \boldsymbol{S}(\boldsymbol{d}_b)$, where $\boldsymbol{S}(\boldsymbol{d}_b)$ is given by (32). Notice that the EIC under the candidate model M in (21) is

$$\mathrm{EIC} = nF(\hat{\boldsymbol{\xi}}|\boldsymbol{S}) + np\log(2\pi) + \frac{n}{B}\sum_{i=1}^{n} \mathrm{tr}\left\{\boldsymbol{M}(\hat{\boldsymbol{\mu}}(\boldsymbol{d}_b), \mathbf{1}_n - \boldsymbol{d}_b)\boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}}(\boldsymbol{d}_b))^{-1}\right\},$$

where $\hat{\boldsymbol{\xi}}$ is given by (28). Substituting (A16) into $\boldsymbol{M}(\boldsymbol{\mu}, \mathbf{1}_n - \boldsymbol{d}_b)$ yields $\boldsymbol{M}(\hat{\boldsymbol{\mu}}(\boldsymbol{d}_b), \mathbf{1}_n - \boldsymbol{d}_b) = \boldsymbol{V}(\boldsymbol{d}_b)$ in (31), which further leads to (29).

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd. International Symposium on Information Theory* (Eds. B. N. Petrov & F. Csáki), pp. 267–281. Akadémiai Kiadó, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

[3] Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, **52**, 317–332.

[4] Balakrishnan, N., & Ambagaspitiya, R. S. (1994). On skew Laplace distribution. *Technical Report, Department of Mathematics & Statistics, McMaster University*, Hamilton, Ontario, Canada.

[5] Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.

[6] Bentler, P. M. (1995). *EQS Structural Equation Program Manual*. Multivariate Software, Inc, Encino, CA.

[7] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

[8] Fujikoshi, Y., Noguchi, T., Ohtaki, M., & Yanagihara, H. (2003). Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models. *Ann. Inst. Statist. Math.*, **55**, 537–553.

[9] Fujisawa, H., & Eguchi, S. (2006). Robust estimation in the normal mixture model. *J. Statist. Plann. Inference*, **136**, 3989–4011.

[10] Hall, P. (1987). Edgeworth expansion for student's t statistic under minimal moment condition. *Ann. Probab.*, **15**, 920–931.

[11] Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Stat. Math.*, **49**, 411–434.

[12] Konishi, S. (1999). Statistical model evaluation and information criteria. In *Multivariate Analysis, Design of Experiments, and Survey Sampling* (Ed. S. Ghosh), pp. 369–399. Marcel Dekker, New York.

[13] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics*, **22**, 79–86.

[14] Lee, S.-Y., & Kontoghiorghes, E. J. (Eds.). (2007). *Handbook of Latent Variable and Related Models*. Elsevier, Amsterdam.

[15] Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd. edition). Springer-Verlag, New York.

[16] Magnus, J. R., & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (revised edition). John Wiley & Sons, New York.

[17] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

[18] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[19] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser.* **B**, **36**, 111–147.

[20] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser.* **B**, **39**, 44–47.

[21] Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Math. Sci.*, **153**, 12–18 (in Japanese).

[22] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.

[23] Yanagihara, H. (2005). Selection of covariance structure models in nonnormal data by using information criterion: an application to data from the survey of the Japanese national character. *Proc. Inst. Statist. Math.*, **53**, 133–157 (in Japanese).

[24] Yanagihara, H. (2007). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. *J. Multivariate Anal.*, **98**, 1–29.

[25] Yanagihara, H., & Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British J. Math. Statist. Psych.*, **58**, 209–237.

[26] Yanagihara, H., Tonda, T., & Matsumoto, C. (2006). Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition. *J. Multivariate Anal.*, **97**, 1965–1975.

[27] Yanagihara, H., Kamo, K., & Tonda, T. (2006). Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *TR No. 06-01, Statistical Research Group, Hiroshima University*, Hiroshima, Japan.

[28] Yuan, K.-H., & Bentler, P. M. (2007). Structural equation modeling. In *Handbook of Statistics 26: Psychometrics* (Eds. C. R. Rao & S. Sinharay), pp. 297–358. North-Holland, Amsterdam.

TABLE 1. Risk of each candidate model

True Distribution	Model 1	Model 2	Model 3
1	466.7	464.7	461.5
2	468.9	467.1	470.6
3	466.0	463.9	457.6
4	469.1	467.2	470.7
5	471.0	469.2	480.3
6	475.7	474.0	493.8

TABLE 2. Relative biases, RMSEs, and selection frequencies of five information criteria

True Dist.	Criterion	Model 1 Bias	Model 1 (RMSE	Model 1 Freq.)	Model 2 Bias	Model 2 (RMSE	Model 2 Freq.)	Model 3 Bias	Model 3 (RMSE	Model 3 Freq.)
1	AIC	0.52	(4.3	0.0)	1.13	(4.4	43.1)	0.99	(4.2	56.9)
	TIC	-0.51	(4.7	1.6)	0.12	(4.8	35.7)	0.62	(4.4	62.7)
	EIC	0.05	(4.3	11.4)	0.08	(4.4	31.6)	0.16	(4.2	57.0)
	CV	-0.08	(4.3	13.5)	-0.10	(4.4	31.8)	-0.16	(4.2	54.6)
	CCV	0.02	(4.3	12.4)	0.02	(4.4	31.5)	0.05	(4.2	56.1)
2	AIC	1.40	(6.3	0.0)	2.04	(6.6	37.2)	4.21	(7.0	62.8)
	TIC	-2.71	(8.8	1.7)	-2.04	(8.8	26.3)	2.25	(6.6	72.0)
	EIC	0.30	(6.5	19.6)	0.36	(6.7	38.6)	1.08	(6.4	41.7)
	CV	-0.04	(6.7	24.4)	-0.07	(6.9	40.2)	-0.19	(7.0	35.4)
	CCV	0.09	(6.7	22.8)	0.09	(6.9	40.0)	0.22	(6.8	37.2)
3	AIC	0.16	(3.1	0.0)	0.74	(3.2	47.0)	-0.50	(3.2	53.0)
	TIC	0.68	(3.1	1.2)	1.29	(3.3	43.8)	0.11	(3.2	55.1)
	EIC	-0.02	(3.0	4.4)	-0.02	(3.1	22.7)	-0.20	(3.1	72.8)
	CV	-0.08	(3.0	3.7)	-0.11	(3.1	20.9)	-0.15	(3.1	75.3)
	CCV	0.01	(3.0	3.3)	0.00	(3.1	20.7)	-0.01	(3.1	76.0)
4	AIC	1.39	(6.4	0.0)	2.02	(6.6	37.8)	4.16	(6.9	62.2)
	TIC	-2.71	(9.6	1.8)	-2.08	(9.5	26.8)	2.37	(6.8	71.5)
	EIC	0.27	(6.8	19.7)	0.30	(6.9	38.5)	1.02	(6.7	41.9)
	CV	-0.11	(7.1	24.7)	-0.15	(7.3	39.3)	-0.33	(7.6	36.0)
	CCV	0.03	(7.0	22.9)	0.02	(7.2	39.2)	0.12	(7.3	37.8)
5	AIC	2.06	(7.7	0.0)	2.74	(8.0	32.2)	7.12	(9.6	67.8)
	TIC	-4.34	(12.8	1.6)	-3.69	(12.7	19.6)	4.46	(8.8	78.8)
	EIC	0.37	(8.4	24.2)	0.46	(8.5	41.3)	1.73	(8.7	34.5)
	CV	-0.25	(9.1	29.4)	-0.25	(9.4	41.1)	-0.55	(10.9	29.5)
	CCV	-0.07	(9.0	27.2)	-0.04	(9.2	41.2)	0.26	(10.0	31.6)
6	AIC	4.07	(10.5	0.0)	4.75	(10.9	27.1)	11.65	(13.8	72.9)
	TIC	-6.10	(20.0	1.1)	-5.45	(19.9	13.4)	8.05	(12.4	85.5)
	EIC	1.22	(12.9	25.6)	1.29	(13.2	42.1)	3.72	(14.4	32.3)
	CV	-0.21	(16.5	30.6)	-0.26	(16.9	42.5)	-0.77	(21.9	26.9)
	CCV	0.13	(15.5	28.7)	0.13	(15.8	42.3)	1.17	(15.9	29.0)

TABLE 3. Averages of frequencies of choosing the model having the smallest risk

Criterion	AIC	TIC	EIC	CV	CCV
Average of Frequencies (%)	40.7	34.0	48.4	48.8	49.1

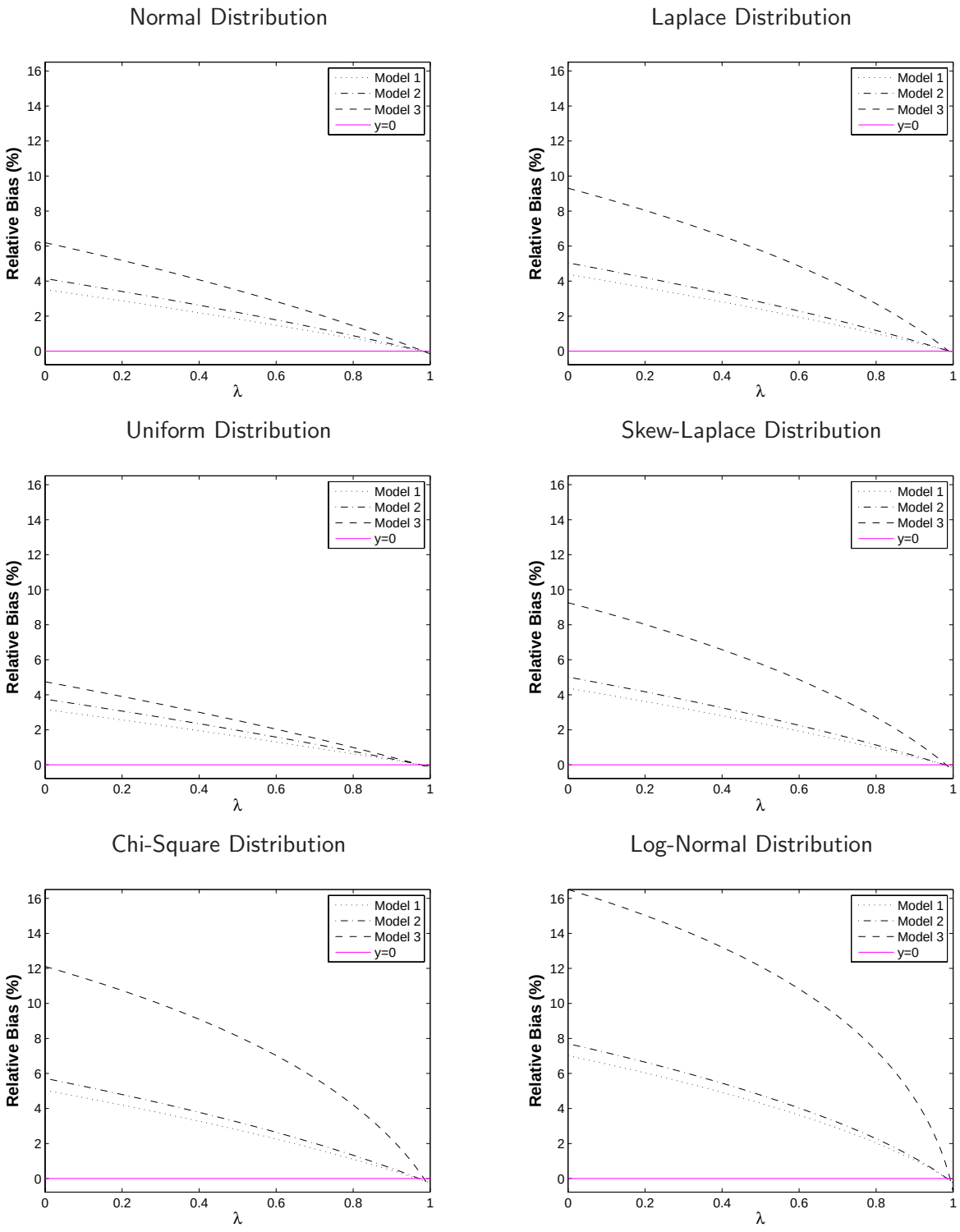

FIGURE 1. Relative biases of cross-validatory model selection criteria

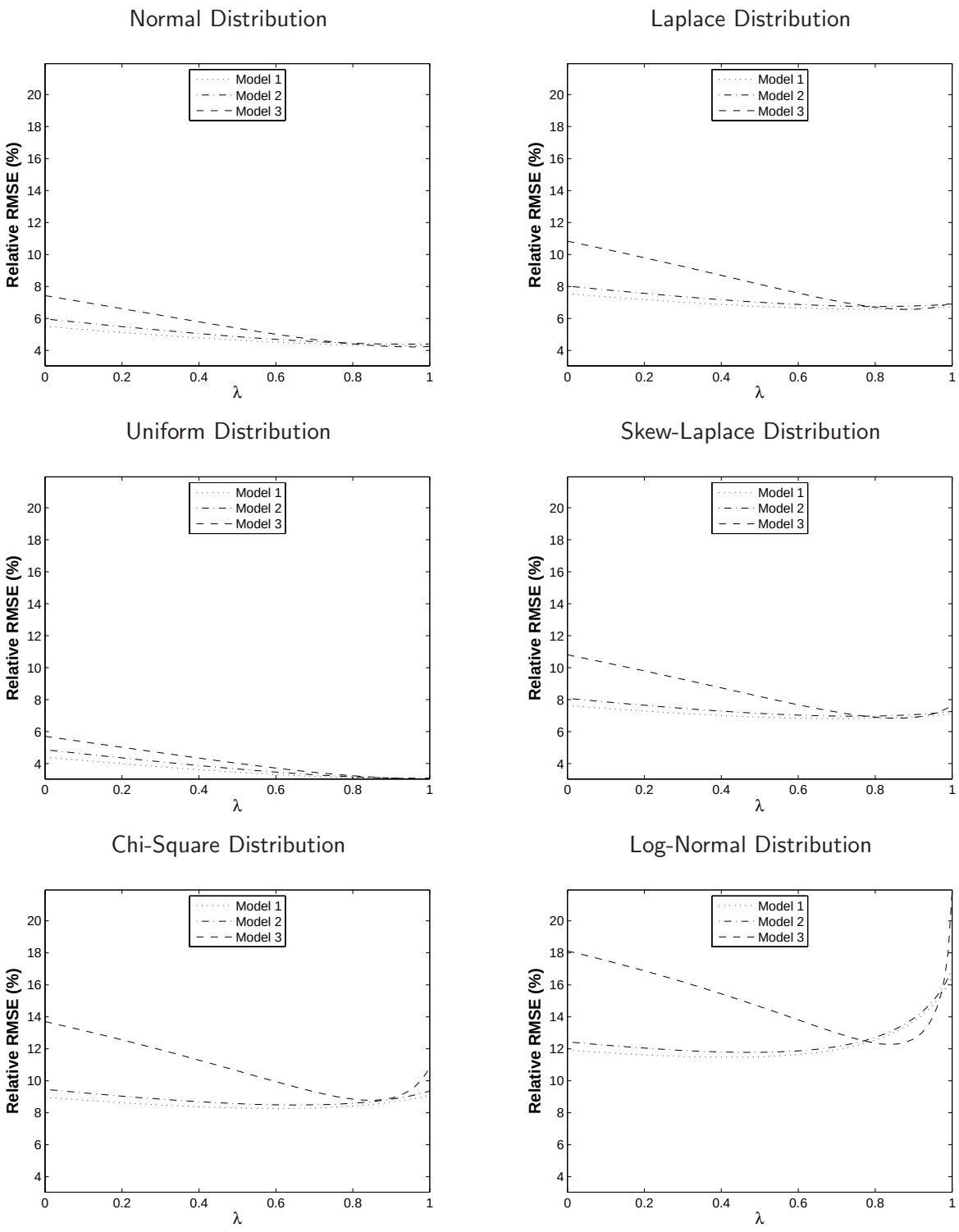

FIGURE 2. Relative RMSEs of cross-validatory model selection criteria

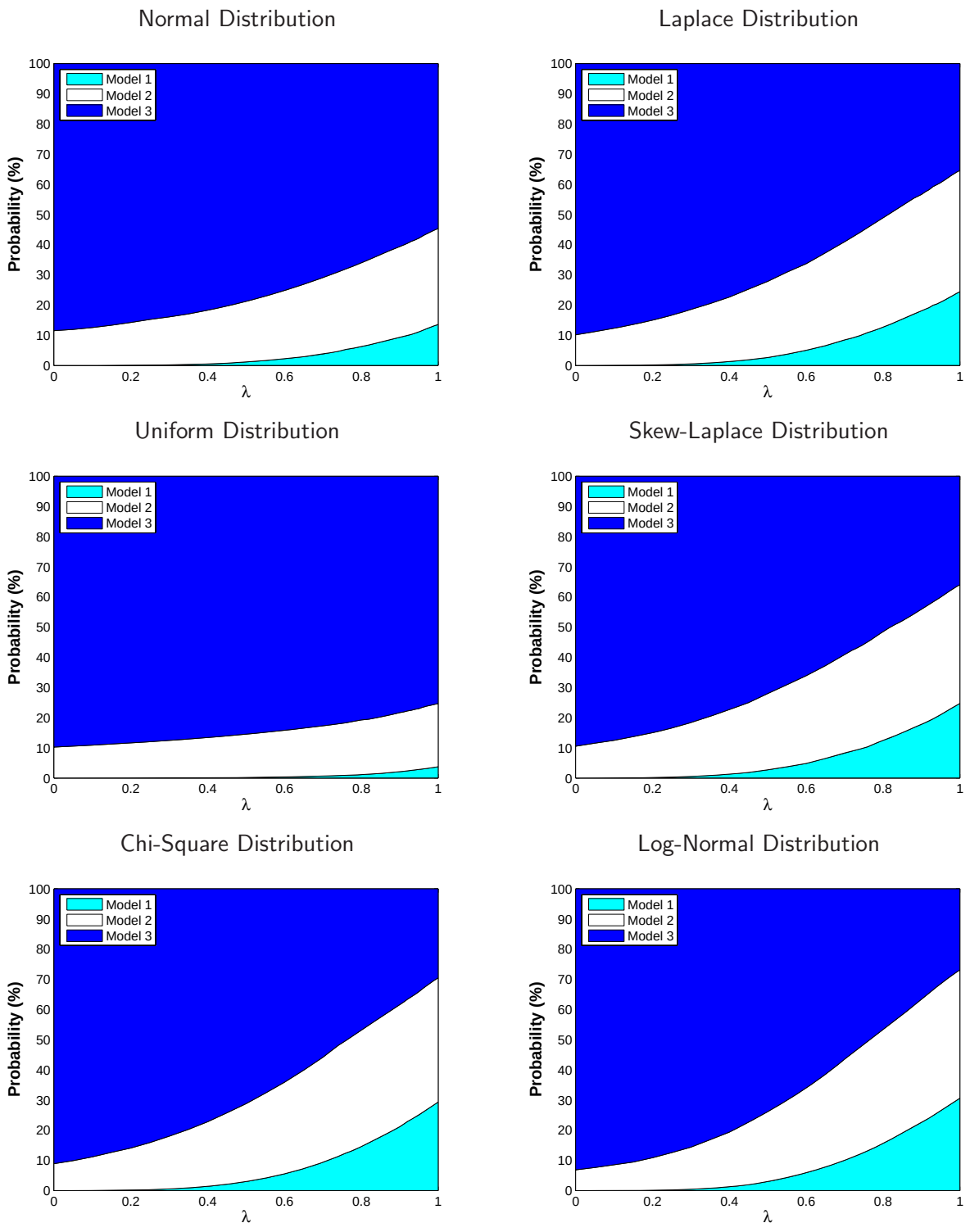

FIGURE 3. Selection frequencies of cross-validatory model selection criteria