Communications for Statistical Applications and Methods
2020, Vol. 27, No. 3, 301–311

https://doi.org/10.29220/CSAM.2020.27.3.301
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Bayesian information criterion accounting for the number of covariance parameters in mixed effects models

Junoh Heo[a], Jung Yeon Lee[b], Wonkuk Kim[1,c]

[a]Department of Statistics, Chung-Ang University, Korea;
[b]Department of Psychiatry, New York University School of Medicine, USA;
[c]Department of Applied Statistics, Chung-Ang University, Korea

**Abstract**

Schwarz's Bayesian information criterion (BIC) is one of the most popular criteria for model selection, that was derived under the assumption of independent and identical distribution. For correlated data in longitudinal studies, Jones (*Statistics in Medicine*, **30**, 3050–3056, 2011) modified the BIC to select the best linear mixed effects model based on the effective sample size where the number of parameters in covariance structure was not considered. In this paper, we propose an extended Jones' modified BIC by considering covariance parameters. We conducted simulation studies under a variety of parameter configurations for linear mixed effects models. Our simulation study indicates that our proposed BIC performs better in model selection than Schwarz's BIC and Jones' modified BIC do in most scenarios. We also illustrate an example of smoking data using a longitudinal cohort of cancer patients.

Keywords: correlated data, effective sample size, Fisher information matrix, longitudinal study, model selection

## 1. Introduction

There are a number of information-based model selection methods such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) to choose the best model in a candidate model space. In 1973, Akaike first developed the AIC by adjusting biased empirical information (Akaike, 1973):

$$\text{AIC} = -2\text{log-likelihood} + 2k,$$

where $k$ is the number of parameters in a candidate model. Schwarz (1978) derived the BIC under Bayesian framework. BIC uses a greater penalty term $k \log n$ as compared to $2k$ in AIC, where $n$ is the number of observations.

$$\text{BIC} = -2\text{log-likelihood} + k \log n,$$

It is known that BIC is asymptotically consistent in many model selection scenarios, as opposed to AIC that is inconsistent (Nishii, 1984; Gassiat, 2002). BIC tends to select a less complex model compared to AIC. BIC behaves different from AIC, since the penalty term of BIC is smaller than

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-Ro, Dongjak-Gu, Seoul 06974, Korea. E-mail: wkim@cau.ac.kr

Published 31 May 2020 / journal homepage: http://csam.or.kr

that of AIC when the sample size is more than 8. In practice, AIC is used to find the best prediction model, whereas BIC is applied to choose the best model for further inferences. There is a wide range of applications of BIC including K-means clustering inverse regression (Ahn and Yoo, 2011), multiple change-points (Kim and Cheon, 2013), dynamic conditional correlation model (Kim, 2014), and growth mixture model (Lee *et al.*, 2019).

Many researchers have worked on improving the information criteria. Hurvich and Tsai (1989) suggested the corrected AIC (AICc) for regression and autoregressive time series models, particularly when the sample size is small or the sample size is inadequate to fit the number of parameters.

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1},$$

that is asymptotically efficient when the true model is infinite dimensional. Hannan and Quinn (1979) proposed Hannan and Quinn information criterion (HQIC) to determine the order of an autoregression model, in which a double logarithm of the sample size is used.

$$\text{HQIC} = -2\text{log-likelihood} + 2k\log(\log n).$$

HQIC is not asymptotically efficient nor is it an estimator of Kullback-Leibler divergence. Burnham and Anderson (1998) pointed out that HQIC, "while often cited, seems to have seen little use in practice." Bozdogan (1987) proposed another asymptotically consistent information criterion, that is called the consistent AIC (CAIC).

$$\text{CAIC} = -2\text{log-likelihood} + k(\log n + 1).$$

Information criteria can often be viewed as a special case of generalized information criterion (GIC) defined by

$$\text{GIC} = -2\text{log-likelihood} + ka_n,$$

where $a_n$ is a positive sequence that depends only on the sample size $n$ and that controls the penalty on model complexity.

Many statistical software programs offer a variety of information criteria. PROC MIXED (Institute Inc, 2008) of SAS produces AIC, AICc, HQIC, CAIC, and Schwarz's BIC. SPSS also offers AIC, AICc, CAIC, and BIC in MIXED procedure. Table 1 summarizes the information criteria that are available in SAS, and Table 2 shows the information criteria that are available in SPSS.

There are several extensions of information criteria to select the best linear mixed effects model when samples are correlated. Vaida and Blanchard (2005) derived their conditional AIC (cAIC) based on the concept of the effective number of parameters $\rho$ (Hodges and Sargent, 2001) in a linear mixed model. The effective number of parameters $\rho$ measures a level of complexity between a fixed effects model with no cluster effect and a corresponding model with fixed cluster effects.

$$\text{cAIC} = -2\text{log-likelihood} + \frac{2(n-p-1)}{n-p-2}\left(\rho\left(\hat{\tau}_R\right) + 1 + \frac{p+1}{n-p-1}\right),$$

where $p$ is the number of parameters for fixed effects, and $\rho(\hat{\tau}_R)$ is the residual maximum likelihood (REML) estimator of $\rho(\tau) = \text{tr}((X'V^{-1}X)^{-1}X'V^{-1}R_iV^{-1}X) + n - \text{tr}(R_iV^{-1})$.

Jennrich and Shluchter (1986) and Diggle (1988) discussed a limitation on applying Schwarz's BIC when choosing an optimal linear mixed effects model. In a space of linear mixed effects models with a fixed covariance structure, Jones (2011) proposed a modified BIC by using the effective

Table 1: Information criteria options in SAS

| Criterion | Formula | Reference |
|---|---|---|
| AIC | $-2\log\text{-likelihood} + 2k$ | Akaike (1973, 1974) |
| AICc | $-2\log\text{-likelihood} + \frac{2nk}{n-k-1}$ | Hurvich and Tsai (1989), Burnham and Anderson (1998) |
| BIC | $-2\log\text{-likelihood} + k\log n$ | Schwarz (1978) |
| CAIC | $-2\log\text{-likelihood} + k(\log n + 1)$ | Bozdogan (1987) |
| HQIC | $-2\log\text{-likelihood} + 2k\log(\log n)$ | Hannan and Quinn (1979) |

AIC = Akaike information criterion; AICc = corrected AIC; BIC = Bayesian information criterion; CAIC = consistent AIC; HQIC = Hannan and Quinn information criterion.

Table 2: Information criteria options in SPSS

| Criterion | Formula | Reference |
|---|---|---|
| AIC | $-2\log\text{-likelihood} + 2k$ | Akaike (1973, 1974) |
| AICc | $-2\log\text{-likelihood} + \frac{2nk}{n-k-1}$ | Hurvich and Tsai (1989), Burnham and Anderson (1998) |
| BIC | $-2\log\text{-likelihood} + k\log n$ | Schwarz (1978) |
| CAIC | $-2\log\text{-likelihood} + k(\log n + 1)$ | Bozdogan (1987) |

AIC = Akaike information criterion; AICc = corrected AIC; BIC = Bayesian information criterion; CAIC = consistent AIC.

sample size. He defined the effective sample size based on the fixed correlation structure and Fisher's information matrix. In our work, we consider a model selection based on BIC when the correlation structure is not fixed so that the number of covariance parameters in linear mixed effects models is not a constant.

The remainder of this paper is organized as follows: Section 2 includes a brief review of the derivation of Schwarz's BIC and Jone's BIC, and presents our derivation of the proposed BIC. Section 3 presents simulation studies and illustrates a real data example. Finally, conclusions and discussion are presented in Section 4.

## 2. Methods

### 2.1. Review of Schwarz's BIC

Suppose $M_1, \ldots, M_k$ are $k$ candidate models, and assume that $j^{th}$ model $M_j$ has a density $p(y;\theta_j)$.

$$M_j = \left\{p(y;\theta_j) : \theta_j \in \Theta_j\right\}, \quad j = 1, \ldots, k.$$

Let $y_1, \ldots, y_n$ be a data set from a density $f$. Let $p_j$ denote the probability that the $j^{th}$ model is true. Suppose the parameter $\theta_j$ has a prior density $\pi_j(\theta_j)$. By Bayes' theorem,

$$p\left(M_j|y_1, \ldots, y_n\right) \propto p\left(y_1, \ldots, y_n|M_j\right) p_j,$$

and

$$p\left(y_1, \ldots, y_n|M_j\right) = \int p\left(y_1, \ldots, y_n|M_j, \theta_j\right)\pi_j\left(\theta_j\right)\, d\theta_j = \int L\left(\theta_j\right)\pi_j\left(\theta_j\right)\, d\theta_j.$$

We choose the model $M_j$ by maximizing $p(M_j|y_1, \ldots, y_n)$. It is equivalent to maximize $\log \int L(\theta_j)$ $\pi_j(\theta_j)d\theta_j + \log p_j$. By Taylor's series expansion at $\theta_j = \hat{\theta}_j$ that is the mode of the density, we have:

$$\log L\left(\theta_j\right)\pi_j\left(\theta_j\right) \approx \log L\left(\hat{\theta}_j\right)\pi_j\left(\hat{\theta}_j\right) - \frac{1}{2}\left(\theta_j - \hat{\theta}_j\right)'\left(-\nabla^2 \log L\left(\hat{\theta}_j\right)\pi_j\left(\hat{\theta}_j\right)\right)\left(\theta_j - \hat{\theta}_j\right). \tag{2.1}$$

By assuming negative definiteness of the Hessian matrix, that is, $-\nabla^2 \log L(\hat{\theta}_j)\pi_j(\hat{\theta}_j) = A > 0$ in Equation (2.1), we have the Laplace approximation:

$$\begin{aligned} p\left(y_1,\ldots,y_n|M_j\right) &= \int p\left(y_1,\ldots,y_n|M_j,\theta_j\right)\pi_j\left(\theta_j\right)\,d\theta_j \\ &\approx \int L\left(\hat{\theta}_j\right)\pi_j\left(\hat{\theta}_j\right)\exp\left(-\frac{1}{2}\left(\theta_j-\hat{\theta}_j\right)'A\left(\theta_j-\hat{\theta}_j\right)\right)d\theta_j \\ &\approx L\left(\hat{\theta}_j\right)\pi_j\left(\hat{\theta}_j\right)(2\pi)^{\frac{d_j}{2}}|A|^{-\frac{1}{2}} \end{aligned}$$

that implies

$$\log p\left(y_1,\ldots,y_n|M_j\right) \approx \log L\left(\hat{\theta}_j\right) + \log \pi_j\left(\hat{\theta}_j\right) + \frac{d_j}{2}\log(2\pi) - \frac{1}{2}\log|A|.$$

By using $\log|A| \approx \log|nA_0| = d_j \log n + \log|A_0|$,

$$\log p\left(y_1,\ldots,y_n|M_j\right) \approx \log L\left(\hat{\theta}_j\right) + \log \pi_j\left(\hat{\theta}_j\right) + \frac{d_j}{2}\log(2\pi) - \frac{d_j}{2}\log n - \frac{1}{2}\log|A_0|,$$

and by dropping terms that do not depend on the sample size, we asymptotically have

$$\log p\left(y_1,\ldots,y_n|M_j\right) \approx \log L\left(\hat{\theta}_j\right) - \frac{d_j}{2}\log n.$$

Maximizing $p(M_j|y_1,\ldots,y_n)$ is equivalent to maximizing $\log p(y_1,\ldots,y_n|M_j) + \log p_j \approx \log L(\hat{\theta}_j) - (d_j/2)\log n + \log p_j$. Since $p_j$ does not depend on $n$, we have only $\log L(\hat{\theta}_j) - (d_j/2)\log n$. Schwarz's BIC is defined as

$$\text{BIC} = -2\log L\left(\hat{\theta}_j\right) + d_j \log n.$$

## 2.2. Review of Jones' BIC based on the effective sample size

A linear mixed model (Laird and Ware, 1982) with Gaussian error for subject $i$ is written as

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \quad \text{for } i = 1,\ldots,m,$$

where $y_i$ is an $n_i \times 1$ column vector of response variables for subject or cluster $i$, $X_i$ is an $n_i \times p$ matrix of observed independent variables, $\beta$ is a $p \times 1$ vector of the regression coefficients of fixed effects, $Z_i$ is an $n_i \times q$ matrix of random effects, $\gamma_i$ is a $q \times 1$ vector which are assumed to be independently distributed across subjects with multivariate normal distribution $\gamma_i \sim N_q(0, G)$, and $\epsilon_i \sim N_{n_i}(0, R_i)$ is an $n_i \times 1$ independent random error vector. Since $E(y_i) = X_i\beta$ and $\text{Cov}(y_i) = V_i = Z_iGZ_i' + R_i$, the marginal density is given by

$$f(y_i) = (2\pi)^{-\frac{n_i}{2}}\left|Z_iGZ_i' + R_i\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(y_i - X_i\beta\right)'\left(Z_iGZ_i' + R_i\right)^{-1}\left(y_i - X_i\beta\right)\right). \tag{2.2}$$

In a linear mixed effects model, Fisher's information matrix for the fixed effects part can be calculated by $X_i$ and $V_i$:

$$-E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\log f(y_i)\right] = X_i'V_i^{-1}X_i = \begin{pmatrix} 1'V_i^{-1}1 & 1'V_i^{-1}X_i^* \\ \left(X_i^*\right)'V_i^{-1}1 & \left(X_i^*\right)'V_i^{-1}X_i^* \end{pmatrix},$$

where 1 is an $n_i$ dimensional vector with all elements one, and $X_i = (1 : X_i^*)$. The subject $i^{th}$ information for the intercept term of the fixed effects is equal to $1'V_i^{-1}1$ (McCullagh and Nelder, 1989). Jones defined the effective sample size by $n_e = \sum_{i=1}^{m} 1'C_i^{-1}1$, where $C_i$ is the correlation matrix obtained from the covariance matrix $V_i$. He proposed to use BIC by replacing the sample size $n$ by the effective sample size $n_e$ when selecting the best linear mixed effects model.

$$\mathrm{BIC}_{\mathrm{Jones}} = -2\text{log-likelihood} + k \log n_e, \tag{2.3}$$

where $k$ is the number of estimated parameters.

## 2.3. A modified BIC adjusting the number of parameters in covariance matrix

In this subsection, we derive our proposed BIC to handle varying number of parameters in the covariance matrix of linear mixed effects models. As seen in the above subsection, the information for the parameters of covariance structure has not been accounted during the derivation of Jones' BIC. From Equation (2.2), the log-likelihood function is written as

$$l(\beta, V_i) = \log f(y_i) = -\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log|V_i| - \frac{1}{2}\mathrm{tr}\left(V_i^{-1}(y_i - X_i\beta)(y_i - X_i\beta)'\right).$$

Suppose the variance-covariance is a function of $l$ parameters, $\varphi = (\varphi_1, \ldots, \varphi_l)$ so that

$$V_i = Z_i G Z_i' + R_i = V_i(\varphi) = V_i(\varphi_1, \ldots, \varphi_l),$$

where $\varphi$ does not depend on the parameter vector of the fixed effects $\beta$. The first derivatives of the log-likelihood $l(\beta, V_i)$ with respect to the parameter vectors are given by

$$\begin{aligned}
&\frac{\partial}{\partial\beta} l(\beta, V_i) = X_i' V_i^{-1}(y_i - X_i\beta),\\
&\frac{\partial}{\partial\varphi_k} l(\beta, V_i) = -\frac{1}{2}\left(\mathrm{tr}\left(V_i^{-1}\frac{\partial V_i}{\partial\varphi_k}\right) - (y_i - X_i\beta)' V_i^{-1}\frac{\partial V_i}{\partial\varphi_k}V_i^{-1}(y_i - X_i\beta)\right).
\end{aligned}$$

The second derivatives are given by

$$\begin{aligned}
-E\left[\frac{\partial^2}{\partial\beta\partial\beta'} l(\beta, V_i)\right] &= X_i' V_i^{-1} X_i,\\
-E\left[\frac{\partial^2}{\partial\beta\partial\varphi_k} l(\beta, V_i)\right] &= -E\left[\frac{\partial^2}{\partial\varphi_k\partial\beta'} l(\beta, V_i)\right] = 0, \quad \text{for all } k,\\
-E\left[\frac{\partial^2}{\partial\varphi_k\partial\varphi_j} l(\beta, V_i)\right] &= \frac{1}{2}\mathrm{tr}\left(V_i^{-1}\frac{\partial V_i}{\partial\varphi_j}V_i^{-1}\frac{\partial V_i}{\partial\varphi_k}\right).
\end{aligned}$$

The Fisher information matrix can be written as

$$I(\beta, V_i) = \begin{pmatrix} X_i' V_i^{-1} X_i & 0 \\ 0 & I_{i\varphi\varphi} \end{pmatrix}, \quad i = 1, \ldots, m,$$

where $I_{i\varphi\varphi} = ((1/2)\mathrm{tr}(V_i^{-1}(\partial V_i/\partial\varphi_j)V_i^{-1}(\partial V_i/\partial\varphi_k)_{1\le j,k\le l}$ is the $l \times l$ submatrix of the subject $i^{th}$ information matrix for the covariance parameters $\varphi$. When selecting the best model among $k$ linear

mixed effects models $\{M_1, \ldots, M_k\}$, we have $p(y_1, \ldots, y_n|M_j) \approx \sum_{i=1}^{m} \log f(y_i) - (1/2)\log|I|$, where $I = \sum_{i=1}^{m} I(\beta, V_i)$ is sum of fisher information matrix of $y_1, \ldots, y_m$. We can decompose

$$\log|I| = \log \sum_{i=1}^{m} \left|X_i' V_i^{-1} X_i\right| + \log \sum_{i=1}^{m} \left|I_{i\varphi\varphi}\right|. \tag{2.4}$$

We can approximate the first term by Jones' effective sample size $n_e$. By the law of large numbers, we may approximate the second term as

$$\log \sum_{i=1}^{m} \left|I_{i\varphi\varphi}\right| \approx \log \left|m \times \mu_{I_{\varphi\varphi}}\right| = l \log m + \log \mu_{I_{\varphi\varphi}},$$

where $\mu_{I_{\varphi\varphi}}$ is the expected value of $I_{i\varphi\varphi}$. Therefore, our proposed BIC, denoted by $\text{BIC}_{lme}$, is written as

$$\text{BIC}_{lme} = -2\text{log-likelihood} + (p \log n_e + l \log m),$$

where $p$ is the number of parameters for fixed effects, $n_e$ is the effective sample size obtained in Jones' BIC, $l$ is the number of parameters for covariance structure, and $m$ is the number of subjects or items in the study.

## 3. Results

In this section, we present our simulation studies and a real data example.

### 3.1. Simulation study

We conducted simulation studies based on 1,000 replicates of samples per each configuration to compare the performance of our proposed BIC to Schwarz's BIC and Jones' BIC. During our simulation work, we fixed the covariance matrix of random error for which each subject is in a form of $R_i = \sigma^2 I_{n_i \times n_i}$ while the covariance matrix $G$ of random effects is one of the three possible scenarios: diagonal, symmetric, or compound symmetric. We considered three possible models for fixed effects by setting the degree of a polynomial for the single covariate "time" from constant to the second degree. In addition, we considered low or high correlation settings given other parameter settings as follow.

For the low correlation configuration with diagonal $G$, we set $\sigma$ and $G$ so that the correlation of $y_i$ becomes

$$\text{corr}(y_i) = \begin{pmatrix} 1 & 0.05 & 0.04 & 0.04 \\ 0.05 & 1 & 0.10 & 0.12 \\ 0.04 & 0.10 & 1 & 0.23 \\ 0.04 & 0.12 & 0.23 & 1 \end{pmatrix},$$

where maximum possible correlation is $\rho = 0.23$. For the high correlation setting, we chose $\sigma$ and $G$ so that $y_i$ has the correlation matrix as

$$\text{corr}(y_i) = \begin{pmatrix} 1 & 0.88 & 0.83 & 0.75 \\ 0.88 & 1 & 0.89 & 0.83 \\ 0.83 & 0.89 & 1 & 0.90 \\ 0.75 & 0.83 & 0.90 & 1 \end{pmatrix},$$

Table 3: Simulation results for the number of subjects $m = 100$

| Correlation | Structure of true model | BIC | $\text{BIC}_{\text{Jones}}$ | $\text{BIC}_{lme}$ |
|---|---|---|---|---|
| Low | Diagonal with Constant | 0.197 | 0.188 | 0.298 |
| | Symmetry with Constant | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Constant | 0.983 | 0.981 | 0.976 |
| | Diagonal with Time | 0.180 | 0.165 | 0.270 |
| | Symmetry with Time | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Time | 0.142 | 0.143 | 0.145 |
| | Diagonal with Time and $\text{Time}^2$ | 0.197 | 0.182 | 0.262 |
| | Symmetry with Time and $\text{Time}^2$ | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Time and $\text{Time}^2$ | 0.998 | 0.996 | 0.996 |
| High | Diagonal with Constant | 0.979 | 0.962 | 0.962 |
| | Symmetry with Constant | 0.022 | 0.070 | 0.073 |
| | Compound symmetry with Constant | 0.979 | 0.963 | 0.963 |
| | Diagonal with Time | 0.982 | 0.966 | 0.967 |
| | Symmetry with Time | 0.019 | 0.061 | 0.066 |
| | Compound symmetry with Time | 0.893 | 0.889 | 0.891 |
| | Diagonal with Time and $\text{Time}^2$ | 1.000 | 0.999 | 0.999 |
| | Symmetry with Time and $\text{Time}^2$ | 0.015 | 0.077 | 0.081 |
| | Compound symmetry with Time and $\text{Time}^2$ | 1.000 | 1.000 | 1.000 |

The last three columns show the empirical rates of correct selection for the three BICs based on 1,000 replicates per each configuration. BIC, $\text{BIC}_{\text{Jones}}$, and $\text{BIC}_{lme}$ denote Schwarz' BIC, Jones' BIC, and the proposed BIC, respectively. BIC = Bayesian information criterion.

where the minimum correlation is 0.75. Similarly, the correlation matrix for low correlation setting with symmetric $G$, where the maximum correlation is 0.25, is given by

$$\text{corr}(y_i) = \begin{pmatrix} 1 & 0.11 & 0.13 & 0.15 \\ 0.11 & 1 & 0.17 & 0.20 \\ 0.13 & 0.17 & 1 & 0.25 \\ 0.15 & 0.20 & 0.25 & 1 \end{pmatrix}.$$

In the high correlation setting with symmetric $G$, where the minimum correlation is 0.79, the correlation matrix of $y_i$ is given by

$$\text{corr}(y_i) = \begin{pmatrix} 1 & 0.82 & 0.81 & 0.79 \\ 0.82 & 1 & 0.89 & 0.88 \\ 0.81 & 0.89 & 1 & 0.93 \\ 0.79 & 0.88 & 0.93 & 1 \end{pmatrix}.$$

For simulation studies with compound symmetric $G$, the low- and high-correlation settings are as follow: the correlation matrix of $y_i$ (maximum correlation is 0.022) for low-correlation setting is

$$\text{corr}(y_i) = \begin{pmatrix} 1 & 0.007 & 0.009 & 0.010 \\ 0.007 & 1 & 0.012 & 0.015 \\ 0.009 & 0.012 & 1 & 0.022 \\ 0.010 & 0.015 & 0.022 & 1 \end{pmatrix},$$

and the correlation matrix of $y_i$ (minimum correlation is 0.82) for high-correlation setting is

$$\text{corr}(y_i) = \begin{pmatrix} 1 & 0.96 & 0.89 & 0.82 \\ 0.96 & 1 & 0.98 & 0.93 \\ 0.89 & 0.98 & 1 & 0.99 \\ 0.82 & 0.93 & 0.99 & 1 \end{pmatrix}.$$

Table 4: Simulation results for the number of subjects $m = 500$

| Correlation | Structure of true model | BIC | $BIC_{Jones}$ | $BIC_{lme}$ |
|---|---|---|---|---|
| Low | Diagonal with Constant | 0.515 | 0.513 | 0.586 |
| | Symmetry with Constant | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Constant | 0.986 | 0.985 | 0.983 |
| | Diagonal with Time | 0.524 | 0.518 | 0.587 |
| | Symmetry with Time | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Time | 0.600 | 0.597 | 0.599 |
| | Diagonal with Time and $Time^2$ | 0.527 | 0.515 | 0.595 |
| | Symmetry with Time and $Time^2$ | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Time and $Time^2$ | 0.994 | 0.992 | 0.987 |
| High | Diagonal with Constant | 0.997 | 0.989 | 0.989 |
| | Symmetry with Constant | 0.676 | 0.818 | 0.825 |
| | Compound symmetry with Constant | 0.996 | 0.985 | 0.985 |
| | Diagonal with Time | 0.993 | 0.989 | 0.989 |
| | Symmetry with Time | 0.667 | 0.820 | 0.833 |
| | Compound symmetry with Time | 0.994 | 0.988 | 0.988 |
| | Diagonal with Time and $Time^2$ | 1.000 | 1.000 | 1.000 |
| | Symmetry with Time and $Time^2$ | 0.685 | 0.832 | 0.841 |
| | Compound symmetry with Time and $Time^2$ | 1.000 | 1.000 | 1.000 |

The last three columns show the empirical rates of correct selection for the three BICs based on 1,000 replicates per each configuration. BIC, $BIC_{Jones}$, and $BIC_{lme}$ denote Schwarz' BIC, Jones' BIC, and the proposed BIC, respectively. BIC = Bayesian information criterion.

Table 5: Simulation results for the number of subjects $m = 1{,}000$

| Correlation | Structure of true model | BIC | $BIC_{Jones}$ | $BIC_{lme}$ |
|---|---|---|---|---|
| Low | Diagonal with Constant | 0.739 | 0.739 | 0.785 |
| | Symmetry with Constant | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Constant | 0.988 | 0.987 | 0.985 |
| | Diagonal with Time | 0.735 | 0.729 | 0.776 |
| | Symmetry with Time | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Time | 0.893 | 0.889 | 0.891 |
| | Diagonal with Time and $Time^2$ | 0.765 | 0.757 | 0.821 |
| | Symmetry with Time and $Time^2$ | 0.000 | 0.000 | 0.000 |
| | Compound symmetry with Time and $Time^2$ | 0.997 | 0.995 | 0.990 |
| High | Diagonal with Constant | 0.981 | 0.990 | 0.990 |
| | Symmetry with Constant | 0.989 | 0.986 | 0.986 |
| | Compound symmetry with Constant | 0.956 | 0.968 | 0.973 |
| | Diagonal with Time | 1.000 | 0.994 | 0.998 |
| | Symmetry with Time | 0.992 | 0.983 | 0.991 |
| | Compound symmetry with Time | 0.966 | 0.960 | 0.979 |
| | Diagonal with Time and $Time^2$ | 1.000 | 1.000 | 1.000 |
| | Symmetry with Time and $Time^2$ | 0.998 | 0.994 | 0.994 |
| | Compound symmetry with Time and $Time^2$ | 1.000 | 1.000 | 1.000 |

The last three columns show the empirical rates of correct selection for the three BICs based on 1,000 replicates per each configuration. BIC, $BIC_{Jones}$, and $BIC_{lme}$ denote Schwarz' BIC, Jones' BIC, and the proposed BIC, respectively. BIC = Bayesian information criterion.

Tables 3, 4, and 5 summarize the results of simulation studies. The results show that the proposed BIC performs better than or similar to Schwarz's BIC and Jones' BIC in most scenarios of our simulation settings. In particular, the proposed BIC shows the best performance for high correlation with symmetric covariance. When the true covariance structure is compound symmetric, Schwarz's BIC shows better performances compared to Jones' BIC and our proposed BIC whereas it does not perform well for symmetric covariance settings. Schwarz's BIC tends to choose a model with compound

Table 6: The scores of Schwarz's BIC, Jones' BIC, and the proposed BIC for three non-nested linear mixed effects models

| Model Structure | No. parameters | Schwarz's BIC | Jones' BIC | Proposed BIC |
|---|---|---|---|---|
| M1 | 9 (2, 7) | 3055.84 (2) | 3051.61 (1) | 3045.19 (1) |
| M2 | 7 (3, 4) | 3057.63 (3) | 3052.98 (2) | 3050.09 (3) |
| M3 | 9 (4, 5) | 3044.08 (1) | 3056.41 (3) | 3049.48 (2) |

The numbers in parenthesis at the second column are (no. parameters of fixed effects, no. parameters of covariance structure). The number in parenthesis at the last three columns is its rank of each BIC. BIC = Bayesian information criterion.

symmetric covariance, that has less number of parameters than the models with symmetric covariance matrices. None of the three BICs works properly when the correlation setting is low and the covariance structure is symmetric. It appears that our proposed BIC and Jones' BIC are less conservative than Schwarz's BIC.

### 3.2. Example with real data

Caroll *et al.* (2019) examined the bidirectional longitudinal associations between smoking and its affect among cancer patients using varenicline to quit smoking. The study participants were 119 cancer patients at 4 different time waves; week 0 (pre-quit), week 1 (target quit week), week 4, and week 12. Patients were smokers with a diagnosis of cancer and were recruited for a 24-week trial of extended duration varenicline plus behavioral counseling (Schnoll, 2018). Smoking was assessed via self-reported number of cigarettes in the past 24 hours. The negative affect was assessed using the Positive and Negative Affect Scale (PANAS) which is the measure for positive and negative affect in Psychology. The variables used in the study were the number of cigarettes per day, panas pos time, panas neg time, age, gender, ftnd score which is a score of Fagerstrom test for nicotine dependence, race, education status, marital status, employment status, and income status. To illustrate how differently Schwarz's BIC, Jones' BIC and the proposed BIC select models, we considered three models including gender as a fixed effect and other configurations are: (M1) symmetric covariance matrix for random components with intercept fixed effect, (M2) diagonal covariance matrix for random components with linear fixed effects in time, and (M3) compound symmetric covariance matrix for random components with quadratic fixed effects in time. Table 6 summarizes the BIC scores of the three approaches.

Although the models M1 and M3 have the same total number of parameters, Schwarz's BIC has the smallest BIC score for M3 which is the model with the compound symmetric covariance for random components and the fixed effects of gender and the $2^{nd}$ degree of polynomial in time. However, Jones' BIC and the proposed BIC have the smallest BIC scores for M1 which is the model with symmetric covariance for random components and gender fixed effects. The proposed BIC has the second smallest BIC for M3, but Jones' BIC has the second smallest score for M2.

## 4. Conclusions and discussion

Bayesian Information Criterion plays an important role in model selection. Although BIC is one of the most useful criteria in model selection tasks, the performance of Schwarz's BIC may be poor when the observations are correlated, such as those in longitudinal studies. Jones proposed a modified BIC based on the effective sample size for longitudinal and clustered data under a fixed type of covariance structure. However, it is limited to same covariance structures with equal number of parameters in the covariance structure. In this paper, we extended Jones' BIC that can be applied to linear mixed effects

models when the number of parameters in the covariance matrix may differ.

Our simulation studies present that the proposed BIC outperformed Schwarz's BIC and Jones' BIC in most scenarios, particularly when correlations are high and sample sizes are relatively small or moderate, that is, $n$ =100 or 500. For example, when the covariance matrix of random effects is symmetric with high correlation, the fixed effect is linear in a single covariate, and the sample size is moderate ($n$ = 500), our proposed BIC selects true models with 83.3% while Schwarz's BIC and Jones' BIC select true models with 66.7% and 82%, respectively. Schwarz's BIC sometimes performs better than the proposed BIC, but it is only about 2%. Another interesting finding in our simulation studies is that none of the three BICs performs properly in selecting a true model when the covariance matrix is symmetric with low correlation even though the sample size is large enough (i.e., $n$ = 1,000).

There is an increased use of cross-validation approaches to select an optimal model. However, many researchers in social and behavioral sciences rely on information criteria when they choose the best model in the space of complex models such as growth mixture model, mixture hidden Markov model, and structural equation model. The present study focuses on the model selection of linear mixed effects; however, it can be extended to the model selection of mixture-based models.

## Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07050012) and supported by the Chung-Ang University Graduate Research Scholarship in 2018.

## References

Ahn JH and Yoo JK (2011). A short note on empirical penalty term study of BIC in K-means clustering inverse regression, *Communications for Statistical Applications and Methods*, **18**, 267–275.

Akaike H (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Petrov BN and Csaki F eds, pp. 267–281), Akademia Kiado, Budapest.

Akaike H (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

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

Burnham KP and Anderson DR (1998). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer-Verlag, New York.

Carroll AJ, Kim K, Miele A, Olonoff M, Leone FT, Schnoll RA, and Hitsman B (2019). Longitudinal associations between smoking and affect among cancer patients using varenicline to quit smoking, *Addictive Behaviors*, **95**, 206–210.

Diggle PJ (1988). An approach to the analysis of repeated measurements, *Biometrics*, **44**, 959–971.

Gassiat G (2002). Likelihood ratio inequalities with applications to various mixtures, *Annales De L Institut Henri Poincare-Probabilites Et Statistiques*, **38**, 897–906.

Hannan EJ and Quinn BG (1979). The determination of the order of an autoregression, *Journal of the Royal Statistical Society. Series B*, **41**, 190–195.

Hodges JS and Sargent DJ (2001). Counting degrees of freedom in hierarchical and other richly parameterized models, *Biometrika*, **88**, 367–379.

Hurvich CM and Tsai CL (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297–307.

Jennrich RI and Schluchter MD (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics*, **42**, 805–820.

Jones RH (2011). Bayesian information criterion for longitudinal and clustered data, *Statistics in Medicine*, **30**, 3050–3056.

Kim J and Cheon S (2013). Bayesian multiple change-point estimation and segmentation, *Communications for Statistical Applications and Methods*, **20**, 439–454.

Kim W (2014). Time-varying comovement of KOSPI 200 sector indices returns, *Communications for Statistical Applications and Methods*, **21**, 335–347.

Laird NM and Ware JH (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

Lee JY, Kim W, and Brook JS (2019). Triple comorbid trajectories of alcohol, cigarette, and marijuana use from adolescence to adulthood predict insomnia in adulthood, *Addictive Behaviors*, **90**, 437–443.

McCullagh P and Nelder JA (1989). *Generalized Linear Models* (2nd ed), London and Boca Raton, Florida.

Nishii R (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *The Annals of Statistics*, **12**, 758–765.

Schnoll R (2018). Bidirectional Longitudinal Associations between Smoking and Affect among Cancer Patients Using Varenicline to Quit Smoking, *Inter-University Consortium for Political and Social Research*.

Schwarz G (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Vaida F and Blanchard S (2005). Conditional Akaike information for mixed-effects models, *Biometrika*, **92**, 351–370.

*Received December 3, 2019; Revised December 13, 2019; Accepted December 18, 2019*