Far East Journal of Theoretical Statistics
Volume 25, Number 1, 2008, Pages 51-72
Published Online: May 21, 2008
This paper is available online at http://www.pphmj.com

SELECTION CRITERIA BASED ON MONTE CARLO SIMULATION AND CROSS VALIDATION IN MIXED MODELS

JUNFENG SHANG

Department of Mathematics and Statistics
450 Math Science Building
Bowling Green State University
Bowling Green, OH 43403, U. S. A.
e-mail: jshang@bgnet.bgsu.edu

Abstract

In the mixed modeling framework, Monte Carlo simulation and cross validation are employed to develop an "improved" Akaike information criterion, AICi, and the predictive divergence criterion, PDC, respectively, for model selection. The selection and the estimation performance of the criteria is investigated in a simulation study. Our simulation results demonstrate that PDC outperforms AIC and AICi in choosing an appropriate mixed model as a selection criterion, and AICi is less biased than AIC and PDC in estimating the Kullback-Leibler discrepancy between the true model and a fitted candidate model.

1. Introduction

The Akaike [1, 2] information criterion, AIC, has received widespread recognition as a model selection criterion. AIC provides a method of finding the "best" approximation to the generating model or true model, by providing an asymptotic estimator of the expected Kullback-Leibler discrepancy. This discrepancy is a measure of difference between a fitted candidate model and the true model. AIC is composed of a "goodness of

2000 Mathematics Subject Classification: 62F99.

Keywords and phrases: Akaike information criterion (AIC), improved AIC (AICi), predictive divergence criterion (PDC), Kullback-Leibler discrepancy.

Received October 17, 2007

fit" term and a penalty term. The "goodness of fit" term serves as a biased estimator of the expected Kullback-Leibler discrepancy and reflects how well the model fits the observed data; the penalty term, twice the overall number of estimated parameters, estimates the biased adjustment evaluated by the difference between the expected Kullback-Leibler discrepancy and the expected "goodness of fit" term. Originally justified in large-sample instances, AIC is applicable in a broad array of modeling frameworks. However, in small-sample applications, AIC is likely to choose unsuitably high dimensional models. This limits its effectiveness as a model selection criterion. To overcome this limitation of AIC, the "corrected" AIC, AICc, has been proposed.

AICc (Sugiura [10] and Hurvich and Tsai [6]) is the best known corrected version of AIC. The advantage of using AICc relies on its superior performance to AIC as a selection criterion in small-sample applications. However, because the justification of AICc necessitates the distribution from the candidate model, AICc is less generally applicable than AIC.

An "improved" version of AIC, AICi, has been developed by Hurvich et al. [5] for Gaussian autoregressive model selection. For the estimation of the same discrepancy of AIC, the criterion consists of the same "goodness of fit" term as AIC, yet features a penalty term. Under appropriate conditions, the bias adjustment of the discrepancy asymptotically converges to twice the overall number of estimated parameters, indicating that the bias adjustment is asymptotically independent of the true parameter. To approximately estimate the bias adjustment, the penalty term of AICi therefore is assessed by utilizing Monte Carlo samples generated at an arbitrary parameter value.

Apart from AIC variants, the predictive divergence criterion (PDC) has been proposed by Davies et al. [4] based on cross validation approach in the linear regression framework. PDC serves as an unbiased estimator of an expected discrepancy constructed for gauging the adequacy of a candidate model by assessing how effectively each case-deleted fitted model predicts the deleted case. This expected discrepancy is asymptotically equivalent to the expected Kullback-Leibler discrepancy, the target measure for AIC, AICc and AICi.

In the mixed modeling framework, different covariance structures formulate a rigorous problem for clarifying the distribution of the candidate model, developing a variant of AIC whose justification requires the distribution of the candidate model therefore confronts a formidable challenge. In longitudinal data analysis, a corrected variant of AIC has recently been justified by Azari et al. [3] for comparing models having different mean structures yet the same covariance structure.

To improve effectiveness of mixed model selection, using computing techniques combined with asymptotic theories is an alternative approach to building up selection criteria. Shang and Cavanaugh [8] have proposed two bootstrap-corrected variants of AIC for the joint selection of the fixed and random components of a linear mixed model. These variants are justified by extending the asymptotic theory of Shibata [9]. They can be easily applied under nonparametric, semiparametric, and parametric bootstrapping.

Motivated from the considerations in the previous ideas, we propose versions of AICi and PDC suitable for mixed model applications in this paper. The selection and the estimation performance of AIC, AICi, and PDC is investigated in a simulation study. Our simulation results demonstrate that PDC outperforms AIC and AICi in choosing an appropriate mixed model as a selection criterion, and AICi is less biased than AIC and PDC in estimating the Kullback-Leibler discrepancy between the true model and a fitted candidate model.

The format of the paper is arranged as follows. In Section 2, we present the model and notation. In Section 3, we propose and discuss the criteria. Our simulation study is presented and summarized in Section 4. Concluding remarks are given in Section 5.

2. Model and Notation

For $i = 1, ..., m$, let y_i denote an $n_i \times 1$ vector of responses observed on the ith subject or case, and let b_i denote a $q \times 1$ vector of associated random effects. Assume the vectors b_i are independently distributed as $N(0, D)$. Let $N = \sum_{i=1}^{m} n_i$ denote the total number of response measurements.

The general linear mixed model can be represented as

$$Y = X\beta + Zb + \varepsilon, \tag{2.1}$$

where Y denotes the $N \times 1$ response vector $(y_1', ..., y_m')'$, X is an $N \times (p+1)$ design matrix of full column rank, Z is an $N \times mq$ block diagonal design matrix comprised of m blocks, where each block is an $n_i \times q$ matrix, β is the $(p+1) \times 1$ fixed effects parameter vector, b is the $mq \times 1$ random effects vector $(b_1', ..., b_m')'$, and ε is the $N \times 1$ error vector. We assume $b \sim N(0, G)$ and $\varepsilon \sim N(0, \sigma^2 R)$, with b and ε distributed independently. Here, R and G are positive definite block diagonal matrices and G is $mq \times mq$ and comprised of m identical blocks, each of which is D.

Let θ denote the unknown parameter vector, consisting of the elements of the vector β, the matrix D, and the scalar σ^2. Let $V = ZGZ' + \sigma^2 R$. Note that V represents the covariance matrix of Y and that V is positive definite.

Let $\hat{\theta}$ denote the MLE of θ, consisting of the elements of the vector $\hat{\beta}$, the matrix $\hat{D}$, and the scalar $\hat{\sigma}^2$. Note that $\hat{G}$ is a positive definite block diagonal matrix and is comprised of m identical blocks, each of which is $\hat{D}$. For a given set of estimates $\hat{\theta}$, the covariance matrix of Y is given by $\hat{V} = Z\hat{G}Z' + \hat{\sigma}^2 R$.

Suppose the generating model or the true model, which presumably gave rise to the data, is given by

$$Y = X_o\beta_o + Z_o b_o + \varepsilon_o, \tag{2.2}$$

where X_o is an $N \times (p_o + 1)$ design matrix of full rank, Z_o is an $N \times mq_o$ block diagonal design matrix, where each block is an $n_i \times q_o$ matrix. Here, the other terms are similarly defined as those in the model (2.1). For brevity, similar notations are not repeatedly depicted here. We only need to add the subscript "o" to the terms related to the generating model including the parameters arising from the model.

3. Selection Criteria

3.1 An improved Akaike information criterion (AICi) based on Monte Carlo simulation

Suppose that a candidate or approximating model is a model that could potentially be used to describe the data and a fitted model is a candidate model that has been fit to the data. A candidate class contains all the candidate models of interest. The Kullback-Leibler discrepancy between the generating model or the true model and a candidate or approximating model is defined as

$$d(\theta, \theta_o) = E_o\{-2 \ln f(Y \mid \theta)\},$$

where E_o denotes the expectations under the generating model, and $f(Y \mid \theta)$ represents the probability density function corresponding to the approximating model.

For a given set of estimates $\hat{\theta}$,

$$d(\hat{\theta}, \theta_o) = E_o\{-2 \ln f(Y \mid \theta)\}|_{\theta=\hat{\theta}} \tag{3.1}$$

would provide a useful measure of separation between the fitted approximating model and the generating model. The overall Kullback-Leibler discrepancy is the expected value of (3.1),

$$\begin{aligned}\delta_{\mathrm{AIC}}(k, \theta_o) &= E_o[d(\hat{\theta}, \theta_o)] \\ &= E_o\{E_o\{-2 \ln f(Y \mid \theta)\}|_{\theta=\hat{\theta}}\} \\ &= E_o[-2 \ln f(Y \mid \hat{\theta})] + B(k, \theta_o),\end{aligned} \tag{3.2}$$

where the bias adjustment $B(k, \theta_o)$ is evaluated as

$$B(k, \theta_o) = E_o\{E_o\{-2 \ln f(Y \mid \theta)\}|_{\theta=\hat{\theta}}\} - E_o\{-2 \ln f(Y \mid \hat{\theta})\}. \tag{3.3}$$

We identify a candidate model having the same structure as the true model as correctly specified, a candidate model which has a more simplistic structure than the true model (e.g., includes fewer parameters, explanatory variables, effects, etc.) as underspecified, and a candidate model which has a more complex structure than the true model (e.g., includes more parameters, explanatory variables, effects, etc.) as

overspecified. When a candidate is fit to the data, correspondingly, we will have a correctly specified, underfit, or overfit model.

Akaike's [1] original AIC is given by

$$\mathrm{AIC} = -2\ln f(\hat{\theta}\,|\,Y) + 2k,$$

where $f(\hat{\theta}\,|\,Y)$ is the maximized likelihood function, and k represents the dimension of estimated parameter $\hat{\theta}$ under the given candidate model. Here, the "goodness of fit" term, $-2\ln f(\hat{\theta}\,|\,Y)$, gauges how well the model fits the data, and the penalty term, $2k$, measures the complexity that compensates for the bias in the lack of fit when the maximum likelihood estimators are used. The success of AIC depends on its approximation to the bias adjustment (3.3) by $2k$ for large samples. In Akaike's justification of AIC, two fairly general assumptions are involved:

(a) The fitted model is either correctly specified or overfit, i.e., θ_o is a member of the candidate class parameters θ.

(b) A set of regularity conditions holds to ensure the conventional asymptotic properties of the maximum likelihood estimator $\hat{\theta}$.

These two assumptions imply that AIC only applies to "correctly specified or overfit" candidate models and the regularity conditions must hold among "correctly specified or overfit" models, yet these assumptions never limit a wide range application of AIC since it has a simple form and is easily computed. However, as k increases relative to the sample size, AIC becomes increasingly negatively biased. The negative bias of AIC in small-sample applications often results in severe overfitting.

To overcome this problem of AIC, AICi was proposed in the work advanced by Hurvich et al. [5] in the context of univariate Gaussian autoregressive models.

The idea of AICi is motivated by the meaning of AIC. Under the conditions (a) and (b), the bias adjustment (3.3) converges to $2k$, indicating that the dependence of the bias adjustment on the true parameter θ_o diminishes as the sample size tends to infinity. As a result, in large-sample applications, to estimate the bias adjustment (3.3), we

can use an arbitrary parameter value instead of the true parameter θ_o. Evaluating the bias adjustment (3.3) is not possible since doing so requires the knowledge of the true parameter θ_o. If an arbitrary parameter value is utilized to evaluate (3.3), the problem will be solved. We can estimate the bias adjustment (3.3) by an estimator with an arbitrary "true" parameter.

With an arbitrary chosen "true" parameter, estimating (3.3) relies upon how to estimate the first term $E_o\{E_o\{-2\ln f(Y\mid\theta)\}|_{\theta=\hat{\theta}}\}$. Note that in this term the inner expectation is taken with respect to the distribution of Y at an arbitrarily chosen "true" parameter and is easily accessed, and the outer expectation is taken with respect to the sampling distribution of the MLE $\hat{\theta}$. By the strong law of large numbers, the outer expectation can be approximately estimated by averaging over a large collection of the MLE's of Monte Carlo samples generated from the true model at an arbitrarily chosen "true" parameter. Hence, AICi will be constructed by a "goodness of fit" term same as that of AIC, and a penalty term evaluated by Monte Carlo samples.

For the purpose of simplifying the computation, we are naturally inclined to choose convenient values for θ_o if possible. Although this simulated approximation to the bias correction originates from asymptotic instances, in small to moderate-sample applications, the approximation should provide a more accurate estimate for the bias adjustment in (3.3) than $2k$. This claim can be supported both by the simulation results that follow and by those reported in Hurvich et al. [5]. For the clarification of the notation, we denote the true model parameter by θ_o, yet denote the arbitrary choice of the true parameter for Monte Carlo simulation by θ_a.

To evaluate the AICi in the mixed modeling framework, we first note that

$$-2\ln f(Y\mid\theta) = \ln|V| + (Y - X\beta)'V^{-1}(Y - X\beta), \tag{3.4}$$

and we then have

$$E_o\{-2\ln f(Y\mid\theta)\} = \ln|V| + \operatorname{tr}(V_oV^{-1}) + (X_o\beta_o - X\beta)'V^{-1}(X_o\beta_o - X\beta).$$

(In the preceding relations and throughout the paper, we have neglected the additive constant $n \ln 2\pi$.) For a given $\hat{\theta}$, we can arrive at

$$E_o\{-2\ln f(Y\,|\,\theta)\}|_{\theta=\hat{\theta}} = \ln|\hat{V}| + \mathrm{tr}\,(V_o\hat{V}^{-1}) + (X_o\beta_o - X\hat{\beta})'\hat{V}^{-1}(X_o\beta_o - X\hat{\beta}). \tag{3.5}$$

Again, since the bias adjustment is asymptotically independent of the true parameter, using the arbitrary choice of the true parameter for Monte Carlo simulation, it can be rewritten as

$$B(k,\,\theta_a) = E_a\{E_a\{-2\ln f(Y_*\,|\,\theta)\}|_{\theta=\hat{\theta}_*}\} - E_a\{-2\ln f(Y_*\,|\,\hat{\theta}_*)\}, \tag{3.6}$$

where Y_* is a data vector generated from the true model at parameter θ_a and $\hat{\theta}_*$ is the MLE of θ based on maximizing $f(Y_*\,|\,\theta_a)$; the sampling distribution of $\hat{\theta}_*$ is governed by the density function $f(Y_*\,|\,\theta_a)$; and the expectation $E_a\{\cdot\}$ is taken with respect to $f(Y_*\,|\,\theta_a)$. Note that $B(k,\,\theta_o)$ and $B(k,\,\theta_a)$ are asymptotically equivalent when the sample size is large enough. Also note that the crucial difference between θ_o and θ_a is that θ_o is unknown and that θ_a is known. Certainly, the convenient chosen values for θ_a may simplify the computation. However, provided that computation allows, the parameter θ_a could be any value. Since the justification of AICi requires that the candidate models are correctly specified or overfit, we need to limit θ_a in the parameter space of the candidate class. In practice, since the true parameter is unknown, we can choose an arbitrary parameter θ_a including less parameters so that the candidate models are more likely to be correctly specified or overfit ones. Once θ_a is chosen, we can generate Monte Carlo samples from the model with the parameter θ_a.

Let $\{Y_*(j),\ j = 1, ..., K\}$ represent a set of K Monte Carlo samples generated from the model with the density function $f(Y_*\,|\,\theta_a)$. Let $\{\hat{\theta}_*(j),\ j = 1, ..., K\}$ represent a set of the MLEs corresponding to $\{Y_*(j),\ j = 1, ..., K\}$, respectively.

Now by the strong law of large numbers, as $K \to \infty$, one can argue that

$$\frac{1}{K}\sum_{j=1}^{K} E_a\{-2\ln f(Y_* \mid \theta)\}\big|_{\theta=\hat{\theta}_*(j)} \to E_a\{E_a\{-2\ln f(Y_* \mid \theta)\}\big|_{\theta=\hat{\theta}_*}\} \text{a.s.},$$

and

$$\frac{1}{K}\sum_{j=1}^{K} \{-2\ln f(Y_*(j) \mid \hat{\theta}_*(j))\} \to E_a\{-2\ln f(Y_* \mid \hat{\theta}_*)\} \text{a.s.} \tag{3.7}$$

Expressions (3.6) and (3.7) result in the following large-sample estimator of the bias adjustment:

$$\hat{B}(k, \theta_a) = \frac{1}{K}\sum_{j=1}^{K} [E_a\{-2\ln f(Y_* \mid \theta)\}\big|_{\theta=\hat{\theta}_*(j)} - \{-2\ln f(Y_*(j) \mid \hat{\theta}_*(j))\}]. \tag{3.8}$$

As previously discussed, AICi is thereby defined as

$$\text{AICi} = -2\ln f(Y \mid \hat{\theta}) + \hat{B}(k, \theta_a). \tag{3.9}$$

Since $\hat{B}(k, \theta_a)$ is an asymptotically unbiased estimator of the bias adjustment (3.3), we can easily verify that AICi in (3.9) serves as an asymptotically unbiased estimator of the Kullback-Leibler discrepancy in (3.2).

For the further development of AICi, we need to introduce some notation related to θ_a in the sense of the true model for the data vector Y_*. Similar to the unknown parameter θ_o in model (2.2), let θ_a denote the set of parameters for the generating model of Y_*, i.e., θ_a consists of the elements of the vector β_a, the matrix D_a, and the scalar σ_a^2. Under the true model, the covariance matrix of Y_* can be written as $V_a = Z_a G_a Z_a' + \sigma_a^2 R_a$, where Z_a is a design matrix for the random effects under this model, R_a is a known matrix, and G_a is a positive definite block diagonal matrix and is comprised of m identical blocks, each of which is D_a. We generate $2K$ Monte Carlo samples $\{Y_*(1), ..., Y_*(K),$

$Y_*(K+1), ..., Y_*(2K)\}$ from the true model at the chosen parameter value and solve for the K sets of corresponding MLEs $\{(\hat{D}_*(1), \hat{\beta}_*(1), \hat{\sigma}^2_*(1)), ..., (\hat{D}_*(K), \hat{\beta}_*(K), \hat{\sigma}^2_*(K))\}$ under a candidate model (2.1) based on maximizing $\{f(Y_*(K+1)|\theta), ..., f(Y_*(2K)|\theta)\}$, respectively. Doing so ensures to make full use of the information from Monte Carlo samples. By the expressions of (3.8), (3.4) and (3.5), we can further develop $\hat{B}(k, \theta_a)$ in the mixed model setting as

$$\begin{aligned}\hat{B}(k, \theta_a) &= \frac{1}{K}\sum_{j=1}^{K}[\ln|\hat{V}_*(j)| + \mathrm{tr}(V_a\hat{V}_*^{-1}(j)) \\ &\quad + (X_a\beta_a - X\hat{\beta}_*(j))'\hat{V}_*^{-1}(j)(X_a\beta_a - X\hat{\beta}_*(j))] \\ &\quad - \frac{1}{K}\sum_{j=1}^{K}[\ln|\hat{V}_*(j)| + (Y_*(j) - X\hat{\beta}_*(j))'\hat{V}_*^{-1}(j)(Y_*(j) - X\hat{\beta}_*(j))] \\ &= \frac{1}{K}\sum_{j=1}^{K}[\mathrm{tr}(V_a\hat{V}_*^{-1}(j)) + (X_a\beta_a - X\hat{\beta}_*(j))'\hat{V}_*^{-1}(j)(X_a\beta_a - X\hat{\beta}_*(j))] \\ &\quad - \frac{1}{K}\sum_{j=1}^{K}[(Y_*(j) - X\hat{\beta}_*(j))'\hat{V}_*^{-1}(j)(Y_*(j) - X\hat{\beta}_*(j))]. \qquad (3.10)\end{aligned}$$

Note that in the preceding expression, $X_a\beta_a$ is the mean of Y_* under its generating model. Expression (3.10) is the penalty term of AICi, and is utilized to estimate the bias adjustment for compensating the lack of fit from the biased estimator.

3.2. The predictive divergence criterion (PDC) based on cross validation

To assess predictive ability of candidate models, the predictive divergence criterion (PDC) has been recently justified by Davies et al. [4] in the linear regression framework. Although the target discrepancy upon which PDC is based is not the Kullback-Leibler discrepancy, it essentially measures the dissimilarity between the generating model and

a fitted candidate model. Furthermore, the target overall PDC discrepancy is asymptotically equivalent to the Kullback-Leibler discrepancy and PDC is therefore an asymptotically unbiased estimator of the Kullback-Leibler discrepancy.

Since PDC makes use of cross-validation approach, we need to introduce the notation in the cross validation setting. For the candidate model, let $\hat{\theta}^i$ represent the maximum likelihood estimator of θ based on the data set removing the ith case y_i from the vector Y, and let $f_i(y_i \mid \theta)$ represent the individual probability density function of the ith case for $i = 1, \ldots, m$. The likelihoods corresponding to the generating model and the candidate model can then be expressed as

$$L(\theta_o \mid Y) = \prod_{i=1}^{m} f_i(y_i \mid \theta_o)$$

and

$$L(\theta \mid Y) = \prod_{i=1}^{m} f_i(y_i \mid \theta),$$

respectively.

We notice that

$$d_i(\hat{\theta}^i, \theta_o) = E_o\{-2 \ln f_i(y_i \mid \theta)\}\big|_{\theta=\hat{\theta}^i}$$

would measure an individual discrepancy between the case-deleted fitted model and the true model for the deleted case. Since each case y_i is generated from the true model, in some sense this individual discrepancy reflects how well each case-deleted fitted model predicts the deleted case. For the predictive divergence criterion (PDC), the overall predictive discrepancy for the m independent cases can therefore be defined by

$$\begin{aligned} d_{\mathrm{PDC}}(Y, \theta_o) &= \sum_{i=1}^{m} d_i(\hat{\theta}^i, \theta_o) \\ &= \sum_{i=1}^{m} E_o\{-2 \ln f_i(y_i \mid \theta)\}\big|_{\theta=\hat{\theta}^i}. \end{aligned} \tag{3.11}$$

The expected overall PDC discrepancy corresponding to (3.11) is

$$\delta_{\mathrm{PDC}}(k, \theta_o) = \sum_{i=1}^{m} E_o\{E_o\{-2 \ln f_i(y_i \mid \theta)\}|_{\theta=\hat{\theta}^i}\}. \tag{3.12}$$

Note that the overall discrepancy $d_{\mathrm{PDC}}(Y, \theta_o)$ is not a statistic since the evaluation of $d_{\mathrm{PDC}}(Y, \theta_o)$ requires knowledge of θ_o. Therefore, $d_{\mathrm{PDC}}(Y, \theta_o)$ cannot be used to estimate $\delta_{\mathrm{PDC}}(k, \theta_o)$.

However, the log-likelihood measure

$$\sum_{i=1}^{m} -2 \ln f_i(y_i \mid \hat{\theta}^i)$$

is a statistics and thus can be used to estimate $\delta_{\mathrm{PDC}}(k, \theta_o)$. Moreover, this statistics is exactly unbiased for $\delta_{\mathrm{PDC}}(k, \theta_o)$, since

$$\begin{aligned}
\delta_{\mathrm{PDC}}(k, \theta_o) &= E_o\{d_{\mathrm{PDC}}(Y, \theta_o)\} \\
&= \sum_{i=1}^{m} E_o\{E_o\{-2 \ln f_i(y_i \mid \theta)\}|_{\theta=\hat{\theta}^i}\} \\
&= E_o\left\{\sum_{i=1}^{m} -2 \ln f_i(y_i \mid \hat{\theta}^i)\right\}.
\end{aligned}$$

Note that in the preceding expression, for the second expectation on the right-hand side, the inner expectation is taken with respect to the distribution of y_i (under the true model), and the outer expectation is taken with respect to the sampling distribution of $\hat{\theta}^i$. See Davies et al. [4] for details. Thus, we define the model selection criterion

$$\mathrm{PDC} = \sum_{i=1}^{m} -2 \ln f_i(y_i \mid \hat{\theta}^i).$$

PDC serves as an exactly unbiased estimator of $\delta_{\mathrm{PDC}}(k, \theta_o)$ regardless of the sample size, the relationship between the likelihoods $L(\theta_o \mid Y)$ and $L(\theta \mid Y)$, or the distribution of the underlying data.

For the mixed modeling framework, PDC is therefore defined as

$$\text{PDC} = \ln|\hat{V}^i| + (Y - X\hat{\beta}^i)'(\hat{V}^i)^{-1}(Y - X\hat{\beta}^i),$$

where β^i and V^i are the corresponding MLE's with the ith case deleted from the data under the given candidate model.

For the m independent cases, the targeted measure of AIC in (3.2) can be re-expressed as

$$\delta_{\text{AIC}}(k, \theta_o) = E_o\{E_o\{-2 \ln L(\theta \mid Y)\}|_{\theta=\hat{\theta}}\}$$

$$= \sum_{i=1}^{m} E_o\{E_o\{-2 \ln f_i(y_i \mid \theta)\}|_{\theta=\hat{\theta}}\}$$

and the target measure of PDC in (3.12) can again be written as

$$\delta_{\text{PDC}}(k, \theta_o) = \sum_{i=1}^{m} E_o\{E_o\{-2 \ln f_i(y_i \mid \theta)\}|_{\theta=\hat{\theta}^i}\}.$$

When the sample size m approaches infinity, Davies et al. [4] have proved that $\delta_{\text{PDC}}(k, \theta_o) \approx \delta_{\text{AIC}}(k, \theta_o)$ holds in the linear regression framework. This claim also holds in the mixed model setting, which is supported by the simulation results in what follows. As a result, PDC serves as an exactly unbiased estimator of its own discrepancy $\delta_{\text{PDC}}(k, \theta_o)$, and is also an asymptotically unbiased estimator of $\delta_{\text{AIC}}(k, \theta_o)$.

In the development of AIC, we assume that the candidate model (2.1) is correctly specified or overfit, or the candidate family subsumes the generating model. Yet cross validation procedures do not require this assumption and are applicable in a wider range of settings than AIC.

4. Simulation Study

The goal of our simulations is to search among a class of candidate models described as what follows for the fitted model which serves as the best approximation to the specified true model. We investigate the

effectiveness of AIC, AICi and PDC at choosing this optimal fitted candidate model whose corresponding criterion is minimized among the candidate models.

In many applications data have a clustered structure, the mixed effects model treats clustered data adequately and assumes two sources of variation, within cluster and between clusters. The compound covariance structure exactly involves two variations: between subject (or cluster, location) variance and within subject (or cluster, location). We therefore consider the mixed model with the compound covariance structure both as the true model and as a type of the candidate models.

The generating model

We assume that data arises from a generating model in the form of (2.2) with the compound covariance structure, leading to the simply specified random effects. Particularly, in model (2.2), for the random effects vector b_o, the dimension $q_o = 1$, the covariance $G_o = \sigma^2_{\tau o} I$, i.e., $D_o = \sigma^2_{\tau o}$, and Z_o is an $N \times m$ block diagonal design matrix comprised of m blocks, where each block is an $n_i \times 1$ vector consisting of all 1's; for the error vector, the variance is $\sigma^2_o I$, and the positive definite matrix $R_o = I$. The covariance matrix of Y is given by $V_o = Z_o Z_o' \sigma^2_{\tau o} + \sigma^2_o I$.

The true model parameter vector θ_o can be defined as $(\beta_o', \sigma_{\tau o}, \sigma_o)'$. Note that k_o, the dimension of θ_o, is $p_o + 3$.

To generate the simulated data, we choose the parameters $\beta_o = (1, 1, 1, 1, 1, 1, 1)'$, $\sigma^2_{\tau o} = 2$, and $\sigma^2_o = 1$ for the generating model.

To compute AICi, we choose $\beta_a = (10, 10)'$, $\sigma^2_{\tau a} = 9$, and $\sigma^2_a = 5$ for the evaluation of the penalty term. For the fixed effects, a two-dimension parameter is chosen. One-dimension is for the intercept, and the other one is for one explanatory variable. Choosing one explanatory variable in the "true" model will ensure that the candidate models containing this one or more explanatory variables are correctly specified or overfit. Hence, the candidate class subsumes the chosen "true" model, which satisfies the condition of constructing AICi. Note that the chosen

parameters are not close to those for the generating model in order to show that AICi can perform quite effectively even though the choice is sightless.

To obtain the stable penalty term of AICi, the number of Monte Carlo samples K is chosen as 500. With a larger value of K, the penalty term of AICi does not change much, so the number of 500 is large enough for calculating a stable penalty term and is small enough for avoiding a waste of computation.

The candidate models

When specifying the fixed effects, we limit the candidate models using nested design matrices X. The nested models are often used in simulation studies for model selection criteria so that large candidate models may be considered without making the number of models in the candidate class excessively high. (See, for instance, McQuarrie and Tsai [7].) Suppose that P regressor variables of interest are considered, then P candidate models are reflected on a sequence of design matrices X of ranks 2, 3, ..., $(P+1)$. Each successive design matrix contains all of the regressors in its predecessors. We refer to p, the number of regressors in the candidate model, as the order of the model, and to p_o as the true order.

In our simulation study, one-dimensional random effects are inclusive or exclusive.

Matching the combination of the fixed effects and random effects, we consider two types of candidate models for modeling data Y arising from model (2.1). The first type of candidate models includes one-dimensional random effects, i.e., it has the same "compound symmetric" covariance structure as the generating model and the generating model is included in the candidate class. For the random effects vector b, the dimension $q = 1$, and the covariance $G = \sigma_\tau^2 I$, i.e., $D = \sigma_\tau^2$, Z is an $N \times m$ block diagonal design matrix comprised of m blocks, where each block is an $n_i \times 1$ vector consisting of all 1's; for the error term, the positive definite matrix $R = I$. The covariance matrix of Y is represented by $V = ZZ'\sigma_\tau^2 + \sigma^2 I$.

The candidate model parameter vector θ can be defined as $(\beta', \sigma_\tau, \sigma)'$. Note that k, the dimension of θ, is $p + 3$. The MLE's of the parameters, $\hat{\theta} = (\hat{\beta}', \hat{\sigma}_\tau, \hat{\sigma})'$, can be found via the EM algorithm. The MLE of V is given by $\hat{V} = ZZ'\hat{\sigma}_\tau^2 + \hat{\sigma}^2 I$.

The second type of candidate models excludes one-dimensional random effects, i.e., there is no within-case variation in model (2.1). That is, for the random effects vector b, the covariance $G = 0$, i.e., $D = 0$; for the error term, the positive definite matrix $R = I$. The covariance matrix of Y is represented by $V = \sigma^2 I$.

The candidate model parameter θ can be defined as $(\beta', \sigma)'$. Note that k, the dimension of θ, is $p + 2$. The MLE's of the parameters, $\hat{\theta} = (\hat{\beta}', \hat{\sigma})'$, can be easily found via ordinary least squares. The MLE of V is given by $\hat{V} = \hat{\sigma}^2 I$.

For each design matrix with the order p, we consider the candidate models both with and without the random effects in model (2.1). As a result, the candidate class takes account of the first type of candidate models, i.e., the P candidate models involving both a sequence of design matrices X of ranks 2, 3, ..., $(P + 1)$ and the random effects, and the second type of candidate models, i.e., the P candidate models only involving a sequence of design matrices X of ranks 2, 3, ..., $(P + 1)$. Again, the generating model is included in the first type of candidate class. Therefore, in the first type of candidate class, the model of order p_o is correctly specified $(1 \le p_o \le P)$. Fitted models for which $p < p_o$ are underfit, and those for which $p > p_o$ are overfit. We randomly generate all regressors as independent, identically distributed variates from a standard normal distribution.

Simulation results

To inspect the performance of the criteria AICi and PDC, the simulation is completed in four sets for $m = 15, 20, 30,$ and 50 with

$n = 3$ observations for each case. For each simulation set, 100 samples consisting of $N = m \times n$ observations are generated from the specified true model of order $p_o = 6$. The maximum order of the candidate class is set at $P = 12$. For every sample, each candidate model is fit to the data, the criteria AIC, AICi, PDC, and the simulated values of $\delta_{\mathrm{AIC}}(\hat{\theta}, \theta_o)$ and $\delta_{\mathrm{PDC}}(\hat{\theta}, \theta_o)$ are evaluated, and the fitted model favored by each criterion or by each simulated discrepancy is recorded. Over the 100 samples, the distribution of model selections is tabulated for each of the criteria and for each of the discrepancies.

For each simulation set $(m = 15, 20, 30, 50)$, the distributions of selections by AIC, AICi, PDC, and the discrepancies $\delta_{\mathrm{AIC}}(\hat{\theta}, \theta_o)$ and $\delta_{\mathrm{PDC}}(\hat{\theta}, \theta_o)$ are compiled over the 100 samples.

To explore the effectiveness of the criteria as asymptotically unbiased estimators of $\delta_{\mathrm{AIC}}(k, \theta_o)$, the average values of the criterion or discrepancy are computed for each of the two types of candidate models over the orders 1 through P on the 100 samples. Then we plot the averages of $\delta_{\mathrm{AIC}}(k, \theta_o)$ and $\delta_{\mathrm{PDC}}(k, \theta_o)$ along with the averages for AIC, AICi, and PDC against the orders from 1 to P.

The order selections for AIC, AICi, PDC, and two discrepancies are reported in Table 1. Over all four sets, PDC obtains the most correct model selections as a selection criterion. In the sets where the sample size is small $(m = 15$ or $m = 20)$ or moderate $(m = 30)$, AICi and PDC both outperform AIC as a selection criterion. However, in the set where the sample size is large $(m = 50)$, only PDC significantly outperforms AIC in choosing the correct model. In this set, AICi and AIC obtain a comparable number of correct model selections, although AICi tends to choose more parsimonious models.

Table 1. Model selections for simulations

m		With random effects			Without random effects		
		Underfit	Correctly specified	Overfit	Underfit	Correctly specified	Overfit
15	AIC	0	61	37	0	1	1
	δ_{AIC}	4	87	8	0	1	0
	AICi	15	76	9	0	0	0
	PDC	0	84	12	0	4	0
	δ_{PDC}	1	95	0	0	4	0
20	AIC	0	57	43	0	0	0
	δ_{AIC}	0	92	7	0	1	0
	AICi	0	78	22	0	0	0
	PDC	0	84	15	0	1	0
	δ_{PDC}	0	99	0	0	1	0
30	AIC	0	73	27	0	0	0
	δ_{AIC}	0	96	4	0	0	0
	AICi	0	78	22	0	0	0
	PDC	0	88	12	0	0	0
	δ_{PDC}	0	100	0	0	0	0
50	AIC	0	66	34	0	0	0
	δ_{AIC}	0	93	7	0	0	0
	AICi	0	64	36	0	0	0
	PDC	0	73	27	0	0	0
	δ_{PDC}	0	99	1	0	0	0

Figures 1-4 demonstrate how effectively the criteria serve as approximately unbiased estimators of $\delta_{AIC}(k, \theta_o)$. As the sample size increases, the average curves for AICi and PDC tend to grow closer, both approaching the simulated $\delta_{AIC}(k, \theta_o)$ curve. This implies that AICi and PDC are all asymptotically unbiased estimators of $\delta_{AIC}(k, \theta_o)$.

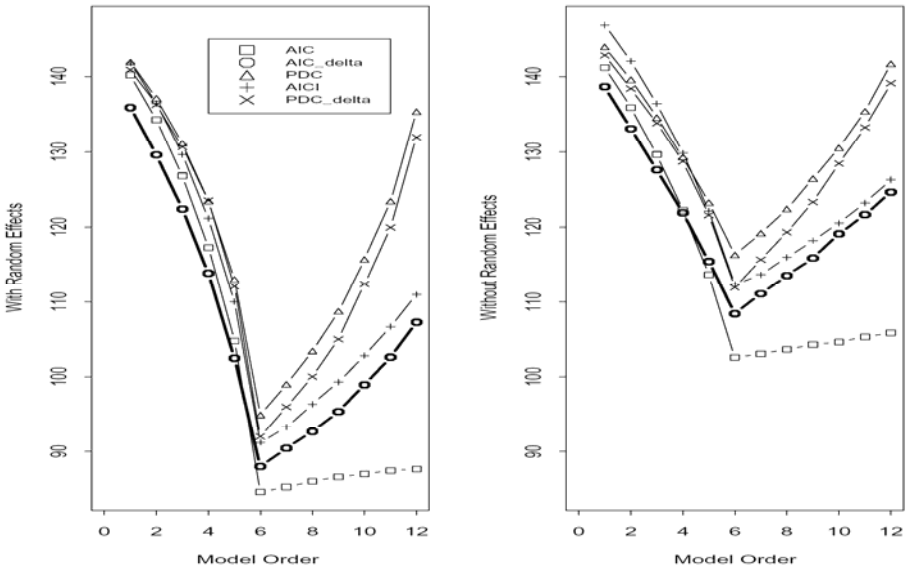

Figure 1. Averages of criteria and simulated discrepancies ($m = 15$).

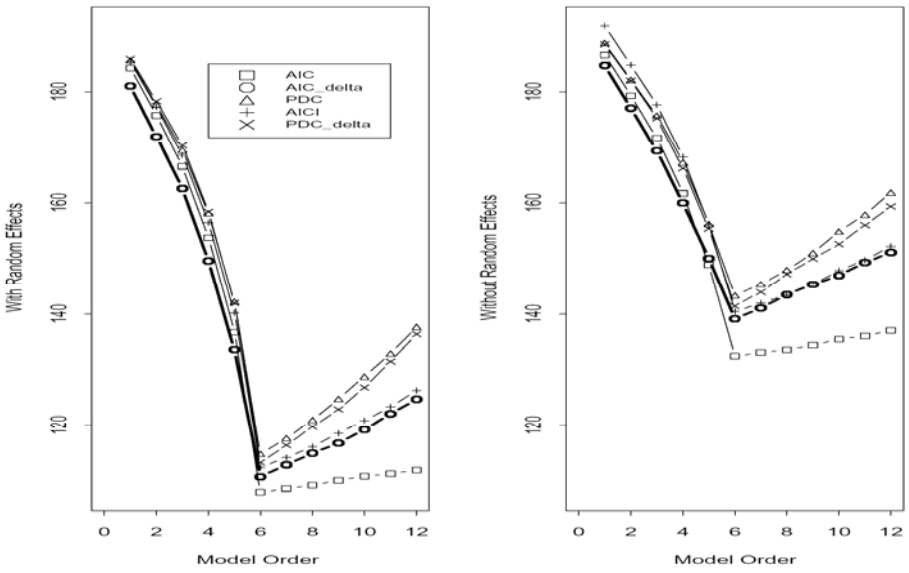

Figure 2. Averages of criteria and simulated discrepancies ($m = 20$).

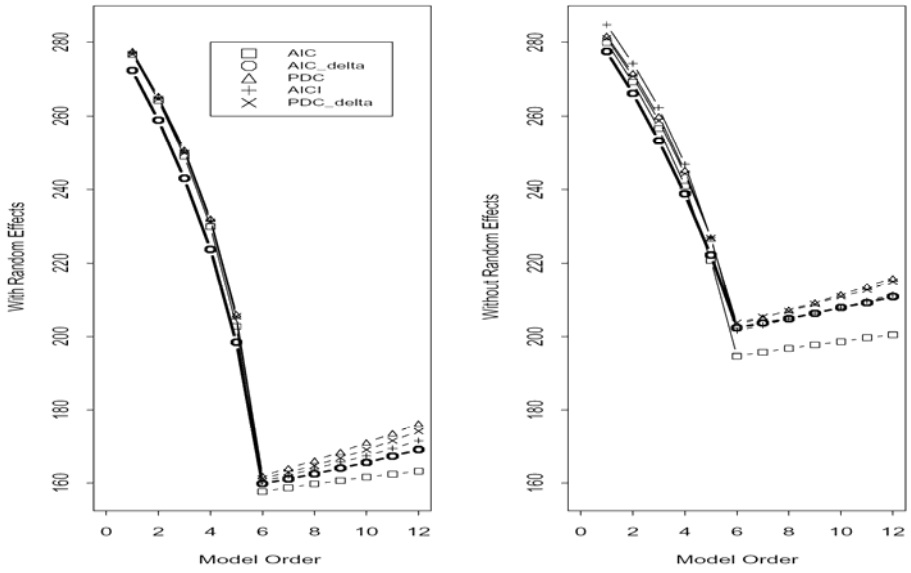

Figure 3. Averages of criteria and simulated discrepancies ($m = 30$).

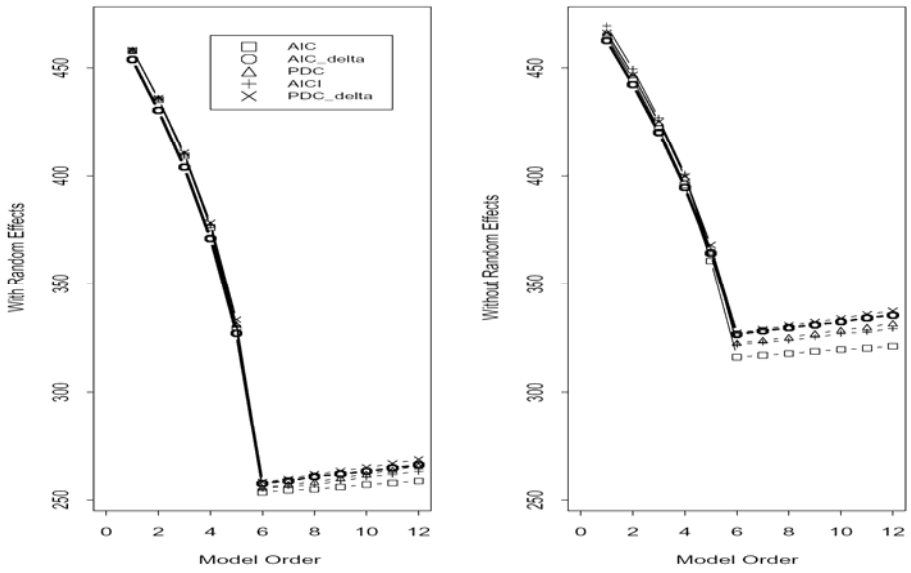

Figure 4. Averages of criteria and simulated discrepancies ($m = 50$).

Figures 1-4 also exhibit that the criterion PDC is more accurate as an estimator of its target discrepancy $\delta_{\mathrm{PDC}}(k, \theta_o)$ than as an estimator of $\delta_{\mathrm{AIC}}(k, \theta_o)$. In each simulation set (except for the set with $m = 50$), the average curve of PDC tracks the curve of $\delta_{\mathrm{PDC}}(k, \theta_o)$ more closely than the curve of $\delta_{\mathrm{AIC}}(k, \theta_o)$. This tendency can be explained by our previous justification, which claims that PDC is an unbiased estimator of its target discrepancy $\delta_{\mathrm{PDC}}(k, \theta_o)$ and is an asymptotically unbiased estimator of the target discrepancy of AIC, $\delta_{\mathrm{AIC}}(k, \theta_o)$.

The asymptotic equivalence of the two criterion discrepancies is further shown by Figures 1-4. With an increasing sample size, the simulated average curves for $\delta_{\mathrm{AIC}}(k, \theta_o)$ and $\delta_{\mathrm{AIC}}(k, \theta_o)$ approach each other. In large sample set $(m = 50)$, we can see the $\delta_{\mathrm{AIC}}(k, \theta_o)$ and $\delta_{\mathrm{AIC}}(k, \theta_o)$ curves are almost the same. This trend confirms that the target discrepancy of PDC, $\delta_{\mathrm{PDC}}(k, \theta_o)$, and the target discrepancy of AIC, $\delta_{\mathrm{AIC}}(k, \theta_o)$ are asymptotically equivalent.

For correctly specified or overfit models, the average AICi curve follows the simulated $\delta_{\mathrm{AIC}}(k, \theta_o)$ curve more closely than either the average AICi or PDC curve. This reveals that AICi is less biased than AIC and PDC in estimating the expected Kullback-Leibler discrepancy $\delta_{\mathrm{AIC}}(k, \theta_o)$.

5. Conclusion

Under suitable conditions in the general linear mixed model, AICi serves as an asymptotically unbiased estimator of the expected discrepancy $\delta_{\mathrm{AIC}}(k, \theta_o)$ between the generating model and a fitted approximating model. PDC provides an exactly unbiased estimator for its targeted discrepancy and acts as an asymptotically unbiased estimator of the expected discrepancy $\delta_{\mathrm{AIC}}(k, \theta_o)$ as well.

The simulation study indicates that AICi and PDC perform effectively in selecting a mixed model with an appropriate mean and covariance structure. PDC exhibits a higher success rate in identifying the correct

model than either AIC and AICi. In small sample applications, both AICi and PDC outperform AIC in selecting the correct model.

In addition to the model selection, the simulation results demonstrate that the AICi provides considerably less unbiased estimates of the expected discrepancy $\delta_{\mathrm{AIC}}(k, \theta_o)$ than AIC and PDC.

PDC is developed in the context of a general model formulation and a nonrestrictive set of conditions; whereas AICi is justified under certain conditions. From this point of view, PDC can be applied in a wider range of settings.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, 2nd International Symposium on Information Theory, B. N. Petrov and F. Csaki, eds., pp. 267-281, Akademia Kiado, Budapest, 1973.

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

[3] R. Azari, L. Li and C. L. Tsai, Longitudinal data model selection, Comput. Statist. Data Anal. 50 (2006), 3053-3066.

[4] S. L. Davies, A. A. Neath and J. E. Cavanaugh, Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy, Stat. Methodol. 2 (2005), 249-266.

[5] C. M. Hurvich, R. H. Shumway and C. L. Tsai, Improved estimators of Kullback-Leibler information for auto regressive model selection in small samples, Biometrika 77 (1990), 709-719.

[6] C. M. Hurvich and C. L. Tsai, Regression and time series model selection in small samples, Biometrika 78 (1989), 499-509.

[7] A. D. R. McQuarrie and C. L. Tsai, Regression and Time Series Model Selection, World Scientific, River Edge, New Jersey, 1998.

[8] J. Shang and J. E. Cavanaugh, Bootstrap variants of the Akaike information criterion for mixed model selection, Comput. Statist. Data Anal. 52 (2008), 2004-2021.

[9] R. Shibata, Bootstrap estimate of Kullback-Leibler information for model selection, Statist. Sinica 7 (1997), 375-394.

[10] N. Sugiura, Further analysis of the data by Akaike's information criterion and the finite corrections, Commun. Statist. Theory Methods 7 (1978), 13-26.

■