

Bootstrap Variants of the Akaike Information Criterion for Mixed Model Selection

Junfeng Shang^{1,*} and Joseph E. Cavanaugh^{2,†}

¹Bowling Green State University, USA

²The University of Iowa, USA

Abstract: Two bootstrap-corrected variants of the Akaike information criterion are proposed for the purpose of small-sample mixed model selection. These two variants are asymptotically equivalent, and provide asymptotically unbiased estimators of the expected Kullback-Leibler discrepancy between the true model and a fitted candidate model. The performance of the criteria is investigated in a simulation study where the random effects and the errors for the true model are generated from a Gaussian distribution. The parametric bootstrap is employed. The simulation results suggest that both criteria provide effective tools for choosing a mixed model with an appropriate mean and covariance structure. A theoretical asymptotic justification for the variants is presented in the Appendix.

Key words: AIC, Kullback-Leibler information, model selection criteria

*Corresponding author. Phone: (419) 372-7457. Fax: (419) 372-6092.
e-mail: jshang@bgsu.edu. Department of Mathematics and Statistics, 450 Math Science Building,
Bowling Green State University, Bowling Green, OH 43403.

†e-mail: joe-cavanaugh@uiowa.edu. Department of Biostatistics, The University of Iowa, IA 52242.

1. Introduction

The first model selection criterion to gain widespread acceptance was the Akaike (1973, 1974) information criterion, AIC. AIC was developed to estimate the expected Kullback-Leibler (1951) discrepancy between the generating model and a fitted candidate model. AIC is applicable in a broad array of modeling frameworks, since its large-sample justification only requires conventional asymptotic properties of maximum likelihood estimators. However, in settings where the sample size is small, AIC may underestimate the expected discrepancy for fitted candidate models of high dimension. As a result, the criterion may choose such a model even when the expected discrepancy for the model is relatively large (Hurvich and Tsai, 1989). This limits the effectiveness of AIC as a model selection criterion. To adjust for this weakness, the “corrected” AIC, AICc, has been proposed.

AICc (Sugiura, 1978; Hurvich and Tsai, 1989) has proven to be one of the most effective model selection criteria in an increasingly crowded field (McQuarrie and Tsai, 1998, p. 2). Originally developed for linear regression, AICc has been extended to a number of additional frameworks, including autoregressive moving-average modeling (Hurvich, Shumway, and Tsai, 1990), vector autoregressive modeling (Hurvich and Tsai, 1993), and multivariate regression modeling (Bedrick and Tsai, 1994). In small-sample applications, AICc often dramatically outperforms AIC as a selection criterion. Since the basic form of AICc is identical to that of AIC, the improvement in selection performance comes without an increase in computational cost. However, AICc is less generally applicable than AIC since its justification relies upon the structure of the candidate model.

In the mixed modeling framework, developing a corrected variant of AIC that is appropriate for comparing models having both different mean and different covariance structures presents a formidable challenge. In the setting of mixed models amenable to longitudinal data analysis, AICc has recently been justified by Azari, Li, and Tsai (2006) for comparing models having different mean structures yet the same covariance structure. Yet the authors

remark (p. 3065) "...finding a selection criterion to jointly select regression variables and covariance parameters would be a potential research area for further study."

With this motivation, we propose 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 (1997). They can be easily applied under nonparametric, semiparametric, and parametric bootstrapping. We investigate the performance of the bootstrap criteria in a simulation study.

The idea of using the bootstrap to improve the performance of a model selection rule was introduced by Efron (1983, 1986), and is extensively discussed by Efron and Tibshirani (1993, pp. 237-253). Ishiguro and Sakamoto (1991) advocated a bootstrap variant of AIC, WIC, which is based on Efron's methodology. Ishiguro, Morita, and Ishiguro (1991) used this variant successfully in an aperture synthesis imaging problem. Cavanaugh and Shumway (1997) proposed a bootstrap variant of AIC, AICb, for state-space model selection. Shibata (1997) has established the asymptotic equivalence of AICb and WIC under a general set of assumptions, and has indicated the existence of other asymptotically equivalent bootstrap-corrected AIC variants. Shibata's framework, however, applies only to independent data, and therefore does not accommodate the type of data arising in mixed modeling applications.

In Section 2, we present the mixed model and briefly review the motivation behind AIC. We then present the bootstrap AIC variants. In Sections 3 and 4, the performance of the variants is investigated in a simulation study that also evaluates AIC. Section 5 concludes. A theoretical asymptotic justification for the criteria is presented in the Appendix.

2. The Bootstrap AIC Variants

For $i = 1, \dots, m$, let y_i denote an $n_i \times 1$ vector of responses observed on the i th 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

$$Y = X\beta + Zb + \varepsilon, \quad (2.1)$$

where Y denotes the $N \times 1$ response vector $(y_1', \dots, y_m)'$; X is an $N \times (p+1)$ design matrix of full column rank; Z is an $N \times mq$ block diagonal design matrix comprised on 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 $(b_1', \dots, b_m)'$; and ε is the $N \times 1$ error vector. We assume $b \sim N(0, D)$ and $\varepsilon \sim N(0, \sigma^2 R)$, with b and ε distributed independently. Here, R and D are positive definite block diagonal matrices and D 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 $V = ZDZ' + \sigma^2 R$. Note that V represents the covariance matrix of Y and that V is positive definite.

A well-known measure of separation between two models is given by the non-normalized Kullback-Leibler information, also known as the cross entropy or discrepancy. Let θ_o represent the set of parameters for the “true” or generating model and θ represent the set of parameters for a candidate or approximating model. The Kullback-Leibler discrepancy between the models is defined as

$$d(\theta, \theta_o) = E_o\{-2 \log L(\theta | Y)\},$$

where E_o denotes the expectation under the generating model, and $L(\theta | Y)$ represents the likelihood corresponding to the approximating model.

For a given set of estimates $\hat{\theta}$, the overall Kullback-Leibler discrepancy

$$d(\hat{\theta}, \theta_o) = E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}} \quad (2.2)$$

would provide a useful measure of separation between the generating model and the fitted approximating model. Yet evaluating (2.2) is not possible since doing so requires the knowledge of θ_o .

Akaike (1973), however, noted that $-2 \log L(\hat{\theta} | Y)$ serves as a biased estimator of the expected value of (2.2), and that the bias adjustment

$$E_o\{E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}}\} - E_o\{-2 \log L(\hat{\theta} | Y)\} \quad (2.3)$$

can often be asymptotically estimated by twice the dimension of $\hat{\theta}$.

Thus, if we let k represent the dimension of $\hat{\theta}$, then under appropriate conditions, the expected value of

$$\text{AIC} = -2 \log L(\hat{\theta} | Y) + 2k$$

should be asymptotically close to the expected value of (2.2), say

$$\begin{aligned} \Delta(k, \theta_o) &= E_o\{d(\hat{\theta}, \theta_o)\} \\ &= E_o\{E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}}\} \\ &= E_o\{-2 \log L(\hat{\theta} | Y)\} \\ &\quad + [E_o\{E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}}\} - E_o\{-2 \log L(\hat{\theta} | Y)\}]. \end{aligned} \quad (2.4)$$

Note that the “goodness of fit” term in AIC, $-2 \log L(\hat{\theta} | Y)$, estimates the first of the terms in (2.4), whereas the “penalty” term in AIC, $2k$, estimates the bias expression (2.3).

AIC provides us with an asymptotically unbiased estimator of $\Delta(k, \theta_o)$ in settings where the sample size is large and k is comparatively small. In settings where the sample size is small and k is comparatively large, $2k$ is often much smaller than the bias adjustment (2.3), making AIC substantially negatively biased as an estimator of $\Delta(k, \theta_o)$ (Hurvich and Tsai, 1989). If AIC severely underestimates $\Delta(k, \theta_o)$ for higher dimensional fitted models in the candidate set, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large.

To adjust for this weakness of AIC in the setting of mixed models, we utilize the bootstrap to develop two variants of AIC. These criteria are formulated by constructing approximately unbiased estimators of the expected discrepancy $\Delta(k, \theta_o)$. Specifically, we propose

two bootstrap-based estimators for the bias adjustment (2.3), which in small-sample settings should estimate (2.3) more accurately than $2k$.

Let Y^* represent a bootstrap sample, based on resampling on a case-by-case basis (i.e., based on resampling from $\{y_1, \dots, y_m\}$). Let E_* represent the expectation with respect to the bootstrap distribution (i.e., with respect to the distribution of Y^*). Let $\{\hat{\theta}^*(i), i = 1, \dots, W\}$ represent a set of W bootstrap replicates of $\hat{\theta}$. We use $\hat{\theta}$ to denote the maximum likelihood estimator of θ based on maximizing $L(\theta | Y)$. Accordingly, $\hat{\theta}^*$ represents the maximum likelihood estimator of θ based on maximizing $L(\theta | Y^*)$.

The bootstrap sample size is taken to be the same as the size of the observed sample Y (i.e., m). The properties of the bootstrap when the bootstrap sample size is equal to the original sample size are discussed by Efron and Tibshirani (1993).

Let

$$\begin{aligned} \text{b1} &= E_*\{-2 \log L(\hat{\theta}^* | Y) - \{-2 \log L(\hat{\theta}^* | Y^*)\}\}, \text{ and} \\ \text{b2} &= 2E_*\{-2 \log L(\hat{\theta}^* | Y) - \{-2 \log L(\hat{\theta} | Y)\}\}. \end{aligned} \quad (2.5)$$

In the Appendix, we establish that under suitable conditions, b1 and b2 in (2.5) serve as consistent estimators of the bias adjustment (2.3), and are asymptotically equivalent.

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

$$\begin{aligned} \frac{1}{W} \sum_{i=1}^W -2 \log L(\hat{\theta}^*(i) | Y) &\rightarrow E_*\{-2 \log L(\hat{\theta}^* | Y)\} \text{ a.s.}, \text{ and} \\ \frac{1}{W} \sum_{i=1}^W -2 \log L(\hat{\theta}^*(i) | Y^*(i)) &\rightarrow E_*\{-2 \log L(\hat{\theta}^* | Y^*)\} \text{ a.s.} \end{aligned} \quad (2.6)$$

Expressions (2.5) and (2.6) therefore lead us to the following large-sample estimators of $\Delta(k, \theta_o)$:

$$\begin{aligned} \text{AICb1} &= -2 \log L(\hat{\theta} | Y) + \frac{1}{W} \sum_{i=1}^W -2 \log \frac{L(\hat{\theta}^*(i) | Y)}{L(\hat{\theta}^*(i) | Y^*(i))}, \text{ and} \\ \text{AICb2} &= -2 \log L(\hat{\theta} | Y) + 2 \left\{ \frac{1}{W} \sum_{i=1}^W -2 \log \frac{L(\hat{\theta}^*(i) | Y)}{L(\hat{\theta} | Y)} \right\}. \end{aligned} \quad (2.7)$$

Note that AICb1 and AICb2 are composed of two terms. The “goodness of fit” term, $-2 \log L(\hat{\theta} | Y)$, estimates the first term in (2.4). The “penalty” terms in AICb1 and AICb2, the second terms in (2.7), estimate the bias expression (2.3).

We note that the criterion AICb1 was originally introduced by Efron (1983, 1986). This criterion was named WIC by Ishiguro and Sakamoto (1991), who further promoted its use. The criterion AICb2 was originally introduced as AICb by Cavanaugh and Shumway (1997) in the context of Gaussian state-space model selection. Shibata (1997) considered AICb1 and AICb2 in addition to three additional bootstrap-corrected AIC variants, AICb3 – AICb5. Our simulation results have indicated that AICb4 has similar bias properties as AICb2, yet performs less effectively at selecting a model of appropriate form. AICb3 often exhibits similar selection patterns as AICb1, yet has inferior bias properties. AICb5 is characterized by both poor bias properties (comparable to those of AICb3) and deficient selection tendencies. Thus, we limit our consideration to only AICb1 and AICb2.

In the Appendix, we extend the development of Shibata (1997) to provide an asymptotic justification of AICb1 and AICb2 for candidate models of the form (2.1). In our justification, two points must be emphasized.

First, unlike the asymptotic justification of AIC and the small-sample validations of AICc, our asymptotic justification of the bootstrap criteria does not require the assumption that the candidate model subsumes the true model. Thus, our justification applies to candidate models which are underspecified as well as to those which are correctly or overspecified.

Second, our justification of the bootstrap criteria holds regardless of normality. Even though normality is referenced in presenting the candidate models via (2.1), our theoretical development does not require this assumption. We have therefore investigated the behavior of AICb1 and AICb2 under nonparametric and semiparametric bootstrapping, where normality is not employed, along with parametric bootstrapping, where normality is utilized in generating the bootstrap sample. However, for the sake of brevity, only parametric boot-

strapping results are presented in our simulation study. Results based on semiparametric and nonparametric bootstrapping are briefly described.

We encourage the use of the bootstrap criteria in two instances: (1) when the sample size is small enough to cast doubt on the efficacy of AIC, or (2) when bootstrapping is used as part of the overall analysis in a mixed modeling application. In small-sample settings, we will illustrate the improved performance of the bootstrap criteria over AIC in our simulation study.

3. Description of Simulations

Consider a setting in which data arises from a generating model of the following form:

$$Y = X_o\beta_o + Z_o\tau_o + \varepsilon_o, \quad (3.1)$$

where Y denotes the $N \times 1$ response vector $(y_1', \dots, y_m')'$; X_o is an $N \times (p_o + 1)$ design matrix of full column rank; Z_o is an $N \times m$ block diagonal design matrix comprised on m blocks, where each block is an $n_i \times 1$ vector consisting of all 1's; β_o is a $(p_o + 1) \times 1$ fixed effects parameter vector; τ_o is a $m \times 1$ random effects vector; and ε_o is an $N \times 1$ error vector. Assume $\tau_o \sim N(0, \sigma_{\tau_o}^2 I)$ and $\varepsilon_o \sim N(0, \sigma_o^2 I)$, with τ_o and ε_o distributed independently. The covariance matrix of Y is given by $V_o = Z_o Z_o' \sigma_{\tau_o}^2 + \sigma_o^2 I$.

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 $p_o + 3$.

Assume that two types of candidate models are considered for modeling data Y arising from model (3.1). First, we entertain a mixed model having the same covariance structure as the generating model: i.e., the ‘‘compound symmetric’’ covariance structure. Such a candidate model can be represented in the format of model (2.1) by writing

$$Y = X\beta + Z\tau + \varepsilon, \quad (3.2)$$

where Y is as defined as previously; X is an $N \times (p + 1)$ design matrix of full column rank; Z is an $N \times m$ block diagonal design matrix having the same format as Z_o in (3.1); β is a

$(p + 1) \times 1$ fixed effects parameter vector; τ is a $m \times 1$ random effects vector; and ε is an $N \times 1$ error vector. We assume $\tau \sim N(0, \sigma_\tau^2 I)$ and $\varepsilon \sim N(0, \sigma^2 I)$, with τ and ε distributed independently. 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 model is the traditional normal linear regression model: i.e., a fixed effects model that treats within-case correlations as zero. Such a model is a special case of model (3.2), and can be written as

$$Y = X\beta + \varepsilon. \tag{3.3}$$

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 $\hat{V} = \hat{\sigma}^2 I$.

In the simulation setting outlined, we assume that data arises from the mixed model (3.1) having a compound symmetric covariance structure. We consider modeling such data using candidate models both with and without the appropriate covariance structure. Model selection requires the determination of a suitable set of regressors for the design matrix X , and a decision as to whether the random effects should be included (as in (3.2)) or excluded (as in (3.3)). For the specification of the regressor set, supposing we have P explanatory variables of interest, we might consider candidate models of the form (3.2) and (3.3) corresponding to design matrices X of ranks $p + 1$, where $p = 1, 2, \dots, P$. The smallest models ($p = 1$) would be based only one regressor, and the largest model ($p = P$) would contain all P regressors.

To envision how such a setting might arise in practice, consider a biomedical study in which data is collected on patients in a multi-center clinical trial. A response variable and a

collection of covariates are measured on each patient, and the biostatistician must determine which covariates are appropriate for inclusion in a model designed to explain or predict the response. Additionally, since groups of patients are associated with multiple centers, the inclusion of a random center effect should be investigated. The random effect would account for clustering that may occur among responses from the same center. If the random effect is included, the model treats responses for patients within the same center as correlated and responses for patients in different centers as independent; if the random effect is excluded, the model treats all responses as independent.

In our simulations, we examine the behavior of AIC, AICb1, and AICb2 in settings where the criteria are used to select a fitted model from the candidate class. In each set, 100 samples are generated from the true model. For every sample, candidate models of the form (3.2) and (3.3) are fit to the data, the criteria are evaluated, and the fitted model favored by each criterion is recorded. Over the 100 samples, the distribution of model selections is tabulated for each of the criteria.

For a given set of estimates $\hat{\theta}$, the discrepancy $d(\hat{\theta}, \theta_o)$ is evaluated via

$$d(\hat{\theta}, \theta_o) = \log |\hat{V}| + \text{tr}(\hat{V}^{-1}V_o) + (X_o\beta_o - X\hat{\beta})'\hat{V}^{-1}(X_o\beta_o - X\hat{\beta}).$$

Since AIC and the bootstrap variants of AIC serve as proxies for $d(\hat{\theta}, \theta_o)$, the performance of $d(\hat{\theta}, \theta_o)$ as a selection rule serves as an appropriate “gold standard” for assessing the performance of the criteria. Averaging values of $d(\hat{\theta}, \theta_o)$ over various samples allows us to evaluate $\Delta(k, \theta_o)$, the expected discrepancy.

In configuring our simulations, we require that one of the candidate models in our class is correctly specified, i.e., contains the same regressors and features the same covariance structure as the true model. We randomly generate all regressors as independent, identically distributed variates from a standard normal distribution.

Our simulation study consists of two parts. In the first part, the candidate models are based on nested design matrices X . Thus, with P regressor variables of interest, we entertain

P candidate models based on a sequence of design matrices X of ranks $2, 3, \dots, (P+1)$. 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. For the candidate models of the form (3.2), the model of order p_o is correctly specified ($1 \leq p_o \leq P$). Fitted models for which $p < p_o$ are underfit, and those for which $p > p_o$ are overfit.

In the second part, the candidate models are based on design matrices X corresponding to all possible subsets of the regressor variables. With P regressor variables of interest, a total of $2^P - 1$ design matrices may be constructed. (Intercept-only models are excluded.) We again entertain candidate models based on design matrices X of ranks $2, 3, \dots, (P+1)$; however, for each regressor subset size p ($p = 1, 2, \dots, P$), we must consider $\binom{P}{p}$ design matrices representing various combinations of p regressors. For the candidate models of the form (3.2), one of the models having a design matrix X based on p_o regressors is correctly specified ($1 \leq p_o \leq P$). Fitted models corresponding to design matrices that do not contain all of the regressors in the true model are underfit. For such models, the column space of the design matrix for the fitted model, $C(X)$, does not contain the column space of the design matrix for the true model, $C(X_o)$; i.e., $C(X_o) \not\subseteq C(X)$. Fitted models corresponding to design matrices for which $p > p_o$ that contain all of the regressors in the true model are overfit. For such models, the column space of the design matrix for the true model is a proper subset of the column space of the design matrix for the fitted model; i.e., $C(X_o) \subset C(X)$.

For practical applications, the all possible regressions (APR) framework is more realistic than the nested models (NM) framework. However, the latter setting is 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, 1998.) When P is large and the sample size is relatively small, criteria with penalty terms justified asymptotically are often outperformed by variants with

penalty terms refined for finite samples. Thus, simulations based on nested models may allow us to better assess the efficacy of the bootstrap-corrected criteria.

Our goal is to search among a class of candidate models of the form (3.2) and (3.3) for the fitted model which serves as the best approximation to the true model (3.1). Since the true model is included in the candidate class, and since $\Delta(k, \theta_o)$ is minimized when the structure of the candidate model corresponds to that of the true model, the optimal fitted candidate model is correctly specified. We investigate the effectiveness of AIC, AICb1, and AICb2 at choosing this optimal fitted candidate model.

In the computation of the penalty terms for AICb1 and AICb2, the parametric bootstrap is employed. Simulation results for semiparametric and nonparametric bootstrapping have been compiled, and will be briefly described for the sets based on nested models. However, such results are not featured here.

In the context of the mixed model (3.2), the algorithm for parametric bootstrap can be outlined as follows.

Let y_i denote the $n_i \times 1$ response vector for case i , let X_i denote the $n_i \times p$ design matrix for case i , and let z_i denote an $n_i \times 1$ vector consisting of all 1's.

- (1) Fit the candidate mixed model (3.2) to the data to obtain the estimators $\hat{\beta}$, $\hat{\sigma}_\tau^2$, and $\hat{\sigma}^2$.
- (2) Generate the bootstrap sample on a case-by-case using the fitted model

$$y_i^* = X_i \hat{\beta} + z_i \tau_i^* + \varepsilon_i^*, \quad i = 1, \dots, m,$$

where τ_i^* and ε_i^* are generated from $N(0, \hat{\sigma}_\tau^2)$ and $N(0, \hat{\sigma}^2 I)$ distributions, respectively.

- (3) Fit the candidate mixed model (3.2) to the bootstrap data, thereby obtaining the bootstrap MLE's $\hat{\beta}^*$, $\hat{\sigma}_\tau^{2*}$, and $\hat{\sigma}^{2*}$.
- (4) Repeat steps (2)-(3) W times.

Note that the same algorithm may be applied in the context of the traditional regression model (3.3), yet the $z_i\tau_i^*$ model term and its associated variance component $\hat{\sigma}_\tau^2$ are omitted. For our simulation sets, the EM algorithm was employed as a fitting procedure when the algorithm was applied to the mixed model (3.2).

4. Presentation of Simulation Results

4.1 Nested Models

To generate the simulated data, we choose the parameters $\beta_o = (1, 1, 1, 1, 1, 1, 1)'$, $\sigma_{\tau_o}^2 = 2$, and $\sigma_o^2 = 1$ in model (3.1). We consider sample sizes of $m = 15, 20, 30$, 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 $P = 12$. In the computation of AICb1 and AICb2, $W = 500$ bootstrap replicates are used.

For each simulation set ($m = 15, 20, 30, 50$), the distributions of selections by AIC, AICb1, AICb2, and the discrepancy $d(\hat{\theta}, \theta_o)$ are compiled over the 100 samples.

For AIC, AICb1, and AICb2, over the 100 samples, the average criterion value is computed for each of the candidate models (3.2) and (3.3) over the orders 1 through P . The value of $\Delta(k, \theta_o)$ is approximated by averaging values of $d(\hat{\theta}, \theta_o)$. To explore the effectiveness of the criteria as asymptotically unbiased estimators of $\Delta(k, \theta_o)$, for each candidate model type, we plot $\Delta(k, \theta_o)$ along with the averages for AIC, AICb1, and AICb2 against the orders from 1 to P .

The order selections for AIC, AICb1, AICb2, and $d(\hat{\theta}, \theta_o)$ are reported in Tables 4.1–4.4. Over all four sets, AICb2 obtains the most correct model selections. In the sets where the sample size is small ($m = 15$ or $m = 20$) or moderate ($m = 30$), AICb1 and AICb2 both outperform AIC as a selection criterion. However, in the set where the sample size is large ($m = 50$), only AICb2 outperforms AIC in choosing the correct model. In this set, AICb1

and AIC obtain a comparable number of correct model selections, although AICb1 tends to choose more parsimonious models.

Figures 4.1–4.4 demonstrate how effectively the bootstrap criteria serve as approximately unbiased estimators of $\Delta(k, \theta_o)$. With an increase in sample size, the average curves for AICb1 and AICb2 tend to grow closer, both approaching the simulated $\Delta(k, \theta_o)$ curve. The figures illustrate the large-sample theory derived in the Appendix: the bootstrap criteria are not only asymptotically unbiased estimators of $\Delta(k, \theta_o)$, but they are also asymptotically equivalent.

For correctly specified or overfit models, the average AICb1 curve follows the simulated $\Delta(k, \theta_o)$ curve more closely than either the average AIC or AICb2 curve. The figures also reveal that AICb1 and AICb2 are less biased estimators of $\Delta(k, \theta_o)$ than AIC.

As mentioned previously, the preceding results are based on parametric bootstrapping. We have also compiled results under semiparametric and nonparametric bootstrapping. With semiparametric bootstrapping, the model selections for AICb1 and AICb2 are similar to those obtained for parametric bootstrapping. When figures such as 4.1–4.4 are constructed based on the criterion averages, the curves for AICb1 and AICb2 reflect the general shape of the simulated $\Delta(k, \theta_o)$ curve. However, in terms of location, the AICb1 and AICb2 curves are separated from one another and from the $\Delta(k, \theta_o)$ curve. A possible explanation for this tendency is provided by the results of Morris (2002), who demonstrated both mathematically and by simulation that the semiparametric bootstrap for mixed models (based on resampling BLUP's) will consistently underestimate the variation of the parameter estimates.

With nonparametric bootstrapping, the model selections for AICb1 and AICb2 are again similar to those obtained for parametric bootstrapping. In small-sample settings, however, unusually large criterion values occasionally result from certain samples. Thus, when figures such as 4.1–4.4 are constructed based on the criterion averages, the curves for AICb1 and AICb2 are quite dissimilar from the simulated $\Delta(k, \theta_o)$ curve when the sample size is small

(e.g., $m = 15$). The presence of the atypically large criterion values implies that nonparametric bootstrapping in small-sample applications may not lead to accurate estimators of $\Delta(k, \theta_o)$.

4.2 All Possible Regressions

To generate the simulated data, we choose the parameters $\beta_o = (1, 1, 1, 1)'$, $\sigma_{\tau_o}^2 = 2$, and $\sigma_o^2 = 1$ in model (3.1). We again consider sample sizes of $m = 15, 20, 30$, 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 containing $p_o = 3$ regressor variables. The largest model in the candidate class is based on $P = 5$ regressors. In the computation of AICb1 and AICb2, $W = 500$ bootstrap replicates are used.

Again, for each simulation set ($m = 15, 20, 30, 50$), the distributions of selections by AIC, AICb1, AICb2, and the discrepancy $d(\hat{\theta}, \theta_o)$ are compiled over the 100 samples. The parametric bootstrap is employed.

The model selections for AIC, AICb1, AICb2, and $d(\hat{\theta}, \theta_o)$ are reported in Tables 4.5–4.8. The selections are grouped according to (a) whether the chosen model includes or excludes the random effects, and (b) whether the mean structure is correctly specified ($C(X_o) = C(X)$), underspecified ($C(X_o) \not\subseteq C(X)$), or overspecified ($C(X_o) \subset C(X)$).

Again, over all four sets, AICb2 obtains the most correct model selections. In the set based on the smallest sample size ($m = 15$), AICb1 and AICb2 both outperform AIC as a selection criterion. However, in the remaining sets ($m = 20, 30, 50$), AICb1 and AIC obtain a comparable number of correct model selections. Thus, the selection patterns in the APR setting are similar to those in the nested model setting, although the propensity for AIC to choose an overfit model is reduced in the APR sets because fewer models in the candidate class are overspecified in the mean structure.

We close this section by commenting on how to choose the number of bootstrap samples W used in the evaluation of the bootstrap criteria. As W increases, the averages which

comprise the penalty terms of the bootstrap criteria stabilize. Choosing a value of W which is too small may result in inaccurate estimation of the bias expression (2.3), yet choosing a value of W which is too large will waste computational time. The value 500 for W was chosen since smaller values seemed to marginally diminish the number of correct model selections for the bootstrap criteria while larger values did not appreciably improve the performance of the bootstrap criteria.

5. Conclusion and Further Directions

We have focused on model selection for the general linear mixed model. Under suitable conditions, the bootstrap criteria AICb1 and AICb2 serve the same objective as traditional AIC, in that they provide asymptotically unbiased estimators of the expected discrepancy $\Delta(k, \theta_o)$ between the generating model and a fitted approximating model. The two bootstrap criteria are asymptotically equivalent.

Our simulation results indicate that AICb1 and AICb2 perform effectively in choosing a mixed model with an appropriate mean and covariance structure. AICb2 exhibits a higher success rate in identifying the correct model than either AIC or AICb1. In small-sample applications, both bootstrap criteria outperform AIC in selecting the correct model.

Akaike criteria are designed to serve as estimators of the expected Kullback-Leibler discrepancy (2.4). Figures 4.1–4.4 illustrate the bias characteristics of AIC, AICb1, and AICb2 as estimators of $\Delta(k, \theta_o)$. Our investigations indicate that the criteria are quite comparable in terms of variance; hence, the accuracy of the criteria as estimators of $\Delta(k, \theta_o)$ is largely governed by their bias properties. In small to moderate sample-size settings, the bias of AICb1 is substantially less than that of AIC, which tends to underestimate $\Delta(k, \theta_o)$ for correctly specified and overfit models. The bias of AICb1 is also less than that of AICb2, which tends to overestimate $\Delta(k, \theta_o)$ for such models. Thus, as an estimator of $\Delta(k, \theta_o)$, AICb1 is superior to AIC and AICb2 in terms of accuracy. However, the tendency for AICb2 to marginally overestimate $\Delta(k, \theta_o)$ arises from a penalty term that imposes a greater penal-

ization for overfitting than the bias adjustment (2.3) dictates. As a result, the difference between AICb2 values for an overfit model and the correctly specified fitted model tends to exceed the difference between AICb1 values or AIC values. Thus, AICb2 tends to be more sensitive towards detecting overspecification than either of its competitors.

AICb1 and AICb2 can be justified in the context of a general model formulation under a nonrestrictive set of conditions. Our justification and simulations focus on the framework of the mixed model, yet the criteria have potential applicability in a large array of practical modeling frameworks. Also, although AICb1 and AICb2 are more computationally expensive to evaluate than AIC, they have simplistic forms, and should be convenient to compute as part of an overall bootstrap-based analysis.

Apart from their utility in small-sample applications, we also hope to investigate other advantages of using bootstrap criteria. We wish to further explore the effectiveness of the criteria in selecting an appropriate covariance structure for the random effects, even when the true covariance structure is not represented among the models in the candidate class. We wish to also develop bootstrap-based procedures for constructing confidence intervals for the expected discrepancy $\Delta(k, \theta_o)$. Such intervals could be used to determine whether values of $\Delta(k, \theta_o)$ statistically differ; point estimates of $\Delta(k, \theta_o)$ alone cannot address this issue.

Acknowledgments

The authors would like to express their appreciation to the referees for providing thoughtful and insightful comments which helped to improve the original version of this manuscript.

Table 4.1: Model Selections for Simulation Set 1 ($m = 15$)

Model Structure		Selections				
Random	Mean					
Effects	Order	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2	
With	1-4	0	1	0	0	
	5	0	7	1	1	
	6	56	84	78	81	
	7	6	5	7	3	
	8	7	2	3	3	
	9	5	0	1	1	
	10	6	0	1	0	
	11	7	0	0	0	
	12	10	0	0	0	
	Without	1-5	0	0	0	0
		6	0	1	6	8
		7	0	0	2	2
8-12		3	0	1	1	

Table 4.2: Model Selections for Simulation Set 2 ($m = 20$)

Model Structure		Selections			
Random	Mean				
Effects	Order	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	1-5	0	1	0	0
	6	62	94	77	83
	7	11	3	13	10
	8	7	2	5	3
	9	5	0	0	0
	10	5	0	1	0
	11	1	0	0	0
	12	7	0	0	0
Without	1-12	2	0	4	4

Table 4.3: Model Selections for Simulation Set 3 ($m = 30$)

Model Structure		Selections			
Random	Mean				
Effects	Order	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	1-5	0	0	0	0
	6	70	97	77	88
	7	12	1	15	7
	8	7	1	6	3
	9	2	1	0	1
	10	3	0	1	0
	11	3	0	0	0
	12	3	0	1	1
Without	1-12	0	0	0	0

Table 4.4: Model Selections for Simulation Set 4 ($m = 50$)

Model Structure		Selections			
Random	Mean				
Effects	Order	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	1-5	0	0	0	0
	6	74	95	72	81
	7	8	4	15	10
	8	9	0	6	5
	9	3	1	3	1
	10	1	0	1	1
	11	3	0	2	2
	12	2	0	1	0
Without	1-12	0	0	0	0

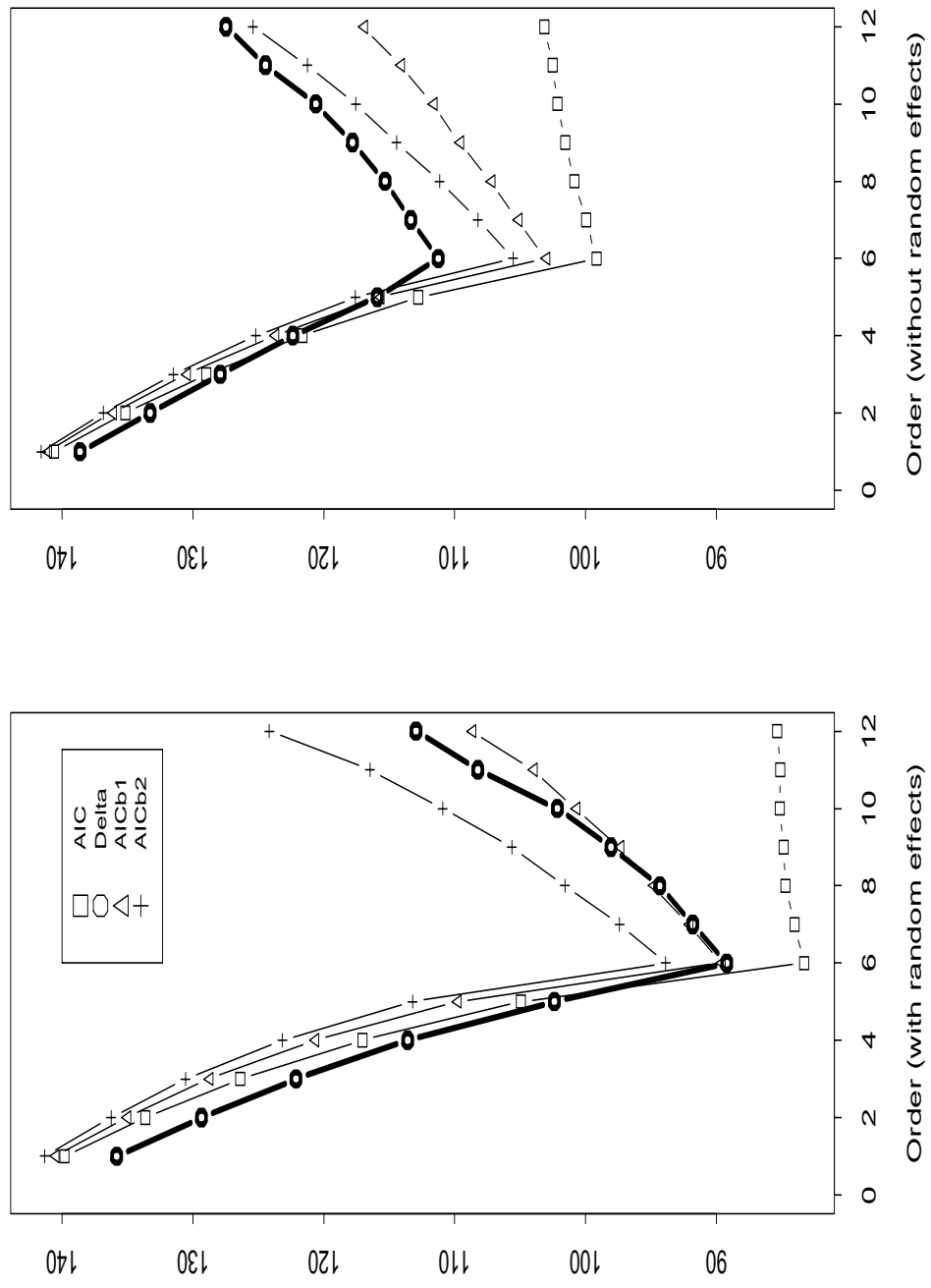


Figure 4.1: Criterion Averages and Simulated $\Delta(k, \theta_o)$ (Set 1, $m = 15$)

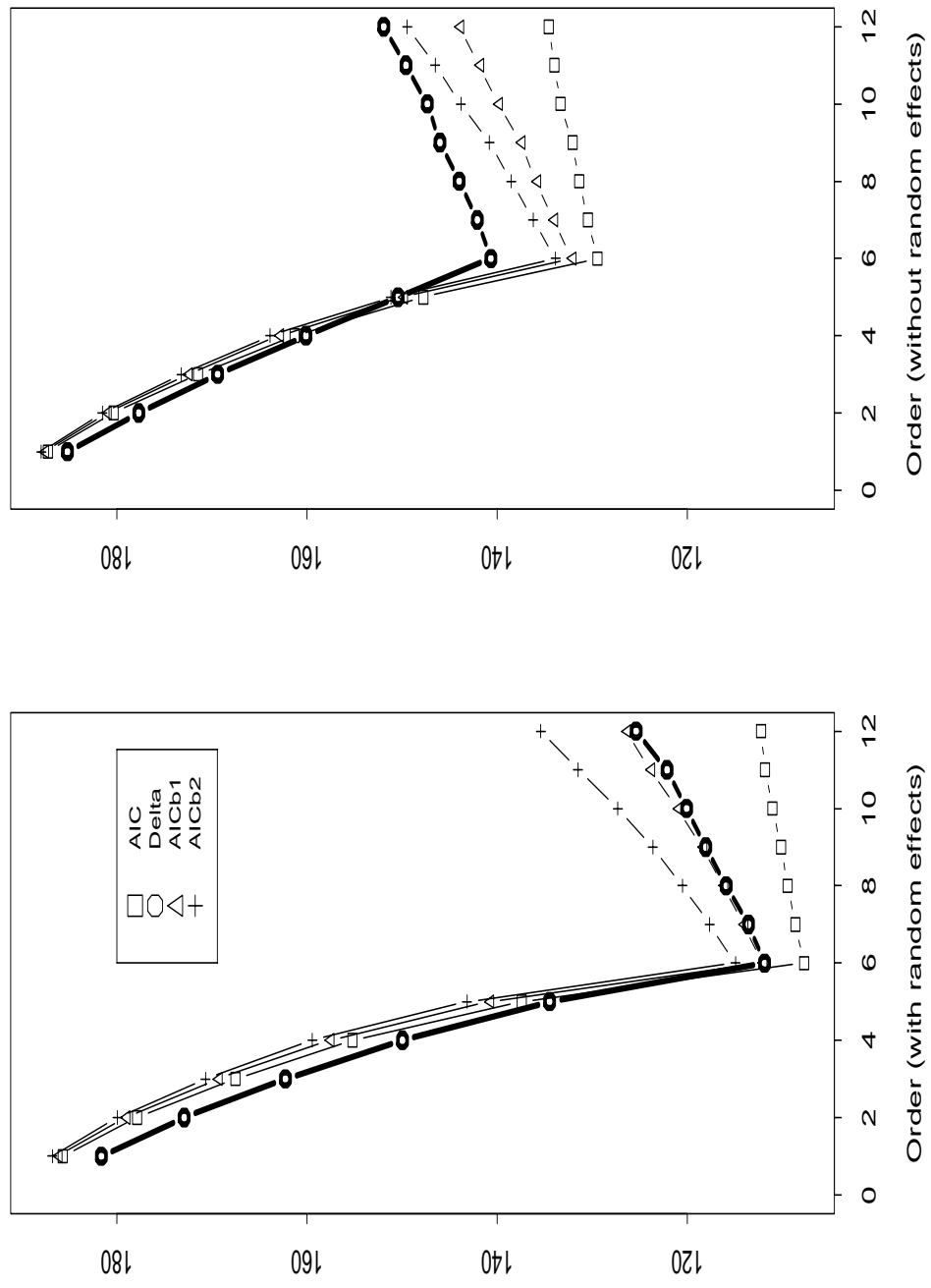


Figure 4.2: Criterion Averages and Simulated $\Delta(k, \theta_o)$ (Set 2, $m = 20$)

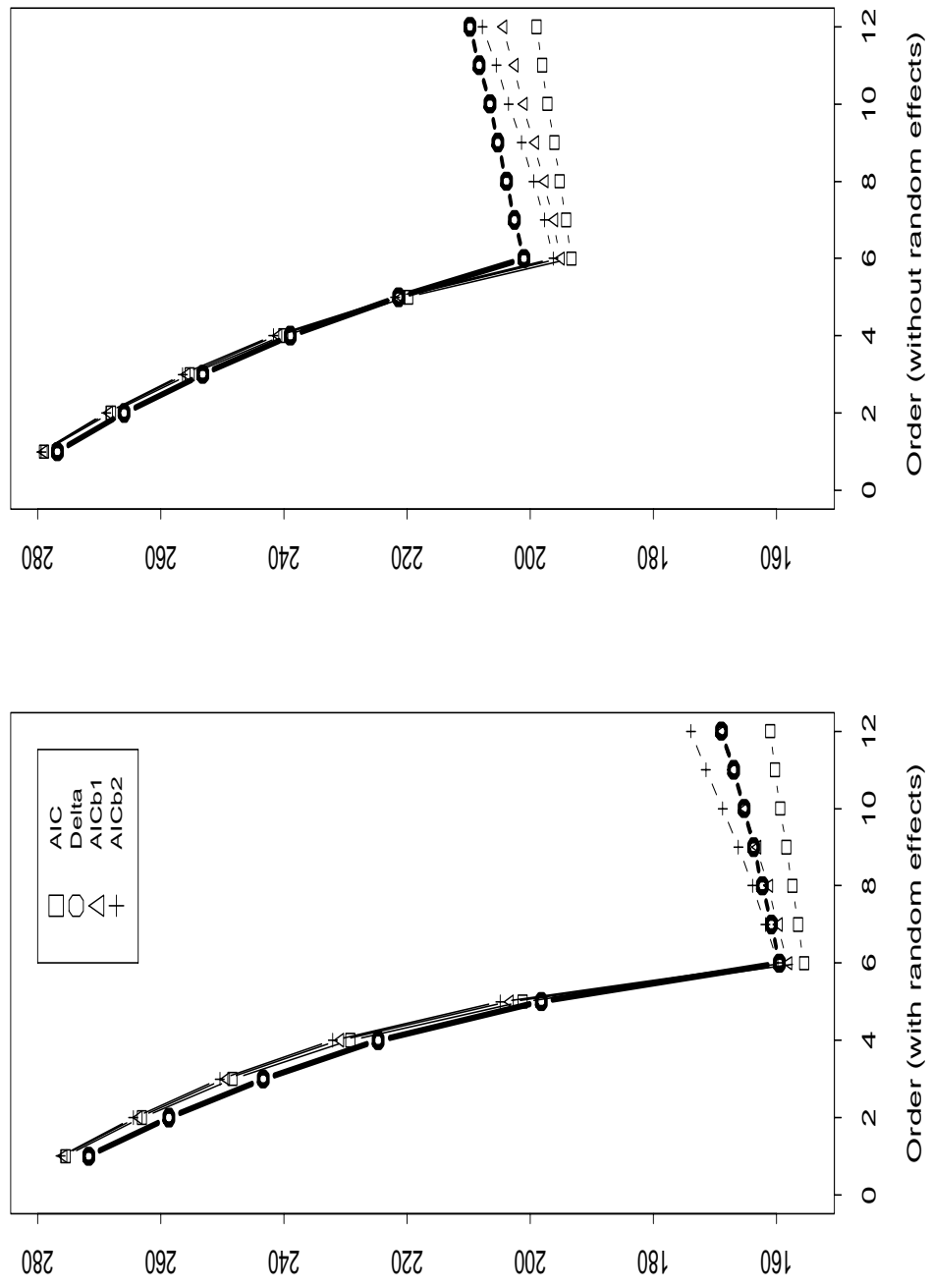


Figure 4.3: Criterion Averages and Simulated $\Delta(k, \theta_o)$ (Set 3, $m = 30$)

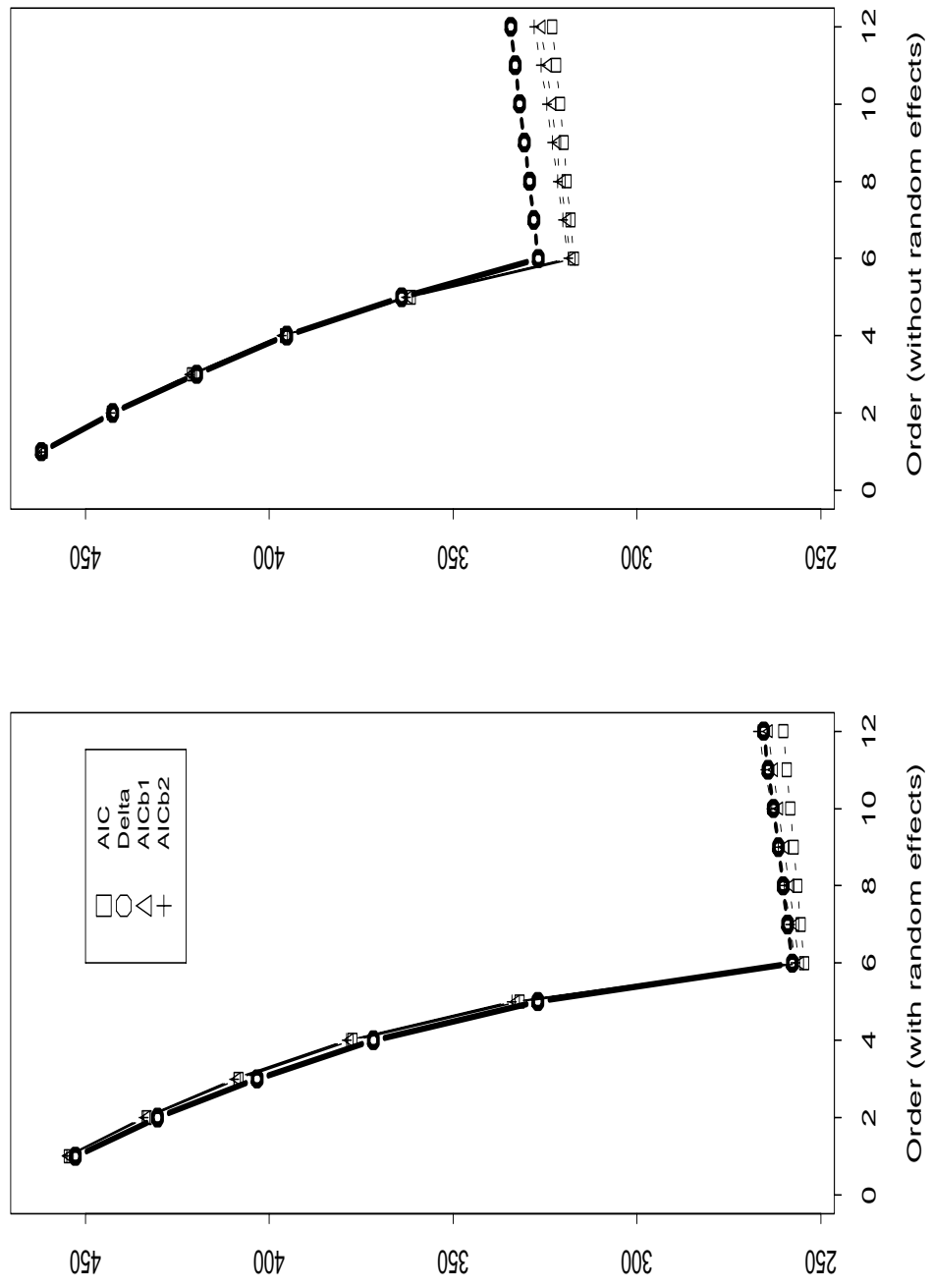


Figure 4.4: Criterion Averages and Simulated $\Delta(k, \theta_o)$ (Set 4, $m = 50$)

Table 4.5: Model Selections for Simulation Set 5 ($m = 15$)

Model Structure		Selections			
Random Effects	Mean Structure	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	Underfit	0	1	0	0
Without		0	0	0	0
With	Correctly Specified	67	83	78	83
Without		1	0	0	3
With	Overfit	32	16	20	13
Without		0	0	2	1

Table 4.6: Model Selections for Simulation Set 6 ($m = 20$)

Model Structure		Selections			
Random Effects	Mean Structure	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	Underfit	0	0	0	0
Without		0	0	0	0
With	Correctly Specified	73	87	73	79
Without		0	0	0	1
With	Overfit	27	13	27	20
Without		0	0	0	0

Table 4.7: Model Selections for Simulation Set 7 ($m = 30$)

Model Structure		Selections			
Random Effects	Mean Structure	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	Underfit	0	0	0	0
Without		0	0	0	0
With	Correctly Specified	74	93	74	79
Without		0	0	0	0
With	Overfit	26	7	26	21
Without		0	0	0	0

Table 4.8: Model Selections for Simulation Set 8 ($m = 50$)

Model Structure		Selections			
Random Effects	Mean Structure	AIC	$d(\hat{\theta}, \theta_o)$	AICb1	AICb2
With	Underfit	0	0	0	0
Without		0	0	0	0
With	Correctly Specified	74	92	71	81
Without		0	0	0	0
With	Overfit	26	8	29	19
Without		0	0	0	0

Appendix

Here, we present a formal justification of AICb1 and AICb2 as asymptotically unbiased estimators of $\Delta(k, \theta_o)$.

We first establish that b1 and b2 are asymptotically equivalent (as both $W \rightarrow \infty$ and $m \rightarrow \infty$).

Let Θ represent the parameter space for θ , the set of parameters for a candidate model. Let $\bar{\theta}$ represent the θ corresponding to the global maximum of $E_o\{\log L(\theta | Y)\}$, or equivalently, the θ for which $E_o\{-2 \log L(\theta | Y)\}$ is minimized. We assume that $\bar{\theta}$, the “pseudo” true parameter, exists and is unique.

To prove the asymptotic equivalence of b1 and b2, we must establish the consistency of both $\hat{\theta}$ and $\hat{\theta}^*$. Here, consistency means that the estimator converges to $\bar{\theta}$ almost surely as the sample size m approaches infinity.

We establish consistency by presenting a set of fundamental conditions that will allow us to appeal to Lemma 1 of Shibata (1997).

Assumption 1

- (i) The parameter space Θ is a compact subset of k -dimensional Euclidean space.
- (ii) Derivatives of the log likelihood up to order three exist with respect to θ , and are continuous and bounded over Θ .
- (iii) $\bar{\theta}$ is an interior point of Θ .

Regarding the likelihood ratio, we provide relevant notation. Let $f_i(y_i | \theta)$ denote the marginal density for case i , and consider the log-likelihood ratio statistic defined for a neighborhood U in Θ by

$$R_i(y_i, \theta, U) = \inf_{\tilde{\theta} \in U} \log \frac{f_i(y_i | \theta)}{f_i(y_i | \tilde{\theta})}.$$

We assume that the limit

$$\bar{I}(\bar{\theta}, U) = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m E_o\{R_i(y_i, \bar{\theta}, U)\}$$

exists and is finite in a neighborhood $U = U_\theta$ for any θ in Θ .

From the Lebesgue monotone convergence theorem,

$$\lim_{q \rightarrow \infty} \bar{I}(\bar{\theta}, U_\theta^{(q)}) = \bar{I}(\bar{\theta}, \theta) = \lim_{m \rightarrow \infty} \frac{1}{m} E_o \{ \log L(\bar{\theta} | Y) - \log L(\theta | Y) \} \quad (\text{A.1})$$

holds true for a monotone decreasing sequence of neighborhoods $U_\theta^{(q)}$, $q = 1, 2, \dots$, converging to a parameter θ . The preceding is valid provided that $\log L(\theta | Y)$ is continuous with respect to θ , that is,

$$\lim_{\bar{\theta} \rightarrow \theta} \{ \log L(\bar{\theta} | Y) \} = \log L(\theta | Y)$$

for any $\theta \in \Theta$. Note that the right hand side of (A.1) is nonnegative by the definition of $\bar{\theta}$.

We use analogous notation for the bootstrap sample $Y^* = (y_1^*, \dots, y_m^*)'$. We have

$$\bar{I}_B(\bar{\theta}, U) = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m E_o E_* \{ R_i(y_i^*, \bar{\theta}, U) \}$$

and

$$\lim_{q \rightarrow \infty} \bar{I}(\bar{\theta}, U_\theta^{(q)}) = \bar{I}_B(\bar{\theta}, \theta) = \lim_{m \rightarrow \infty} \frac{1}{m} E_o E_* \{ \log L(\bar{\theta} | Y^*) - \log L(\theta | Y^*) \}.$$

Assumption 2

(i) Both $\frac{1}{m} \sum_{i=1}^m R_i(y_i, \bar{\theta}, U_\theta)$ and $\frac{1}{m} \sum_{i=1}^m R_i(y_i^*, \bar{\theta}, U_\theta)$ almost surely converge to $\bar{I}(\bar{\theta}, U_\theta)$ and $\bar{I}_B(\bar{\theta}, U_\theta)$, respectively, in a neighborhood U_θ for any $\theta \in \Theta$.

(ii) $\bar{I}(\bar{\theta}, \theta) > 0$ and $\bar{I}_B(\bar{\theta}, \theta) > 0$ for any $\theta \in \Theta$ where $\theta \neq \bar{\theta}$.

The assumption (i) means that the average of the log-likelihood ratio statistics has a limit in a neighborhood $U_{\bar{\theta}}$ of $\bar{\theta}$. The assumption (ii) is an identifiability condition.

Assumptions 1 and 2 allow us to apply Lemma 1 of Shibata (1997). Thus, we can assert that $\hat{\theta}$ and $\hat{\theta}^*$ almost surely converge to $\bar{\theta}$ as m tends to infinity. For any neighborhood $U_{\bar{\theta}}$ of $\bar{\theta}$, the estimators $\hat{\theta}$ and $\hat{\theta}^*$ lie in the neighborhood for sufficiently large m .

Assumption 3

$E_*\{\log L(\hat{\theta} | Y^*)\} = \log L(\hat{\theta} | Y)$, where $\hat{\theta}$ is the MLE based on maximizing $\log L(\theta | Y)$.

We explore this assumption in the context of the general linear model (2.1).

With parametric bootstrapping, a bootstrap sample Y^* is produced via the formula

$$Y^* = X\hat{\beta} + \xi^*,$$

where $\hat{\beta}$ is the maximum likelihood estimate of β , and the bootstrap residuals ξ^* are generated from a normal distribution with mean vector 0 and covariance matrix $\hat{V} = Z\hat{D}Z' + \hat{\sigma}^2R$. Note that \hat{V} is the maximum likelihood estimate of V under model (2.1).

Neglecting the constant in $\log L(\hat{\theta} | Y^*)$, we have

$$\begin{aligned} E_*\{\log L(\hat{\theta} | Y^*)\} &= E_*\left\{-\frac{1}{2}\log|\hat{V}| - \frac{1}{2}(Y^* - X\hat{\beta})'\hat{V}^{-1}(Y^* - X\hat{\beta})\right\} \\ &= -\frac{1}{2}\log|\hat{V}| - \frac{1}{2}E_*\|\hat{V}^{-\frac{1}{2}}\xi^*\|^2 \\ &= -\frac{1}{2}\log|\hat{V}| - \frac{1}{2}\text{tr}(\hat{V}^{-1}\hat{V}) \\ &= -\frac{1}{2}\log|\hat{V}| - \frac{1}{2}N. \end{aligned}$$

Recall that N denotes the total number of observations in the vector Y .

Similarly,

$$\log L(\hat{\theta} | Y) = -\frac{1}{2}\log|\hat{V}| - \frac{1}{2}(Y - X\hat{\beta})'\hat{V}^{-1}(Y - X\hat{\beta}).$$

From Christensen (1996, pp. 271-272), for the MLE's \hat{V} and $\hat{\beta}$, we have

$$(Y - X\hat{\beta})'\hat{V}^{-1}(Y - X\hat{\beta}) = N.$$

Thus,

$$\log L(\hat{\theta} | Y) = -\frac{1}{2}\log|\hat{V}| - \frac{1}{2}N.$$

Therefore, Assumption 3 holds true under parametric bootstrapping.

For semiparametric and nonparametric bootstrapping, we can verify that the relation $E_*\{\log L(\theta | Y^*)\} = \log L(\theta | Y)$ holds for any $\theta \in \Theta$, although this demonstration is nontrivial. Thus, Assumption 3 holds as a special case where $\theta = \hat{\theta}$.

Let $\hat{\mathcal{J}}(Y, \theta) = -\frac{\partial^2}{\partial\theta\partial\theta'}\{\log L(\theta | Y)\}$ and $\hat{\mathcal{J}}(Y^*, \theta) = -\frac{\partial^2}{\partial\theta\partial\theta'}\{\log L(\theta | Y^*)\}$ denote the observed Fisher information matrices under the original sample Y and under the bootstrap sample Y^* , respectively. We now present and discuss our fourth and final assumption.

Assumption 4

$\hat{\mathcal{J}}(Y, \theta)/m$ and $\hat{\mathcal{J}}(Y^*, \theta)/m$ almost surely converge to positive definite matrices $\bar{\mathcal{J}}(\theta)$ and $\bar{\mathcal{J}}_B(\theta)$, respectively. Furthermore, $\bar{\mathcal{J}}(\bar{\theta}) = \bar{\mathcal{J}}_B(\bar{\theta})$.

First, the regularity conditions under Assumption 1 allow us to assert that both $\hat{\mathcal{J}}(Y, \theta)/m$ and $\hat{\mathcal{J}}(Y^*, \theta)/m$ have limits. Specifically, let

$$\bar{\mathcal{J}}(\theta) = \lim_{m \rightarrow \infty} \frac{1}{m} E_o \{\hat{\mathcal{J}}(Y, \theta)\}, \text{ and } \bar{\mathcal{J}}_B(\theta) = \lim_{m \rightarrow \infty} \frac{1}{m} E_o E_* \{\hat{\mathcal{J}}(Y^*, \theta)\}.$$

Then, we may claim that

$$\lim_{m \rightarrow \infty} \frac{1}{m} \hat{\mathcal{J}}(Y, \theta) = \bar{\mathcal{J}}(\theta) \text{ a.s.}, \text{ and } \lim_{m \rightarrow \infty} \frac{1}{m} \hat{\mathcal{J}}(Y^*, \theta) = \bar{\mathcal{J}}_B(\theta) \text{ a.s.}$$

Under parametric bootstrapping, $\bar{\mathcal{J}}(\theta) = \bar{\mathcal{J}}_B(\theta)$ may not hold for any $\theta \in \Theta$; however, $\bar{\mathcal{J}}(\theta) = \bar{\mathcal{J}}_B(\theta)$ holds when $\theta = \bar{\theta}$. By Assumptions 1 and 3, along with the consistency of $\hat{\theta}$, we have

$$\begin{aligned} \lim_{m \rightarrow \infty} \frac{1}{m} \hat{\mathcal{J}}(Y^*, \hat{\theta}) &= \bar{\mathcal{J}}_B(\bar{\theta}) \text{ a.s.} \\ &= \lim_{m \rightarrow \infty} \frac{1}{m} E_o E_* \{\hat{\mathcal{J}}(Y^*, \hat{\theta})\} \\ &= \lim_{m \rightarrow \infty} \frac{1}{m} E_o \left\{ -\frac{\partial^2}{\partial\theta\partial\theta'} \{\log L(\theta | Y)\} \Big|_{\theta=\hat{\theta}} \right\} \\ &= \lim_{m \rightarrow \infty} \frac{1}{m} E_o \{\hat{\mathcal{J}}(Y, \hat{\theta})\} = \bar{\mathcal{J}}(\bar{\theta}) \text{ a.s.} \end{aligned}$$

Therefore, under parametric bootstrapping, $\bar{\mathcal{J}}(\bar{\theta}) = \bar{\mathcal{J}}_B(\bar{\theta})$.

Through Assumption 3, in the context of the general linear mixed model (2.1), we can prove that $\bar{\mathcal{J}}(\theta) = \bar{\mathcal{J}}_B(\theta)$ for any $\theta \in \Theta$ under semiparametric and nonparametric bootstrapping. Clearly, $\bar{\mathcal{J}}(\bar{\theta}) = \bar{\mathcal{J}}_B(\bar{\theta})$ holds as a special case.

We now make use of Assumptions 1-4 to establish the asymptotic equivalence of b1 and b2.

First, consider a second-order expansion of $-2 \log L(\hat{\theta} | Y^*)$ about $\hat{\theta}^*$. We have

$$-2 \log L(\hat{\theta} | Y^*) = -2 \log L(\hat{\theta}^* | Y^*) + (\hat{\theta} - \hat{\theta}^*)' \hat{\mathcal{J}}(Y^*, \theta^*)(\hat{\theta} - \hat{\theta}^*). \quad (\text{A.2})$$

Here, θ^* is a random vector which lies between $\hat{\theta}$ and $\hat{\theta}^*$.

Taking expectations with respect to both sides of (A.2), and using the consistency of $\hat{\theta}$ and $\hat{\theta}^*$ along with Assumption 4, we obtain

$$\begin{aligned} & E_* \{-2 \log L(\hat{\theta} | Y^*) - \{-2 \log L(\hat{\theta}^* | Y^*)\}\} \\ &= E_* \{m(\hat{\theta} - \hat{\theta}^*)' \bar{\mathcal{J}}_B(\bar{\theta})(\hat{\theta} - \hat{\theta}^*)\} (1 + o(1)) \text{ a.s.} \end{aligned} \quad (\text{A.3})$$

as $m \rightarrow \infty$.

Next, consider a second-order expansion of $-2 \log L(\hat{\theta}^* | Y)$ about $\hat{\theta}$. We have

$$-2 \log L(\hat{\theta}^* | Y) = -2 \log L(\hat{\theta} | Y) + (\hat{\theta}^* - \hat{\theta})' \hat{\mathcal{J}}(Y, \theta^{**})(\hat{\theta}^* - \hat{\theta}). \quad (\text{A.4})$$

Here, θ^{**} is a random vector which lies between $\hat{\theta}^*$ and $\hat{\theta}$.

Taking expectations with respect to both sides of (A.4), and using the consistency of $\hat{\theta}$ and $\hat{\theta}^*$ along with Assumption 4, we obtain

$$\begin{aligned} & E_* \{-2 \log L(\hat{\theta}^* | Y) - \{-2 \log L(\hat{\theta} | Y)\}\} \\ &= E_* \{m(\hat{\theta}^* - \hat{\theta})' \bar{\mathcal{J}}(\bar{\theta})(\hat{\theta}^* - \hat{\theta})\} (1 + o(1)) \text{ a.s.} \end{aligned} \quad (\text{A.5})$$

as $m \rightarrow \infty$.

Since $\bar{\mathcal{J}}_B(\bar{\theta}) = \bar{\mathcal{J}}(\bar{\theta})$, we know that

$$E_* \{(\hat{\theta} - \hat{\theta}^*)' \bar{\mathcal{J}}_B(\bar{\theta})(\hat{\theta} - \hat{\theta}^*)\} = E_* \{(\hat{\theta} - \hat{\theta}^*)' \bar{\mathcal{J}}(\bar{\theta})(\hat{\theta} - \hat{\theta}^*)\}.$$

Considering (A.3) and (A.5), as $m \rightarrow \infty$, we have

$$\begin{aligned} & E_*\{-2 \log L(\hat{\theta} | Y^*) - \{-2 \log L(\hat{\theta}^* | Y^*)\}\} \\ &= E_*\{-2 \log L(\hat{\theta}^* | Y) - \{-2 \log L(\hat{\theta} | Y)\}\}(1 + o(1)) \text{ a.s.} \end{aligned} \quad (\text{A.6})$$

With reference to (2.5), by (A.6) and Assumption 3, as $m \rightarrow \infty$, we can assert

$$\begin{aligned} \text{b1} &= E_*\{-2 \log L(\hat{\theta}^* | Y) - \{-2 \log L(\hat{\theta}^* | Y^*)\}\} \\ &= 2 E_*\{-2 \log L(\hat{\theta}^* | Y) - \{-2 \log L(\hat{\theta} | Y)\}\}(1 + o(1)) \text{ a.s.} \\ &= \text{b2} (1 + o(1)) \text{ a.s.} \end{aligned} \quad (\text{A.7})$$

Thus, b1-b2 in (2.5) are asymptotically equivalent to each other.

Next, we will justify that b1-b2 in (2.5) are consistent estimators of the bias adjustment (2.3).

Let

$$Q_B = E_*\{m(\hat{\theta} - \hat{\theta}^*)' \bar{\mathcal{J}}_B(\bar{\theta})(\hat{\theta} - \hat{\theta}^*)\} = E_*\{m(\hat{\theta} - \hat{\theta}^*)' \bar{\mathcal{J}}(\bar{\theta})(\hat{\theta} - \hat{\theta}^*)\}.$$

By (A.5) and (A.7), we can conclude that b1 and b2 are asymptotically equivalent to $2Q_B$. Hence, our goal can be modified to prove that $2Q_B$ is a consistent estimator of the bias adjustment (2.3).

We begin our justification by obtaining a useful expression for the limit of Q_B . The derivation of this expression will involve the matrix

$$\hat{\mathcal{I}}(\theta, Y) = \left\{ \frac{\partial}{\partial \theta} \log L(\theta | Y) \frac{\partial}{\partial \theta'} \log L(\theta | Y) \right\}.$$

The regularity conditions under Assumption 1 imply that $\hat{\mathcal{I}}(\theta, Y)/m$ has a limit. Specifically, let

$$\bar{\mathcal{I}}(\theta) = \lim_{m \rightarrow \infty} \frac{1}{m} E_o \left\{ \hat{\mathcal{I}}(\theta, Y) \right\}.$$

We may then claim that

$$\lim_{m \rightarrow \infty} \frac{1}{m} \hat{\mathcal{I}}(\theta, Y) = \bar{\mathcal{I}}(\theta) \text{ a.s.} \quad (\text{A.8})$$

As shown in what follows, the limit of Q_B is based on the matrices $\bar{\mathcal{I}}(\theta)$ and $\bar{\mathcal{J}}(\theta)$.

Lemma 1

$$\lim_{m \rightarrow \infty} Q_B = \text{tr} \{ \bar{\mathcal{I}}(\bar{\theta}) \bar{\mathcal{J}}(\bar{\theta})^{-1} \} \quad a.s.$$

Proof:

We expand $\frac{\partial}{\partial \theta} \log L(\hat{\theta} | Y)$ around $\hat{\theta}^*$ to obtain

$$\hat{\theta} - \hat{\theta}^* = \left\{ -\frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_\alpha | Y) \right\}^{-1} \frac{\partial}{\partial \theta} \log L(\hat{\theta}^* | Y). \quad (\text{A.9})$$

Here, θ_α is a random vector between $\hat{\theta}$ and $\hat{\theta}^*$.

Substituting (A.9) for $\hat{\theta}^* - \hat{\theta}$ into Q_B , and utilizing Assumption 4 along with the consistency of $\hat{\theta}$ and $\hat{\theta}^*$, we can establish that

$$\begin{aligned} & \lim_{m \rightarrow \infty} Q_B \\ = & \lim_{m \rightarrow \infty} E_* \left\{ \frac{1}{m} \left\{ \left[\frac{\partial}{\partial \theta'} \log L(\hat{\theta}^* | Y) \right] \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \bar{\mathcal{J}}(\bar{\theta}) \left\{ \bar{\mathcal{J}}(\bar{\theta})^{-1} \left[\frac{\partial}{\partial \theta} \log L(\hat{\theta}^* | Y) \right] \right\} \right\} \quad a.s. \\ = & \lim_{m \rightarrow \infty} E_* \text{tr} \left\{ \left\{ \frac{1}{m} \left[\frac{\partial}{\partial \theta} \log L(\hat{\theta}^* | Y) \frac{\partial}{\partial \theta'} \log L(\hat{\theta}^* | Y) \right] \right\} \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \quad a.s. \\ = & \lim_{m \rightarrow \infty} E_* \text{tr} \left\{ \left\{ \frac{1}{m} \hat{\mathcal{I}}(\hat{\theta}^*, Y) \right\} \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \quad a.s. \end{aligned} \quad (\text{A.10})$$

Now using (A.8) and the consistency of $\hat{\theta}^*$, we have

$$\lim_{m \rightarrow \infty} \frac{1}{m} \hat{\mathcal{I}}(\hat{\theta}^*, Y) = \bar{\mathcal{I}}(\bar{\theta}). \quad (\text{A.11})$$

Thus, by (A.10) and (A.11), we can argue

$$\lim_{m \rightarrow \infty} Q_B = \text{tr} \{ \bar{\mathcal{I}}(\bar{\theta}) \bar{\mathcal{J}}(\bar{\theta})^{-1} \} \quad a.s. \quad (\text{A.12})$$

Therefore, the lemma is established by (A.12).

Lemma 2

$2Q_B$ is a consistent estimator of the bias adjustment (2.3).

Proof:

First, consider expanding $E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}}$ about $\bar{\theta}$ to obtain

$$\begin{aligned}
E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}} &= E_o\{-2 \log L(\bar{\theta} | Y)\} \\
&\quad + (\hat{\theta} - \bar{\theta})' E_o \left\{ -\frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_\gamma | Y) \right\} (\hat{\theta} - \bar{\theta}) \\
&= E_o\{-2 \log L(\bar{\theta} | Y)\} \\
&\quad + (\hat{\theta} - \bar{\theta})' E_o \left\{ \hat{\mathcal{J}}(Y, \theta_\gamma) \right\} (\hat{\theta} - \bar{\theta}). \tag{A.13}
\end{aligned}$$

Here, θ_γ is a random vector which lies between $\hat{\theta}$ and $\bar{\theta}$.

Next, we expand $-2 \log L(\bar{\theta} | Y)$ about $\hat{\theta}$ and take expectations of both sides of the resulting expression to obtain

$$\begin{aligned}
E_o\{-2 \log L(\bar{\theta} | Y)\} &= E_o\{-2 \log L(\hat{\theta} | Y)\} \\
&\quad + E_o \left\{ (\bar{\theta} - \hat{\theta})' \left\{ -\frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_\delta | Y) \right\} (\bar{\theta} - \hat{\theta}) \right\} \\
&= E_o\{-2 \log L(\hat{\theta} | Y)\} \\
&\quad + E_o \left\{ (\hat{\theta} - \bar{\theta})' \hat{\mathcal{J}}(Y, \theta_\delta) (\hat{\theta} - \bar{\theta}) \right\}. \tag{A.14}
\end{aligned}$$

Here, θ_δ is a random vector which lies between $\hat{\theta}$ and $\bar{\theta}$.

With regard to (A.13) and (A.14), we arrive at

$$\begin{aligned}
&E_o\{-2 \log L(\theta | Y)\} |_{\theta=\hat{\theta}} - E_o\{-2 \log L(\hat{\theta} | Y)\} \\
&= (\hat{\theta} - \bar{\theta})' E_o \left\{ \hat{\mathcal{J}}(Y, \theta_\gamma) \right\} (\hat{\theta} - \bar{\theta}) + E_o \left\{ (\hat{\theta} - \bar{\theta})' \hat{\mathcal{J}}(Y, \theta_\delta) (\hat{\theta} - \bar{\theta}) \right\}. \tag{A.15}
\end{aligned}$$

Again, θ_γ and θ_δ are random vectors which lie between $\hat{\theta}$ and $\bar{\theta}$.

Therefore, from (A.15), we note that the expected value of

$$(\hat{\theta} - \bar{\theta})' E_o \left\{ \hat{\mathcal{J}}(Y, \theta_\gamma) \right\} (\hat{\theta} - \bar{\theta}) + (\hat{\theta} - \bar{\theta})' \hat{\mathcal{J}}(Y, \theta_\delta) (\hat{\theta} - \bar{\theta}) \tag{A.16}$$

should be close to the bias adjustment (2.3) provided that m is large.

Now, we will show that as $m \rightarrow \infty$, the limits of (A.16) and $2Q_B$ are identical. Thus, $2Q_B$ should be asymptotically close to the bias adjustment (2.3). The lemma will therefore be established.

By Assumption 4 and the consistency of $\hat{\theta}$,

$$\lim_{m \rightarrow \infty} \left\{ \frac{1}{m} \hat{\mathcal{J}}(Y, \theta_\delta) \right\} = \lim_{m \rightarrow \infty} E_o \left\{ \frac{1}{m} \hat{\mathcal{J}}(Y, \theta_\gamma) \right\} = \bar{\mathcal{J}}(\bar{\theta}) \text{ a.s.} \quad (\text{A.17})$$

Expanding $\frac{\partial}{\partial \theta} \log L(\hat{\theta} | Y)$ about $\bar{\theta}$, we have

$$\hat{\theta} - \bar{\theta} = \left\{ -\frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_\epsilon | Y) \right\}^{-1} \frac{\partial}{\partial \theta} \log L(\bar{\theta} | Y). \quad (\text{A.18})$$

Here, θ_ϵ is a random vector between $\hat{\theta}$ and $\bar{\theta}$.

By the consistency of $\hat{\theta}$, the limits (A.8) and (A.17), result (A.18), and Lemma 1, the limit of (A.16) reduces as follows:

$$\begin{aligned} & \lim_{m \rightarrow \infty} \left\{ (\hat{\theta} - \bar{\theta})' E_o \left\{ \hat{\mathcal{J}}(Y, \theta_\gamma) \right\} (\hat{\theta} - \bar{\theta}) + (\hat{\theta} - \bar{\theta})' \hat{\mathcal{J}}(Y, \theta_\delta) (\hat{\theta} - \bar{\theta}) \right\} \\ &= 2 \lim_{m \rightarrow \infty} \left\{ \frac{1}{m} \left\{ \left[\frac{\partial}{\partial \theta'} \log L(\bar{\theta} | Y) \right] \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \bar{\mathcal{J}}(\bar{\theta}) \left\{ \bar{\mathcal{J}}(\bar{\theta})^{-1} \left[\frac{\partial}{\partial \theta} \log L(\bar{\theta} | Y) \right] \right\} \right\} \text{ a.s.} \\ &= 2 \lim_{m \rightarrow \infty} \text{tr} \left\{ \left\{ \frac{1}{m} \left[\frac{\partial}{\partial \theta} \log L(\bar{\theta} | Y) \frac{\partial}{\partial \theta'} \log L(\bar{\theta} | Y) \right] \right\} \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \text{ a.s.} \\ &= 2 \lim_{m \rightarrow \infty} \text{tr} \left\{ \left\{ \frac{1}{m} \hat{\mathcal{I}}(Y, \bar{\theta}) \right\} \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \text{ a.s.} \\ &= 2 \text{tr} \left\{ \bar{\mathcal{I}}(\bar{\theta}) \bar{\mathcal{J}}(\bar{\theta})^{-1} \right\} \text{ a.s.} \\ &= 2 \lim_{m \rightarrow \infty} Q_B \text{ a.s.} \end{aligned} \quad (\text{A.19})$$

Hence, the lemma is established by (A.19).

Lemma 2 allows us to assert that b1 and b2 are consistent estimators of the bias adjustment (2.3). Therefore, the criteria AICb1 and AICb2 as given in (2.7) provide us with asymptotically unbiased estimators of $\Delta(k, \theta_o)$.

References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov and F. Csaki (Eds.), Second International Symposium on Information Theory. Akademia Kiado, Budapest, 267–281.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–723.
- Azari, R., Li, L., Tsai C.L., 2006. Longitudinal data model selection. *Computational Statistics and Data Analysis* 50, 3053–3066.
- Bedrick, E.J., Tsai, C.L., 1994. Model selection for multivariate regression in small samples. *Biometrics* 50, 226–231.
- Cavanaugh, J.E., Shumway, R.H., 1997. A bootstrap variant of AIC for state-space model selection. *Statistica Sinica* 7, 473–496.
- Christensen, R., 1996. *Plane Answers to Complex Questions: The Theory of Linear Models* (Second Edition). Springer, New York.
- Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 78, 316–331.
- Efron, B., 1986. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* 81, 461–470.
- Efron, B., Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Hurvich, C.M., Shumway, R.H., Tsai, C.L., 1990. Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika* 77, 709–719.
- Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. *Biometrika* 78, 499–509.

- Hurvich, C.M., Tsai, C.L., 1993. A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series* 14, 271–279.
- Ishiguro, M., Morita, K.I., Ishiguro, M., 1991. Application of an estimator-free information criterion (WIC) to aperture synthesis imaging. In: T.J. Cornell and R.A. Perley (Eds.), *Radio Interferometry: Theory, Techniques, and Applications*. Astronomical Society of the Pacific, San Francisco, 243–248.
- Ishiguro, M., Sakamoto, Y., 1991. WIC: An estimation-free information criterion. Research memorandum. Institute of Statistical Mathematics, Tokyo.
- Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 76–86.
- McQuarrie, A.D.R., Tsai, C.L., 1998. *Regression and Time Series Model Selection*. World Scientific, River Edge, New Jersey.
- Morris, J.S., 2002. The BLUPs are not “best” when it comes to bootstrapping. *Statistics & Probability Letters* 56, 425–430.
- Shibata, R., 1997. Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica* 7, 375–394.
- Sugiura, N., 1978. Further analysis of the data by Akaike’s information criterion and the finite corrections. *Communications in Statistics—Theory and Methods* 7, 13–26.