Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/sjos.12179

Selection Strategy for Covariance Structure of Random Effects in Linear Mixed-effects Models

XINYU ZHANG
Chinese Academy of Sciences

HUA LIANG
George Washington University

ANNA LIU
University of Massachusetts

DAVID RUPPERT
Cornell University

GUOHUA ZOU
Chinese Academy of Sciences and Capital Normal University

ABSTRACT. Linear mixed-effects models are a powerful tool for modelling longitudinal data and are widely used in practice. For a given set of covariates in a linear mixed-effects model, selecting the covariance structure of random effects is an important problem. In this paper, we develop a joint likelihood-based selection criterion. Our criterion is the approximately unbiased estimator of the expected Kullback–Leibler information. This criterion is also asymptotically optimal in the sense that for large samples, estimates based on the covariance matrix selected by the criterion minimize the approximate Kullback–Leibler information. Finite sample performance of the proposed method is assessed by simulation experiments. As an illustration, the criterion is applied to a data set from an AIDS clinical trial.

Key words: asymptotic optimality, covariance structure, Kullback–Leibler information, longitudinal data

1. Introduction

In clinical trials, as well as other biological and biomedical applications, subjects are often measured repeatedly over a given period. Measurements obtained from one subject are generally correlated, while measurements obtained from different subjects can be independent. One useful tool for such longitudinal data is mixed-effects modelling, in which within-subject and between-subject variation are both considered. Linear mixed-effects (LME) models (Laird & Ware, 1982) are a powerful technique for the analysis of longitudinal data and have been studied and applied widely (Pinheiro & Bates, 2000; Verbeke & Molenberghs, 2009; Vonesh & Chinchilli, 1996). As in conventional linear regression, statistical analysis of longitudinal data involves model selection, and selecting the most desirable model is typically the first problem encountered in data analysis. The Akaike information criteria (AIC) (and some other related criteria such as Bayesian information criterion (BIC)) have been used for model selection in LME models (Ngo & Brand, 2002; Pinheiro & Bates, 2000). However, it is well known that for small samples, AIC can be highly biased in estimating expected Kullback–Leibler information under linear models (Hurvich & Tsai, 1989), and similar biases of AIC have been found in LME models when the marginal likelihood function is used (Greven & Kneib, 2010). To reduce these biases, a natural idea is to apply Hurvich & Tsai's (1989) AIC_c to the LME models by considering the marginal likelihood function, that is, the marginal density of the observed

data with the unobserved random effects integrated out. This leads one to use the marginal AIC_c in the LME model selection.

The marginal AIC (mAIC) and marginal AIC_c generally apply to the case of a population focus where interest lies predominantly in estimation of the population parameters, that is, the fixed effects and variance components. On the other hand, for the LME models, we are often interested in prediction of the cluster-specific random effects, not just the population parameters (Vaida & Blanchard, 2005). Vaida & Blanchard (2005) observed in the analysis of a cadralazine study that had a cluster focus; the mAIC and marginal AIC_c are not appropriate for selecting an LME model. Instead, Vaida & Blanchard (2005) proposed using the conditional likelihood function, given the random effects, which leads to the conditional AIC (cAIC) and conditional AIC_c; Liang *et al.* (2008), Greven & Kneib (2010) and Yu & Yau (2012) also study cAIC.

Both the fixed effects and random effects are estimated/predicted by the empirical best linear unbiased predictions (BLUPs), which are the BLUPs with the unknown covariance matrix of the random effects replaced by its estimate. Therefore, a good estimate for this matrix is important both for a population focus and for a cluster focus. In this paper, we are concerned with selection of a model for this covariance matrix. For selection of this model, the cAIC (or conditional AIC_c) is not appropriate, because the conditional likelihood does not explicitly depend on the variance components, as mentioned by Vaida & Blanchard (2005). Marginal likelihood may be used for identifying the covariance matrix of the random effects, because this matrix is part of the marginal covariance matrix of the response. However, as discussed by Vaida & Blanchard (2005), marginal likelihood is not appropriate for a cluster focus; this point is discussed further at the end of Section 2.1.

For identifying the covariance structure of the random effects, we find that the joint likelihood of the data and the random effects is an appropriate tool. We develop a selection criterion based on the joint likelihood and call it joint AIC (jAIC). Because the joint density is the product of the conditional density of the data given the random effects and the marginal density of the latter, jAIC uses the same conditional likelihood that is the basis of cAIC, and so jAIC retains the cluster focus of cAIC. Moreover, because jAIC includes the marginal density of the random effects, it is better than cAIC at selecting a model for the random effects. In our simulation study, both mAIC and jAIC greatly outperform cAIC when selecting the covariance model.

To develop our selection criterion, we utilize a similar method to that of Hurvich & Tsai's (1989) linear model criterion, that is, deriving an (approximately) unbiased estimator of the expected Kullback–Leibler information. Technically, our derivation is by no means straightforward for the following reasons: (i) in Hurvich and Tsai's approach, the (approximate) distributions of the residual sum of squares (RSS) and the variance estimators should be known. For the LME models, such distributions may not be easy to find, because closed forms for variance estimators are not available. (ii) The RSS and the variance estimators are required to be independent in Hurvich and Tsai's approach. This cannot be true for the LME models.

This paper is organized as follows. In Section 2, we derive approximately unbiased estimators of the expected Kullback–Leibler information and then propose a jAIC for the selection of the covariance matrix of the random effects. We prove asymptotic optimality of jAIC in the sense that the estimators based on the selected covariance matrix that minimizes our proposed criterion also minimize approximate Kullback–Leibler information in large samples. In Section 3, we present Monte Carlo simulation results to illustrate the performance of the proposed criterion. We apply our criterion to a real data set in Section 4. Some discussions are given in Section 5. The details of derivation and proof are presented in Appendix and Supporting Information.

2. Joint AIC for LME models

2.1. The LME model

The general form of the LME model is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i,\ i = 1, \ldots, n$$

with $\mathbf{b}_i \sim N(0, \mathbf{D})$ and $\varepsilon_i \sim N(0, \sigma^2\mathbf{I}_{m_i})$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed regression coefficients; $\mathbf{b}_i$ is a $k \times 1$ vector of random coefficients specific to the subject i; $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$, $\mathbf{X}_i\,(m_i \times p)$ and $\mathbf{Z}_i\,(m_i \times k)$ denote the response variables, known covariate matrices for the fixed and random effects of full column rank, respectively, of the ith subject; $\varepsilon_i\,(m_i \times 1)$ is an error vector independent of $\mathbf{b}_i$; σ^2 is an unknown parameter; and $\mathbf{I}_{m_i}$ is an $m_i \times m_i$ identity matrix. The covariance matrix $\mathbf{D}$ may have special structures. Let $N = \sum_{i=1}^{n} m_i$ be the total number of observations and $\boldsymbol{\theta}$ be the vector of parameters in the model, including $\boldsymbol{\beta}$, σ^2 and the parameters in $\mathbf{D}$. Clearly, the previous LME model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \varepsilon,\ \mathbf{b} \sim N(0, \mathbf{G}), \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector of observations, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$ is an $N \times p$ matrix of rank p, $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is an $N \times r$ block-diagonal matrix of rank $r = nk$, $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_n')'$, $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_n')'$ and $\mathbf{G} = \mathrm{diag}(\mathbf{D}, \ldots, \mathbf{D})$ is an $r \times r$ block-diagonal matrix.

Suppose that the data $\mathbf{y}$ comes from the following true LME model:

$$\begin{cases} \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}_0\mathbf{b}_0 + \varepsilon_0, \\ \mathbf{b}_0 \sim N(0, \mathbf{G}_0),\ \varepsilon_0 \sim N(0, \sigma_0^2\mathbf{I}_N), \end{cases} \tag{2}$$

where $\boldsymbol{\beta}_0$ is a $p_0 \times 1$ vector of fixed effects, $\mathbf{b}_0$ is an $r_0 \times 1$ vector of random effects, $\mathbf{X}_0$ and $\mathbf{Z}_0$ are the $N \times p_0$ and $N \times r_0$ matrices, respectively, ε_0 is the $N \times 1$ disturbance, and $\mathbf{G}_0 = \mathrm{diag}(\mathbf{D}_0, \ldots, \mathbf{D}_0)$ is an $r_0 \times r_0$ block-diagonal matrix.

In applications, we use the model (1) to fit the observed data $\mathbf{y}$. The purpose of this paper is to choose the appropriate matrix $\mathbf{G}$ for random effects given covariate matrices $\mathbf{X}$ and $\mathbf{Z}$. The setup of models (1) and (2) implies that when we consider the choice of covariance structure, the fitting covariate matrices $\mathbf{X}$ and $\mathbf{Z}$ are allowed to be different from the true covariate matrices $\mathbf{X}_0$ and $\mathbf{Z}_0$, which is a more practical and general framework.

Let $\Sigma = \sigma^2\mathbf{I}_N + \mathbf{Z}\mathbf{G}\mathbf{Z}'$. Under model (1), for given $\mathbf{G}$ and σ^2, the maximum likelihood (ML) estimators of $\boldsymbol{\beta}$ and $\mathbf{b}$ are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y} \tag{3}$$

and

$$\widehat{\mathbf{b}} = \mathbf{G}\mathbf{Z}'\Sigma^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{4}$$

respectively (e.g. Laird & Ware, 1982). The unknown parameters in $\mathbf{G}$ and σ^2 can be estimated by using the ML or restricted ML methods (Davidian & Giltinan, 1995). Denote their estimators by $\widehat{\mathbf{G}}$ and $\widehat{\sigma}^2$. In the following content, when using $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{b}}$, $\widehat{\mathbf{G}}$ and $\widehat{\sigma}^2$ have been plugged in. Let $\widehat{\boldsymbol{\theta}}$ be the estimator of $\boldsymbol{\theta}$.

2.2. Joint Akaike information and selection criterion

For avoiding the impact of reparameterization on the criterion that we will propose,[1] we first scale the random effect $\mathbf{b}$ by defining $\mathbf{b}^* = \mathbf{G}^{-1/2}\mathbf{b}$.

Let $\mathbf{b}^* = \mathbf{G}^{-1/2}\mathbf{b}$, and then $\mathbf{b}^*$ is a vector of independent standard normal variables. As some studies on selecting random effect such as Chen & Dunson (2003) and Kinney & Dunson (2007), we focus on $\mathbf{b}^*$ instead of $\mathbf{b}$ when considering likelihood functions. Denote the density of $\mathbf{y}$ under the true model by (2) by $f_{\text{true}}(\mathbf{y}, \mathbf{b}_0^*|\boldsymbol{\beta}_0, \mathbf{G}_0)$, the conditional density function of $\mathbf{y}$ by $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{b}^*)$ and the density of $\mathbf{b}^*$ by $p(\mathbf{b}^*|\boldsymbol{\theta})$. Rewrite the estimators $\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{b}}$ as $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, $\widehat{\mathbf{G}}(\mathbf{y})$ and $\widehat{\mathbf{b}}(\mathbf{y})$, respectively, and let $\widehat{\mathbf{b}}^* = \widehat{\mathbf{b}}^*(\mathbf{y}) = \widehat{\mathbf{G}}(\mathbf{y})^{-1/2}\widehat{\mathbf{b}}(\mathbf{y})$ and $\mathbf{b}_0^* = \mathbf{G}_0^{-1/2}\mathbf{b}_0$. Then we define the joint Akaike information as follows.

Definition 1. The joint Akaike information (or expected Kullback–Leibler information) is defined to be

$$\text{jAI} = 2E_{\mathbf{y},\mathbf{b}_0^*} f_{\text{true}}(\mathbf{y}, \mathbf{b}_0^*|\boldsymbol{\beta}_0, \mathbf{G}_0) - 2E_{\mathbf{y},\mathbf{b}_0^*}\left[E_{\widetilde{\mathbf{y}},\widetilde{\mathbf{b}}^*}\left\{\log(f(\widetilde{\mathbf{y}}|\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\mathbf{b}}^*(\mathbf{y})) \cdot p(\widehat{\mathbf{b}}^*(\mathbf{y})|\widehat{\boldsymbol{\theta}}(\mathbf{y})))\right\}\right],$$

where $\widetilde{\mathbf{y}}$ and $\widetilde{\mathbf{b}}^*$ have the same distributions as $\mathbf{y}$ and $\mathbf{b}^*$, and are independent of $\mathbf{y}$ and $\mathbf{b}^*$, respectively.

Here, we consider the discrepancy between the joint density functions of the data and random effects (which are, of course, unobservable) under the candidate and true models. Clearly, if we investigate the deviation of the likelihood under the candidate model from that under the true model by making use of the marginal or conditional density function, then the mAIC or cAIC will be resulted in. This quantity has been used in literature (e.g. Li *et al.*, 2002; Wu & Zhang, 2006, pp. 19).

Using model (1), the distinctions between cluster and population focus can be stated more analytically. The cluster focus captured by cAIC is on the model $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ for the conditional mean of $\mathbf{y}$. The population focus captured by mAIC is on the marginal mean $\mathbf{X}\boldsymbol{\beta}$ and marginal covariance matrix $\Sigma = \sigma^2\mathbf{I}_N + \mathbf{Z}\mathbf{G}\mathbf{Z}'$ of $\mathbf{y}$. In contrast, jAIC focuses on the conditional mean of $\mathbf{y}$ and the marginal covariance matrix of $\mathbf{b}$. Thus, jAIC measures both the goodness-of-fit of $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ to $\mathbf{y}$ *and* how well the covariance matrix of $\mathbf{b}$ is fit by $\mathbf{G}$, while cAIC does only the former. One premise of this paper is that it is often best to measure the goodness-of-fit of the conditional, rather than the marginal, model for the mean of $\mathbf{y}$; therefore, we prefer jAIC to mAIC. Another premise is that one should not neglect the model for the distribution of the random effects, which leads us to prefer jAIC to cAIC.

Define $\mathbf{Z}^* = \mathbf{Z}\widehat{\mathbf{G}}^{1/2}$. We use E_0 to represent $E_{\mathbf{y},\mathbf{b}_0^*}$ for simplicity. Noting that

$$\begin{aligned}
&\log f((\widetilde{\mathbf{y}}|\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\mathbf{b}}^*(\mathbf{y})) \cdot p(\widehat{\mathbf{b}}^*(\mathbf{y})|\widehat{\boldsymbol{\theta}}(\mathbf{y})))\\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\widehat{\sigma}^2 - \frac{1}{2\widehat{\sigma}^2}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}^*\widehat{\mathbf{b}}^*\|^2 - \frac{r}{2}\log(2\pi) - \frac{1}{2}(\widehat{\mathbf{b}}^*)'\widehat{\mathbf{b}}^*\\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\widehat{\sigma}^2 - \frac{1}{2\widehat{\sigma}^2}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{b}}\|^2 - \frac{r}{2}\log(2\pi) - \frac{1}{2}\widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}} \equiv \log g_{\widehat{\boldsymbol{\theta}}}(\widetilde{\mathbf{y}}, \widehat{\mathbf{b}}),
\end{aligned}$$

we obtain

[1] If we do not scale the random effect, there will be an additional term $\log|\widehat{G}|$ in our criterion, which is related to reparameterization of $\mathbf{Z}$ (i.e., replace $\mathbf{Z}$ by multiple of $\mathbf{Z}$).

$$
\begin{aligned}
&2E_0 f_{\text{true}}(\mathbf{y}, \mathbf{b}_0^* | \boldsymbol{\beta}_0, \mathbf{G}_0) - 2\left[E_{\widetilde{\mathbf{y}},\widetilde{\mathbf{b}}^*}\left\{\log f(\widetilde{\mathbf{y}}|\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\mathbf{b}}^*(\mathbf{y})) \cdot p(\widehat{\mathbf{b}}^*(\mathbf{y})|\widehat{\boldsymbol{\theta}}(\mathbf{y}))\right\}\right] \\
&= N\log\frac{\widehat{\sigma}^2}{\sigma_0^2} + \frac{1}{\widehat{\sigma}^2}\left\{\|\mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{b}}\|^2 + \text{tr}(\mathbf{Z}_0\mathbf{G}_0\mathbf{Z}_0')\right\} + \frac{N\sigma_0^2}{\widehat{\sigma}^2} \\
&\quad + \widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}} + (r - r_0)\log(2\pi) - N - r_0 \equiv \Delta(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}}),
\end{aligned}
$$

where $\|\mathbf{a}\|^2 = \mathbf{a}'\mathbf{a}$ and the notation 'tr' means taking trace. It is seen that $\text{jAI} = E_0\Delta(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$. For the convenience in the subsequent analysis, we rewrite $\Delta(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$ as

$$
\begin{aligned}
\Delta(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}}) &= N\log\frac{\widehat{\sigma}^2}{\sigma_0^2} + \frac{1}{\widehat{\sigma}^2}\|\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}_0\mathbf{b}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{b}}\|^2 \\
&\quad + \frac{1}{\widehat{\sigma}^2}\left\{\text{tr}(\mathbf{Z}_0\mathbf{G}_0\mathbf{Z}_0') - \mathbf{b}_0'\mathbf{Z}_0'\mathbf{Z}_0\mathbf{b}_0 - 2\mathbf{b}_0'\mathbf{Z}_0'(\mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{b}})\right\} \\
&\quad + \frac{N\sigma_0^2}{\widehat{\sigma}^2} + \widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}} + (r - r_0)\log(2\pi) - N - r_0.
\end{aligned} \tag{5}
$$

Observing that $E_0[\mathbf{b}_0'\mathbf{Z}_0'\mathbf{Z}_0\mathbf{b}_0 + 2\mathbf{b}_0'\mathbf{Z}_0'(\mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{y}}) - \mathbf{Z}\widehat{\mathbf{b}}(\widetilde{\mathbf{y}}))] = \text{tr}(\mathbf{Z}_0\mathbf{G}_0\mathbf{Z}_0')$, we see that under the true model, the right-hand side of (5) can be approximated by

$$
\begin{aligned}
&N\log\frac{\widehat{\sigma}^2}{\sigma_0^2} + \frac{1}{\widehat{\sigma}^2}\|\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}_0\mathbf{b}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{b}}\|^2 + \frac{N\sigma_0^2}{\widehat{\sigma}^2} + \widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}} \\
&\quad + (r - r_0)\log(2\pi) - N - r_0 \equiv \Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}}).
\end{aligned} \tag{6}
$$

Therefore, a reasonable measure representing the discrepancy between the candidate and true models would be $E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$.

Next, we derive an (approximately) unbiased estimator of $E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$ because such an estimator can be used to define a feasible selection criterion for the covariance structure of random effects. Define $\widehat{\boldsymbol{\gamma}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$, $\Phi(\mathbf{y}) = \widehat{\sigma}^{-2}\text{tr}\left(\partial\widehat{\boldsymbol{\gamma}}'/\partial\mathbf{y}\right) + (\widehat{\boldsymbol{\gamma}} - \mathbf{y})'\partial(\widehat{\sigma}^{-2})/\partial\mathbf{y}, \Psi(\mathbf{y}) = \text{tr}\left\{\partial^2(\widehat{\sigma}^{-2})/(\partial\mathbf{y}\partial\mathbf{y}')\right\}$ and

$$
\text{jAIC}_0 = -2\log g_{\widehat{\boldsymbol{\theta}}}(\mathbf{y}, \widehat{\mathbf{b}}) + 2\sigma_0^2\Phi(\mathbf{y}) + \sigma_0^4\Psi(\mathbf{y}) - N\log 2\pi - N\log\sigma_0^2 - N - r_0 - r_0\log 2\pi. \tag{7}
$$

From derivation in Appendix A.1, we know jAIC_0 is an unbiased estimator of $E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$, that is,

$$
E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}}) = E_0\text{jAIC}_0. \tag{8}
$$

Note that jAIC_0 still depends on unknown σ_0^2. In Section S.1 of Supporting Information, we derive an exactly unbiased estimator of σ_0^2 under some conditions. Here, for simplicity, we estimate σ_0^2 by $\widehat{\sigma}^2$. In Section S.3 of Supporting Information, we can see that the selection results by using σ_0^2 and $\widehat{\sigma}^2$ are very similar.

Thus, we propose to select the best covariance matrix of random effects that minimizes the following quantity:

$$
\text{jAIC}^* = -2\log g_{\widehat{\boldsymbol{\theta}}}(\mathbf{y}, \widehat{\mathbf{b}}) + 2\widehat{\sigma}^2\Phi(\mathbf{y}) + \widehat{\sigma}^4\Psi(\mathbf{y}) = -2\log g_{\widehat{\boldsymbol{\theta}}}(\mathbf{y}, \widehat{\mathbf{b}}) + 2\rho_0(\mathbf{y}) + \rho_1(\mathbf{y}), \tag{9}
$$

where $\rho_0(\mathbf{y}) = \text{tr}\left(\partial\widehat{\boldsymbol{\gamma}}'/\partial\mathbf{y}\right)$ and $\rho_1(\mathbf{y}) = 2\widehat{\sigma}^2(\widehat{\boldsymbol{\gamma}} - \mathbf{y})'\partial(\widehat{\sigma}^{-2})/\partial\mathbf{y} + \widehat{\sigma}^4\text{tr}\left\{\partial^2(\widehat{\sigma}^{-2})/(\partial\mathbf{y}\partial\mathbf{y}')\right\}$.

It is interesting to observe that the expectation of $\rho_0(\mathbf{y})$ in (9), conditional on $\mathbf{b}_0$, is just the generalized degrees of freedom defined by Ye (1998) for the LME models. Also, the penalty term $2\rho_0(\mathbf{y})$ is exactly the same as that in cAIC with the known error variance (Liang *et al.*,

2008). The third term, $\rho_1(\mathbf{y})$ in (9), is an extra penalty due to the variability of estimating the unknown error variance. Compared with the cAIC of Liang *et al.* (2008) and Greven & Kneib (2010), our jAIC utilizes joint likelihood function that contains the distribution information of the random effect $\mathbf{b}$, but the cAIC does not directly depend on the distribution of $\mathbf{b}$ (of course, when estimating $\boldsymbol{\theta}$, the distribution information of $\mathbf{b}$ is used) and thus the estimated covariance $\widehat{\mathbf{G}}$. As a result, a new term $\widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}}$ that measures the goodness-of-fit related to the random effects appears in the jAIC. In addition, our simulation results in Section 3 also show that the cAIC is not appropriate in selecting covariance structure of random effects.

2.3. Asymptotic optimality

Our foregoing discussion focuses on the finite sample justification of the proposed criterion. We now consider the large sample asymptotic optimality of our approach. Noting that $\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$ can be regarded as approximate Kullback–Leibler information between the candidate model and true model, in this subsection, we will illustrate that the estimators based on the selected covariance structure of random effects that minimizes our proposed criterion also minimize $\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$ in large samples.

Let $\{1, \ldots, S\}$ be the index set denoting the candidate LME models that have the same forms as (1) but different covariance structures of random effects, and $\mathbf{D}_s$, $\mathbf{G}_s$, σ_s^2 and Σ_s be the values of $\mathbf{D}$, $\mathbf{G}$, σ^2 and Σ under the sth candidate model, respectively. We further write $\widehat{\mathbf{b}}_s$ and $\widehat{\boldsymbol{\theta}}_s$ as the corresponding versions of $\widehat{\mathbf{b}}$ and $\widehat{\boldsymbol{\theta}}$, respectively. Denote $\boldsymbol{\omega}_s$ as the unknown parameter vector in $\mathbf{G}_s$, and $\widehat{\boldsymbol{\omega}}_s$ and $\widehat{\sigma}_s^2$ as the estimators of $\boldsymbol{\omega}_s$ and σ_s^2, respectively. Let $\boldsymbol{\eta}_s = (\boldsymbol{\omega}_s', \sigma_s^2)'$ with finite J_s elements, and $\widehat{\boldsymbol{\eta}}_s = (\widehat{\boldsymbol{\omega}}_s', \widehat{\sigma}_s^2)'$ that is obviously a part of $\widehat{\boldsymbol{\theta}}_s$. Denote $\widehat{\Sigma}_s = \Sigma_s(\widehat{\boldsymbol{\eta}}_s) = \widehat{\sigma}_s^2\mathbf{I}_N + \mathbf{Z}\mathbf{G}_s(\widehat{\boldsymbol{\omega}}_s)\mathbf{Z}' = \widehat{\sigma}_s^2\mathbf{I}_N + \mathbf{Z}\widehat{\mathbf{G}}_s\mathbf{Z}'$, $\widehat{\mathbf{V}}_s = \widehat{\Sigma}_s^{-1/2}\mathbf{X}(\mathbf{X}'\widehat{\Sigma}_s^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\Sigma}_s^{-1/2}$, and $\widehat{\mathbf{P}}_s = \mathbf{I}_N - \widehat{\sigma}_s^2\widehat{\Sigma}_s^{-1/2}(\mathbf{I}_N - \widehat{\mathbf{V}}_s)\widehat{\Sigma}_s^{-1/2}$. Then from (3) and (4), we observe that under the sth candidate model, the estimator of $\boldsymbol{\gamma}$ is given by $\widehat{\boldsymbol{\gamma}}_s = \mathbf{X}\widehat{\boldsymbol{\beta}}_s + \mathbf{Z}\widehat{\mathbf{b}}_s = \widehat{\mathbf{P}}_s\mathbf{y}$. Recalling (6) and (9), the values of $\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$ and jAIC^* under the sth candidate model can be written as

$$\begin{aligned}\Delta_s^*(\widehat{\boldsymbol{\theta}}_s, \widehat{\mathbf{b}}_s) = N\left(\frac{\sigma_0^2}{\widehat{\sigma_s^2}} + \log\frac{\widehat{\sigma_s^2}}{\sigma_0^2} - 1\right) + \widehat{\sigma}_s^{-2}\|\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}_0\mathbf{b}_0 - \widehat{\boldsymbol{\gamma}}_s\|^2 \\ + \widehat{\mathbf{b}}_s'\widehat{\mathbf{G}}_s^{-1}\widehat{\mathbf{b}}_s + (r - r_0)\log(2\pi) - r_0\end{aligned} \tag{10}$$

and

$$\begin{aligned}\mathrm{jAIC}_s^* = N\log(2\pi) + N\log\widehat{\sigma_s^2} + \widehat{\sigma}_s^{-2}\|\mathbf{y} - \widehat{\boldsymbol{\gamma}}_s\|^2 + 2\mathrm{tr}\frac{\partial\widehat{\boldsymbol{\gamma}}_s}{\partial\mathbf{y}'} + r\log 2\pi + \widehat{\mathbf{b}}_s'\widehat{\mathbf{G}}_s^{-1}\widehat{\mathbf{b}}_s \\ + 2\widehat{\sigma}_s^{-2}(\mathbf{y} - \widehat{\boldsymbol{\gamma}}_s)\frac{\partial\widehat{\sigma_s^2}}{\partial\mathbf{y}} - \mathrm{tr}\frac{\partial^2\widehat{\sigma_s^2}}{\partial\mathbf{y}\partial\mathbf{y}'} + 2\widehat{\sigma}_s^{-2}\frac{\partial\widehat{\sigma_s^2}}{\partial\mathbf{y}'}\frac{\partial\widehat{\sigma_s^2}}{\partial\mathbf{y}},\end{aligned} \tag{11}$$

respectively.

Let $\widehat{s} = \underset{s\in\{1,\ldots,S\}}{\arg\min}\ \mathrm{jAIC}_s^*$, the model selected by minimizing the criterion jAIC_s^*. Under some reasonable conditions, the following theorem shows the asymptotic optimality of the jAIC_s^* in the sense of making $\Delta_s^*(\widehat{\boldsymbol{\theta}}_s, \widehat{\mathbf{b}}_s)$ as small as possible.

Theorem 1. *Under conditions 1–8 in Appendix A.2, we have*

$$\frac{\Delta_{\widehat{s}}^*(\widehat{\boldsymbol{\theta}}_{\widehat{s}}, \widehat{\boldsymbol{b}}_{\widehat{s}})}{\min\limits_{s\in\{1,\ldots,S\}}\left\{\Delta_s^*(\widehat{\boldsymbol{\theta}}_s, \widehat{\boldsymbol{b}}_s)\right\}} \xrightarrow{p} 1 \quad \textit{as } N \to \infty. \tag{12}$$

The proof is given in Appendix A.2.

2.4. Implementation of jAIC

In fact, the penalty term in jAIC*, $2\rho_0(\mathbf{y}) + \rho_1(\mathbf{y})$, is the same as $2\Phi_1$ of Greven & Kneib (2010) except that $\widetilde{\sigma}^2$ is replaced by its estimate $\widehat{\sigma}^2$ (see formula (9) of Greven & Kneib, 2010). When σ^2 is known, the penalty term can be simplified to $2\rho_0(y)$, that is, $2\Phi_0$ in Liang *et al.* (2008) and Greven & Kneib (2010). But the calculation of Φ_0 and Φ_1 requires additional N and $2N$ model fits, respectively, which leads to large calculation burden. Exhilaratingly, Greven & Kneib (2010) develop an analytic representation of Φ_0 (see their theorem 3) and provide an R package for implementation. They also show the close agreement between using Φ_1 and $\Phi_0 + 1$ for model selection by simulations. Therefore, for the implementation of jAIC* to be convenient, we propose to use

$$\text{jAIC} = -2\log g_{\widehat{\boldsymbol{\theta}}}(\mathbf{y}, \widehat{\mathbf{b}}) + 2\Phi_0, \tag{13}$$

where $\Phi_0 = \widehat{\rho} + \sum_{j=1}^{s} \mathbf{e}_j' \widehat{\mathbf{B}}_*^{-1} \widehat{\mathbf{G}}_* \widehat{\mathbf{A}}_* \widehat{\mathbf{W}}_{*,j} \widehat{\mathbf{A}}_* \mathbf{y}$ and the definitions of $\widehat{\rho}$, s, $\mathbf{e}_j$, $\widehat{\mathbf{B}}_*$, $\widehat{\mathbf{G}}_*$, $\widehat{\mathbf{A}}_*$ and $\widehat{\mathbf{W}}_{*,j}$ can be found in theorem 3 of Greven & Kneib (2010).

3. Simulation study

In this section, we investigate the finite sample performance of the proposed procedure and compare it with some existing methods by Monte Carlo simulations.

We simulate data from the following LME model:

$$y_{ij} = (1, t_j)\boldsymbol{\beta}_0 + (1, t_j)\mathbf{b}_{0i} + \varepsilon_{ij}, \qquad i = 1, \ldots, 10, \qquad j = 1, \ldots, 20,$$

with $\mathbf{b}_{0i} \sim N(0, \mathbf{D})$, $t_j = 5(j-1)$, $(\beta_0, \beta_1) = (3, 0.2)$, and ε_{ij}s are normally distributed with mean 0 and variance 0.5^2. That is, we have 10 subjects with a positive definite covariance matrix $\mathbf{D}$. In the simulations, we vary a complexity parameter (ν or τ); when this parameter increases from its baseline value, the structure of $\mathbf{D}$ diverges from a simple model. As in Greven & Kneib (2010), we report the probability of selecting the more complex model by the various criteria. The squared Frobenius norm differences between the true and estimated covariance matrices, which are defined by $\text{tr}\{(\mathbf{D} - \widehat{\mathbf{D}})'(\mathbf{D} - \widehat{\mathbf{D}})\}$, are also presented.

Four cases are considered as follows.

Case I: In this case, the choice is between a multiple of the identity structure ($\nu = 1$) and a general diagonal structure ($\nu > 1$) for $\mathbf{D}$, where $\mathbf{D} = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6\nu \end{pmatrix}$. We fit the simulated data using the R function lme with the options that the random effects covariance matrix would be multiple of the identity or a diagonal matrix. We then obtain the mAIC and the marginal BIC (mBIC) values under these two covariance structures. Specifically, the mAIC and mBIC are defined by $-2\text{logLik}_s + 2\psi_s$ and $-2\text{logLik}_s + \log(N)\psi_s$, respectively, where logLik_s and ψ_s are the marginal log-likelihood and the number of unknown parameters in the sth candidate model, respectively. Additionally, we also include the adjusted BIC (aBIC) proposed by Delattre *et al.* (2014) in this simulation, which is defined by $-2\text{logLik}_s + \log(n)\psi_{R,s} + \log(N)\psi_{F,s}$, where $\psi_{F,s}$ is the number of fixed effects and $\psi_{R,s}$ is the sum of the number of random effects and the number of unknown parameters in $\mathbf{D}_s$. Based on 1000 replications, we compute the proportion selecting the diagonal structure and the average Frobenius norm difference between the true and estimated covariance matrix. The results are presented in Fig. 1A and B. First, when $\nu = 1$, which means that $\mathbf{D}$ is exactly a multiple of the identity, mBIC leads to the smallest proportion of selecting the diagonal structure and all AIC correspond to about 0.3 estimated probability of wrongly choosing the diagonal structure. As the value of ν increases

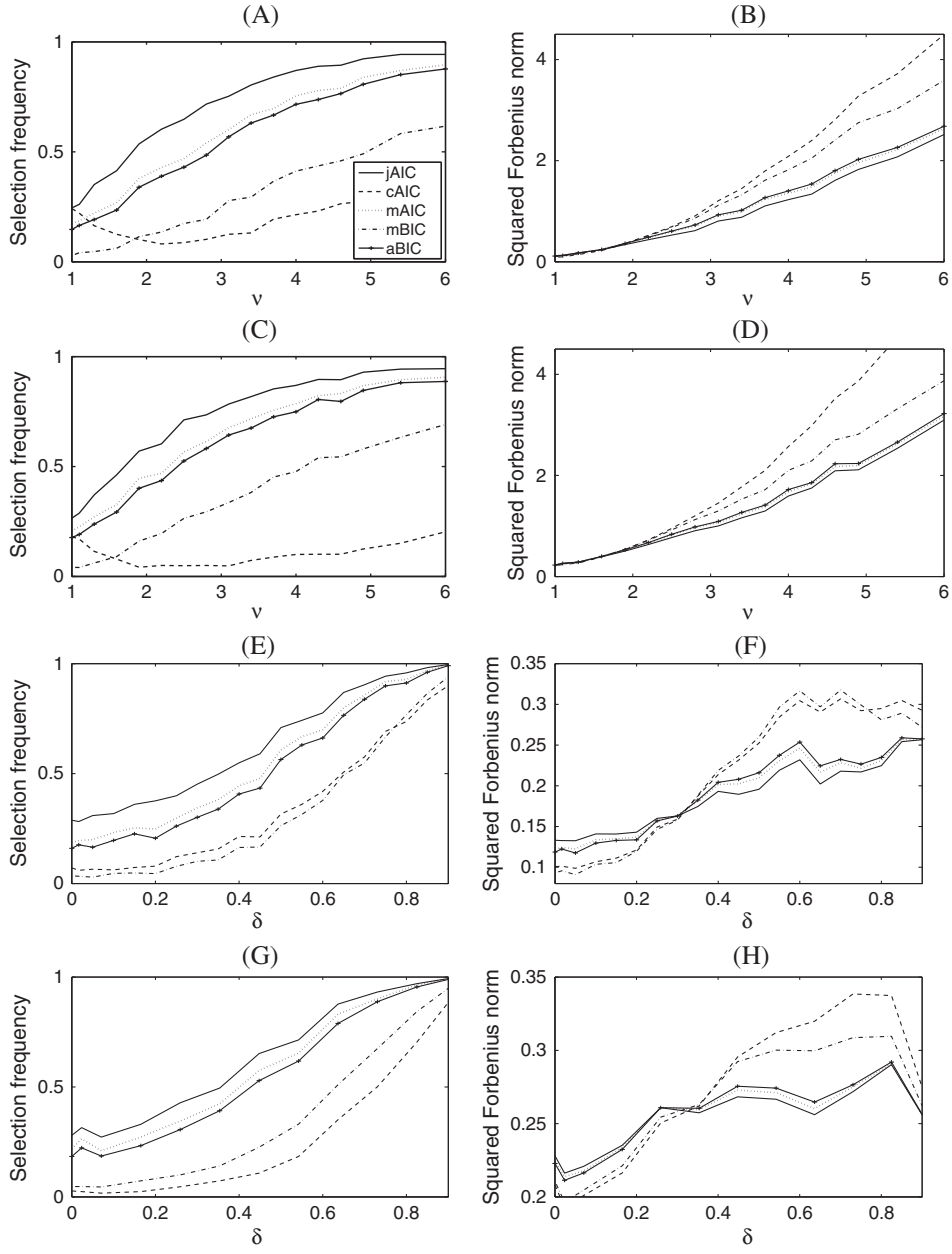

Fig. 1. Left column: proportions of 1000 replications where the more complex model is selected. Right column: the values of the average Frobenius distance between estimated and true covariance matrices. Joint Akaike information criteria (jAIC): solid line; conditional AIC (cAIC): dashed line; marginal AIC (mAIC): dotted line; marginal BIC (mBIC): dashed-dotted line; and adjusted BIC (aBIC): solid line with +. Cases I–IV are in rows 1–4, respectively.

from 1, that is, the structure of **D** moves away from being a multiple of the identity to a general diagonal structure, the probability of each criterion choosing a diagonal structure increases as well, except that cAIC leads first to a decreasing probability before increasing. However, cAIC and mBIC have relatively low probabilities of selecting the general diagonal structure when

$\nu > 1$; their probabilities are even less than 0.5 when $\nu \geq 4$, which implies very different diagonal elements. It is worth noting that the probability of selecting the general diagonal structure is always somewhat higher for jAIC than for mAIC, which means that jAIC tends to favour a more complex structure, but the mAIC tends to select a simpler structure. This performance can be partly explained by the finding in theorem 1 of Greven & Kneib (2010) that mAIC is not an asymptotically unbiased estimator of the Akaike information and favours smaller models without random effects. aBIC performs between mAIC and mBIC and closely to mAIC. Figure 1B shows that when ν is small, all selection criteria estimate $\mathbf{D}$ with similar accuracy, but when ν is larger, jAIC performs the best. In a high proportion of circumstance, jAIC is found to be the preferred criterion.

Case II: We select between compound-symmetry ($\nu = 1$) and general positive definite (general PD) ($\nu > 1$) structures of $\mathbf{D}$. We set $\mathbf{D} = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6\nu \end{pmatrix}$. The results shown in Fig. 1C and D are similar to those in Case I.

Case III: In this case, we select between a multiple of the identity structure and a compound-symmetry structure of $\mathbf{D}$. We set $\mathbf{D} = \begin{pmatrix} 0.6 & \tau \\ \tau & 0.6 \end{pmatrix}$, so the correlation coefficient $\delta = \tau/0.6$. Figure 1E displays the proportions of selecting compound-symmetry structure of all criteria. When δ is very small (say < 0.1), which means that the multiple of the identity structure is appropriate, the mBIC and cAIC obviously support this structure (the estimated probability of selecting it is above 0.9), and jAIC, mAIC and aBIC lead to somewhat smaller probabilities (about 0.7–0.8). When δ is big (say 0.6), that is, the compound-symmetry structure is more appropriate, jAIC leads to the highest estimated probability of selecting the compound-symmetry structure, while the probabilities for cAIC and mBIC are still lower than 0.5. Figure 1F shows that when δ is small (say < 0.3), the criteria that favour a simpler structure provide better estimates of the covariance matrix. In contrast, when δ is moderate to large (say > 0.4), the criteria that favour a more complex structure provide better estimates of the covariance matrix.

Case IV: We select between diagonal and general PD structures of $\mathbf{D}$. We set $\mathbf{D} = \begin{pmatrix} 0.6 & \tau \\ \tau & 0.3 \end{pmatrix}$. The correlation coefficient $\delta = \tau/\sqrt{0.18}$. The results shown in Fig. 1G and H are similar to those in Case III.

In conclusion, jAIC and mAIC are preferred to cAIC in selection for covariance structure. Compared with mAIC, mBIC and aBIC, jAIC tends to favour a more complex structure regardless of the values of ν or δ; as a result, jAIC leads to a better estimate of covariance matrix of the random effects when ν or δ is large, but a worse estimate when ν or δ is small. In most situations we considered, jAIC performs best.

All results presented in Fig. 1 are based on ML estimation. The results based on restricted ML estimation are similar to those we have shown and omitted to save space but available upon request from the authors.

4. Example: decay rate of viral response

An understanding of the pathogenesis of HIV infection plays an important role in the evaluation of antiviral therapies for AIDS/HIV. Recent research indicates that the decay rate of the first phase of the viral response (number of copies of HIV RNA in the plasma or viral load)

is a useful marker for antiviral potency (Wei *et al.*, 1995; Ho *et al.*, 1995). LME has become a standard tool for estimation of the first decay rate (Wu & Ding, 1999). In this section, we present an analysis of a subset of an AIDS clinical trial group (ACTG 315) study. In this study, viral load was scheduled to be measured on days 2, 7, 10, 14, 21 and 28 and weeks 8, 12, 24 and 48 after initiation of an antiviral therapy. We used the data from the first 2 weeks, because week 2 is the time when the second decay appears. Our data set is comprised of 36 patients, with the number of observations per patient varying from 3 to 5. We present the scatterplot of these observations in Fig. 2. See Lederman *et al.* (1998) for the details about this study.

We now fit the following model to the data:

$$y_{ij} = (1, t_j)\boldsymbol{\beta}_0 + (1, t_j)\mathbf{b}_{0i} + \varepsilon_{ij},$$

where $i = 1, \ldots, 36$, $j = 1, \ldots, m_i$ and y_{ij} is the viral load (log scale) of patient i at measurement time t_{ij}. We consider the following four covariance structures: (a) multiple of an identity, (b) diagonal, (c) general PD, (d) compound-symmetry, because the package nlme in R implements for these four commonly used structures. All of the AIC and BIC criteria disfavour structures (a) and (d) but give ambiguous results for structures (b) and (c). Therefore, in what follows, we present the details for structures (b) and (c). It is seen in Table 1 that cAIC prefers to general PD structure. On the other hand, the results based on other criteria indicate

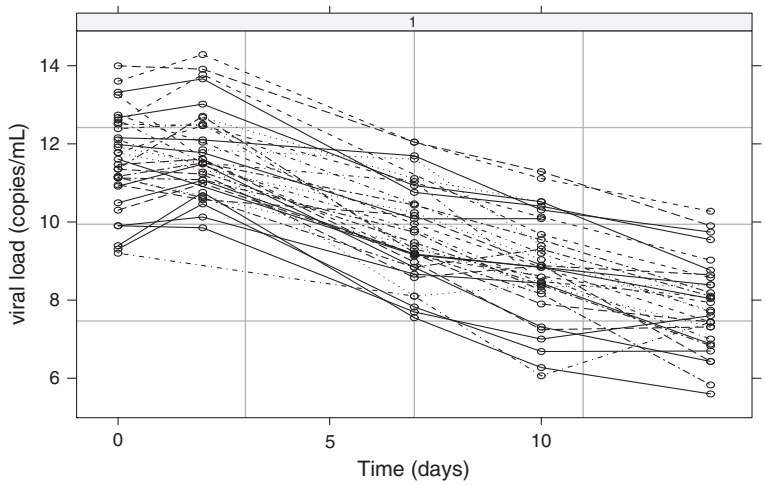

Fig. 2. The scatter plot of viral load (log scale) against time for 36 AIDS patients from the AIDS Clinical Trials Group 315 study.

Table 1. *The AIC and BIC values of the various criteria for ACTG 315 data*

Structure	aBIC	mBIC	mAIC	cAIC	jAIC
Diagonal	435.47	443.61	428.11	558.42	601.59
General PD	437.21	446.86	428.26	555.74	602.74

AIC, Akaike information criteria; BIC, Bayesian information criterion; ACTG, AIDS Clinical Trials Group; aBIC, adjusted BIC; mBIC, marginal BIC; mAIC, marginal AIC; cAIC, conditional AIC; jAIC, joint AIC; PD, positive definite.

Table 2. *The estimated values and se of the intercept and slope using LME with the diagonal and general PD covariance structures for ACTG 315 data*

Structure	Intercept (se)	Slope (se)	Mean prediction error (se)
Diagonal	11.8816(0.1892)	−0.2912(0.0117)	0.6518 (0.0844)
General PD	11.8828(0.2010)	−0.2916(0.0124)	0.7982 (0.1049)

se, standard errors; LME, linear mixed-effects; PD, positive definite; ACTG, AIDS Clinical Trials Group.

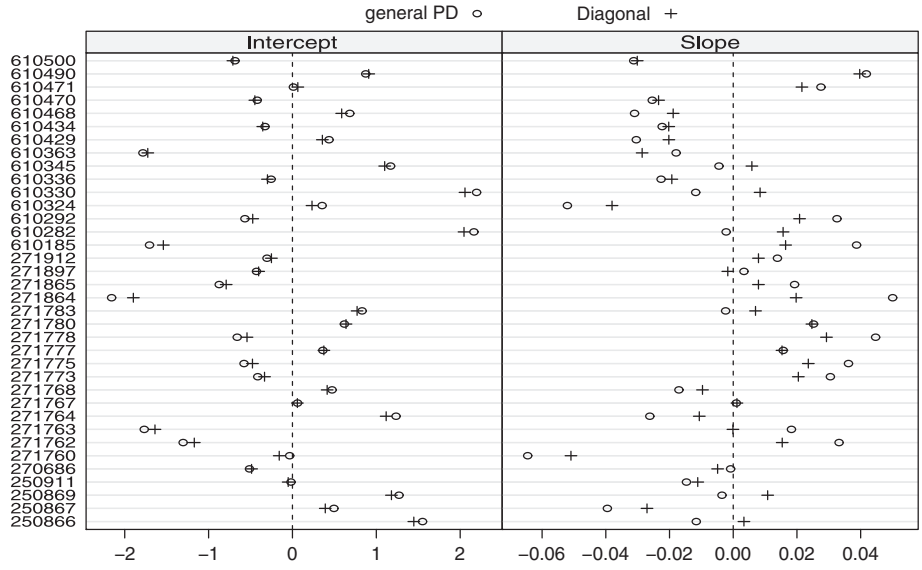

Fig. 3. The random estimates of the intercept (left panel) and slope (right panel) with the diagonal (+) and general positive definite (PD) (∘) covariance matrices from a linear mixed-effects fit of the AIDS data. The numbers in the left margin are 'ids' of patients.

that the diagonal structure is preferable. Recalling the simulation performance that shows that the cAIC is not appropriate in selection of covariance structure of random effects, we suggest that the diagonal structure be selected. Now let us see the estimated values of the intercept and slope along these two structures, which we present in the middle two columns of Table 2. The estimated values with these two structures are similar, but the standard errors based on the diagonal structure are smaller than those based on the general PD structure, which indicates that the fixed estimates based on the diagonal covariance matrix are more efficient than those based on the general PD structure. Furthermore, it is interesting to observe that the predictors of random intercept and slope based on the diagonal structure are mostly closer to zero than those based on the general PD structure (shown in Fig. 3). Last, we examine the predictive power of the two models with diagonal and general PD random effect covariance matrices, respectively. Specifically, we exclude the last observation of each patient from the original sample as the testing sample and do estimation based on the left observations. The last column of Table 2 shows the mean prediction errors over the testing sample and their standard errors, from which we see that the model with diagonal structure has more precise prediction than the model with general PD structure. This supports the choice of diagonal structure in this example again.

5. Discussion and summary

To obtain a more appropriate selection criterion for the covariance structure of random effects in the LME models, we have generalized Hurvich & Tsai (1989) approach and developed a new criterion – jAIC based on the joint likelihood of data and random effects. The jAIC takes the variance components of the random effects fully into account and can well help distinguish between models with different covariance structures of the random effects. Our criterion is nearly unbiased for estimating the expected Kullback–Leibler information and has asymptotic optimality. The proposed method has also been shown to be promising by a simulation study. In that study, we found that jAIC is much better than cAIC at selecting a model for the covariance matrix of the random effects. It is worth noting that the jAIC is developed to select structure of random effects but cannot be used to test the significance of random effects.

As in Liang *et al.* (2008), we made use of the integration by part technique, which has been utilized to obtain risk-unbiased estimators before (Lu & Berger, 1989; Stein, 1981), to derive the selection criterion for the covariance matrix of random effects. We expect that our method is applicable to model selection in other contexts such as non-parametric regression (Hurvich *et al.*, 1998) and single-index models (Naik & Tsai, 2001). For generalized mixed-effects models, the jAIC can be developed using the technique of Saefken *et al.* (2014). These warrant our future research. Variable selection for the LME models is also an important topic. When focusing on the choice of random effects, the jAIC can be regarded as the cAIC with the additional penalty term $\widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}}$ and thus should have the tendency to choose models with less random effects than the cAIC. Note that Greven & Kneib (2010) mentioned that mAIC chooses to incorporate random effects rarely. So it is expected that jAIC performs between cAIC and mAIC when focusing on the choice of random effects. Presently, we are applying the ideas in this paper to address the variable selection for the LME models. Last, extending our method to LME models with missing data is an interesting problem and remains our further research.

Acknowledgements

We thank the editor Holger Rootzen, the associate editor and two anonymous referees for their constructive comments and suggestions that greatly improved the original manuscript. This research was partially supported by the National Natural Science Foundation of China (NNSFC) (grant nos. 11228103 and 11529101). Zhang's work was partially supported by the NNSFC (grant nos. 11471324 and 71522004). Liang's work was partially supported by the NSF grants DMS-1440121 and DMS-1418042. Zou's work was partially supported by the NNSFC (grant nos. 11331011 and 11271355) and a grant from the Beijing High-level Talents Program.

Appendix

A.1. Derivation of formula (8)

The following lemma will be used in the proof of formula (8).

Lemma (Stein, 1981): Let a be an $N(0,1)$ random variable and $g : \mathcal{R}$ to $\mathcal{R}$ be an indefinite integral of the Lebesgue measurable function g', essentially the derivative of g. Suppose also that $E|g'(a)| < \infty$. Then $E[g'(a)] = E(ag(a))$.

We first write $\boldsymbol{\gamma}_0 = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}_0\mathbf{b}_0$ and $\widehat{\boldsymbol{\lambda}} = (\mathbf{y} - \widehat{\boldsymbol{\gamma}})/\widehat{\sigma}^2$. Recall $\widehat{\boldsymbol{\gamma}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$. From formula (6), we have

$$E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}}) = E_0\left\{N\log\widehat{\sigma}^2 + \frac{\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|^{2\,2}}{\widehat{\sigma}} + \frac{N\sigma_0^2}{\widehat{\sigma}^2} + r\log(2\pi) + \widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}}\right\} - N\log\sigma_0^2 - r_0\log 2\pi - r_0 - N. \tag{A.1}$$

A commonly used approach to obtain an (approximately) unbiased estimator of $E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}})$ is to make use of the (approximate) distributions of $\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|^2/\widehat{\sigma}^2$ and $\widehat{\sigma}^2$. See, for example, Hurvich & Tsai (1989, 1995), Hurvich *et al.* (1998) and Vaida & Blanchard (2005). For the LME models, however, such distributions might not be easy to find. Instead, we propose to use the following integration by part method. It is clear that

$$\begin{aligned}
&E_0\left[\frac{(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}})'(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}})}{\widehat{\sigma}^2}\right] \\
&= E_0\left[\frac{\|\mathbf{y} - \widehat{\boldsymbol{\gamma}}\|^2}{\widehat{\sigma}^2}\right] - 2\cdot\sum_{i=1}^{N} E_0\left[\frac{(Y_i - \gamma_{0i})(Y_i - \widehat{\gamma}_i)}{\widehat{\sigma}^2}\right] + \sum_{i=1}^{N} E_0\left[\frac{(Y_i - \gamma_{0i})^2}{\widehat{\sigma}^2}\right] \\
&= E_0\left[\frac{\|\mathbf{y} - \widehat{\boldsymbol{\gamma}}\|^2}{\widehat{\sigma}^2}\right] - 2\cdot\sum_{i=1}^{N} E_{\mathbf{b}_0}E_{\mathbf{y}|\mathbf{b}_0}\left[\frac{(Y_i - \gamma_{0i})(Y_i - \widehat{\gamma}_i)}{\widehat{\sigma}^2}\right] + \sum_{i=1}^{N} E_{\mathbf{b}_0}E_{\mathbf{y}|\mathbf{b}_0}\left[\frac{(Y_i - \gamma_{0i})^2}{\widehat{\sigma}^2}\right],
\end{aligned}$$

where Y_i, γ_{0i} and $\widehat{\gamma}_i$ denote the ith components of $\mathbf{y}$, $\boldsymbol{\gamma}_0$ and $\widehat{\boldsymbol{\gamma}}$, respectively. Note that conditionally given $\mathbf{b}_0$, $\mathbf{y}$ follows the distribution of $N(\boldsymbol{\gamma}_0, \sigma_0^2\mathbf{I}_N)$. Assuming that $\widehat{\lambda}_i = (Y_i - \widehat{\gamma}_i)/\widehat{\sigma}^2$ is a continuous function with piecewise continuous partial derivatives with respect to $\mathbf{y}$, it can be shown from Stein's lemma that

$$E_{\mathbf{y}|\mathbf{b}_0}\left[\frac{(Y_i - \gamma_{0i})(Y_i - \widehat{\gamma}_i)}{\widehat{\sigma}^2}\right] = E_{\mathbf{y}|\mathbf{b}_0}\left[(Y_i - \gamma_{0i})\widehat{\lambda}_i\right] = \sigma_0^2\cdot E_{\mathbf{y}|\mathbf{b}_0}\left[\frac{\partial\widehat{\lambda}_i}{\partial Y_i}\right],$$

provided the expectation on the right-hand side exists.

Similarly, assuming that $\partial(\widehat{\sigma}^{-2})/\partial Y_i$ $(i = 1, \ldots, N)$ are continuous functions with piecewise continuous partial derivatives and the corresponding expectations exist, we have

$$E_{\mathbf{y}|\mathbf{b}_0}\left[\frac{(Y_i - \gamma_{0i})^2}{\widehat{\sigma}^2}\right] = \sigma_0^2\cdot E_{\mathbf{y}|\mathbf{b}_0}\left(\frac{1}{\widehat{\sigma}^2}\right) + \sigma_0^4\cdot E_{\mathbf{y}|\mathbf{b}_0}\left[\frac{\partial^2(\widehat{\sigma}^{-2})}{\partial Y_i^2}\right].$$

So,

$$E_0\left[\frac{\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|^2}{\widehat{\sigma}^2}\right] = E_0\left[\frac{\|\mathbf{y} - \widehat{\boldsymbol{\gamma}}\|^2}{\widehat{\sigma}^2} - 2\sigma_0^2\sum_{i=1}^{N}\frac{\partial\widehat{\lambda}_i}{\partial Y_i} + \frac{N\sigma_0^2}{\widehat{\sigma}^2} + \sigma_0^4\sum_{i=1}^{N}\frac{\partial^2(\widehat{\sigma}^{-2})}{\partial Y_i^2}\right].$$

Substituting this formula in (A.1) and after some calculations, we obtain

$$E_0\Delta^*(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{b}}) = E_0\left\{N\log(2\pi) + N\log\widehat{\sigma}^2 + \frac{\|\mathbf{y} - \widehat{\boldsymbol{\gamma}}\|^2}{\widehat{\sigma}^2} + 2\sigma_0^2\cdot\Phi(\mathbf{y}) + \sigma_0^4\cdot\Psi(\mathbf{y}) + r\log(2\pi) + \widehat{\mathbf{b}}'\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}} - N\log 2\pi - N\log\sigma_0^2 - N - r_0 - r_0\log 2\pi\right\}. \tag{A.2}$$

In a consequence, formula (8) follows.

A.2. Conditions and proof of theorem 1

In what follows, $\max_i$ ($\min_i$), $\max_s$ ($\min_s$) and $\max_j$ ($\min_j$) indicate the maximization (minimization) over $i \in \{1, \ldots, n\}$, $s \in \{1, \ldots, S\}$ and $j \in \{1, \ldots, J_s\}$, respectively. Assume that as $N \to \infty$, $\widehat{\boldsymbol{\eta}}_s \xrightarrow{p} \boldsymbol{\eta}_s^*$ and the limit of $\widehat{\sigma}_s^2$ is ${\sigma_s^2}^* > 0$. Denote $\widehat{\eta}_{s,j}$ and $\eta_{s,j}^*$ as the jth elements of $\widehat{\boldsymbol{\eta}}_s$ and $\boldsymbol{\eta}_s^*$, respectively. For $s = 1, \ldots, S$ and $j = 1, \ldots, J_s$, we can write $\partial \widehat{\eta}_{s,j}/\partial \mathbf{y} = \widehat{\mathbf{T}}_{s,j}\mathbf{y}$ almost surely, where $\widehat{\mathbf{T}}_{s,j}$ can be random and depend on $\mathbf{y}$. Write $\widehat{\Omega}_s = \partial^2 \widehat{\sigma}_s^2/(\partial \mathbf{y} \partial \mathbf{y}')$ and $\Sigma_0 = \sigma_0^2 \mathbf{I}_N + \mathbf{Z}_0 \mathbf{G}_0 \mathbf{Z}_0'$ that is the covariance matrix of $\mathbf{y}$. Let $\widehat{\mathbf{M}}_s = \mathbf{I}_N - \widehat{\mathbf{P}}_s$, $\mathbf{V}_s$ have the same form as $\widehat{\mathbf{V}}_s$ except that the notation $\widehat{\ }$ is removed and $\mathbf{V}_s^* = \mathbf{V}_s \mid_{\boldsymbol{\eta} = \boldsymbol{\eta}^*}$. Similarly, we can define $\mathbf{P}_s$, $\mathbf{M}_s$, $\mathbf{G}_s^*$, Σ_s^*, $\mathbf{P}_s^*$ and $\mathbf{M}_s^*$. Further, for $j \in \{1, \ldots, J_s\}$, denote $\mathbf{W}_{s,j} = \frac{\partial \mathbf{G}_s}{\partial \omega_{s,j}}$ where $\omega_{s,j}$ is the jth element of $\boldsymbol{\omega}_s$, $\widehat{\mathbf{W}}_{s,j} = \mathbf{W}_{s,j} \mid_{\boldsymbol{\omega}_s = \widehat{\boldsymbol{\omega}}_s}$, $\boldsymbol{\gamma}_s^* = \mathbf{P}_s^* \mathbf{y}$, $H_s = E_0 \|\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_s^*\|^2$, and $\xi = \min_s H_s$. $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum singular values for a matrix A. All limiting processes discussed here are with respect to $N \to \infty$. c, $\bar{c}$, $\tilde{c}$, c^* and $c^\star$ are all generic positive constants. The following conditions are imposed to obtain theorem 1.

A.2.1. Conditions

Condition 1. $\sum_{s=1}^{S} H_s^{-1} \to 0$.

Condition 2. $r_0 \xi^{-1} \to 0$, $(p + r)\xi^{-1} \to 0$, and $N\xi^{-2} \to 0$.

Condition 3. $\|\mathbf{y}\|^2 = O_p(N)$.

Condition 4. $N\xi^{-1} \max_s \lambda_{\max}(\mathbf{P}_s^* - \widehat{\mathbf{P}}_s) \xrightarrow{p} 0$.

Condition 5. $N\xi^{-1} \max_s \max_j \lambda_{\max}(\widehat{\mathbf{T}}_{s,j}) \xrightarrow{p} 0$ and $N\xi^{-1} \max_s \lambda_{\max}(\widehat{\Omega}_s) \xrightarrow{p} 0$.

Condition 6. $\max_s \max_{j \in \{1, \ldots, J_s - 1\}} \lambda_{\max}(\widehat{\mathbf{W}}_{s,j}) \le \bar{c} < \infty$.

Condition 7. $\min_s \lambda_{\min}(\widehat{\mathbf{G}}_s) \ge \tilde{c} > 0$ or $\max_s \left[\lambda_{\max}(\mathbf{Z}\mathbf{Z}')\right] \le c^\star < \infty$.

Condition 8. $\max_s \lambda_{\max}(\Sigma_s^{*-1/2} \Sigma_0 \Sigma_s^{*-1/2}) \le c^* < \infty$.

Condition 1 is a standard condition for asymptotic optimality of model selection, and the similar conditions are used in the literature like (A.3) in Li (1987) and (2.6) in Shao (1997). Condition 2 contains restrictions on the increasing rates of the numbers of fixed effects and random effects and ξ as $N \to \infty$. Restrictions similar to them can be found in Shibata (1980) and Newey (1997). Condition 3, which concerns the sum of y_i^2 with $i \in \{1, .., N\}$, is quite common and reasonable.

Condition 4 requires that $\widehat{\boldsymbol{\eta}}_s$ converge to $\boldsymbol{\eta}_s^*$ at a rate such that $\xi^{-1} \max_s \lambda_{\max}(\mathbf{P}_s^* - \widehat{\mathbf{P}}_s)$ converges to 0 quicker than $N \to \infty$. For $s \in \{1, \ldots, S\}$ and $j_1, j_2 \in \{1, \ldots, J_s\}$, we define Υ_s as an $N \times N$ matrix with the $i_1 i_2$th element $\Upsilon_{s,i_1,i_2} = \sum_{j_1=1}^{J_s} \sum_{j_2=1}^{J_s} (\widehat{\eta}_{s,j_1} - \eta_{s,j_1}^*)(\widehat{\eta}_{s,j_2} - \eta_{s,j_2}^*) \partial^2 \mathbf{P}_{s,i_1,i_2} / (\partial \eta_{s,j_1} \partial \eta_{s,j_2}) \mid_{\boldsymbol{\eta}_s = \widetilde{\boldsymbol{\eta}}_s^{i_1,i_2}}$, where $\mathbf{P}_{s,i_1,i_2}$ is the $i_1 i_2$th element of $\mathbf{P}_s$ and $\widetilde{\boldsymbol{\eta}}_s^{i_1,i_2}$ is a $J_s \times 1$ vector between $\widehat{\boldsymbol{\eta}}_s$ and $\boldsymbol{\eta}_s^*$. From the formulas (S.13) and (S.25)–(S.27) in the proof of

theorem 1 in Supporting Information, we see that when conditions 2, 6 and 7 hold, condition 4 is implied by $\lambda_{\max}(\Upsilon_s) = O_p(N^{-1/2})$ and $\sqrt{N}(\widehat{\eta}_{s,j} - \eta^*_{s,j}) = O_p(1)$, which are the common convergence rates.

Condition 5 places constraints on the robustness of the estimators $\widehat{\boldsymbol{\eta}}_s$. Take the last element of $\widehat{\boldsymbol{\eta}}_s$, $\widehat{\sigma}^2_s$, as an example. In the process of ML estimation, $\widehat{\sigma}^2_s$ is calculated by $\widehat{\sigma}^2_s = \mathbf{y}'\mathbf{M}'_s\mathbf{M}_s\mathbf{y}/N$. Then the restrictions related to $\widehat{\sigma}^2_s$ in condition 5 are obviously implied by conditions 2 and 3.

Condition 6 is on the derivative of $\widehat{\mathbf{G}}_s$. Consider a very common case with $\omega_{s,j}$'s directly being the elements of $\mathbf{G}_s$. In this situation, $\mathrm{tr}\left(\frac{\partial \mathbf{D}_s}{\partial \omega_{s,j}} \frac{\partial \mathbf{D}'_s}{\partial \omega_{s,j}}\right)$ is finite, and so condition 6 holds.

The first part of condition 7 excludes degenerate estimated distribution of $\mathbf{b}$, which is analogous to condition (A.2*) of Andrews (1995) and condition (A.1) of Hansen & Racine (2012). The second part of condition 7 holds, for example, in a situation with m_i being bounded and $\lambda_{\max}(\mathbf{Z}_i\mathbf{Z}'_i) = O(m_i)$ uniformly for $i \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, S\}$. Condition 8 places constraint on the relation between Σ_s's and the true covariance matrix of $\mathbf{y}$. Analogous conditions have been imposed by Yang (2004) and Yuan & Yang (2005) for linear regression.

A.2.2. Proof of theorem 1.

Let $R_s = \widehat{\sigma}^{-2}_s H_s$,

$$\Pi_s = R_s + N\left(\frac{\sigma^2_0}{\widehat{\sigma^2_s}} + \log\frac{\widehat{\sigma^2_s}}{\sigma^2_0} - 1\right) + \widehat{\mathbf{b}}'_s\widehat{\mathbf{G}}^{-1}_s\widehat{\mathbf{b}}_s + r\log 2\pi,$$

$$\mathbf{U}_s = 2\widehat{\sigma}^{-2}_s(\mathbf{y} - \widehat{\boldsymbol{\gamma}}_s)\frac{\partial\widehat{\sigma^2_s}}{\partial\mathbf{y}} - \mathrm{tr}\frac{\partial^2\widehat{\sigma^2_s}}{\partial\mathbf{y}\partial\mathbf{y}'} + 2\widehat{\sigma}^{-2}_s\frac{\partial\widehat{\sigma^2_s}}{\partial\mathbf{y}'}\frac{\partial\widehat{\sigma^2_s}}{\partial\mathbf{y}},$$

and

$$\widehat{\mathrm{jAIC}}^*_s = \mathrm{jAIC}^*_s - N\log 2\pi - N\log\sigma^2_0 - N - r_0\log 2\pi - r_0,$$

where $-N\log 2\pi - N\log\sigma^2_0 - N - r_0\log 2\pi - r_0$ has nothing to do with s. So, $\widehat{s} = \underset{s\in\{1,\ldots,S\}}{\arg\min}\ \widehat{\mathrm{jAIC}}^*_s$. From (10) and (11), we have

$$\begin{aligned}
&\Delta^*_s(\widehat{\boldsymbol{\theta}}_s, \widehat{\mathbf{b}}_s) - \widehat{\mathrm{jAIC}}^*_s \\
&= -N\log 2\pi + N\sigma^2_0\widehat{\sigma}^{-2}_s - N\log\sigma^2_0 - N - r_0\log 2\pi - r_0 \\
&\quad + \widehat{\sigma}^{-2}_s\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_s\|^2 - \widehat{\sigma}^{-2}_s\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_s + \varepsilon_0\|^2 - 2\mathrm{tr}\frac{\partial\widehat{\boldsymbol{\gamma}}_s}{\partial\mathbf{y}'} - \mathbf{U}_s \\
&\quad + N\log 2\pi + N\log\sigma^2_0 + N + r_0\log 2\pi + r_0 \\
&= \widehat{\sigma}^{-2}_s(N\sigma^2_0 - \|\varepsilon_0\|^2) - 2\widehat{\sigma}^{-2}_s(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_s)'\varepsilon_0 - 2\mathrm{tr}\frac{\partial\widehat{\boldsymbol{\gamma}}_s}{\partial\mathbf{y}'} - \mathbf{U}_s
\end{aligned} \tag{A.3}$$

and

$$\Delta^*_s(\widehat{\boldsymbol{\theta}}_s, \widehat{\mathbf{b}}_s) - \Pi_s = \widehat{\sigma}^{-2}_s\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_s\|^2 - R_s - r_0(1 + \log 2\pi). \tag{A.4}$$

From conditions 1–8, we can show that (see Section S.2 of Supporting Information for detailed proofs)

$$\max_s\frac{r_0}{\Pi_s} \xrightarrow{p} 0, \tag{A.5}$$

$$\max_s \frac{\widehat{\sigma}_s^{-2}\left|N\sigma_0^2 - \|\varepsilon_0\|^2\right|}{\Pi_s} \xrightarrow{p} 0, \tag{A.6}$$

$$\max_s \frac{\widehat{\sigma}_s^{-2}\left|(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_s)'\varepsilon_0\right|}{\Pi_s} \xrightarrow{p} 0, \tag{A.7}$$

$$\max_s \frac{\widehat{\sigma}_s^{-2}\left|\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}_s\|^2 - H_s\right|}{\Pi_s} \xrightarrow{p} 0, \tag{A.8}$$

$$\max_s \frac{|\mathbf{U}_s|}{\Pi_s} \xrightarrow{p} 0, \tag{A.9}$$

and

$$\max_s \frac{|\mathrm{tr}(\partial\widehat{\boldsymbol{\gamma}}_s/\partial\mathbf{y})|}{\Pi_s} \xrightarrow{p} 0. \tag{A.10}$$

Now, by (A.3)–(A.10), we have

$$\max_s \frac{|\Delta_s^*(\widehat{\boldsymbol{\theta}}_s, \widehat{\mathbf{b}}_s) - \widetilde{\mathrm{jAIC}}_s^*|}{\Pi_s} \xrightarrow{p} 0 \quad \text{and} \quad \max_s \frac{|\Delta_s^*(\widehat{\boldsymbol{\theta}}_s, \widehat{\mathbf{b}}_s) - \Pi_s|}{\Pi_s} \xrightarrow{p} 0, \tag{A.11}$$

which, along with the proof of theorem 2.1 in Li (1987), imply (12).

References

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

Andrews, D. W. K. (1995). Asymptotic optimality of generalized c_l, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* **4**, 359–377.

Chen, Z. & Dunson, D. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.

Davidian, M. & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*, Chapman and Hall, New York.

Delattre, M., Lavielle, M. & Poursat, M. A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics* **8**, 456–475.

Greven, S. & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773–789.

Hansen, B. & Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics* **167**, 38–46.

Ho, D., Neumann, A., Perelson, A., Chen, W., Leonard, J. & Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–126.

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

Hurvich, C. & Tsai, C. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51**, 1077–1084.

Hurvich, C. M., Simonoff, J. S. & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271–293.

Kinney, S. & Dunson, D. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63**, 690–698.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lederman, M., Connick, E., Landay, A., Kuritzkes, D., Spritzkes, J., St Clair, M., Kotzin, B., Fox, L., Chiozzi, M., Leonard, J., Rousseau, F., Wade, M., Roe, J., Martinez, A. & Kessler, H. (1998). Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS clinical trials group protocol 315. *The Journal of Infectious Diseases* **178**, 70–79.

Li, K. C. (1987). Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* **15**, 958–975.
Li, L., Brown, M., Lee, K. & Gupta, S. (2002). Estimation and inference for a spline-enhanced nonlinear population pharmacokinetic model. *Biometrics* **58**, 601–611.
Liang, H., Wu, H. L. & Zou, G. H. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95**, 773–778.
Lu, K. L. & Berger, J. O. (1989). Estimation of normal means: frequentist estimation of loss. *The Annals of Statistics* **17**, 890–906.
Naik, P. & Tsai, C. (2001). Single-index model selections. *Biometrika* **61**, 821–832.
Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**, 147–168.
Ngo, L. & Brand, R. (2002). Model selection in linear mixed effects models using SAS Proc Mixed. *SAS Global Forum* **22**.
Pinheiro, J. & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*, Springer Science, New York.
Saefken, B., Kneib, T., van Waveren, C. & Greven, S. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics* **8**, 201–225.
Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–264.
Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics* **8**, 147–164.
Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9**, 1135–1151.
Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–370.
Verbeke, G. & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*, Springer, New York.
Vonesh, E. & Chinchilli, V. (1996). *Linear and nonlinear models for the analysis of repeated measurements*, Marcel Dekker, New York.
Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifsonparallel, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. & Shaw, G. M. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117–122.
Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and Its Applications* **5**, 302–305.
Wu, H. & Ding, A. (1999). Population HIV-1 dynamics *in vivo*: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410–418.
Wu, H. & Zhang, J. T. (2006). *Nonparametric regression methods for longitudinal data analysis. Wiley Series in Probability and Statistics. Hoboken*, Wiley-Interscience [John Wiley & Sons], NJ.
Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory* **20**, 176–222.
Ye, J. M. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93**, 120–131.
Yu, D. & Yau, K. K. W. (2012). Conditional Akaike information criterion for generalized linear mixed models. *Computational Statistics & Data Analysis* **56**, 629–644.
Yuan, Z. & Yang, Y. (2005). Combining linear regression models: when and how? *Journal of the American Statistical Association* **100**, 1202–1214.

Received September 2014, in final form July 2015

Xinyu Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
E-mail: xinyu@amss.ac.cn

Supporting information

Additional supporting information may be found in the online version of this article at the publishers web site.