Biometrika (2014), **101**, 1, *pp.* 205–218
Printed in Great Britain
doi: 10.1093/biomet/ast052
Advance Access publication 7 February 2014

Model averaging and weight choice in linear mixed-effects models

BY XINYU ZHANG, GUOHUA ZOU

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
xinyu@amss.ac.cn ghzou@amss.ac.cn

AND HUA LIANG

Department of Statistics, George Washington University, Washington, District of Columbia 20052, U.S.A.
hliang@gwu.edu

SUMMARY

This article studies model averaging for linear mixed-effects models. We establish an unbiased estimator of the squared risk for the model averaging, and use the estimator as a criterion for choosing weights. The resulting model average estimator is proved to be asymptotically optimal under some regularity conditions. Simulation experiments show it is superior or comparable to estimators based on the final models selected by the commonly-used methods and some existing averaging procedures. The proposed procedure is applied to data from an AIDS clinic trial.

Some key words: Asymptotic optimality; Conditional Akaike information criterion; Model averaging; Squared loss.

1. INTRODUCTION

Model selection procedures generally ignore the uncertainty involved in the selection step. The corresponding results may be misleading (Buckland et al., 1997; Hjort & Claeskens, 2003; Claeskens & Hjort, 2008), and also may ignore useful information from the form of the relationship between response and covariates (Bates & Granger, 1969). Furthermore, different selection criteria may yield different final models or several models with very close criterion values (Miller, 2002). Various attempts have been made to solve these problems. Remarkable results of these attempts are the work on model averaging such as Bayesian (Raftery et al., 1997; Hoeting et al., 1999) and frequentist model averaging (Buckland et al., 1997; Yang, 2001, 2003; Hansen, 2007). Hjort & Claeskens (2003) systematically discussed the advantages of weighting estimators across several models, proposed a general framework for frequentist model averaging, and established large-sample properties of estimators. Following their work, efforts have been invested to study model averaging in, for example, semiparametric models (Claeskens & Carroll, 2007) and generalized partially linear additive models (Zhang & Liang, 2011).

The choice of weight in model averaging is fundamental and important because it determines whether the resulting estimator has superior performance. Buckland et al. (1997) suggested taking weights based on AIC and BIC scores of candidate models. A similar strategy was advocated in Hjort & Claeskens (2003) for their model-averaging procedure based on the focused information criterion. Hansen (2007) suggested weights that minimize the Mallows criterion and proved

that the resulting model average estimator minimizes the squared error in large samples. More recently, under linear regression models, Liang et al. (2011) introduced a general random weight form that includes smoothed AIC, smoothed BIC and many other weights as special cases.

Most existing work on model averaging focuses on cross-sectional observations, but we face the concerns with model selection mentioned above for longitudinal data using linear mixed-effects models (Laird & Ware, 1982). The situation is actually more serious because both fixed and random effects and their covariance structure need to be determined. In this paper, we develop a data-driven model-averaging procedure, and show that it is asymptotically optimal in the sense that the corresponding squared error is asymptotically identical to that of the infeasible best possible model average estimator. We also illustrate numerically that the proposed method is superior or comparable to commonly-used model averaging or selection methods.

2. MODEL AVERAGING AND WEIGHT CHOICE

Consider the linear mixed-effects model

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \qquad b_i \sim N(0, D), \quad \varepsilon_i \sim N(0, R_i) \quad (i = 1, \ldots, n), \tag{1}$$

where β is a $p \times 1$ vector of fixed regression coefficients, b_i is a $k \times 1$ vector of random coefficients specific to the cluster i, ε_i is an $m_i \times 1$ error vector independent of b_i, and y_i, X_i, and Z_i are the response vector and covariate matrices for the fixed and random effects related to the ith cluster, respectively. Both the covariance matrices D and R_i may have special structures, and in this paper we assume that $R_i = \sigma^2 I_{m_i}$, where I_{m_i} is an $m_i \times m_i$ identity matrix. Let $N = \sum_{i=1}^{n} m_i$ be the total number of observations. Write the model in matrix notation:

$$y = \mu + \varepsilon = X\beta + Zb + \varepsilon, \quad b \sim N(0, G), \tag{2}$$

where $y = (y_1', \ldots, y_n')'$ is an $N \times 1$ vector of observations with mean μ conditional on b, $X = (X_1', \ldots, X_n')'$ is an $N \times p$ matrix, $Z = \mathrm{diag}(Z_1, \ldots, Z_n)$ is an $N \times r$ block-diagonal matrix, $r = nk$, $b = (b_1', \ldots, b_n')'$, $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_n')'$, and $G = \mathrm{diag}(D, \ldots, D)$ is an $r \times r$ block-diagonal matrix.

Assume that a series of candidate linear mixed-effects models, of form

$$y = X_{(s)}\beta_{(s)} + Z_{(s)}b_{(s)} + \varepsilon, \quad b_{(s)} \sim N(0, G_{(s)}) \qquad (s = 1, \ldots, S), \tag{3}$$

are used to approximate (2), where $X_{(s)}$ and $Z_{(s)}$ are covariate matrices in the sth candidate model, $\beta_{(s)}$ and $b_{(s)}$ are the corresponding fixed and random coefficients, $G_{(s)} = \mathrm{diag}(D_{(s)}, \ldots, D_{(s)})$ with $D_{(s)}$ the $k_s \times k_s$ covariance matrix of random effect for the cluster i, and $X_{(s)}$ has full column rank.

Given σ^2 and $G_{(s)}$, the fixed effects $\beta_{(s)}$ and random effects $b_{(s)}$ under the sth candidate model can be estimated by the best linear unbiased estimators (Robinson, 1991),

$$\widetilde{\beta}_{(s)} = (X_{(s)}'\Sigma_{(s)}^{-1}X_{(s)})^{-1}X_{(s)}'\Sigma_{(s)}^{-1}y, \qquad \widetilde{b}_{(s)} = G_{(s)}Z_{(s)}'\Sigma_{(s)}^{-1}(y - X_{(s)}\widetilde{\beta}_{(s)}), \tag{4}$$

where $\Sigma_{(s)} = \sigma^2 I_N + Z_{(s)}G_{(s)}Z_{(s)}'$. Now let $\hat{D}_{(s)}$ be an estimator of $D_{(s)}$. Write $\hat{G}_{(s)} = \mathrm{diag}(\hat{D}_{(s)}, \ldots, \hat{D}_{(s)})$, $\hat{V}_{(s)} = \hat{\Sigma}_{(s)}^{-1/2}X_{(s)}(X_{(s)}'\hat{\Sigma}_{(s)}^{-1}X_{(s)})^{-1}X_{(s)}'\hat{\Sigma}_{(s)}^{-1/2}$, $\hat{\Sigma}_{(s)} = \sigma^2 I_N + Z_{(s)}\hat{G}_{(s)}Z_{(s)}'$, and $\hat{P}_{(s)} = I_N - \sigma^2\hat{\Sigma}_{(s)}^{-1/2}(I_N - \hat{V}_{(s)})\hat{\Sigma}_{(s)}^{-1/2}$. From (4), when $G_{(s)}$ is unknown,

$\beta_{(s)}$ and $b_{(s)}$ can be estimated by

$$\hat{\beta}_{(s)} = (X'_{(s)}\hat{\Sigma}^{-1}_{(s)}X_{(s)})^{-1}X'_{(s)}\hat{\Sigma}^{-1}_{(s)}y, \qquad \hat{b}_{(s)} = \hat{G}_{(s)}Z'_{(s)}\hat{\Sigma}^{-1}_{(s)}(y - X_{(s)}\hat{\beta}_{(s)}),$$

and thus μ can be estimated by $\hat{\mu}_{(s)} = X_{(s)}\hat{\beta}_{(s)} + Z_{(s)}\hat{b}_{(s)}$. Direct manipulation yields

$$\hat{\mu}_{(s)} = \hat{P}_{(s)}y, \qquad \hat{P}_{(s)}X_{(s)} = X_{(s)}.$$

From (A9) of the Appendix, we see that $I_N - \sigma^2\hat{\Sigma}^{-1}_{(s)}$ is nonnegative-definite, so $\hat{P}_{(s)} = \sigma^2\hat{\Sigma}^{-1/2}_{(s)}\hat{V}_{(s)}\hat{\Sigma}^{-1/2}_{(s)} + I_N - \sigma^2\hat{\Sigma}^{-1}_{(s)}$ is also nonnegative-definite. Let $\hat{A}_{(s)} = I_N - \hat{P}_{(s)}$. Then $\hat{A}_{(s)} = \sigma^2\hat{\Sigma}^{-1/2}_{(s)}(I_N - \hat{V}_{(s)})\hat{\Sigma}^{-1/2}_{(s)}$ is also nonnegative-definite because $\hat{V}_{(s)}$ is idempotent.

Given a weight vector $w = (w_1, \ldots, w_S)'$ belonging to the set $\mathcal{W} = \{w \in [0, 1]^S : \sum_{s=1}^S w_s = 1\}$, the corresponding model average estimator of μ can be expressed as

$$\hat{\mu}(w) = \sum\nolimits_{s=1}^S w_s\hat{\mu}_{(s)} = \sum\nolimits_{s=1}^S w_s\hat{P}_{(s)}y \equiv \hat{P}(w)y.$$

Remark 1. If there are many candidate models, the computational burden of model averaging will be large and a model screening step prior to model averaging will be desirable. Screening using AIC and BIC has been advocated by Yuan & Yang (2005), and stepwise screening has been used in Claeskens et al. (2006) and Zhang et al. (2012).

When the focus is on clusters, any future prediction takes place in the same clusters as the observed data, and the random effects for these clusters are held constant (Donohue et al., 2011). Thus, we define the squared loss and risk of $\hat{\mu}(w)$ as $L(w) = \|\hat{\mu}(w) - \mu\|^2$ and $R(w) = E_{y|b}\{\|\hat{\mu}(w) - \mu\|^2\}$, respectively. Let $\hat{A}(w) = I_N - \hat{P}(w)$. Noting that $\varepsilon \sim N(0, \sigma^2 I_N)$, by integration by parts as in the derivation of Theorem 1 of Liang et al. (2008), we have

$$R(w) = E_{y|b}\{\|\hat{\mu}(w) - \mu\|^2\} = E_{y|b}\{\|\hat{A}(w)y\|^2 + 2\sigma^2 w'\rho - n\sigma^2\}, \tag{5}$$

where ρ is an $S \times 1$ vector with sth element $\rho_s = \mathrm{tr}\{\partial(\hat{P}_{(s)}y)/\partial y'\}$.

Motivated by the expression given in (5), we propose the following criterion for weight choice:

$$\hat{C}(w) = \|\hat{A}(w)y\|^2 + 2\sigma^2 w'\rho. \tag{6}$$

Let $\hat{w} = \mathrm{argmin}_{w\in\mathcal{W}}\hat{C}(w)$. Then the resulting model average estimator is $\hat{\mu}(\hat{w})$. When the $\Sigma_{(s)}$ is known, write $P_{(s)} = \hat{P}_{(s)}\,|_{\hat{\Sigma}_{(s)}=\Sigma_{(s)}}$, $P(w) = \sum_{s=1}^S w_s P_{(s)}$, and $A(w) = I_N - P(w)$. In this case, $\hat{C}(w)$ simplifies to $\|A(w)y\|^2 + 2\sigma^2\mathrm{tr}\{P(w)\}$, which has the same form as Mallows criterion in linear regression models (Hansen, 2007).

The derivatives $\partial(\hat{P}_{(s)}y)/\partial y'$ $(s = 1, \ldots, S)$ appear in $\hat{C}(w)$. We may use numerical methods to calculate these derivatives. Such calculations are cumbersome, so we propose an alternative using results from Greven & Kneib (2010).

Let $\theta_{(s)} = (\theta_{(s),1}, \ldots, \theta_{(s),J_{(s)}})'$ be a $J_{(s)} \times 1$ vector containing all unknown parameters in $D_{(s)}$ and $\hat{\theta}_{(s)}$ be its estimator. It follows from Theorem 3 of Greven & Kneib (2010) that

$$\rho_s = \mathrm{tr}(\hat{P}_{(s)}) + \sum_{j=1}^{J_{(s)}} \frac{\partial\hat{\theta}_{(s),j}}{\partial y'}\hat{A}_{(s)}\hat{W}_{(s),j}\hat{A}_{(s)}y, \tag{7}$$

where $\hat{\theta}_{(s),j}$ is the jth element of $\hat{\theta}_{(s)}$ and $\hat{W}_{(s),j} = \partial\Sigma_{(s)}(\theta_{(s)})/\partial\theta_{(s),j}\,|_{\theta_{(s)}=\hat{\theta}_{(s)}}$.

We now consider a common case where $\theta_{(s)}$ consists of the elements of matrix $D_{(s)}$. We use maximum likelihood to estimate $\theta_{(s)}$. Furthermore, assume that some elements of $\hat{\theta}_{(s)}$, say the last u elements, are zeros. Denote $\hat{\eta}_{(s)} = (\hat{\theta}_{(s),1}, \ldots, \hat{\theta}_{(s),J_{(s)}-u})'$. Using Theorem 3 of Greven & Kneib (2010), we have

$$\frac{\partial \hat{\theta}_{(s),j}}{\partial y'} = 0, \ \ j = J_{(s)} - u + 1, \ldots, J_{(s)}, \ \ \frac{\partial \hat{\eta}_{(s)}}{\partial y'} = H_{(s)}^{-1} \widetilde{H}_{(s)}, \tag{8}$$

where $H_{(s)}$ is a $(J_{(s)} - u) \times (J_{(s)} - u)$ matrix with (i, j)th element

$$\begin{aligned} &- \operatorname{tr}(\hat{\Sigma}_{(s)}^{-1} \hat{W}_{(s),i} \hat{\Sigma}_{(s)}^{-1} \hat{W}_{(s),j}) - N y' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} y y' \hat{A}_{(s)} \hat{W}_{(s),j} \hat{A}_{(s)} y (y' \hat{A}_{(s)} y)^{-2} \\ &\quad + 2N y' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} \hat{W}_{(s),j} \hat{A}_{(s)} y (y' \hat{A}_{(s)} y)^{-1}, \end{aligned} \tag{9}$$

and $\widetilde{H}_{(s)}$ is a $(J_{(s)} - u) \times N$ matrix with ith row

$$\widetilde{H}_{(s),i\cdot} = 2N \left\{ -(y' \hat{A}_{(s)} y)^{-2} y' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} y y' \hat{A}_{(s)} + (y' \hat{A}_{(s)} y)^{-1} y' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} \right\}. \tag{10}$$

Expressions (7)–(10) Give Explicit formulae for calculating ρ_s, so the minimization of $\hat{C}(w)$ is easily managed.

The following theorem shows the asymptotic optimality of $\hat{\mu}(\hat{w})$. All limiting processes discussed in this section and the Appendix are with respect to $N \to \infty$, which applies to the case where $n \to \infty$ and m_i is bounded uniformly for $i \in \{1, \ldots, n\}$.

THEOREM 1. *Under Assumption* A1 *in the Appendix,* $L(\hat{w})/\{\inf_{w \in \mathcal{W}} L(w)\} \to 1$ *in probability.*

Theorem 1 indicates that the squared loss based on the weight vector $\hat{w}$ is asymptotically identical to that obtained using the infeasible optimal weight vector, even if (2) is not among the candidate models in (3).

There remains an unknown parameter σ^2 in the criterion (6). One may develop an unbiased estimator of the squared risk by taking the estimation of σ^2 into account; see Liang et al. (2008). An alternative is that we insert an estimator $\hat{\sigma}^2$ of σ^2, obtained using the candidate model with the largest degree of freedom, like the Mallows criterion of Hansen (2007). This is computationally easier, and Greven & Kneib (2010) have shown a close agreement between these two options for model selection. We therefore adopt the second and propose the following criterion for weight choice,

$$\widehat{C}_{\hat{\sigma}}(w) = \| \hat{A}_{\hat{\sigma}}(w) y \|^2 + 2\hat{\sigma}^2 w' \hat{\rho},$$

where $\hat{A}_{\hat{\sigma}}(w) = \hat{A}(w)|_{\sigma^2 = \hat{\sigma}^2}$ and $\hat{\rho}$ is an $S \times 1$ vector with sth element $\hat{\rho}_s = \operatorname{tr}\{\partial(\hat{P}_{(s)} y)/\partial y' |_{\sigma^2 = \hat{\sigma}^2}\}$. Write $\widetilde{w} = \operatorname{argmin}_{w \in \mathcal{W}} \widehat{C}_{\hat{\sigma}}(w)$, which is the weight vector we choose.

Let $\widetilde{\mu}_{(s)} = \hat{P}_{(s)}|_{\sigma^2 = \hat{\sigma}^2} y$, $\widetilde{\mu}(w) = \sum_{s=1}^{S} w_s \widetilde{\mu}_{(s)}$, and $\widetilde{L}(w) = \| \widetilde{\mu}(w) - \mu \|^2$.

The following theorem shows the asymptotic optimality of $\widetilde{\mu}(\widetilde{w})$.

THEOREM 2. *Under Assumption* A2 *in the Appendix,* $\widetilde{L}(\widetilde{w})/\{\inf_{w \in \mathcal{W}} \widetilde{L}(w)\} \to 1$ *in probability.*

3. NUMERICAL EXAMPLES

We investigate finite-sample properties of our estimator first by conducting simulation studies to compare it with estimators from traditional subset selection methods based on conditional

AIC (Vaida & Blanchard, 2005), AIC, and BIC; model average estimators based on smoothed AIC and smoothed BIC weights (Buckland et al., 1997); jackknife model averaging, as proposed by Hansen & Racine (2012); and the maximum likelihood estimator under the model including all variables and using the most complex structure of D. The smoothed AIC uses weights proportional to the exponents of the AICs of the candidate models, that is,

$$w_{\mathrm{AIC,S}} = \exp(-C_{\mathrm{AIC,S}}/2) / \sum_{s=1}^{S} \exp(-C_{\mathrm{AIC,S}}/2),$$

where $C_{\mathrm{AIC,S}}$ is the AIC score of the sth candidate model. The weights of the smoothed BIC method are defined analogously. Second, we compare the proposed model averaging method with penalized regression methods: the mixed-model adaptive lasso of Bondell et al. (2010) and a moment-based method with a sandwich-type soft-thresholding penalty (Ahn et al., 2012). Some other works such as Ibrahim et al. (2011) also focused on the variable selection of linear mixed-effects model by penalization method. Third, we use our method to analyse a real dataset. Maximum likelihood estimators are used for each candidate model.

Example 1. We generated data from the model (1) where $\beta = (1, 0{\cdot}2)^{\mathrm{T}}$, the jth row of X_i is $(1, x_{i,j_2})$, the jth row of Z_i is $(1, z_{i,j_2}, z_{i,j_3}, z_{i,j_4})$, x_{i,j_2} and z_{i,j_k} for $k = 2, 3, 4$ independently generated from $N(0, 1)$, $n = 20$, $m_i = 10$, and $\sigma \in \{0{\cdot}3, 0{\cdot}8\}$. Four specifications for D were considered: $D_1 = 1{\cdot}1 I_4$, $D_2 = \mathrm{diag}(1{\cdot}4, 1{\cdot}2, 0{\cdot}4, 1)$,

$$D_3 = \begin{pmatrix} 1{\cdot}4 & 0{\cdot}4 & 0 & 0{\cdot}8 \\ 0{\cdot}4 & 1{\cdot}2 & 0 & 0 \\ 0 & 0 & 0{\cdot}4 & 0{\cdot}2 \\ 0{\cdot}8 & 0 & 0{\cdot}2 & 1 \end{pmatrix}, \qquad D_4 = \begin{pmatrix} 1{\cdot}4 & 0{\cdot}4 & 0{\cdot}6 & 0{\cdot}8 \\ 0{\cdot}4 & 1{\cdot}2 & 0{\cdot}2 & 0{\cdot}6 \\ 0{\cdot}6 & 0{\cdot}2 & 0{\cdot}4 & 0{\cdot}2 \\ 0{\cdot}8 & 0{\cdot}6 & 0{\cdot}2 & 1 \end{pmatrix}.$$

Thus, we had eight configurations, for each of which 100 independent sets of data were generated. All approximating models include the four random effects and fixed intercept, but possibly contain $x_{i,j2}$. The candidate covariance structure was chosen from a multiple of an identity matrix, a diagonal matrix, and a general positive definite matrix. Therefore, we combined $S = 6$ candidate models in all.

Evaluation of estimator performance was based on averaged squared loss: $100^{-1} \sum_{a=1}^{100} \|\hat{\mu}^{(a)} - \mu^{(a)}\|^2$, where $\mu^{(a)}$ is the value of μ in the ath replication and $\hat{\mu}^{(a)}$ is the estimate of $\mu^{(a)}$ obtained by a model selection or averaging method. In each replication, we subtracted the theoretically optimal squared loss, $\inf_{w \in \mathcal{W}} \| \sum_{s=1}^{S} w_s \tilde{\mu}_{(s)}^{(a)} - \mu^{(a)} \|^2$, from squared losses of the model selection and averaging methods, where $\tilde{\mu}_{(s)}^{(a)}$ is the estimate of $\mu^{(a)}$ under model s.

Table 1 presents the squared losses and the largest standard errors of losses in each configuration. Our method performs better than the jackknife model averaging for five of the eight configurations. The implementation of our method is more convenient than that of jackknife model averaging, because the latter requires $N - 1$ more estimations under each model than our method. Compared with other competitors, the proposed method performs best in most configurations. When $D \neq D_1$, our method performs best except for the configuration $(D, \sigma) = (D_2, 0{\cdot}3)$ where our method produces a slightly bigger loss than conditional AIC but still a smaller loss than other methods; even for the configurations with $D = D_4$ where the structure of D is most complex, our method produces smaller loss than maximum likelihood. When $D = D_1$, i.e., the structure of D is simplest, BIC and smoothed BIC lead to smaller losses than our method for small σ, while

Table 1. *Simulation results for Example* 1*: averaged squared losses and the largest standard errors of the losses in each row*

D	σ	AOMA	CAIC	AIC	BIC	SAIC	SBIC	JMA	ML	Largest s.e.
D_1	0·3	0·093	0·096	0·099	0·088	0·095	0·088	0·111	0·126	0·014
	0·8	1·384	1·361	1·201	1·978	1·217	1·525	1·188	2·256	0·204
D_2	0·3	0·114	0·113	0·126	0·126	0·119	0·123	0·156	0·148	0·015
	0·8	1·912	2·331	2·234	2·985	2·053	2·432	1·781	2·739	0·252
D_3	0·3	0·150	0·162	0·170	0·183	0·168	0·178	0·182	0·170	0·020
	0·8	1·637	2·580	2·922	4·107	2·935	3·325	1·800	2·571	0·220
D_4	0·3	0·170	0·218	0·199	0·228	0·199	0·214	0·171	0·199	0·030
	0·8	2·029	3·732	2·492	4·931	2·612	3·854	1·900	2·338	0·313

AOMA, asymptotically optimal model averaging; CAIC, conditional AIC; SAIC, smoothed AIC; SBIC, smoothed BIC; JMA, jackknife model averaging; ML, maximum likelihood; s.e., standard error.

Table 2. *Simulation results for Example* 2*: averaged squared losses and the largest standard errors of the losses in each row*

σ	AOMA	MALASSO	MM	ML	Largest s.e.
0·5	1·466	1·910	1·607	2·891	0·110
1	4·933	5·303	5·366	9·351	0·768
2	17·513	17·880	17·745	30·249	1·750

AOMA, asymptotically optimal model averaging; MALASSO, mixed-model adaptive lasso; MM, moment-based method; ML, maximum likelihood; s.e., standard error.

conditional AIC, AIC and smoothed AIC lead to smaller losses than our method for relatively big σ. As for comparison of model selection, conditional AIC can be better or worse than AIC and BIC.

Example 2. In this example, we compared our method with penalized regression methods. The data-generating process is the same as in Example 1 of Bondell et al. (2010). Specifically, the jth row of X_i is $(x_{i,j1}, \ldots, x_{i,j_9})$, and the jth row of Z_i is $(1, z_{i,j_2}, z_{i,j_3}, z_{i,j_4})$, $x_{i,jl}$ for $l = 1, \ldots, 9$ and z_{i,j_k} for $k = 2, 3, 4$ were independently generated from uniform $(-2, 2)$, $\beta = (1, 1, 0, \ldots, 0)'$, $n = 30$, $m_i = 5$, $\sigma = 1$, and

$$D = \begin{pmatrix} 9 & 4{\cdot}8 & 0{\cdot}6 & 0 \\ 4{\cdot}8 & 4 & 1 & 0 \\ 0{\cdot}6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which means that the last random effect is unimportant. We also set $\sigma = 0{\cdot}5$ and 2. We used the mixed-model adaptive lasso method of Bondell et al. (2010) to select variables, in which the tuning parameter was selected from $0{\cdot}05|\hat{\phi}_{\text{full}}|$ to $|\hat{\phi}_{\text{full}}|$, where $|\hat{\phi}_{\text{full}}|$ is the sum of all absolute estimates under the full model, by BIC. Our model averaging method was implemented based on the selected models by the mixed-model adaptive lasso using different candidate tuning parameters. The moment-based method of Ahn et al. (2012) was also performed with the tuning parameter selected by BIC. Other settings are the same as those of Example 1.

Table 2 presents averaged squared losses. For all configurations, our method performs best and maximum likelihood under the full model performs worst.

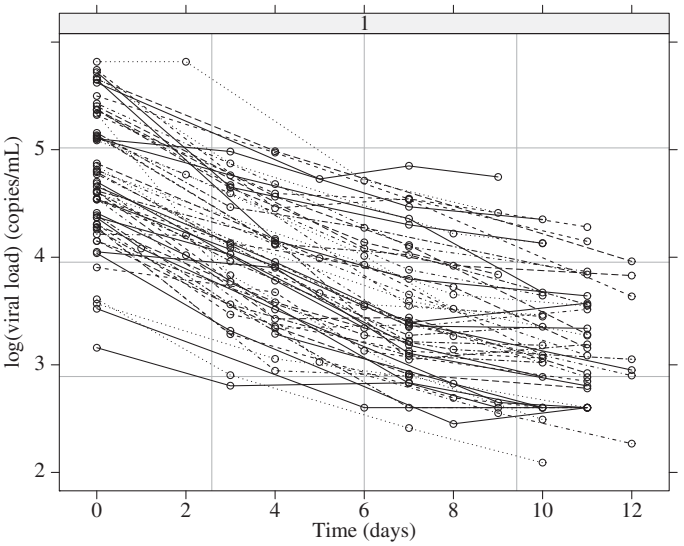

Fig. 1. Example 3. Scatter-plot of viral load (log-scale) against time for 60 patients from the AIDS clinical trial group study.

Example 3. Fitting viral load dynamics and understanding the pathogenesis of HIV infection play an important role in assessment of the treatment of antiviral therapy for AIDS/HIV patients. The decay rate of the viral load at the beginning stage is a useful marker (Ho et al., 1995; Wei et al., 1995). The linear mixed-effects model has become a helpful tool for estimating the first decay rate (Wu & Ding, 1999). Below we present an analysis of a subset of an AIDS clinical trial group study, in which viral load was scheduled to be measured on days $1, \ldots, 12$, and weeks 8, 12, 24 and 48 after initiation of an antiviral therapy. We analysed the data for the first two weeks, since the second decay appeared in week 2. The dataset comprises 60 patients, with the number of observations per patient varying from 2 to 5. We present the scatter-plot of these observations in Fig. 1. See McMahon et al. (2001) for the details of this study.

We adopted fixed intercept x_1, measure days x_2, natural logarithm of CD4 cell count x_3, and natural logarithm of CD8 cell count x_4 as covariates for fixed effects and $z_k = x_k$ for $k = 1, \ldots, 4$ as covariates for random effects. Sixteen linear mixed-effects models were combined; each includes the first k fixed effects and the first l random effects, $k, l \in \{1, \ldots, 4\}$, (k, l) indicates a candidate model. The structure of D in each candidate model was set to be positive definite. Take the last observation of each patient as the testing sample. The estimation was based on the remaining observations.

We used AIC, BIC or conditional AIC to select a final model. Such a selection strategy caused us difficulty because both AIC and BIC support model (2, 1), while conditional AIC supports model (3, 2). We also performed the mixed-model adaptive lasso and moment-based method as in Example 2 with the tuning parameter selected by BIC. The former supports the model with all fixed effects and the random effects z_1, z_3, and z_4, and the latter also chooses these random effects, but only x_1 and x_2 as covariates for fixed effects. Instead of arguing which model is favourable, we applied our procedure because we were primarily interested in prediction accuracy.

Table 3 presents the values of conditional AIC, AIC and BIC for some candidate models, and weights of three model averaging methods. We list only the models whose biggest weights for three model averaging methods are not smaller than 0·001. In each row showing criterion values, the smallest value is indicated by an asterisk. Smoothed AIC, smoothed BIC, jackknife model

Table 3. *The values of various criteria and weights of model averaging methods in the AIDS clinical trial group study. The asterisk indicates the selected model*

Models	CAIC	AIC	BIC	Weights of SAIC	Weights of SBIC	Weights of JMA	Weights of AOMA
(1,2)	78·6	318·3	333·9	0·000	0·000	0·006	0·000
(1,3)	59·4	316·1	341·1	0·000	0·000	0·000	0·303
(2,1)	40·1	161·5*	173·9*	0·431	0·866	0·994	0·697
(2,2)	38·0	165·2	183·9	0·066	0·006	0·000	0·000
(3,1)	42·0	162·4	178·0	0·266	0·113	0·000	0·000
(3,2)	36·7*	165·9	187·8	0·046	0·001	0·000	0·000
(4,1)	44·1	163·4	182·2	0·161	0·014	0·000	0·000
(4,2)	37·3	166·8	191·8	0·030	0·000	0·000	0·000

The abbreviations are the same as in Table 1.

Table 4. *The mean squared prediction errors for the AIDS clinical trial group study*

	AOMA	MALASSO	MM	CAIC	AIC	BIC	SAIC	SBIC	JMA	ML
MSPE	0·187	0·224	0·222	0·209	0·222	0·222	0·221	0·220	0·222	0·214
s.e.	0·022	0·026	0·026	0·024	0·026	0·026	0·025	0·026	0·026	0·027

The abbreviations are the same as in Tables 1 and 2. MSPE, mean squared prediction error.

averaging and our model averaging methods all put the largest weights on the model (2, 1). Table 4 presents the mean squared prediction errors under testing samples. The prediction error using our method, 0·187, is the smallest among all prediction errors based on the model selection and averaging methods.

ACKNOWLEDGEMENT

The authors are grateful to Drs. Mihye Ahn, Thomas Kneib and Hongtu Zhu for providing computational codes, and the editor, the associate editor and two referees for their constructive comments. This work was supported by the National Natural Science Foundation of China and the National Science Foundation, U.S.A. The work of Zhang was performed during his visit to Dr. Raymond J. Carroll at Texas A&M University, whose support is gratefully appreciated.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes a simple example for illustrating Condition (A1a), and a verification of Condition (A2b) from Conditions (A1b), (A1c) and (A3).

APPENDIX

Assumptions

For any $s \in \{1, \ldots, S\}$, suppose that there exists a vector $\theta^*_{(s)}$ such that $\hat{\theta}_{(s)} \to \theta^*_{(s)} = (\theta^*_{(s),1}, \ldots, \theta^*_{(s),J_{(s)}})'$ in probability as $N \to \infty$. Let $P^*_{(s)} = \hat{P}_{(s)} \mid_{\hat{\theta}_{(s)}=\theta^*_{(s)}}$, $P^*(w) = \hat{P}(w) \mid_{\hat{\theta}_{(s)}=\theta^*_{(s)}}$, $A^*(w) = \hat{A}(w) \mid_{\hat{\theta}_{(s)}=\theta^*_{(s)}}$, and $L^*(w) = \|\hat{\mu}(w) - \mu\|^2 \mid_{\hat{\theta}_{(s)}=\theta^*_{(s)}}$. Then we have $R^*(w) = E_{y|b}L^*(w) = \|A^*(w)\mu\|^2 + \sigma^2 \mathrm{tr}\{P^{*2}(w)\}$. Furthermore, define $\xi_N = \inf_{w\in\mathcal{W}} R^*(w)$, $K_{(s),j} = \partial G_{(s)}(\theta_{(s)})/\partial\theta_{(s),j}$, and $\hat{K}_{(s),j} = K_{(s),j} \mid_{\theta_{(s)}=\hat{\theta}_{(s)}}$, and let $\lambda_{\max}(\Phi)$ and $\lambda_{\min}(\Phi)$ denote the biggest and smallest singular values of a matrix Φ. Let w^0_s be a weight

vector with component unity for model s and zero for other models, $\max_j$, $\min_j$ indicate maximization, minimization over $j \in \{1, \ldots, J_{(s)}\}$, and $\sup_s$ indicates a supremum over $s \in \{1, \ldots, S\}$. Assume that $J_{(s)}$ are bounded uniformly for $s \in \{1, \ldots, S\}$. For $j \in \{1, \ldots, J_{(s)}\}$, we can write $\partial \hat{\theta}_{(s),j}/\partial y = \hat{T}_{(s),j} y$ almost surely, where $\hat{T}_{(s),j}$ can be random and depend on y.

Assumption A1. As $N \to \infty$, there exists an integer $1 \leqslant \kappa < \infty$ such that

$$S\xi_N^{-2\kappa} \sum_{s=1}^{S} \{R^*(w_s^0)\}^\kappa \to 0 \qquad \text{almost surely,} \tag{A1}$$

$$\sup_s \left\{ \max_j \lambda_{\max}\left(\hat{K}_{(s),j}\right) \right\} = O_p(1), \tag{A2}$$

$$\sup_s \left\{ \lambda_{\min}^{-1}(\hat{G}_{(s)}) \right\} = O_p(N^{-1/2}) \quad \text{or} \quad \sup_s \left\{ \lambda_{\max}(Z_{(s)} Z'_{(s)}) \right\} = O(1), \tag{A3}$$

and

$$N\xi_N^{-1} \sup_s \left\{ \max_j \lambda_{\max}(\hat{T}_{(s),j}) \right\} = o_p(1), \tag{A4}$$

$$N\xi_N^{-1} \sup_s \left\{ \lambda_{\max}(\hat{P}_{(s)} - P^*_{(s)}) \right\} = o_p(1), \tag{A5}$$

$$N^{-1}\|\mu\|^2 = O_p(1). \tag{A6}$$

These conditions are standard technical conditions in the model averaging literature. Condition (A1) is a straightforward extension of Condition (8) of Wan et al. (2010), which was used to prove asymptotic optimality of model averaging under linear regression models. A similar condition has also been imposed to study model selection for the linear mixed-effects models, like Condition (17) in Pu & Niu (2006). A simple example in § S.1 of the Supplementary Material sheds further light on Condition (A1).

Condition (A2) requires the largest singular value of $\hat{K}_{(s),j}$ to be bounded in probability as the sample size increases. When $\theta_{(s)}$ consists of the elements of $D_{(s)}$, each element of $\partial D_{(s)}/\partial \theta_{(s),j}$ is either 0 or 1, and in this case, Condition (A2) holds automatically.

The first part of Condition (A3) requires the inverse of minimum singular value of $\hat{G}_{(s)}$ to be bounded in probability as the sample size increases. The second part of Condition (A3) holds, for example, in a situation with m_i being bounded and $\lambda_{\max}(Z_{(s),i} Z'_{(s),i}) = O(m_i)$ uniformly for $i \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, S\}$, where $Z_{(s),i}$ is the ith diagonal component of $Z_{(s)}$.

Condition (A4) requires that $\lambda_{\max}(\hat{T}_{(s),j})$ converge to 0 at some rate with the sample size increasing. Considering a simple case with only random intercept, in each iterative step of maximum likelihood estimation, $\hat{\theta}_{(s),j}$ has a form of $y'\Psi_{(s),j} y/N$ with a finite $\lambda_{\max}(\Psi_{(s),j})$ (Hsiao, 2003), then $\lambda_{\max}(\hat{T}_{(s),j}) = O_p(1/N)$.

Condition (A5) requires that $\hat{\theta}_{(s)}$ converges to $\theta^*_{(s)}$ at a rate such that $\xi_N^{-1} \sup_s \lambda_{\max}(\hat{P}_{(s)} - P^*_{(s)})$ converges to 0 at a rate quicker than $N \to \infty$. For $s \in \{1, \ldots, S\}$ and $j_1, j_2 \in \{1, \ldots, J_{(s)}\}$, we define $\Upsilon_{(s)}$ as an $N \times N$ matrix with (i_1, i_2)th element $\Upsilon_{(s),i_1,i_2} = \sum_{j_1=1}^{J_{(s)}} \sum_{j_2=1}^{J_{(s)}} (\hat{\theta}_{(s),j_1} - \theta^*_{(s),j_1})(\hat{\theta}_{(s),j_2} - \theta^*_{(s),j_2}) \partial^2 P_{(s),i_1,i_2}/(\partial\theta_{(s),j_1}\partial\theta_{(s),j_2})\,|_{\theta_{(s)}=\tilde{\theta}^{i_1,i_2}_{(s)}}$, where $P_{(s),i_1,i_2}$ is the (i_1, i_2)th element of $P_{(s)}$ and $\tilde{\theta}^{i_1,i_2}_{(s)}$ is a $J_{(s)} \times 1$ vector between $\hat{\theta}_{(s)}$ and $\theta^*_{(s)}$. When Conditions (A2) and (A3) hold, Condition (A5) is implied by

$$\lambda_{\max}\left(\Upsilon_{(s)}\right) = o_p(1), \quad N^{1/2}(\hat{\theta}_{(s),j} - \theta^*_{(s),j}) = O_p(1), \quad N\xi_N^{-2} \to 0, \tag{A7}$$

uniformly for $s \in \{1, \ldots, S\}$ and $j \in \{1, \ldots, J_{(s)}\}$. See § S.2 of the Supplementary Material. The second part of (A7) is a common convergence rate and the third part of (A7) is implied by Condition (A1) when there exists an $s \in \{1, \ldots, S\}$ such that $NR^*(w_s^0)^{-1} = O(1)$. Condition (A6), concerning the sum of μ_j, is a commonly-used condition in linear regression models (Liang et al., 2011).

When σ^2 is unknown, the estimator of $\theta_{(s)}$ is different from that for known σ^2, but in the following assumptions, we still use the notation $\hat{\theta}_{(s)}$ as the estimator and the notation $\theta^*_{(s)}$ as its limit for simplicity.

Assumption A2. As $N \to \infty$, Conditions (A1)–(A4) and (A6) hold, Condition (A5) holds with $\hat{P}_{(s)}$ being replaced by $\hat{P}_{(s)}\,|_{\sigma^2=\hat{\sigma}^2}$, $\hat{\sigma}^2 \to \sigma^2$ in probability, and

$$N\xi_N^{-1}(\hat{\sigma}^2 - \sigma^2) = o_p(1). \tag{A8}$$

When $N^{1/2}(\hat{\sigma}^2 - \sigma^2) = O_p(1)$, Condition (A8) is implied by the third part of (A7).

Proof of Theorem 1

First, it is straightforward to verify that

$$\lambda_{\max}(\hat{\Sigma}_{(s)}^{-1}) = \lambda_{\max}\{(\sigma^2 I_N + Z_{(s)}\hat{G}_{(s)}Z'_{(s)})^{-1}\} \leqslant \sigma^{-2},\ s = 1, \ldots, S, \tag{A9}$$

and that for any two $N \times N$ matrices Φ_1 and Φ_2 (Li, 1987),

$$\lambda_{\max}(\Phi_1\Phi_2) \leqslant \lambda_{\max}(\Phi_1)\lambda_{\max}(\Phi_2), \qquad \lambda_{\max}(\Phi_1 + \Phi_2) \leqslant \lambda_{\max}(\Phi_1) + \lambda_{\max}(\Phi_2). \tag{A10}$$

Because $\hat{V}_{(s)}$ is symmetric and idempotent, it follows from (A9) and (A10) that

$$\lambda_{\max}(\hat{P}_{(s)}) = \lambda_{\max}\{I_N - \sigma^2\hat{\Sigma}_{(s)}^{-1/2}(I_N - \hat{V}_{(s)})\hat{\Sigma}_{(s)}^{-1/2}\} \leqslant 2,\ s = 1, \ldots, S, \tag{A11}$$

and thus for any weight vector $w \in \mathcal{W}$, we obtain that

$$\lambda_{\max}\{\hat{P}(w)\} \leqslant 2. \tag{A12}$$

By the fact that $\varepsilon_{i,j} \sim N(0, \sigma^2)$, we see that for any positive integer κ, there exists a constant c such that

$$E(\varepsilon_{i,j}^{4\kappa}) \leqslant c < \infty. \tag{A13}$$

Let $h_s = \sum_{j=1}^{J_{(s)}} \partial\hat{\theta}_{(s),j}/\partial y'\,\hat{A}_{(s)}\hat{W}_{(s),j}\hat{A}_{(s)}y$ and $h(w) = \sum_{s=1}^{S} w_s h_s$. From (7) and the definitions of $L(w)$, $\hat{C}(w)$ and $R^*(w)$, we have

$$\begin{aligned} L(w) - \hat{C}(w) &= \|\varepsilon\|^2 + \|\hat{A}(w)y\|^2 - 2\varepsilon'\hat{A}(w)y - \|\hat{A}(w)y\|^2 - 2\sigma^2\mathrm{tr}\{\hat{P}(w)\} - 2\sigma^2 h(w) \\ &= \|\varepsilon\|^2 - 2\varepsilon' A^*(w)\mu - 2\varepsilon'\{P^*(w) - \hat{P}(w)\}\mu - 2\varepsilon'\{P^*(w) - \hat{P}(w)\}\varepsilon \\ &\quad + 2\sigma^2[\mathrm{tr}\{P^*(w)\} - \mathrm{tr}\{\hat{P}(w)\}] + 2[\varepsilon' P^*(w)\varepsilon - \sigma^2\mathrm{tr}\{P^*(w)\}] - 2\sigma^2 h(w) \end{aligned}$$

and

$$\begin{aligned} L(w) - R^*(w) &= \|\hat{P}(w)\varepsilon - \hat{A}(w)\mu\|^2 - \|A^*(w)\mu\|^2 - \sigma^2\mathrm{tr}\{P^{*2}(w)\} \\ &= \mu'\{P^*(w) - \hat{P}(w)\}\{A^*(w) + \hat{A}(w)\}\mu - 2\mu' A^*(w)P^*(w)\varepsilon \\ &\quad + 2\mu'\{P^*(w) - \hat{P}(w)\}\varepsilon - 2\mu' P^*(w)\{P^*(w) - \hat{P}(w)\}\varepsilon \\ &\quad - 2\mu'\{P^*(w) - \hat{P}(w)\}\hat{P}(w)\varepsilon - \varepsilon'\{P^*(w) + \hat{P}(w)\}\{P^*(w) - \hat{P}(w)\}\varepsilon \\ &\quad + \varepsilon' P^{*2}(w)\varepsilon - \sigma^2\mathrm{tr}\{P^{*2}(w)\}. \end{aligned}$$

So, similar to the proof of Theorem 2.1 of Li (1987), in order to prove Theorem 1, we need only to verify that

$$\sup_{w\in\mathcal{W}}\left\{\left|\varepsilon' A^*(w)\mu\right|/R^*(w)\right\}=o_p(1), \tag{A14}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\varepsilon' P^*(w)\varepsilon-\sigma^2\mathrm{tr}\{P^*(w)\}\right|/R^*(w)\right]=o_p(1), \tag{A15}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\mu' A^*(w)P^*(w)\varepsilon\right|/R^*(w)\right]=o_p(1), \tag{A16}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\varepsilon' P^{*2}(w)\varepsilon-\sigma^2\mathrm{tr}\{P^{*2}(w)\}\right|/R^*(w)\right]=o_p(1), \tag{A17}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\varepsilon'\{P^*(w)+\hat{P}(w)\}\{P^*(w)-\hat{P}(w)\}\varepsilon\right|/R^*(w)\right]=o_p(1), \tag{A18}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\varepsilon'\{P^*(w)-\hat{P}(w)\}\varepsilon\right|/R^*(w)\right]=o_p(1), \tag{A19}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\mu'\{P^*(w)-\hat{P}(w)\}\varepsilon\right|/R^*(w)\right]=o_p(1), \tag{A20}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\mu' P^*(w)\{P^*(w)-\hat{P}(w)\}\varepsilon\right|/R^*(w)\right]=o_p(1), \tag{A21}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\mu'\{P^*(w)-\hat{P}(w)\}\hat{P}(w)\varepsilon\right|/R^*(w)\right]=o_p(1), \tag{A22}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\mu'\{P^*(w)-\hat{P}(w)\}\{A^*(w)+\hat{A}(w)\}\mu\right|/R^*(w)\right]=o_p(1), \tag{A23}$$

$$\sup_{w\in\mathcal{W}}\left[\left|\mathrm{tr}\{P^*(w)\}-\mathrm{tr}\{\hat{P}(w)\}\right|/R^*(w)\right]=o_p(1), \tag{A24}$$

and

$$\sup_{w\in\mathcal{W}}\{|h(w)|/R^*(w)\}=o_p(1). \tag{A25}$$

From (A11), (A13) and Condition (A1), the equations (A14)–(A17) can be shown by using the same steps as in the proof of Theorem 1′ of Wan et al. (2010).

For proving (A18), by (A10) and (A13), it is seen that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\left[\left|\varepsilon'\{P^*(w)+\hat{P}(w)\}\{P^*(w)-\hat{P}(w)\}\varepsilon\right|/R^*(w)\right]\\
&\quad\leqslant\xi_N^{-1}2^{-1}\sup_{w\in\mathcal{W}}\left|\varepsilon'[\{P^*(w)+\hat{P}(w)\}\{P^*(w)-\hat{P}(w)\}\right.\\
&\qquad\left.+\{P^*(w)-\hat{P}(w)\}\{P^*(w)+\hat{P}(w)\}]\varepsilon\right|\\
&\quad\leqslant\xi_N^{-1}2^{-1}\|\varepsilon\|^2\sup_{w\in\mathcal{W}}\lambda_{\max}[\{P^*(w)+\hat{P}(w)\}\{P^*(w)-\hat{P}(w)\}\\
&\qquad+\{P^*(w)-\hat{P}(w)\}\{P^*(w)+\hat{P}(w)\}]\\
&\quad\leqslant\xi_N^{-1}\|\varepsilon\|^2\sup_{w\in\mathcal{W}}[\lambda_{\max}\{P^*(w)+\hat{P}(w)\}\lambda_{\max}\{P^*(w)-\hat{P}(w)\}]\\
&\quad\leqslant\xi_N^{-1}\|\varepsilon\|^2\sup_{w\in\mathcal{W}}([\lambda_{\max}\{P^*(w)\}+\lambda_{\max}\{\hat{P}(w)\}]\sum\nolimits_{s=1}^{S}w_s\lambda_{\max}(P^*_{(s)}-\hat{P}_{(s)}))\\
&\quad\leqslant\xi_N^{-1}4\|\varepsilon\|^2\sup_{w\in\mathcal{W}}\left\{\sum\nolimits_{s=1}^{S}w_s\lambda_{\max}(P^*_{(s)}-\hat{P}_{(s)})\right\}
\end{aligned}$$

$$\leqslant 4\frac{\|\varepsilon\|^2}{N} N\xi_N^{-1} \sup_s\{\lambda_{\max}(\hat{P}_{(s)} - P^*_{(s)})\}$$
$$= o_p(1),$$

where the fifth inequality is from (A12) and the last step is from Condition (A5). Similarly, we can prove (A19). For (A20),

$$\sup_{w\in\mathcal{W}}\left[\left|\mu'\{P^*(w) - \hat{P}(w)\}\varepsilon\right| / R^*(w)\right] \leqslant \xi_N^{-1} \sup_{w\in\mathcal{W}}[\|\mu\|^2\varepsilon'\{P^*(w) - \hat{P}(w)\}^2\varepsilon]^{1/2}$$
$$\leqslant \|\mu\|/N^{1/2}(\|\varepsilon\|^2/N)^{1/2}\xi_N^{-1} N \sup_s\{\lambda_{\max}(\hat{P}_{(s)} - P^*_{(s)})\}$$
$$= o_p(1), \tag{A26}$$

where the last step is from Conditions (A5) and (A6). From (A10) and (A12), we have $\lambda_{\max}\{P^*(w) + \hat{P}(w)\} \leqslant 4$ and $\lambda_{\max}\{A^*(w) + \hat{A}(w)\} \leqslant 6$, equations (A21)–(A23) can be proved by similar steps used in (A26).

For (A24),

$$\sup_{w\in\mathcal{W}}\left|\mathrm{tr}\{P^*(w)\} - \mathrm{tr}\{\hat{P}(w)\}\right| / R^*(w) \leqslant \xi_N^{-1} N \sup_s\{\lambda_{\max}(P^*_{(s)} - \hat{P}_{(s)})\} = o_p(1),$$

where the equality is from Condition (A5).

Let $U_{(s),j} = \hat{\Sigma}_{(s)}^{-1/2} Z_{(s)} \hat{K}_{(s),j} Z'_{(s)} \hat{\Sigma}_{(s)}^{-1/2}$. Using (A9)–(A10), we obtain that

$$\sup_s\{\max_j \lambda_{\max}(U_{(s),j})\} \leqslant \sup_s\{\max_j \lambda_{\max}(\hat{K}_{(s),j})\lambda_{\max}(\hat{\Sigma}_{(s)}^{-1/2} Z_{(s)} Z'_{(s)} \hat{\Sigma}_{(s)}^{-1/2})\}$$
$$\leqslant \sigma^{-2} \sup_s\{\max_j \lambda_{\max}(\hat{K}_{(s),j})\} \sup_s\{\lambda_{\max}(Z_{(s)} Z'_{(s)})\}, \tag{A27}$$

and when $\lambda_{\min}^{-1}(\hat{G}_{(s)})$ is bounded in probability,

$$\sup_s\{\max_j \lambda_{\max}(U_{(s),j})\} \leqslant \sup_s\{\max_j \lambda_{\max}(\hat{K}_{(s),j})\lambda_{\min}^{-1}(\hat{G}_{(s)})\lambda_{\max}(\hat{\Sigma}_{(s)}^{-1/2} Z_{(s)} \hat{G}_{(s)} Z'_{(s)} \hat{\Sigma}_{(s)}^{-1/2})\}$$
$$= \sup_s[\max_j \lambda_{\max}(\hat{K}_{(s),j})\lambda_{\min}^{-1}(\hat{G}_{(s)})\lambda_{\max}\{(\sigma^2 I_N + Z_{(s)} \hat{G}_{(s)} Z'_{(s)})^{-1/2}$$
$$\times\ Z_{(s)} \hat{G}_{(s)} Z'_{(s)} (\sigma^2 I_N + Z_{(s)} \hat{G}_{(s)} Z'_{(s)})^{-1/2}\}]$$
$$\leqslant \sup_s\{\max_j \lambda_{\max}(\hat{K}_{(s),j})\} \sup_s\{\lambda_{\min}^{-1}(\hat{G}_{(s)})\}. \tag{A28}$$

It follows from (A27)–(A28) and Conditions (A2)–(A3) that

$$\sup_s\{\max_j \lambda_{\max}(U_{(s),j})\} = O_p(1). \tag{A29}$$

Now, using (A9), (A10) and (A29), we have

$$\sup_s\{\max_j \lambda_{\max}(\hat{A}_{(s)} \hat{W}_{(s),j} \hat{A}_{(s)})\} = \sigma^4 \sup_s[\max_j \lambda_{\max}\{\hat{\Sigma}_{(s)}^{-1/2}(I_N - \hat{V}_{(s)})U_{(s),j}(I_N - \hat{V}_{(s)})\hat{\Sigma}_{(s)}^{-1/2}\}]$$
$$= O_p(1). \tag{A30}$$

By (A13), (A30) and Conditions (A4) and (A6), it is observed that

$$\begin{aligned}\sup_{w\in\mathcal{W}}\{|h(w)|/R^*(w)\} &\leqslant \xi_N^{-1}\sup_s J_{(s)}\sup_s(\max_j|y'T_{(s),j}\hat{A}_{(s)}\hat{W}_{(s),j}\hat{A}_{(s)}y|)\\ &\leqslant \xi_N^{-1}\sup_s J_{(s)}\|y\|^2\sup_s\{\max_j\lambda_{\max}(T_{(s),j}\hat{A}_{(s)}\hat{W}_{(s),j}\hat{A}_{(s)})\}\\ &\leqslant \sup_s J_{(s)}N^{-1}\|y\|^2N\xi_N^{-1}\sup_s\{\max_j\lambda_{\max}(T_{(s),j})\}\sup_s\{\max_j\lambda_{\max}(\hat{A}_{(s)}\hat{W}_{(s),j}\hat{A}_{(s)})\}\\ &= o_p(1),\end{aligned}$$

so (A25) holds. This completes the proof of Theorem 1.

Proof of Theorem 2

Based on the proof of Theorem 1, to prove Theorem 2, we need only to show that

$$\hat{\sigma}^2\sup_{w\in\mathcal{W}}\{|h(w)|_{\sigma^2=\hat{\sigma}^2}|/R^*(w)\}=o_p(1) \tag{A31}$$

and

$$|\sigma^2-\hat{\sigma}^2|\sup_{w\in\mathcal{W}}[\mathrm{tr}\{P^*(w)\}/R^*(w)]=o_p(1). \tag{A32}$$

When σ^2 is replaced by $\hat{\sigma}^2$, the formulae (A29) and (A30) still hold, and thus (A31) is true. On the other hand, from the argument of (A11), we have $\lambda_{\max}(P^*_{(s)})\leqslant 2$, so

$$\begin{aligned}|\sigma^2-\hat{\sigma}^2|\sup_{w\in\mathcal{W}}[\mathrm{tr}\{P^*(w)\}/R^*(w)] &\leqslant \xi_N^{-1}|\sigma^2-\hat{\sigma}^2|\sup_s\{\mathrm{tr}(P^*_{(s)})\}\\ &\leqslant \xi_N^{-1}|\sigma^2-\hat{\sigma}^2|\sup_s\{\mathrm{rank}(P^*_{(s)})\lambda_{\max}(P^*_{(s)})\}\\ &\leqslant 2N\xi_N^{-1}|\sigma^2-\hat{\sigma}^2|,\end{aligned}$$

which together with Condition (A8) implies (A32). This completes the proof of Theorem 2.

References

Ahn, M., Zhang, H. H. & Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statist. Sinica* **22**, 1539–62.

Bates, J. M. & Granger, C. W. J. (1969). The combination of forecasts. *Oper. Res. Quart.* **20**, 451–68.

Bondell, H. D., Krishna, A. & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069–77.

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603–18.

Claeskens, G. & Carroll, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249–65.

Claeskens, G., Croux, C. & van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62**, 972–9.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Donohue, M. C., Overholser, R., Xu, R. & Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika* **98**, 685–700.

Greven, S. & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773–89.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica* **75**, 1175–89.

Hansen, B. E. & Racine, J. (2012). Jackknife model averaging. *J. Economet.* **167**, 38–46.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators. *J. Am. Statist. Assoc.* **98**, 879–99.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. & Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–6.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. & VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14**, 382–17.
HSIAO, C. (2003). *Analysis of Panel Data*, vol. 34 of *Econometric Society Monographs*. Cambridge: Cambridge University Press, 2nd ed.
IBRAHIM, J. G., ZHU, H., GARCIA, R. I. & GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67**, 495–503.
LAIRD, N. M. & WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.
LI, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958–75.
LIANG, H., WU, H. & ZOU, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95**, 773–8.
LIANG, H., ZOU, G., WAN, A. T. K. & ZHANG, X. (2011). Optimal weight choice for frequentist model average estimators. *J. Am. Statist. Assoc.* **106**, 1053–66.
MCMAHON, D., LEDERMAN, M., HAAS, D. W., HAUBRICH, R., STANFORD, J., COONEY, E., HORTON, J., KELLEHER, D., ROSS, L., CUTRELL, A., LEE, D., SPREEN, W. & MELLORS, J. W. (2001). Antiretroviral activity and safety of abacavir in combination with selected HIV-1 protease inhibitors in therapy-naive HIV-1-infected adults. *Antiviral Therap.* **6**, 105–14.
MILLER, A. J. (2002). *Subset Selection in Regression*, 2nd ed. London: Chapman and Hall.
PU, W. & NIU, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *J. Mult. Anal.* **97**, 733–58.
RAFTERY, A. E., MADIGAN, D. & HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Am. Statist. Assoc.* **92**, 179–91.
ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6**, 15–51.
VAIDA, F. & BLANCHARD, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–70.
WAN, A. T. K., ZHANG, X. & ZOU, G. (2010). Least squares model averaging by Mallows criterion. *J. Economet.* **156**, 277–83.
WEI, X., GHOSH, S. K., TAYLOR, M. E., JOHNSON, V. A., EMINI, E. A., DEUTSCH, P., LIFSON, 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**, 117–22.
WU, H. L. & DING, A. A. (1999). Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410–18.
YANG, Y. (2001). Adaptive regression by mixing. *J. Am. Statist. Assoc.* **96**, 574–88.
YANG, Y. (2003). Regression with multiple candidate models: Selecting or mixing? *Statist. Sinica* **13**, 783–809.
YUAN, Z. & YANG, Y. (2005). Combining linear regression models: When and how? *J. Am. Statist. Assoc.* **100**, 1202–14.
ZHANG, X. & LIANG, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.* **39**, 174–200.
ZHANG, X., WAN, A. T. K. & ZHOU, S. Z. (2012). Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *J. Bus. Econ. Statist.* **30**, 132–42.

[*Received November* 2012. *Revised September* 2013]