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 works 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 that the resulting model average estimator minimizes the squared error in large samples. 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

$$y = X_1 \beta + Z_1 b_1 + \epsilon_1, \quad b_1 \sim N(0, D_1), \quad \epsilon_1 \sim N(0, R_1) \quad (i = 1, \ldots, n),$$

where $\beta$ is a $p \times 1$ vector of fixed regression coefficients, $b_1$ is a $k \times 1$ vector of random coefficients specific to the cluster $i$, $\epsilon_1$ is an $m_i \times 1$ error vector independent of $b_1$, and $y_i, X_i, Z_i$ are the response vector and covariate matrices for the fixed and random effects related to the $i$th cluster, respectively. Both the covariates 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:

$$y = \mu + \varepsilon = X_\beta + Zb + \varepsilon, \quad b \sim N(0, G),$$

where $y = (y'_1, \ldots, y'_n)'$ is an $N \times 1$ vector of observations with mean $\mu$ conditional on $b$, $X = (X'_1, \ldots, X'_n)'$ is an $N \times p$ matrix, $Z = \text{diag}(Z_1, \ldots, Z_n)$ is an $N \times r$ block-diagonal matrix, $r = nk, b = (b'_1, \ldots, b'_n)'$, $\varepsilon = (\varepsilon'_1, \ldots, \varepsilon'_n)'$, and $G = \text{diag}(D, \ldots, D)$ is an $r \times r$ block-diagonal matrix.

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

$$y = X(s)\beta(s) + Z(s)b(s) + \varepsilon, \quad b(s) \sim N(0, G(s)) \quad (s = 1, \ldots, S),$$

are used to approximate (2), where $X(s)$ and $Z(s)$ are covariate matrices in the $s$th candidate model, $\beta(s)$ and $b(s)$ are the corresponding fixed and random coefficients, $G(s) = \text{diag}(D(s), \ldots, D(s))$ 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 $\sigma^2$ and $G(s)$, the fixed effects $\beta(s)$ and random effects $b(s)$ under the $s$th candidate model can be estimated by the best linear unbiased estimators (Robinson, 1991),

$$\hat{\beta}(s) = (X'_s \Sigma^{-1}_s X(s))^{-1}X'_s \Sigma^{-1}_s y, \quad \hat{b}(s) = G(s)Z'_s \Sigma^{-1}_s (y - X(s)\hat{\beta}(s)), \quad (4)$$

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 $\hat{G}(s) = \text{diag}(\hat{D}(s), \ldots, \hat{D}(s)), \hat{\nu}(s) = \hat{\Sigma}^{-1/2}_s X(s)X'_s \hat{\Sigma}^{-1/2}_s X(s)' \hat{\Sigma}^{-1/2}_s, \hat{\Sigma}_s = \sigma^2 I_N + Z(s)\hat{G}(s)Z'_s, \text{ and } \hat{P}(s) = I_N - \sigma^2 \hat{\Sigma}^{-1/2}_s (I_N - \hat{\nu}(s)) \hat{\Sigma}^{-1/2}_s$. From (4), when $G(s)$ is unknown,
\[ \beta_{(s)} \text{ and } b_{(s)} \text{ can be estimated by} \]
\[ \hat{\beta}_{(s)} = (X'_{(s)} \hat{\Sigma}_{(s)}^{-1} X_{(s)})^{-1} X'_{(s)} \hat{\Sigma}_{(s)}^{-1} y, \]
\[ \hat{b}_{(s)} = \hat{G}_{(s)} Z'_{(s)} \hat{\Sigma}_{(s)}^{-1} (y - X_{(s)} \hat{\beta}_{(s)}), \]
and thus \( \mu \) can be estimated by \( \hat{\mu}_{(s)} = X_{(s)} \hat{\beta}_{(s)} + Z_{(s)} \hat{b}_{(s)} \). Direct manipulation yields
\[ \hat{\mu}_{(s)} = \hat{P}_{(s)} y, \quad \hat{P}_{(s)} X_{(s)} = X_{(s)}. \]

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

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

**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 \( \text{AIC} \) and \( \text{BIC} \) 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 \( R(w) = E_y | \hat{\mu}(w) - \mu |^2 \) respectively. Let \( \hat{A}(w) = I_N - \hat{P}(w) \). Noting \( \varepsilon \sim N(0, \sigma^2 I_N) \), by integration by parts as in the derivation of Theorem 1 of Liang et al. (2008), we have
\[ R(w) = E_y | \hat{\mu}(w) - \mu |^2 = E_y | \hat{A}(w) y |^2 + 2 \sigma^2 w' \rho - n \sigma^2, \tag{5} \]
where \( \rho \) is an \( S \times 1 \) vector with \( s \)th element \( \rho_s = \text{tr} \{ \partial(\hat{P}_{(s)} y) / \partial y' \} \).

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 \( \hat{w} = \text{argmin}_{w \in W} \hat{C}(w) \). Then the resulting model average estimator is \( \hat{\mu}(\hat{w}) \). When the \( \Sigma_{(s)} \) is known, write \( P_{(s)} = \hat{P}_{(s)} | \Sigma_{(s)} = \Sigma_{(s)} \), \( P(w) = \sum_{s=1}^S w_s P_{(s)} \), and \( A(w) = I_N - P(w) \). In this case, \( \hat{C}(w) \) simplifies to \( \| A(w) y \|^2 + 2 \sigma^2 \text{tr} \{ P(w) \} \), 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
\[ \rho_s = \text{tr}(\hat{P}_{(s)}) + \sum_{j=1}^{J(s)} \frac{\partial \hat{\theta}_{(s),j}}{\partial y'} \hat{A}_{(s)} \hat{W}_{(s),j} \hat{A}_{(s),y}, \tag{7} \]
where \( \hat{\theta}_{(s),j} \) is the \( j \)th 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 \( \theta(s) \), say the last \( u \) elements, are zeros. Denote \( \hat{\theta}(s) = (\hat{\theta}(s)_1, \ldots, \hat{\theta}(s)_{J(s) - u})' \). Using Theorem 3 of Greven & Kneib (2010), we have

\[
\begin{align*}
\frac{\partial \hat{\theta}(s)_j}{\partial y'} &= 0, \quad j = J(s) - u + 1, \ldots, J(s), \quad \frac{\partial \hat{\theta}(s)_j}{\partial y'} = H_{-1} \tilde{H}(s),
\end{align*}
\]

where \( H(s) \) is a \((J(s) - u) \times (J(s) - u)\) matrix with \((i, j)\)th element

\[
- \text{tr}(\hat{\Sigma}_{(s)}^{-1} \hat{W}_{(s),i} \hat{\Sigma}_{(s)}^{-1} \hat{W}_{(s),j}) - Ny' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} y (y' \hat{A}_{(s)} y)^{-2} + 2Ny' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} y (y' \hat{A}_{(s)} y)^{-1},
\]

and \( \tilde{H}(s) \) is a \((J(s) - u) \times N\) matrix with \(i\)th row

\[
\tilde{H}(s)_{:,i} = 2N \left\{ -y' \hat{A}_{(s)} y^{-2} y' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} y + (y' \hat{A}_{(s)} y)^{-1} y' \hat{A}_{(s)} \hat{W}_{(s),i} \hat{A}_{(s)} \right\}.
\]

Expressions (7)–(10) give explicit formulae for calculating \( \rho_s \), so the minimization of \( \hat{C}(w) \) is easily managed.

The following theorem shows the asymptotic optimality of \( \hat{\mu}(\hat{\omega}) \). 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, \( L(\hat{\omega})/\inf_{w \in W} L(w) \to 1 \) in probability.

Theorem 1 indicates that the squared loss based on the weight vector \( \hat{\omega} \) 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 \( \sigma^2 \) in the criterion (6). One may develop an unbiased estimator of the squared risk by taking the estimation of \( \sigma^2 \) into account; see Liang et al. (2008). An alternative is that we insert an estimator \( \hat{\sigma}^2 \) of \( \sigma^2 \), 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,

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

where \( \hat{A}_{\hat{\sigma}}(w) = \hat{A}(w) \mid_{\sigma^2 = \hat{\sigma}^2} \) and \( \hat{\rho} \) is an \( S \times 1 \) vector with \( s \)th element \( \hat{\rho}_s = \text{tr}(\partial(\hat{P}(s)y)/\partial y'|_{\sigma^2 = \hat{\sigma}^2}) \). Write \( \tilde{w} = \arg\min_{w \in W} \hat{C}_{\hat{\sigma}}(w) \), which is the weight vector we choose.

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

The following theorem shows the asymptotic optimality of \( \tilde{\mu}(\hat{\omega}) \).

**Theorem 2.** Under Assumption A2 in the Appendix, \( L(\hat{\omega})/\inf_{w \in W} \tilde{L}(w) \to 1 \) 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
**Linear mixed-effects models**

\(\text{AIC} \) (Vaida & Blanchard, 2005), \(\text{AIC} \), and \(\text{BIC} \); model average estimators based on smoothed \(\text{AIC} \) and smoothed \(\text{BIC} \) 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 \(\text{AIC} \) uses weights proportional to the exponents of the \(\text{AICs} \) of the candidate models, that is,

\[
w_{\text{AIC},s} = \exp(-C_{\text{AIC},s}/2)/\sum_{s=1}^{S} \exp(-C_{\text{AIC},s}/2),
\]

where \(C_{\text{AIC},s}\) is the \(\text{AIC} \) score of the \(s\)th candidate model. The weights of the smoothed \(\text{BIC} \) 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 \(\beta = (1, 0.2)^T\), the \(j\)th row of \(X_i\) is \((1, x_i, j2, j)\), the \(j\)th row of \(Z_i\) is \((1, z_i, j2, j, z_i, j4, z_i, j6, z_i, j8)\), \(x_i, j2\) and \(z_i, jk\) for \(k = 2, 3, 4\) independently generated from \(N(0, 1)\), \(n = 20\), \(m_j = 10\), and \(\sigma \in \{0.3, 0.8\}\). Four specifications for \(D\) were considered: \(D_1 = 1 \cdot I_4\), \(D_2 = \text{diag}(1, 4, 1, 2, 0, 4, 1)\),

\[
D_3 = \begin{pmatrix}
1.4 & 0.4 & 0 & 0.8 \\
0.4 & 1.2 & 0 & 0 \\
0 & 0 & 0.4 & 0.2 \\
0.8 & 0 & 0.2 & 1
\end{pmatrix}, \quad D_4 = \begin{pmatrix}
1.4 & 0.4 & 0.6 & 0.8 \\
0.4 & 1.2 & 0.2 & 0.6 \\
0.6 & 0.2 & 0.4 & 0.2 \\
0.8 & 0.6 & 0.2 & 1
\end{pmatrix}.
\]

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 \(\mu\) in the \(a\)th 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, \(\inf_{w \in Y} \sum_{s=1}^{S} w_s \|\hat{\mu}^{(s)} - \mu^{(a)}\|^2\), from squared losses of the model selection and averaging methods, where \(\bar{\mu}^{(s)}\) 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 \((D, \sigma) = (D_2, 0.3)\) where our method produces a slightly bigger loss than conditional \(\text{AIC} \) 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, \(\text{BIC} \) and smoothed \(\text{BIC} \) lead to smaller losses than our method for small \(\sigma\), while...
Table 1. Simulation results for Example 1: averaged squared losses and the largest standard errors of the losses in each row

<table>
<thead>
<tr>
<th>$D$</th>
<th>$\sigma$</th>
<th>AOMA</th>
<th>CAIC</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>JMA</th>
<th>ML</th>
<th>Largest s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>0.3</td>
<td>0.093</td>
<td>0.096</td>
<td>0.099</td>
<td>0.088</td>
<td>0.095</td>
<td>0.088</td>
<td>0.111</td>
<td>0.126</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>1.381</td>
<td>1.361</td>
<td>1.201</td>
<td>1.978</td>
<td>1.217</td>
<td>1.525</td>
<td>1.188</td>
<td>2.256</td>
<td>0.204</td>
</tr>
<tr>
<td>$D_2$</td>
<td>0.3</td>
<td>0.114</td>
<td>0.113</td>
<td>0.126</td>
<td>0.126</td>
<td>0.119</td>
<td>0.123</td>
<td>0.156</td>
<td>0.148</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>1.912</td>
<td>2.331</td>
<td>2.234</td>
<td>2.985</td>
<td>2.053</td>
<td>2.432</td>
<td>1.781</td>
<td>2.739</td>
<td>0.252</td>
</tr>
<tr>
<td>$D_3$</td>
<td>0.3</td>
<td>0.150</td>
<td>0.162</td>
<td>0.170</td>
<td>0.183</td>
<td>0.168</td>
<td>0.178</td>
<td>0.182</td>
<td>0.170</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>1.637</td>
<td>2.580</td>
<td>2.922</td>
<td>4.107</td>
<td>2.935</td>
<td>3.325</td>
<td>1.800</td>
<td>2.571</td>
<td>0.220</td>
</tr>
<tr>
<td>$D_4$</td>
<td>0.3</td>
<td>0.170</td>
<td>0.218</td>
<td>0.199</td>
<td>0.228</td>
<td>0.199</td>
<td>0.214</td>
<td>0.171</td>
<td>0.199</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>2.029</td>
<td>3.732</td>
<td>2.492</td>
<td>4.931</td>
<td>2.612</td>
<td>3.854</td>
<td>1.900</td>
<td>2.338</td>
<td>0.313</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>AOMA</th>
<th>MALASSO</th>
<th>MM</th>
<th>ML</th>
<th>Largest s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.466</td>
<td>1.910</td>
<td>1.607</td>
<td>2.891</td>
<td>0.110</td>
</tr>
<tr>
<td>1</td>
<td>4.933</td>
<td>5.303</td>
<td>5.366</td>
<td>9.351</td>
<td>0.768</td>
</tr>
<tr>
<td>2</td>
<td>17.513</td>
<td>17.880</td>
<td>17.745</td>
<td>30.249</td>
<td>1.750</td>
</tr>
</tbody>
</table>

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 $\sigma$. 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 $X_l$ is $(x_{i,j1}, \ldots, x_{i,j9})$, and the jth row of $Z_l$ is $(1, z_{i,j2}, \ldots, z_{i,j4})$, $x_{i,j}$ for $l = 1, \ldots, 9$ and $z_{i,j}$ for $k = 2, 3, 4$ were independently generated from uniform $(-2, 2)$, $\beta = (1, 1, 0, \ldots, 0)^T$, $n = 30$, $m_l = 5$, $\sigma = 1$, and

$$D = \begin{pmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which means that the last random effect is unimportant. We also set $\sigma = 0.5$ 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 $0.05 \hat{\phi}_{\text{full}}$ 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.
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, 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 $\text{AIC}$, $\text{BIC}$ or conditional $\text{AIC}$ to select a final model. Such a selection strategy caused us difficulty because both $\text{AIC}$ and $\text{BIC}$ support model $(2, 1)$, while conditional $\text{AIC}$ 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 $\text{BIC}$. 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 $\text{AIC}$, $\text{AIC}$ and $\text{BIC}$ 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 $\text{AIC}$, smoothed $\text{BIC}$, jackknife model
The abbreviations are the same as in Table 1.

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

<table>
<thead>
<tr>
<th>Models</th>
<th>AIC</th>
<th>BIC</th>
<th>Weights of AIC</th>
<th>Weights of BIC</th>
<th>Weights of JMA</th>
<th>Weights of AOMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>78.6</td>
<td>318.3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.006</td>
<td>0.000</td>
</tr>
<tr>
<td>(1,3)</td>
<td>59.4</td>
<td>316.1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.303</td>
</tr>
<tr>
<td>(2,1)</td>
<td>40.1</td>
<td>161.5*</td>
<td>0.431</td>
<td>0.866</td>
<td>0.994</td>
<td>0.697</td>
</tr>
<tr>
<td>(2,2)</td>
<td>38.0</td>
<td>165.2</td>
<td>0.066</td>
<td>0.006</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(3,1)</td>
<td>42.0</td>
<td>162.4</td>
<td>0.266</td>
<td>0.113</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(3,2)</td>
<td>36.7*</td>
<td>165.9</td>
<td>0.046</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(4,1)</td>
<td>44.1</td>
<td>163.4</td>
<td>0.161</td>
<td>0.014</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(4,2)</td>
<td>37.3</td>
<td>166.8</td>
<td>0.030</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The abbreviations are the same as in Table 1.

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 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 \( \theta^* \) such that \( \hat{\theta}^{*}_{(s)} \rightarrow \theta^* = (\theta^*_1, \ldots, \theta^*_K) \)' in probability as \( N \rightarrow \infty \). Let \( P^*_{(s)} = \hat{P}(\theta) \bigg| \hat{\theta}^{(s)} = \theta^*_s \), \( P^*(w) = \hat{P}(\theta) \bigg| \hat{\theta}^{(s)} = \theta^*_s \), \( A^*(w) = \hat{A}(\theta) \bigg| \hat{\theta}^{(s)} = \theta^*_s \), and \( L^*(w) = \| \hat{\mu}_p(w) - \mu \|^2 \bigg| \hat{\theta}^{(s)} = \theta^*_s \). Then we have \( P^* \rightarrow E_{\theta^*}L^* \equiv \| A^* \mu \|^2 + \sigma^2 \text{tr}(P^{**}) \). Furthermore, define \( \xi_N = \inf_{w \in \mathcal{W}} R^*(w) \), \( K_{(s), j} = \partial G(x) / \partial \theta(s), j \), and \( \hat{K}_{(s), j} = K_{(s), j} \big| \hat{\theta}^{(s)} = \theta^*_s \), and let \( \lambda_{\text{max}}(\Phi) \) and \( \lambda_{\text{min}}(\Phi) \) denote the biggest and smallest singular values of a matrix \( \Phi \). Let \( w^*_z \) be a weight
vector with component unity for model $s$ and zero for other models, max, min indicate maximization, minimization over $j \in \{1, \ldots, J(s)\}$, and sup 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 $\partial \hat{\theta}(s) / \partial y = \hat{T}(s,j)y$ almost surely, where $\hat{T}(s,j)$ can be random and depend on $y$.

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

$$S_N^{-2\kappa} \sum_{s=1}^{S} \{R^*(u_{s}^0)\}^\kappa \to 0 \quad \text{almost surely},$$

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

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

and

$$N_N^{-1} \sup_s \left\{ \max_j \lambda_{\max}(\hat{\Theta}(s,j)) \right\} = o_p(1),$$

$$N_N^{-1} \sup_s \left\{ \lambda_{\max}(\hat{P}(s) - P^*(s)) \right\} = o_p(1),$$

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

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 $\hat{\theta}(s)$ consists of the elements of $D(s)$, each element of $\partial D(s) / \partial \hat{\theta}(s,j)$ is either 0 or 1, and in this case, Condition (A2) holds automatically.

The first part of Condition (A3) requires the inverse of maximum 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 $\lambda_{\max}(Z(s)Z(s)^*) = O(m_i)$ uniformly for $i \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, S\}$, where $Z(s,j)$ is the $j$th 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 $y^*\Psi(s,j)/N$ 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 $\theta_{\text{opt}}(s)$ at a rate such that $N_N^{-1} \sup \lambda_{\max}(\hat{P}(s) - P^*(s))$ 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 $Y(s)$ as an $N \times N$ matrix with $(i_1, i_2)$th element $Y(s, i_1, i_2) = \sum_{j_1=1}^{J(s)} \sum_{j_2=1}^{J(s)} (\hat{\theta}(s, j_1) - \theta^*_o(j_1)) (\hat{\theta}(s, j_2) - \theta^*_o(j_2)) \partial^2 P(s, i_1, i_2) / \partial \theta(s, j_1) \partial \theta(s, j_2)$, where $P(s, i_1, i_2)$ is the $(i_1, i_2)$th element of $P(s)$ and $\theta^*_o(j_i)$ is a $J(s) \times 1$ vector between $\hat{\theta}(s)$ and $\theta^*_o(s)$. When Conditions (A2) and (A3) hold, Condition (A5) is implied by

$$\lambda_{\max}(Y(s)) = o_p(1), \quad N^{1/2} (\hat{\theta}(s,j) - \theta^*_o(s)) = o_p(1), \quad N_N^{-2} \to 0,$$

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 $NR^*(w^0_s)^{-1} = O(1)$. Condition (A6), concerning the sum of $\mu_j$, is a commonly-used condition in linear regression models (Liang et al., 2011).
When \( \sigma^2 \) is unknown, the estimator of \( \theta_{(1)} \) is different from that for known \( \sigma^2 \), but in the following assumptions, we still use the notation \( \hat{\theta}_{(1)} \) as the estimator and the notation \( \theta^*_{(1)} \) as its limit for simplicity.

Assumption A2. As \( N \to \infty \), Conditions (A1)–(A4) and (A6) hold, Condition (A5) holds with \( \hat{P}_{(1)} \) being replaced by \( \hat{P}_{(1)} |_{\sigma^2=\hat{\sigma}^2} \), \( \hat{\sigma}^2 \to \sigma^2 \) in probability, and

\[
N \xi^{-1}_N (\hat{\sigma}^2 - \sigma^2) = o_p(1). \tag{A8}
\]

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

\[
\lambda_{\max}(\hat{\Sigma}_{(1)}^{-1}) = \lambda_{\max}\{(\sigma^2 I_N + Z(\theta) \hat{G}_{(1)} Z'(\theta))^{-1}\} \leq \sigma^{-2}, \quad s = 1, \ldots, S, \tag{A9}
\]

and that for any two \( N \times N \) matrices \( \Phi_1 \) and \( \Phi_2 \) (Li, 1987),

\[
\lambda_{\max}(\Phi_1 \Phi_2) \leq \lambda_{\max}(\Phi_1) \lambda_{\max}(\Phi_2), \quad \lambda_{\max}(\Phi_1 + \Phi_2) \leq \lambda_{\max}(\Phi_1) + \lambda_{\max}(\Phi_2). \tag{A10}
\]

Because \( \hat{V}_{(1)} \) is symmetric and idempotent, it follows from (A9) and (A10) that

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

and thus for any weight vector \( w \in W \), we obtain that

\[
\lambda_{\max}\{\hat{P}(w)\} \leq 2. \tag{A12}
\]

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

\[
E(\varepsilon_{i,j}^\kappa) \leq c < \infty. \tag{A13}
\]

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

\[
L(w) - \hat{C}(w) = \|\varepsilon\|^2 + \|\hat{A}(w)y\|^2 - 2\varepsilon\hat{A}(w)y - \|\hat{A}(w)y\|^2 - 2\sigma^2 \text{tr}[\hat{P}(w)] - 2\sigma^2 h(w)
\]

\[
= \|\varepsilon\|^2 - 2 \varepsilon' A'(w) \mu - 2 \varepsilon'[P^\ast(w) - \hat{P}(w)]\mu - 2 \varepsilon'[P^\ast(w) - \hat{P}(w)]\varepsilon
\]

\[
+ 2\sigma^2[\text{tr}[P^\ast(w)] - \text{tr}[\hat{P}(w)]\] + 2[\varepsilon'P^\ast(w)\varepsilon - \sigma^2 \text{tr}[P^\ast(w)]] - 2\sigma^2 h(w)
\]

and

\[
L(w) - R^\ast(w) = \|\hat{P}(w)\varepsilon - \hat{A}(w)\mu\|^2 - \|A^\ast(w)\mu\|^2 - 2\sigma^2 \text{tr}[P^\ast^2(w)]
\]

\[
= \mu' [P^\ast(w) - \hat{P}(w)] A^\ast(w) + \hat{A}(w)\mu - 2 \mu' A^\ast(w) P^\ast(w)\varepsilon
\]

\[
+ 2 \mu' [P^\ast(w) - \hat{P}(w)] \varepsilon - 2 \mu' P^\ast(w) [P^\ast(w) - \hat{P}(w)] \varepsilon
\]

\[
- 2 \mu' [P^\ast(w) - \hat{P}(w)] \hat{P}(w) \varepsilon - \varepsilon' [P^\ast(w) + \hat{P}(w)] [P^\ast(w) - \hat{P}(w)] \varepsilon
\]

\[
+ \varepsilon' P^\ast^2(w) \varepsilon - \sigma^2 \text{tr}[P^\ast^2(w)].
\]
So, similar to the proof of Theorem 2.1 of Li (1987), in order to prove Theorem 1, we need only to verify that

\[
\sup_{w \in W} \left\{ \frac{\epsilon'A'(w)\mu}{R^*(w)} \right\} = o_p(1),
\]

(A14)

\[
\sup_{w \in W} \left[ \frac{\epsilon'P^*(w)\epsilon - \sigma^2\text{tr}(P^*(w))}{R^*(w)} \right] = o_p(1),
\]

(A15)

\[
\sup_{w \in W} \left[ \frac{\mu'A'(w)P^*(w)\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A16)

\[
\sup_{w \in W} \left[ \frac{\epsilon'P^2(w)\epsilon - \sigma^2\text{tr}(P^2(w))}{R^*(w)} \right] = o_p(1),
\]

(A17)

\[
\sup_{w \in W} \left[ \frac{\epsilon'(P^*(w) + \hat{P}(w))P^*(w) - \hat{P}(w)\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A18)

\[
\sup_{w \in W} \left[ \frac{\mu'(P^*(w) - \hat{P}(w))\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A19)

\[
\sup_{w \in W} \left[ \frac{\mu'P^*(w)\{P^*(w) - \hat{P}(w)\}\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A20)

\[
\sup_{w \in W} \left[ \frac{\mu'P^*(w)(P^*(w) - \hat{P}(w))\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A21)

\[
\sup_{w \in W} \left[ \frac{\mu'(P^*(w) - \hat{P}(w))\hat{P}(w)\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A22)

\[
\sup_{w \in W} \left[ \frac{\mu'(P^*(w) - \hat{P}(w))(A^*(w) + \hat{A}(w))\epsilon}{R^*(w)} \right] = o_p(1),
\]

(A23)

\[
\sup_{w \in W} \left[ \frac{\text{tr}(P^*(w)) - \text{tr}(\hat{P}(w))}{R^*(w)} \right] = o_p(1),
\]

(A24)

and

\[
\sup_{w \in W} \left\{ |h(w)| / R^*(w) \right\} = o_p(1).
\]

(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

\[
\sup_{w \in W} \left[ \frac{\epsilon'(P^*(w) + \hat{P}(w))\{P^*(w) - \hat{P}(w)\}\epsilon}{R^*(w)} \right] 
\]

\[
\leq \xi_N^{-1}2^{-1} \left( \sup_{w \in W} \left[ \frac{\epsilon'\{P^*(w) + \hat{P}(w)\}\{P^*(w) - \hat{P}(w)\}\epsilon}{R^*(w)} \right] 
\right.
\]

\[
\leq \xi_N^{-1}2^{-1} \|\epsilon\|^2 \sup_{w \in W} \lambda_{\max}\left[ \{P^*(w) + \hat{P}(w)\}\{P^*(w) - \hat{P}(w)\} \right] 
\]

\[
\leq \xi_N^{-1}2^{-1} \|\epsilon\|^2 \sup_{w \in W} \lambda_{\max}\{P^*(w) + \hat{P}(w)\} 
\]

\[
\leq \xi_N^{-1}4\|\epsilon\|^2 \sup_{w \in W} \left\{ \sum_{s=1}^S w_s \lambda_{\max}(P^*_s - \hat{P}(s)) \right\} 
\]

\[
\leq \xi_N^{-1}4\|\epsilon\|^2 \sup_{w \in W} \left\{ \sum_{s=1}^S w_s \lambda_{\max}(P^*_s - \hat{P}(s)) \right\} 
\]
where the fifth inequality is from (A12) and the last step is from Condition (A5). Similarly, we can prove (A19). For (A20),

\[
\sup_{w \in W} \left| \mu'(P^*(w) - \hat{P}(w)) \right| / R^*(w) \leq \xi_N^{-1} N \sup_{w \in W} \left[ \| \mu' \|_2 \| e' (P^*(w) - \hat{P}(w)) \| e \right]^{1/2} \\
\leq \| \mu' \| / N^{1/2} \left( \| e' \| / N \right)^{1/2} \xi_N^{-1} N \sup_s \lambda_{\max}(\hat{P}(s) - P^*_s) \\
= o_p(1),
\]

(A26)

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

For (A24),

\[
\sup_{w \in W} \left| \text{tr}(P^*(w) - \text{tr}(\hat{P}(w))) / R^*(w) \right| \leq \xi_N^{-1} N \sup_s \lambda_{\max}(P^*_s - P^*_s) = o_p(1),
\]

where the equality is from Condition (A5).

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

\[
\sup_s \max_j \lambda_{\max}(U_{(s),j}) \leq \sup_s \max_j \lambda_{\max}(\hat{K}_{(s),j}) \lambda_{\max}(\tilde{\Sigma}_j^{-1/2} Z_{(s),j} Z_{(s),j}' \tilde{\Sigma}_j^{-1/2}) \\
\leq \sigma^{-2} \sup_s \max_j \lambda_{\max}(\hat{K}_{(s),j}) \sup_s \lambda_{\max}(Z_{(s),j} Z_{(s),j}')
\]

(A27)

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

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

(A28)

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

\[
\sup_s \max_j \lambda_{\max}(U_{(s),j}) = O_p(1).
\]

(A29)

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

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

(A30)
By (A13), (A30) and Conditions (A4) and (A6), it is observed that
\[
\sup_{w \in W} \{ |h(w)| / R^*(w) \} \leq \xi_N^{-1} \sup_{x} \{ J\sup_{x} \{ \max_{y} |y^T T_{x,j} \hat{A}_{x,j} \hat{W}_{x,j} \hat{A}_{x,j} y| \} \\
\leq \xi_N^{-1} \sup_{x} \{ J\sup_{x} \{ \max_{\lambda_{\max}}(T_{x,j}, \hat{A}_{x,j} \hat{W}_{x,j} \hat{A}_{x,j}) \} \\
\leq \sup_{x} J_{x} N^{-1} \|\| y \|\|^2 N \xi_N^{-1} \sup_{x} \{ \max_{\lambda_{\max}}(T_{x,j}) \} \sup_{x} \{ \max_{\lambda_{\max}}(\hat{A}_{x,j} \hat{W}_{x,j} \hat{A}_{x,j}) \} \\
= o_p(1),
\]
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
\[
\hat{\sigma}^2 \sup_{w \in W} \{ |h(w)|_{\hat{\sigma}^2=\hat{\sigma}^2} / R^*(w) \} = o_p(1) \tag{A31}
\]
and
\[
|\sigma^2 - \hat{\sigma}^2| \sup_{w \in W} \{ \max_{\lambda_{\max}}(P^*_x) \}/R^*(w) = o_p(1). \tag{A32}
\]

When $\sigma^2$ 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 $\lambda_{\max}(P^*_x) \leq 2$, so
\[
|\sigma^2 - \hat{\sigma}^2| \sup_{w \in W} \{ \max_{\lambda_{\max}}(P^*_x) \}/R^*(w) \leq \xi_N^{-1} |\sigma^2 - \hat{\sigma}^2| \sup_{x} \{ \max_{\lambda_{\max}}(P^*_x) \}
\leq \xi_N^{-1} |\sigma^2 - \hat{\sigma}^2| \sup_{x} \{ \max_{\lambda_{\max}}(P^*_x) \}
\leq 2N\xi_N^{-1} |\sigma^2 - \hat{\sigma}^2|,
\]
which together with Condition (A8) implies (A32). This completes the proof of Theorem 2.

**References**


[Received November 2012. Revised September 2013]