Noname manuscript No.
(will be inserted by the editor)

Combining models in longitudinal data analysis

Song Liu · Yuhong Yang

Abstract

Model selection uncertainty in longitudinal data analysis is often much more serious than that in simpler regression settings, which challenges the validity of drawing conclusions based on a single selected model when model selection uncertainty is high. We advocate the use of appropriate model selection diagnostics to formally assess the degree of uncertainty in variable/model selection as well as in estimating a quantity of interest. We propose a model combining method with its theoretical properties examined. Simulations and real data examples demonstrate its advantage over popular model selection methods.

Keywords: Adaptive regression by mixing; Longitudinal data; Model combining; Model selection; Model selection diagnostics; Model selection uncertainty

Song Liu
Consumer Banking JPMorgan Chase, 1111 Polaris Parkway, Columbus, OH 43240 USA
E-mail: song.x.liu@jpmchase.com

Yuhong Yang
School of Statistics, The University of Minnesota, 313 Ford Hall 224 Church Street SE, Minneapolis, MN 55455 USA
E-mail: yyang@stat.umn.edu

1 Introduction

Longitudinal data arise frequently in many scientific studies where each of independent subjects is measured repeatedly over a time period. A variety of modeling approaches have been proposed for handling such data. Linear models, as is the focus of this paper, are commonly used for continuous response (see, e.g., Diggle, *et al*, 2002; Fitzmaurice, *et al*, 2004); semi-parametric and nonparametric models are useful for modeling more flexible structures (see e.g., Lin and Ying 2001; Ruppert, *et al* 2003). When multiple models are considered (which is almost always the case), model comparison is a critical step for reaching reliable conclusions.

For longitudinal data, model/variable selection issues have not been much addressed. A few exceptions include Pan (2001), Cantoni *et al.* (2007), Yafune *et al.* (2005), where familiar model selection methods are adapted to the longitudinal case. Fitzmaurice *et al.* (2004) give a general strategy to obtain a sensible choice of models for both the covariance and the mean: first select an appropriate model for the covariance and then select a model for the mean. In a semi-parametric setting, Fan and Li (2004) propose a penalized weighted least squares procedure and establish an asymptotic efficiency property for selecting significant variables. Huang *et al.* (2006) propose nonparametric methods for covariance matrix selection and estimation by modified Cholesky decomposition. Wang and Qu (2009) derive a consistent model selection rule in the estimation equations approach.

The problem of model selection uncertainty is now well known (see, e.g., Draper, 1995; Chatfield, 1995; Breiman, 1996; Hoeting *et al.*, 1999). In the context of longitudinal data analysis, the issue can be substantially more serious due to the uncertainty in both mean and covariance modeling, which makes it usually much harder to choose

the "best" combination of the mean and variance models when compared with the cross-sectional data case.

Clearly, in medical and other applications, interpretations are desirable and important. The approach of selecting a single model and then basing all inference on the model is natural and often helpful for that purpose. In case of high uncertainty in model selection, a convenient interpretation/conclusion on which variables are important for explaining the response may be seriously misleading. Going beyond being aware of the uncertainty in model selection, a proper assessment of the uncertainty in finding the best model as well as in estimating a quantity of interest is desirable. It is quite possible that the model selection uncertainty has a substantial impact on some quantities of interest but not so on others. It is thus important to differentiate situations between severe uncertainty and negligible ones for estimating a parameter of scientific significance.

Methods that try to address the problem of model selection uncertainty have been proposed from different perspectives, including Baysian model averaging (e.g., Hoeting *et al*, 1999), some non-Baysian approaches (e.g., Buckland *et al.*, 1997; Breiman, 1996; Hjort and Claeskens, 2003). Yang (2001, 2003) propose a model combining method ARM based on information-theoretic tools. In addition to its applicability to combine both parametric and nonparametric methods, another advantage of this approach is that the good performance of the combined estimator is theoretically characterized with non-asymptotic risk bounds. In our opinion, non-asymptotic characterizations of performance are preferred to asymptotic expressions when model selection uncertainty is high due to the often lack of reliability of asymptotic arguments. To our knowledge, no model combining methods with risk properties have been derived for longitudinal data analysis.

The objective of this work is mainly two-fold. First, we propose model selection diagnostic measures to assess reliability of model selection for longitudinal data. When the measures indicate there is not much model selection uncertainty, conclusion and interpretations based on a properly selected model are sound. When there is much evidence of severe model selection uncertainty, however, results based on a single selected model may not be trusted. A sensible alternative is to focus on estimating the regression function by model combination. Second, we study a model combination method and obtain oracle inequalities for combining longitudinal models.

The rest of the paper is organized as follows. We set up the problem in Section 2. Model selection diagnostics to assess the uncertainty in model selection are proposed in Section 3. In Section 4, we propose the ARM algorithm for longitudinal data and give theoretical results on the combined estimator. In Section 5, we compare ARM with model selection methods via fair and informative simulations. An application on real data sets is presented in Section 6. Concluding remarks are in Section 7. The proofs of the theoretical results are in an appendix.

2 Problem setup

Let $Y_{ij}, j = 1, ..., n_i$ be the sequence of observed measurements on the i-th subject, $i = 1, ..., m$; t_{ij} be the time when the j-th measurement for the i-th subject is taken. Associated with each Y_{ij} there are p explanatory variables, X_{ijk}, $k = 1, ..., p$. Consider the regression model:

$$Y_{ij} = f(X_{ij}) + e_{ij},\ j = 1, 2..., n_i, i = 1, ...m,$$

where f is the true regression function, $X_{ij} = (X_{ij1}, ..., X_{ijp})'$ is the vector of predictors, and the e_{ij} are Gaussian errors conditional on the predictors. We expect the errors to be correlated within subjects. Throughout this paper, $X_i = (X_{i1}, ..., X_{in_i})'$

are assumed to be independent of each other for different subjects and $E(e_{ij}|X_{ij}) = 0$ for all i and j. Using vector and matrix notation, the model is

$$Y_i = f(X_i) + e_i,\ i = 1, ..., m, \tag{1}$$

where $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^T$, $f(X_i) = (f(X_{i1}), f(X_{i2}), ..., f(X_{in_i}))^T$ and e_i is conditionally Gaussian with mean $\mathbf{0}$ and covariance matrix V_i. In the mixed-effect model framework, e_i can be further divided into two parts, $e_i = Z_i b_i + \epsilon_i$, where b_i is the random effect from $N(\mathbf{0}, G)$ and ϵ_i is the within-subject random errors from $N(\mathbf{0}, R_i)$ for some positive definite matrices G and R_i.

For estimating f, linear combinations of the predictors are considered. Suppose that K such linear models are considered as candidates for fitting the data. The k-th model is

$$Y_i = f_k(X_i; \beta_k) + e_i \text{ and } e_i \sim N(\mathbf{0}, V_i(\alpha_k)),\ i = 1, ..., m;$$

where f_k is a linear combination of a subset of the predictors, and β_k and α_k are the parameters for the mean function and covariance function respectively. Here V_i denotes a specific covariance structure. For instance if an exchangeable covariance structure is used for model k, then V_i denotes the form and α_k is the correlation parameter (usually denoted by ρ) between any two observations within the same subject. For a given model, let $\widehat{\beta}_k$ and $\widehat{\alpha}_k$ be the maximum likelihood or other appropriate estimators.

Although the focus of this work is on linear models for simplicity, the methodology and theory work more generally. The explanatory variables can be time-invariant or time-variant.

When there is much uncertainty in model selection, finding the "true model" is not realistic. Prediction, as emphasized by Geisser (1993, Chapter 1), is perhaps of more interest, and it serves as a means to assess different statistical models/methods. We

consider the loss function of the form

$$(Y_i - \widehat{Y}_i)^T V_i^{-1} (Y_i - \widehat{Y}_i), \tag{2}$$

which is equivalent in expectation to $\{f(X_i) - \widehat{f}(X_i)\}^T V_i^{-1} \{f(X_i) - \widehat{f}(X_i)\}$, a suitable loss for estimating f. For theoretical results, we will consider the risk

$$\sum_i E\{f(X_i) - \widehat{f}(X_i)\}^T V_i^{-1} \{f(X_i) - \widehat{f}(X_i)\}.$$

For empirical comparison of model selection and model combining methods using real data, we will consider random data splittings and use an average prediction error as the performance measure (see Section 6 for details).

3 Model selection diagnostics through instability measures

Whereas some researchers favor model combining over model selection, we do not believe simply combining the candidate models is always the solution. First, if the quantities of interest are minimally affected by model selection uncertainty, why should one bother to use a much more complicated approach? Second, unlike statements under idealized assumptions (e.g., Burnham and Anderson, 2004, p. 293), in reality any model combining method can do much worse than model selection, especially when the best model is readily identified (see Juditsky and Nemirovski, 2000; Yang, 2004; Tsybakov, 2003, for results that quantify the "price" of combining procedures). We will see in Section 6 that combining by ARM is certainly not always superior to model selection. A natural question then is: when should model combining be favored? This is our motivation of considering instability measures to describe model selection uncertainty. In our opinion, model selection diagnostics are at least as important as model diagnostics. It is perfectly possible that a selected model passes the usual model diagnostics yet the model selection diagnostics show that the process is highly uncertain and thus

the selected model is not trustworthy. Without model selection diagnostic measures, presenting only the estimates from the selected model is potentially misleading.

In this section, we propose to use simple and intuitive instability measures for selecting a model as well as for estimating a quantity of interest. The measures, based on well-known bootstrap and data perturbation ideas, in our opinion, provide critical information regarding reliability of a selected model or an estimate based on it. An application on real data will be given in Section 6.

3.1 Bootstrap instability in selection (BIS)

Breiman (1996) uses nonparametric bootstrap methods (e.g., Efron and Tibshirani, 1993) to perform *Bagging* to improve an estimator. Note, in our context, in order to keep the correlation structure, resampling should be applied on subjects. The bootstrap instability in selection (BIS) is based on parametric bootstrap resampling as follows.

Step 1. Apply a model selection strategy on the original data to select the "best" model (both mean and covariance).

Step 2. Generate bootstrap samples from the selected model. Let $\tilde{Y}_i = \widehat{f}(X_i) + \tilde{e}_i$, $i = 1, ..., m$, where $\tilde{e}_i$ is generated from $N(\mathbf{0}, \widehat{V}_i)$ distribution and $\widehat{f}$ is based on the selected model with estimated parameters, and let $(\tilde{Y}_i, X_i)_{i=1}^m$ be the bootstrap sample.

Step 3. Apply the model selection strategy again on the bootstrap sample and choose the "best" model.

Step 4. Repeat Steps 2 and 3 a large number of times (say B). Then BIS is defined to be the fraction of times that the original "best" model is not selected.

If the model selection is stable for the current problem, we expect that most of the time we will choose the same mean and covariance models as in Step 1. Thus if BIS is

larger than 50%, we need to honestly admit the infeasibility of finding the best model by the model selection method.

3.2 Bootstrap instability in estimation (BIE)

Instability can also be measured in terms of estimation of the regression function or other unknown quantities.

Suppose a one-dimensional model-independent estimand is Q. Then BIE of Q is calculated in the same way as BIS except that after a model is chosen we obtain an estimate of Q, denoted by $\widehat{Q}$ and $\widehat{Q}^{(b)}$ from the original data and bootstrap data respectively and $\widehat{Q}$ is assumed not to be zero. BIE based on B bootstrap samples is given by

$$\mathrm{BIE}(\widehat{Q}) = \frac{\frac{1}{B}\sum_b |\widehat{Q} - \widehat{Q}^{(b)}|}{|\widehat{Q}|}.$$

A large BIE value, say greater than 0.5, indicates severe instability in estimation of Q.

It should be noticed that the above instability in estimation in fact consists of two sources of uncertainty: the parameter estimation uncertainty and the instability due to model selection. Even if no model selection procedure is performed, the parameter estimates from the bootstrap samples are different from the original one and a high value in BIE may well be mainly from the parameter estimation. We propose BIE_s as follows.

First let BIE_p be the pure bootstrap instability in estimation (without model selection) calculated in a similar way as BIE, i.e.,

$$\mathrm{BIE}_p(\widehat{Q}) = \frac{\frac{1}{B}\sum_b |\widehat{Q} - \check{Q}^{(b)}|}{|\widehat{Q}|},$$

where $\check{Q}^{(b)}$ is estimated from the bootstrap sample based on the originally selected model. Then we define the bootstrap instability in estimation due to model selection

as

$$\mathrm{BIE}_s(\widehat{Q}) = \mathrm{BIE}(\widehat{Q})/\mathrm{BIE}_p(\widehat{Q}) - 1.$$

If the instability in estimation indeed mainly comes from the model selection uncertainty, we expect a large BIE_s value.

3.3 Perturbation instability in estimation

Data perturbation is an alternative to bootstrap resampling for measuring instability, and it can bring in additional information about the model selection process. Some related previous uses of perturbation are in Breiman (1996), Ye (1998), Shen and Ye (2002), and Yuan and Yang (2005). Our construction of perturbation instability has two distinct features. One is that we measure the perturbation instability due to model selection (not the overall instability) and the other is that we perturb two components, which provides information on sources of high instability that is not available in BIE.

We focus on the mixed-effect model case, where there are between-subject and within-subject variabilities corresponding to two kinds of randomness, random effects and (within-subject) random errors. We generate a new set of perturbation random effects $\tilde{b}_i$ from $N(\mathbf{0}, \tau_1^2 \widehat{G})$ and a new set of perturbation random errors $\tilde{\epsilon}_i$ from $N(\mathbf{0}, \tau_2^2 \widehat{R}_i)$, where τ_1 and τ_2 are perturbation sizes between 0 and 1, and $\widehat{G}$ and $\widehat{R}_i$ are estimates for G and R_i based on the selected model. Consider $\tilde{Y}_i = Y_i + Z_i\tilde{b}_i + \tilde{\epsilon}_i$ for $i = 1, 2, ..., m$ and apply the model selection procedure under examination on the perturbed data $(\tilde{Y}_i, X_i)$ for $i = 1, 2, ..., m$. In general, the random effects $\tilde{b}_i$ can be a vector, but we only consider random intercept model and thus it is a scalar here. At each permutation size (τ_1, τ_2), we do this a large number of time (say M) and compute the average deviation of the perturbed estimates of the quantity of interest:

$$I(\tau_1, \tau_2) = \frac{1}{M} \sum_{r=1}^{M} |\widehat{Q} - \tilde{Q}^{(r)}|,$$

where $\tilde{Q}^{(r)}$ is obtained from the r-th perturbed data.

From the definition, how fast I changes in τ_1 and τ_2 is a suitable instability measure. Let $I'_{\tau_1} = \frac{\partial I(\tau_1,\tau_2)}{\partial \tau_1}|_{\tau_1=0,\tau_2=0}$ and $I'_{\tau_2} = \frac{\partial I(\tau_1,\tau_2)}{\partial \tau_2}|_{\tau_1=0,\tau_2=0}$. Then PIE for estimating Q based on the model selection procedure can be defined as

$$\text{PIE}(\widehat{Q}) = \sqrt{\{I'_{\tau_1}\}^2 + \{I'_{\tau_2}\}^2},$$

and we define the perturbation instability in estimation *due to model selection* as:

$$\text{PIE}_s(\widehat{Q}) = \text{PIE}(\widehat{Q})/\text{PIE}_p(\widehat{Q}) - 1,$$

where PIE_p is the pure perturbation instability in estimation without model selection.

To understand the main source of instability between the two components of randomness, namely random effects and random errors, the vector $\mathbf{PIE} = (I'_{\tau_1}, I'_{\tau_2})^T$ gives useful information. It can provide better insight into the problem and may be helpful for planning a future study. If a large instability is mainly due to random errors, then the improvement for a more accurate estimation should focus on instrumental devices, for instance designing a better questionnaire for a survey study or implementing a more precise measuring device for a plant-growth experiment; if the instability is mainly due to subject variation, then perhaps selecting more homogeneous subjects can help for estimating some quantities of interest (other than the random effect).

To estimate PIE, we first fix $\tau_2 = 0$ and consider equally spaced τ_1 values from 0 to 1 with width 0.1, and obtain the corresponding values $I(\tau_1, 0)$. Then use simple linear regression ($I(\tau_1, 0)$ versus τ_1) to estimate the first term in PIE, I'_{τ_1}, by the slope of the linear regression. Then fix $\tau_1 = 0$ and let τ_2 vary from 0 to 1 to estimate the second term, I'_{τ_2}, in the same way. Combining the two terms leads to a reasonable estimate of PIE. Note that the perturbation plot of I versus τ_1 or τ_2 (with τ_2 or τ_1 fixed at zero) is typically linear in τ_1 or τ_2 (see, e.g., Figures 1 and 2 in a simpler regression context

in Yuan and Yang, 2005), and thus the derivative of I can be estimated by the slope of linear regression.

3.4 On using the model-selection diagnostic measures

We have proposed three instability measures as model-selection diagnostic values to assess the uncertainty due to model selection for longitudinal data analysis. They allow us to make a proper judgment regarding the model selection reliability. Based on our numerical investigations, we suggest the following for application.

1. First look at the BIS value. If BIS is no larger than 0.5, we are reasonably comfortable about the selected model and then work with it for inference or prediction. However, if BIS is greater than 0.5, we know that model selection uncertainty is too high and the selected model cannot be taken as the best (or true) with reasonable confidence. We then continue as follows.
2. For an estimand of interest, we find its BIE or PIE. If the values are no larger than 0.5, the model selection uncertainty does not seem to have much effect on estimating the quantity of interest and the estimate of the estimand from the selected model can be reasonably trusted. Otherwise (i.e., if BIE ≥ 0.5 or PIE ≥ 0.5), continue as follows.
3. Compute BIE_s or PIE_s. Lack of reliability due to model selection is announced if BIE_s or PIE_s is bigger than 1, which roughly means that the instability due to model selection alone is worse than that of parameter estimation. Then model averaging should be considered for the estimation task. Otherwise (i.e., BIE_s and PIE_s are no larger than 0.5), uncertainty due to model selection does not cause much additional trouble beyond uncertainty of parameter estimation (which cannot be avoided) and thus may be tolerated.

Of course, the above suggestions are not meant to be taken rigidly. Specific context and subject area knowledge/judgment are relevant.

3.5 Pre-combining Strategies

For longitudinal data analysis, a large number of combinations of covariance models and mean models can be explored to better capture the complexity of the data. For the mean, besides variable selection, one often considers transformations as well. The computational cost can be very high for combining all these models. To alleviate the burden, we consider some pre-combining strategies to reduce the number of models to be combined.

We first consider a large or full (including all the terms) model for the mean and screen on covariance models. We randomly split the data into two parts and use the first part to do estimation and the second to compute the prediction errors. We then compute weights for the different covariance models. The weighting is similar to the one we use in the ARM algorithm to be proposed in Section 5. Based on the weights, we decide which covariance models to be removed. Once the covariance specifications are narrowed down, one can use various model selection methods to screen out less important variables. A demonstration will be given in the data example section.

4 Combining longitudinal models by ARM

4.1 The algorithm of ARM for longitudinal data

For combining longitudinal models, the correlation between observations within subjects must be taken into account. We propose the following ARM algorithm. For simplicity, assume m is even in computing the weights for the models.

1. Randomly split the data into two parts $Z^{(1)} = (X_i, Y_i)$, $1 \le i \le m/2$ and $Z^{(2)} = (X_i, Y_i)$, $m/2+1 \le i \le m$. Note that we are splitting on subjects, the natural sampling unit in the context of longitudinal data.

2. Estimate β_k and $V_i(\alpha_k)$ by MLE (or other sensible methods) on $Z^{(1)}$. Let $\widehat{f}_{k,Z^{(1)}}(x) = \widehat{f}_{k,Z^{(1)}}(x; \widehat{\beta}_{k,Z^{(1)}})$.

3. Assess the accuracies of the models using the second part of the data $Z^{(2)}$. For each k, for $m/2+1 \le i \le m$, predict Y_i by $\widehat{f}_{k,Z^{(1)}}(X_i, \widehat{\beta}_{k,Z^{(1)}})$. Compute the overall measure of discrepancy:

$$D_k = \sum_{i=m/2+1}^{m} \{Y_i - \widehat{f}_{k,Z^{(1)}}(X_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k,Z^{(1)}})^{-1} \{Y_i - \widehat{f}_{k,Z^{(1)}}(X_i)\}. \tag{3}$$

Ideally, we would use the true covariance matrix, V_i, in the above equation, but in practice an estimate, $\widehat{V}_i$, is used.

4. Compute the weight for model k:

$$W_k = \frac{\mathit{\Pi}_{i=m/2+1}^{m} |\widehat{V}_i(\widehat{\alpha}_{k,Z^{(1)}})|^{-1/2} \exp(-D_k/2)}{\sum_{l=1}^{K} \mathit{\Pi}_{i=m/2+1}^{m} |\widehat{V}_i(\widehat{\alpha}_{l,Z^{(1)}})|^{-1/2} \exp(-D_l/2)}, \tag{4}$$

where $|\widehat{V}_i|$ denotes the determinant of $\widehat{V}_i$.

5. Repeat the above steps $P-1$ times and let $W_{k,p}$ denote the weight of model k at the p-th permutation for $0 \le p \le P-1$. Let $\widehat{W}_k = \frac{1}{P} \sum_{p=0}^{P-1} W_{k,p}$ be the final weight of model k.

6. The ARM estimator of the true regression function f is $\widehat{f}(x) = \sum_{k=1}^{K} \widehat{W}_k \widehat{f}_{k,Z}(x; \widehat{\beta}_{k,Z})$.

In the above algorithm, data splitting ratio is 1 : 1. Our experience suggests that half-half splitting works well, which results in optimal estimation of the regression function in rate of convergence, as well be seen in the next section. It should be pointed out that, although our focus in this paper is on linear models, the ARM method does not require the candidate models to be linear or parametric. In fact, it works for combining general linear, nonlinear and nonparametric/semi-parametric models.

4.2 A theoretical result for ARM

Risk bounds will be given in this section under two difference losses, Kullback-Leibler divergence and squared L_2 loss. In statistics and machine learning literature, besides the risk at a given sample size, cumulative risks are often of interest as well. Suppose we begin the estimation process at the sample size m_0 and continue the task sequentially as more observations come in one by one. Let $\widehat{f}_{k,l}(x) = \widehat{f}_{k,l}(x;\widehat{\beta}_{k,l})$ be the estimate of the true regression function and $\widehat{V}_i(\widehat{\alpha}_{k,l})$ be the estimate of V_i by model k based on data $(Y_i, X_i)_{i=1}^l$. For $i = m_0+1$, let $W_{k,i} = 1/K$ and for $m_0+1 < i \le m$, let $W_{k,i}$ be

$$\frac{\Pi_{l=m_0+1}^{i-1}|\widehat{V}_l(\widehat{\alpha}_{k,l-1})|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\sum_{l=m_0+1}^{i-1}\{Y_l-\widehat{f}_{k,l-1}(X_l)\}^T\widehat{V}_l(\widehat{\alpha}_{k,l-1})^{-1}\{Y_l-\widehat{f}_{k,l-1}(X_l)\}\right]}{\sum_{k=1}^K\Pi_{l=m_0+1}^{i-1}|\widehat{V}_l(\widehat{\alpha}_{k,l-1})|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\sum_{l=m_0+1}^{i-1}\{Y_l-\widehat{f}_{k,l-1}(X_l)\}^T\widehat{V}_l(\widehat{\alpha}_{k,l-1})^{-1}\{Y_l-\widehat{f}_{k,l-1}(X_l)\}\right]}. \tag{5}$$

Now for $i = 1, 2, ..., m$, the true density function of Y_i given the predictors is

$$p_i = \frac{1}{(2\pi)^{n_i/2}|V_i|^{1/2}}\exp[-\frac{1}{2}\{y_i - f(x_i)\}^T V_i^{-1}\{y_i - f(x_i)\}];$$

the estimated density function under model k is

$$\widehat{q}_{k,i} = \frac{1}{(2\pi)^{n_i/2}|\widehat{V}_i(\widehat{\alpha}_{k,i-1})|^{1/2}}\exp[-\frac{1}{2}\{y_i-\widehat{f}_{k,i-1}(x_i)\}^T\widehat{V}_i(\widehat{\alpha}_{k,i-1})^{-1}\{y_i-\widehat{f}_{k,i-1}(x_i)\}];$$

and the combined estimate is

$$\widehat{g}_i = \sum_k \frac{1}{(2\pi)^{n_i/2}|\widehat{V}_i(\widehat{\alpha}_{k,i-1})|^{1/2}}W_{k,i}\exp[-\frac{1}{2}\{y_i-\widehat{f}_{k,i-1}(x_i)\}^T\widehat{V}_i(\widehat{\alpha}_{k,i-1})^{-1}\{y_i-\widehat{f}_{k,i-1}(x_i)\}].$$

Let $D(p||g) = \int p(x)\log\frac{p(x)}{g(x)}dx$ denote the Kullback-Leibler (K-L) divergence between p and g. We have the following oracle inequality. Proofs of the theorems in this section are given in Web Appendix D.

Theorem 1 *The cumulative risk of the combined procedure by ARM satisfies*

$$\sum_{i=m_0+1}^m ED(p_i||\widehat{g}_i) \le \log K + \inf_k \sum_{i=m_0+1}^m ED(p_i||\widehat{q}_{k,i}).$$

Thus in terms of the cumulative K-L risk, our combined estimator achieves the smallest among the candidates up to a penalty $\log K$ (which is negligible as $m \to \infty$). To derive a risk bound under the squared L_2 loss, more conditions are required.

Condition 1: *There exists a positive constant τ, such that for all $i > 1$, with probability one,*

$$\sup_k |\{f(x) - \widehat{f}_{k,i-1}(x)\}^T V_i^{-1} \{f(x) - \widehat{f}_{k,i-1}(x)\}| \leq \tau.$$

Condition 2: *There exist positive constants $0 < \eta_1 \leq 1$ and $1 \leq \eta_2 < \infty$, such that with probability one for all $i > 1$ and k,*

$$\eta_1 V_i \leq \widehat{V}_i(\widehat{\alpha}_{k,i-1}) \leq \eta_2 V_i,$$

where for two matrices A and B, $A \leq B$ means $B - A$ is nonnegative definite.

A sufficient condition for Condition 2 is that the eigenvalues of each covariance matrix family $\{V_i(\alpha_k)\}$, $i = 1, 2, ..., m$ and $k = 1, ..., K$, are uniformly bounded away from zero and infinity, which is typically satisfied for parametric covariance models if the parameter α_k is bounded away from boundaries.

Conditions 1 and 2 require that the individual estimates are not too far away from the true values, as are commonly used for deriving risk bounds (not asymptotic expressions) for function estimation. The constants involved are not required to be known. We give below a simple example that satisfies both conditions. Let $n^* = \max\{n_{m_0+1}, ..., n_m\}$.

1. $||f||_\infty \leq A < \infty$ for some constant A and the estimates $\widehat{f}_{k,i-1}$ are restricted to be in the range;

2. Assume $V_i = \sigma^2 V_{0i}$, $i \geq 1$, where the correlation matrix V_{0i} is known and σ^2 is unknown, but known to be between $\underline{\sigma}^2 > 0$ and $\overline{\sigma}^2 < \infty$. The estimates of σ^2 are bounded accordingly;

3. $n^* \leq B < \infty$ *for some positive constant* B*.*

We define a loss function to gauge performance of covariance matrix estimation as

$$L(\widehat{V}, V) = \mathrm{tr}\{V(\widehat{V}^{-1} - V^{-1})\} - \log\{\det(V\widehat{V}^{-1})\}.$$

Note that the same loss function can be found in Huang *et al.* (2006).

Theorem 2 *Assume that Conditions 1 and 2 are satisfied, then the risk of the combined regression estimator satisfies:*

$$\begin{aligned}
&\sum_{i=m_0+1}^{m} E\{f(X_i) - \sum_k W_{k,i}\widehat{f}_{k,i-1}(X_i)\}^T V_i^{-1}\{f(X_i) - \sum_k W_{k,i}\widehat{f}_{k,i-1}(X_i)\} \\
&\leq \frac{(2+\xi)n^* + 2.5\tau}{\eta_1}\left(2\eta_1 \log K + \inf_k \sum_{i=m_0+1}^{m}\left[E\left\{f(X_i) - \widehat{f}_{k,i-1}(X_i)\right\}^T V_i^{-1}\left\{f(X_i) - \widehat{f}_{k,i-1}(X_i)\right\}\right.\right. \\
&\quad \left.\left. + \eta_1 EL\{\widehat{V}_i(\widehat{\alpha}_{k,i-1}), V_i\}\right]\right),
\end{aligned}$$

where $\xi = \max\{|\eta_2 - 1|, |\eta_1 - 1|\}$*. If further* (Y_i, X_i) *are i.i.d. and* V_i*,* $i = 1, 2, ..., m$*, are identical with* V *(of dimension* n_0*), then for the combined estimator* $\tilde{f} = \sum_{k=1}^{K}(\frac{2}{m}\sum_{i=m/2+1}^{m} W_{k,i})\widehat{f}_{k,m/2}$ *with* $W_{k,i}$ *computed by* (5) *with* $\widehat{f}_{k,l} = \widehat{f}_{k,m/2}$ *and* $\widehat{V}(\widehat{\alpha}_{k,l}) = \widehat{V}(\widehat{\alpha}_{k,m/2})$*, we have*

$$E||f - \tilde{f}||_V^2 \leq \frac{1}{\eta_1}\{(2+\xi)n_0 + 5\tau/2\}\left(\frac{4\eta_1 \log K}{m} + \inf_k\left[E||f - \widehat{f}_{k,m/2}||_V^2 + \eta_1 EL\{\widehat{V}(\widehat{\alpha}_{k,m/2}), V\}\right]\right),$$

where $||f - g||_V^2$ *denotes* $E\{f(x) - g(x)\}^T V^{-1}\{f(x) - g(x)\}$*.*

In the above risk bound, the covariance estimation risk is involved. If the estimators of the parameters in V_i are $m^{\frac{1}{2}}$-consistent (see Liang and Zeger (1986)), it does not affect the rate of convergence. For example, under the sufficient conditions given before the theorem, if the estimators of σ^2 of the best model converge in the usual parametric rate, then the combined estimator retains its rate-optimality. Also note that no

assumptions about the forms of the true f and the true covariance matrix are made (except the boundedness conditions). Theorems 1 and 2 point out that the combined estimators perform optimally in rate of convergence among all candidates even if all the candidate models are only approximations, which is more realistic in practice. The risk bounds cannot be improved in order of magnitude, although the multiplicative constant has the potential to be reduced for the squared error loss case.

5 Simulation results

In this section, we compare ARM and model selection methods through simulations in R with the *gls* function. In what follows, we consider linear models, which has two interpretations, either linear models with correlated errors (without the random effects) or linear mixed-effect models.

For 50 subjects, each with five measurements, we consider the marginal mean model $\beta_0 + X_{ij}^T\beta$. Assume $X_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ij5})^T \sim N(0, \Sigma_X)$ with Σ_X having the pp'-th element $0.5^{|p-p'|}$ for $p, p' = 1, 2, ..., 5$ (the case with independent predictors gives similar results, but will not be reported in this work). The error correlation within subjects is exchangeable.

The model selection approach that is to be compared with ARM involves two steps, the first for choosing an appropriate covariance model based on the full model for the mean and the second for choosing the mean model given the covariance model. Here we use AIC or BIC to select both the covariance model and the mean model.

For comparison, we report two types of estimation risks. The first one is based on the squared L_2 loss, which is computed by $\frac{1}{5m_{test}} \sum_{i=1}^{m_{test}} \sum_{j=1}^{5} \{f(X_{ij}) - \widehat{f}(X_{ij})\}^2$ at $m_{test} = 500$ new independently generated X_{ij} from the same distribution. The second is computed by $\frac{1}{m_{test}} \sum_{i=1}^{m_{test}} \{f(X_i) - \widehat{f}(X_i)\}^T V_i^{-1} \{f(X_i) - \widehat{f}(X_i)\}$. The permutation size (in step 5) for ARM is set to be 40. The results summarized in the following

tables are the simulated global estimation risks (based on 100 replications) with the corresponding standard errors given in the parentheses. There, risk reduction (denoted by RR in the tables) is defined as the risk reduction of ARM compared to the *best* of the two model selection methods under each of the two aforementioned loss functions.

5.1 Combining models with the true covariance structure

The true model is

$$Y_i = \beta_0 + X_i\beta + e_i, \tag{6}$$

where $X_i = (X_{i1}, X_{i2}, ..., X_{i5})^T$, $\beta_0 = 1$ and $\beta = (1, 0.5, 0.2, 0, 0)^T$, and the normal error vector e_i has a compound symmetry covariance matrix with entries $\rho\sigma^2$, $1 \leq i \neq j \leq 5$ and σ^2, $1 \leq i = j \leq 5$.

For model selection, we fix the covariance model to be exchangeable and select the mean model from all subset models by AIC/BIC. For ARM, we combine all the candidate mean models (all under the same covariance structure). Note that the true mean includes two small coefficients, which are difficult to identify when σ^2 is not too small.

Case 5.1.1 (small correlation) First consider a small correlation case with $\rho = 0.1$. It is clear from Table 1 that combining is superior over selection in risk when $\sigma^2 \geq 1$. The improvement of combining over selection is increasing in σ^2 in the studied range.

table 1 about here

Case 5.1.2 (moderate correlation) Here $\rho = 0.5$. From Table 2, although combining still beats selection when $\sigma^2 \geq 1$, the difference is not as substantial as the previous case.

table 2 about here

5.2 Combining models with multiple covariance structures

We consider three different covariance structures, independence, exchangeable correlation (or compound symmetry), and exponential spatial correlation. The true model is still the same but with $\beta = (0.4, 0.4, 0.2, 0.2, 0.1)^T$. The selection procedure first chooses a covariance structure among the three candidate structures and then chooses a mean structure based on the chosen covariance. In this case, unlike the case in Section 5.1, AIC has an advantage over BIC because the true model is the full model and hence AIC cannot overfit. We combine all the subset models with the covariance matrices, the total number of which is $31 \times 3 = 93$. The results are presented in graphs.

Case 5.2.1 (small correlation) Here $\rho = 0.1$. Again, combining beats selection when σ^2 is not small. The contrast is more evident than the previous cases.

figure 1 about here

Case 5.2.2 (moderate correlation) Here $\rho = 0.5$. Comparisons can be done in two aspects, in terms of the effect of correlation coefficient ρ or the effect of the number of candidate covariance models. When compared to Figure 1, combining still does a better job than selection when σ^2 is not small, but the advantage becomes somewhat smaller as the correlation is increased. When compared to Case 5.1.2 where only one covariance structure is included, the advantage of ARM becomes larger when σ^2 is bigger than 1.

figure 2 about here

In conclusion of our simulation, we have observed that model combining improves over model selection in terms of estimation accuracy when σ^2 is not small. However, the amount of improvement varies, with large improvement taking place when model selection is difficult. While our proposed combining method beats the better one of AIC and BIC when σ^2 is relatively large, it rarely loses to both of them.

To be fair, it should be pointed out that model combining does bring in difficulty in interpretation, and when estimation/prediction is not the dominating interest, a marginal gain in predictive accuracy may not be enough to abandon the selected model for a complicated convex weighting of the candidate models. However, when the instability measures indicate a very serious issue of model selection uncertainty, it is the reliability of the inference based on the selected model (rather than the loss of accuracy in prediction) that becomes the main concern for a sound statistical description of the messages in the data.

6 Data examples

6.1 Description of the data and candidate models

A CD4 count data set (http://biosun1.harvard.edu/ ∼fitzmaur/ala/), taken from a randomized and double-blinded study of AIDS patients (Henry *et al.*, 1998), is used. Patients were randomly assigned to four treatment groups. Measurements of CD4 counts were collected at baseline and at 8-week intervals for a total of 40 weeks. There are four covariates: *time* (in weeks), *group* (*treatment*), *age* and *gender*. The response variable is the log transformed CD4 counts, $\log(CD4+1)$, available on 1309 patients. Fitzmaurice *et al.* (2004, Chapter 8) provide a fairly complete analysis on $\log(CD4+1)$, which suggests that the mean response function has a change point at week 16. They used a piecewise linear spline with a knot at week 16 to model the effect of time. In order to demonstrate the utility of our proposed model selection uncertainty measures and model combination method in a more focused way, we consider only the observations after week 16 so that it is adequate to consider a linear term of week in the mean function. For random effects, again based on their results, we only consider a random subject effect, i.e., a random intercept. Four different covariance models, independence, exchangeable correlation, Gaussian spatial correlation and ex-

ponential spatial correlation are considered. A screening indicates that the latter two ($corr(\epsilon_{i,j}, \epsilon_{i,j'}) = \exp\{-\frac{(j-j')^2}{\rho}\}$ and $corr(\epsilon_{i,j}, \epsilon_{i,j'}) = \exp\{-\frac{|j-j'|}{\rho}\}$) are promising. Since the number of predictors is small, we did not screen on the mean models. As for the mean, we consider 25 subset models that arise from two specifications: 1) the predictor *time* must be included; 2) only the predictors themselves and the cross-product terms are considered. Thus overall we have 50 linear mixed models.

6.2 Alarmingly high model selection uncertainty and its different effects on estimation

We compute the bootstrap instability in selection (BIS) for the model selection strategy of Fitzmaurice *et al.* (2004) with AIC (minus twice the maximized likelihood plus twice the number of parameters) as the selection criterion.

The BIS value based on 100 bootstrap samples is 0.68! It is clear that an alarmingly high selection uncertainty exists in modeling the mean and the covariance. In such a case, it is plainly wrong to simply draw conclusions based solely on the selected model.

For the example, of interest are Q_1, the difference between the average effect of the first three treatments (zidovudine alternating monthly with 400mg didanosine, zidovudine plus 2.25mg of zalcitabine and zidovudine plus 400mg of didanosine) and that of the fourth one (zidovudine plus 400mg of didanosine plus 400mg of nevirapine) on the CD4 counts (this describes two-drug effect versus three-drug effect), and Q_2, the difference between the first and the third (zidovudine or didanosine versus zidovudine and didanosine). If the treatment effects are denoted by μ_1, μ_2, μ_3 and μ_4, then the quantities of interest are estimated by $\widehat{Q}_1 = \frac{\widehat{\mu}_1+\widehat{\mu}_2+\widehat{\mu}_3}{3} - \widehat{\mu}_4$ and $\widehat{Q}_2 = \widehat{\mu}_3 - \widehat{\mu}_1$, where $\widehat{\mu}_i$ are based on the selected model (note that due to randomization, μ_1,...,μ_4 are well defined here).

How does model selection uncertainty affect those two quantities? The two measures of instability in estimation are BIE($\widehat{Q}_1$) = 0.29, BIE($\widehat{Q}_2$) = 0.66, PIE($\widehat{Q}_1$) = 0.17, and

PIE($\widehat{Q}_2$) = 0.50. The two types of instability measure agree with each other: selection uncertainty has a mild effect on Q_1 but a large effect on Q_2. Note that the estimates for Q_1 and Q_2 (and their standard errors) from the originally selected model by AIC are -0.33 (0.066) and 0.24 (0.082), respectively. It seems from the standard error of Q_2 that the estimate of 0.24 is reasonably accurate (it would change by 27% in an average sense under normality of the estimator); however, according to BIE, Q_2 may easily change by 66%. The BIE$_s$ and PIE$_s$ values are 1.2 and 6.1 respectively. Therefore, the standard error of Q_2 is misleading and the estimate is not reliable. Consequently, the conclusion that the treatment with both zidovudine and didanosine is significantly more effective than alternating those two is not as confirmative as it appears to be. The model selection based estimates will be compared with the combined estimates by our method later from a prediction perspective.

In conclusion, 1) model selection uncertainty is a high for the CD4 count data; 2) the quantities of interest can be affected by model selection uncertainty to very different degrees.

For understanding the possible source of instability, we have $\mathbf{PIE}(\widehat{Q}_1) = (0.17, 0.05)^T$ and $\mathbf{PIE}(\widehat{Q}_2) = (0.31, 0.40)^T$. The instability in estimating Q_1 is really due to randomness of errors; while the instability in estimating Q_2 comes from both randomness of errors and that of random intercept.

6.3 Model combination reduces bias and instability in estimation

Recall that the estimates of Q_1 and Q_2 by AIC are -0.33 (0.066) and 0.24 (0.082) respectively. The estimates by our method ARM are -0.45 and 0.03 respectively. The ARM estimate of Q_1 falls into the corresponding 95% confidence interval obtained based on AIC, and both estimates indicate that the three-drug effect is larger than the two-drug effect. However the two estimates of Q_2 are very different: the ARM estimate

does not fall into the AIC confidence interval and it suggests that there is virtually no improvement by using zidovudine and didanosine together. This example clearly shows that model selection can give a much inflated confidence on an estimate and it illustrates the necessity of conducting model selection diagnostics. Besides reducing instability, we will see that ARM also outperforms model selection in prediction.

It is worth pointing out that combining the models reduces the instability in estimation. The PIE values of ARM are 0.05 and 0.35 for Q_1 and Q_2 respectively, a significant reduction from the PIE values of AIC of 0.17 and 0.50 respectively. Note that for estimating Q_1, the pure perturbation instability (without model selection) is also 0.05, thus ARM leads to a very stable estimate of Q_1. For Q_2, ARM does not reduce the instability to the level of the pure perturbation instability (0.07).

6.4 Comparing predictive performances

The comparison of the selection and combining procedures is done as follows.

1. Randomly permute the order of the subjects and then split the data into two parts, with the first part (m_t subjects) as a training/estimation set, and the second part ($m - m_t$ subjects) as the validation/testing set.

2. Based on the training set, a model is selected by AIC or BIC. The estimator based on ARM is also obtained. For the purpose of comparison, we also consider an "optimal" model (Opt) as follows: fit all the candidate models using the training set and then find the one with the best predictive performance on the validation set. So it has the smallest possible prediction error among the models.

3. Compute two types of prediction errors based on the validation set:

$$PE_1 = \frac{\sum_{i=m_t+1}^{m}\sum_{j=1}^{n_i}(Y_{i,j} - \widehat{Y}_{i,j})^2}{\sum_{i=m_t+1}^{m} n_i}, \quad PE_2 = \frac{\sum_{i=m_t+1}^{m}(Y_i - \widehat{Y}_i)^T \widehat{V}_i^{-1}(Y_i - \widehat{Y}_i)}{m - m_t}.$$

4. Repeat the above steps 500 times and obtain the average PE, APE, for each procedure. The error reduction of ARM relative to the better one of AIC and BIC presented in Table 3 is computed by

$$\frac{\{APE(AIC/BIC) - APE(Opt)\} - \{APE(ARM) - APE(Opt)\}}{APE(AIC/BIC) - APE(Opt)}, \tag{7}$$

where $APE(AIC/BIC) = \min\{APE(AIC),\ APE(BIC)\}$.

table 3 about here

Note that the above quantity measures the improvement of our method ARM over the better one of AIC and BIC relative to the best model. Such a measure provides the ability to differentiate the competing methods when the estimation problem is difficult due to a high noise level which is the case for the CD4 data set. The numbers in the parentheses in Table 3 are the standard deviations of PE over the splittings, divided by $\sqrt{500}$.

The results show that ARM outperforms AIC and BIC. The relative error reductions are 23% and 38% respectively. ARM beats both AIC and BIC 77% and 60% of the time among the 500 replications for PE_1 and PE_2 respectively.

7 Concluding remarks

As was already pointed out in the literature, there exist gaps in model selection theory and practice for longitudinal data analysis. While establishing model selection theories in this context is indeed in need, we believe model selection uncertainty presents another challenge. Without a proper assessment of the influence of model selection on the estimation of quantities of interest, interpretations based on a single final model, no mater how good it looks like on its own, are potentially misleading.

We proposed several ways to do model-selection diagnostics from an instability perspective, including BIS, BIE and PIE, which can guide on deciding whether selection

is trustworthy. A high value of BIS indicates that finding the true or best model is unrealistic. While some quantities of interest are sensitive to the model selection process, others are not. In the latter case, the uncertainty is not a serious concern. Our numerical investigations suggest that the model selection uncertainty can be very serious in longitudinal data analysis, and ARM significantly reduces risks in estimation.

In conclusion, model selection diagnostics should be done when model selection is involved and model combining should be applied only when it is necessary.

Appendix D: Proofs of the theorems

Proof of Theorem 1. Let $m_1 = m_0$ and $m_2 = m - m_0$, and

$$p^{m_2} = \Pi_{i=m_1+1}^{m} \frac{1}{(2\pi)^{n_i/2}|V_i|^{1/2}} \exp[-\frac{1}{2}\{y_i - f(x_i)\}^T V_i^{-1}\{y_i - f(x_i)\}],$$

be the joint density of y_i, $m_1 < i \leq m$ (conditional on x_i, $m_1 < i \leq m$), and

$$\widehat{q}^{m_2} = \sum_k \frac{1}{K} \Pi_{i=m_1+1}^{m} \frac{1}{(2\pi)^{n_i/2}|\widehat{V}_i(\widehat{\alpha}_{k,i-1})|^{1/2}} \exp[-\frac{1}{2}\{y_i - \widehat{f}_{k,i-1}(x_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k,i-1})^{-1}\{y_i - \widehat{f}_{k,i-1}(x_i)\}]$$

be an estimate of the joint density based on the candidate models. It follows from the definitions of $W_{k,i}$, p_i and $\widehat{g}_i$ that $\log \frac{p^{m_2}}{\widehat{q}^{m_2}} = \sum_{i=m_1+1}^{m} \log \frac{p_i}{\widehat{g}_i}$. Then

$$\sum_{i=m_1+1}^{m} E \log \frac{p_i}{\widehat{g}_i} = E \log \frac{p^{m_2}}{\widehat{q}^{m_2}}. \tag{8}$$

The left side of (9) just equals $\sum_{i=m_1+1}^{m} ED(p_i||\widehat{g}_i)$. For the right side of (9), since $\log(x)$ is an increasing function, for any model k_0 we have

$$E \log \frac{p^{m_2}}{\widehat{q}^{m_2}} \leq E \int p^{m_2} \log \frac{p^{m_2}}{\frac{1}{K}\widehat{q}_{k_0}^{m_2}} = \log K + E \int p^{m_2} \log \frac{p^{m_2}}{\widehat{q}_{k_0}^{m_2}},$$

where

$$\widehat{q}_{k_0}^{m_2} = \Pi_{i=m_1+1}^{m} \frac{1}{(2\pi)^{n_i/2}|\widehat{V}_i(\widehat{\alpha}_{k_0,i-1})|^{1/2}} \exp[-\frac{1}{2}\{y_i - \widehat{f}_{k_0,i-1}(x_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1}\{y_i - \widehat{f}_{k_0,i-1}(x_i)\}].$$

Observe that for each k, as in Barron (1987), $E\int p^{m_2}\log\frac{p^{m_2}}{\widehat{q}_k^{m_2}} = \sum_{i=m_1+1}^{m} ED(p_i||\widehat{q}_{k,i})$. Thus we have for any k_0, $\sum_{i=m_1+1}^{m} ED(p_i||\widehat{g}_i) \le \log K + \sum_{i=m_1+1}^{m} ED(p_i||\widehat{q}_{k_0,i})$. This completes the proof. □

The following lemma will be used for proving Theorem 2. It generalizes a one-dimensional result in Yang (2004).

Lemma 1. Let p and q be two probability densities on R^n with respect to a measure ν, with mean vector μ_p and μ_q, and variance matrix Σ_p and Σ_q respectively, then

$$\int(\sqrt{p}-\sqrt{q})^2 d\nu \ge \frac{(\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q)}{2\{n+tr(\Sigma_q\Sigma_p^{-1})+\frac{1}{2}(\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q)\}}.$$

Proof. Let a be any length n vector of constants. By matrix manipulations and applying Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&(\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q)\\
&=(\Sigma_p^{-1/2}\mu_p-\Sigma_p^{-1/2}\mu_q)^T(\Sigma_p^{-1/2}\mu_p-\Sigma_p^{-1/2}\mu_q)\\
&=\left\{\int(\Sigma_p^{-1/2}x-a)p(x)\nu(dx)-\int(\Sigma_p^{-1/2}x-a)q(x)\nu(dx)\right\}^T\\
&\quad\left\{\int(\Sigma_p^{-1/2}x-a)p(x)\nu(dx)-\int(\Sigma_p^{-1/2}x-a)q(x)\nu(dx)\right\}\\
&=\left\{\int(\Sigma_p^{-1/2}x-a)(p-q)d\nu\right\}^T\left\{\int(\Sigma_p^{-1/2}x-a)(p-q)d\nu\right\}\\
&\le\left\{\int(|\Sigma_p^{-1/2}x-a||\sqrt{p}+\sqrt{q}|)|\sqrt{p}-\sqrt{q}|d\nu\right\}^T\left\{\int(|\Sigma_p^{-1/2}x-a||\sqrt{p}+\sqrt{q}|)|\sqrt{p}-\sqrt{q}|d\nu\right\}\\
&\le 2\int(\sqrt{p}-\sqrt{q})^2d\nu\int(\Sigma_p^{-1/2}x-a)^T(\Sigma_p^{-1/2}x-a)(p+q)d\nu\\
&=2\int(\sqrt{p}-\sqrt{q})^2d\nu\left\{\mu_p^T\Sigma_p^{-1}\mu_p+\mathrm{tr}(\Sigma_p\Sigma_p^{-1})-a^T\Sigma_p^{-1/2}\mu_p-\mu_p^T\Sigma_p^{-1/2}a+a^Ta+\right.\\
&\qquad\left.\mu_q^T\Sigma_p^{-1}\mu_q+\mathrm{tr}(\Sigma_q\Sigma_p^{-1})-a^T\Sigma_p^{-1/2}\mu_q-\mu_q^T\Sigma_p^{-1/2}a+a^Ta\right\}.
\end{aligned}$$

Now taking $a=\Sigma_p^{-1/2}\frac{\mu_p+\mu_q}{2}$, with some simplifications we have

$$(\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q)\le 2\int(\sqrt{p}-\sqrt{q})^2d\nu\left\{n+\mathrm{tr}(\Sigma_q\Sigma_p^{-1})+\frac{1}{2}(\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q)\right\}.$$

The conclusion follows. □

Proof of Theorem 2. Following the earlier notations, for each fixed k_0, we have,

$$\begin{aligned}
\log \frac{p^{m_2}}{\widehat{q}^{m_2}} &\le \log K + \\
&\log \frac{\mathit{\Pi}_{i=m_1+1}^{m} \frac{1}{(2\pi)^{n_i/2}|V_i|^{1/2}} \exp[-\frac{1}{2}\{y_i - f(x_i)\}^T V_i^{-1}\{y_i - f(x_i)\}]}{\mathit{\Pi}_{i=m_1+1}^{m} \frac{1}{(2\pi)^{n_i/2}|\widehat{V}_i(\widehat{\alpha}_{k_0,i-1})|^{1/2}} \exp[-\frac{1}{2}\{y_i - \widehat{f}_{k_0,i-1}(x_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1}\{y_i - \widehat{f}_{k_0,i-1}(x_i)\}]} \\
&= \log K + \frac{1}{2}\sum_{i=m_1+1}^{m} [\{y_i - \widehat{f}_{k_0,i-1}(x_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1}\{y_i - \widehat{f}_{k_0,i-1}(x_i)\} \\
&\quad - \{y_i - f(x_i)\}^T V_i^{-1}\{y_i - f(x_i)\} + \log \frac{|\widehat{V}_i(\widehat{\alpha}_{k_0,i-1})|}{|V_i|}]. \qquad (9)
\end{aligned}$$

Now under Conditions 1 and 2, taking expectation conditioned on the first part of data, denoted by E_{m_1}:

$$\begin{aligned}
&E_{m_1}[\{Y_i - \widehat{f}_{k_0,i-1}(X_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1}\{Y_i - \widehat{f}_{k_0,i-1}(X_i)\} \\
&-\{Y_i - f(X_i)\}^T V_i^{-1}\{Y_i - f(X_i)\} + \log \frac{|\widehat{V}_i(\widehat{\alpha}_{k_0,i-1})|}{|V_i|}] \\
&= E_{m_1}[\{f(X_i) - \widehat{f}_{k_0,i-1}(X_i)\}^T \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1}\{f(X_i) - \widehat{f}_{k_0,i-1}(X_i)\} \\
&+e_i^T \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1} e_i - e_i^T V_i^{-1} e_i + \log \frac{|\widehat{V}_i(\widehat{\alpha}_{k_0,i-1})|}{|V_i|}] \\
&\le E_{m_1}[\frac{1}{\eta_1}\{f(X_i) - \widehat{f}_{k_0,i-1}(X_i)\}^T V_i^{-1}\{f(X_i) - \widehat{f}_{k_0,i-1}(X_i)\} \\
&+ \mathrm{tr}\{V_i \widehat{V}_i(\widehat{\alpha}_{k_0,i-1})^{-1} - I_{n_i}\} + \log \frac{|\widehat{V}_i(\widehat{\alpha}_{k_0,i-1})|}{|V_i|}]. \qquad (10)
\end{aligned}$$

Conditioned on the first part of data and x_i as denoted by E'_{m_1}, we have

$$E'_{m_1} \log \frac{p_i}{\widehat{g}_i} = \int p_i \log \frac{p_i}{\widehat{g}_i} dy_i \ge \int (\sqrt{p_i} - \sqrt{\widehat{g}_i})^2 dy_i \ge \frac{\frac{1}{n_i}\{f(x_i) - \tilde{\mu}_i\}^T V_i^{-1}\{f(x_i) - \tilde{\mu}_i\}}{2(2 + \xi + \frac{5\tau}{2n_i})},$$

where for the last inequality we apply Lemma 1 with $\tilde{\mu}_i = \sum_k W_{k,i} \widehat{f}_{k,i-1}(x_i)$. Combining this with (10) and (11), we obtain

$$\begin{aligned}
&\sum_{i=m_1+1}^{m} \frac{1}{n^*} E\{f(X_i) - \sum_k W_{k,i}\widehat{f}_{k,i-1}(X_i)\}^T V_i^{-1}\{f(X_i) - \sum_k W_{k,i}\widehat{f}_{k,i-1}(X_i)\} \\
&\le \frac{1}{\eta_1}(2 + \xi + \frac{5\tau}{2n^*}) \left(2\eta_1 \log K + \sum_{i=m_1+1}^{m} \inf_k \Big[E\{f(X_i) - \widehat{f}_{k,i-1}(X_i)\}^T V_i^{-1}\{f(X_i) - \widehat{f}_{k,i-1}(X_i)\} \right. \\
&\quad \left. + \eta_1 E L\{\widehat{V}_i(\widehat{\alpha}_{k,i-1}), V_i\}\Big] \right).
\end{aligned}$$

This completes the proof of the theorem. □

Acknowledgment

The work of the second author is partially supported by NSF grant DMS-0706850.

References

1. Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350-2383.
2. Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, **53**, 603-618.
3. Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, **33** (2), 261-304.
4. Cantoni, E., Field, C., Flemming, J. M., and Ronchetti, E. (2007). Longitudinal variable selection by cross-validation in the case of many covariates. *Statistics in Medicine*, **26**, 919-930.
5. Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society*, **Series A, 158**, 419-466.
6. Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press.
7. Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, **Series B, 57**, 45-70.
8. Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
9. Fan, J., and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710-723.
10. Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. New York: Wiley. http://biosun1.harvard.edu/ fitzmaur/ala/.
11. Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, New York.
12. Henry, K., Erice, A., Tierney, C., Balfour, H. H. Jr, Fischl, M. A., Kmack, A., Liou, S. H., Kenton, A., Hirsch, M. S., Phair, J., Martinez, A., and Kahn, J. O. for the AIDS Clinical Trial Group 193A Study Team (1998). A randomized, controlled, double-blinded study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug,two-drug,and alternative drug)for the treatment of advanced AIDS. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **19**, 339-349.
13. Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, **14**, 382-417.
14. Hjort, N.L., and Claeskens, G. (2003). Frequentist model average estimators. *JASA*, **98**, 879-899.
15. Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, **93**, 85-98.
16. Juditsky, A., and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *The Annals of Statistics*, **28**, 681-712.
17. Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.
18. Lin, D. Y., and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*, **96**, 103-126.
19. Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.
20. Ruppert, D., Wand, M. P., and Raymond, C. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.
21. Shen, X., and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, **97**, 210-221.
22. Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture notes in Artificial Intelligence*, **v. 2777**, 303-313. Springer-Verlag, Heidelberg.

23. Wang, L., and Qu, A. (2009) Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society,* **Series B: 71**, 177-190.
24. Wu, W., and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**, 831-844.
25. Yafune, A., Funatogawa, T., and Ishiguro, M. (2005). Extended information criterion (EIC) approach for linear mixed effects models under restricted maximum likelihood (REML) estimation. *Statistics in Medicine*, **24**, 3417-3429.
26. Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, **96**, 574-588.
27. Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, **13**, 783-809.
28. Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, **20**, 176-222.
29. Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120-131.
30. Yuan, Z., and Yang, Y. (2005). Combining linear regression models: when and how? *Journal of the American Statistical Association*, **100**, 1202-1214.

	σ	0.3	0.7	1	1.5	2
AIC	l1	.0021	.0112	.0247	.0572	.0993
		(.00015)	(.00067)	(.00156)	(.00376)	(.00548)
	l2	.099	.105	.113	.120	.118
		(.0068)	(.0062)	(.0074)	(.0079)	(0.0064)
BIC	l1	.0018	.0110	.0271	.0633	.0973
		(.00013)	(.00086)	(.00173)	(.00401)	(.00583)
	l2	.083	.103	.129	.134	.115
		(.0057)	(.0084)	(.0084)	(.0085)	(.0070)
ARM	l1	.0020	.0119	.0220	.0513	.0843
		(.00013)	(.00062)	(.00122)	(.00364)	(.00463)
	l2	.092	.112	.102	.107	.098
		(.0056)	(.0056)	(.0055)	(.0076)	(.0051)
RR	l1	-11%	-8%	8%	10%	13%
	l2	-11%	-9%	10%	11%	15%

Table 1 Comparing ARM with selection by AIC/BIC where the true model is specified by (6) with $\rho = 0.1$. RR refers to risk reduction of ARM compared to the best of AIC and BIC. l1 and l2 refer to the estimation risks based on the squared L_2 loss without or with normalization by the covariance matrix respectively.

	σ	0.3	0.7	1	1.5	2
AIC	l1	.0025	.0123	.0255	.0574	.1055
		(.00019)	(.00098)	(.00202)	(.00463)	(.00839)
	l2	.111	.110	.118	.118	.124
		(.0070)	(.0070)	(.0084)	(.0073)	(.0081)
BIC	l1	.0023	.0114	.0252	.0640	.1178
		(.00019)	(.00102)	(.00212)	(.00454)	(.00903)
	l2	.093	.095	.115	.142	.149
		(.0064)	(.0091)	(.0098)	(.0064)	(.0105)
ARM	l1	.0023	.0123	.0249	.0557	.1016
		(.00019)	(.00097)	(.00195)	(.00456)	(.00858)
	l2	.098	.110	.112	.111	.115
		(.0058)	(.0064)	(.0061)	(.0067)	(.0079)
RR	l1	0%	-8%	1%	3%	4%
	l2	-5%	-15%	3%	6%	7%

Table 2 Comparing ARM with selection by AIC/BIC where the true model is specified by (6) with $\rho = 0.5$. RR refers to risk reduction of ARM compared to the best of AIC and BIC. l1 and l2 refer to the estimation risks based on the squared L_2 loss without or with normalization by the covariance matrix respectively.

	APE_1	APE_2
AIC	0.946(0.0025)	2.414(0.0122)
BIC	0.945(0.0025)	2.411(0.0122)
ARM	0.940(0.0025)	2.408(0.0123)
Optimal	0.932(0.0026)	2.398(0.0122)
RER	38%	23%

Table 3 Comparing of ARM with AIC and BIC on real data. Relative Error Reduction (denoted by RER) is computed by (7).

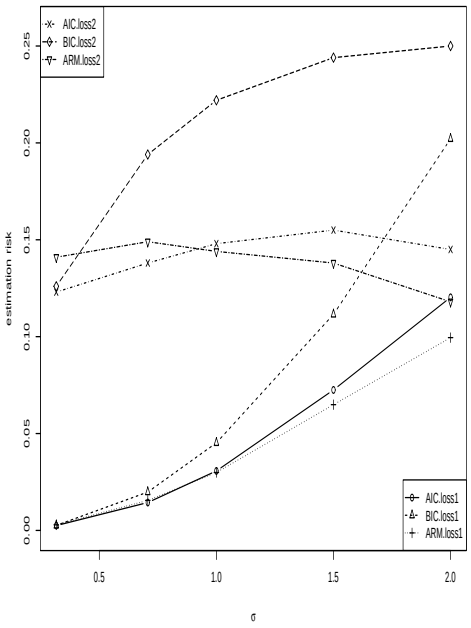

Fig. 1 Comparing ARM with selection by AIC/BIC where the true model is specified by (6) with $\rho = 0.1$ RR refers to risk reduction of ARM compared to the best of AIC and BIC. l1 and l2 refer to the estimation risks based on the squared L_2 loss without or with normalization by the covariance matrix respectively.

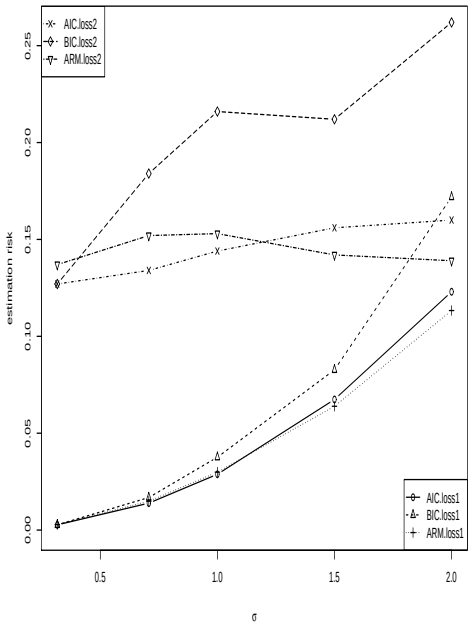

Fig. 2 Comparing ARM with selection by AIC/BIC where the true model is specified by (6) with $\rho = 0.5$. RR refers to risk reduction of ARM compared to the best of AIC and BIC. l1 and l2 refer to the estimation risks based on the squared L_2 loss without or with normalization by the covariance matrix respectively.