Published in final edited form as:
Comput Stat Data Anal. 2012 March 1; 56(3): 574–586. doi:10.1016/j.csda.2011.09.001.

Generalized Degrees of Freedom and Adaptive Model Selection in Linear Mixed-Effects Models

Bo Zhang[1], **Xiaotong Shen**[2], and **Sunni L. Mumford**[1]

[1]Division of Epidemiology, Statistics, and Prevention Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD 20892, U.S.A.

[2]School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Abstract

Linear mixed-effects models involve fixed effects, random effects and covariance structure, which require model selection to simplify a model and to enhance its interpretability and predictability. In this article, we develop, in the context of linear mixed-effects models, the generalized degrees of freedom and an adaptive model selection procedure defined by a data-driven model complexity penalty. Numerically, the procedure performs well against its competitors not only in selecting fixed effects but in selecting random effects and covariance structure as well. Theoretically, asymptotic optimality of the proposed methodology is established over a class of information criteria. The proposed methodology is applied to the BioCycle study, to determine predictors of hormone levels among premenopausal women and to assess variation in hormone levels both between and within women across the menstrual cycle.

Keywords

Adaptive penalty; linear mixed-effects models; loss estimation; generalized degrees of freedom

1 Introduction

In clinical or epidemiologic studies, linear mixed-effects models (LMMs) (Laird and Ware 1982; Longford 1993) are commonly used in analyzing clustered data (repeated measures data, longitudinal data) with continuous outcomes and multiple covariates. LMMs are attractive because they can effectively model the dependence structure that arises from repeated measures for the same cluster by appropriately using random effects and covariance structure. Despite a large body of literature on LMMs, the issue of selecting their fixed effects, random effects or covariance structure has not received much attention. Incorrect inclusion of fixed or random effects, or incorrect specification of covariance structure can result in biased results and false interpretation. Therefore, accurate model assessment and precise model selection procedures are essential for improving the performance of LMMs. In this article, we focus on model selection in LMMs and develop a competitive methodology for selecting LMMs.

Specialized model selection methods for LMMs were rarely proposed in the literature, although information criteria have been extensively applied in data analysis where LMMs

are used. Most information criteria select the optimal model $\hat{M}$ among candidate models $\{M_\gamma, \gamma \in \Gamma\}$ by minimizing a model selection criterion of the form

$$-\ell_{M_\gamma} + \lambda(n, k_{M_\gamma}) \cdot k_{M_\gamma}, \tag{1}$$

where n is the total number of observations, k_{M_γ} is the number of independent parameters in a candidate model M_γ, and ℓ_{M_γ} is the log-likelihood given by M_γ. Akaike's information criterion (AIC) in Akaike (1973) uses the expected Kullback-Leibler information with $\lambda(n, k_{M_\gamma}) = 1$; a bias-corrected version of AIC, called AICc (Hurvich and Tsai 1989), with $\lambda(n, k_{M_\gamma}) = nk_{M_\gamma}/(n - k_{M_\gamma} - 1)$ estimates the expected Kullback-Leibler information directly in a regression model where a second order bias adjustment is made; Bayesian information criterion (BIC) in Schwarz (1978) uses an asymptotic Bayes factor and advocates $\lambda(n, k_{M_\gamma}) = \log(n)/2$; risk inflation criterion (RIC) in Foster and George (1994) is based on the minimax principle, and adjusts the penalization parameter to be $\lambda(n, k_{M_\gamma}) = \log(p)$, where p is the number of available covariates; covariance inflation criterion (CIC) in Tibshirani and Knight (1999) with $\lambda = 2\sum_{l=1}^{k_{M_\gamma}} \log(n/l)/k_{M_\gamma}$ adjusts the prediction error by the average covariance of the predictions and responses when the prediction rule is applied to permute the data set; and many others are available. The total number of unknown parameters k_{M_γ} in the information criteria characterize the model complexity that the information criteria intend to penalize on. Increasing number of unknown parameter in either random effects or the variance-covariance structure of random effects or within-cluster errors in LMM indeed increases the model complexity. Therefore, as suggested by Pinheiro and Bates (2000), Diggle *et al.* (2002), and Wolfinger (1997), the total number of unknown parameters k_{M_γ} used to compute information criteria includes not only the parameters introduced by fixed effects but also the ones introduced by random effects and variance-covariance structure. However, in (1), the penalization parameter $\lambda(n, k_{M_\gamma})$ penalizes an increase in the size of a model only through a *fixed* penalization parameter, in the sense that it is pre-determined by n and k_{M_γ}, and therefore it is not adaptive to various model structures. The model selection procedures with form (1) are hereby referred as *nonadaptive* selection procedures. The nonadaptive model selection procedures with a large penalty often yield an optimal model whose size is small, and the nonadaptive procedures with a small penalty often yield an optimal model whose size is large. Consequently, a large penalty is likely to perform well when the true model has a parsimonious representation, and is likely to perform poorly otherwise. This feature of nonadaptive model selection procedures results in large selection bias in LMMs. Shen and Ye (2002) and Shen, Huang, and Ye (2004) confirmed the disadvantages of those nonadaptive procedures in linear regression, logistic regression and Poisson regression. They showed that, with the inflexibility of the penalization parameter, information criteria ignore the uncertainty of data and fail to adjust the penalization parameter for better performance. The need is compelling for a *data-adaptive* model selection procedure that can reduce the selection bias and essentially performs well over a variety of situations.

In this article, we derive the generalized degrees of freedom (GDF) (Ye 1998; Efron 2004) in LMMs, and discuss how to use data perturbation (Shen, Huang, and Ye 2004; Shen and Huang 2006) to estimate the GDF. Through the GDF, we extend the adaptive model selection procedure proposed by Shen and Ye (2002) and Shen, Huang, and Ye (2004) to the context of LMMs. In simulations, we evaluate the finite sample performance of the proposed methodology in selecting fixed effects, random effects and covariance structure of LMMs. We establish the large-sample asymptotic optimality of the proposed procedure. The asymptotic properties are in agreement with our numerical examples that the proposed methodology approximates the best performance over a class of information criteria with

form (1). Finally, we apply the proposed adaptive model selection procedure to the BioCycle study, to determine factors that influence hormone levels of premenopausal women and to assess variation in hormone levels both between and within women across the menstrual cycle.

The rest of the article is organized as follows. Section 2 presents the GDF and the adaptive model selection for LMMs. Section 3 discusses the data perturbation estimation for the GDF and establishes the asymptotic optimality of adaptive model selection. In Section 4, numerical studies with small sample simulation datasets are performed to demonstrate advantages of the proposed method over information criteria. In Section 5, we demonstrate the adaptive model selection by applying it to hormone levels data in the BioCycle study. The last section is devoted to a discussion and technical proofs is in appendix.

2 Generalized Degrees of Freedom and Adaptive Model Selection

2.1 Generalized degrees of freedom

Suppose that data are collected from m independent clusters (or subjects in longitudinal data) with response variable Y_{ij}, covariates $X_{ij,1},\cdots, X_{ij,p}$ that are associated with fixed effects, and covariates $Z_{ij,1}\cdots, Z_{ij,q}$ that are associated with random effects b_i, where $i = 1, 2,\cdots, m$ indicates clusters and $j = 1, 2,\cdots, n_i$ indicates observations within the ith cluster. LMMs specify the response vector Y_{ij} as

$$Y_{ij}=X_{ij}^{'}\beta+Z_{ij}^{'}b_i+\epsilon_{ij}, \tag{2}$$

where $X_{ij} = (X_{ij,1},\cdots, X_{ij,p})'$ is a fixed-effects covariate vector, β is a fixed-effects coefficient vector, $Z_{ij} = (Z_{ij,1},\cdots, Z_{ij,q})'$ is a random-effects covariate vector, and $\epsilon_i = (\epsilon_{i1},\cdots, \epsilon_{in_i})'$ is a within-cluster error vector that follows $N(0, \sigma^2\Lambda_i)$. The random effects b_i are independent and identically distributed and follow $N(0,\Psi)$. The within-cluster errors ϵ_i are assumed to be independent for different i. The random effects b_i and the within-cluster errors ϵ_i are assumed to be independent. LMMs assume the Gaussian continuous response to be a linear function of covariates with regression coefficients that vary over individuals, which reflects natural heterogeneity due to unmeasured factors. They allow flexible correlation structures by assuming appropriate random-effects covariates Z_{ij} and covariance matrices Ψ and Λ_i's.

In (2), the Kullback-Leibler (KL) loss can be used to measure the accuracy of maximum likelihood method. The KL loss measures the deviation of the estimated likelihood from the true likelihood as if the truth were known. Let $\psi = \text{vech}(\Psi)$ be a vector that contains the distinct components of Ψ, let $\varphi = \text{vech}(\Lambda_1,\cdots,\Lambda_m)$ be a vector that contains the distinct components of $\Lambda_1,\cdots, \Lambda_m$, and let $\xi = (\beta',\psi', \sigma, \varphi')'$ be the vector that contains all the parameters in (2). Let $Y_i = (Y_{i1},\cdots, Y_{in_i})'$, $X_i = (X_{i1},\cdots, X_{in_i})'$ and $Z_i = (Z_{i1},\cdots, Z_{in_i})'$. The performance of estimator $\hat{\xi}$ in the ith cluster can be evaluated by its closeness to ξ in terms of the clusterwise KL loss of ξ versus $\hat{\xi}$: $\int p(y_i|\xi)[\log p(y_i|\xi) - \log p(y_i|\hat{\xi}))]dy_i$, where $p(y_i|\xi)$ is the likelihood function of the observations in the ith cluster. This yields the total KL loss for all independent clusters:

$$\mathscr{L}(\xi,\widehat{\xi})=\sum_{i=1}^{m}\int p(y_i|\xi)[\log p(y_i|\xi) - \log p(y_i|\widehat{\xi}))]dy_i, \tag{3}$$

which, after dropping the terms that are only related to ξ, reduces to

$$\mathscr{L}(\xi,\widehat{\xi})=\frac{1}{2}\sum_{i=1}^{m}\left[(\widehat{\mu}_i-\mu_i)'\widehat{\sum}_i^{-1}(\widehat{\mu}_i-\mu_i)+\log\left|\widehat{\sum}_i\right|+\mathrm{tr}(\sum_i\widehat{\sum}_i^{-1})\right], \tag{4}$$

where $\mu_i = X_i\beta$ and $\sum_i = Z_i\Psi Z'_i + \sigma^2\Lambda_i$ are the clusterwise mean and covariance matrix, respectively; $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are the corresponding estimates with ξ replaced by $\hat{\xi}$. The KL loss $\mathscr{L}(\xi, \hat{\xi})$ compares different LMM estimations in virtue of the true parameter value ξ. If ξ were known, then we could select the optimal model by minimizing (4) with respect to candidate models.

Motivated by the information criteria (1), we now consider a class of KL loss estimators of the form

$$-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi})+\hbar. \tag{5}$$

Members of this class penalize an increase in the size of a model used in estimation, with $\hbar$ controlling the degree of penalization. Clearly, different choices of $\hbar$ yield different model selection criteria; for instance, when $\hbar$ is the number of parameters, (5) becomes AIC.

Theorem 1—(optimal KL loss estimation). *The optimal $\hbar$ that minimizes* $E[\mathscr{L}(\xi,\widehat{\xi})-(-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi})+\hbar)]^2$, *the expected ℓ_2 distance between the KL loss (4) and the class of loss estimators (5), is*

$$\mathcal{G}=\frac{1}{2}\sum_{i=1}^{m}E\left[2(Y_i-\mu_i)'\widehat{\sum}_i^{-1}\widehat{\mu}_i-Y_i^T\widehat{\sum}_i^{-1}Y_i+\mu_i^T\widehat{\sum}_i^{-1}\mu_i+\mathrm{tr}(\sum_i\widehat{\sum}_i^{-1})\right]. \tag{6}$$

The optimal $\hbar$, the $\mathcal{G}$ in (6), which measures the degrees of freedom cost in model selection or statistical uncertainty of model selection, is thereby defined as the *generalized degrees of freedom* or the GDF of LMMs. When LMMs (2) degenerate into linear models by discarding random effects, the proposed GDF (6) becomes the GDF discussed in Ye (1998), which is a generalization of the degrees of freedom of fit in linear models (Weisberg 2005). In the rest of the article, we use $\mathcal{G}$ to denote the GDF, and denote as $\mathcal{G}(M)$ the GDF of a specified model M. Moreover, by Theorem 1, the performance of M can be assessed through its optimal KL loss "estimator"

$$-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_M)+\mathcal{G}(M), \tag{7}$$

and, for any M,

$$E[\mathscr{L}(\xi,\widehat{\xi}_M)]=E\left[-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_M)+\mathcal{G}(M)\right]. \tag{8}$$

The optimal KL loss estimator (7) measures the divergence between the true likelihood and the estimated likelihood by M. It can be used to assess and compare various models. In Section 2.2, (7) will be used to develop our adaptive model selection procedure in LMMs.

The optimal KL loss estimator (7) would be an unbiased estimator of $\mathcal{K}(\xi, \hat{\xi}_M)$ if it were independent of the true parameters. However, it usually depends on unknown parameters through $\mathcal{G}(M)$, and therefore needs to be estimated by data. We will develop in Section 3 the data perturbation estimate of $\mathcal{G}(M)$, denoted by $\hat{\mathcal{G}}(M)$, in order to fully realize adaptive model selection. Prior to that, we will describe the procedure of adaptive model selection in Section 2.2 assuming $\hat{\mathcal{G}}(M)$ is available.

2.2 Adaptive model selection

Now consider a class of model selection criteria in the form of

$$-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}M_\gamma)+\lambda\cdot k_{M_\gamma},\ \lambda\in(0,\infty), \tag{9}$$

for selecting the best model from a class of candidate models $\{M_\gamma, \gamma \in \Gamma\}$. To achieve the goal of adaptive selection, we choose the optimal λ from data by selecting the optimal model selection procedure from a class of information criteria (9) indexed by $\lambda \in (0, \infty)$. First, for each fixed $\lambda \in (0, \infty)$, one model, denoted by $\hat{M}(\lambda)$, is selected from candidate models such that it minimizes (9). Let the parameter estimates for $\hat{M}(\lambda)$ be $\hat{\xi}_{\hat{M}(\lambda)}$, and let the estimated GDF for $\hat{M}(\lambda)$ be $\hat{\mathcal{G}}(\hat{M}(\lambda))$. Second, the optimal λ, denoted by $\hat{\lambda}$, is obtained such that it minimizes the estimated loss of $\hat{M}(\lambda)$

$$-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_{\widehat{M}(\lambda)})+\widehat{\mathcal{G}}(\widehat{M}(\lambda)) \tag{10}$$

with respect to $\lambda \in (0, \infty)$. Finally, inserting $\hat{\lambda}$ back into (9) yields the adaptive model selection procedure: the adaptive selection procedure chooses the optimal model $\hat{M}(\hat{\lambda})$ by minimizing the *adaptive model selection* criterion

$$-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_{M_\gamma})+\widehat{\lambda}\cdot k_{M_\gamma} \tag{11}$$

over the candidate models $\{M_\gamma, \gamma \in \Gamma\}$. In (11), the $\hat{\lambda}$ is data-dependent as well as our selection procedure. The adaptive penalty $\hat{\lambda}$ estimates the ideal optimal penalization parameter over the class (9); its value varies depending on the data and the size of the true model. Therefore, it permits an approximation to the best performance of the class of model selection criteria (9).

3 Estimation of Generalized Degrees of Freedom

3.1 Estimation through data perturbation

This section estimates the GDF for LMMs through the data perturbation technique (Shen, Huang and Ye 2004; Shen and Huang 2006). Data perturbation assesses sensitivity of the estimated parameter through the pseudo response vector

$$Y_i^*=(Y_{i1}^*, Y_{i2}^*, \cdots, Y_{in_i}^*)=Y_i+\tau(\tilde{Y}_i - Y_i),\ i=1, \cdots, m, \tag{12}$$

which is generated from the original response vector Y_i and a perturbed vector $\tilde{Y}_i$ with perturbation size $\tau \in (0,1]$. To generate $\tilde{Y}_i$, data perturbation samples $\tilde{Y}_i$ are taken from the distribution of Y_i with the unknown distribution mean replaced by Y_i; that is, if we denote as $p_{Yi}(\cdot|\mu)$ the distribution of Y_i with distribution mean μ, $\tilde{Y}_i$ is sampled from $p_{Yi}(\cdot|Y_i)$, for $i = 1,2\cdots, m$. In the logistic models and the Poisson models, data perturbation can be implemented directly because of the absence of dispersion parameters (Shen, Huang and Ye 2004). But in LMMs, the distribution of Y_i depends on unknown dispersion parameters ψ, σ, and φ, besides μ. Thus, more precisely, we may denote the distribution of Y_i as $p_{Yi}(\cdot|\mu, \psi, \sigma, \varphi)$. To sample $\tilde{Y}_i$ in LMMs, we suggest to use the most complex model among all candidate models to obtain estimates $\tilde{\psi}$, $\tilde{\sigma}$, and $\tilde{\varphi}$, and then sample $\tilde{Y}_i$ from $p_{Yi}(\cdot|Y_i, \tilde{\psi}, \tilde{\sigma}, \tilde{\varphi})$.

To estimate the GDF, we can rewrite the GDF in (6) to be the summation of the difference of two covariance penalties (Efron 2004):

$\mathcal{G}=\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}[\mathrm{cov}(\widehat{\sigma}_{ijk}(Y)\mu_{ij}(Y), Y_{ik}) - \mathrm{cov}(\widehat{\sigma}_{ijk}(Y), Y_{ij}Y_{ik})/2]$, where $Y=(Y_1,\cdots, Y_m)'$ denotes the response vector, $\hat{\sigma}_{ijk}(Y)$ is the jkth element of $\widehat{\sum_i^{-1}}(Y)$, and $\hat{\mu}_{ij}(Y)$ is the jth element of $\hat{\mu}_i(Y)$. The response vector Y in the parenthesis indicates that the estimates depend on the response vector Y. With perturbed Y_i^* with perturbation size τ, let E^*, var*, and cov* denote the conditional mean, variance, and covariance, respectively, given Y_i. For any combination of i j, and k, note that the first covariance penalty term $\mathrm{cov}(\hat{\sigma}_{ijk}(Y)\mu_{ij}(Y),Y_{ik})$ equals $\tau^{-2}(E^*\mathrm{var}^*Y_{ik}^*)E\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(Y)\widehat{\mu}_{ij}(Y)$, which can be approximated by $\tau^{-2}E^*\frac{\partial}{\partial Y_{ik}^*}\widehat{\sigma}_{ijk}(Y^*)\widehat{\mu}_{ij}(Y^*)\mathrm{var}^*(Y_{ik}^*)=\tau^{-2}\mathrm{cov}^*(\widehat{\sigma}_{ijk}(Y^*)\widehat{\mu}_{ij}(Y^*), Y_{ik}^*)$; whereas the second covariance penalty $\mathrm{cov}(\hat{\sigma}_{ijk}(Y), Y_{ij}Y_{ik})/2$ equals $E^*\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)E\frac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk}(Y)\mathrm{var}(Y_{ij}Y_{ik})/2\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)$, which can be approximated by $(\widehat{\pi}_{ijk}^2/2\mathrm{var}^*(Y_{ij}^*Y_{ik}^*))E^*\frac{\partial}{\partial Y_{ij}^*Y_{ik}^*}\widehat{\sigma}_{ijk}(Y^*)\mathrm{var}(Y_{ij}^*Y_{ik}^*)=(\widehat{\pi}_{ijk}^2/2\mathrm{var}^*(Y_{ij}^*Y_{ik}^*))\mathrm{cov}^*(\widehat{\sigma}_{ijk}(Y^*), Y_{ij}^*Y_{ik}^*)$, where $\widehat{\pi}_{ijk}^2=\widehat{\mathrm{var}}(Y_{ij}Y_{ik})$ is the estimated variance of $Y_{ij}Y_{ik}$. For our implementation, we use a Monte Carlo numerical approximation. We sample $Y^{*d}=(Y_1^{*d}, Y_2^{*d}, \cdots, Y_m^{*d})$, $d = 1, 2, \cdots, D$, independently from the distribution of $Y^*=(Y_1^*, Y_2^*, \cdots, Y_m^*)$ as described earlier. Note that Y_i^{*d} follows the conditional distribution of Y_i^* given Y_i, $i = 1, 2,\cdots, m$ and $d = 1, 2,\cdots, D$. Then the GDF is approximated by

$$\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}\left[\tau^{-1}D^{-1}\left[\widehat{\sigma}_{ijk}(Y^{*d})\widehat{\mu}_{ij}(Y^{*d}) - D^{-1}\sum_{d=1}^{D}\widehat{\sigma}_{ijk}(Y^{*d})\widehat{\mu}_{ij}(Y^{*d})\right]\left[Y_{ik}^{*d} - D^{-1}\sum_{d=1}^{D}Y_{ik}^{*d}\right]+\widehat{\pi}_{ijk}^2D^{-1}.\left[\widehat{\sigma}_{ijk}(Y^{*d}) - D^{-1}\sum_{d=1}^{D}\widehat{\sigma}_{ijk}(Y^{*d})\right]\left[Y_{ij}^{*d}Y_{ik}^{*d} - D\right.\right.$$

where D is chosen to be sufficiently large to ensure an adequate Monte Carlo approximation. By the law of large numbers, the proposed Monte Carlo approximation of the GDF via data perturbation converges to the true GDF as $D \to 0$. However, both our simulation studies and Shen, Huang and Ye (2004) found that the Monte Carlo approximation of $\mathcal{G}(M)$ is sufficiently accurate if we choose D to be greater or equal to the number of observations.

Therefore, we recommended that D be at least $\sum_{i=1}^{m} n_i$ for the model selection problems that we consider in LMMs. We will choose D to be $\sum_{i=1}^{m} n_i$ in our simulations and data analysis.

In data perturbation (DP) technique, the parameter τ with $\tau \in (0,1]$ is called *perturbation size*. In the literature, the choise of τ and the sensitivity of the performance of the adaptive model selection (or model assessment) to τ has been thoroughly investigated in linear models, logistic regression, Poisson regression. Please refer to Ye (1998), Shen and Ye (Shen2002), Shen, Huang and Ye (2004), and Shen and Huang (2006) for more details. We have performed the sensitivity study to τ in LMMs for the adaptive model selection procedure, and find that the choice of τ does not affect the performance of the adaptive selection procedure in selecting either fixed-effects covariates, random effects, or covariance structures. We follow Shen and Huang (2006) and use $\tau = 0.5$ in our simulations and data analysis.

3.2 Asymptotic optimality

In what follows, we investigate theoretical aspects of $\hat{M}(\hat{\lambda})$, the optimal model selected by adaptive model selection criterion, based on properties of data perturbation. Particularly, the asymptotic optimality of $\hat{M}(\hat{\lambda})$ is established in Theorem 2; that is, $\hat{M}(\hat{\lambda})$ approximates the best performance among all models selected by the procedures with form (9).

Theorem 2—(asymptotic optimality). *Assume that: (1) (integrability) for some* $\delta > 0$ and $\lambda \in (0, \infty)$, $E\sup_{\tau \in (0,\delta)} |\hat{\mathcal{G}}(\hat{M}(\lambda))| < \infty$*; (2) (identifiability)* $\inf_{\lambda \in (0,\infty)} |\mathcal{R}(\xi, \hat{\xi}_{\hat{M}(\lambda)})| > 0$*; (3) (finite variance estimation) for any i, j, and k,* $\lim_{m,n_i \to \infty} \lim_{\tau \to 0+} E\left[(\widehat{\mathrm{var}}(Y_{ij}Y_{ik}) - \mathrm{var}(Y_{ij}Y_{ik}))\mathrm{cov}^*(\widehat{\sigma}_{ijk}(Y^*), Y^*_{ij}Y^*_{ik})/\mathrm{var}(Y^*_{ij}Y^*_{ik})\right] = 0$*. Let* $\hat{\lambda}$ *be the minimizer of (10), then*

$$\lim_{m,n_i \to \infty} \lim_{\tau \to 0+} \frac{E(\mathcal{L}(\xi, \widehat{\xi}_{\widehat{M}(\widehat{\lambda})}))}{\inf_{\lambda \in (0,\infty)} E(\mathcal{L}(\xi, \widehat{\xi}_{\widehat{M}(\lambda)}))} = 1. \tag{13}$$

If it is further assumed that (4) (loss and risk) $\lim_{m,ni \to \infty} \sup_{\lambda \in (0,\infty)} |\mathcal{R}(\xi, \hat{\xi}_{\hat{M}(\lambda)})/E(\mathcal{R}(\xi, \hat{\xi}_{\hat{M}(\lambda)})) - 1| = 0$*, then*

$$\lim_{m,n_i \to \infty} \lim_{\tau \to 0+} \frac{\mathcal{L}(\xi, \widehat{\xi}_{\widehat{M}(\widehat{\lambda})})}{\inf_{\lambda \in (0,\infty)} \mathcal{L}(\xi, \widehat{\xi}_{\widehat{M}(\lambda)})} = 1. \tag{14}$$

Theorems 2 establishes asymptotic optimality of the proposed adaptive model selection procedure in LMMs. The selected model $\hat{M}(\hat{\lambda})$ by the adaptive selection procedure is optimal in the sense that $\hat{M}(\hat{\lambda})$, the model that minimizes (11) with a data-adaptive $\hat{\lambda}$, asymptotically achieves the minimal loss among all models selected by procedures with form (9).

4 Simulation Studies

In this section, we access the finite sample performance of the proposed selection procedure for LMMs through simulations. The numerical studies focus on three aspects in selecting LMMs: (1) selecting fixed-effects covariates in the mean structure, (2) selecting random effects for the covariates and (3) selecting covariance structure for within-cluster errors. By

repeating each simulation 500 times, we compare the performance of AIC, BIC, and adaptive model selection procedure. AIC and BIC are directly applicable. The proposed adaptive procedure is implemented as described above, with $\hat{\lambda}$ in (11) computed through a fine grid search of a large interval $(0, C)$ instead of $(0, \infty)$, where C is sufficiently large. In our simulations, we use $C=3\log(\sum_{i=1}^{m} n_i)$.

Example 1 (fixed effects selection for longitudinal data). This simulation example considers the LMM

$$Y_{ij}=(\beta_0+b_{i0})+X_{ij,1}(\beta_1+b_{i1})+X_{ij,2}\beta_2+\sum_{k=3}^{10}X_{ij,k}\beta_k+\epsilon_{ij}, \tag{15}$$

which contains a "within-cluster time-covariate" $X_{ij,1}$, a "cluster-level covariate" $X_{i,2}$, and other 8 covariates $X_{ij,3},\ldots,X_{ij,10}$. The within-cluster time-covariate $X_{ij,1} = x_{ij}$ takes values $x_{ij} = (j-1)/n_i, j = 1,\ldots,n_i$. The covariate $X_{i,2} = x_i$ takes binary values 0 or 1 with equal probabilities. The covariates $(X_{ij,3},\ldots,X_{ij,10})'$ follow the multivariate normal distribution with zero mean and covariance between the k_1th and k_2th element being $\rho^{|k1-k2|}$, $k_1,k_2 = 3,\ldots, 10$. Three values of ρ are examined: 0.5, 0, and -0.5. The random effects b_{i0}, b_{i1} and within-cluster errors ϵ_{ij} are mutually independent and follow the standard normal distribution. The simulation data are generated from (15) with $m = 20$ clusters and $n_i = 5$ observations for each cluster. In this example, two cases are examined: (1) $\beta_0 = 2.5$, $\beta_1 = 3.75$, $\beta_2 = 1.5$, $\beta_7 = 2$, and $\beta_k = 0$ otherwise; and (2) $\beta_0 = 2.5$, $\beta_1 = 3.75$, $\beta_4 = \beta_5 = \beta_6 = -1.25$, $\beta_8 = \beta_9 = \beta_{10} = 1.75$, and $\beta_2 = \beta_3 = \beta_7 = 0$.

Based on the criteria of AIC, BIC, and the adaptive model selection, we perform backward stepwise selection to select fixed-effects covariates in the simulated datasets generated from each case. The random effects b_{i0} and b_{i1} are forced into each candidate model as we are examining the performance on selecting fixed-effects covariates. The simulation results are summarized in Table 1. In the first case, AIC with $\lambda = 1$ selects the highest number of fixed-effects covariates on average. Because of the large number of incorrectly selected covariates, the proportion of correct fit (exactly selecting the fixed-effects covariates in true model, see Zou and Li 2008) by AIC is the lowest. In contrast, BIC has fewer both correctly and incorrectly selected covariates and produces 0.6 correct-fit rate. Our approach with flexible data-adaptive penalization parameter in (1), however, acts between AIC and BIC in terms of correctly selected fixed-effects covariates and performs the best in terms of the correct-fit rate, the KL loss, and the number of incorrectly selected covariates. It introduces 0.8 correct-fit rate and the lowest averaged KL loss. In the second case, three criteria correctly identify the true nonzero fixed-effects coefficients. Unsurprisingly, AIC collects the most incorrect nonzero coefficients, BIC does less than AIC, and the adaptive procedure collects almost no incorrect fixed-effects covariates.

Example 2 (fixed effects selection for longitudinal data). This simulation example considers the LMM

$$Y_{ij}=(\beta_0+b_{i0})+X_{ij,1}(\beta_1+b_{i1})+\sum_{k=2}^{30}X_{ij,k}\beta_k+\epsilon_{ij}, \tag{16}$$

which contains one "within-cluster time-covariate" $X_{ij,1}$ taking values $x_{ij} = (j-1)/n_i$, $j = 1,\cdots,n_i$ and other 29 covariates $X_{ij,2},\cdots,X_{ij,30}$. The covariates $(X_{ij,2},\cdots,X_{ij,30})'$ follow the

multivariate normal distribution with zero mean and covariance between the k_1th and k_2th element being $\rho^{|k1-k2|}$, k_1, $k_2 = 2, \cdots, 30$. Three values of ρ are considered: 0.5, 0, and −0.5. The random effects b_{i0}, b_{i1} and within-cluster errors ϵ_{ij} are mutually independent and follow the standard normal distribution. The simulation data are generated from Model (16) with $m = 20$ clusters and $n_i = 5$ observations from each cluster. In this example, two cases are examined: (1) $\beta_0 = 2.5$, $\beta_1 = 3.75$, $\beta_{10} = -\beta_{20} = \beta_{30} = 1.25$, and $\beta_k = 0$ otherwise; and (2) $\beta_0 = 2.5$, $\beta_1 = 3.75$, $\beta_2 = \cdots = \beta_9 = 1.25$, $\beta_{11} = \cdots = \beta_{19} = -1.25$, $\beta_{21} = \cdots = \beta_{29} = 1.25$ and $\beta_{10} = \beta_{20} = \beta_{30} = 0$.

Based on the criteria of AIC, BIC, and the adaptive model selection, we perform backward stepwise selection to select fixed-effects covariates to examine the performance of AIC, BIC, and the adaptive model selection. The random effects b_{i0} and b_{i1} are forced into each candidate model. The simulation results are summarized in Table 2. In the first case, all three criteria are able to identify the true covariates. However, AIC adds several incorrect fixed-effects covariates in the selected model, whereas BIC collects fewer than AIC. The adaptive procedure achieves exact selection in every simulation replication. Due to the non-adaptive penalization parameters, AIC and BIC have less than 0.03 and 0.5 correct-fit rate, respectively. In the second case, the adaptive model selection procedure performs the best by offering the lowest KL loss and the highest correct-fit rate. BIC performs better than AIC when correlation coefficient $\rho = 0.5$ or $\rho = 0$. When $\rho = -0.5$, BIC selects much less than 27 fixed-effects covariates in some simulation replications, and therefore dramatically reduces the numbers of incorrect and correct selected covariates.

Example 3 (random effects selection for clustered data). This simulation example considers the LMM

$$Y_{ij} = (\beta_0 + b_{i0}) + \sum_{k=1}^{10} X_{ij,k}(\beta_k + \delta_k b_{ik}) + \epsilon_{ij}, \tag{17}$$

which contains 10 covariates $X_{ij,1}, \cdots, X_{ij,10}$ with possible random effects $b_{i1}, \cdots, b_{i,10}$. Whether or not covariate $X_{ij,k}$ has random effect is determined by the corresponding indicator variable δ_k, which takes values either 0 or 1. The covariates $(X_{ij,1}, \cdots, X_{ij,10})'$ follow the multivariate normal distribution with zero mean and covariance between the k_1th and k_2th element being $\rho^{|k1-k2|}$, k_1, $k_2 = 1, \cdots, 10$, where ρ takes 0.5, 0, and −0.5. The random effects b_{ik} follow a normal distribution with mean zero and standard deviation 0.5, the within-subject errors ϵ_{ij} follow the standard normal distribution, and b_{ik} and ϵ_{ij} are mutually independent. The simulation data are generated from (17) with $m = 10$ clusters and $n_i = 25$ observations from each cluster. In this example, four cases are examined: (1) $\delta_1 = \delta_2 = 1$, and $\delta_k = 0$ otherwise; (2) $\delta_1 = \cdots = \delta_4 = 1$, and $\delta_k = 0$ otherwise; (3) $\delta_1 = \cdots = \delta_6 = 1$, and $\delta_k = 0$ otherwise; and (4) $\delta_1 = \cdots = \delta_8 = 1$, and $\delta_k = 0$ otherwise. Throughout the four cases, the fixed-effects coefficients are assigned values as $\beta_0 = 1.5$, $\beta_1 = \cdots = \beta_{10} = 1.25$.

We conduct the best subset search for 10 random effects $b_{i1}, \cdots, b_{i,10}$. We always include the fixed-effects coefficients of ten covariates and the random intercept in the candidate models. The simulation results are summarized in Tables 3 and 4. Our proposed method achieves the best performance in all cases in terms of the number of correctly selected random effects, the KL loss and the proportion of correct fit (exactly selecting the random effects in true model). In Cases 1 and 2, BIC with fixed penalization parameter λ=2.76 selects smaller number of both correct and incorrect random effects than AIC, and it also produces higher correct-fit rate. As a comparison, AIC with fixed penalization parameter λ=1 does much better than BIC in terms of the proportion of correct fit in Cases 3 and 4, because the large

number of random effects in the true models prefers relatively small penalties. However, the adaptive model selection outperforms AIC and BIC in all cases.

Example 4 (covariance structure selection). This simulation example considers the LMM

$$Y_{ij}=(\beta_0+b_{i0})+X_{ij}(\beta_1+b_{i1})+\epsilon_{ij}, \tag{18}$$

which contains a "within-cluster time-covariate" $X_{ij,1}$ taking values $x_{ij} = (j-1)/n_i$, the random effects b_{i0} and b_{i1} independently following $N(0,0.5)$, and correlated within-cluster errors that are generated from a mixed autoregressive-moving average (ARMA) model (Box and Jenkins 1994)

$$\epsilon_{ij}=\sum_{l_1=1}^{r_1}\phi l_1\epsilon_{i,j-l_1}+\sum_{l_2=1}^{r_2}\theta_{l_2}a_{i,j-l_2}+a_{ij},$$

with homoscedastic noise a_{ij} independently following $N(0, 0.5)$, r_1 autoregressive parameters ϕ_{l1}, and r_2 moving average parameters θ_{l2}. The simulation data are generated from (18) with $m = 5$ clusters and $n_i = 50$ observations from each cluster. The goal of the correlation structure selection is to determine parameters r_1 and r_2. Five correlation structures are examined: $ARMA(s,5)$, $s = 1, 2, \cdots, 5$. For implementation, r_1 and r_2 are assumed to potentially take values $0, 1, \cdots, 10$. We conduct an exhaustive search for LMMs with all possible $ARMA(r_1,r_2)$ structures. Throughout the simulation study, $\beta_0 = \beta_1 = 1$ and $\phi_{l1} = \theta_{l2} = 0.5$ for any l_1 and l_2. The simulation results are summarized in Table 5. The proposed procedure yields the best performance in five situations in terms of the KL loss, the average r_1 and r_2 in the selected models, and the proportion of correctly selecting the true covariance structure. It is evident from Table V that AIC and BIC, with a nonadaptive penalty, cannot simultaneously perform well for both large and small s.

5 Application: Estradiol Levels in the BioCycle Study

5.1 The BioCycle study

The BioCycle study is an epidemiologic study of menstrual cycle function among healthy, regularly menstruating women. It was conducted by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development and the State University of New York at Buffalo from 2005 to 2007. One of the objectives was to study endogenous reproductive hormone levels and their association with other covariates across the menstrual cycle. The BioCycle Study followed 259 regularly menstruating premenopausal women from New York for up to two menstrual cycles. The study population, materials, and methods have been previously described in detail (Wactawski-Wende *at el.* 2009). In summary, healthy women between the ages of 18–44 had to be regularly menstruating (self-reported cycle length between 21 and 35 days for each menstrual cycle in the past 6 months) in order to participate. Women with conditions known to affect menstrual cycle function such as polycystic ovary disease, uterine fibroids, or current use of hormonal contraception (i.e., 3 months prior to study entry) were excluded. Eligible participants visited the study clinic 8 times during each menstrual cycle, at which time fasting serum samples were collected. The visits were scheduled to occur during specific phases of the menstrual cycle, based on an approximate 28-day cycle length. Hormones levels, including estradiol levels, and other biological markers, including insulin and lipoprotein cholesterol levels, were measured from serum samples collected at each visit. Participants were asked to complete standardized

questionnaires at the baseline visit on lifestyle, physical activity, and reproductive history. Dietary intake was assessed four times per cycle using the 24-hour dietary recall methodology and the Nutrition Data System for Research software version 2005 developed by the Nutrition Coordinating Center, University of Minnesota, Minneapolis, MN. Physical and anthropometric measures were done according to standardized protocols and included height and weight for the calculation of body mass index.

Due to the considerable variability in hormone levels both between women and within a woman from cycle to cycle, LMMs with random effects and complex within-cluster covariance structure have typically been used to account for the correlations between and within women in the analysis of factors associated with hormone levels and menstrual cycle function (Schisterman *at el.* 2010; Mumford *at el.* 2010). One of the challenges during the data analysis was the demand of precisely selecting covariates in LMMs as well as their random effects and covariance structure. However, traditional model selection methods for LMMs have major drawbacks and sometimes fail to identify the associated factors and correlation structure, which can induce inaccurate prediction of hormone levels and incorrect interpretation of the association between hormone levels and the predictors. This motivated us to propose a novel model selection procedure with better performance, so as to benefit not only the BioCycle study, but also other clinical or epidemiologic studies in the future that will use LMMs.

5.2 Modeling estradiol levels

One of the hormones that is of particular interest is estradiol, as estradiol is the primary estrogen secreted by the ovary and the predominant sex hormone present in females. Estradiol plays a key role in reproductive function, as well as in the development and recurrence of breast cancer and other chronic diseases. Understanding the factors associated with estradiol levels may aid in understanding an individual's susceptibility to disease, as well as offer potential strategies for disease prevention. It has also been argued that differences in breast cancer incidence between populations could be due to differences in demographic characteristics associated with estradiol levels. Here we are interested in identifying factors that are associated with estradiol levels and accessing variation in estradiol levels both between women and within a women across the menstrual cycle. The main outcome of interest is the logarithm of estradiol levels as measured in fasting serum samples in the BioCycle study. Potential biological factors that might influence estradiol levels include age (X_{age}), body mass index (X_{bmi}), race (white, black X_{rac1}, other X_{rac2}), past use of oral contraceptives (yes or no, X_{oc}), age at menarche (X_{men}), parity (X_{par}), marital status (married/living as married or single/separated/divorced, X_{mar}), physical activity (low, moderate X_{phy1}, high X_{phy2}), smoking status (ever or never, X_{smo}), insulin levels (X_{ins}), the logarithm of total cholesterol levels (X_{cho}), the logarithm of luteinizing hormone levels (X_{lh}), dietary fat intake (X_{fat}), dietary fiber intake (X_{fib}), and total energy intake (X_{ene}). We consider LMMs that include an intercept and a subset of the 19 covariates: aforementioned 17 covariates plus the standardized cycle day (X_{Day}) and the quadratic term of standardized cycle day (X^2_{Day}). We allow the intercept, X_{age}, X_{lh}, X_{Day}, and X^2_{Day} to potentially have random effects and allow correlated within-cluster errors modeled by $ARMA(r_1,r_2)$ with r_1 and r_2 possibly taking values 0, 1, · · ·, 5.

We implement the best subset selection with AIC, BIC, and our proposed procedure. Selected fixed effects, random effects, autoregressive parameters, and moving average parameters by the three selection procedures are summarized in Table 6. Figure 1 shows the numbers of selected fixed effects, selected random effects, selected autoregressive and moving average parameters by the information criteria (1) with the penalization parameter λ changing from 0.1 to 10.0. All three procedures select fixed effects X_{age}, X_{lh}, X_{Day}, X^2_{Day} and

the intercept. AIC selects three more fixed effects, namely X_{par}, X_{fib}, and X_{phy1}, while BIC selects only one more fixed effect, namely X_{phy1}. The adaptive model selection procedure with penalization parameter $\lambda = 2.25$, however, selects X_{fib}, and X_{phy1}. The previous literature scientifically supports the conclusions of the proposed method. In particular, the lack of association between parity and estradiol is consistent with Westhoff *et al.* (1996). Moreover, high fiber intake has been associated with lower levels of estradiol in many studies (Bagga *et al* 1995; Gann *et al* 2003; Goldin *et al.* 1994) presumably due to a reduction in β-glucuronidase activity in feces in response to fiber intake, which subsequently leads to a decline in the reabsorption of estrogen in the colon. For random effects selection, all three methods select a random coefficient for the intercept. BIC and the proposed method select an additional random effect for X^2_{Day}, while AIC selects additional random effects for both X^2_{Day} and X_{lh}. For within-cluster covariance structure selection, the proposed method, AIC, and BIC select *ARMA*(1,1), *ARMA*(2, 2), and *ARMA*(0,1), respectively. From the simulation studies and theoretical properties shown in the previous sections, the LMM selected by the proposed methodology has the best prediction performance and the highest probability of correct-fit.

6 Discussion

In the analysis of biomedical data, LMMs are useful models. For data like hormone levels in the BioCycle study, linear mixed models provide an attractive framework. Sophisticated model selection procedures can help LMMs to improve their interpretability and predictability. This article develops the concept of GDF for LMMs, as well as a data perturbation estimation procedure of the GDF of LMMs. As a model complexity measurement of LMMs, the GDF permits adaptive model selection, in which the penalization parameter is estimated from data. We show the adaptive model selection procedure approximates the best performance of nonadaptive alternatives within the class (1) in model selection. We examine performance of the proposed model selection procedure in fixed effects selection, random effects selection and covariance structure selection. Numerical examples suggest that it performs well against information criteria with form (1) in terms of the KL loss and correct-fit rate. As seen from the simulations and theoretical results, the adaptive model selection has advantages over its nonadaptive counterparts in LMMs.

We have concentrated on developing the adaptive model selection procedure for LMMs. The idea of data-adaptive selection can be extended to other models such as survival models and generalized linear mixed models. Such extension works may require deriving the GDF and the optimal loss estimators for those models.

Appendix

The proof of the Theorem 1 is straightforward. Before we present the proof of Theorem 2, a lemma is presented.

Lemma 1

Under the Assumptions (1) and (3) in Theorem 2, the data perturbation estimator $\hat{g}(\hat{M}(\lambda))$ of the GDF of $\hat{M}(\lambda)$ in Section 3.1 satisfies

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0+} E\widehat{\mathcal{G}}(\widehat{M}(\lambda)) = \lim_{m,n_i\to\infty}\mathcal{G}(\widehat{M}(\lambda)),\ \forall\lambda\in(0,\infty). \tag{19}$$

Proof of Lemma 1

First note that the left hand side of (19) is equal to

$$\begin{aligned}
&\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\left[\frac{\mathrm{var}(Y_{ik})}{\mathrm{var}^*(Y_{ik}^*)}E^*\widehat{\sigma}_{ijk}(Y^*)\widehat{\mu}_{ij}(Y^*)(Y_{ik}^*-E^*Y_{ik}^*)\right]\\
&\quad+\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\left[\frac{\mathrm{var}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}\mathrm{cov}^*(\widehat{\sigma}_{ijk}(Y^*),Y_{ij}^*Y_{ik}^*)\right]\\
&\quad+\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\left[\frac{\widehat{\mathrm{var}}(Y_{ij}Y_{ik})-\mathrm{var}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}\mathrm{cov}^*(\widehat{\sigma}_{ijk}(Y^*),Y_{ij}^*Y_{ik}^*)\right].
\end{aligned}\tag{20}$$

By the Assumption (3), the last term in (20) can be dropped. By assumption (1) and dominated convergence theorem, the first term in (20) is equal to

$$\begin{aligned}
&\lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}\lim_{\tau\to 0+}E\left[\tfrac{\mathrm{var}(Y_{ik})}{\mathrm{var}^*(Y_{ik}^*)}E^*\tfrac{\partial}{\partial Y_{ik}^*}\widehat{\sigma}_{ijk}(Y^*)\widehat{\mu}_{ij}(Y^*))\mathrm{var}^*(Y_{ik}^*)\right]\\
&=\lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\left[\mathrm{var}(Y_{ik})\left(E^*\mathrm{var}(Y_{ik}^*)\tfrac{\partial}{\partial Y_{ik}^*}\widehat{\sigma}_{ijk}(Y^*)\widehat{\mu}_{ij,\lambda}(Y^*)\right)\right]\\
&=\lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\left[\mathrm{var}(Y_{ik})\tfrac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(Y)\widehat{\mu}_{ij}(Y)\right]\\
&=\lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\tfrac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(Y)\widehat{\mu}_{ij}(Y)(Y_{ik}-EY_{ik}).
\end{aligned}$$

Similarly, the second term in (20) is equal to

$$\begin{aligned}
&\lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}\lim_{\tau\to 0+}E\left[\tfrac{\mathrm{var}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}E^*\tfrac{\partial}{\partial Y_{ij}^*Y_{ik}^*}\widehat{\sigma}_{ijk}(Y^*)\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)\right]\\
&=\lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j,k=1}^{n_i}E\tfrac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk}(Y)(Y_{ij}Y_{ik}-EY_{ij}Y_{ik}).
\end{aligned}$$

Therefore, (19) holds.

Proof of Theorem 2

Suppose λ_{opt} minimizes $\mathscr{L}(\xi,\hat{\xi}_{\widehat{M}(\lambda)})$ in terms of $\lambda_{opt}=\inf_{\lambda\in(0,\infty)}E(\mathscr{L}(\xi,\hat{\xi}_{\widehat{M}(\lambda)}))$. By the definition of λ_{opt}, we have

$$\begin{aligned}
&\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}E(\mathscr{L}(\xi,\widehat{\xi}_{\widehat{M}(\widehat{\lambda})}))\\
&\qquad=\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}E\left[-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_{\widehat{M}(\widehat{\lambda})})+\mathcal{G}(\widehat{M}(\widehat{\lambda}))\right]\\
&=\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}E\left[-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_{\widehat{M}(\widehat{\lambda})})+\widehat{\mathcal{G}}(\widehat{M}(\widehat{\lambda}))-\widehat{\mathcal{G}}(\widehat{M}(\widehat{\lambda}))+\mathcal{G}(\widehat{M}(\widehat{\lambda}))\right]\le\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}E\left[-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_{\widehat{M}(\lambda_{opt})})+\widehat{\mathcal{G}}(\widehat{M}(\lambda_{opt}))-\widehat{\mathcal{G}}(\widehat{M}(\widehat{\lambda}))+\mathcal{G}(\widehat{M}\right.\\
&=\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}E\left[-\sum_{i=1}^{m}\log p(Y_i|\widehat{\xi}_{\widehat{M}(\widehat{\lambda})})+\mathcal{G}(\widehat{M}(\lambda_{opt}))-\mathcal{G}(\widehat{M}(\lambda_{opt}))+\widehat{\mathcal{G}}(\widehat{M}(\lambda_{opt}))-\widehat{\mathcal{G}}(\widehat{M}(\widehat{\lambda}))+\mathcal{G}(\widehat{M}(\widehat{\lambda}))\right]\\
&=\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}E(\mathscr{L}(\xi,\widehat{\xi}_{\widehat{M}(\lambda_{opt})}))+\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}(E\widehat{\mathcal{G}}(\widehat{M}(\lambda_{opt}))-\mathcal{G}(\widehat{M}(\lambda_{opt})))+\lim_{m,n_i\to\infty}\lim_{\tau\to 0+}(E\widehat{\mathcal{G}}(\widehat{M}(\widehat{\lambda}))-\mathcal{G}(\widehat{M}(\widehat{\lambda})))
\end{aligned}$$

By Lemma 1, $\lim_{m,ni\to\infty}\lim_{\tau\to 0^+}(E\,\hat{\mathcal{G}}(\hat{M}(\lambda_{opt}))-\mathcal{G}(\hat{M}(\lambda_{opt}))) = 0$, and $\lim_{m,ni\to\infty}\lim_{\tau\to 0^+}(E\,\hat{\mathcal{G}}(\hat{M}(\hat{\lambda}) - \mathcal{G}(\hat{M}(\hat{\lambda})) = 0$. Therefore,

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0+} E(\mathcal{L}(\xi,\widehat{\xi}_{\widehat{M}(\hat{\lambda})}))=\lim_{m,n_i\to\infty}\lim_{\tau\to 0+} E(\mathcal{L}(\xi,\widehat{\xi}_{\widehat{M}(\lambda_{opt})})),$$

which implies (13). With Assumption (4), (14) can be further concluded.

Acknowledgments

The authors would like to sincerely thank Editor, Associate Editor and two anonymous referees for their insightful comments that have led to significant improvement of this paper. Bo Zhang and Sunni Mumford's research was supported by the Intramural Research Program of the National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Xiaotong Shen's research was supported in part by NIH grant 1R01GM081535–01, and NSF grants DMS–0604394 and DMS–0906616. We thank the Center for Information Technology, the National Institutes of Health, for providing access to the high performance computational capabilities of the Biowulf Linux cluster.

References

Akaike, H. Information theory and the maximum likelihood principle. In: Petrov, V.; Csáki, F., editors. International Symposium on Information Theory. Budapest: Akademiai Kiádo; 1973. p. 267-281.

Box, GEP.; Jenkins, GM.; Reinsel, GC. Time Series Analysis: Forecasting and Control. 3rd. Holden-Day; San Francisco: 1994.

Diggle, P.; Heagerty, P.; Liang, K.; Zeger, S. Analysis of Longitudinal Data. 2nd. Oxford University Press; Oxford: 2002.

Efron B. The estimation of prediction error: covariance penalties and cross-validation. Journal of the American Statistical Association. 2004; 99:619–642.

Gann PH, Chatterton RT, Gapstur SM, Liu K, Garside D, Giovanazzi S, Thedford K, Van Horn L. The effects of a low-fat/high-fiber diet on sex hormone levels and menstrual cycling in premenopausal women: a 12-month randomized trial (the diet and hormone study). Cancer. 2003; 98:1870–1879. [PubMed: 14584069]

George EI, Foster DP. The risk inflation criterion for multiple regression. The Annals of Statistics. 1994; 22:1947–1975.

Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika. 1989; 76:297–307.

Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]

Longford, NT. Random Coefficient Models. Oxford; Clarendon: 1993.

Mumford SL, Schisterman EF, Siega-Riz AM, Browne RW, Gaskins AJ, Trevisan M, Steiner AZ, Daniels JL, Zhang C, Perkins NJ, Wactawski-Wende J. A longitudinal study of serum lipoproteins in relation to endogenous reproductive hormones during the menstrual cycle: findings from the biocycle study. Journal of Clinical Endocrinology and Metabolism. 2010; 95:E80–E85. [PubMed: 20534764]

Pinheiro, JC.; Bates, DM. Mixed-effects models in S and S-PLUS. Springer-Verlag; New York: 2000.

Schisterman EF, Gaskins AJ, Mumford SL, Browne RW, Yeung E, Trevisan M, Hediger M, Zhang C, Perkins NJ, Hovey K, Wactawski-Wende J. Influence of endogenous reproductive hormones on F_2-isoprostane levels in premenopausal women: the BioCycle study. American Journal of Epidemiology. 2010; 172:430–439. [PubMed: 20679069]

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6:461–464.

Shen X, Huang H, Ye J. Adaptive model selection and assessment for exponential family. Technometrics. 2004; 46:306–317.

Shen X, Huang H. Optimal model assessment, selection, and combination. Journal of the American Statistical Association. 2006; 102:554–568.

Shen X, Ye J. Adaptive model selection. Journal of the American Statistical Association. 2002; 97:210–221.

Tibshirani R, Knight K. The covariance inflation criterion for model selection. Journal of the Royal Statistical Society, Ser B. 1999; 61:529–546.

Wactawski-Wende J, Schisterman EF, Hovey KM, Howards PP, Browne RW, Hediger M, Liu A, Trevisan M. BioCycle study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. Paediatric and Perinatal Epidemiology. 2009; 23:171–184. [PubMed: 19159403]

Weisberg, S. Applied Linear Regression. 3rd. Wiley/Interscience; New York: 2005.

Westhoff C, Gentile G, Lee J, Zacur H, Helbig D. Predictors of ovarian steroid secretion in reproductive-age women. American Journal of Epidemiology. 1996; 144:381–388. [PubMed: 8712195]

Wolfinger RD. An example of using mixed models and proc mixed for longitudinal data. Journal of Biopharmaceutical Statistics. 1997; 7:481–500. [PubMed: 9358325]

Ye J. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association. 1998; 93:120–131.

Zou H, Li R. One-step sparse estimates in nonconcave penlaized likelihood models. Annals of Statistics. 2008; 36:1509–1533. [PubMed: 19823597]

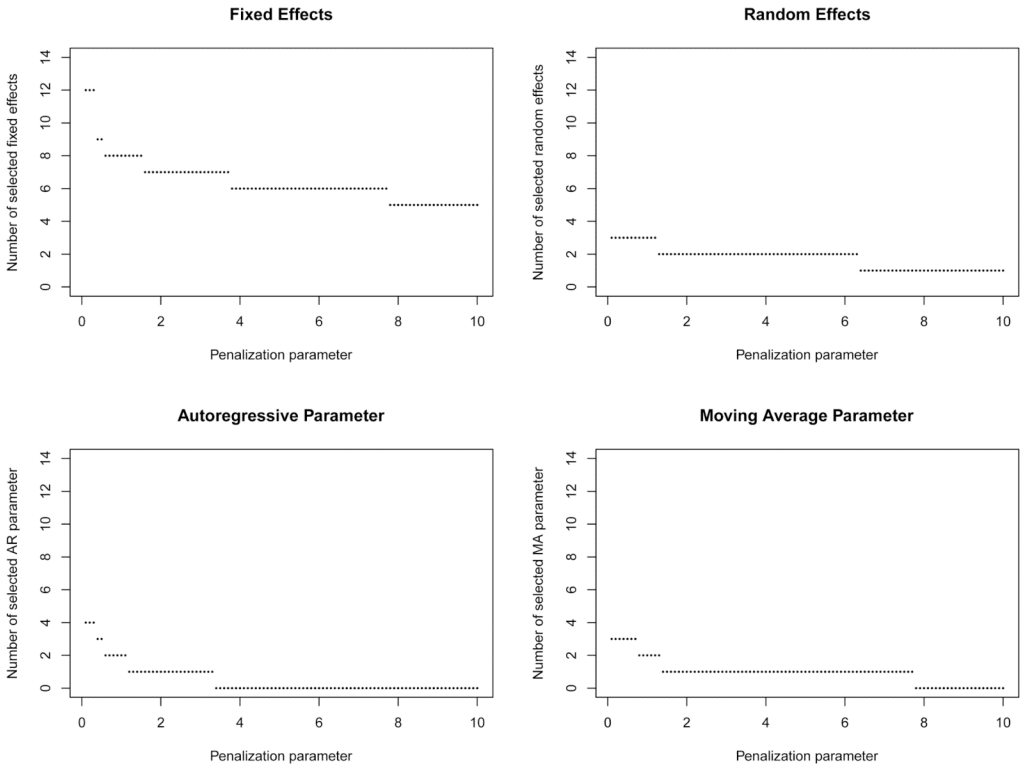

Figure 1.
Numbers of selected fixed effects, numbers of selected random effects, numbers of autoregressive (AR) parameter and numbers of moving average (MA) parameter, by information criteria (1) with the penalization parameter λ changing from 0.1 to 10.0 (grid length 0.1).

Table 1

The performance of adaptive model selection procedure, AIC and BIC in selecting fixed-effects covariates in Example 1. “GDF” refers to the averaged GDF of selected models; “KL” refers to the averaged KL loss of selected models with standard error in the parenthesis; “C” refers to the average number of correctly selected fixed-effects covariates; “IC” refers to the average number of incorrectly selected fixed-effects covariates; “Correct-fit” refers to the proportion, in 500 simulation replications, of the selection procedure exactly selecting the true fixed-effects covariates.

		GDF	KL	C	IC	Correct-fit
			Case 1			
$\rho = 0.5$	AIC	13.42	79.63(4.71)	2.93	1.14	0.32
	BIC	10.60	78.51(4.34)	2.72	0.24	0.61
	Adaptive	8.92	77.25(3.61)	2.88	0.10	0.80
$\rho = 0$	AIC	12.65	79.40(4.63)	2.94	1.14	0.30
	BIC	9.78	78.21(4.15)	2.76	0.25	0.63
	Adaptive	8.25	77.15(3.24)	2.86	0.11	0.77
$\rho = -0.5$	AIC	13.00	79.49(4.93)	2.93	1.13	0.31
	BIC	9.95	78.12(4.03)	2.76	0.24	0.65
	Adaptive	8.37	77.08(3.43)	2.86	0.09	0.79
			Case 2			
$\rho = 0.5$	AIC	15.86	80.94(5.02)	7.00	0.56	0.56
	BIC	14.33	80.02(4.64)	7.00	0.14	0.87
	Adaptive	13.05	79.23(3.89)	7.00	0.00	1.00
$\rho = 0$	AIC	15.86	80.93(5.05)	7.00	0.61	0.50
	BIC	14.13	79.90(4.72)	7.00	0.15	0.86
	Adaptive	12.87	79.12(3.88)	7.00	0.00	1.00
$\rho = -0.5$	AIC	15.92	80.97(4.97)	7.00	0.57	0.54
	BIC	14.33	80.03(4.76)	7.00	0.13	0.88
	Adaptive	13.33	79.40(4.11)	7.00	0.01	0.99

Table 2

The performance of adaptive model selection procedure, AIC and BIC in selecting fixed-effects covariates in Example 2. “GDF” refers to the averaged GDF of selected models; “KL” refers to the averaged KL loss of selected models with standard error in the parenthesis; “C” refers to the average number of correctly selected fixed-effects covariates; “IC” refers to the average number of incorrectly selected fixed-effects covariates; “Correct-fit” refers to the proportion, in 500 simulation replications, of the selection procedure exactly selecting the true fixed-effects covariates.

		GDF	KL	C	IC	Correct-fit
			Case 1			
$\rho = 0.5$	AIC	33.02	92.07(12.10)	4.00	4.57	0.02
	BIC	16.29	81.43(6.58)	4.00	0.92	0.42
	Adaptive	8.73	77.02(2.92)	4.00	0.00	1.00
$\rho = 0$	AIC	37.98	95.16(13.82)	4.00	5.55	0.01
	BIC	18.17	82.42(6.95)	4.00	1.12	0.32
	Adaptive	9.25	77.17(3.09)	4.00	0.00	1.00
$\rho = -0.5$	AIC	34.13	92.33(11.63)	4.00	4.66	0.02
	BIC	17.49	81.81(6.94)	4.00	0.94	0.42
	Adaptive	9.78	77.25(3.32)	4.00	0.00	1.00
			Case 2			
$\rho = 0.5$	AIC	56.11	107.91(16.15)	27.00	0.75	0.42
	BIC	53.29	105.91(15.20)	27.00	0.23	0.78
	Adaptive	51.54	104.64(14.18)	27.00	0.16	0.86
$\rho = 0$	AIC	57.69	108.90(16.48)	27.00	0.90	0.34
	BIC	54.54	106.81(16.42)	26.97	0.30	0.73
	Adaptive	53.03	105.39(14.73)	27.00	0.20	0.83
$\rho = -0.5$	AIC	56.96	108.01(14.63)	27.00	0.81	0.40
	BIC	31.09	133.62[#](29.44)	17.78[#]	0.23[#]	0.37[#]
	Adaptive	53.64	105.67(13.30)	27.00	0.42	0.68

[#]BIC selects much less than 27 fixed-effects covariates in some simulation replications when correlation coefficient $\rho = -0.5$.

Table 3

The performance of adaptive model selection procedure, AIC and BIC in selecting random effects in Example 3. “GDF” refers to the averaged GDF of selected models; “KL” refers to the averaged KL loss of selected models with standard error in the parenthesis; “C” refers to the average number of correctly selected random effects; “IC” refers to the average number of incorrectly selected random effects; “Correct-fit” refers to the proportion, in 500 simulation replications, of the selection procedure exactly selecting the true random effects.

		GDF	KL	C	IC	Correct-fit
			Case 1			
$\rho = 0.5$	AIC	17.91	171.20(5.14)	1.92	0.45	0.59
	BIC	17.83	171.72(5.99)	1.83	0.05	0.79
	Adaptive	16.53	170.28(4.40)	1.98	0.07	0.92
$\rho = 0$	AIC	17.33	172.29(5.19)	1.98	0.45	0.61
	BIC	16.83	172.35(6.53)	1.94	0.03	0.91
	Adaptive	15.95	171.41(4.73)	1.99	0.02	0.97
$\rho = -0.5$	AIC	15.38	170.08(4.83)	1.96	0.48	0.60
	BIC	15.22	170.57(6.38)	1.87	0.09	0.83
	Adaptive	13.85	169.17(4.17)	2.00	0.07	0.93
			Case 2			
$\rho = 0.5$	AIC	22.82	192.26(9.14)	3.67	0.27	0.62
	BIC	25.22	195.65(11.40)	3.33	0.05	0.45
	Adaptive	21.00	190.45(7.21)	3.91	0.38	0.88
$\rho = 0$	AIC	22.72	194.55(8.69)	3.90	0.29	0.68
	BIC	23.93	196.47(10.87)	3.73	0.02	0.74
	Adaptive	21.46	193.54(7.83)	3.97	0.31	0.80
$\rho = -0.5$	AIC	21.74	192.15(8.69)	3.68	0.25	0.60
	BIC	24.17	195.58(11.04)	3.33	0.03	0.47
	Adaptive	19.91	190.40(7.32)	3.92	0.31	0.87

Table 4

The performance of adaptive model selection procedure, AIC and BIC in selecting random effects in Example 3. “GDF” refers to the averaged GDF of selected models; “KL” refers to the averaged KL loss of selected models with standard error in the parenthesis; “C” refers to the average number of correctly selected random effects; “IC” refers to the average number of incorrectly selected random effects; “Correct-fit” refers to the proportion, in 500 simulation replications, of the selection procedure exactly selecting the true random effects.

		GDF	KL	C	IC	Correct-fit
			Case 3			
$\rho = 0.5$	AIC	28.16	211.11(10.15)	5.50	0.18	0.58
	BIC	33.19	217.66(12.96)	4.81	0.04	0.24
	Adaptive	25.63	208.37(7.18)	5.90	0.34	0.73
$\rho = 0$	AIC	25.84	214.91(9.48)	5.78	0.26	0.64
	BIC	29.02	219.16(13.06)	5.44	0.02	0.58
	Adaptive	23.77	212.95(7.00)	5.97	0.32	0.73
$\rho = -0.5$	AIC	25.72	210.79(9.27)	5.43	0.11	0.57
	BIC	31.11	217.75(12.76)	4.70	0.01	0.25
	Adaptive	23.08	207.90(6.44)	5.91	0.31	0.79
			Case 4			
$\rho = 0.5$	AIC	31.54	229.83(11.37)	7.13	0.08	0.40
	BIC	38.75	239.15(15.62)	6.05	0.02	0.09
	Adaptive	27.85	225.75(8.28)	7.81	0.30	0.60
$\rho = 0$	AIC	27.28	233.58(9.96)	7.68	0.09	0.69
	BIC	33.58	241.31(15.14)	7.01	0.01	0.35
	Adaptive	25.23	231.45(7.55)	7.93	0.21	0.77
$\rho = -0.5$	AIC	31.44	230.81(11.84)	7.07	0.09	0.36
	BIC	38.46	239.99(14.61)	5.97	0.02	0.07
	Adaptive	27.36	226.32(8.49)	7.81	0.31	0.62

Table 5

The performance of adaptive model selection procedure, AIC and BIC in selecting the covariance structures in Example 4. "GDF" refers to the averaged GDF of selected models; "KL" refers to the averaged KL loss of selected models with standard error in the parenthesis; "r_1" refers to the average number of autoregressive parameters in selected models; "r_2" refers to the average number of moving average parameters in selected models; "Correct-fit" refers to the proportion, in 500 simulation replications, of the selection procedure exactly selecting the true covariance structure.

True structure		GDF	KL	r_1	r_2	Correct-fit
ARMA(1, 5)	AIC	12.61	68.16(4.66)	4.33	5.57	0.15
	BIC	8.25	44.96(4.46)	1.33	5.40	0.73
	Adaptive	7.48	37.70(2.99)	1.29	5.10	0.78
ARMA(2, 5)	AIC	13.63	79.28(5.52)	5.29	5.88	0.17
	BIC	9.76	54.06(2.62)	2.02	4.86	0.77
	Adaptive	8.89	48.32(3.43)	2.09	4.86	0.72
ARMA(3, 5)	AIC	14.31	87.52(4.07)	4.20	6.08	0.28
	BIC	10.31	65.77(2.55)	2.75	4.82	0.78
	Adaptive	9.10	63.24(4.46)	2.95	5.18	0.82
ARMA(4, 5)	AIC	15.13	96.74(3.87)	4.22	5.63	0.33
	BIC	11.75	79.76(4.30)	3.43	5.90	0.79
	Adaptive	10.82	70.27(3.00)	3.72	5.14	0.85
ARMA(5, 5)	AIC	16.86	107.34(4.16)	6.21	6.21	0.40
	BIC	12.92	92.55(2.91)	4.17	2.87	0.57
	Adaptive	11.49	82.16(4.59)	4.55	4.92	0.73

Table 6

Estimated fixed-effects coefficients (the estimated standard deviations in parentheses), estimated standard deviations of random effects, estimated autoregressive and moving average parameters from selected LMMs for hormone levels data in the BioCycle study via different methods.

	Adaptive($\hat{\lambda}$ = 2.2500)	**AIC(λ= 1.0000)**	**BIC(λ = 3.7826)**
	Fixed effects coefficients		
Intercept	$2.5995_{(0\ 1132)}$	$2.4953_{(0.\ 1262)}$	$2.4599_{(0\ 0994)}$
X_{Day}	$0.1558_{(0.0069)}$	$0.1502_{(0.0067)}$	$0.1572_{(0.0070)}$
X^2_{Day}	$-0.0042_{(0.0002)}$	$-0.0040_{(0.0002)}$	$-0.0042_{(0.0002)}$
X_{lh}	$0.2521_{(0.0172)}$	$0.2598_{(0.0172)}$	$0.2502_{(0.0173)}$
X_{age}	$0.0112_{\ (0.0029)}$	$0.0169_{\ (0.0040)}$	$0.0105_{\ (0.0030)}$
X_{cho}	–	–	–
X_{mar}	–	–	–
X_{par}	–	$-0.0494_{(0.0280)}$	–
X_{par}	–	–	–
X_{fat}	–	–	–
X_{ene}	–	–	–
X_{smo}	–	–	–
X_{oc}	–	–	–
X_{bmi}	–	–	–
X_{men}	–	–	–
X_{fib}	$-0.0109_{(0.0041)}$	$-0.0111_{(0.0042)}$	–
X_{phy1}	–	–	–
X_{phy2}	–	–	–
X_{rac1}	$0.2623_{(0.0624)}$	$0.2811_{(0.0629)}$	$0.2885_{(0.0616)}$
X_{rac2}	–	–	–
	Random effects standard deviations		
$\sigma_{\text{Intercept}}$	0.2945	0.3013	0.2865
σ_{Day}	–	–	–
$\sigma_{\text{Day}}{}^2$	0.0004	0.0004	0.0003
σ_{lh}	–	0.0460	–
σ_{scr}	–	–	–
	Autoregressive parameters		
φ_1	−0.3163	−0.7187	–
φ_2	–	−0.7239	–
φ_3	–	–	–
φ_4	–	–	–
φ_5	–	–	–
	Moving average parameters		
θ_1	0.5730	0.9266	0.2949

	Adaptive($\hat{\lambda}$ = 2.2500)	**AIC(λ= 1.0000)**	**BIC(λ = 3.7826)**
θ_2	–	0.7172	–
θ_3	–	–	–
θ_4	–	–	–
θ_5	–	–	–