Statistica Sinica **24** (2014), 1161-1178
doi:http://dx.doi.org/10.5705/ss.2011.230

MINIMUM DESCRIPTION LENGTH PRINCIPLE FOR LINEAR MIXED EFFECTS MODELS

Li Li, Fang Yao, Radu V. Craiu and Jialin Zou

University of Toronto

Abstract: The minimum description length (MDL) principle originated from data compression literature and has been considered for deriving statistical model selection procedures. Most of the existing methods that use the MDL principle focus on models with independent data, particularly in the context of linear regression. This paper considers data with repeated measurements and studies the selection of fixed effect covariates for linear mixed effect models. We propose a class of MDL procedures that incorporate the dependence structure within individual or cluster and use data-adaptive penalties that suit both finite and infinite dimensional data generating mechanisms. Theoretical justifications are provided from both data compression and statistical perspectives, where the covariance of random effects is treated as known or estimated by maximum likelihood. Numerical experiments are conducted to demonstrate the usefulness of the proposed MDL procedure and the influence of the estimated covariance, and an application to U.S. EPA data for air quality control is provided.

Key words and phrases: AIC, BIC, data compression, linear mixed effects, minimum description length.

1. Introduction

The concept of a "true model" responsible for generating a given set of data is usually introduced solely for theoretical purposes since, in practice, the mechanisms producing the data are often much more complex than the models that are being contemplated. From a model selection perspective, a more reasonable aim is to detect, in a class of models, the one that best approximates, or describes, the observed data — this is the approach adopted in this paper. The initial model selection criteria were established from a "true model" perspective for independent data (Schwarz (1978)), or for "best" prediction within a class of candidate models (Akaike (1973)). However, a model selection criterion that performs optimally under a wide variety of scenarios has been elusive so far, this area of research continuing to grow rapidly even after more than 30 years of development. For instance, there has been an increasing demand for criteria applicable for correlated data models such as those encountered in longitudinal studies (Pan (2001); Vaida and Blanchard (2005)).

For a model under consideration and given a sample $\mathcal{Y}_n$ of size n, both the Akaike Information Criterion (AIC, Akaike (1973)) and the Bayesian Information Criterion (BIC, Schwarz (1978)) take the form of a penalized log-likelihood $-2l(\theta|\mathcal{Y}_n) + \mathcal{R}(n, p)$, where $\theta \in \Theta$ is the vector of parameters under focus and p is the dimension of the parameter space Θ. Under independence, the AIC and BIC penalties are, respectively, $\mathcal{R}_{AIC} = 2k$ and $\mathcal{R}_{BIC} = k \log(n)$, where k represents the number of free parameters to be estimated in the model. Both AIC and BIC have become widely accepted concepts and a large literature has been devoted to the study of their statistical properties and to the development of alternative formulations. Modifications of the AIC have been proposed to account for small sample sizes (AIC$_c$, Hurvich and Tsai (1989)), for overdispersion in count data (QAIC, Burnham and Anderson (2002)). For fitting linear mixed effects (LME) models, Vaida and Blanchard (2005) distinguished between model selection for marginal and conditional models. In the latter case, they propose the conditional AIC (cAIC) for selection of the random effects covariates. Similarly, Pauler (1998) proposed a modification of BIC for correlated data models, while Pan (2001) proposed a modification of AIC (named QIC) for generalized estimating equation models. Also relevant are likelihood ratio tests in LME models, most of which aim for testing random effects based on restricted likelihoods (Morrel (1998); Crainiceanu and Ruppert (2004); Wiencierz, Greven and Küchenhoff (2011), among others).

It is well known that the AIC and BIC address different goals of model selection — the former achieves asymptotic optimality and the latter possesses selection consistency (e.g., Shibata (1981); Nishii (1984)). However, Yang (2005) has shown that the main properties of AIC and BIC cannot be shared. This motivates our proposed model selection criterion for models with correlated data built upon the minimum description length (MDL) principle, as it attempts to find a good balance between AIC and BIC. Our exposition of MDL principle differs from the shrinkage-type variable selection based on penalized likelihoods.

The MDL, introduced by Rissanen (1978), originated from Shannon's coding theory (Shannon (1948)), as a general principle for statistical model selection based on the code length needed to describe the data. From the MDL standpoint, any probability distribution is considered for its ability to describe the data and does not have to be identical to the distribution underlying the data-generating mechanism. The connection between MDL and statistical analysis has matured with the work of Barron, Rissanen and Yu (1998), Hansen and Yu (2001), Lee (2000, 2001); Lu, Lund and Lee (2010) and Hansen and Yu (2003). Hansen and Yu (2001, 2003) presented various frameworks in which the MDL principle can be applied to (generalized) linear models with independent data. They also pointed out numerous connections between MDL and AIC/BIC. In this work, we focus on the selection of fixed effects. The main contribution is to derive valid model selection criteria based on the MDL principle for widely used LME models. The proposed criteria systematically take into account the dependence structure by interweaving the estimation of variance-covariance structure with the code length calculation. This does not follow straightforwardly from the independent case and has large effects on the performance of the criterion. The methods developed for LME models are justified as a "valid" description length in the sense of achieving the smallest redundancy in terms of Kullback-Leibler divergence. Moreover, the proposed criteria possess the selection consistency of BIC for finite-dimensional models, while the data-adaptive penalties are observed to mimic the behavior of AIC as the number of covariates increases.

A brief general description of MDL, followed by the proposed criteria for LME models, is presented in Section 2, with theoretical justifications in Section 3. Numerical experiments are conducted in Section 4 to illustrate the performance of the proposal, and an application to the U.S. EPA data is presented in Section 5. The paper ends with concluding remarks and suggestions for future developments.

2. MDL for Linear Mixed Effects Model

2.1. A general description of MDL principle

The MDL principle relies on the length of the code used for data description (or transmission) based on a given model. In general, Cover and Thomas (1991) uncovered the correspondence between the description length function $L(\cdot)$ and the distribution function. Suppose the data string $y = (y_1^\top, \ldots, y_n^\top)^\top$ is modeled by $\mathcal{M} = \{f(y|\theta) : \theta \in \Theta\}$, a class of models known up to θ. If θ is given, then the description length of the data can be found using the density function indexed by θ, $L(y|\theta) = -\log f_\theta(y)$. However, since the parameter needs to be estimated, it is necessary to transmit the estimator $\hat{\theta}$ too. A simple two-stage framework consists of transmitting the estimate $\hat{\theta}$, followed by encoding the data sequence with the distribution $f_{\hat{\theta}}$ indexed by $\hat{\theta}$. Then the resulting code length $L(y) = L(y|\hat{\theta}) + L(\hat{\theta}) = -\log f_{\hat{\theta}}(y) + (k/2)\log(n)$ is the same as BIC, if the responses are independent. The term $\log(n)/2$ reflects the precision used to encode each parameter with a uniform encoder. However, the coding rule suggests assigning short codewords to a common symbol and long codewords to a rare symbol. If we believe that the parameter follows a distribution other than uniform, then the minimum code length should be different from $k\log(n)/2$.

Following the idea of *mixture description length* suggested by Hansen and Yu (2003), we assume for the data a mixture distribution induced by the user-defined probability distribution $\omega(\theta)$ on the parameter space Θ. The mixture description length of y is then $-\log m(y) = -\log \int f_\theta(y)\omega(\theta)d\theta$. If λ is a hyperparameter, $\omega(\theta) = \omega(\theta|\lambda)$, then the code length used to transmit λ should be added. Rissanen (1989) emphasizes that $\omega(\theta)$ is not a Bayesian prior, but "an artificial device to minimize the description length". One can see that the mixture MDL may take into account the "believed" structure of the parameter space. This is a desirable feature that can be carried over to correlated data models which are our primary interest. Due to the complex nature of the models considered here, it is possible that not all parameters can be assigned suitable distributions that lead to closed form calculation of $m(y)$. In such situations, we are often able to utilize appropriate distributions for the parameters of primary interest, e.g., regression coefficients, while the estimates of the nuisance parameters (e.g., variance components) are plugged in and encoded accordingly. From a model selection perspective, the challenge arises from appropriately assessing the information contained in the dependent data.

2.2. Proposed MDL criteria for linear mixed effects models

The general framework of a LME model (Laird and Ware (1982)) is

$$y_i = X_i\beta + Z_i\mathbf{b}_i + \varepsilon_i,$$

where $y_i = (y_{i1}, \ldots, y_{in_i})^\top$ is the $n_i \times 1$ response vector for the i-th subject, $1 \le i \le n$, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ design matrices for fixed and random effects, respectively, β is the $p \times 1$ vector of fixed effects parameters, $\mathbf{b}_i$ is the $q \times 1$ vector of random effects, and n is the number of subjects. It is often assumed that the vector of random effects $\mathbf{b}_i$ is $N(\mathbf{0}, D)$, the $n_i \times 1$ vector of residuals, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^\top$, is $N(\mathbf{0}, R_i)$, and $\mathbf{b}_i, \varepsilon_{i'}$ are mutually independent for all $1 \le i, i' \le n$. Throughout we use the canonical form $R_i = \sigma^2 I_{n_i}$ for a clear exposition. As a consequence, the cluster-specific response vectors y_i are independent $y_i \sim N(X_i\beta, Z_iDZ_i^\top + \sigma^2 I_{n_i})$, $i = 1, \ldots, n$.

We focus on a marginal approach with the selection of the fixed effects covariates β, while the variance-covariance components of the random effects are treated as nuisance parameters. We begin with the simple case in which only β is unknown. To suitably transmit the unknown parameters, we apply the mixture MDL principle. A convenient distribution choice for assigning code length to β is the multivariate normal with covariance hyperparameter V, say $\beta \sim N(0, V)$, a conjugate prior in the Bayesian context. Then the marginal distribution of y has description length $L(y|V) = -\log m(y|V)$. Here V needs to be specified by the user since the length function depends on it. Since we usually use the generalized least square estimator of β in practice, we assume that $V = c\mathrm{Var}(\hat\beta_{GLS}) = c(\sum_{i=1}^n X_i^\top \Sigma_i^{-1} X_i)^{-1}$, where $\Sigma_i = Z_iDZ_i^\top + \sigma^2 I_{n_i}$ and $c \ge 0$ is a scalar. This leads to a simplification of the code length expression, which remains only a function of the hyperparameter c. The MDL principle suggests minimizing the length function with respect to $c \geq 0$, yielding the estimate $\hat{c}$ which is plugged into the code length function and leads to the lMDL$_0$ criterion

$$\begin{cases} \frac{1}{2}\Big\{\sum_{i=1}^{n} y_i^\top \Sigma_i^{-1} y_i - FSS_\sigma + p\Big[1 + \log\Big(\frac{FSS_\sigma}{p}\Big)\Big] + \log n\Big\}, & \text{if } FSS_\sigma > p, \\ \frac{1}{2}\sum_{i=1}^{n} y_i^\top \Sigma_i^{-1} y_i, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $FSS_\sigma = (\sum_{i=1}^n y_i^\top \Sigma_i^{-1} X_i)(\sum_{i=1}^n X_i^\top \Sigma_i^{-1} X_i)^{-1}(\sum_{i=1}^n X_i^\top \Sigma_i^{-1} y_i)$ and $(\log n)/2$ is the length of code needed for transmitting $\hat{c}$. The detailed derivation is deferred to the Appendix. When $FSS_\sigma \leq p$, the estimate $\hat{c} = 0$ and the "device" distribution of β becomes a point mass. This corresponds to the null model with all fixed effects zero.

The first term of the lMDL$_0$ criterion, $(\sum_{i=1}^n y_i^\top \Sigma_i^{-1} y_i - FSS_\sigma)$, is the log-likelihood. From (2.1), lMDL$_0$ has the same form of a penalized likelihood as AIC and BIC, but its penalty is data-adaptive; in lMDL$_0$ both the data size and its dependence structure are taken into account. For instance, the penalty term depends on FSS_σ which involves the covariance matrices Σ_i, $i = 1, \ldots, n$.

The criterion (2.1) is usually impractical, as in most applications all the parameters, $\{\beta, \sigma^2, D\}$, are unknown. Since for modeling purpose the focus is on β, the variance-covariance components are treated as nuisance parameters. Nevertheless, one still needs to consider the impact of the dependence structure on the derivation of MDL criteria. We assume a normal distribution for β but we modify the variance specification to include the parameter σ^2. Thus $\Sigma_i = \sigma^2 W_i$, where $W_i = Z_i D^* Z_i^\top + I_{n_i}$ and $D = \sigma^2 D^*$. Then $\beta \sim N(0, \sigma^2 V)$, where $V = c(\sum_{i=1}^n X_i^\top W_i^{-1} X_i)^{-1}$ with a slight abuse of notation. Put $\tau = \sigma^2$ and assume an inverse gamma distribution as the coding device, $\tau \sim \mathrm{InvGamma}(a, 3/2)$, as suggested by Hansen and Yu (2003). In this formulation the criterion depends on D^*. However introducing a device to encode the $p \times p$ covariance matrix D^* does not yield a closed form criterion, nor is amenable to efficient numerical computation. We thus adopt a two-stage principle to treat D^* by plugging in its consistent estimate and increasing the code length by $s\log(n)/2$ (empirically supported by simulations), where s is the number of distinct parameters used for modeling D^*.

The marginal distribution of y now includes hyperparameters a and c. Following the MDL principle, we jointly estimate them by minimizing the length function (details deferred to the Appendix), yielding the MDL criterion, denoted by lMDL,

$$\begin{cases} \frac{1}{2}\Big\{\sum_{i=1}^{n}\log|\widehat{W}_i|+N\log\Big(\frac{N\cdot RSS_{\widehat{W}}}{N-p}\Big)+p\log\Big[\frac{(N-p)FSS_{\widehat{W}}}{pRSS_{\widehat{W}}}\Big]+(s+2)\log n\Big\}, & \\ & \text{if } \frac{FSS_{\widehat{W}}/RSS_{\widehat{W}}}{p/(N-p)}>1, \\ \frac{1}{2}\Big\{\sum_{i=1}^{N}\log|\widehat{W}_i|+N\log\sum_{i=1}^{n}y_i^\top\widehat{W}_i^{-1}y_i+(s+1)\log n\Big\}, & \text{otherwise,} \end{cases} \tag{2.2}$$

where $N=\sum_{i=1}^{n}n_i$,

$$\begin{aligned} FSS_W &= \Big(\sum_{i=1}^{n}y_i^\top W_i^{-1}X_i\Big)\Big(\sum_{i=1}^{n}X_i^\top W_i^{-1}X_i\Big)^{-1}\Big(\sum_{i=1}^{n}X_i^\top W_i^{-1}y_i\Big), \\ RSS_W &= \sum_{i=1}^{n}y_i^\top W_i^{-1}y_i - FSS_W, \end{aligned} \tag{2.3}$$

and "ˆ" is generic notation for those quantities obtained when replacing D^* with its ML estimate. Similar to $l\text{MDL}_0$, the second form is taken when $\hat{c}=0$ leading to the null model. Note that without the constant part, (2.2) is

$$-\frac{1}{2}\sum_{i=1}^{n}\log|\widehat{W}_i|-\frac{N}{2}\log\frac{N\cdot RSS_{\widehat{W}}}{N-p}$$

which is the log-likelihood calculated using the estimate of D^*. The remaining terms play the role of the penalty on model complexity and show the distinction from AIC or BIC by incorporating the dependence structure inherent to the data.

3. Theoretical Justification

In this section we show that the MDL procedures not only possess the desirable property from data compression perspective, but also enjoy the consistency of BIC and the asymptotic optimality of AIC for different underlying model situations due to the data-adaptive penalty forms. The latter also improve the criterion's finite sample performance, as illustrated by the simulation studies in Section 4.

We begin our exposition from the point of view of data compression. Suppose that $f_\theta(y)$ is the true density function and $Q(y)$ is any other density function with the corresponding code length $(-\log Q)$. To encode the data string y, some extra code length is needed to transmit the parameters. The expected value of this extra code length is called the *redundancy* of Q, defined as $R(Q)=E_\theta\{-\log Q(y)-[-\log f_\theta(y)]\}$, which coincides with the Kullback-Leibler divergence between Q and f_θ. Hansen and Yu (2001) pointed out that if Q can achieve the "smallest" redundancy possible for all members in $\mathcal{M} = \{f_\theta(y) : \theta \in \Theta\}$, then $(-\log Q)$ is a valid description length for the data string based on models from the class $\mathcal{M}$. Rissanen (1986) has shown that, if a $\sqrt{n}$-rate estimator $\hat{\theta}(y)$ exists and it has uniformly summable tail probabilities, $P_\theta\{\sqrt{n}||\hat{\theta}(y) - \theta|| \geq \log(n)\} \leq \delta_n$ for all θ and $\sum_n \delta_n \leq \infty$, then the redundancy for any density Q satisfies, for all $\theta \in \Theta$ except on a set with Lebesgue measure zero,

$$\liminf_{n\to\infty} \frac{E_\theta log[f_\theta(y)/Q(y))]}{(k/2)\log n} \geq 1, \tag{3.1}$$

where k is finite and represents the number of free parameters to be estimated. This implies that one needs at least $(k/2)\log n$ additional bits to encode the data without knowing the true distribution f_θ. If $R(Q) = (k/2)\log n[1 + o(1)]$, we say that the redundancy of Q achieves the lower bound. If one focuses on the primary parameters β, the following property can be obtained, see the Appendix for details.

Theorem 1. *If σ^2 and D are known, conditional on $\{X_1, \ldots, X_n\}$, p is finite, and $n^{-1}\sum_{i=1}^n X_i^\top \Sigma_i^{-1} X_i \rightarrow C$ as $n \rightarrow \infty$ for some positive definite matrix C, $\max_{1\leq i\leq n} n_i < \infty$, then the redundancy of the density that induces $lMDL_0$ achieves the lower bound.*

Although the MDL principle is well motivated from data compression perspective, it is also of interest to assess whether the proposed MDL formulations lead to statistically sensible model selection procedures. AIC and BIC cannot share each other's advantage. AIC achieves asymptotic optimality when the true underlying model is infinite-dimensional, while BIC possesses selection consistency when the "true" model is finite-dimensional. We show that the MDL criteria asymptotically behaves similar to BIC, and we illustrate that the MDL procedures can also adjust the data-adaptive penalties to include more covariates when the model dimension becomes large, which mimics the AIC behavior. For technical convenience, we consider the balanced design with $n_i = m < \infty$, for $i = 1, \ldots, n$, and that all individuals are independent realizations. As argued by Breiman and Freedman (1983, Lemma 2.1), there is no loss of generality by assuming that the columns of the $m \times p$ design matrix, X_{ij}, are independent and identically distributed (i.i.d.), since the column space and σ-field do not differ from those given by a general setting that assumes imperfectly correlated covariates. We thus take $\sigma_0^2 = \sum_{j=1}^\infty \beta_j^2 < \infty$ and $\sigma_p^2 = \sum_{j=p+1}^\infty \beta_j^2$ for simplicity. When the "true" model is finite-dimensional, there exists a p_0, such that $\sigma_p^2 = 0$ for $p \geq p_0$.

Theorem 2. *Under the random design assumptions,*

$$FSS_\sigma = \frac{N}{\sigma^2}\Big[c_0\sigma_0^2 - c_p\sigma_p^2 + \frac{p}{N(c_p\sigma_p^2+\sigma^2)}\Big](1+o_p(1)).$$

$$\frac{(N-p)FSS_W}{pRSS_W} = \Big[\frac{N}{p}\frac{c_0\sigma_0^2 - c_p\sigma_p^2}{c_p\sigma_p^2+\sigma^2}+1\Big](1+o_p(1)).$$

for some positive constants c_0 *and* c_p*, where the* $o_p(1)$ *in both statements are uniform over* $1 \le p \le n/2$*,* FSS_σ *is as in* (2.1) *and* FSS_W *and* RSS_W *are as in (2.3).*

From this theorem we can see that, when the model is finite-dimensional, $FSS_\sigma = O_p(N)$ and $(N-p)FSS_{\widehat{W}}(pRSS_{\widehat{W}})^{-1} = O_p(N)$. Then lMDL$_0$ and lMDL share with BIC the consistency of selection. However, for moderate samples, it is not clear which method performs better and we compare their performances via simulations in the next section. If the model is infinite-dimensional in the sense that p may increase with sample size, it is easy to see that the penalty in lMDL$_0$ is more affected by p with $(c_0\sigma_0^2 - c_p\sigma_p^2)$ decreasing. For lMDL, the factor $[c_0\sigma_0^2 - c_p\sigma_p^2][p(c_p\sigma_p^2+\sigma^2)]^{-1}$ reflects "the average signal to noise ratio for the fitted model", which balances between N and p and makes the penalty relatively small with large p, thus making it more likely to include more covariates and to mimic the behavior of AIC. This self-adjustment property of the proposed MDL criteria makes them suitable for different data generating mechanisms.

Notice the connection between model selection based on information criteria and the likelihood ratio test (LRT), where the penalty difference between the considered full and reduced models can be compared to a LRT rejection region. Unlike AIC/BIC, due to data-adaptive feature of MDL penalties, it is not easy to appreciate such connections in an explicit form. The theoretical justification with estimated D deserves further investigation due to allowing p to vary in the range $1 \le p \le n/2$.

4. Simulation Studies

Previous comparisons have shown that MDL has a more balanced performance than AIC and BIC as its capability is close to the best of the two in a wide range of scenarios for independent data (see, for example, Hansen and Yu (2003); Craiu and Lee (2005)), and we expect similar findings for correlated data models.

To demonstrate the application of MDL criterion for the linear mixed effects models, we simulated data for $n = 50$ (respectively $n = 80$) clusters/individuals each containing $n_i = 4$ repeated measurements, $1 \le i \le n$. The fixed covariates contain an intercept and up to 12 more predictors generated from multivariate normal distributions with unit variances, where the autoregressive (AR) and compound symmetric (CS) correlation structures were used when simulating these predictors. We considered AR(0.3) and CS(0.15) for the weaker correlation, AR(0.5) and CS(0.4) for stronger cases. The whole set of fixed coefficients were $(.5, .6, .6, .6, .5, .6, .5, .5, .5, .4, .5, .5, .5)^\top \in R^{13}$. We considered four underlying models with increasing dimensions: the first four ($x_1 \sim x_4$), the first seven ($x_1 \sim x_7$), the first ten ($x_1 \sim x_{10}$), and all ($x_1 \sim x_{13}$) fixed predictors, respectively. In order to maintain the study computationally feasible, we examined 13 candidate models, beginning with an intercept and subsequently including an additional variable. Two random effect covariates were included, one of which coincided with the fixed covariate and the second independently generated from standard normal, $z_{ij1} = x_{ij1} = 1$ and $z_{ij2} \sim N(0, 1)$. The set of random effects also include a random intercept and were multivariate normal with mean zero and covariances $\text{cov}(b_{i1}, b_{i2}) = 0.8$, $\text{cov}(b_{i1}, b_{i3}) = 0.5$, $\text{cov}(b_{i2}, b_{i3}) = 0.4$. An unstructured covariance was used in estimation throughout the simulation studies. The error vector ε_i was independently simulated from $N(0, 2I_5)$, implying a relatively high noise level.

Our comparison focused on lMDL, as this is the criterion most likely to be used in a realistic data analysis. The lMDL with REML estimate of D yielded suboptimal results in our simulations (not reported for conciseness), thus we focused our comparisons on the lMDL with ML estimate. Of interest is assessing the incurred effect on the model criterion accuracy when D is estimated. Therefore, in addition to lMDL based on the ML estimate of D, we also include, for reference, the $l\text{MDL}_D$ calculated based on the true value. We compare the selection results with those of AIC and BIC based on 500 Monte Carlo replicates for sample sizes $n = 50$ and $n = 80$ in Table 1 and Table 2, respectively. There we see that the lMDL and $l\text{MDL}_D$ yield comparable results, providing empirical support for using the ML-estimated covariance. When the number of fixed covariates is relatively low, e.g. the models contain only the first 4 predictors, AIC always chooses larger models while BIC selects the correct model more often. The proposed lMDL's performance tends to be closer to BIC in this case. When p is increased to 7, the gap between lMDL and BIC becomes narrower, but AIC still underperforms. For the case of 10 predictors, lMDL outperforms both BIC and AIC in all settings. The last model containing all fixed covariates clearly points to AIC as the winner, nevertheless, lMDL yields close results and outperforms BIC. In this numerical study, we confirm the known fact that the AIC and BIC work in opposite directions in terms of over-/under-selection across various model dimensions. More importantly, we see that the performance of the proposed lMDL criterion is more balanced and stable compared to AIC and BIC for different model complexities, and may outperform both for some intermediate cases. This suggests the MDL proposal can be used as a safe alternative in

Table 1. Comparison of lMDL criterion with AIC and BIC for LME with the sample size $n = 50$ and cluster size 4, where lMDL is calculated with the ML estimate of D, while $l\text{MDL}_D$ uses the true values of D. Shown are the selection percentages (%) out of 500 Monte Carlo runs, F<T or F>T corresponds to the final models that are under-selected or over-selected, respectively, while F=T indicates that the true model is correctly identified.

True (T)		$x_1 \sim x_4$				$x_1 \sim x_7$				$x_1 \sim x_{10}$				$x_1 \sim x_{13}$			
Method		AIC	BIC	lMDL	$l\text{MDL}_D$	AIC	BIC	lMDL	$l\text{MDL}_D$	AIC	BIC	lMDL	$l\text{MDL}_D$	AIC	BIC	lMDL	$l\text{MDL}_D$
AR(.3)	F<T	0.6	3.2	1	0.8	1.8	10	4.8	6	7.6	30.6	15.6	16.8	1.8	11.8	4.8	6.2
	F=T	63.2	93.2	82.8	81.6	66	85.8	84	83.8	65.2	67	74.4	72.8	98.2	88.2	95.2	93.8
	F>T	36.2	3.6	16.2	17.6	32.2	4.2	11.2	10.2	27.2	2.4	10	10.4	0	0	0	0
AR(.5)	F<T	0.4	6.6	1.8	1.2	2.6	17.2	7.2	6.8	12	38.6	27.6	29.2	5.2	19.4	11.8	12.2
	F=T	67.6	90.2	83.2	83.6	65.8	81	86.6	84.6	60.6	58.4	66.8	63.6	92.6	80.6	88.2	87.8
	F>T	32	3.2	15	15.2	31.6	1.8	6.2	8.6	27.4	3	5.6	7.2	0	0	0	0
CS(.15)	F<T	0.6	3.4	0.6	0.4	4	7.2	2.6	3.2	6.8	30.6	18.4	18.8	2.2	13.8	7.8	9
	F=T	71.2	93.2	85.2	84.4	67.8	89	85.8	83.2	65.4	67	74.2	72.8	97.8	86.2	92.2	91
	F>T	28.2	3.4	14.2	15.2	28.2	3.8	11.6	13.6	27.8	2.4	7.4	8.4	0	0	0	0
CS(.4)	F<T	1.2	7.4	1.8	1.2	6	24.8	14.4	15.2	6	24.6	16	16.8	2.4	15.6	11.2	13.6
	F=T	64	88.6	84.6	86.2	65.6	72.8	77.2	75.8	63.2	72.4	76.8	74.6	97.6	84.4	88.8	86.4
	F>T	34.8	4	13.6	12.6	28.4	2.4	8.4	9	30.8	3	7.2	8.6	0	0	0	0

Table 2. Comparison of lMDL criterion with AIC and BIC for LME with the sample size $n = 80$ using the same setting as in Table 1.

True (T)		$x_1 \sim x_4$				$x_1 \sim x_7$				$x_1 \sim x_{10}$				$x_1 \sim x_{13}$			
Method		AIC	BIC	lMDL	$l\text{MDL}_D$	AIC	BIC	lMDL	$l\text{MDL}_D$	AIC	BIC	lMDL	$l\text{MDL}_D$	AIC	BIC	lMDL	$l\text{MDL}_D$
AR(.3)	F<T	0	0.4	0	0	0	2.6	0.8	0.6	0.8	12	5	8.6	0	2.8	1.4	1.8
	F=T	70.6	96.6	88.4	89.8	76	96.4	92.8	90.4	73.6	85.2	86.2	85.8	100	97.2	98.6	98.2
	F>T	29.4	3	11.6	10.2	24	1	6.4	9	25.6	2.8	8.8	4.6	0	0	0	0
AR(.5)	F<T	0	0.4	0	0	0.6	5	2	2.6	3	18.8	10.8	11.8	0.6	5.8	3	3.6
	F=T	72.2	97.6	91	89.6	66.8	92.2	86.8	88.2	72.6	80.4	85.8	84.2	99.4	94.2	97	96.4
	F>T	27.8	2	9	10.4	32.6	2.8	11.2	9.2	24.4	0.8	3.4	4	0	0	0	0
CS(.15)	F<T	0	0.2	0	0	0.4	2.4	1	2.4	2	12.6	7.4	7.8	0.2	1.6	0.8	1.4
	F=T	72.6	97.6	89.8	90.2	67.8	96.6	92.4	91.6	71.6	85.4	87.6	86.8	99.8	98.4	99.2	98.6
	F>T	27.4	2.2	10.2	9.8	31.8	1	6.6	6	26.4	2	5	5.4	0	0	0	0
CS(.4)	F<T	0	0.8	0.2	0	0.8	7	3.4	2.8	1	9	5.6	5.8	0.2	2.4	1.4	0.8
	F=T	69	96.4	88.6	87.8	70.4	90.8	89.8	88.4	71.2	88.8	89.6	90.2	99.8	97.6	98.6	99.2
	F>T	31	2.8	11.2	12.2	28.8	2.2	6.8	8.8	27.8	2.2	4.8	4	0	0	0	0

practical data analysis where one has little idea about the underlying model and is uncertain about the use of BIC or AIC.

5. Application to the U.S. EPA Data

We applied the proposed MDL procedure to the U.S. EPA data which have been recently used to study the association between total nitrate concentration in the atmosphere and a set of measured predictors that can serve as surrogates for different nitrate formation and loss pathways (Ghosh et al. (2010)). The data are from the U.S. EPA Clean Air Status and Trends Network (CASTNet), consisting of multiple sites with repeated measurements of pollution and meteorological variables. We based our analysis on the previous study conducted by Bondell, Krishna and Ghosh (2010), and used the same subset of 15 sites in the eastern portion of U.S. for the period of 2000 to 2004, with monthly average as observations. The response was log(TNO_3) as taken by Ghosh et al. (2010), and the nine meteorological predictors were sulfate (SO_4), ammonium (NH_4), ozone (O_3), temperature (T), dew point temperature (T_d), relative humidity (RH), solar radiation (SR), wind speed (WS), and precipitation (P). To describe the liner and seasonal and trends, we also included $l(t) = t$, $s_j = \sin(2\pi jt/12)$, and $c_j = \cos(2\pi jt/12)$, $j = 1, 2, 3$. The response was centered and the predictors standardized to remove the fixed intercept.

Since practitioners usually consider few possible random effects for the meteorological predictors, we set the random effects according to those obtained by Bondell, Krishna and Ghosh (2010), including an intercept, a linear trend $l(t)$, and seasonal trends $s_1(t)$, $c_1(t)$, and $c_2(t)$. The emphasis of Bondell, Krishna and Ghosh (2010) was the joint selection of fixed and random effects in LME by adopting shrinkage penalty via the adaptive LASSO idea. Our goal was to choose fixed predictors while keeping the same random effects. For estimation we use an unstructured covariance for all models considered. We used AIC, BIC, and the proposed lMDL procedures, conducting backward selection starting with a full model containing all 16 fixed predictors. To assess the performance of different methods, we calculated the 5-fold cross-validation likelihood based on 1,000 random splits (van der Laan, Dudoit and Keles (2004)). From Table 3, we see that the lMDL differs from the proposal of Bondell, Krishna and Ghosh (2010) by one variable, T, while the AIC (resp. BIC) selects more (resp. less) variables, as expected. Although the cross-validated (CV) likelihood values slightly favor BIC, the lMDL is comparable to that obtained by Bondell, Krishna and Ghosh (2010) and outperforms AIC. It is known that the performance of linear models can be heavily impacted by strong collinearity. Therefore we conducted a second analysis by first inspecting the 16 predictors, and removing the NH_4 that had the strongest collinearity with all other variables. Then we applied AIC, BIC and lMDL analogously to arrive at models shown in Table 4 with drastically improved CV likelihoods. It is interesting to see that the lMDL now yields the most favorable result in terms of CV likelihood, followed by AIC and then BIC. The model chosen by lMDL is more parsimonious than the one by AIC, and is more desirable in practice. This indicates that the performance of selection criteria in this application is indeed confounded by collinearity, so that none of them is able to yield the "best model" without dropping NH_4. To conclude this section, we also list the models chosen by lMDL based on the REML estimate of D; they are identical to ones chosen by the recommended ML-based lMDL, where the CV likelihood is no longer applicable.

Table 3. Selection of fixed effects for the CASTNet data obtained via different model selection criteria. All models include the random effects $\{1, l(t), c_1(t), s_1(t), c_2(t), s_2(t)\}$.

Method	Fixed Effects	CV Likelihood
Bondell et al.	SO_4, NH_4, O_3, T, RH, P, $l(t), s_1(t), c_1(t), s_2(t)$	-697.29 (0.839)
AIC	SO_4, NH_4, T, O_3, RH, P, WS, $l(t), s_1(t), c_1(t), s_2(t), s_3(t)$	-700.06 (0.912)
BIC	SO_4, NH_4, O_3, P, $l(t), s_1(t), c_1(t), s_2(t)$	-695.53 (0.875)
lMDL	SO_4, NH_4, O_3, T, RH, P, $l(t), s_1(t), c_1(t), s_2(t)$	-696.92 (0.926)
$l\text{MDL}_R$	SO_4, NH_4, O_3, T, RH, P, $l(t), s_1(t), c_1(t), s_2(t)$	—

Table 4. Comparison of fixed effects selected for the CASTNet data after excluding the NH_4 covariate that exhibits strong collinearity with all other variables. All models include the random effects $\{1, l(t), c_1(t), s_1(t), c_2(t), s_2(t)\}$.

Method	Fixed Effects	CV Likelihood
AIC	SO_4, T, O_3, RH, P, WS, $l(t), s_1(t), c_1(t), s_2(t), s_3(t)$	-641.24 (0.892)
BIC	T, O_3, RH, P, $l(t), s_1(t), c_1(t), s_2(t)$	-644.83 (0.864)
lMDL	SO_4, O_3, T, RH, P, WS, $l(t), s_1(t), c_1(t), s_2(t)$	-640.42 (0.903)
$l\text{MDL}_R$	SO_4, O_3, T, RH, P, WS, $l(t), s_1(t), c_1(t), s_2(t)$	—

6. Conclusion

In this work we motivate the use of MDL principle for LME models, and provide theoretical justifications from both information and statistical perspectives. These results partially explain its advantage across different model dimensions, as evidenced by our numerical studies, and suggests that the MDL principle may be used as a viable alternative, especially when one has little information about the underlying model. In spite of many documented instances where the MDL principle yields reliable model selection procedures, their use in the statistical literature has been rather infrequent. We hope that our contribution to MDL criterion for LME models will stimulate the growth and usage of this exciting area of statistics. Theoretical investigation of the unknown random effects covariance in the MDL context is challenging and deserves further study. Extensions to a variety of complex data models, e.g., categorical response with generalized estimating equation approach and spline representation models for functional data, are potentially useful. We are also interested in adopting the MDL principle in choosing tuning parameters in shrinkage-type variable selection approaches for high-dimensional data, in which very few methods have been utilized besides AIC and BIC.

Acknowledgements

We thank the editor, an associate editor, and three referees for their insightful suggestions that have improved the paper. This research was partially supported by Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix: Derivations and Technical Proofs

A.1. Derivations of MDL formulations

Derivation of lMDL$_0$ (2.1). The marginal distribution of y is

$$m(y|V) = \int \Big(\prod_{i=1}^{n} \frac{1}{(2\pi)^{n_i/2}|\Sigma_i|^{1/2}}\Big) \exp\Big\{ -\sum_{i=1}^{n} \frac{1}{2}(y_i - X_i\beta)^\top \Sigma_i^{-1}(y_i - X_i\beta)\Big\} \\ \times \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\Big\{ -\frac{1}{2}\beta^\top V^{-1}\beta\Big\} d\beta.$$

Taking the form $V = c(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i)^{-1}$, the description length $L(y|c)$ is given by

$$\frac{p}{2}\log(1+c) + \frac{1}{2}\Bigg[\sum_{i=1}^{n} y_i^\top \Sigma_i^{-1} y_i \\ - \frac{c}{1+c}\Big(\sum_{i=1}^{n} y_i^\top \Sigma_i^{-1} X_i\Big)\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big)^{-1}\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} y_i\Big)\Bigg].$$

Minimizing the above expression w.r.t. $c \geq 0$ yields

$$\hat{c} = \max\Bigg(\Big(\sum_{i=1}^{n} y_i^\top \Sigma_i^{-1} X_i\Big)\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big)^{-1}\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} y_i\Big)p^{-1} - 1, 0\Bigg).$$

Substituting $\hat{c}$ with an additional length $(\log n)/2$ for the hyperparameter leads to (2.1).

Derivation of lMDL (2.2). The marginal distribution of y is

$$m(y|a,c,D^*) = \iint f_{\beta,\tau}(y)\omega(\beta,\tau|D^*)d\beta d\tau = \int \Big\{\int f_{\beta,\tau^2}(y)\omega_{D^*}(\beta|\tau)d\beta\Big\}\omega(\tau)d\tau \\ = \left(\prod_{i=1}^{n} \frac{1}{(2\pi)^{n_i/2}|W_i|^{1/2}}\right) \frac{|(\sum_{i=1}^{n} X_i^\top W_i^{-1} X_i + V^{-1})^{-1}|^{1/2}}{|V|^{1/2}} \\ \times \frac{\sqrt{a}}{\sqrt{2\pi}}\left(\frac{RSS_V + a}{2}\right)^{-(N+1)/2} \Gamma\left(\frac{N+1}{2}\right),$$

where $N = \sum_{i=1}^{n} n_i$ is the total number of observations and $RSS_V = \sum_{i=1}^{n} y_i^\top W_i^{-1} y_i - (\sum_{i=1}^{n} y_i^\top W_i^{-1} X_i)(\sum_{i=1}^{n} X_i^\top W_i^{-1} X_i + V^{-1})^{-1}(\sum_{i=1}^{n} X_i^\top W_i^{-1} y_i)$. Ignoring the terms that do not depend on the choice of the model, we have the code length function

$$L(y|a,c,D^*) = \frac{1}{2}\sum_{i=1}^{n}\log|W_i| + \frac{1}{2}\log|\sum_{i=1}^{n} X_i^\top W_i^{-1}X_i + V^{-1}|$$
$$+\frac{1}{2}\log|V| - \frac{1}{2}\log a + \frac{N+1}{2}\log(RSS_V + a).$$

Minimizing this with respect to a yields the solution $\hat{a} = RSS_V/N$. If one assumes that $V = c(\sum_{i=1}^{n} X_i^\top W_i^{-1}X_i)^{-1}$, the description length function is

$$L(y|c,D^*) = \frac{1}{2}\sum_{i=1}^{n}\log|W_i| + \frac{p}{2}\log(1+c) + \frac{N}{2}\log\Big(\sum_{i=1}^{n} y_i^\top W_i^{-1}y_i - \frac{c}{1+c}FSS_W\Big).$$

In turn, we find the value of c to minimize this,

$$\hat{c} = \max\left(\frac{(N-p)FSS_W}{pRSS_W} - 1, 0\right).$$

Insert the expression of $\hat{c}$ here to get the shortest description length

$$L(y|D^*) = \frac{1}{2}\sum_{i=1}^{n}\log|W_i| + \frac{p}{2}\log\frac{(N-p)FSS_W}{pRSS_W} + \frac{N}{2}\log\frac{N\cdot RSS_W}{N-p}.$$

Substituting an estimate of D^* with an additional length $s\log n/2$, where s is the number of unknown parameters in D^*, achieves the lMDL (2.2).

A.2. Technical proofs of main theorems

Proof of Theorem 1. In our case $Q(y) = m(y)$ and $\theta = \beta$. Recall (2.1) and $FSS_\sigma = (\sum_{i=1}^{n} y_i^\top \Sigma_i^{-1}X_i)(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1}X_i)^{-1}(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1}y_i)$. By Jensen's Inequality, the numerator in (3.1) is bounded by

$$2E_\beta[\log f_\beta(y) - \log m(y)] = E_\beta\Big[l\mathrm{MDL}_1 - \sum_{i=1}^{n}(y_i - X_i\beta)^\top \Sigma_i^{-1}(y_i - X_i\beta)\Big]$$
$$= pE_\beta\Big[\log\Big(\frac{FSS_\sigma}{p}\Big)\Big] \le p\log\Big[\frac{E_\beta(FSS_\sigma)}{p}\Big].$$

We then simplify $EFFS_\sigma$ as follows, where $\mathrm{tr}(\cdot)$ denotes the trace,

$$E_\beta\Big[\hat{\beta}_{GLS}^\top\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1}X_i\Big)\hat{\beta}_{GLS}\Big]$$
$$= E_\beta\Big\{\mathrm{tr}\Big[\hat{\beta}_{GLS}^\top\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1}X_i\Big)\hat{\beta}_{GLS}\Big]\Big\}$$

$$= \mathrm{tr}\Big[\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big) E_\beta(\hat\beta_{GLS}\hat\beta_{GLS}^\top)\Big]$$
$$= \mathrm{tr}\Big[\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big)\Big(\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big)^{-1} + \beta\beta^\top\Big)\Big]$$
$$= \mathrm{tr}\Big[I_p + \Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big)\beta\beta^\top\Big] = p + \beta^\top\Big(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i\Big)\beta.$$

The assumption $n^{-1}\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i \to C$ for some positive definite matrix C implies that $\beta^\top(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i)\beta = O(n)$. Then

$$\liminf_{n\to\infty} \frac{E_\beta[\log f_\beta(y) - \log m(y)]}{(p/2)\log n}$$
$$\le \liminf_{n\to\infty} \frac{(p/2)\log(1 + \beta^\top(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i)\beta/p)}{(p/2)\log n} = 1,$$

together with (3.1), leads to the proposition.

Proof of Theorem 2. Since our interest is the marginal distribution of y, we simplify the model to $y_i = X_i\beta + \epsilon_i$, where $\epsilon_i = Z_i\mathbf{b}_i + \varepsilon_i$. Since W_i is positive definite, we can find $M_i = M_i^\top$ such that $W_i = M_i^2$. Then we can consider the problem as a simple linear regression of $M_i^{-1}X_i$ on $M_i^{-1}y_i$, $M_i^{-1}y_i = M_i^{-1}X_i\beta + M_i^{-1}\epsilon_i$. Let $y_i^* = M_i^{-1}y_i$, $X_i^* = M_i^{-1}X_i$ and $\epsilon_i^* = M_i^{-1}\epsilon_i$. Here $\mathrm{var}(\epsilon_i^*) = \sigma^2 I_{n_i}$. Let $W = \mathrm{diag}(W_1, \ldots, W_n)$, $M = \mathrm{diag}(M_1, \ldots, M_n)$, $X^* = (X_1^{*\top}, \ldots, X_n^{*\top})^\top$, and $y^* = (y_1^{*\top}, \ldots, y_n^{*\top})^\top$. Following Breiman and Freedman (1983), let

$$R_{np} = \frac{1}{N-p}\sum_{i=1}^{n}\sum_{k=1}^{m}(y_{ik}^* - \hat y_{ik}^*)^2 = \frac{RSS_W}{N-p},$$

where $\hat y_{ik}^* = \sum_{j=1}^{p} x_{ijk}^*\hat\beta_j$. Furthermore, let $\delta_i^* = \sum_{j=p+1}^{\infty} X_{ij}^*\beta_j$ and $S = S(n,p)$ $= ||(I-H)y^*||^2 = (N-p)R_{np}$, where $H = X^*(X^{*T}X^*)^{-1}X^{*T}$. Then

$$S = ||(I-H)(\epsilon^* + \delta^*)||^2 = S_1 + S_2 + S_3,$$

where

$$S_1 = ||(I-H)\epsilon^*||^2, \quad S_2 = ||(I-H)\delta^*||^2, \quad S_3 =< (I-H)\epsilon^*, (I-H)\delta^* > .$$

Since ϵ^* and δ^* are independent of each other and also independent of $(I-H)$, the cross term S_3 can be ignored. It is obvious that $(I-H)$ is an idempotent matrix with trace equal to $(N-p)$. We follow the arguments for Lemma 2.2–2.5 in Breiman and Freedman (1983). We have $S_1 = \sigma^2\sum_{j=1}^{N-p} Z_j^2 \doteq (N-p)\sigma^2$, where Z_j's are i.i.d. standard normal, and "$\doteq$" signifies asymptotic equivalence in the sense of Breiman and Freedman (1983). Without loss of generality, the

X_{ij} can be regarded as i.i.d. standard normal. If the jth column of X is $X^{(j)}$, the second term can be written as

$$S_2 = ||(I-H)\delta^*||^2 = \delta^{*T}(I-H)\delta^* = \sum_{j=p+1}^{\infty} \beta_j^2 X^{(j)T} M^{-1}(I-H)M^{-1}X^{(j)}.$$

Since M is of full rank and $\mathrm{rank}(I-H) = N-p$, $S_2 \doteq \sigma_p^2 \sum_{j=i}^{N-p} \tau_j Z_j^2$, where τ_j's are the eigenvalues of $M^{-1}(I-H)M^{-1}$. Given $(N-p) = O(n)$ and the fact that all the matrices are blocked and each block possesses identical distributions, one has $\mathrm{tr}\{M^{-1}(I-H)M^{-1}\} \doteq (N-p)c_p$ for some $c_p > 0$. Consequently $S_2/(N-p) \doteq c_p\sigma_p^2$. As for the term $FSS_W = ||Hy^*||^2$,

$$\begin{aligned} FSS_W = ||y^*||^2 + ||Hy^*||^2 - ||y^*||^2 &= \Big\| \sum_{j=1}^{\infty} \beta_j^2 X^{(j)*} + \epsilon^* \Big\|^2 - (N-p)RSS_W \\ &\doteq Nc_0\sigma_0^2 + N\sigma^2 - (N-p)\sigma^2 - (N-p)c_p\sigma_p^2 \\ &= p\sigma^2 + Nc_o\sigma_0^2 - Nc_p\sigma_p^2 + pc_p\sigma_p^2, \end{aligned}$$

where $\mathrm{tr}(W^{-1}) \doteq Nc_0$ for some $c_0 > 0$. Therefore, we have

$$\frac{(N-p)FSS_W}{pRSS_W} = \Big[\frac{N}{p}\frac{c_0\sigma_0^2 - c_p\sigma_p^2 + (p/N)(c_p\sigma_p^2+\sigma^2)}{c_p\sigma_p^2+\sigma^2}\Big](1+o_p(1)).$$

Noticing $FSS_\sigma = FSS_W/\sigma^2$ leads to

$$FSS_\sigma = \frac{N}{\sigma^2}\Big[c_0\sigma_0^2 - c_p\sigma_p^2 + \frac{p}{N}(c_p\sigma_p^2+\sigma^2)\Big](1+o_p(1)),$$

and this completes the proof.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267-281. Akademiai Kiado, Budapest.

Barron, A., Rissanen, J. and Yu, B. (1998). Minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44**, 2743-2760.

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069-1077.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78**, 131-136.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach.* Springer, New York.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* Wiley, New York.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B* **66**, 165-185.

Craiu, R. and Lee, T. C. (2005). Model selection for the competing risks model with and without masking. *Technometrics* **47**, 457-467.

Ghosh, S. K., Bhave, P. V., Davis, J. M. and Lee, H. (2010). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *J. Amer. Statist. Assoc.* **105**, 538-557.

Hansen, M. and Yu, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.* **96**, 746-774.

Hansen, M. and Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. In *Statistics and Science: A Festschrift for Terry Speed*, 145-163.

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

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

Lee, T. C. (2000). A minimum description length based image segmentation procedure, and its comparison with a cross-validation based segmentation procedure. *J. Amer. Statist. Assoc.* **95**, 259-270.

Lee, T. C. (2001). An introduction to coding theory and the two-part minimum description length principle. *Internat. Statist. Rev.* **69**, 169-183.

Lu, Q., Lund, R. and Lee, T. C. (2010). An MDL approach to the climate segmentation problem. *Ann. Appl. Statist.* **4**, 299-319.

Morrel, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* **54**, 1560-1568.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-765.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13-27.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465-471.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080-1100.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore.

Schwarz, G. (1978). Estimating the dimension of model. *Ann. Statist.* **6**, 461-464.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379-423.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-370.

van der Laan, M., Dudoit, S. and Keles, S. (2004). Asymptotic optimality of likelihood based cross-validation. *Statist. Appl. in Genetics and Molecular Biology* **3**, Article 4.

Wiencierz, A., Greven, S. and Küchenhoff (2011). Restricted likelihood ratio testing in linear mixed models with general error covariance structure. *Electronic J. Statist.* **5**, 1718-1734.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937-950.

Department of Statstics, University of Toronto, Toronto, Ontario M5S 3G3, Canada.
E-mail: lili@utstat.toronto.edu

Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada.
E-mail: fyao@utstat.toronto.edu

Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada.
E-mail: craiu@utstat.toronto.edu

Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada.
E-mail: jialin@utstat.toronto.edu

(Received September 2011; accepted July 2013)