The E-MS Algorithm: Model Selection with Incomplete Data

JIMING JIANG, THUAN NGUYEN AND J. SUNIL RAO
University of California, Davis, Oregon Health and Science University and University of Miami

We propose a procedure associated with the idea of the E-M algorithm for model selection in the presence of missing data. The idea extends the concept of parameters to include both the model and the parameters under the model, and thus allows the model to be part of the E-M iterations. We develop the procedure, known as the E-MS algorithm, under the assumption that the class of candidate models is finite. Some special cases of the procedure are considered, including E-MS with the generalized information criteria (GIC), and E-MS with the adaptive fence (AF; Jiang *et al.* 2008). We prove numerical convergence of the E-MS algorithm as well as consistency in model selection of the limiting model of the E-MS convergence, for E-MS with GIC and E-MS with AF. We study the impact on model selection of different missing data mechanisms. Furthermore, we carry out extensive simulation studies on the finite-sample performance of the E-MS with comparisons to other procedures. The methodology is also illustrated on a real data analysis involving QTL mapping for an agricultural study on barley grains.

Key Words. backcross experiments, conditional sampling, consistency, convergence, missing data mechanism, model selection, regression

1 Introduction

The missing-data problem has a long history (e.g., Afifi and Elashoff 1966, Hartley and Hocking 1971). While there is an extensive literature on statistical analysis with missing or incomplete data (e.g., Rubin 1976, Dempster *et al.* 1977, Robins *et al.* 1995, Rotnitzky

et al. 1998, Little & Rubin 2002), the literature on model selection in the presence of missing data is relatively sparse. Existing model selection procedures face special challenges when confronted with missing or incomplete data. Obviously, the naive complete-data-only strategy is inefficient, sometimes even unacceptable by the practitioners due to the overwhelmingly wasted information. For example, in a study of backcross experiments (e.g., Lander and Botstein 1989, Zeng 1993, Jansen 1993, Broman and Speed 2002), a data set was obtained by researchers at UC-Riverside (personal communications; see Zhan *et al.* 2011 for a related work). Out of the 150 or so subjects, only 4 have complete data record. Situations like this are, unfortunately, the reality that we often have to deal with, and the main motivation for this research project.

Fuchs (1982) proposed to use the E-M algorithm (Dempster *et al.* 1977) for the ML estimation under a log-linear model with missing data, and then test for goodness-of-fit based on the ML estimation in order to choose an appropriate model. Motivated by the predictive divergence for incomplete observation models (PDIO; Shimodaira 1994), Cavanaugh and Shumway (1998) derived an AIC for model selection in the presence of incomplete data. A similar approach was considered by Seghouane *et al.* (2005), in which the authors obtained an unbiased estimator of the complete-data Kullback-Leibler symmetric divergence. Bueso *et al.* (1999) used the E-M algorithm to compute the minimum description length (MDL; Rissanen 1983) for model selection, when only incomplete data are available. Sebastiani and Ramoni (2001) discussed a Bayesian approach for the selection of decomposable models by maximizing the posterior probability of a candidate model, and showed how to do this with incomplete data. Hens *et al.* (2006) considered a modification of the AIC based on reweighting incomplete and design-based samples. Claeskens and Consentino (2008) proposed some variations on the AIC based on the output of the E-M algorithm. The method is applicable to model selection problems with missing covariates, but the response variable is assumed to be fully observed. Schomaker *et al.* (2010) considered two approaches of

handling the missing data in determining the weights in frequentist model averaging. The first is based on adjusting an existing criterion; while the second uses the unadjusted criterion but with the missing data replaced by their imputed values. Verbeke *et al.* (2008) offered a review of formal and informal model selection strategies with incomplete data, but the focus is on model comparison, instead of model selection. As noted by Ibrahim *et al.* (2008), while model comparisons "demonstrate the effect of assumptions on estimates and tests, they do not indicate which modeling strategy is best, nor do they specifically address model selection for a given class of models". The latter authors further proposed a class of model selection criteria based on the output of the E-M algorithm. Also see Garcia *et al.* (2010). A potential drawback with the E-M approach of Ibrahim *et al.* (2008) is that the conditional expectation in the E-step is taken under the assumed (candidate) model, rather than an objective (true) model. Note that the complete-data log-likelihood is also based on the assumed model. Thus, by taking the conditional expection, again, under the assumed model, it may bring false supporting evidence for an incorrect model. The problem is sometimes referred to as "double-dipping". We illustrate this with an example.

Example 1. Suppose that one attempts to select a logistic model, $\mathrm{logit}(p_i) = x_i'\beta$, where $p_i = \mathrm{P}(Y_i = 1)$, $Y_1, \ldots, Y_n$ being independent, binary, observations, and x_i is a vector of covariates to be selected. Suppose that $y_1, \ldots, y_5$ are observed, and the rest of the y_i's are missing. Also, for simplicity, assume that all the x_i's are observed. The derivation below in this paragraph is based on MAR (Rubin 1976) for simplicity. Let M_0 denote the intercept only model and suppose that the true model is not M_0. The complete-data log-likelihood under M_0 is $l = \sum_{i=1}^{n}\{y_i \log(p_0) + (1-y_i)\log(1-p_0)\}$, where $p_0 = e^{\beta_0}/(1+e^{\beta_0})$ and β_0 is the intercept. Note that, under M_0, we have $\mathrm{E}(l|y_1, \ldots, y_5, \text{ all } x_i's) = \sum_{i=1}^{5}\{y_i \log(p_0) + (1-y_i)\log(1-p_0)\} + (n-5)\{p_0 \log(p_0) + (1-p_0)\log(1-p_0)\}$. If $y_i = 1, 1 \le i \le 5$, then, as $p_0 \to 1$, we have $\mathrm{E}(l|y_1, \ldots, y_5, \text{ all } x_i's) \to 0$. On the other hand, under any other model, M, the corresponding log-likelihood is $l = \sum_{i=1}^{n}\{y_i \log(p_i) + (1-y_i)\log(1-p_i)\} \le 0$,

hence $\mathrm{E}(l|y_1,\ldots,y_5,\ \text{all}\ x_i's) \leq 0$, under M. This means that the maximized conditional expectation of l under M_0 (which is 0) is greater than or equal to the maximized conditional expectation of l under M (which is less than or equal to 0). Thus, the first term of any information criterion under M_0 is less than or equal to that under M. On the other hand, M_0 certainly has the smallest dimension. Therefore, M_0 will be selected as the optimal model by the IC criteria of Ibrahim *et al.* (2008), which, of course, is an incorrect model.

To further illustrate numerically, we carry out a simulation study under the following specific setting. Suppose that the candidate covariates include a continuous variable, x_1, whose values are generated from the standard normal distribution, and a binary indicator, x_2, whose values are generated from the Bernoulli(0.5) distribution. The following candidate models are considered: Model 0: $x_i'\beta = \beta_0$, Model j: $x_i'\beta = \beta_0 + \beta_j x_{ji}, j = 1, 2$, and Model 3: $x_i'\beta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. Two scenarios are considered. In the first scenario, Model 1 is the true underlying model with the true parameters $\beta_0 = \beta_1 = 1$; in the second scenario, Model 3, which is the full model, is the true underlying model with the true parameters $\beta_0 = \beta_1 = 1, \beta_2 = -1$. Furthermore, the missing data indicators, M_i, which is 1 if y_i is missing, and 0 otherwise, are generated either under an *ignorable* mechanism, in which case $\mathrm{P}(M_i = 1|y) = 0.5$ (case A), or under a *non-ignorable* mechanism, in which case $\mathrm{P}(M_i = 1|y) = h(\psi_0 + \psi_1 y_i)$ with $h(x) = e^x/(1 + e^x)$ and the true parameters $\psi_0 = 0.5$ and $\psi_1 = 0.2$ (case B). See Section 6 for more details. We apply the method of Ibrahim *et al.* (2008) with the BIC penalty, denoted by IZT, under two different sample sizes, $n = 50$ and $n = 100$. A comparing method, which is what we are going to propose in this paper, called E-MS (to be introduced in the next section), here in conjunction with the BIC, is also applied to the same simulated data. Results of the empirical true positive (TP, i.e., the selected model is exactly the true underlying model) rates, based on 1,000 simulations, are reported in Table 1. It is seen that IZT performs considerably worse than E-MS under all scenarios, cases, and sample sizes. Note that both methods perform

Table 1: **Empirical TP for Logistic Model Selection**

Missing Data	Sample	True Model $=$ Model 1		True Model $=$ Model 3	
Mechanism	Size	E-MS	IZT	E-MS	IZT
Case A	$n = 50$	0.787	0.483	0.213	0.136
	$n = 100$	0.965	0.738	0.467	0.216
Case B	$n = 50$	0.837	0.395	0.169	0.097
	$n = 100$	0.970	0.607	0.459	0.160

much worse under Model 3 than under Model 1, which is not surprising–the BIC is known to over-penalize "larger" models, especially the full model (e.g., Jiang *et al.* 2008). Furthermore, the performance of E-MS does not seem to be affected by the different missing data mechanisms (see Section 6 for more discussion), while IZT appears to perform worse under the non-ignorable missing data setting (case B).

2 Outline of our main contributions

The strategic failure as illustrated by Example 1 is due to the double use of the assumed model, once in the measure of lack-of-fit (i.e., the negative log-likelihood) and once in the conditional expectation of this measure. Note that the assumed model is not necessarily the true model, so the conditional expectation under the assumed model is not necessarily the true conditional expectation. As mentioned, this may bring false evidence in favor of an incorrect model, and, by doing so, the E-M loses its "updating power" when applied to model selection problems. In fact, the assumed model should be treated the same way as the unknown parameters (the model and the parameters under the model together completely specify "the model"), so it is not reasonable to update only the parameters.

Note that the double usage of the assumed model has been shown in the literature to have serious consequences. For example, Copas and Eguchi (2005) discuss a similar issue that they term as *incomplete-data bias*, in which the maximum likelihood estimators can be (sometimes severely) biased when incomplete data are present, and an incorrect model is being fit, and yet still appears to give a good fit to the available data. Jiang *et al.* (2011a) showed that if one derives the parameter estimators by evaluating the best predictor (BP) under the assumed model, say, M, using the distribution also under M, the resulting predictor is not robust in the sense that it may perform poorly when M is not the true model. Here, the failure of the BP is due to a similar double-dipping strategy, that is, (1) the measure of lack-of-fit (sum of squared prediction errors), is for the BP under M; and (2) the distribution under which the measure of lack-of-fit is evaluated is also under on M.

In this paper, we propose a general strategy for model selection in the presence of incomplete or missing data that can be used with any existing model selection procedure that is designed for a complete data situation. Our strategy is based on the E-M idea; however, unlike Ibrahim *et al.* (2008), the conditional expectation is evaluated under an objective model, which is the same for all the candidate models. A key idea is to include the model, as well as the parameters, in the E-M iteration, and the objective model, under which the conditional expectations are evaluated in the E-step, is the current model. Another main contribution of the current paper is that we establish theoretical properties of the proposed E-MS algorithm, including the (numerical) convergence of the algorithm, and consistency of the limiting model of the E-MS convergence in terms of model selection. We also investigate, from a theoretical standpoint, the impact of the missing data mechanism (MDM, e.g., Little & Rubin 2002) on the performance of the E-MS. Furthermore, we provide empirical evidence, in terms of simulation studies and real data analysis, that support the theoretical findings. More specifically, the simulation results compare the finite-sample performance of the E-MS with existing, ad-hoc, or "ideal" procedures. We consider various scenarios

in our simulation studies, such as different types of MDMs, and the situation that the true model is not among the candidate models.

It should be noted that, for the most part, there are three major approaches for model selection, namely, the information criteria or, more generally, generalized information criteria (GIC; e.g., Nishii 1984, Shibata 1984), the shrinkage methods (Tibshirani 1996, Fan & Li 2001, among others), and the fence methods (Jiang *et al.* 2008). See, for example, a recent review by Müller, Scealy and Welsh (2013). However, for the shrinkage methods, E-MS is the same as the E-M algorithm. This is because the shrinkage methods combine variable selection with estimation of the corresponding coefficients (the variables with zero estimated coefficients are dropped from the current model). Thus, updating the model is the same as updating the parameter estimates; or, from another point of view, the model does not change with the iteration–it is always the full model. Therefore, in the subsequent development we shall use GIC and the fence as main examples to illustrate our method.

It should also be pointed out that the current development is under the assumption that the class of candidate models is finte. Therefore, the methodology may not be applicable if the model space is infinite dimensional, such as in semi-parametric modeling.

Following the general convention, throughout this paper we use capital letters, e.g., Y, for a random variable, or random vector, and small letters, e.g., y, for the observed, or realized, value of Y (the only exception is when the observed values or realized values are entries of a matrix, which, as usual, is denoted with a capital letter).

3 The E-MS algorithm

The E-M is well known for parameter estimation in the presence of missing data. On the other hand, model selection, as another component of model identification, may also be viewed as parameter estimation, with the parameter being [the identification (ID) number

of] the model and the parameter space being the (ID numbers of the) model space. Namely, we combine the parameters with the model under which the parameters are defined. So, at the current stage of the iteration, we have the current model, $M_{\rm c}$, as well as the current estimates of the parameters, $\hat{\theta}_{\rm c}$, under $M_{\rm c}$. Let $Q(M) = Q(Y, M, \theta_M)$ be a measure of lack-of-fit, where Y represents the complete data, M a candidate model, and θ_M the vector of parameters under M. We take the conditional expectation of $Q(M)$ under $M_{\rm c}$, with the parameters under $M_{\rm c}$, $\theta_{M_{\rm c}} \equiv \theta_{\rm c}$, being $\hat{\theta}_{\rm c}$, given the observed data, $y_{\rm o}$, denoted by $\mathrm{E}_{\rm c}\{Q(M)|y_{\rm o}\}$. This is the E-step.

In the next step, we carry out model selection using $\mathrm{E}_{\rm c}\{Q(M)|y_{\rm o}\}$ as the measure of lack-of-fit. To do so, we first find $\hat{Q}_{\rm c}(M) = \inf_{\theta_M \in \Theta_M} \mathrm{E}_{\rm c}\{Q(M)|y_{\rm o}\}$, where Θ_M is the parameter space under M. We can use $\hat{Q}_{\rm c}(M)$ in a GIC setting, in which the optimal model, $\hat{M}_{\rm opt}$, is found by minimizing $\hat{Q}_{\rm c}(M) + \lambda_n |M|$ over $M \in \mathcal{M}$, the class of candidate models, where λ_n is a penalty that depends on the sample size, n, and $|M|$ is the dimension of M. Alternatively, we may use the fence method (Jiang *et al.* 2008) based on $\hat{Q}_{\rm c}(M)$. This is the MS-step, where MS stands for "model selection". We then replace $M_{\rm c}$ by $\hat{M}_{\rm opt}$, found in the MS-step, and $\hat{\theta}_{\rm c}$ by $\hat{\theta}_{\rm opt}$, where $\hat{\theta}_{\rm opt}$ is the parameter vector under $\hat{M}_{\rm opt}$ corresponding to the minimizer of $\mathrm{E}_{\rm c}\{Q(\hat{M}_{\rm opt})|y_{\rm o}\}$ over $\theta_{\hat{M}_{\rm opt}} \in \Theta_{\hat{M}_{\rm opt}}$, and return to the E-step. We illustrate the E-MS procedure with some examples.

Example 2 (Backcross experiments). Quantitative trait loci (QTL) mapping in genetics has been extensively studied (e.g., Lander and Botstein 1989, Zeng 1993, Jansen 1993). More recently, Broman and Speed (2002) modified the BIC and applied it to QTL mapping in backcross experiments. The method is for complete-data analysis only. In practice, however, missing data are often present. For example, as mentioned earlier, in the data set obtained for backcross experiments by the researchers at UC-Riverside, less than 3% of the data have the complete records, that is, without the missing values.

Following Broman and Speed (2002), we have a conditional linear regression model

for the phenotype variable, Y, such that, given the marker indicators, x, we have $Y_i = \sum_{k=1}^{r} \sum_{j \in M_k} \beta_{jk} x_{ijk} + \epsilon_i$, where r is the number of chromosomes, M_k is a subset of $\{1, \ldots, q\}$ and q is the number of markers on each chromosome, and ϵ_i is a normal error, with mean zero and unknown variance σ^2. The ϵ_i's are uncorrelated and also independent with the X_{ijk}'s. Furthermore, the marker indicators, X_{ijk}, are assumed to be a Markov chain within each chromosome with $\mathrm{P}(X_{i1k} = 0) = \mathrm{P}(X_{i1k} = 1) = 1/2$ (Mendel's rule) and $\mathrm{P}(X_{i,j+1,k} = 1 | X_{ijk} = 0) = \mathrm{P}(X_{i,j+1,k} = 0 | X_{ijk} = 1) = \theta$, where θ is the *recombination fraction*. The problem of interest is to identify the subset $M = (M_1, \ldots, M_r)$, which is viewed as a model selection problem as in Broman and Speed (2002).

We consider the E-MS in conjunction with the BIC procedure. Due to the high dimensionality, we consider the forward/backward (FW/BW) BIC procedure of Broman and Speed (2002). A detailed description of the latter is given in the Supplementary Material (Section A.3). The log-likelihood, under a given model, M, can be expressed as $l_M = l_{M,y|x} + l_x$, where l_x does not depend on the model, $l_{M,y|x} = c - 0.5\{n \log \sigma^2 + \sigma^{-2} \sum_{i=1}^{n} (y_i - - x'_{M,i} \beta_M)^2\}$, c being a constant. Thus, we have $\mathrm{BIC}(M) = -2\hat{l}_M + |M| \log(n)$, where $\hat{l}_M$ is the maximized l_M (over the parameters). It is easy to show that the MLE of θ, $\hat{\theta}$, is the same as the maximizer of l_x, which does not depend on M. Thus, we have

$$\mathrm{BIC}(M) = -2\hat{l}_{M,y|x} + |M| \log(n) - 2\hat{l}_x \propto -2\hat{l}_{M,y|x} + |M| \log(n). \tag{1}$$

In addition, the FW/BW requires evaluation of $\mathrm{RSS}(y, X) = \min_\beta \mathrm{RSS}(y, X, \beta)$, where

$$\mathrm{RSS}(y, X, \beta) \;=\; \sum_{i=1}^{n} (y_i - x'_i \beta)^2 \tag{2}$$

with $X = (x'_i)_{1 \le i \le n}$. Because both (1) and (2) involve missing data, we replace them by their conditional expectations under the current model, M_c, and the current parameter

estimates under $M_{\rm c}$, before the minimization/maximization. This leads to

$$\mathrm{RSS}_{\rm c}(M|y_{\rm o}, x_{\rm o}) \equiv \min_{\beta} \mathrm{E}_{\rm c}\{\mathrm{RSS}(Y, X, \beta)|y_{\rm o}, x_{\rm o}\}, \tag{3}$$

$$\mathrm{BIC}_{\rm c}(M|y_{\rm o}, x_{\rm o}) \equiv -2 \max_{\beta_M, \sigma^2} \mathrm{E}_{\rm c}(l_{M,Y|X}|y_{\rm o}, x_{\rm o}) + |M| \log(n), \tag{4}$$

both of which have closed-form expressions (see Subsection A.4.1 of the Supplementary Material), where $x_{\rm o}$ and $y_{\rm o}$ denote the observed x's and y's, respectively.

In summary, given $M_{\rm c}$ and the current parameter estimates, the FW/BW, based on (3), is used to generate a sequence of models; the BIC, based on (4), is then applied to the sequence generated by the FW/BW to update the model as well as parameter estimates.

A reasonable initial model is the full model, $M_{\rm f}$. A reasonable initial estimator for θ is $\hat{\theta}_0$ = proportion of observed cases in which x_{ijk} and $x_{i,j+1,k}$ are different. As for the initial estimator of $\beta_{\rm f}$, the vector of regression coefficients under $M_{\rm f}$, note that the idea of least squares (LS) fit in regression is to find the parameter estimates that minimizes $\sum_{i=1}^{n}\{y_i - \mathrm{E}_{\rm f}(Y_i|x)\}^2$, where $\mathrm{E}_{\rm f}$ denotes expectation under $M_{\rm f}$. Due to the missing data, it is natural to replace this by $\sum_{i\in I_{\rm o}}\{y_i - \mathrm{E}_{\rm f}(Y_i|x_{\rm o})\}^2$, where $I_{\rm o}$ denotes the subset of indexes i so that y_i is observed. Furthermore, we have $\mathrm{E}_{\rm f}(Y_i|x_{\rm o}) = \sum_{k=1}^{r}\sum_{j=1}^{q}\beta_{jk}\mathrm{E}_{\rm f}(X_{ijk}|x_{{\rm o},i})$, where $x_{{\rm o},i}$ denotes the observe x's for the ith subject; $\mathrm{E}_{\rm f}(X_{ijk}|x_{{\rm o},i}) = x_{ijk}$ if the latter is observed, and an expression of the conditional expectation can be easily obtained, with θ replaced by $\hat{\theta}_0$, if x_{ijk} is missing. We then run the LS with $y_i, i \in I_{\rm o}$ as the responses and $\mathrm{E}_{\rm f}(X_{ijk}|x_{{\rm o},i})$'s, $i \in I_{\rm o}$, as the predictors, to obtain the initial estimator $\hat{\beta}_{{\rm f},0}$, for $\beta_{\rm f}$. The initial estimator for σ^2, $\hat{\sigma}_0^2$, is the RSS of this LS fit divided by $|I_{\rm o}| - qr$.

Example 3 (Linear regression). The classical linear regression is a conditional model, in which the distribution of the covariates (or predictors) is not specified. As is well known, such a model may not directly work with the E-M algorithm, if some of the covariates are also missing. Little and Rubin (2002) proposed the following model for the joint distribution of the response and covariates in a linear regression. Suppose that the candidate

predictors can be listed as $x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q}$ such that $x_1, \ldots, x_p$ are continuous and $x_{p+1}, \ldots, x_{p+q}$ are discrete or categorical (in case there is an intercept, the corresponding constant, 1, is considered as the first discrete/categorical predictor). Furthermore, let $v_1, \ldots, v_s$ be all the possible (vector-valued) values for $x_{\rm d} = (x_{p+1}, \ldots, x_{p+q})'$. Let $x_{i,{\rm d}}$ be the $x_{\rm d}$ corresponding to the ith observation, and $x_{i,{\rm c}}$ be the vector $(x_1, \ldots, x_p)'$ corresponding to the ith observation, and $x_i = (x'_{i,{\rm c}}, x'_{i,{\rm d}})'$. The assumptions are: (i) $Y_i, X_i, i = 1, \ldots, n$ are independent; (ii) for each i, $X_{i,{\rm d}}$ has the probability distribution $\mathrm{P}(X_{i,{\rm d}} = v_r) = \pi_r, 1 \le r \le s$, where the π_r's are unknown probabilities such that $\sum_{r=1}^s \pi_r = 1$; (iii) given $X_{i,{\rm d}} = v_r$, $X_{i,{\rm c}}$ has a multivariate normal distribution with mean μ_r and covariance matrix Ω, where $\mu_r, 1 \le r \le s$ are unknown vectors, and Ω is an unknown covariance matrix that does not depend on r; and (iv) given x_i, Y_i is normal with mean $x_i'\beta$ and variance σ^2, where β is an unknown $(p+q)$-dimensional vector of regression coefficients, and σ^2 is an unknown variance. These assumptions are for the full model. More generally, we are interested in a model, M, for the conditional distribution (iv). Write $x_{i,M} = (x'_{i,M,{\rm c}}, x'_{i,M,{\rm d}})'$, and $\beta_M = (\beta'_{M,{\rm c}}, \beta'_{M,{\rm d}})'$. Then, under M, (iv) is replaced by (iv-M) given x_i, $Y_i \sim N(x'_{i,M}\beta_M, \sigma^2)$. The parts (i)–(iii) of the model are unchanged.

Let $y, x, x_{\rm c}, x_{\rm d}$ denote the data for the $y_i, x_i, x_{i,{\rm c}}, x_{i,{\rm d}}$, respectively, across $1 \le i \le n$. Then, it can be shown that the complete-data log-likelihood has the expression

$$\begin{aligned} l &= c - \frac{n}{2}(\log\sigma^2 + \log|\Omega|) + \sum_{r=1}^s n_r \log\pi_r - \frac{1}{2}\sum_{r=1}^s\sum_{i=1}^n 1_{(x_{i,{\rm d}}=v_r)} \\ &\quad \times (x_{i,{\rm c}} - \mu_r)'\Omega^{-1}(x_{i,{\rm c}} - \mu_r) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x'_{i,M}\beta_M)^2, \end{aligned} \tag{5}$$

where c is a constant. Note that the maximum likelihood is a constrained maximization problem, namely, $\max l$ subject to $\sum_{r=1}^s \pi_r = 1$. Define $\mathcal{L} = l + \lambda(\sum_{r=1}^s \pi_r - 1)$. Then, the MLE of the parameters, plus the Lagrange multiplier λ, is a stationary point of $\mathcal{L}$.

In Example 2 we considered the E-MS with BIC. To see an alternative, let us now consider the E-MS in conjunction with the adaptive fence (AF; Jiang *et al.* 2008). See

Jiang (2014) for a recent review on the fence methods. Take the initial model, M_0, as the full model, $M_{\rm f}$, and let $\beta_{\rm f}$ be the β under $M_{\rm f}$. Let $\mathrm{E_f}$ denote the conditional expectation under $M_{\rm f}$ and the current estimates of parameters, under $M_{\rm f}$, including $\beta_{\rm f}$, σ^2, $\mu_r, \pi_r, 1 \le r \le s$, and Ω. Let $y_{\rm o}, x_{\rm o}$ denote the observed y, x, respectively. By (5), with $M = M_{\rm f}$, we have

$$\bar{\mathcal{L}}_{\rm f} \equiv \mathrm{E_f}(\mathcal{L}|y_{\rm o}, x_{\rm o}) = c - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathrm{E_f}\left\{(Y_i - X_{i,\rm f}'\beta_{\rm f})^2|y_{\rm o}, x_{\rm o}\right\}, \quad (6)$$

where c does not depend on $\beta_{\rm f}$ and σ^2. From (6), we obtain the updates for $\beta_{\rm f}$ and σ^2,

$$\hat{\beta}_{\rm f} = S_2^{-1}S_1, \;\; \hat{\sigma}^2 = n^{-1}\{S_0 - S_1'S_2^{-1}S_1\}, \quad (7)$$

$S_0 = \sum_{i=1}^n \mathrm{E_f}(Y_i^2|y_{\rm o}, x_{\rm o})$, $S_1 = \sum_{i=1}^n \mathrm{E_f}(X_{i,\rm f}Y_i|y_{\rm o}, x_{\rm o})$, $S_2 = \sum_{i=1}^n \mathrm{E_f}(X_{i,\rm f}X_{i,\rm f}'|y_{\rm o}, x_{\rm o})$. Furthermore, we have (see Subsection A.4.2 of the Supplementary Material)

$$\hat{\mu}_r = \frac{\sum_{i=1}^n \mathrm{E_f}\{1_{(X_{i,\rm d}=v_r)}X_{i,\rm c}|y_{\rm o}, x_{\rm o}\}}{\sum_{i=1}^n \mathrm{P_f}(X_{i,\rm d} = v_r|y_{\rm o}, x_{\rm o})}, \; \hat{\pi}_r = \frac{\mathrm{E_f}(N_r|y_{\rm o}, x_{\rm o})}{\sum_{t=1}^s \mathrm{E_f}(N_t|y_{\rm o}, x_{\rm o})}, \; 1 \le r \le s, \quad (8)$$

$$\hat{\Omega} = \frac{1}{n}\sum_{r=1}^{s}\sum_{i=1}^{n}\mathrm{E_f}\{1_{(X_{i,\rm d}=v_r)}(X_{i,\rm c} - \hat{\mu}_r)(X_{i,\rm c} - \hat{\mu}_r)'|y_{\rm o}, x_{\rm o}\}. \quad (9)$$

It remains to evaluate the conditional expectations involved in (7)–(9). Let $y_{\rm m}$, $x_{\rm m}$, $x_{\rm c,m}$, and $x_{\rm d,m}$ denote the missing parts of y, x, $x_{\rm c}$, and $x_{\rm d}$, respectively. Although it is possible to obtain the conditional density $f_M(y_{\rm m}, x_{\rm m}|y_{\rm o}, x_{\rm o})$, the result is not a common distribution (e.g., normal), under which the conditional expectations can be easily obtained analytically. Alternatively, one may consider sampling from the conditional distribution, and use the Monte Carlo method to compute the conditional expectations. To do so, first note that it is easy to show that one can sample from the joint conditional distribution by sampling independently from the conditional distribution for each subject. To sample from the subject conditional distribution, note that $f_{M,i}(y_{i,\rm m}, x_{i,\rm m}|y_{i,\rm o}, x_{i,\rm o}) \propto f_{M,i}(y_i, x_i) \propto$

$$\exp\left[\sum_{r=1}^{s}1_{(x_{i,\rm d}=v_r)}\left\{\log\pi_r - \frac{1}{2}(x_{i,\rm c} - \mu_r)'\Omega^{-1}(x_{i,\rm c} - \mu_r)\right\} - \frac{(y_i - x_{i,M}'\beta_M)^2}{2\sigma^2}\right],$$

where $\propto$ means that the expression is up to a function of $y_{i,\mathrm{o}}, x_{i,\mathrm{o}}$, which is considered constant during the sampling of $y_{i,\mathrm{m}}, x_{i,\mathrm{m}}$. Next, we employ the Metropolized independence sampler (MIS, e.g., Liu 2004, p. 115), which is a special case of the Metropolis-Hastings algorithm. We refer the details to Subsection A.4.2 of the Supplementary Material.

The initial estimates of $\mu_r, 1 \le r \le s, \Omega, \pi_r, 1 \le r \le s$ are $\hat{\mu}_r^{(0)} = n_{r,\mathrm{o}}^{-1} \sum_{i \in I_{r,\mathrm{o}}} x_{i,c}$, $1 \le r \le s$, where $I_{r,\mathrm{o}} = \{1 \le i \le n : x_i \text{ is observed and } x_{i,\mathrm{d}} = v_r\}$, and $n_{r,\mathrm{o}} = |I_{r,\mathrm{o}}|$; $\hat{\Omega}^{(0)} = n_{\mathrm{o}}^{-1} \sum_{r=1}^{s} \sum_{i \in I_{r,\mathrm{o}}} \{x_{i,\mathrm{c}} - \hat{\mu}_r^{(0)}\}\{x_{i,\mathrm{c}} - \hat{\mu}_r^{(0)}\}'$, where $I_{\mathrm{o}} = \cup_{r=1}^{s} I_{r,\mathrm{o}}$ and $n_{\mathrm{o}} = |I_{\mathrm{o}}|$, and $\hat{\pi}_r^{(0)} = \#\{1 \le i \le n : x_{i,\mathrm{d}} \text{ observed and } x_{i,\mathrm{d}} = v_r\} / \#\{1 \le i \le n : x_{i,\mathrm{d}} \text{ observed}\}$, $1 \le r \le s$. Furthermore, the initial estimate of β_{f} is the LS estimate based on the all-observed data, that is, $\hat{\beta}_{\mathrm{f}}^{(0)} = (X_{\mathrm{ao}}' X_{\mathrm{ao}})^{-1} X_{\mathrm{ao}}' y_{\mathrm{ao}}$ (assuming, without loss of generality, that $X_{\mathrm{ao}}' X_{\mathrm{ao}}$ is nonsingular), where $X_{\mathrm{ao}} = (x_{\mathrm{f},i}')_{i \in I_{\mathrm{ao}}}$ with $I_{\mathrm{ao}} = \{1 \le i \le n : x_{\mathrm{f},i}, y_i \text{ observed}\}$, and $y_{\mathrm{ao}} = (y_i)_{i \in I_{\mathrm{ao}}}$. The initial estimate of σ^2 is $(\hat{\sigma}^2)^{(0)} = |y_{\mathrm{ao}} - X_{\mathrm{ao}} \hat{\beta}_{\mathrm{f}}^{(0)}|^2 / (|I_{\mathrm{ao}}| - p - q)$.

For any candidate model M, let $Q(M) = S_0 - S_1' S_2^{-1} S_1$, where S_0 is the same as that below (7), and $S_j, j = 1, 2$ are the same as those below (7) with $x_{i,\mathrm{f}}$ replaced by $x_{i,M}$. Note that the conditional expectation, E_{f}, will be done by the conditional sampling method mentioned above, with $M = M_0$. Run the AF, with $Q(M)$ being the measure of lack-of-fit. Denote the model selected by AF by $\hat{M}$. Let $\hat{\beta} = S_2^{-1} S_1$, where $S_j, j = 1, 2$ are given below (7) with $x_{i,\mathrm{f}}$ replaced by $x_{i,\hat{M}}$. Next, let $\hat{\sigma}^2$ be given by (7), where $S_j, j = 0, 1, 2$ are given below (7) with $x_{i,\mathrm{f}}$ replaced by $x_{i,\hat{M}}$. Also, let $\hat{\mu}_r, 1 \le r \le s, \hat{\Omega}, \hat{\pi}_r, 1 \le r \le s$ be given by (8), (9) (note that these depend only on $M_0 = M_{\mathrm{f}}$, but not on $\hat{M}$).

Replace M_0 by $\hat{M}$, and the initial estimates by $\hat{\beta}, \hat{\sigma}^2, \hat{\mu}_r, 1 \le r \le s, \hat{\Omega}, \hat{\pi}_r, 1 \le r \le s$, and repeat the process. Note that, after this iteration, the E_{f} is replaced by $\mathrm{E}_{\hat{M}}$, evaluated by the conditional sampling method with $M = \hat{M}$.

Keep updating the model and parameters iteratively until convergence (see below).

Note. The AF procedure is potentially time-consuming due to the need for bootstrapping (Jiang *et al.* 2008). In this regard, we refer to some recent development on improving

the computational efficiency of the AF. See Pang *et al.* (2013).

The convergence of the E-MS algorithm, as mentioned above, is a key theoretical issue that we address in the next section.

4 Convergence and consistency of E-MS

In this section, we state the results regarding two importantxi theoretical properties of the E-MS: The numerical convergence and consistency, in terms of model selection, of the limit of the E-MS convergence. We term the latter as consistency of the E-MS. The details, including proofs and interpretation of conditions, are deferred to Subsection A.1.3 of the Supplementary Material. Also, we shall focus on E-MS with GIC, and defer similar results for E-MS with AF to the same subsection in Supplementary Material.

The GIC, which include AIC, BIC, and other information criteria, is defined as

$$c(M, \theta, Y) \quad = \quad Q(M, \theta, Y) + p(M), \tag{10}$$

where Q is a measure of lack-of-fit that depends on M, a candidate model, θ, the parameter vector under M (strictly speaking, it should be denoted by θ_M; we suppress the subscript for notation simplicity), and Y, the vector of complete data, and $p(\cdot)$ is a penalty function on the complexity of M. If Y were observed, the model selection would be done by minimizing $c(M, \theta, Y)$, first over $\theta \in \Theta_M$, the parameter space under M, and then over $M \in \mathcal{M}$, the space of candidate models. Note that, we have

$$\min_{M\in\mathcal{M}} \min_{\theta\in\Theta_M} c(M, \theta, Y) = \min_{M\in\mathcal{M}} \left\{ \min_{\theta\in\Theta_M} Q(M, \theta, Y) + p(M) \right\} = \min_{M,\theta} c(M, \theta, Y), \tag{11}$$

where in the right side minimization, θ is confined to Θ_M. Because Y contains missing values, we cannot really do (11). Instead, we replace (10) by its conditional expectation, given the vector of observed data, y_{o}, under the current model, $M^{(t)}$, and the current parameter

vector, $\theta^{(t)}$, which is defined under $M^{(t)}$, that is,

$$\mathrm{E}\left\{ c(M,\theta,Y)|\, y_{\mathrm{o}}, M^{(t)}, \theta^{(t)} \right\} = \mathrm{E}\left\{ Q(M,\theta,Y)|\, y_{\mathrm{o}}, M^{(t)}, \theta^{(t)} \right\} + p(M). \tag{12}$$

(11) is then carried out with $c(M,\theta,Y)$ replaced by the right side of (12), or $Q(M,\theta,Y)$ replaced by $\mathrm{E}\{Q(M,\theta,Y)|y_{\mathrm{o}}, M^{(t)}, \theta^{(t)}\}$, resulting the minimizer $M^{(t+1)}$ and $\theta^{(t+1)}$.

Suppose that there is an observed version of (10), $g(M,\theta,y_{\mathrm{o}}) = Q_{\mathrm{o}}(M,\theta,y_{\mathrm{o}}) + p(M)$. Denote $\psi = (M,\theta)$, where θ is understood as the parameter vector under M. Let Ψ denote the model/parameter space for ψ. We assume the following regularity conditions.

A1. The model space $\mathcal{M}$ is finite; the parameter space Θ_M is compact for any $M \in \mathcal{M}$.

A2. For any fixed $M_j \in \mathcal{M}, j = 0, 1$, as $\theta_j, \tilde{\theta}_j \in \Theta_{M_j}$ and $\tilde{\theta}_j \to \theta_j$, $j = 0, 1$, we have $\mathrm{E}\{Q(M_1,\tilde{\theta}_1,Y) - Q(M_1,\theta_1,Y)|y_{\mathrm{o}}, M_0, \tilde{\theta}_0\} \to 0$ and $\mathrm{E}\{Q(\psi_1,Y)|y_{\mathrm{o}}, M_0, \tilde{\theta}_0\} - \mathrm{E}\{Q(\psi_1,Y)|y_{\mathrm{o}}, M_0, \theta_0\} \to 0$.

A3. For any $M, \tilde{M}$, we have

$\mathrm{E}\{Q(M,\theta,Y) - Q_{\mathrm{o}}(M,\theta,y_{\mathrm{o}})|y_{\mathrm{o}}, M, \theta\} \le \mathrm{E}\{Q(\tilde{M},\tilde{\theta},Y) - Q_{\mathrm{o}}(\tilde{M},\tilde{\theta},y_{\mathrm{o}})|y_{\mathrm{o}}, M, \theta\}$.

A4. $\{\Psi \setminus \Psi_0\} \cap \Psi_1 = \emptyset$, where $\Psi_0 = \mathrm{argmin}_{\psi\in\Psi}\{Q_{\mathrm{o}}(\psi,y_{\mathrm{o}}) + p(M)\}$ and $\Psi_1 = \{\psi_1 \in \Psi : \psi_1 \in a(\psi_1)\}$ with $a(\psi_1) = \mathrm{argmin}_{\psi\in\Psi}[\mathrm{E}\{Q(\psi,Y)|y_{\mathrm{o}}, \psi_1\} + p(M)]$.

A5. $|\Psi_0| = 1$, where $|\cdot|$ denotes cardinality.

Theorem 1. Under assumptions *A1–A5*, the E-MS with GIC converges globally.

Note. The assumption about the parameter spaces being compact in *A1* may be removed, with a probability statement being added to the conclusion of Theorem 1. This is because one can often consider a compact subspace of the parameter space, if the latter is not compact, and let the subspace expand as the sample size increases (similar to the method of sieves; e.g., Jiang 1997). Meanwhile, the other assumptions of Theorem 1 are expected to hold with probability tending to one, as the sample size increases, under regularity conditions. Thus, by applying Theorem 1, we conclude that, with any initial point, the probability that the E-MS converges goes to one as the sample size increases. We show

this with an example in the Supplementary Material (see Section A.2).

Following the classical assumptions for consistency of model selection, we assume the existence of an optimal model, $M_{\rm opt} \in \mathcal{M}$, which is a true model that has the minimum dimension among all true models in $\mathcal{M}$. The corresponding true parameter vector is denoted by $\theta_{\rm opt}$. Suppose that $\mathcal{M}$ is divided into subclasses, $\mathcal{M}_{\rm u}$ and $\mathcal{M}_{\rm o}$, such that $\mathcal{M} = \mathcal{M}_{\rm u} \cup \{M_{\rm opt}\} \cup \mathcal{M}_{\rm o}$. Here the subscripts u and o stand for "underfit" and "overfit", respectively. We use w.p.$\to 1$ for "with probability tending to one".

Theorem 2. Under the assumptions of Theorem 1, if, in addition, we have

A6. for any $M \in \mathcal{M}_{\rm u}$, we have w.p.$\to 1$ that $Q_{\rm o}(M, Y_{\rm o}) > Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})$, and $\{p(M) - p(M_{\rm opt})\}\{Q_{\rm o}(M, Y_{\rm o}) - Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})\}^{-1} = o_{\rm P}(1)$, where $Q_{\rm o}(M, y_{\rm o}) = \inf_{\theta \in \Theta_M} Q_{\rm o}(M, \theta, y_{\rm o})$; and

A7. for any $M \in \mathcal{M}_{\rm o}$, we have w.p.$\to 1$ that $p(M) - p(M_{\rm opt}) > Q_{\rm o}(M_{\rm opt}, Y_{\rm o}) - Q_{\rm o}(M, Y_{\rm o})$,

then, we have, w.p.$\to 1$, that the limiting model of the E-MS convergence is $M_{\rm opt}$. In other words, the E-MS with GIC is consistent.

5 More simulation study

We have carried out a number of simulation studies to evaluate the finite-sample performance of E-MS as well as its comparison with other strategies. One study is presented in this section. More studies are presented in Section A.6 of the Supplementary Material.

We consider the backcross experiment model, described in Example 2, Section 3, with $q = 6$ and $r = 5$, so there are 5 chromosomes with 6 markers on each chromosome. There are 6 true QTLs, which are located at markers 1, 2, 3 on chromosome 1, markers 1, 2 on chromosome 2, and marker 1 on chromosome 3. The coefficients at the true markers are equal, and the value varies according to Table 2; so does the true value of σ. The true value for θ is 0.2. The complete data are generated as follows: First generate the Markov

chain $X_{\rm f}$ with $\theta = 0.2$; then generate e from $N(0, I_n)$; let $Y = \beta X_{\rm opt}(1,1,1,1,1,1)' + e$, where $X_{\rm opt}$ has 6 columns corresponding to the true QTLs. Next, we randomly assign 10% of the values in each column of the data matrix as missing. This leaves less than 4% of the complete-data records, on average (similar to the backcross experiment data obtained by the researchers at UC-Riverside; see Example 2). Let $I_{\rm o} = \{1, \ldots, n\} \setminus I_{\rm m}$ and $O_{jk} = \{1, \ldots, n\} \setminus M_{jk}, 1 \le k \le r, 1 \le j \le q$. The subsets I's, M's and O's are fixed throughout the simulations. The observed data are $y_i, i \in I_{\rm o}$, and $x_{ijk}, i \in O_{jk}, 1 \le k \le r, 1 \le j \le q$.

We study the performance of E-MS with BIC, as described in Example 2. The full model $M_{\rm f}$ was used as the initial model. The result is compared with the complete-data BIC (CDBIC), that is, the BIC result using the complete data. The latter is not available, of course, in practice, but the goal was to see how much loss of efficiency there is in the presence of missing data. As another comparison, we have included results of same-data comparison with a standard imputation-based approach (IM), working in conjunction with the BIC (IMBIC). A description of the IM is provided in Subsection A.6.1 (also see Subsection A.6.4) of the Supplementary Material. Part of the IMBIC results are included in Table 2, and part of the results are deferred to Subsection A.6.4 of the Supplementary Material due to the space limitation. We consider the following measures of performance: TP – empirical probability of correct identification of exactly all the true QTLs (and nothing else); MC – empirical mean number of correctly identified true QTLs (s.d.); and MIC – empirical mean number of incorrectly identified "QTLs" (s.d.). In addition, we compute the percentage ratio (% Ratio) of the TP of E-MS over the TP of CDBIC as a measure of relative efficiency of the E-MS in terms of model selection. The % Ratio for IMBIC is computed in a similar way. The results, based on 100 simulation runs, are presented in Table 2. It is seen that the E-MS results improve when either the sample size increases, or the value of β (the signal) increases, or the value of σ (the noise) decreases, by all of the performance measures. This makes sense because larger n means more information about

Table 2: **Summary of Performance: Backcross Experiment**

n	β	σ	Method	TP	MC (s.d.)	MIC (s.d.)	% Ratio
250	1	1	E-MS	0.51	5.99 (0.10)	0.71 (0.91)	82%
			IMBIC	0.38	5.82 (0.41)	0.75 (0.74)	61%
			CDBIC	0.62	6.00 (0.00)	0.47 (0.66)	
100	1	1	E-MS	0.12	5.22 (0.62)	1.59 (1.70)	52%
			IMBIC	0.09	4.68 (0.96)	1.42 (1.32)	39%
			CDBIC	0.23	5.49 (0.64)	1.22 (1.37)	
250	0.5	1	E-MS	0.08	4.50 (0.90)	1.12 (1.07)	67%
			IMBIC	0.00	3.98 (0.80)	1.12 (1.07)	0%
			CDBIC	0.12	4.73 (0.80)	0.64 (0.78)	
250	1	0.1	E-MS	0.53	6.00 (0.00)	0.66 (0.87)	85%
			IMBIC	0.20	6.00 (0.00)	1.03 (0.70)	32%
			CDBIC	0.62	6.00 (0.00)	0.47 (0.66)	
500	1	1	E-MS	0.57	6.00 (0.00)	0.60 (0.82)	85%
			IMBIC	0.34	5.98 (0.14)	0.86 (0.83)	51%
			CDBIC	0.67	6.00 (0.00)	0.46 (0.72)	

the true underlying model; larger β (or stronger signal) makes it easier to detect the true underlying model; and smaller σ (or weaker noise) makes the sample size more effective and signal relatively stronger. The IMBIC results are not quite comparable to the E-MS, especially in terms of the % Ratio. In particular, unlike the E-MS results, the IMBIC results do not seem to improve when n increases from 250 to 500 (with the same β and σ).

More results of simulation studies are presented in the next section. Furthermore, we have carried out simulation studies on the performance of E-MS in terms of parameter

estimation. The results are presented in Subsection A.6.5 of the Supplementary Material.

6 Missing data mechanism

In a way, there are three cases that the MDM may be involved. The first case, case I, is that the MDM is known, which is rarely the case in practice; the second case, case II, is that the MDM is also of interest, and subject to model selection; the third case, case III, is that the MDM is unknown, but is not of interest; in other words, in case III, there is an underlying MDM, but the latter is something that one wishes to avoid dealing with. In our experience, the third case is encountered most frequently in practice.

The presented E-MS method applies to cases I and II without any change. This is because, in those cases, the observed data include both $y_{\rm obs}$, which is what we normally call "the data" without considering the MDM, and the missing data indicators, $m_{\rm ind}$. In other words, the full (observed) data is $(y_{\rm obs}, m_{\rm ind})$. Under either case I or case II, one has a complete specification of the distribution of $(Y_{\rm obs}, M_{\rm ind})$, that is,

$$f(y_{\rm obs}, m_{\rm ind}|\theta, \psi) \;\; = \;\; \int f(y|\theta) f(m_{\rm ind}|y, \psi) dy_{\rm mis}. \tag{13}$$

The first factor inside the integral on the right side of (13) corresponds to the distribution of the complete data, $Y = (Y_{\rm obs}, Y_{\rm mis})$, where $Y_{\rm mis}$ represents the missing data; the second factor, $f(m_{\rm ind}|y, \psi)$, corresponds to the MDM. Here θ and ψ denote the parameter vectors that are involved in the distribution of Y and the MDM, respectively. Therefore, from a methodology point of view, there is nothing new and (13) is just a special case to which the E-MS applies, that is, a set of data and a distribution for the data under an assumed model, a part of which is the MDM. Note that, sometimes, the integration in (13) can be computed either analytically, or numerically fairly easily. In such cases, the E-MS is not needed; in other words, the model selection can be carried out by directly using the likelihood function

based on the full data, given by (13), which yields the same result as the converged E-MS, had the latter been carried out, at least asymptotically (Theorems 1 & 2).

A more challenging case seems to be case III, in which one is interested in the model on Y only, and would avoid dealing with the MDM if possible. As noted, this case is encountered most frequently in practice. Of course, one may always consider some candidate models for the MDM, and treat the case the same way as case II; once a joint model is selected, one simply takes the part regarding the distribution of Y, which is of main interest. The question is: How does the latter approach compare to the E-MS that focuses on the Y model only? Another related question is: How is the performance of the E-MS, which ignores the MDM, affected by the true underlying MDM? In this section, we address these questions from both empirical and theoretical standpoints.

We refer to Rubin (1976) and Little & Rubin (2002) for the well known theory about missing data, including the notions of MCAR, MAR, NMAR; and ignorable and non-ignorable MDM. According to Little & Rubin (2002, sec. 6.2), the frequentist's methods of inference that ignore the MDM are still valid, even if the MDM is non-ignorable, although there may be a loss of efficiency. It follows that the E-MS, as a frequentist's method, is valid even without considering the MDM; on the other hand, there may be a loss of efficiency in terms of model selection performance. Furthermore, if the true MDM is ignorable, there is no loss of efficiency in any likelihood-based inference, including model selection, by ignoring the MDM. Therefore, the case of interest is when the MDM is non-ignorable.

6.1 Empirical studies

Let us begin by considering a simple model of the analysis of covariance (ANCOVA) with two treatment groups and a control variable. The model can be expressed as

$$Y_{ij} \;=\; \mu_i + \beta x_{ij} + \epsilon_{ij}, \tag{14}$$

$i = 1, 2, j = 1, \ldots, k$, where Y_{ij} is the response; μ_i is the unknown effect for group i; β is an unknown coefficient; x_{ij} is a covariate used as the control variable; and ϵ_{ij} is the error. The ϵ_{ij}'s are assumed to be independent $N(0, \sigma^2)$, where σ^2 is an unknown variance, and independent with the X_{ij}'s. Our interest is in selecting a model for Y_{ij}. There are four candidate models:

I. (14) with $\mu_1 = \mu_2 = \mu$ and $\beta = 0$. The true parameters are $\mu = \sigma^2 = 1$.

II. (14) with $\mu_1 = \mu_2 = \mu$. The true parameters are $\mu = \beta = \sigma^2 = 1$.

III. (14) with $\beta = 0$. The true parameters are $\mu_1 = 1$, $\mu_2 = -1$, and $\sigma^2 = 1$.

IV. (14) with no restriction. The true parameters are $\mu_1 = 1$, $\mu_2 = -1$, and $\beta = \sigma^2 = 1$.

Again, we consider the E-MS with BIC. We assume that the distribution of X_{ij} does not depend on the above models or parameters. Thus, as far as the BIC is concerned, only the conditional log-likelihood, $l_{y|x}$, matters. In each simulation run, the x_{ij}'s are generated from the standard normal distribution; the ϵ_{ij}'s are then generated, and the Y_{ij} obtained under the true model.

We first investigate the impact of different MDMs on the performance of E-MS. Assume that there are no missing x_{ij}'s but some of the responses, Y_{ij}, are missing. Define $M_{\text{ind},ij} = 1$ if Y_{ij} is missing, and $M_{\text{ind},ij} = 0$ if Y_{ij} is observed. It is assumed that the $M_{\text{ind},ij}$'s are independent given Y. Furthermore, the following MDMs are considered:

A. $\mathrm{P}(M_{\text{ind},ij} = 1|y, \psi) = \psi$. The true ψ is 0.5.

B. $\mathrm{P}(M_{\text{ind},ij} = 1|y, \psi) = h(\psi_0 + \psi_1 x_{ij})$, where $h(u) = e^u/(1 + e^u)$. The true parameters are $\psi_0 = 0.5, \psi_1 = 0.2$.

C. $\mathrm{P}(M_{\text{ind},ij} = 1|y, \psi) = h(\psi_0 + \psi_1 \mu_i)$, where the μ_i's are the same group effect introduced above. The true ψ's are the same as in B.

D. $\mathrm{P}(M_{\text{ind},ij} = 1|y, \psi) = h(\psi_0 + \psi_1 y_{ij})$. The true ψ's are the same as in B.

In a way, the models are motivated by the examples considered in Little & Rubin (2002, ch. 6). The basic idea is to consider different types of MDMs including ignorable and

Table 3: **E-MS (BIC) under Different MDMs**: n - total sample size. Reported are empirical probabilities of true-positive (TP) based on 1,000 simulation runs.

	A		B		C		D	
n	10	50	10	50	10	50	10	50
I	0.678	0.745	0.650	0.711	0.641	0.650	0.601	0.542
II	0.727	0.985	0.701	0.988	0.720	0.979	0.736	0.992
III	0.415	0.764	0.256	0.531	0.265	0.570	0.291	0.591
IV	0.327	0.942	0.227	0.798	0.214	0.841	0.239	0.850

non-ignorable missingness. It is clear that both A and B are ignorable. On the other hand, C is a case of MCAR, but no distinctness of parameters, and therefore non-ignorable; D is a case of NMAR, and hence non-ignorable. As mentioned, one expects no loss of efficiency for E-MS under A or B, but the purpose is to see the difference under different situations.

The results, based on 1,000 simulation runs for each combination of the model and MDM, and for two different sample sizes, $n = 10$ and $n = 50$, where $n = 2k$ is the total number of observations, are reported in Table 3. As we can see, the performance of E-MS depends heavily on the underlying true model, but to a much lesser extent on the MDM. More specifically, when model I is the true model, the performance of E-MS somehow decreases as the MDM gets more complex. On the other hand, when the true model is III, or IV, there is a significant drop in the performance once the MDM moves away from A, but not much of a difference between B, C, D. Finally, when model II is the true model, the performance of the E-MS is fairly stable across all the MDMs.

Another aspect of the performance that seems to be affected by the MDM is the improvement as the sample size increases. In almost all the cases the performance of E-MS improves as the sample size gets larger; however, the improvement is much more signifi-

cant under II, III and IV than under I. In fact, in one case under I when the MDM is NMAR, the performance even gets worse as n gets larger. One explanation is that the MDM is, in this case, confounded with some of the candidate models that leads to incorrect model selections. In general, missing data reduces the effective sample size. However, additional covariate data are available under II, III and IV, namely, the x_{ij}'s (under II and IV) and the group indicators (as another covariate, under III and IV), which are not affected by the missing data. The covariate information helps to improve the performance as the sample size increases. In fact, the largest improvement is seen under IV, which has both of the covariates (x_{ij} and the group indicator) under the true model.

In our next simulation study, we focus on the efficiency of E-MS (in model selection), and compare its performance with the approach based on the full-data-likelihood (13). To make a fair comparison, both procedures are based on the BIC. The candidate models for $f(y|\theta)$ are the same as above. The candidate MDMs are A–C plus
E. $\mathrm{P}(M_{\mathrm{ind},ij} = 1|y, \psi) = h(\psi_0 + \psi_1\mu_i + \psi_2 x_{ij})$, where the μ_i's are the same as in (14).
A motivation for not using model D as a candidate MDM is that we would like to see what happens when a NMAR missingness (that is, model D) is not considered as a candidate MDM, but is actually at play. Model E also has the features that (i) it is non-ignorable, and (ii) it is a full model when considered together with A, B, C. Let us term the E-MS with BIC as E-MS, and the full-data BIC as FBIC. Note that, in this case, the FBIC can be carried out directly without using the E-MS, as noted earlier. We compare the E-MS with FBIC for two cases where the true underlying MDM is among the candidates, namely, II-B and IV-B, in which case the FBIC would be considered efficient, and two cases where the true underlying MDM is not among the candidates: namely, II-D and IV-D, in which case the FBIC may not be efficient. Note that IV is a full model for $f(y|\theta)$. We increase the sample size slightly from the previous simulation, namely, $n = 40$ and $n = 80$ now. Results based on 500 simulation runs are reported in Table 4.

Table 4: **Comparison of E-MS and FBIC**: n - total sample size. Reported are empirical probabilities of true-positive (TP) based on 500 simulation runs.

	II-B		II-D		IV-B		IV-D	
n	E-MS	FBIC	E-MS	FBIC	E-MS	FBIC	E-MS	FBIC
40	0.992	0.830	0.992	0.782	0.726	0.878	0.758	0.848
80	0.996	0.950	0.996	0.934	0.934	0.996	0.978	0.986

Before the results are revealed, one might speculate that E-MS would outperform FBIC when the true MDM is not among the candidates, that is, II-D and IV-D, and the pattern would reverse when the true MDM is among the candidates, that is, II-B and IV-B. Thus, the way that the results turn out to be might have surprised someones, including ourselves. However, there are some explanations. First, in FBIC, one first targets the joint model then marginalize to the model of interest, that is, $f(y|\theta)$. This is not necessarily a better approach than targeting directly the model of interest. See, for example, Claeskens and Hjort (2003). Another example, in the context of parameter estimation, is the restricted maximum likelihood (REML; e.g., Jiang 2007), which targets the parameters of direct interest, that is, the variance components. This often works better than the straight maximum likelihood, which estimates all the parameters, some of which may be considered nuisance.

Secondly, the BIC is known to have the tendency of over-penalizing "larger" models, and this is especially the case when the full model is the true underlying model (e.g., Jiang *et al.* 2008). For E-MS, model IV is, simply, the full model, therefore, the BIC-based E-MS suffers from over-penalizing. However, model IV is not necessarily (part of) the full model for FBIC. This is because the full model for FBIC is the joint model (IV,E). For example, suppose that (IV,B) is selected by the FBIC, then, obviously, it is not the full model, even though it is "full" for the first component. The point being made is that the E-MS would

suffer more from over-penalizing than the FBIC once IV is the true underlying model.

Thirdly, the true underlying MDM can affect the performance of E-MS in positive or negative ways, as shown by the earlier simulation result. In fact, if the MDM works in the right direction, the E-MS can have a "super-performance", as shown in the next study.

In Section 5, the missing data indexes were generated randomly independent of the data; thus, the MDM was ignorable. We now repeat the simulation study but with the missing data indexes generated according to the following two scenarios. Let $M_{\text{ind},i}$ be the missing data indicator for Y_i, and $m_{\text{ind},ijk}$ that for x_{ijk}. Scenario MA: Given the data Y and x, (i) generate the $M_{\text{ind},i}$'s independently with $\mathrm{P}(M_{\text{ind},i} = 1) = 0.1$; (ii) generate the $m_{\text{ind},ijk}$'s independently so that $\mathrm{P}(m_{\text{ind},ijk} = 1) = 0.05$ if $x_{ijk} = 0$, and $\mathrm{P}(m_{\text{ind},ijk} = 1) = 0.1$ if $x_{ijk} = 1$. Scenario MB: Given the data Y and x, (i) generate the $M_{\text{ind},i}$'s independently with $\mathrm{P}(M_{\text{ind},i} = 1) = h(\psi_0 + \psi_1 Y_i)$, where $h(u) = e^u/(1 + e^u)$, $\psi_0 = -2.5$, and $\psi_1 = 0.1$; (ii) generate the $m_{\text{ind},ijk}$'s the same way as Scenario MA. It is clear that both scenarios are non-ignorable. Scenario MA is MCAR in terms of the Y data, but NMAR in terms of the Y, x data; Scenario MB is NMAR in terms of both Y and x data. Thus, in a way, Scenario MB has a more serious non-ignorable MDM than Scenario MA. Due to the space limitation, the simulation results are presented in Subsection A.6.3 of the Supplementary Material. Comparing with the results reported in Table 2, it is seen that, in some cases (5 out of 10), the E-MS performed worse, but in some cases (5 out of 10) the E-MS performed better (note that these simulations used the same random seeds, so the results are completely comparable). In particular, there are a couple of cases of super-performance, in which the E-MS actually outperformed the CDBIC. An interpretation is that the missing data indicators may carry additional information to the complete data, which the E-MS is able to make use of (while the CDBIC cannot), if the MDM functions in the right way.

The apparent interaction between the E-MS and MDM observed in the simulation studies is quite interesting. To demonstrate this theoretically, we explore the connection be-

tween E-MS and MDM from a large sample point of view.

6.2 Large sample consideration

For simplicity, let us assume that the observations Y_i are independent Gaussian with mean $\mathrm{E}_{M,\theta_M}(Y_i)$, where M indicates the assumed model for the mean, and θ_M the vector of parameters under M, and unknown variance σ^2. Consider selection of M using the E-MS with BIC, which, at the current iteration, amounts to minimize $n \log\{\mathrm{E_c}(Q_M|y_{\mathrm{obs}})\} + \log(n)|M|$, where $Q_M = \sum_{i=1}^n \{Y_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2$, $|M|$ is the dimension of θ_M, and $\mathrm{E_c}$ denotes conditional expectation under the current model and parameters under the current model. Because the penalty term, $\log(n)|M|$, is not affected by the MDM, we can focus on the first term, which, eventually, leads to consideration of $\mathrm{E_c}(Q_M|y_{\mathrm{obs}})$. The derivation below requires, of course, some regularity conditions (e.g., Jiang, Lahiri & Wan 2002); however, we shall bypass these technical conditions and focus on the insight of the result.

Let $m_{\mathrm{ind},i}$ denote the missing data indicator. Then, we have

$$\begin{aligned} \mathrm{E_c}(Q_M|y_{\mathrm{obs}}) &= \sum_{i=1}^n \{y_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(m_{\mathrm{ind},i}=0)} \\ &\quad + \sum_{i=1}^n \mathrm{E_c}\{Y_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(m_{\mathrm{ind},i}=1)}. \end{aligned} \tag{15}$$

Suppose that the current model is correct, but not necessarily optimal. For example, if the space of candidate models includes a true model, then the full model, M_{f}, is correct, but not necessarily optimal in that it may include extraneous variables. Furthermore, suppose that the current estimator of parameters is consistent. Then, the conditional expectation, $\mathrm{E_c}$, can be replaced by the true conditional expectation, E, resulting a difference that is of lower order. Another situation is when the E-MS results in consistent model selection (see Theorem 2). Then, asymptotically, one can replace $\mathrm{E_c}$ by E. Furthermore, by Theorem 2 of Jiang *et al.* (2011a), the minimizer of (15), with $\mathrm{E_c}$ replaced by E, $\hat{\theta}_M$, converges

in probability to some limiting vector, say, θ_M, and this is true regardless whether M is a correct model. Thus, by considering the leading term, we can focus on (15) with $\mathrm{E_c}$ replaced by E, and θ_M being the limiting vector. Let $\mathrm{P}(M_{\mathrm{ind},i} = 1|y) = 1 - h(y_i)$ be the true underlying MDM; in other words, $h(y_i) = \mathrm{P}(M_{\mathrm{ind},i} = 0|y)$, where y is the complete data. Define $c_i = \mathrm{E}\{Y_i h(Y_i)\}/\mathrm{E}\{h(Y_i)\}$ (again, E without subscript represents the true expectation). It is shown in Seciton A.5 of the Supplementary Material that

$$\begin{aligned}\mathrm{E}\{\mathrm{E_c}(Q_M|Y_{\mathrm{obs}})\} &= \sum_{i=1}^{n}\{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2\mathrm{E}\{h(Y_i)\} \\ &\quad + \sum_{i=1}^{n}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}^2[1 - \mathrm{E}\{h(Y_i)\}] + \delta, \qquad (16)\end{aligned}$$

where δ consists of lower-order terms, or terms that do not depend on M. Let M_{opt} denote the optimal model. Then, for $M = M_{\mathrm{opt}}$, the second term on the right side of (16) disappears. Thus, we have (again, see Section A.5 of the Supplementary Material)

$$\begin{aligned}&\text{difference in (15) between } M \text{ and } M_{\mathrm{opt}} \\ &= \mathrm{E}\{\mathrm{E_c}(Q_M|Y_{\mathrm{obs}})\} - \mathrm{E}\{\mathrm{E_c}(Q_{M_{\mathrm{opt}}}|Y_{\mathrm{obs}})\} + \delta_1 \\ &= 2\sum_{i=1}^{n}\mathrm{cov}\{Y_i, h(Y_i)\}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\} + \delta_2, \qquad (17)\end{aligned}$$

where δ_1 denotes terms of lower-order, and δ_2 consists of terms of lower-order, or terms that do not depend on the MDM. (17) is a key result that shows how the performance of the E-MS is influenced by the MDM through its leading term, namely, the larger this term (i.e., more positive), the easier to distinguish a non-optimal model from the optimal one. It is interesting to note that the leading term is a sum of products, where the first factor of the product, $\mathrm{cov}\{Y_i, h(Y_i)\}$, depends on the MDM but not on M, while the second factor of the product, $\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)$, depends on M but not on the MDM.

Expression (17) may help to explain, for example, the interesting pattern observed in Table 3. Note that $h(y_i)$ is the probability that y_i is observed. Therefore, among the four

MDMs considered, case D is likely the case that the covariance, $\mathrm{cov}\{Y_i, h(Y_i)\}$, is largest in absolute value, but the sign is negative because $h(y_i)$ is decreasing with y_i in this case. Thus, if we denote the difference $\mathrm{E}_{M,\theta_M}(Y_i) - \mathrm{E}(Y_i)$ by d_M, the summand in (17) can be written as the product of the positive covariance, $\mathrm{cov}\{Y_i, 1 - h(Y_i)\}$, and d_M. Note that d_M is likely to be much larger when M is underfitting than overfitting. Now look at Table 3, case D, with $n = 50$ to imitate the large sample behavior. Under model I, none of the candidate models are underfitting; thus, their d_M contributions are likely to be relatively small, hence it is more difficult to identify a non-optimal model. Similarly, under model III, none of the other models appear to be underfitting. On the other hand, under model II, models I and III are underfitting; under model IV, all of the other models are underfitting. This explains why the empirical TPs are much higher under models II and IV. It should be noted that, as is well known, a BIC-based approach tends to suffer when the full model is the underlying model, which may explain why the empirical TPs under model II are higher than those under model IV. Similar explanations also apply to cases C and B. The behavior under case A is somewhat different, and there is, again, an explanation. Note that, under case A, the probability of missing is a constant. It follows that $\mathrm{cov}\{Y_i, h(Y_i)\} = 0$. Therefore, in this case, the leading term in (17) has disappeared.

7 Real data example

Recall the data set obtained by the UC-Riverside researchers mentioned in Section 1. The gene expression data were originally published by Luo *et al.* (2007). The phenotypic values of eight quantitative traits of barley were published by Hayes *et al.* (1993). Detailed description of the experiment can be found in the latter reference, which involved 150 double haploid (DH) lines derived from the cross of two spring barley varieties, Morex and Steptoe. The DH lines are considered as the subjects here. In all there were 495

SNP markers on seven chromosomes that are under investigation. As mentioned, there are significant missing values in the data so that only 4 of the 150 subjects have complete genotype records. On the other hand, there are no missing values in the phenotypic data.

We consider a Markov-chain model as in Example 2. However, the high-dimensional nature of the data presents a problem for the direct application of the E-MS, because the total number of markers (495) is much larger than the sample size ($n = 150$). More specifically, the least squares (LS) fit is unfeasible when the number of predictors is larger than the sample size. To overcome this difficulty, we use the following idea of *conditional modeling*, described under a more general setting.

Suppose that, conditional on $X = (x_i')_{1\le i\le n}$, one has a linear regression $Y = X\beta + \epsilon$, where $Y = (Y_i)_{1\le i\le n}$ are the observations, and $\epsilon = (\epsilon_i)_{1\le i\le n}$ are the errors such that the components of ϵ are independent with mean 0, and ϵ is independent of X. Furthermore, suppose that $X = [X_{(1)}\ X_{(2)}]$ with $X_{(r)} = (X_{ir}')_{1\le i\le n}, r = 1, 2$ such that $X_{(1)}, X_{(2)}$ are independent [e.g., Broman & Speed (2002)]. Then, it is easy to show that $X_{(1)}$ is independent of $[X_{(2)}, \epsilon]$. Note that we can express the regression model as $Y = X_{(1)}\beta_1 + X_{(2)}\beta_2 + \epsilon$. Without loss of generality, we assume that $X_{(1)}\beta_1$ does not involve an intercept [which, if exist, belongs to $X_{(2)}\beta_2$].

Now suppose that $X_{i2}, i = 1, \ldots, n$ are independent, and that $\mathrm{E}(X_{i2})$ does not depend on i. Then, $\mathrm{E}(X_{i2}'\beta_2 + \epsilon_i) = \mathrm{E}(X_{i2})'\beta_2$ is a constant, say, β_0. Let $e_i = X_{i2}'\beta_2 + \epsilon_i - \beta_0$. It is easy to show that $e_i, i = 1, \ldots, n$ are independent with $\mathrm{E}(e_i) = 0$, and $Y = [1_n\ X_{(1)}](\beta_0\ \beta_1')' + e$, e being independent of $[1_n\ X_{(1)}]$. In other words, conditional on $X_{(1)}$, we, once again, have a standard linear regression model (i.e., the errors are independent with mean zero, and independent with the predictors).

The point is that $X_{(1)}$ can be of much lower dimension than X. For the barley cross data, we can let $X_{(1)}$ correspond to markers on any particular chromosome. The number of markers on the 7 chromosomes are 60, 78, 81, 60, 93, 56 and 67, respectively, all of which

are smaller than the sample size 150. Within each chromosome, we apply the E-MS in conjuction with the IF (Jiang *et al.* 2011b; also see Jiang 2014). The number of bootstrap samples is chosen as $B = 100$.

It is known that, for high-dimensional data the IF may suffer from the so-called dominant factor effect (Jiang *et al.* 2011b, sec. 3.3). For the most part, this means that the IF frequency (i.e., the empirical probability of the most frequently selected model; e.g., Jiang 2014) tends to be in favor of a lower dimensional model than the true model, if the "signals" are relatively weak due to the limited sample size. This problem is dealt with naturally by the E-MS. First we apply the IF, under the full model, that is, all the markers on a given chromosome, to obtain the IF frequencies at different dimensions, say, $p_1^*, p_2^*, \ldots, p_q^*$, where p_j^* is the IF frequency at dimension j, and q is the total number of markers, for the chromosome. If the frequencies show a "peak", that is, there is a $1 < j < q$ such that $p_j^* > p_{j-1}^*$ and $p_j^* > p_{j+1}^*$, the E-MS shall continue; otherwise, we conclude that there is no more than one QTL on the chromosome. In the latter case, the highest IF frequecy must take place at the boundary, that is, either at dimension one or at the highest dimension corresponding to all the markers on the chromosome. However, it is unlikely that all the markers are QTLs; therefore, dimension one is chosen, and the E-MS stops.

If the frequency plot show a "peak", and therefore the E-MS is to continue, we first look for the last peak, that is, the highest dimension that corresponds to a peak in order to be conservative. This is similar to the AF (Jiang *et al.* 2009), where the first significant peak is chosen in order to determine the cut-off for the fence (e.g., Jiang 2014). The first peak for the AF corresponds to the last peak for the IF. The markers corresponding to the last peak are selected, the current model is updated, and the updated model is treated as the (new) full model for the next step of iteration. The procedure is repeated until either the updated model is identical to the current model, or no peak is found during the current step; in both cases, the current model is chosen as the final model. For the latter case, when

no peak is found, we choose the highest dimension, instead of dimension one as above in the initial step. This is because, at this stage, we have already determined that there are more than one QTLs on the chromosome (the E-MS would not have continued otherwise); furthermore, the highest dimension possibly has been updated, so it no longer corresponds to all of the markers on the chromosome.

The results for the grain protein phenotype are presented in Table A.10 of the Supplementary Material. The results show some consistency with the foundings of Zhan *et al.* (2011). For example, the latter authors found that chromosomes 2, 3, 5 "seem to control more genes than other chromosomes". According to our results, those three chromosomes contain nearly 60% of all the QTLs found. In particular, chromosomes 3 and 5 are the top two according to the number of QTLs found. It should be noted that the number of QTLs found on a chromosome is not the only thing that represents the relative importance of the chromosome; the magnitude of the QTL effect is also important. In this application, however, our focus is identification of the QTLs, rather than estimation of the QTL effects.

8 Discussion

George Box once famously said that "essentially, all models are wrong, but some are useful" (Box 1979). Practical use of statistical modeling involves using the model as an approximation to the real-life problem, rather than the truth for the problem. Thus, model selection, correspondingly, should be understood as finding the optimal model that most efficiently approximates the problem of practical interest. Although, in the simulation studies presented in this paper, we have looked at cases where there is a true model among the candidate models, we have, indeed, considered situations where there is no true model among the candidate models. More specifically, Nguyen *et al.* (2013) considered a situation where the true underlying model is not among those considered as candidate models. Namely, all

of the candidate models assume that the true QTLs are at the exact locations of some of the markers under consideration. In practice, however, this may not be true; in other words, the true QTLs may be at locations between the markers. The authors considered the case where the true QTLs are located in the middle of their flanking markers; therefore, the true underlying model is not a candidate model. Nevertheless, the goal was to identify, among the candidate models, the one that best approximates the true model in the sense that the identified markers are closest to the true QTLs. We consider a setting similar to those of Nguyen *et al.* (2013). Our simulation results show that E-MS, here in conjunction with the FW/BW BIC (see Section A.3 of the Supplementary Material), is still capable of identifying the best approximating model in case that the true model is not among the candidates. See Subsection A.6.2.1 of the Supplementary Material for detail.

Our investigation on the E-MS has revealed other interesting properties of the procedure that deserve further studies. In particular, there have been studies on adjusting the penalty parameter in the information criteria to make the latter "more aggressive", in some sense. For example, Mueller and Welsh (2005) finds that a modified BIC procedure with the penalty $2 \log n$ instead of $\log n$ works better in some cases. Similar findings were reported in Broman & Speed (2002). Mueller and Welsh (2010) treats the problem through a unified approach by considering the selection curves in GIC, in which the criterion function is viewed as a linear function of the penalty parameter. On the other hand, the fence methods (e.g., Jiang 2014) is able to avoid dealing with the penalty by letting the data speak on how to choose a cut-off, or a tuning parameter. Potentially, another way of letting the data speak in choosing the tuning parameter is through the E-MS, as suggested by the simulation study in Subsection A.6.1 of the Supplementary Material. Namely, as the E-MS iteration proceeds, one is having a clearer picture about the data-generating mechanism. This would help one in knowing whether one should be "more aggressive", or "less aggressive", in choosing the penalty.

There has been recent interest in (joint) selection of fixed and random effects in mixed effects models. See Bondell *et al.* (2010), Ibrahim *et al.* (2011). The authors of latter references used Cholesky-type decompositions, which allow them to use the approach of shrinkage selection methods. The E-M algorithm is used (in both references) to deal with the fact that the random effects are not observable. See Jiang (2014) for further discussion. As noted (see second to last paragraph of Section 2), for shrinkage methods, the E-MS and E-M are the same. Alternatively, one may treat the problem as joint selection of the fixed effects and variance-covariance structure of the random effects, as in Mou (2012). The point is that one may treat the random effects as incomplete data, as in the traditional approach of mixed model analysis via the E-M algorithm (e.g., Jiang 2007, sec. 4.1.1). The E-MS procedure developed in the current paper seems to fit naturally to the latter approach. This would be a very interesting problem of future studies.

Acknowledgements. The research of Jiming Jiang is partially supported by the NSF grant SES-1121794. The research of Thuan Nguyen is partially supported by the NSF grant SES-1118469. The research of J. Sunil Rao is partially supported by the NSF grant SES-1122399. The research of all three authors are partially supported by the NIH grant R01-GM085205A1. The authors are grateful to a Co-Editor, an Associate Editor and two referees for their valuable comments.

References

[1] Afifi, A. and Elashoff, R. (1966), Missing observations in multivariate statistics: I. Review of the literature, *J. Amer. Statist. Assoc.* 61, 595-604.

[2] 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.

[3] Booth, J. G. and Hobert, J. P. (1999), Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm, *J. Roy. Statist. Soc. Ser. B* 61, 265–285.

[4] Box, G. E. P. (1979), Some problems of statistics of everyday life, *J. Amer. Statist. Assoc.* 74, 1-4.

[5] Broman, K. W. and Speed, T. P. (2002), A model selection approach for the identification of quantitative trait loci in experiemental crosses, *J. Roy. Statist. Soc. Ser. B* 64, 641-656.

[6] Bueso, M. C., Qian, G., and Angulo, J. M. (1999), Stochastic complexity and model selection from incomplete data, *J. Statist. Planning Inference* 76, 273-284.

[7] van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2005), Fully conditional specification in multivariate imputation, *J. Statist. Comput. Simulation* 76, 1049-1064.

[8] Cavanaugh, J. E. and Shumway, R. H. (1998), An Akaike information criterion for model selection in the presence of incomplete data, *J. Statist. Planning Inference* 67, 45-65.

[9] Claeskens, G. and Consentino, F. (2008), Variable selection with incomplete covariate data, *Biometrics* 64, 1062-1069.

[10] Claeskens, G. and Hjort, N. L. (2003), The focused information criterion (with discussion), *J. Amer. Statist. Assoc.* 98, 900-945.

[11] Copas, J. and Eguchi, S. (2005), Local model uncertainty and incomplete-data bias (with discussion), *J. Roy. Statist. Soc. B*, 4, 459-513.

[12] Dempster, A., Laird, N., and Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. B* 39, 1-38.

[13] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed., Oxford Univ. Press.

[14] van Dyk, D. A. (2000), Nesting EM algorithms for computational efficiency, *Statist. Sinica* 10, 203-225.

[15] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96, 1348-1360.

[16] Fuchs, C. (1982), Maximum likelihood estimation and model selection in contingency tables with missing data, *J. Amer. Statist. Assoc.* 77, 270-278.

[17] Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010), Variable selection for regression models with missing data, *Statist. Sinica* 20, 149-165.

[18] Hartley, H. O. and Hocking, R. (1971), The analysis of incomplete data, *Biometrics* 27, 783-823.

[19] Hayes, P. M. *et al.* (1993), Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm, *Theor. Appl. Genet.* 87, 392-401.

[20] Hens, N., Aerts, M., and Molenberghs, G. (2006), Model selection for incomplete and design-based samples, *Statist. Med.* 25, 2502-2520.

[21] Ibrahim, J. G., Zhu, H., and Tang, N. (2008), Model selection criteria for missing-data problems using the EM algorithm, *J. Amer. Statist. Assoc.* 103, 1648-1658.

[22] Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011), Fixed and random effects selection in mixed effects models, *Biometrics* 67, 495-503.

[23] Jansen, R. C. (1993), Interval mapping of multiple quantitative trait loci, *Genetics*, 135, 205-211.

[24] Jiang, J. (1997), Wald consistency and the method of sieves in REML estimation, *Ann. Statist.* 25, 1781-1803.

[25] Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

[26] Jiang, J. (2014), The fence methods, in *Advances in Statistics*, Hindawi Publishing Corp., Cairo, in press.

[27] Jiang, J., Lahiri, P. and Wan, S. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782-1810.

[28] Jiang, J., Nguyen, T. and Rao, J. S. (2009), A simplified adaptive fence procedure, *Statist. Probab. Letters* 79, 625-629.

[29] Jiang, J., Nguyen, T. and Rao, J. S. (2010), Fence method for nonparametric small area estimation, *Survey Methodology* 36, 3-11.

[30] Jiang, J., Nguyen, T. and Rao, J. S. (2011a), Best predictive small area estimation, *J. Amer. Statist. Assoc.* 106, 732-745.

[31] Jiang, J., Nguyen, T. and Rao, J. S. (2011b), Invisible fence method and the identification of differentially expressed gene sets, *Statist. Interface* 4, 403-415.

[32] Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008), Fence methods for mixed model selection, *Ann. Statist.* 36, 1669-1692.

[33] Lander, E. S., and Botstein, D.(1989), Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, 121, 185-199.

[34] Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd ed., Wiley, New York.

[35] Liu, J. S. (2004), *Monte Carlo Strategies in Scientific Computing*, Springer, New York.

[36] Luo, Z. W. *et al.* (2007), SFP genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators, *Genetics* 176, 789-800.

[37] Mou, J. (2012), Two-stage fence methods in selecting covariates and covariance for longitudinal data, Ph. D. dissertation, Dept. of Statist., Univ. of Calif., Davis, CA.

[38] Müller, S., Scealy, J. L., and Welsh, A. H. (2013), Model selection in linear mixed models, *Statist. Sci.* 28, 135-167.

[39] Müller, S. and Welsh, A. H. (2005), Outlier robust model selection in linear regression, *J. Amer. Statist. Assoc.* 100, 1297-1310.

[40] Müller, S. and Welsh, A. H. (2010), On model selection curves, *International Statist. Rev.* 78, 240-256.

[41] Nguyen, T., Peng, J., and Jiang, J. (2013), Fence methods for backcross experiments, *J. Statist. Comput. Simulation*, in press.

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

[43] Pang, Z., Lin, B., and Jiang, J. (2013), Regularization parameter selections with divergent and NP-dimensionality via bootstrapping, *Australian & New Zealand J. Statist.*, under revision.

[44] Rissanen, J. (1983), A universal prior for integers and estimation by minimum description length, *Ann. Statist.* 11, 416-431.

[45] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995), Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *J. Amer. Statist. Assoc.* 90, 106-121.

[46] Rotnitzky, A., Robins, J. M. and Scharfstein, D. (1998), Semiparametric regression for repeated outcomes with nonignorable nonresponses, *J. Amer. Statist. Assoc.* 93, 1321-1339.

[47] Rubin, D. B. (1976), Inference and missing data, *Biometrika* 63, 581-592.

[48] Sebastiani, P. and Ramoni, M. (2001), Bayesian selection of decomposable models with incomplete data, *J. Amer. Statist. Assoc.* 96, 1375-1386.

[49] Seghouane, A.-K., Bekara, M., and Fleury, G. (2005), A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence, *Signal Process.* 85, 1405-1417.

[50] Shibata, R. (1984), Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika* 71, 43-49.

[51] Shimodaira, H. (1994), A new criterion for selecting models from partially observed data, in P. Cheeseman and R. W. Oldford, eds., *Selecting Model from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics* 89, Spriner, New York, 21-29.

[52] Schomaker, M., Wan, A. T. K., and Heumann, C. (2010), Frequentist Model Averaging with missing observations, *Comput. Statist. Data Anal.* 54, 3336-3347.

[53] Tibshirani, R. J. (1996), Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B* 16, 385-395.

[54] Verbeke, G., Molenberghs, G., and Beunckens, C. (2008), Formal and informal model selection with incomplete data, *Statist. Sci.* 23, 201-218.

[55] Zhan, H., Chen, X., and Xu, S. (2011), A stochastic expectation and maximization algorithm for detecting quantitative trait-associated genes, *Bioinformatics* 27, 63-69.

[56] Zeng, Z.-B. (1993), Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci, *Proc. Nat. Acad. Sci. USA*, 90, 10972-10976.