Appendix to 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

Throughout this Supplementary Material, the paper, "The E-MS Algorithm: Model Selection with Incomplete Data" by Jiang, Nguyen & Rao, is referred to as JNR.

A.1 Convergence and consistency of E-MS

In this section, we provide detailed results regarding the convergence and consistency of E-MS, reported in Section 4 of JNR, and their extensions.

First we would like to point out a key idea for the proof of the (numerical) convergence, which is based on a well-known result in numerical analysis, known as the *global convergence theorem* (GCT). First introduce a few terms in numerical analysis. An algorithm is defined as a map, a, that assigns to every point $x \in \mathcal{X}$ a subset $a(x) \subset \mathcal{X}$. In particular, $a(x)$ may consist of a single point, in which case the definition of a map is consistent the traditional concept. To see an example, suppose that $a(x)$ is defined as the solution(s), y, to the equation $g(x, y) = 0$. Given x, if the solution is unique, then $a(x)$ is a single point; if the solutions exist but are not unique, then $a(x)$ is a subset; and, if the solution does not exist, then $a(x) = \emptyset$. Operated iteratively, the algorithm initiated at $x_0 \in \mathcal{X}$ would generate a sequence, $x_k, k = 0, 1, 2, \ldots$, defined by

$$x_{k+1} \in a(x_k). \tag{A.1}$$

The map a is said to be closed at $x \in \mathcal{X}$ if $x_k \to x, x_k \in \mathcal{X}$ and $y_k \to y, y_k \in a(x_k)$ imply $y \in a(x)$. The algorithm defined by a is said to converge globally if, with any initial point x_0, the sequence $x_k, k = 1, 2, \ldots$ converges to the same point $x^* \in \mathcal{X}$.

Global Convergence Theorem (e.g., Luenberger 1984). Suppose that the sequence x_k is generated by an algorithm a via (A.1), and there is a continuous function, g, such that

the following conditions (a)–(c) hold. Then, the limit of any convergent subsequence of x_k must be a solution to the following optimization problem:

$$\text{minimize} \quad g(x) \quad \text{subject to} \quad x \in \mathcal{X}. \tag{A.2}$$

(a) all points of x_k are contained in a compact subset $S \subset \mathcal{X}$; (b) a is closed at any x that is not a solution to (A.2); (c) if x is not a solution to (A.2), then $g(y) < g(x)$ for all $y \in a(x)$, and if x is a solution to (A.2), then $g(y) \leq g(x)$ for all $y \in a(x)$.

The GCT was used in the proof of the convergence of the E-M algorithm (Wu 1983). Also see, for example, Jiang (2000) for another application of the GCT. In Wu (1983), the inequalities in (c) are reversed because, therein, the author considered maximum likelihood. Clearly, this is equivalent to our version of (c) if, instead, (A.2) is considered. Among the three conditions, the key is to show (c), because the rest of the conditions are relatively easy to verify, or reasonable to assume (such as the compactness of the parameter/model space). Thus, we will focus on condition (c). Also note that, typically, the strict inequality, $<$, in (c) holds under some regularity conditions that rule out some trivial cases, once the inequality $\leq$ or, in other words, the monotonic property of g, is established. It should also be noted that, in numerical analysis, convergence of an algorithm is reached if the distance between the current point and the updated one is less than a threshold that is set up in advance (e.g., 10^{-6}). However, if the model space is discrete, the threshold condition is met if and only if the updated model is identical to the current model; and we use this as the definition of convergence in an iterative model selection procedure.

To verify the key condition (c) of the GCT, it essentially amounts to show that there is a function, g, so that $g(M^{(t+1)}, \theta^{(t+1)}) \leq g(M^{(t)}, \theta^{(t)})$. For example, in the E-M algorithm, g is the negative log-likelihood, which satisfies $g(\theta^{(t+1)}) \leq g(\theta^{(t)})$. But now we have to find a g that involves not just θ, but also M. Recall the observed version of (10) of JNR, introduced three lines below (12) of JNR, where Q_{o} is some observed version of Q. A key condition for Theorem 1 of JNR is assumption *A3*. This condition may be interpreted as that the expected difference in the measure of lack-of-fit under the correct model is no more

than that under an incorrect model. We illustrate with some examples (the numbering of the examples follows the sequence of JNR).

Example 4 (negative log-likelihood). Consider $Q(M,\theta,Y) = -\log f_{M,\theta}(Y)$ and define $Q_{\rm o}(M,\theta,y_{\rm o}) = -\log f_{M,\theta}(y_{\rm o})$, where $f_{M,\theta}(y)$ and $f_{M,\theta}(y_{\rm o})$ are the pdfs of Y and $Y_{\rm o}$, respectively, with respect to some σ-finite measure ν, under M and θ. Then, we have $Q(M,\theta,Y) - Q_{\rm o}(M,\theta,y_{\rm o}) = -\log\{f_{M,\theta}(Y)/f_{M,\theta}(y_{\rm o})\} = -\log f_{M,\theta}(Y_{\rm m}|y_{\rm o})$, where $Y_{\rm m}$ denotes the vector of missing data. It follows that

$$\begin{aligned}
&{\rm E}\{Q(\tilde{M},\tilde{\theta},Y) - Q_{\rm o}(\tilde{M},\tilde{\theta},y_{\rm o})|y_{\rm o},M,\theta\} - {\rm E}\{Q(M,\theta,Y) - Q_{\rm o}(M,\theta,y_{\rm o})|y_{\rm o},M,\theta\}\\
&= \int\{\log f_{M,\theta}(y_{\rm m}|y_{\rm o})\}f_{M,\theta}(y_{\rm m}|y_{\rm o})d\nu - \int\{\log f_{\tilde{M},\tilde{\theta}}(y_{\rm m}|y_{\rm o})\}f_{M,\theta}(y_{\rm m}|y_{\rm o})d\nu\\
&= -\int \log\{f_{\tilde{M},\tilde{\theta}}(y_{\rm m}|y_{\rm o})/f_{M,\theta}(y_{\rm m}|y_{\rm o})\}f_{M,\theta}(y_{\rm m}|y_{\rm o})d\nu\\
&\geq -\log\int\{f_{\tilde{M},\tilde{\theta}}(y_{\rm m}|y_{\rm o})/f_{M,\theta}(y_{\rm m}|y_{\rm o})\}f_{M,\theta}(y_{\rm m}|y_{\rm o})d\nu = 0,
\end{aligned}$$

using Jensen's inequality. Thus, *A3* of JNR is satisfied.

Example 5. Consider selecting the covariates in a linear regression, $Y_i = x_i'\beta + \epsilon_i, i = 1,\ldots,n$, where the errors ϵ_i are independent with mean 0 and variance σ^2. The components of x_i are subject to selection, with β being the corresponding vector of regression coefficients. Assume, for simplicity, that σ^2 is known, and that no x_i's are missing. Thus, we can treat the x_i's as fixed, and drop them from the condition in the conditional expectation. Here, a model M corresponds to a set of specified covariates, x, and $\theta = \beta$. Suppose that $y_1,,\ldots,y_m$ are observed, while $y_{m+1},\ldots,y_n$ are missing. As in Example 1, we consider MAR for simplicity. Let $Q(M,\theta,y) = Q(x,\beta,y) = \sum_{i=1}^m(y_i - x_i'\beta)^2 + \delta_0\sum_{i=m+1}^n(y_i - x_i'\beta)^2$, where δ_0 is positive and $\mathcal{F}(y_{\rm o})$ measurable, and, natually, consider $Q_{\rm o}(x,\beta,y_{\rm o}) = \sum_{i=1}^m(y_i - x_i'\beta)^2$. Then, we have ${\rm E}\{Q(\tilde{x},\tilde{\beta},Y)|y_{\rm o},x,\beta\} = \sum_{i=1}^m(y_i - \tilde{x}_i'\tilde{\beta})^2 + \delta_0\{(n-m)\sigma^2 + \sum_{i=m+1}^n(x_i'\beta - \tilde{x}_i'\tilde{\beta})^2\}$. Thus, we have ${\rm E}\{Q(\tilde{x},\tilde{\beta},Y) - Q_{\rm o}(\tilde{x},\tilde{\beta},y_{\rm o})|y_{\rm o},x,\beta\} = \delta_0\{(n-m)\sigma^2 + \sum_{i=m+1}^n(x_i'\beta - \tilde{x}_i'\tilde{\beta})^2\}$, hence ${\rm E}\{Q(x,\beta,Y) - Q_{\rm o}(x,\beta,y_{\rm o})|y_{\rm o},x,\beta\} = \delta_0(n-m)\sigma^2$. It follows that *A3* of JNR is satisfied.

Proof of Theorem 1:

We verify conditions (a)–(c) of GCT. Note that here $x = \psi$, and $\mathcal{X} = \Psi$; the algorithm

a is defined below (12) and repeated in *A4* of JNR.

Condition (a) clearly holds under *A1*.

For condition (b), let $\tilde{\psi}^{(t)} \in a(\psi^{(t)})$. Suppose that $\psi^{(t)} \to \psi_0$ and $\tilde{\psi}^{(t)} \to \tilde{\psi}_0$, as $t \to \infty$. We know that, for any $\psi \in \Psi$, we have

$$\mathrm{E}\{Q(\tilde{\psi}^{(t)}, Y)|y_{\mathrm{o}}, \psi^{(t)}\} + p(\tilde{M}^{(t)}) \leq \mathrm{E}\{Q(\psi, Y)|y_{\mathrm{o}}, \psi^{(t)}\} + p(M) \qquad (A.3)$$

[definition of $\tilde{\psi}^{(t)}$]. On the other hand, when t is large, we have $M^{(t)} = M_0$ and $\tilde{M}^{(t)} = \tilde{M}_0$. This is because $\mathcal{M}$ is a discrete space, i.e., the ID numbers of the candidate models, such as $1, 2, \ldots, K$. Thus, when t is large, the left side of (A.3) is equal to

$$\begin{aligned}&\mathrm{E}\{Q(\tilde{\psi}_0, Y)|y_{\mathrm{o}}, \psi_0\} + p(\tilde{M}_0) + \mathrm{E}\{Q(\tilde{\psi}_0, Y)|y_{\mathrm{o}}, M_0, \theta^{(t)}\} - \mathrm{E}\{Q(\tilde{\psi}_0, Y)|y_{\mathrm{o}}, M_0, \theta_0\} \\ &+\mathrm{E}\{Q(\tilde{M}_0, \tilde{\theta}^{(t)}, Y) - Q(\tilde{M}_0, \tilde{\theta}_0, Y)|y_{\mathrm{o}}, M_0, \theta^{(t)}\}.\end{aligned}$$

The last two differences of the above expression go to zero by *A2*. On the other hand, when t is large, the right side of (A.3) is equal to

$$\mathrm{E}\{Q(\psi, Y)|y_{\mathrm{o}}, \psi_0\} + p(M) + \mathrm{E}\{Q(\psi, Y)|y_{\mathrm{o}}, M_0, \theta^{(t)}\} - \mathrm{E}\{Q(\psi, Y)|y_{\mathrm{o}}, M_0, \theta_0\}.$$

Again, the last difference of the above expression goes to zero by *A2*. Thus, by letting $t \to \infty$ on both sides of (A.3), we get $\mathrm{E}\{Q(\tilde{\psi}_0, Y)|y_{\mathrm{o}}, \psi_0\} + p(\tilde{M}_0) \leq \mathrm{E}\{Q(\psi, Y)|y_{\mathrm{o}}, \psi_0\} + p(M)$, for any $\psi \in \Psi$. Therefore, by the definition of a, we have $\tilde{\psi}_0 \in a(\psi_0)$.

For condition (c), we have $g(M^{(t+1)}, \theta^{(t+1)}, y_{\mathrm{o}}) =$

$$\begin{aligned}&= \mathrm{E}\{Q(M^{(t+1)}, \theta^{(t+1)}, Y) + p(M^{(t+1)})|y_{\mathrm{o}}, M^{(t)}, \theta^{(t)}\} \\ &\quad +\mathrm{E}\{Q_{\mathrm{o}}(M^{(t+1)}, \theta^{(t+1)}), y_{\mathrm{o}}) - Q(M^{(t+1)}, \theta^{(t+1)}, Y)|y_{\mathrm{o}}, M^{(t)}, \theta^{(t)}\} \\ &\leq \mathrm{E}\{Q(M^{(t)}, \theta^{(t)}, Y) + p(M^{(t)})|y_{\mathrm{o}}, M^{(t)}, \theta^{(t)}\} \\ &\quad +\mathrm{E}\{Q_{\mathrm{o}}(M^{(t)}, \theta^{(t)}, y_{\mathrm{o}}) - Q(M^{(t)}, \theta^{(t)}, Y)|y_{\mathrm{o}}, M^{(t)}, \theta^{(t)}\}\end{aligned}$$

$= g(M^{(t)}, \theta^{(t)}, y_{\mathrm{o}})$. The inequality above is due to *A3* of JNR and the definition of $M^{(t+1)}$, $\theta^{(t+1)}$. Also note that $M^{(t)}$ and $\theta^{(t)}$ are functions of y_{o}, conditional on which, $M^{(t)}, \theta^{(t)}$ are

considered as fixed models and parameters for every t. This proves the second part of (c). As for the first part, if $\psi_1 \in \Psi \setminus \Psi_0$, then, by *A4* of JNR, $\psi_1 \notin \Psi_1$. Let $\tilde{\psi}_1 \in a(\psi_1)$. Because $\psi_1 \notin a(\psi_1)$, we have $\mathrm{E}\{Q(\psi_1, Y)|y_\mathrm{o}, \psi_1\} + p(M_1) > \min_{\psi\in\Psi}[\mathrm{E}\{Q(\psi, Y)|y_\mathrm{o}, \psi_1\} + p(M)] = \mathrm{E}\{Q(\tilde{\psi}_1, Y)|y_\mathrm{o}, \psi_1\} + p(\tilde{M}_1)$. Thus, combined with *A3* of JNR, we have

$$\begin{aligned} g(\tilde{\psi}_1, y_\mathrm{o}) &= Q_\mathrm{o}(\tilde{\psi}_1, y_\mathrm{o}) + p(\tilde{M}_1) \\ &= \mathrm{E}\{Q(\tilde{\psi}_1, Y)|y_\mathrm{o}, \psi_1\} + p(\tilde{M}_1) + \mathrm{E}\{Q_\mathrm{o}(\tilde{\psi}_1, y_\mathrm{o}) - Q(\tilde{\psi}_1, Y)|y_\mathrm{o}, \psi_1\} \\ &< \mathrm{E}\{Q(\psi_1, Y)|y_\mathrm{o}, \psi_1\} + p(M_1) + \mathrm{E}\{Q_\mathrm{o}(\psi_1, y_\mathrm{o}) - Q(\psi_1, Y)|y_\mathrm{o}, \psi_1\} \\ &= Q_\mathrm{o}(\psi_1, y_\mathrm{o}) + p(M_1) \\ &= g(\psi_1, y_\mathrm{o}). \end{aligned}$$

Thus, we have verified conditions (a)–(c) of GCT. Furthermore, by *A5*, the solution to $\min_{\psi\in\Psi} g(\psi, y_\mathrm{o})$ is unique; therefore, the limit of any convergent subsequence of $\psi^{(t)}, t = 0, 1, 2, \ldots$ is the same, hence the entire sequence converges and the limit does not depend on the initial point. In other words, $\psi^{(t)}, t = 0, 1, 2, \ldots$ converges globally. ∎

Before proving Theorem 2, we interpret the additional assumptions, *A6* and *A7* of JNR. A nice feature of the latter is that they are very much the same as those required for the consistency of the GIC with the complete data (e.g., Jiang & Rao 2003). Intuitively, assumption *A6* states that, w.p.$\to 1$, an underfit model has a larger measure of lack-of-fit, and the difference in the measure of lack-of-fit is of higher order than that in the penalty; assumption *A7* states that, w.p.$\to 1$, the difference in the penalty between an overfit model and the optimal model dominates the that in the measure of lack-of-fit between these two models. These assumptions, of course, make sense. Again, we illustrate with an example.

Example 5 (continued). In this case, we have $Q_\mathrm{o}(M, \theta, y_\mathrm{o}) = Q_\mathrm{o}(x, \beta, y_\mathrm{o}) = |y_\mathrm{o} - X_\mathrm{o}\beta|^2$, where $y_\mathrm{o} = (y_i)_{1\le i\le m}$ and $X_\mathrm{o} = (x_i')_{1\le i\le m}$, and $p(M) = \lambda_n \dim(\beta)$. Let $x_j, j = 1, \ldots, p$ be the candidate x variables. We make the classical assumptions that p is bounded, and that the true variables are a subset of the candidate variables. Let $M_\mathrm{opt} = \{1 \le j \le p, \beta_{\mathrm{f},j} \ne 0\}$, where $\beta_{\mathrm{f},j}$ is the jth true coefficient under the full model. Then, M_opt is the

optimal model. A model M corresponds to a subset of $\{1, \ldots, p\}$; $M \in \mathcal{M}_{\rm u}$ iff $M_{\rm opt} \not\subseteq M$, and $M \in \mathcal{M}_{\rm o}$ iff $M_{\rm opt} \subset M$ (i.e., $M_{\rm opt} \subseteq M, M_{\rm opt} \neq M$).

If $M \in \mathcal{M}_{\rm u}$, we have $Q_{\rm o}(M, Y_{\rm o}) = Y_{\rm o}' P_{X_{\rm o}^\perp} Y_{\rm o} = |a|^2 + 2a'\epsilon + \epsilon' P_{X_{\rm o}^\perp}\epsilon$, where $a = P_{X_{\rm o}^\perp} X_{\rm o,opt}\beta_{\rm opt}$, with $X_{\rm o,opt}$ and $\beta_{\rm opt}$ being the $X_{\rm o}$ corresponding to $M_{\rm opt}$ and the true β corresponding to $X_{\rm o,opt}$, respectively. On the other hand, we have $Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o}) = \epsilon'\epsilon = \epsilon' P_{X_{\rm o}^\perp}\epsilon + \epsilon' P_{X_{\rm o}}\epsilon$. Thus, $Q_{\rm o}(M, Y_{\rm o}) > Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})$ iff $|a|^2 + 2a'\epsilon > \epsilon' P_{X_{\rm o}}\epsilon$. However, it is easy to show that $a'\epsilon = O_{\rm P}(|a|)$ and $\epsilon' P_{X_{\rm o}}\epsilon = O_{\rm P}(r)$ with $r = \mathrm{rank}(X_{\rm o})$. Thus, provided that $|a| \to \infty$ as $n \to \infty$, *A6* of JNR is is satisfied.

If $M \in \mathcal{M}_{\rm o}$, then $\mathcal{L}(X_{\rm o,opt}) \subset \mathcal{L}(X_{\rm o})$, where $\mathcal{L}(H)$ denotes the linear space spanned by the columns of matrix H, and $\Delta = \dim(\beta) - \dim(\beta_{\rm opt}) > 0$, where β corresponds to $X_{\rm o}$. It follows that $P_{X_{\rm o}} - P_{X_{\rm o,opt}}$ is a projection matrix. Thus, we have $Q_{\rm o}(M_{\rm opt}, Y_{\rm o}) - Q_{\rm o}(M, Y_{\rm o}) = Y_{\rm o}'(P_{X_{\rm o,opt}^\perp} - P_{X_{\rm o}^\perp})Y_{\rm o} = \epsilon'(P_{X_{\rm o,opt}^\perp} - P_{X_{\rm o}^\perp})\epsilon = \epsilon'(P_{X_{\rm o}} - P_{X_{\rm o,opt}})\epsilon = \sigma^2\chi_\Delta$, where χ^2_Δ has the χ^2 distribution with Δ degrees of freedom (e.g., Jiang 2007, p. 238). It follows that *A7* of JNR is satisfies as long as $\lambda_n \to \infty$, as $n \to \infty$.

Proof of Theorem 2:

Let (M_0, θ_0) denote the limit of the E-MS convergence. According to GCT, and the proof of Theorem 1, (M_0, θ_0) must be a minimizer of $g(M, \theta, y_{\rm o})$. Define $g(M, y_{\rm o}) = \inf_{\theta \in \Theta_M} g(M, \theta, y_{\rm o})$. Then, it is easy to show that $g(M_0, \theta_0, y_{\rm o}) = g(M_0, y_{\rm o})$.

Consider the event $\{M_0 = M\}$. If $M \in \mathcal{M}_{\rm u}$, by *A6* of JNR, we have

$$
\begin{aligned}
& g(M_0, \theta_0, Y_{\rm o}) - g(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o}) \\
= \;& g(M_0, Y_{\rm o}) - g(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o}) \\
= \;& Q_{\rm o}(M_0, Y_{\rm o}) + p(M_0) - Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o}) - p(M_{\rm opt}) \\
= \;& \{Q_{\rm o}(M_0, Y_{\rm o}) - Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})\} \left\{1 + \frac{p(M_0) - p(M_{\rm opt})}{Q_{\rm o}(M_0, Y_{\rm o}) - Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})}\right\} \\
= \;& \{Q_{\rm o}(M, Y_{\rm o}) - Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})\} \left\{1 + \frac{p(M) - p(M_{\rm opt})}{Q_{\rm o}(M, Y_{\rm o}) - Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o})}\right\},
\end{aligned}
$$

which is positive w.p.$\to 1$. Therefore, w.p.$\to 1$, M_0 cannot be M, that is, $\mathrm{P}(M_0 = M) \to 0$. On the other hand, if $M \in \mathcal{M}_{\rm o}$, by *A7* of JNR, we have, w.p.$\to 1$, $p(M_0) - p(M_{\rm opt}) =$

$p(M) - p(M_{\rm opt}) > Q_{\rm o}(M_{\rm opt}, Y_{\rm o}) - Q_{\rm o}(M, Y_{\rm o}) = Q_{\rm o}(M_{\rm opt}, Y_{\rm o}) - Q_{\rm o}(M_0, Y_{\rm o})$; therefore, there is $\tilde{\theta}_{\rm opt} \in \Theta_{M_{\rm opt}}$ such that $p(M_0) - p(M_{\rm opt}) > Q_{\rm o}(M_{\rm opt}, \theta_{\rm opt}, Y_{\rm o}) - Q_{\rm o}(M_0, \theta_0, Y_{\rm o})$, or $g(M_0, \theta_0, Y_{\rm o}) > g(M_{\rm opt}, \tilde{\theta}_{\rm opt}, Y_{\rm o})$, which contradicts the minimization property of (M_0, θ_0) as noted above. Thus, again, w.p.$\to 1$, M_0 cannot be M, that is, $\mathrm{P}(M_0 = M) \to 0$. The arguments, and *A5* of JNR, have shown that $\mathrm{P}(M_0 = M_{\rm opt}) \to 1$. ∎

Next, we extend the convergence and consistency results to with AF. To be specific, assume that the minimum-dimension criterion is used to select the model within the fence (e.g., Jiang 2014). Again, assume the existence of $M_{\rm opt} \in \mathcal{M}$. Also assume that model-based (or parametric) bootstrap is used in the AF procedure, with the bootstrap sample size B. Let $M^{(t)}$ be the current model. We assume that, given a model, the method of parameter estimation is determined (e.g., maximum likelihood with the E-M algorithm), so that we can simply use $P^*(\cdot|M)$ for the bootstrap empirical probability under model M. Let $M^{(t)}$ be the current model. Then, for any cut-off c among a grid of values, $\mathcal{C}$, the updated model, $M^{(t+1)}$, is such that, for some $c^* \in \mathcal{C}$, $\mathrm{P}^*\{M_{c^*} = M^{(t+1)}|M^{(t)}\} = \max_{M\in\mathcal{M}} \mathrm{P}^*\{M_{c^*} = M|M^{(t)}\} = \max_{c\in\mathcal{C}} \max_{M\in\mathcal{M}} \mathrm{P}^*\{M_c = M|M^{(t)}\}$, where M_c is the model selected by the fence for the cut-off c. Note that $\mathcal{C}$ is usually chosen so that the fence does not yield a trivial solution, that is, the minimum model, M_0, or the full model, $M_{\rm f}$. The following theorem assumes consistency of the AF, for which sufficient conditions are given, e.g., in Jiang *et al.* (2008) (again, the numbering of the theorems follows the sequence of JNR).

Theorem 3. Provided that the AF is consistent when bootstrapping under either $M_{\rm f}$ or $M_{\rm opt}$, as $n, B \to \infty$, then, w.p.$\to 1$ as $n, B \to \infty$, the E-MS with AF converges within two iterations when starting with $M_{\rm f}$. Furthermore, the limit of the convergence is $M_{\rm opt}$; in other words, the E-MS with AF is consistent.

A nice feature of Theorem 3 is that it links the convergence of E-MS with AF to the consistency of AF, and shows that the convergence can be very fast (in two iterations) with the limit being the optimal model. In our simulation study (see Subsection A.6.1), the E-MS with AF converges in 2-3 iterations in all of our simulation runs. On the other hand,

unlike Theorem 1, the convergence in Theorem 3 is not global, because the starting model is assumed to be $M_{\rm f}$. In fact, as is seen in the proof below, the GCT is not used in the proof of Theorem 3. The starting model may be replaced by any overfitting model, but not by an arbitrary one.

Proof: Let $\mathcal{P}$ denote the joint probability distribution of the data and bootstrap samples. Then, we have $\mathcal{P}\{M^{(1)} = M_{\rm opt}|M^{(0)} = M_{\rm f}\} \to 1$, as $n, B \to \infty$. Also note that, given $M^{(1)}$, the outcome of $M^{(2)}$ does not depend on $M^{(0)}$. Thus, we have $\mathcal{P}\{M^{(2)} = M_{\rm opt}|M^{(1)} = M_{\rm opt}, M^{(0)} = M_{\rm f}\} = \mathcal{P}\{M^{(2)} = M_{\rm opt}|M^{(1)} = M_{\rm opt}\} \to 1$, as $n, B \to \infty$. Therefore, we have $\mathcal{P}\{\text{E-MS converges in 2 iterations}|M^{(0)} = M_{\rm f}\} \geq \mathcal{P}\{M^{(2)} = M^{(1)}|M^{(0)} = M_{\rm f}\} \geq \mathcal{P}\{M^{(2)} = M_{\rm opt}, M^{(1)} = M_{\rm opt}|M^{(0)} = M_{\rm f}\} = \mathcal{P}\{M^{(1)} = M_{\rm opt}|M^{(0)} = M_{\rm f}\}\mathcal{P}\{M^{(2)} = M_{\rm opt}|M^{(1)} = M_{\rm opt}, M^{(0)} = M_{\rm f}\} \to 1$, as $n, B \to \infty$.

Also note that the E-MS stops whenever $M^{(t+1)} = M^{(t)}$ takes place, in which case the $M^{(t+1)}$ is the limit of convergence. Thus, we have $\mathcal{P}\{\text{the E-MS limit is } M_{\rm opt}|M^{(0)} = M_{\rm f}\} \geq \mathcal{P}\{M^{(2)} = M_{\rm opt}, M^{(1)} = M_{\rm opt}|M^{(0)} = M_{\rm f}\} \to 1$, as $n, B \to \infty$, according to the previous argument, if $M_{\rm f} \neq M_{\rm opt}$. On the other hand, if $M_{\rm f} = M_{\rm opt}$, then, again by the previous argument, we have $\mathcal{P}\{\text{the E-MS limit is } M_{\rm opt}|M^{(0)} = M_{\rm f}\} \geq \mathcal{P}\{M^{(1)} = M_{\rm opt}|M^{(0)} = M_{\rm opt}\} \to 1$, as $n, B \to \infty$. This proves the consistency. ∎

A modification of the E-MS with AF, however, can actually achieve the global convergence. This is done by restricting the model space for $M^{(t+1)}$ to be submodels of $M^{(t)}$. This is not unreasonable because the bootstrap samples are drawn under $M^{(t)}$, which would suggest that $M^{(t)}$ is believed to be a true model; therefore, there is no need to look for anything beyond the submodels of $M^{(t)}$. The modified E-MS with AF is also computationally more attractive, because the model space shrinks with each iteration. As for the convergence property, we have the following result.

Theorem 4. If $\mathcal{M}$ is finite, then the modified E-MS with AF converges globally.

Proof: If the convergence is not achieve, then each time the update is a true submodel. As the iteration goes on, this would generate a squence of non-repeating models

$M^{(0)}, M^{(1)}, \ldots$. Then, because $\mathcal{M}$ is finite, the sequence cannot go on forever, so, at some point, the convergence has to take place, and this is regardless of the starting model. ∎

A.2 Example

We verify the conditions *A1–A5* for Example 5 in the previous subsection. We let δ_0 unspecified, for now, and determine it later. Here $M = x$, and $\theta_M = \beta$, the vector of regression coefficients corresponding to x. Thus, *A1* is satisfied if the number of candidate predictors is finite, and the range of any regression coefficient is bounded.

For *A2*, it is easy to show that $\mathrm{E}\{Q(x_1, \tilde{\beta}_1, Y) - Q(x_1, \beta_1, Y)|y_\mathrm{o}, x_0, \tilde{\beta}_0\} = \sum_{i=1}^m d_i + \delta_0 \sum_{i=m+1}^n d_i$, where $d_i = [x'_{1,i}(\tilde{\beta}_1+\beta_1)-2\{y_i 1(i \le m)+x'_{0,i}\tilde{\beta}_0 1_{(i>m)}\}]x'_{1,i}(\tilde{\beta}_1-\beta_1)$, which goes to zero as $\tilde{\beta}_j \to \beta_j, j = 0, 1$. Furthermore, we have $\mathrm{E}\{Q(x_1, \beta_1, Y)|y_\mathrm{o}, x_0, \tilde{\beta}_0\} =$

$$\sum_{i=1}^m (y_i - x'_{1,i}\beta_1)^2 + \delta_0 \left\{ (n-m)\sigma^2 + \sum_{i=m+1}^n (x'_{0,i}\tilde{\beta}_0 - x'_{1,i}\beta_1)^2 \right\},$$

and $\mathrm{E}\{Q(x_1, \beta_1, Y)|y_\mathrm{o}, x_0, \beta_0\}$ has the same expression with $\tilde{\beta}_0$ replaced by β_0. Thus, $\mathrm{E}\{Q(x_1, \beta_1, Y)|y_\mathrm{o}, x_0, \tilde{\beta}_0\} \to \mathrm{E}\{Q(x_1, \beta_1, Y)|y_\mathrm{o}, x_0, \beta_0\}$ as $\tilde{\beta}_j \to \beta_j, j = 0, 1$.

A3 holds according to Example 5.

For *A5*, we assume, for simplicity, that all the x columns are linearly independent, so that $|M| = \#\mathrm{col}(x)$, where $\#\mathrm{col}(x)$ is the number of columns in x, and $p(M) = \lambda|M|$, where λ is a penalty parameter. Then, we have $\Psi_0 = \mathrm{argmin}_{\psi\in\Psi}\{|y_\mathrm{o} - X_{(1)}\beta|^2 + \lambda\#\mathrm{col}(x)\}$, where $y_\mathrm{o} = (y_i)_{1\le i\le m}$ and $X_{(1)} = (x'_i)_{1\le i\le m}$. Thus, *A5* says that there is a unique $\psi_0 = (x_0, \beta_0)$ that minimizes $g(\psi) = |y_\mathrm{o} - X_{(1)}\beta|^2 + \lambda\#\mathrm{col}(x)$ for all $\psi = (x, \beta)$. To see what this means, note that, for fixed x, the minimum of $g(\psi)$ over β is $G(x) = |P_{X^\perp_{(1)}} y_\mathrm{o}|^2 + \lambda\#\mathrm{col}(x)$, where $P_{X^\perp} = I - P_X$ with $P_X = X(X'X)^{-1}X'$. Thus, $g(\psi)$ has a unique minimizer ψ_0 if and only if $G(x)$ has a unique minimizer x_0, in which case β_0 is the least squares (LS) solution corresponding to x_0, i.e., $\beta_0 = \{X'_{0,(1)}X_{0,(1)}\}^{-1}X'_{0,(1)}y_\mathrm{o}$.

As for *A4*, suppose that $\psi_1 = (x_1, \beta_1) \ne \psi_0$. We consider two cases.

Case I: $\beta_1 \neq \{X'_{1,(1)}X_{1,(1)}\}^{-1}X'_{1,(1)}y_{\rm o}$. Consider the minimizer of

$$\begin{aligned} & {\rm E}\{Q(x_1,\beta,Y)|y_{\rm o},x_1,\beta_1\}+\lambda\#{\rm col}(x_1) \\ = \;& |y_{\rm o}-X_{1,(1)}\beta|^2+\delta_0\{(n-m)\sigma^2+|X_{1,(2)}\beta_1-X_{1,(2)}\beta|^2\}+\lambda\#{\rm col}(x_1) \end{aligned}$$

over β, where $X_{(2)}=(x'_i)_{m+1\le i\le n}$. It can be shown that the solution is

$$\hat{\beta}_1=\{X'_{1,(1)}X_{1,(1)}+X'_{1,(2)}X_{1,(2)}\}^{-1}\{X'_{1,(1)}y_{\rm o}+X'_{1,(2)}X_{1,(2)}\beta_1\}.$$

We can assume that, with probability tending to one, all the LS solutions are within a compact parameter space (the compact space may expand with the sample size), which holds under the standard assumptions. If $\hat{\beta}_1=\beta_1$, it follows that $\beta_1=\{X'_{1,(1)}X_{1,(1)}\}^{-1}X'_{1,(1)}y_{\rm o}$, a contradiction. Thus, we must have $\beta_1\neq\hat{\beta}_1$, hence

$$\begin{aligned} & {\rm E}\{Q(x_1,\beta,Y)|y_{\rm o},x_1,\beta_1\}+\lambda\#{\rm col}(x_1)|_{\beta=\hat{\beta}_1} \\ < \;& {\rm E}\{Q(x_1,\beta,Y)|y_{\rm o},x_1,\beta_1\}+\lambda\#{\rm col}(x_1)|_{\beta=\beta_1}, \end{aligned}$$

which implies that $\psi_1\notin a(\psi_1)$.

Case II: $\beta_1=\{X'_{1,(1)}X_{1,(1)}\}^{-1}X'_{1,(1)}y_{\rm o}$. Then, we have

$$|P_{X^{\perp}_{0,(1)}}y_{\rm o}|^2+\lambda\#{\rm col}(x_0)=g(\psi_0)<g(\psi_1)=|P_{X^{\perp}_{1,(1)}}y_{\rm o}|^2+\lambda\#{\rm col}(x_1). \qquad (A.4)$$

Let $\Delta(x_1)$ denote the difference between the right side of (A.4) and the left side, and

$$D(x_1)=|[X_{1,(2)}\{X'_{1,(1)}X_{1,(1)}\}^{-1}X'_{1,(1)}-X_{0,(2)}\{X'_{0,(1)}X_{0,(1)}\}^{-1}X'_{0,(1)}]y_{\rm o}|^2.$$

Then, we have

$$\begin{aligned} & {\rm E}\{Q(x_0,\beta_0,Y)|y_{\rm o},x_1,\beta_1\}+\lambda\#{\rm col}(x_0) \\ = \;& |P_{X^{\perp}_{0,(1)}}y_{\rm o}|^2+\lambda\#{\rm col}(x_0)+\delta_0\{(n-m)\sigma^2+|X_{1,(2)}\beta_1-X_{0,(2)}\beta_0|^2\} \\ = \;& |P_{X^{\perp}_{1,(1)}}y_{\rm o}|^2+\lambda\#{\rm col}(x_1)+\delta_0(n-m)\sigma^2-\Delta(x_1)+\delta_0 D(x_1) \\ \le \;& |P_{X^{\perp}_{1,(1)}}y_{\rm o}|^2+\lambda\#{\rm col}(x_1)+\delta_0(n-m)\sigma^2-\min_{x_1\neq x_0}\Delta(x_1)+\delta_0\max_{x_1\neq x_0}D(x_1). \end{aligned}$$

Thus, if we take $\delta_0 = \{\min_{x_1 \neq x_0} \Delta(x_1)\}\{2 \max_{x_1 \neq x_0} D(x_1)\}^{-1}$, which is positive by *A5*, then, continuing on, we have

$$\begin{aligned} & \mathrm{E}\{Q(x_0, \beta_0, Y)|y_{\mathrm{o}}, x_1, \beta_1\} + \lambda \#\mathrm{col}(x_0) \\ = \; & |P_{X_{1,(1)}^{\perp}} y_{\mathrm{o}}|^2 + \lambda \#\mathrm{col}(x_1) + \delta_0 (n-m)\sigma^2 - \frac{1}{2} \min_{x_1 \neq x_0} \Delta(x_1) \\ < \; & |P_{X_{1,(1)}^{\perp}} y_{\mathrm{o}}|^2 + \lambda \#\mathrm{col}(x_1) + \delta_0 (n-m)\sigma^2 \\ = \; & \mathrm{E}\{Q(x_1, \beta_1, y)|y_{\mathrm{o}}, x_1, \beta_1\} + \lambda \#\mathrm{col}(x_1). \end{aligned}$$

Therefore, again, we have $\psi_1 \notin a(\psi_1)$.

A.3 The FW/BW BIC

We explain the FW/BW BIC procedure (Browman & Speed 2002) using an example. Suppose that there are 30 candidate variables. First carry out the forward selection by selecting the first variable, $x^{(1)}$, that minimizes $\mathrm{RSS}(y, X) = \min_{\beta} \mathrm{RSS}(y, X, \beta)$, where

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

with $X = (x_i')_{1 \leq i \leq n}$, over all X that consists of a single column of x; next, we select the second variable, $x^{(2)}$, that minimizes $\mathrm{RSS}(y, X)$ over all X that consists of two columns of x, with the first column being $x^{(1)}$ (so the selection is for the second column only); we then select the third variable, $x^{(3)}$, that minimizes $\mathrm{RSS}(y, X)$ over all X that consists of three columns, with the first two columns being $x^{(1)}, x^{(2)}$ (so the selection is for the third columm only), and so on. The forward selection continues until 50% of the variables are selected. Thus, we have $x^{(1)}, \ldots, x^{(15)}$ after the forward selection. We then follow with a backward elimination by taking out the first variable, $x_{(1)}$, from the 15 variables, that minimizes $\mathrm{RSS}(y, \{x^{(1)}, \ldots, x^{(15)} \text{ without } x\})$ over $x \in \{x^{(1)}, \ldots, x^{(15)}\}$; we then take out the second variable, $x_{(2)}$, from the remaining 14 variables, that minimizes $\mathrm{RSS}(y, \{x^{(1)}, \ldots, x^{(15)} \text{ without } x_{(1)} \text{ and } x\})$ over $x \in \{x^{(1)}, \ldots, x^{(15)}\} \setminus \{x_{(1)}\}$; and so on. The backward elimination continues until all 15 variables are taken out. We then apply the BIC to the following reduced model space generated by the FW/BW procedure:

$M_1 = \{x^{(1)}\}$, $M_2 = \{x^{(1)}, x^{(2)}\}, \ldots, M_{15} = \{x^{(1)}, \ldots, x^{(15)}\}$, $M_{16} = \{x^{(1)}, \ldots, x^{(15)}\} \setminus \{x_{(1)}\}, \ldots, M_{29} = \{x^{(1)}, \ldots, x^{(15)}\} \setminus \{x_{(1)}, \ldots, x_{(14)}\}$.

A.4 Some derivations in Section 3

A.4.1 Some details regarding Example 2

We first derive the conditional expectations of BIC and RSS under the current model, $M_{\rm c}$. Note that θ is involved only in l_x, hence, the MLE of θ involves only the x data. Therefore, we first update the estimate of θ based on the x data. Let $\hat{\theta}_{\rm c}$ denote the current estimate. Then, we have, by independence,

$$\mathrm{E_c}(l_X|x_{\rm o}) \propto nr(q-1)\log(1-\theta) + \{\log\theta - \log(1-\theta)\}\sum_{i=1}^{n}\mathrm{E_c}\{s(X_i)|x_{{\rm o},i}\},$$

where $X_i = (X_{ijk})_{1\le j\le q, 1\le k\le r}$, $s(X_i) = \sum_{k=1}^{r}\sum_{j=1}^{q-1}(x_{ijk} + x_{i,j+1,k} - 2x_{ijk}x_{i,j+1,k}) = \#$ of cases among $x_{ijk}, 1\le k\le r, 1\le j\le q-1$ such that $|x_{ijk} - x_{i,j+1,k}| = 1$, $\mathrm{E_c}$ means (conditional) expectation under the current estimates, and $x_{{\rm o},i}$ denotes all of the observed x's among $x_{ijk}, 1\le j\le q, 1\le k\le r$. Similarly, let $x_{{\rm m},i}$ denote all of the missing x's among $x_{ijk}, 1\le j\le q, 1\le k\le r$. Let $f_{\rm c}$ denote the (conditional) pmf, or pdf, under the current estimates. It can be shown that

$$f_{\rm c}(x_i) = \frac{(1-\hat{\theta}_{\rm c})^{r(q-1)}}{2^r}\left(\frac{\hat{\theta}_{\rm c}}{1-\hat{\theta}_{\rm c}}\right)^{s(x_i)},$$

where $x_i = (x_{ijk})_{1\le j\le q, 1\le k\le r}$. Thus, we have $f_{\rm c}(x_{{\rm m},i}|x_{{\rm o},i}) = f_{\rm c}(x_i)/\sum_{x_{{\rm m},i}} f_{\rm c}(x_i) = \hat{\gamma}_{\rm c}^{s(x_i)}/\sum_{x_{{\rm m},i}}\hat{\gamma}_{\rm c}^{s(x_i)}$, where $\hat{\gamma}_{\rm c} = \hat{\theta}_{\rm c}/(1-\hat{\theta}_{\rm c})$, and $\sum_{x_{{\rm m},i}}$ is over all the missing x's among $x_{ijk}, 1\le j\le q, 1\le k\le r$ (each taking the value of 0 or 1). It follows that

$$\mathrm{E_c}\{s(X_i)|x_{{\rm o},i}\} = \frac{\sum_{x_{{\rm m},i}} s(x_i)\hat{\gamma}_{\rm c}^{s(x_i)}}{\sum_{x_{{\rm m},i}}\hat{\gamma}_{\rm c}^{s(x_i)}}.$$

Thus, it is easy to obtain the updated for the θ estimate as

$$\hat{\theta} = \frac{1}{nr(q-1)}\sum_{i=1}^{n}\frac{\sum_{x_{{\rm m},i}} s(x_i)\hat{\gamma}_{\rm c}^{s(x_i)}}{\sum_{x_{{\rm m},i}}\hat{\gamma}_{\rm c}^{s(x_i)}}.$$

As a comparison, note that the MLE of θ based on all of the x's is $\sum_{i=1}^{n} s(x_i)/nr(q-1)$.

Now assume that the MLE $\hat{\theta}$ has been obtained via the E-M algorithm. To evaluate the conditional expectations, we need to obtain $f_{\rm c}(y_{{\rm m},i}, x_{{\rm m},i}|y_{{\rm o},i}, x_{{\rm o},i})$. Note that

$$f_{\rm c}(y_{{\rm m},i}, x_{{\rm m},i}|y_{{\rm o},i}, x_{{\rm o},i}) = \frac{f_{\rm c}(y_i, x_i)}{f_{\rm c}(y_{{\rm o},i}, x_{{\rm o},i})} = \frac{f_{\rm c}(y_i|x_i)f_{\rm c}(x_i)}{f_{\rm c}(y_{{\rm o},i}, x_{{\rm o},i})},$$

and $f_{\rm c}(y_i|x_i) = (2\pi\hat{\sigma}_{\rm c}^2)^{-1/2}\exp\{-(y_i - x'_{{\rm f},i}\hat{\beta}_{\rm f,c})^2/2\hat{\sigma}_{\rm c}^2\}$. If y_i is observed, then, we have $f_{\rm c}(y_{{\rm o},i}, x_{{\rm o},i}) = \sum_{x_{{\rm m},i}} f_{\rm c}(y_i, x_i) = \sum_{x_{{\rm m},i}} f_{\rm c}(y_i|x_i)f_{\rm c}(x_i)$. Thus, we have

$$\begin{aligned} f_{\rm c}(y_{{\rm m},i}, x_{{\rm m},i}|y_{{\rm o},i}, x_{{\rm o},i}) &= \frac{f_{\rm c}(y_i|x_i)f_{\rm c}(x_i)}{\sum_{x_{{\rm m},i}} f_{\rm c}(y_i|x_i)f_{\rm c}(x_i)} \\ &= \frac{\exp\{-(2\hat{\sigma}_{\rm c}^2)^{-1}(y_i - x'_{{\rm f},i}\hat{\beta}_{\rm f,c})^2\}\hat{\gamma}^{s(x_i)}}{\sum_{x_{{\rm m},i}} \exp\{-(2\hat{\sigma}_{\rm c}^2)^{-1}(y_i - x'_{{\rm f},i}\hat{\beta}_{\rm f,c})^2\}\hat{\gamma}^{s(x_i)}}, \end{aligned}$$

where $\hat{\gamma} = \hat{\theta}/(1-\hat{\theta})$ (note that $\hat{\gamma}$ does not change with the iteration). Define

$$w_{{\rm o},i}(y_i, x_i) = \frac{\exp\{-(2\hat{\sigma}_{\rm c}^2)^{-1}(y_i - x'_{{\rm c},i}\hat{\beta}_{\rm c})^2\}\hat{\gamma}^{s(x_i)}}{\sum_{x_{{\rm m},i}} \exp\{-(2\hat{\sigma}_{\rm c}^2)^{-1}(y_i - x'_{{\rm c},i}\hat{\beta}_{\rm c})^2\}\hat{\gamma}^{s(x_i)}},$$

and $w_{{\rm m},i}(x_i) = \hat{\gamma}^{s(x_i)}/\sum_{x_{{\rm m},i}}\hat{\gamma}^{s(x_i)}$. It can be shown that $\mathrm{E_c}\{\mathrm{RSS}(Y, X, \beta)|y_{\rm o}, x_{\rm o}\} = n_{\rm m}\hat{\sigma}_{\rm c}^2 + \sum_{i\in I_{\rm o}}\sum_{x_{{\rm m},i}} w_{{\rm o},i}(y_i, x_i)(y_i - x'_i\beta)^2 + \sum_{i\in I_{\rm m}}\sum_{x_{{\rm m},i}} w_{{\rm m},i}(x_i)(x'_{{\rm c},i}\hat{\beta}_{\rm c} - x'_i\beta)^2$, where $\mathrm{E_c}$ represents the conditional expectation under $M_{\rm c}$; $x_{{\rm c},i}$ is the x_i under $M_{\rm c}$; $\hat{\beta}_{\rm c}, \hat{\sigma}_{\rm c}^2$ are the current estimators under $M_{\rm c}$; $I_{\rm o} = \{1 \le i \le n : y_i \text{ is observed}\}$, $I_{\rm m} = \{1 \le i \le n : y_i \text{ is missing}\}$, and $n_{\rm m} = |I_{\rm m}|$ (cardinality). Note that, like $\hat{\gamma}$, $w_{{\rm m},i}$ does not change with the iterations. Thus, by differentiating and letting the derivative equal to 0, we get

$$\begin{aligned} \hat{\beta} &= \left\{\sum_{i\in I_{\rm o}}\sum_{x_{{\rm m},i}} w_{{\rm o},i}(y_i, x_i)x_i x'_i + \sum_{i\in I_{\rm m}}\sum_{x_{{\rm m},i}} w_{{\rm m},i}(x_i)x_i x'_i\right\}^{-1} \\ &\quad \times \left[\sum_{i\in I_{\rm o}}\sum_{x_{{\rm m},i}} w_{{\rm o},i}(y_i, x_i)x_i y_i + \left\{\sum_{i\in I_{\rm m}}\sum_{x_{{\rm m},i}} w_{{\rm m},i}(x_i)x_i x'_{{\rm c},i}\right\}\hat{\beta}_{\rm c}\right]; \end{aligned}$$

and $\mathrm{RSS_c}(M|y_{\rm o}, x_{\rm o}) \equiv \min_\beta \mathrm{E_c}\{\mathrm{RSS}(Y, X, \beta)|y_{\rm o}, x_{\rm o}\}$

$$\begin{aligned} &\propto \sum_{i\in I_{\rm o}} \frac{\sum_{x_{{\rm m},i}} \exp\{-(2\hat{\sigma}_{\rm c}^2)^{-1}(y_i - x'_{{\rm c},i}\hat{\beta}_{\rm c})^2\}\hat{\gamma}^{s(x_i)}(y_i - x'_i\hat{\beta})^2}{\sum_{x_{{\rm m},i}} \exp\{-(2\hat{\sigma}_{\rm c}^2)^{-1}(y_i - x'_{{\rm c},i}\hat{\beta}_{\rm c})^2\}\hat{\gamma}^{s(x_i)}} \\ &+ \sum_{i\in I_{\rm m}} \frac{\sum_{x_{{\rm m},i}} \hat{\gamma}^{s(x_i)}(x'_{{\rm c},i}\hat{\beta}_{\rm c} - x'_i\hat{\beta})^2}{\sum_{x_{{\rm m},i}} \hat{\gamma}^{s(x_i)}}. \end{aligned}$$

An expression proportional to $\mathrm{E_c}(l_{M,Y|X}|y_\mathrm{o}, x_\mathrm{o})$ can be obtained similarly. It follows that $\hat{\beta}_M = \hat{\beta}$ above with x_i replaced by $x_{M,i}$, and $\hat{\sigma}^2 = n^{-1}\{\sum_{i\in I_\mathrm{o}} \sum_{x_{\mathrm{m},i}} w_{\mathrm{o},i}(y_i, x_i)(y_i - x'_{M,i}\hat{\beta}_M)^2 + n_\mathrm{m}\hat{\sigma}^2_\mathrm{c} + \sum_{i\in I_\mathrm{m}} \sum_{x_{\mathrm{m},i}} w_{\mathrm{m},i}(x_i)(x'_{\mathrm{c},i}\hat{\beta}_\mathrm{c} - x'_{M,i}\hat{\beta}_M)^2\}$. It follows that

$$\begin{aligned}\mathrm{BIC_c}(M|y_\mathrm{o}, x_\mathrm{o}) \quad \propto \quad & n\log\left\{ n_\mathrm{m}\hat{\sigma}^2_\mathrm{c} + \sum_{i\in I_\mathrm{o}}\sum_{x_{\mathrm{m},i}} w_{\mathrm{o},i}(y_i, x_i)(y_i - x'_{M,i}\hat{\beta}_M)^2 \right. \\ & \left. + \sum_{i\in I_\mathrm{m}}\sum_{x_{\mathrm{m},i}} w_{\mathrm{m},i}(x_i)(x'_{\mathrm{c},i}\hat{\beta}_\mathrm{c} - x'_{M,i}\hat{\beta}_M)^2 \right\} + |M|\log(n).\end{aligned}$$

A.4.2 Some details regarding Example 3

We derive the conditional expectation assuming that the current model is M_f. The same derivation applies to any current model, M_c, with only notational changes. Let $\mathrm{E_f}$ denote the conditional expectation under M_f and the current estimates of parameters, under M_f, including β_f, σ^2, $\mu_r, \pi_r, 1 \le r \le s$, and Ω. Let $y_\mathrm{o}, x_\mathrm{o}$ denote the observed y, x, respectively. Similarly, let y_m, x_m, $x_{\mathrm{c,m}}$, and $x_{\mathrm{d,m}}$ denote the missing parts of y, x, x_c, and x_d, respectively. Although it is possible to obtain the conditional density $f_M(y_\mathrm{m}, x_\mathrm{m}|y_\mathrm{o}, x_\mathrm{o})$, 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 $f_{M,i}(y_{i,\mathrm{m}}, x_{i,\mathrm{m}}|y_{i,\mathrm{o}}, x_{i,\mathrm{o}}) \propto f_{M,i}(y_i, x_i) \propto$

$$\exp\left[\sum_{r=1}^{s} 1_{(x_{i,\mathrm{d}}=v_r)}\left\{\log\pi_r - \frac{1}{2}(x_{i,\mathrm{c}} - \mu_r)'\Omega^{-1}(x_{i,\mathrm{c}} - \mu_r)\right\} - \frac{(y_i - x'_{i,M}\beta_M)^2}{2\sigma^2}\right], \quad (A.5)$$

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, as follows. Write $z = (y_{i,\mathrm{m}}, x_{i,\mathrm{m}})$, and $f(z) =$ the right side of (A.5) (note that

$y_{i,\mathrm{o}}, x_{i,\mathrm{o}}$ are held fix in y_i, x_i). Given the current state $z^{(t)}$, (a) draw $z \sim g(z)$, where $g(\cdot)$ is a *trial density* whose expression is known, up to a constant, and from which one knows how to sample; (b) simulate $u \sim \mathrm{Uniform}[0,1]$ and let

$$z^{(t+1)} = \begin{cases} z, & \text{if } u \leq \min\{1, w(z)/w(z^{(t)})\}, \\ z^{(t)}, & \text{otherwise,} \end{cases}$$

where $w(z) = f(z)/g(z)$ is the *importance sampling weight*. The algorithm generates a Markov chain that converges (in distribution) to its stationary distribution, which is the target distribution on the left side of (A.5).

It remains to choose the trial density g. Let $z = (x_{i,\mathrm{d,m}}, x_{i,\mathrm{c,m}}, y_{i,\mathrm{m}})$, where, for the moment, assume that all three components of z are non-empty. We proceed as follows:

(I) First draw $x_{i,\mathrm{d,m}}$ from $f(x_{i,\mathrm{d,m}}|x_{i,\mathrm{d,o}})$. Note that, given the missing value pattern for x_i, each vector v_r is partitioned as $v_{r,i,\mathrm{o}}$ and $v_{r,i,\mathrm{m}}$, with the notation being understood in obvious ways. Then, given $x_{i,\mathrm{d,o}} = v_{r,i,\mathrm{o}}$, the possible values of $x_{i,\mathrm{d,m}}$ are $v_{\tilde{r},i,\mathrm{m}}, 1 \leq \tilde{r} \leq s$ such that $v_{\tilde{r},i,\mathrm{o}} = v_{r,i,\mathrm{o}}$. In other words, define $R(v) = \{1 \leq \tilde{r} \leq s : v_{\tilde{r},i,\mathrm{o}} = v\}$. Then, the possible values of $x_{i,\mathrm{d,m}}$ are $v_{\tilde{r},i,\mathrm{m}}, \tilde{r} \in R(v_{r,i,\mathrm{o}})$. Also, for any $\tilde{r} \in R(v_{r,i,\mathrm{o}})$, we have

$$\mathrm{P}(x_{i,\mathrm{d,m}} = v_{\tilde{r},i,\mathrm{m}}|x_{i,\mathrm{d,o}} = v_{r,i,\mathrm{o}}) = \frac{\mathrm{P}(x_{i,\mathrm{d,m}} = v_{\tilde{r},i,\mathrm{m}}, x_{i,\mathrm{d,o}} = v_{\tilde{r},i,\mathrm{o}})}{\mathrm{P}(x_{i,\mathrm{d,o}} = v_{r,i,\mathrm{o}})} = c\mathrm{P}(x_{i,\mathrm{d}} = v_{\tilde{r}}),$$

where c does not depend on $\tilde{r}$. By summing over $\tilde{r} \in R(v_{r,i,\mathrm{o}})$ and noting that $\mathrm{P}(x_{i,\mathrm{d}} = v_{\tilde{r}}) = \pi_{\tilde{r}}$, by assumption (ii), we get $c = \{\sum_{\tilde{r} \in R(v_{r,i,\mathrm{o}})} \pi_{\tilde{r}}\}^{-1}$. It follows that

$$\mathrm{P}(x_{i,\mathrm{d,m}} = v_{\tilde{r},i,\mathrm{m}}|x_{i,\mathrm{d,o}} = v_{r,i,\mathrm{o}}) = \frac{\pi_{\tilde{r}}}{\sum_{r' \in R(v_{r,i,\mathrm{o}})} \pi_{r'}}, \tag{A.6}$$

$\tilde{r} \in R(v_{r,i,\mathrm{o}})$. The conditional density $f(x_{i,\mathrm{d,m}}|x_{i,\mathrm{d,o}})$ is given by the right side of (A.6) with $v_{r,i,\mathrm{o}}$ replaced by $x_{i,\mathrm{d,o}}$ and $\tilde{r}$ being the $\tilde{r} \in R(x_{i,\mathrm{d,o}})$ such that $v_{\tilde{r},i,\mathrm{m}} = x_{i,\mathrm{d,m}}$. The sample $x_{i,\mathrm{d,m}}$ is drawn from the conditional distribution such that it has the probability given by (A.6), with $v_{r,i,\mathrm{o}}$ replaced by $x_{i,\mathrm{d,o}}$, of taking the value $v_{\tilde{r},i,\mathrm{m}}$, $\tilde{r} \in R(x_{i,\mathrm{d,o}})$.

(II) Next, note that, by assumption (iii), we have $X_{i,\mathrm{c}}|x_{i,\mathrm{d}} \sim N(\mu_r, \Omega)$, where r is such

that $x_{i,\mathrm{d}} = v_r$. Write $\Omega = (\omega_{kl})_{1\le k,l\le p}$. Then, we have

$$\left(\begin{array}{c} X_{i,\mathrm{c,m}} \\ X_{i,\mathrm{c,o}} \end{array} \right) \Bigg| x_{i,\mathrm{d}} \;\sim\; N\left[\left(\begin{array}{c} \mu_{r,i,\mathrm{m}} \\ \mu_{r,i,\mathrm{o}} \end{array} \right), \left(\begin{array}{cc} \Omega_{i,\mathrm{mm}} & \Omega_{i,\mathrm{mo}} \\ \Omega_{i,\mathrm{om}} & \Omega_{i,\mathrm{oo}} \end{array} \right) \right],$$

where $\Omega_{i,\mathrm{mm}} = (\omega_{kl})_{k,l\in s_{i,\mathrm{c,m}}}$, $\Omega_{i,\mathrm{mo}} = (\omega_{kl})_{k\in s_{i,\mathrm{c,m}},l\in s_{i,\mathrm{c,o}}}$, $\Omega_{i,\mathrm{om}} = \Omega'_{i,\mathrm{mo}}$, and $\Omega_{i,\mathrm{oo}} = (\omega_{kl})_{k,l\in s_{i,\mathrm{c,o}}}$ with $s_{i,\mathrm{c,o}} = \{1 \le k \le p : x_{i,\mathrm{c},k} \text{ observed}\}$ ($x_{i,\mathrm{c},k}$ is the kth component of $x_{i,\mathrm{c}}$) and $s_{i,\mathrm{c,m}} = \{1,\ldots,p\} \setminus s_{i,\mathrm{c,o}}$. It follows (e.g., Jiang 2007, Appendix C.1), that

$$X_{i,\mathrm{c,m}}|x_{i,\mathrm{c,o}}, x_{i,\mathrm{d}} \sim N\{\mu_{r,i,\mathrm{m}} + \Omega_{i,\mathrm{mo}}\Omega_{i,\mathrm{oo}}^{-1}(x_{i,\mathrm{c,o}} - \mu_{r,i,\mathrm{o}}), \Omega_{i,\mathrm{mm}} - \Omega_{i,\mathrm{mo}}\Omega_{i,\mathrm{oo}}^{-1}\Omega_{i,\mathrm{om}}\}. \quad (A.7)$$

Denote the mean vector and covariance matrix of the multivariate normal distribution on the right side of (A.7) by $\mu_{r,i,\mathrm{c,m}}$ and $\Omega_{i,\mathrm{c,m}}$, respectively. Then, we have

$$f(x_{i,\mathrm{c,m}}|x_{i,\mathrm{c,o}}, x_{i,\mathrm{d}}) \propto \exp\left\{ -\frac{1}{2}(x_{i,\mathrm{c,m}} - \mu_{r,i,\mathrm{c,m}})'\Omega_{i,\mathrm{c,m}}^{-1}(x_{i,\mathrm{c,m}} - \mu_{r,i,\mathrm{c,m}}) \right\},$$

where r is such that $x_{i,\mathrm{d}} = v_r$.

(III) Finally, by assumption (iv), we have $Y_{i,\mathrm{m}}|x_i \sim N(x'_{i,M}\beta_M, \sigma^2)$, hence

$$f_M(y_{i,\mathrm{m}}|x_i) \propto \exp\left\{ -\frac{(y_{i,\mathrm{m}} - x'_{i,M}\beta_M)^2}{2\sigma^2} \right\}.$$

In conclusion, we can choose (after dropping a constant term)

$$\begin{aligned} g(z) \;=\; & \left\{ \frac{\pi_{\tilde{r}}}{\sum_{r'\in R(x_{i,\mathrm{d,o}})}\pi_{r'}} \right\} \exp\left\{ -\frac{1}{2}(x_{i,\mathrm{c,m}} - \mu_{r,i,\mathrm{c,m}})'\Omega_{i,\mathrm{c,m}}^{-1}(x_{i,\mathrm{c,m}} - \mu_{r,i,\mathrm{c,m}}) \right. \\ & \left. -\frac{(y_{i,\mathrm{m}} - x'_{i,M}\beta_M)^2}{2\sigma^2} \right\}, \end{aligned}$$

where $\tilde{r}$ is such that $\tilde{r} \in R(x_{i,\mathrm{d,o}})$ and $v_{\tilde{r},i,\mathrm{m}} = x_{i,\mathrm{d,m}}$, and r is such that $x_{i,\mathrm{d}} = v_r$. The sampling from g consists of three steps: (I) draw $x_{i,\mathrm{d,m}}$ from the distribution that has the probability equal to $\pi_{\tilde{r}}\{\sum_{r'\in R(x_{i,\mathrm{d,o}})}\pi_{r'}\}^{-1}$ of taking the value $v_{\tilde{r},i,\mathrm{m}}$ for $\tilde{r} \in R(x_{i,\mathrm{d,o}})$; (II) given the $x_{i,\mathrm{d,m}}$ drawn, draw $x_{i,\mathrm{c,m}}$ from the multivariate normal distribution in (A.7), where r is such that $x_{i,\mathrm{d}} = v_r$; (III) given the $x_{i,\mathrm{d,m}}, x_{i,\mathrm{c,m}}$ drawn, draw $y_{i,\mathrm{m}}$ from $N(x'_{i,M}\beta_M, \sigma^2)$.

If any of the components $x_{i,\mathrm{d,m}}$, $x_{i,\mathrm{c,m}}$, or $y_{i,\mathrm{m}}$ are empty, we simply skip the corresponding step(s) (I, II, or III).

A.5 Some derivations in Section 6

A.5.1 Derivation of (16)

First note that

$$\begin{aligned}\{Y_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(M_{\mathrm{ind},i}=0)} &= (Y_i - c_i)^2 1_{(M_{\mathrm{ind},i}=0)} \\ &\quad +2(Y_i - c_i)\{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}1_{(M_{\mathrm{ind},i}=0)} \\ &\quad +\{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 \\ &= \xi_i + 2\eta_i + \zeta_i,\end{aligned}$$

where c_i is defined in JNR [above (16)]. We have

$$\begin{aligned}\mathrm{E}(\eta_i) &= \{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}[\mathrm{E}\{Y_i 1_{(M_{\mathrm{ind},i}=0)}\} - c_i \mathrm{P}(M_{\mathrm{ind},i} = 0)] \\ &= \{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}[\mathrm{E}\{Y_i h(Y_i)\} - c_i \mathrm{E}\{h(Y_i)\}] \\ &= 0.\end{aligned}$$

Thus, by the law of large numbers (LLN), we have $\sum_{i=1}^n \eta_i = o_{\mathrm{P}}(n)$ (e.g., Jiang 2010, ch. 3). It follows that

$$\begin{aligned}&\sum_{i=1}^n \{Y_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(M_{\mathrm{ind},i}=0)} \\ =\ &\sum_{i=1}^n (Y_i - c_i)^2 1_{(M_{\mathrm{ind},i}=0)} + \sum_{i=1}^n \{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(M_{\mathrm{ind},i}=0)} + \delta \\ =\ &\sum_{i=1}^n \{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(M_{,\mathrm{ind},i}=0)} + \delta_1,\end{aligned}$$

where r is a lower-order term, and r_1 is a sum of a term that is not model-dependent and a lower-order term. Similarly, it can be shown that

$$\sum_{i=1}^n \mathrm{E_c}\{Y_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(M_{\mathrm{ind},i}=1)} = \sum_{i=1}^n \{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 1_{(M_{\mathrm{ind},i}=1)} + \delta_2,$$

where and δ_2 is a sum of a term that is not model-dependent and a lower-order term. Thus,

by (15) of JNR, we have

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

where δ is a sum of terms that are not model-dependent and lower-order terms. Thus, by taking the expectation, we obtain (16) of JNR. It should be noted that, for the above argument to hold, some regularity conditions are needed that ensure, for example, the expectation of a lower-order term is a lower-order term.

A.5.2 Derivation of (17)

Because (16) holds for every M, by letting $M = M_{\mathrm{opt}}$, we have

$$\mathrm{E}\{\mathrm{E_c}(Q_{M_{\mathrm{opt}}}|Y_{\mathrm{obs}})\} = \sum_{i=1}^{n}\{c_i - \mathrm{E}(Y_i)\}^2\mathrm{E}\{h(Y_i)\} + \delta_{\mathrm{opt}}, \qquad (A.8)$$

where δ_{opt} consists of terms that are not model-dependent and lower-order terms. Note that, when $M = M_{\mathrm{opt}}$, E_{M,θ_M} becomes E, and the second term on the right side of (16) of JNR disappears. By taking the difference between (16) of JNR and (A.8), we get

$$\begin{aligned}&\mathrm{E}\{\mathrm{E_c}(Q_M|Y_{\mathrm{obs}})\} - \mathrm{E}\{\mathrm{E_c}(Q_{M_{\mathrm{opt}}}|Y_{\mathrm{obs}})\} \\ = \;& \sum_{i=1}^{n}\{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2\mathrm{E}\{h(Y_i)\} + \sum_{i=1}^{n}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}^2\mathrm{E}\{g(Y_i)\} \\ & - \sum_{i=1}^{n}\{c_i - \mathrm{E}(Y_i)\}^2\mathrm{E}\{h(Y_i)\} + \delta,\end{aligned}$$

where δ has the same meaning as in (16) of JNR. Furthermore, note that

$$\sum_{i=1}^{n}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}^2\mathrm{E}\{g(Y_i)\} = c - \sum_{i=1}^{n}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}^2\mathrm{E}\{h(Y_i)\},$$

where c does not depend on the MDM. Also note that $\{c_i - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 - \{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}^2 - \{c_i - \mathrm{E}(Y_i)\}^2 = 2\{c_i - \mathrm{E}(Y_i)\}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}$, and

$$\begin{aligned}&\{c_i - \mathrm{E}(Y_i)\}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}\mathrm{E}\{h(Y_i)\} \\ =\;& [\mathrm{E}\{Y_i h(Y_i)\} - \mathrm{E}(Y_i)\mathrm{E}\{h(Y_i)\}]\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\} \\ =\;& \mathrm{cov}\{Y_i, h(Y_i)\}\{\mathrm{E}(Y_i) - \mathrm{E}_{M,\theta_M}(Y_i)\}\end{aligned}$$

(recall the definition of c_i). The second equation in (17) thus follows.

The derivation in Subsection A.5.1, and the LLN, also show that $\mathrm{E}\{\mathrm{E}_{\mathrm{c}}(Q_M|Y_{\mathrm{obs}})\} - \mathrm{E}\{\mathrm{E}_{\mathrm{c}}(Q_{M_{\mathrm{opt}}}|Y_{\mathrm{obs}})\}$ is the leading term for the difference in (15) between M and M_{opt}. The first equation in (17) thus follows.

A.6 Additional simulation results

A.6.1 Linear regression: Comparison of different strategies

We carry out a simulation study under the following linear regression model (see Example 3). Suppose that the candidate covariates consist of two continuous variables and two indicator variables. So, x_1, x_2 are continuous (with $p = 2$) and x_3, x_4 are indictors (0 or 1). Thus, $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$ with $x_{i,\mathrm{c}} = (x_{i1}, x_{i2})'$ and $x_{i,\mathrm{d}} = (x_{i3}, x_{i4})'$. The distinct possible values for $x_{i,\mathrm{d}}$ are $v_1 = (0,0)'$, $v_2 = (0,1)'$, $v_3 = (1,0)'$, and $v_4 = (1,1)'$. Assumption (iii) of Example 3 means that there are 2×1 vectors $\mu_1, \mu_2, \mu_3, \mu_4$ and 2×2 covariance matrix Ω such that, given $x_{i,\mathrm{d}} = v_r$, $x_{i,\mathrm{c}} \sim N(\mu_r, \Omega)$, $r = 1, 2, 3, 4$.

The simulations are run with the sample size $n = 100$ and the true model being x_1 and x_3. The true parameters are $\beta_1 = \beta_3 = \sigma^2 = 1$. The true μ_r is the same as v_r, $r = 1, 2, 3, 4$; and the true Ω is the 2×2 identity matrix. After the complete data is generated, we randomly select a subset of indexes from $\{1, \ldots, n\}$ for the response as well as for each of the candidate predictors, which correspond to the missing data, so that $100p_{\mathrm{m}}$ % of the data are missing for the response and each of the candidate predictors. Here we consider two cases: $p_{\mathrm{m}} = 0.1$ and $p_{\mathrm{m}} = 0.2$.

We study the performance of E-MS with the invisible fence (IF; Jiang *et al.* 2011b; also see Jiang 2014), which, in this case, is equivalent to the AF, but computationally more efficient. On the other hand, it is known that the latter may suffer from the "dominant factor". Namely, although the true coefficients for the true predictors (x_1 and x_3) are both equal to 1, it turns out that the continuous predictor is the dominant factor. Thus, with a moderate sample size, such as in the current case, the IF tends to overwhelmingly select x_1 at dimension 1, leading to a potential underfitting. To overcome such a problem, we consider the

following modification of the IF to make it more "aggressive". Let α be a chosen number between 0 and 1. Let d^* be the selected dimension by IF and p^* be the corresponding maximum empirical probability (that a given model is selected at the dimension). Let p_1^* be the largest maximum empirical probability corresponding to a dimension greater than d^* (thus, $p_1^* < p^*$). Let d_1^* be the corresponding dimension to p_1^*. If $p_1^* \geq (1-\alpha)p^*$, then d_1^* is selected instead of d^*. It is clear that the modified IF is more in favor of a "larger" model, and in this sense it is more aggressive.

We run a same-data comparison of the E-MS with a number of different procedures. The first is to combine the IF with the E-M (not E-MS) algorithm. Namely, we first run the E-M algorithm to obtain the parameter estimates under the full model. We then generate (parametric) bootstrap samples under the full model, as in the IF. The best part of this procedure is that, when one generates the bootstrap samples, one generates complete data rather than data with missing values. We then apply the modified IF, as described above, to the bootstrapped data. We call such a procedure EMIF. In our simulation study, we consider three different values of α: $\alpha = 0$ (corresponding to IF without the modification), $\alpha = 0.1$, and $\alpha = 0.5$, for EMIF as well as each of the comparing procedures described below, except the E-MS.

Of course, this raises a question on what α is the best, which one may not know in practice. On the other hand, the E-MS seems to have some advantage in this regard. The idea is to start with a relatively large α (say, $\alpha = 0.5$), in order to be more conservative in dropping the predictors, and gradually reduces α as the iteration progresses. More specifically, we begin with $\alpha = 0.5$; with each iteration, we reduce α by half, until convergence.

In addition to EMIF and E-MS, two other methods are also included in our comparison. One is IF based on the complete-record-only analysis (CRNIF); the other is IF with the missing data replaced by the imputed data (IMIF). The latter is based on a method of multivariate imputation developed by van Buuren *et al.* (2005), implemented in the R package, aregImpute(). As a comparison, we have also considered IF based on the complete

data that were generated before the missing values were taken out (CDIF). The latter is, of course, not possible in a practical situation, but it would be interesting to see how much loss of efficiency there is for a method, if any, compared to the "gold standard". Note that all the comparing methods are in conjunction with the IF, the only difference being how the missing data are handled. The results based on 100 simulation runs are presented in Tables A.1 & A.2, where OF standards for overfitting, that is, the empirical probability that the selected model includes all the true predictors plus at least one extraneous predictor; UF standards for underfitting, that is, the empirical probability that the selected model misses at least one true predictor [but may include extraneous predictor(s)]; other performance measures are the same as before (corresponding s.d.'s to the right, for MC and MIC).

Table A.1: Summary of Performance ($p_{\mathrm{m}} = 0.1$)

Method	α	TP	OF	UF	MC	s.d.	MIC	s.d.
CRNIF	0	0.67	0.00	0.33	1.67	0.47	0.00	0.00
	0.1	0.85	0.00	0.15	1.85	0.36	0.00	0.00
	0.5	0.73	0.27	0.00	2.00	0.00	0.27	0.45
IMIF	0	0.61	0.00	0.39	1.61	0.49	0.00	0.00
	0.1	0.88	0.00	0.12	1.88	0.33	0.00	0.00
	0.5	0.72	0.28	0.00	2.00	0.00	0.28	0.45
EMIF	0	0.66	0.00	0.34	1.66	0.48	0.00	0.00
	0.1	0.88	0.00	0.12	1.88	0.33	0.00	0.00
	0.5	0.75	0.25	0.00	2.00	0.00	0.25	0.44
CDIF	0	0.71	0.00	0.29	1.71	0.46	0.00	0.00
	0.1	0.92	0.00	0.08	1.92	0.27	0.00	0.00
	0.5	0.69	0.31	0.00	2.00	0.00	0.31	0.46
E-MS		0.98	0.01	0.01	1.99	0.10	0.01	0.10

Table A.2: Summary of Performance ($p_{\mathrm{m}} = 0.2$)

Method	α	TP	OF	UF	MC	s.d.	MIC	s.d.
CRNIF	0	0.56	0.00	0.44	1.56	0.50	0.01	0.10
	0.1	0.73	0.07	0.20	1.80	0.40	0.09	0.29
	0.5	0.68	0.31	0.01	1.99	0.10	0.32	0.47
IMIF	0	0.59	0.00	0.41	1.59	0.49	0.01	0.10
	0.1	0.85	0.00	0.15	1.85	0.36	0.01	0.10
	0.5	0.81	0.19	0.00	2.00	0.00	0.19	0.39
EMIF	0	0.65	0.00	0.35	1.65	0.48	0.00	0.00
	0.1	0.88	0.00	0.12	1.88	0.33	0.00	0.00
	0.5	0.82	0.17	0.01	1.99	0.10	0.18	0.39
CDIF	0	0.75	0.00	0.25	1.75	0.44	0.00	0.00
	0.1	0.94	0.00	0.06	1.94	0.24	0.00	0.00
	0.5	0.74	0.26	0.00	2.00	0.00	0.26	0.44
E-MS		0.95	0.01	0.04	1.96	0.20	0.03	0.17

It is clear in this comparison that, overall, E-MS outperforms not only all of the methods that are practically feasible (CRNIF, IMIF, EMIF), but also the "gold standard" (CDIF) that is practically infeasible. In fact, in terms of the overall performance, the order seems to be (from best to worst) E-MS, CDIF, EMIF, IMIF, CRNIF. Of course, it is not surprising that CRNIF takes the last place, but what seems a little unexpected is that E-MS even (slightly) outperforms CDIF. An explanation for this is that the performance of CDIF still suffers, to some extend, the dominant factor effect (Jiang *et al.* 2011b), but E-MS is able to overcome this (see below). A key to this "super-performance" is α, which may be viewed as a tuning parameter. It appears that the best α for CRNIF, IMIF, EMIF, and CDIF is somewhere between 0.1 and 0.5. Of course, in the simulation study we could explore this best value, but it would not be possible in practice. On the other hand, the E-MS seems to be able to get the best out of the 'α-business" during its iterations. By the way, in all of the simulation runs,

the E-MS converged in 2-3 iterations. Also, it seems that the performance of IMIF, EMIF, and E-MS are not affected much by the increase of $p_{\rm m}$. This is a bit surprising, as larger $p_{\rm m}$ means less observed data. In fact, with $p_{\rm m} = 0.1$, one expects about 59% complete data records; with $p_{\rm m} = 0.2$, the % of the complete data records drops to less than 33%. On the other hand, it takes about twice the computing time to run the E-MS for $p_{\rm m} = 0.2$, compared to $p_{\rm m} = 0.1$. This is reasonable, as more data are missing under $p_{\rm m} = 0.2$; therefore, the conditional expectations, which have to be dealt with via the Monte-Carle method (see Example 3), need to be evaluated more often than under $p_{\rm m} = 0.1$.

A.6.2 Robustness of E-MS

We investigate performance of the E-MS in terms of robustness under different aspects of model misspecification. Both situations are in association with the example of backcross experiment in JNR (see Example 2 and Section 5).

1. A situation where the true model is not among the candidates. 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. More specifically, the authors considered the case where the true QTLs are located in the middle of their flanking markers; thus, 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. Here we consider a setting similar to those of Nguyen *et al.* (2013). There are 6 true QTLs with identical signals, β (see below). The true θ is 0.2, which corresponds to a heritability of approximately 25%. The true QTLs are located in the middle of two flanking markers, as follows. $1^{\rm st}$ chromosome: markers 1 and 2; 3 and 4; 5 and 6. $2^{\rm nd}$ chromosome: markers 1 and 2; 3 and 4. $3^{\rm rd}$ chromosome: markers 1 and 2. Following Nguyen *et al.* (2013) (also see Broman & Speed 2002), the QTL is considered correctly identified if one

of the two flanking markers of the true QTL is identified; in case both of the flanking markers are identified, one is counted as correctly identified, the other incorrectly identified, to avoid double counts. Once again, we compare the performance of the E-MS with BIC with the CDBIC. The results based on 100 simulations are reported in Table A.3 under the same setting as Table 2 of JNR. Overall, there is no obvious trend, in either way of the performance, compared to Table 2 of JNR. For example, the relative efficiency (% Ratio) of the E-MS with respect to CDBIC is very comparable to that reported in Table 2 of JNR. This suggests that E-MS is still capable in detecting the best approximating model in case that the true model is not among the candidates.

Table A.3: Backcross Experiment; QTLs at Middle of Flanking Markers

n	β	σ	Method	TP	MC (s.d.)	MIC (s.d.)	% Ratio
250	1	1	E-MS	0.36	5.62 (0.56)	1.05 (1.04)	86%
			CDBIC	0.45	5.65 (0.54)	0.66 (0.71)	
100	1	1	E-MS	0.22	4.48 (0.85)	1.46 (1.25)	63%
			CDBIC	0.35	4.64 (0.80)	1.11 (1.10)	
250	0.5	1	E-MS	0.41	4.43 (0.81)	1.04 (1.10)	93%
			CDBIC	0.44	4.57 (0.83)	0.74 (0.81)	
250	1	0.1	E-MS	0.41	5.99 (0.10)	0.93 (0.97)	67%
			CDBIC	0.61	5.99 (0.10)	0.51 (0.73)	
500	1	1	E-MS	0.52	5.96 (0.20)	0.70 (0.88)	90%
			CDBIC	0.58	5.98 (0.14)	0.50 (0.69)	

2. A situation of heavy-tailed error distribution. In this study we consider model misspecification in terms of the distribution of the regression errors in the backcross experiment. Namely, the errors ϵ_i are assumed to be normally distributed, but the assumption fails and the errors, instead, have a t-distribution with 6 degrees of freedom (so the errors have finite fifth moment, but no higher moments). However, pretend that one does not know about the t-distribution, and proceeds with the E-MS under the normality assumption. The E-MS is carried out the same way as in Example 2 and Section 5 of JNR and, in particular,

with the FW/BW BIC (see Section A.3). Here we consider the case that the true QTLs are on the markers, as in Section 5 of JNR. The results based on 100 simulations are reported in Table A.4. Again, the setting is the same as Table 2 of JNR. Compared to the latter, the only significant drop in the E-MS performance seems to be the case with weaker signal, that is, $\beta = 0.5$ ($n = 250, \sigma = 1$). On the other hand, the results seem to suggest that the performance of E-MS is robust to the heavy-tailed error as long as the sample size is relatively large, or the signal is relatively strong (compared to the noise).

Table A.4: Backcross Experiment; Heavy-tailed Error Distribution

n	β	σ	Method	TP	MC (s.d.)	MIC (s.d.)	% Ratio
250	1	1	E-MS	0.54	5.98 (0.14)	0.75 (1.02)	73%
			CDBIC	0.74	5.99 (0.10)	0.36 (0.73)	
100	1	1	E-MS	0.17	5.36 (0.76)	1.79 (1.70)	52%
			CDBIC	0.33	5.59 (0.60)	0.98 (1.18)	
250	0.5	1	E-MS	0.03	4.61 (0.75)	0.87 (0.92)	25%
			CDBIC	0.12	4.84 (0.72)	0.43 (0.77)	
250	1	0.1	E-MS	0.63	6.00 (0.00)	0.61 (0.91)	85%
			CDBIC	0.74	6.00 (0.00)	0.36 (0.73)	
500	1	1	E-MS	0.64	6.00 (0.00)	0.46 (0.69)	83%
			CDBIC	0.77	6.00 (0.00)	0.32 (0.68)	

A.6.3 Missing covariates under various MDM

In this subsection we report results of the last simulation study of Subsection 6.1, discussed near the end of the subsection. See Tables A.5 and A.6. 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.

Table A.5: Backcross Experiment with MDM, Scenario MA

n	β	σ	Method	TP	MC (s.d.)	MIC (s.d.)	% Ratio
250	1	1	E-MS	0.52	5.99 (0.10)	0.84 (1.08)	84%
			CDBIC	0.62	6.00 (0.00)	0.47 (0.66)	
100	1	1	E-MS	0.09	5.31 (0.65)	1.64 (1.59)	39%
			CDBIC	0.23	5.49 (0.64)	1.22 (1.37)	
250	0.5	1	E-MS	0.05	4.52 (0.81)	1.07 (1.07)	42%
			CDBIC	0.12	4.73 (0.80)	0.64 (0.78)	
250	1	0.1	E-MS	0.78	6.00 (0.00)	0.23 (0.45)	126%
			CDBIC	0.62	6.00 (0.00)	0.47 (0.66)	
500	1	1	E-MS	0.51	6.00 (0.00)	0.68 (0.79)	76%
			CDBIC	0.67	6.00 (0.00)	0.46 (0.72)	

Table A.6: Backcross Experiment with MDM, Scenario MB

n	β	σ	Method	TP	MC (s.d.)	MIC (s.d.)	% Ratio
250	1	1	E-MS	0.49	5.99 (0.10)	0.78 (0.89)	79%
			CDBIC	0.62	6.00 (0.00)	0.47 (0.66)	
100	1	1	E-MS	0.14	5.27 (0.78)	1.84 (1.78)	61%
			CDBIC	0.23	5.49 (0.64)	1.22 (1.37)	
250	0.5	1	E-MS	0.09	4.55 (0.82)	0.93 (0.92)	75%
			CDBIC	0.12	4.73 (0.80)	0.64 (0.78)	
250	1	0.1	E-MS	0.79	6.00 (0.00)	0.26 (0.54)	127%
			CDBIC	0.62	6.00 (0.00)	0.47 (0.66)	
500	1	1	E-MS	0.54	6.00 (0.00)	0.67 (0.72)	81%
			CDBIC	0.67	6.00 (0.00)	0.46 (0.72)	

A.6.4 More comparison with IMBIC

In this subsection, we present some additional simulation results regarding the comparison of E-MS with imputation-based methods. Unlike in Subsection A.6.1, which involved

the IF method, the comparison here will focus on the BIC method. As in Subsection A.6.1, we use the R package aregImpute() (van Buuren *et al.* 2005) for the imputation based method, IMBIC. A brief description is as follows. By default, aregImpute uses predictive mean matching (which does not work well when fewer than 3 variables are used to predict the target variable, if the "closest" match is chosen). With the "regression" option, aregImpute will use linear extrapolation to obtain a (hopefully) reasonable distribution of imputed values. Both linear and non-linear imputations are considered in our simulation. Our results are based on 100 (multiple) imputations, from which the model with the highest probability (or frequency) is chosen.

The results under the non-linear imputation (four knots, $nk = 4$; "closest " match) are presented in Table 3 of JNR. The results under the linear imputation (no knots) are presented in Table A.7 here. The two sets of results are quite similiar. Overall, the IMBIC results are not comparable to the E-MS results, especially in terms of the % Ratio.

A.6.5 Performance of E-MS in terms of parameter estimation

In this subsection, we consider performance of E-MS in terms of parameter estimation under the backcross experiment. Note that, because some of the parameters, such as the regression coefficients, depend on the selected model, it is not very clear how to compare these parameters under different models. Therefore, in this study, we have focused on parameters that are common under all of the candidate models, namely, the recombination fraction, θ, and the error variance, σ^2. We evaluate performance of the estimators of θ and σ^2 based on the (final) selected model by the E-MS in terms of bias, variance, and mean squared error (MSE), and compare the results with those based on the selected model by CDBIC. The results, based on 100 simulation runs, are reported in Tables A.8 and A.9. Note that, because the estimators of θ only depend on the x data, the results do not change, within the same sample size, as β and σ change (which only affect the y data). The true θ is 0.2. Overall, the results show that, as the conditions improve, that is, either n increases, or β increases, or σ decreases, the performance of E-MS and CDBIC are getting closer in

estimating both parameters.

Table A.7: More Comparison with IMBIC in 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	0.35	5.79 (0.50)	0.92 (0.94)	56%
			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	0.08	4.68 (0.85)	1.24 (1.11)	35%
			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	0.02	4.10 (0.81)	1.00 (0.95)	17%
			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	0.21	5.90 (0.70)	1.03 (0.76)	34%
			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	0.38	5.99 (0.10)	0.79 (0.80)	57%
			CDBIC	0.67	6.00 (0.00)	0.46 (0.72)	

Table A.8: Estimation of θ in Backcross Experiment

n	Method	Bias^2 (10^{-8})	Variance (10^{-5})	MSE (10^{-5})
100	E-MS	162	6.89	7.05
	CDBIC	90.6	6.33	6.42
250	E-MS	14.0	3.24	3.25
	CDBIC	8.48	2.81	2.81
500	E-MS	2.92	1.48	1.48
	CDBIC	2.07	1.28	1.28

Table A.9: Estimation of σ^2 in Backcross Experiment

n	β	σ	Method	Bias^2	Variance	MSE
100	1	1	E-MS	2.0×10^{-2}	3.15×10^{-2}	5.15×10^{-2}
			CDBIC	3.15×10^{-3}	2.07×10^{-2}	2.38×10^{-2}
250	0.5	1.0	E-MS	9.40×10^{-4}	9.54×10^{-3}	10.50×10^{-3}
			CDBIC	8.23×10^{-6}	8.92×10^{-3}	8.93×10^{-3}
250	1.0	1.0	E-MS	11.2×10^{-4}	9.86×10^{-3}	10.98×10^{-3}
			CDBIC	2.72×10^{-5}	8.62×10^{-3}	8.64×10^{-3}
250	1.0	0.1	E-MS	8.17×10^{-3}	8.85×10^{-7}	8.16×10^{-3}
			CDBIC	8.11×10^{-3}	8.62×10^{-7}	8.11×10^{-3}
500	1.0	1.0	E-MS	8.22×10^{-4}	3.90×10^{-3}	4.73×10^{-3}
			CDBIC	1.08×10^{-4}	3.58×10^{-3}	3.69×10^{-3}

A.7 Analysis of protein data

Table A.10: E-MS Results for Grain Protein

Chromosome	Marker ID#				Chromosome	Marker ID#			
1	12	13			5	280	285	332	333
2	65	66			6	379	380		
3	184	186	199	200	7	467	470		
4	176								

Additional References:

Efron, B. and Tibshirani, R. (2007), On testing the significance of sets of genes, *Ann. Appl. Statist.* 1, 107-129.

Jiang, J. (2000), A nonlinear Gauss-Seidel algorithm for inference about GLMM, *Computational Statist.* 15, 220-241.

Jiang, J. (2010), *Large Sample Techniques for Statistics*, Springer, New York.

Jiang, J. and Rao, J. S. (2003), Consistent procedures for mixed linear model selection, *Sankhya* 65 A, 23-42.

Luenberger, D. G. (1984), *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA.

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.

Nguyen, T. and Jiang, J. (2012), Restricted fence method for covariate selection in longitudinal data analysis, *Biostatistics* 13, 303-314.

Wu, C. F. J. (1983), On the convergence properties of the EM algorithm, *Ann. Statist.* 11, 95-103.