NON-RESPONSE MODELS FOR THE ANALYSIS OF NON-MONOTONE IGNORABLE MISSING DATA

JAMES M. ROBINS

Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

AND

RICHARD D. GILL

Department of Mathematics, Utrecht University, Utrecht, The Netherlands

SUMMARY

We discuss a new class of ignorable non-monotone missing data models – the randomized monotone missingness (RMM) models. We argue that the RMM models represent the most general plausible physical mechanism for generating non-monotone ignorable data. We show that there exists ignorable missing data processes that are not RMM. We argue that it may therefore be inappropriate to analyse non-monotone missing data under the assumption that the missingness mechanism is ignorable, if a statistical test has rejected the hypothesis that the missing data process is RMM representable. We use RMM models to analyse data from a case-control study of the effects of radiation on breast cancer.

1. INTRODUCTION

The purpose of this paper is to introduce a new class of ignorable non-monotone missing data processes, the randomized monotone missingness (RMM) processes, and to provide new methods, based on these processes, for analysing studies with non-monotone missing data. The new methods are used to analyse data provided by the conference organizers from a case-control study of the effect of radiation on breast cancer. This class was first introduced by Robins *et al.*,[1] and then extensively investigated by Gill and Robins,[2] Following Gill and Robins,[2] we (i) argue that the RMM processes represent the most general plausible mechanism for generating a non-monotone ignorable process, and (ii) prove there exist ignorable processes that are not RMM. As a consequence, we argue, in Section 8, that it may be inappropriate to analyse non-monotone missing data under the assumption that the missing data mechanism is ignorable if a statistical test has rejected the hypothesis that the missing process is RMM representable.

2. MOTIVATION

Let $L = (L_1, \ldots, L_K)'$ be a random K-vector representing the complete data. Let R_k be the indicator of whether the kth variable L_k was observed. Let $R = (R_1, \ldots, R_K)'$ be the vector of missing data indicators, $L_{(R)}$ be the observed components of L, and $L_{(\bar{R})}$ be the unobserved components. For example, if $K = 4$, $R = (1, 0, 0, 1)$, $L_{(R)} = (L_1, L_4)'$, $L_{(\bar{R})} = (L_2, L_3)'$. Rubin[3] refers to $L_{(R)}$

as L_{obs} and $L_{(\overline{R})}$ as L_{mis}. For each study subject, we observe an independently and identically distributed realization of $(R, L_{(R)})$ with joint density $f(r, \ell_{(r)})$, where we have used lower-case letters to represent realizations of random variables. Further, we let $\mathbf{1}$ denote a vector of 1's, so, $f(\mathbf{1}, \ell)$ is the probability of observing the complete data $L = \ell$. We say the data are missing at random (MAR) if the probability $f(r \mid L)$ of observing $R = r$ given L depends only on the observed data $L_{(r)}$, which can be written as

$$f(r \mid L) = \pi(r, L_{(r)}) \tag{1}$$

for some function $\pi(r, \ell_{(r)})$. Studies with the missing data are commonly analysed under the assumption that the missing process is MAR. This reflects the convenient statistical fact that, under MAR, likelihood-based inference for the parameters θ of a fully parametric model for the law of the complete data L, that is $f(L; \theta)$, 'ignores' the missingness process, provided, as we assume, the parameters γ of the missingness process are distinct from the parameters θ for the model for L. When the parameters are distinct, MAR missing data processes are referred to as ignorable.[3] Formally, the likelihood $\mathscr{L}(\theta, \gamma)$ factors into a θ-part $\mathscr{L}_1(\theta)$ and a γ-part $\mathscr{L}_2(\gamma)$. That is,

$$\mathscr{L}(\theta, \gamma) = \mathscr{L}_1(\theta)\,\mathscr{L}_2(\gamma) \tag{2}$$

where $\mathscr{L}_1(\theta) = \prod_{i=1}^n \int f(L_i; \theta)\,\mathrm{d}L_{i(\overline{R}_i)}$, $\mathscr{L}_2(\gamma) = \prod_{i=1}^n \pi(R_i, L_{i(R_i)}; \gamma)$, and $\pi(r, \ell_{(r)}; \gamma)$ is a model for the MAR missingness process. Hence, it is unnecessary to specify a model $\pi(r, L_{(r)}; \gamma)$ for the missing data process in order to compute the maximum likelihood estimator (MLE) of θ and the observed information matrix for θ. Nonetheless, no matter how convenient the MAR factorization (2) may be for carrying out likelihood-based inference on θ, the MAR assumption should only be entertained if (i) one believes it reasonable to assume that the missingness process is MAR, and (ii) there is no data evidence contradicting the MAR assumption.

Now, if L is a discrete random vector with support (that is, taking values) in a finite set $\mathbf{L}$ and, if the model $f(L; \theta)$ is saturated (that is, non-parametric), Gill et al.[15] prove that there can <u>never</u> be any data evidence contradicting the MAR assumption (1). Formally, they prove theorem 2.1.

Theorem 2.1. (i) If L has a finite support $\mathbf{L}$, then there exists a MAR missing data process $f^\Delta(r \mid \ell) \equiv \pi^\Delta(r, \ell_{(r)})$ and a complete data law $f^\Delta(\ell)$ such that the true distribution $f(r, \ell_{(r)})$ of the observed data is the marginal distribution of $(R, L_{(R)})$ corresponding to the joint law $f^\Delta(r, \ell) = f^\Delta(r \mid \ell) f^\Delta(\ell)$. (ii) Further, if the probability of observing the complete data vector ℓ is non-zero for all $\ell \in \mathbf{L}$,

$$f(\mathbf{1}, \ell) \neq 0 \quad \text{for all} \quad \ell \in \mathbf{L} \tag{3}$$

then $f^\Delta(r \mid \ell)$ and $f^\Delta(\ell)$ are unique. Specifically, $f^\Delta(\ell)$ is the unique law $f^*(\ell)$ maximizing the expected log likelihood $\sum_{\text{all } (r, \ell_{(r)})} f(r, \ell_{(r)}) \log f^*(\ell_{(r)})$ where $f^*(\ell_{(r)})$ is the marginal law of $L_{(r)}$ under $f^*(\ell)$; and $f^\Delta(r \mid \ell)$ is equal to $f(r, \ell_{(r)}) / f^*(\ell_{(r)})$.

When L has continuous components (that is, the support of L is not finite), Gill et al.[15] conjecture that Theorem 2.1 remains true. Recall that $f(r, \ell_{(r)})$ is the marginal distribution of $(R, L_{(R)})$ corresponding to the true but unknown missing data process $f(r \mid \ell)$ and the true but unknown full data law $f(\ell)$. Theorem 2.1 says that even if the true missing process $f(r \mid \ell)$ is not an MAR process, we can never discover that fact from the observed data $(R, L_{(R)})$ if we place no restrictions on $f(L)$ since (i) the observed data are generated from $f(r, \ell_{(r)})$ and (ii) that law

is always perfectly consistent with an MAR process $f^\Delta(r \mid \ell)$ and a marginal law $f^\Delta(\ell)$. Now if $f(r \mid \ell)$ is not an MAR process, then (i) the true law $f(\ell)$ will not equal $f^\Delta(\ell)$ and (ii) an estimate of $f(\ell)$ computed under the incorrect MAR assumption will converge to $f^\Delta(\ell)$ and thus be inconsistent for $f(\ell)$. Hence, since there can be no data evidence against the MAR assumption (2), it becomes doubly important to understand when that assumption is reasonable. A first step in that direction is to understand how we could generate an MAR process algorithmically, say on a computer. Such understanding will help us to determine whether we believe it is reasonable that in an actual study, nature could have generated MAR data.

In the setting of continuous L, there are other, more pragmatic, reasons for wishing to understand how non-monotone MAR processes can be generated and thus can be modelled. As an example, suppose $L = (X', Y)'$ where Y is an outcome variable of interest and $X = (X_0, X_1, \ldots, X_M)'$ is a vector of continuous and discrete explanatory variables with $X_0 \equiv 1$, and we wish to estimate the parameters of the model

$$\text{pr}\left[Y = 1 \mid X\right] = expit\left(\alpha' X\right) \tag{4}$$

where $\text{expit}(z) = e^z / \{1 + e^z\}$.

When missingness is MAR, the EM algorithm is generally used to estimate the parameters of (4). To implement the EM algorithm, it is not necessary to specify a model for the missing data process; however, it is necessary either (i) to specify a parametric model for the marginal distribution of X, or (ii) to compute the expectations given the observed components of X required in the E-step by non-parametric smoothing, which is unsuitable when X is multivariate and continuous. When scientific interest resides in the conditional distribution of Y given X and not in the marginal distribution of X, one may not wish to specify a parametric model for the marginal of X.

In this paper, we consider an alternative approach in which we leave the marginal distribution of X completely unrestricted and instead (i) specify a parametric model $\pi\left(r, \ell_{(r)}; \gamma\right)$ for the missing-at-random process $f(r \mid \ell)$, (ii) obtain estimates $\hat{\gamma}$ of the parameters γ, and (iii) finally estimate the parameters α of (4) by the solution $\hat{\alpha}$ to the inverse probability weighted estimating equation

$$0 = \sum_{i=1}^{n} U_i(\alpha, \hat{\gamma}), \qquad U(\alpha, \gamma) \equiv \{\delta/\pi(\mathbf{1}, L; \gamma)\}\left\{Y - \text{expit}\left(\alpha' X\right)\right\} X \tag{5}$$

where $\delta = I(R = \mathbf{1})$ takes the value 1 if we observe complete data on a subject and is zero otherwise. $\sum_i U_i(\alpha, \gamma)$ is an unbiased estimating function for α since, like the Horvitz-Thompson estimator,[4] it weights each subject with complete data by the inverse of the conditional probability of having complete data. $\hat{\alpha}$ can be obtained by fitting model (4) to the complete cases ($\delta = 1$) using a canned logistic regression program that allows for individual weights $\omega = \pi(\mathbf{1}, L; \hat{\gamma})^{-1}$.

Since neither the marginal distribution of X nor the missing data process $f(r \mid L)$ is of scientific interest, one might argue it is preferable to model the marginal distribution of X than to model $\pi\left(r; \ell_{(r)}\right)$ when the data are MAR, since more efficient estimators of the parameter of scientific interest α can be obtained by modelling the marginal of X. However, we take a different view. Specifically, with missing data, it is rarely appropriate to analyse the data solely under the assumption that the missing process is MAR. One must also explore, at least as a sensitivity analysis, the possibility that the missingness mechanism is non-ignorable. When X is multivariate with continuous components, such a sensitivity analysis generally requires the specification of parametric models for the possibly non-ignorable missing data process $f(r \mid L)$. If we analyse the data using inverse probability weighted estimators, we need not model the marginal distribution of X.[16] Thus, an approach based on specifying a parametric model for $f(r \mid L)$ and constructing

inverse probability weighted estimates of α allows a unified approach to investigating both ignorable and non-ignorable missing data processes, without having to model the marginal of X.

In order to specify a parametric model for an MAR process, we must understand how MAR processes can be generated.

3. ALGORITHMIC GENERATION OF MAR PROCESSES

We say that a missing data pattern is monotone if $R_k = 1$ implies $R_{k-1} = 1$ with probability one for $k = 2, \ldots, K$, that is, the kth variable is observed only if the $k-1$th variable was observed. When the missing data pattern is monotone, it is straightforward to generate MAR processes, since it is easy to show that $f(r \mid L) \equiv \text{pr}(R = r \mid L)$ is a missing at random process that is, equation (1) holds if and only if

$$\text{pr}\left[R_k = 1 \mid R_{k-1} = 1, L\right] = \text{pr}\left[R_k = 1 \mid R_{k-1} = 1, \bar{L}_{k-1}\right] \tag{6}$$

where $\bar{L}_{k-1} = (L_0, L_1, \ldots, L_{k-1})$ and $L_0 \equiv R_0 \equiv 1$. Since we are interested in $f(r \mid \ell)$, we can and do regard the data L on each subject as fixed (non-random) constants. To generate a monotone MAR process first, with probability $p_1 = \text{pr}(R_1 = 1)$, we observe variable L_1, and with probability $(1 - p_1)$ we quit and observe no variables. If we have observed $L_1, \ldots, L_{k-1}$ with conditional probability $p_k(\bar{L}_{k-1}) \equiv pr[R_k = 1 \mid R_{k-1} = 1, \bar{L}_{k-1}]$, we observe L_k and, with conditional probability $1 - p_k(\bar{L}_{k-1})$, we quit and observe no further variables. Thus

$$\text{pr}\left[R_1 = 1, \ldots, R_{k-1} = 1, R_k = 0, \ldots, R_K = 0 \mid L\right] = \left\{\prod_{m=1}^{k-1} p_m(\bar{L}_{m-1})\right\} \left\{1 - p_k(\bar{L}_{k-1})\right\}$$

which only depends on the observed components $\bar{L}_{k-1}$ of L.

A MAR process is said to be a missing-completely-at-random (MCAR) process if $\text{pr}\left[R = r \mid L\right]$ does not depend on any component of L, observed or unobserved. The *above* monotone MAR processes need not be MCAR. In contrast to monotone MAR processes, we now show that the generation of non-monotone MAR processes that are not MCAR is rather subtle. This is because the most straightforward approach, based on simple polytomous logistic regression models, fails to produce MAR processes that are not also MCAR. To see why, suppose for simplicity $L = X = (X_1, X_2)$ and let the multinomial random variable $R^\dagger$ record the missingness pattern: $R^\dagger = 0$ if both variables are observed; $R^\dagger = 1$ if only X_2 is observed; $R^\dagger = 2$ if only X_1 is observed; and let $R^\dagger = 3$ if neither variable is observed. A simple polytomous logistic regression model specifies that, for $k = 1, 2, 3$,

$$\text{pr}\left[R^\dagger = k \mid X_1, X_2\right] = \exp\left[\gamma_{0k} + \gamma_{1k}X_1 + \gamma_{2k}X_2\right] / \left\{1 + \sum_{k=1}^{3} \exp\left[\gamma_{0k} + \gamma_{1k}X_1 + \gamma_{2k}X_2\right]\right\}. \tag{7}$$

Since, for any MAR process, $\text{pr}\left[R^\dagger = 3 \mid X_1, X_2\right] = \text{pr}\left[R^\dagger = 3\right]$, we deduce that a necessary condition for (7) to be MAR is that $\gamma_{1k} = \gamma_{2k} = 0$ for $k = 1, 2, 3$. That is, the model (7) is MAR if and only if it is MCAR.

4. RANDOMIZED MONOTONE MISSINGNESS PROCESSES

We now introduce the class of randomized monotone missing (RMM) processes, a new class of non-monotone non-MCAR MAR processes. It will be convenient to divide the data $L = (X', Y')'$

into a subset X, any of whose components may be missing, and a subset Y whose components are always observed. Suppose $X = (X_1, \ldots, X_M)'$. For simplicity and without loss of generality, we shall describe an RMM process with $M = 3$. The process is generated as follows. We observe Y. Then we observe X_1 with probability p_1, X_2 with probability p_2, X_3 with probability p_3, or quit without observing any components of X with probability $1 - p_1 - p_2 - p_3$, where the probabilities $p_j \equiv p_j(Y)$ can depend on the observed value of Y. If a variable is observed, say, X_2, we next observe variable X_1, with conditional probability p_{21}, X_3 with conditional probability p_{23}, or quit without observing a second variable with probability $1 - p_{21} - p_{23}$, where each of the probabilities $p_{2j} \equiv p_{2j}(X_2, Y)$ can depend on the observed values of X_2 and Y. Suppose a first and second variable have been observed, say, X_2 then X_1. Then, with conditional probability p_{213}, we observe variable X_3, and, with probability $1 - p_{213}$, we quit without observing any further variables. The probability $p_{213} = p_{213}(X_1, X_2, Y)$ can depend on the observed values of X_1, X_2 and Y as well as the order in which X_1 and X_2 were observed. That is, $p_{213}(X_1, X_2)$ may differ from $p_{123}(X_1, X_2)$. Finally, the order of observation of the variables is not recorded for data analysis, so the final data available for analysis is restricted to the data $(R, L_{(R)})$. The data $(R, L_{(R)})$ records which variables were observed and their observed values.

Figure 1 is a graphical representation of an RMM process. In Figure 1, the dependence of the probabilities on the always observed variable Y is suppressed and q is short for "quit". Since the probability of observing a subsequent variable depends only on the values of the variables already observed, we would expect an RMM process to be an MAR process.

We can verify this fact by explicitly computing $\mathrm{pr}(R = r \mid L)$ based on the probabilities in Figure 1 by marginalizing over the possible orders (paths) in which the components of X have been observed. For example, $\mathrm{pr}\left[R = (0, 0, 0, 1) \mid X_1, X_2, X_3, Y\right] = 1 - \{p_1(Y) - p_2(Y) - p_3(Y)\}$ which only depends on Y and thus satisfies MAR. Also, $\mathrm{pr}\left[R = (1, 0, 1, 1) \mid X, Y\right] = p_1(Y)\, p_{13}(X_1, Y)$ $\{1 - p_{132}(X_1, X_3, Y)\} + p_3(Y)\, p_{31}(X_3, Y)\{1 - p_{312}(X_1, X_3, Y)\}$ since one must observe X_1 and X_3 in one of the two possible orderings and then, having observed X_1 and X_3, not observe X_2. This probability only depends on (X_1, X_3, Y) and hence satisfies MAR. Finally, $\mathrm{pr}\left[R = (1, 1, 1, 1) \mid X, Y\right]$ $= p_{13}^* + p_{12}^* + p_{23}^*$ depends on all components of X where, for example, $p_{13}^* \equiv p_{132}(X_1, X_3, Y)\, p_1(Y)$ $p_{13}(X_1, Y) + p_{312}(X_1, X_3, Y)\, p_3(Y)\, p_{31}(X_3, Y)$ is the probability of either the ordering (X_1, X_3, X_2) or the ordering (X_3, X_1, X_2). The probabilities $pr\left[R = (1, 1, 1, 1) \mid X, Y\right]$ are the probabilities that are estimated in the weighted estimating equation (5). Each is the sum of the ordering-specific probabilities of observing $R = \mathbf{1}$ over the $3! = 6$ orderings in which X_1, X_2, X_3 could have been observed.

We say that an MAR process $\mathrm{pr}\left[R = r \mid L\right] = \pi\left(r, L_{(r)}\right)$ is represented by an RMM process if there exists some RMM process for which $\mathrm{pr}\left[R = r \mid L\right]$ is obtained from the probabilities $p_j(Y)$, $p_{\ell j}(X_\ell, Y)$, $p_{j\ell k}(X_\ell, X_j, Y)$ and $p_{\ell jk}(X_\ell, X_j, Y)$ by marginalizing over the possible orderings of the observed components of X. It is a reasonable conjecture that any MAR process can be represented by some RMM process. However, Gill and Robins.[2] show that this conjecture is false. Specifically, the missingness process in Table I is MAR, since $\mathrm{pr}\left[R = (1, 1, 0) \mid X_1 = 0, X_2 = 0, X_3\right] = 1$ does not depend on X_3, $\mathrm{pr}\left[R = (1, 0, 1) \mid X_1 = 1, X_3 = 0, X_2\right] = 1$ does not depend on X_2, and $\mathrm{pr}\left[R = (0, 1, 1) \mid X_1, X_2 = 1, X_3 = 1\right] = 1$ does not depend on X_1. However, the process does not admit an RMM representation: If the process in Table I had such a representation, then p_1 must be 0, since otherwise, for all values of X, X_1 would have a positive probability of being observed, and so $p\left[R = (0, 1, 1) \mid X_1 = 0, X_2 = 1, X_3 = 1\right]$ could not be 1. By analogous arguments, p_2 and p_3 must be zero, which would imply that $p\left[R = (0, 0, 0) \mid X\right] = 1 - p_1 - p_2 - p_3$ would be 1 which is not true in Table I. Results of Gill and Robins.[2] also imply the existence of MAR processes with probabilities other than 0 or 1 that are not representable by an RMM process.
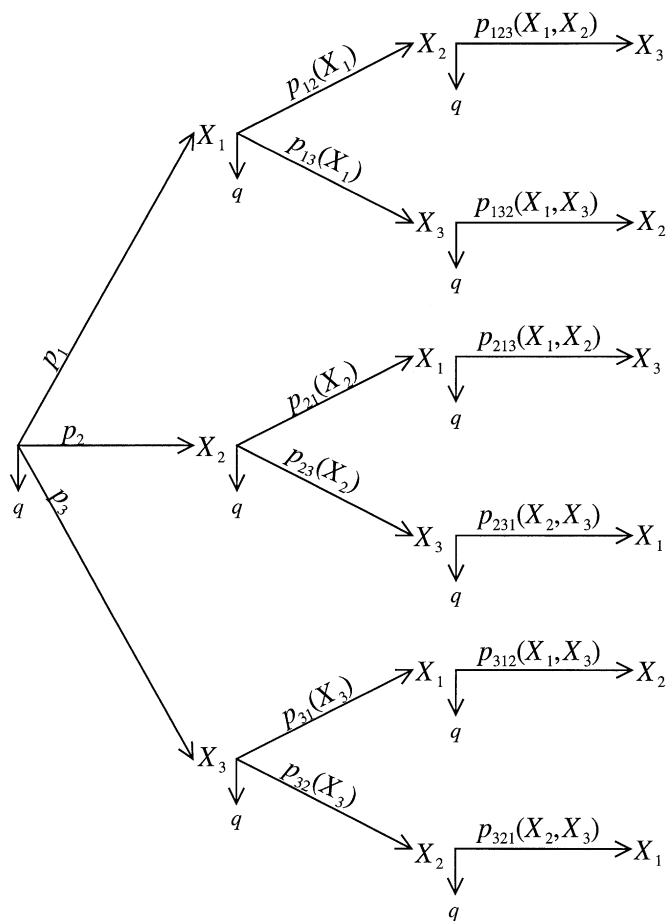
Figure. 1.

Table I. MAR process with no RMM representation

r_1	r_2	r_3	x_1	x_2	x_3	$\mathrm{pr}\,(R = r \mid X = x)$
1	1	0	0	0	0	1·0
1	0	1	1	0	0	1·0
1	1	1	0	1	0	1·0
1	1	0	0	0	1	1·0
1	0	1	1	1	0	1·0
1	1	1	1	0	1	1·0
0	1	1	0	1	1	1·0
0	1	1	1	1	1	1·0

To simulate an RMM process, the computer does not need to use more information about X than that which is finally revealed in the data $(R, L_{(R)})$. That is, if variable k is selected by the RMM algorithm, its value X_k is recorded in the final print statement and is available for data analysis. Only the order of selection is not recorded in the print statement. However, as noted by Gill and Robins,[2] to simulate a MAR mechanism which is not RMM, the computer program

requires, in the course of the procedure, information about X, perhaps even its precise value, which is ultimately hidden and not revealed in the final print statement that outputs $(R, L_{(R)})$. For example, to simulate the MAR mechanism of Table I, we could proceed as follows.

  Observe X_3.
  If $X_3 = 1$, observe X_2; if $X_2 = 1$, do not observe X_1. If $X_2 = 0$, observe X_1. If $X_1 = 0$, hide X_3.
  If $X_3 = 0$, observe X_1; if $X_1 = 1$, do not observe X_2. If $X_1 = 0$, observe X_2. If $X_2 = 0$, hide X_3.
  Finally, print the values of the observed and unhidden components of X.

  When the computer print statement prints out the observation $X_1 = 0$, $X_2 = 0$ and X_3 missing, the computer knows but has hidden the value of X_3. Gill and Robins[2] describe the fact that there are MAR mechanisms whose computer implementation requires that the computer needs to know more about X than it is willing to output in its final print statement by the slogan that 'MAR is more than it seems.'

  We have been unable to conceive of a plausible social, economic, physical or biological process that would generate MAR processes that are not RMM representable, due to the subtle and precise manner in which the data must be 'hidden' to insure that the process is MAR. That is, we believe that natural missing data processes that are not representable as RMM processes will be non-ignorable.

  When equation (3) is true, according to theorem 2.1, there is exactly one MAR law $f^\Delta (r \mid \ell)$ consistent with the observed data law $f(r, \ell_{(r)})$. If $f^\Delta (r \mid \ell)$ is not representable by an RMM process, then the observed data distribution could not have been generated by an RMM process, and, based on our prior beliefs, we might choose to reject the hypothesis that the data were generated by any MAR process. Hence, it would be important to test whether the observed data distribution could have been generated by an RMM process. We describe such a test in Section 7.2.

5. MARKOV RMM

A Markov randomized monotone missingness process is the special case of a randomized monotone missingness process in which the probabilities of observing a given variable can depend on the variables that have previously been observed but not on the order in which they have been observed. That is, in the notation of Figure 1, $p_{231}(X_2, X_3, Y) = p_{321}(X_2, X_3, Y)$, which we henceforth abbreviate to $p_1(X_2, X_3, Y)$. Markov randomized monotone missingness processes can be represented as in Figure 2, since once a set of variables has been observed, the conditional probability of observing a subsequent variable does not depend on the order in which the previous variables were observed. Again, the dependence on Y has been suppressed in Figure 2. We say a MAR process is represented by a Markov RMM process if it can be represented by an RMM process that is Markov. Gill *et al.*[2] prove that every MAR process that can be represented as an RMM process can also be represented as a Markov RMM process, so there is no loss and considerable computational advantage in restricting attention to Markov RMM processes.

6. PARAMETRIC MARKOV RANDOMIZED MONOTONE MISSINGNESS MODELS

In this section, we discuss the specification and fitting of parametric Markov RMM models. In Figure 2, at stage m, $m = 1, 2, \ldots, M + 1$, there are $\binom{M}{m-1} = M! / \{m-1\}! [M - (m-1)]!$ groups of $m - 1$ variables X^{mk}, $k = 1, \ldots, \binom{M}{m-1}$. For example, at stage $m = 3$, we have $3!/(2!1!) = 3$
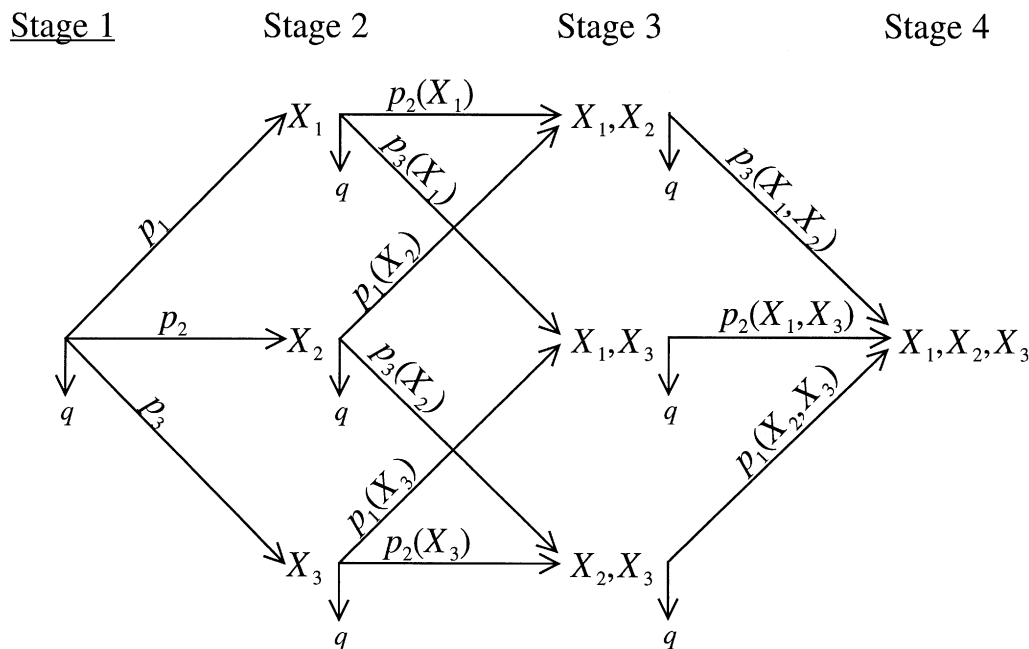
<u>Stage 1</u>　　　　　　Stage 2　　　　　　　Stage 3　　　　　　　Stage 4

Figure. 2.

groups of $m-1=2$ variables. Each group $X^{mk}, m \leqslant M$, is connected by arrows to the $M-(m-1)$ groups $\{X_j, X^{mk}\}$ at stage $(m+1)$ with $X_j \notin X^{mk}$. For example, if $X^{21} = \{X_1\}$, X^{21} is connected to the $3-(2-1)=2$ groups $\{X_1, X_2\}$ and $\{X_1, X_3\}$. The probabilities $p_j(X^{mk}, Y)$ on Figure 2 are the conditional probabilities that the variable $X_j, X_j \notin X^{mk}$, will be observed in the next stage conditional on the observed values of Y and of the variables X^{mk} that have been observed through stage m. $p_-(X^{mk}, Y) \equiv 1 - \sum_{\{j; X_j \notin X^{mk}\}} p_j(X^{mk}, Y)$ is the conditional probability of quitting without proceeding to the next stage.

We now discuss how we can fit a parametric Markov RMM model $p_j(X^{mk}, Y; \gamma_{mk})$ depending on a group X^{mk} specific parameter vector γ_{mk} for the $M-(m-1)+1 = M-m+2$ probabilities $p_j(X^{mk}, Y)$, and $p_-(X^{mk}, Y)$. Without loss of generality, we will assume that we specify a polytomous logistic model for the dependence of the $M-m+2$ probabilities on (X^{mk}, Y). In the worked example of Section 7, we use the polytomous logistic model

$$p_j(X^{mk}, Y; \gamma_{mk}) = \phi_{jmk} \Big/ \left\{ 1 + \sum_{\{j; X_j \notin X^{mk}\}} \phi_{jmk} \right\}$$

where

$$\phi_{jmk} \equiv \exp\left[\sum_{\{p; X_p \in X^{mk}\}} \gamma_{1pjmk} X_p + \gamma_{2jmk} Y + \gamma_{3jmk} \right] \tag{8}$$

with $\gamma_{mk} \equiv \{\gamma_{1jmk}, \gamma_{2jmk}, \gamma_{3jmk}; X_j \notin X^{mk}\}$ and $\gamma_{1jmk} = \{\gamma_{1pjmk}; X_p \in X^{mk}\}$. There are $\sum_{m=1}^{M} \binom{M}{m-1} = 2^M - 1$ separate parameter vectors γ_{mk} and thus models. Fitting these polytomous

logistic Markov RMM models would be straightforward if, for each subject, we had observed the order of observation of the variables, (that is, we had observed the subject's path through the graph). We would then fit the $2^M - 1$ models separately by finding $\hat{\gamma}_{mk}$ that maximizes the polytomous logistic likelihood for γ_{mk} where a subject contributes a weight 1 to the likelihood for γ_{mk} if, at stage m, he had the observed variables in X^{mk} and a weight 0 otherwise (that is, if the group X^{mk} lies on the subject's path). The dimensionality could be reduced by placing cross-model restrictions on the γ_{mk}. In that case, the $2^M - 1$ model-specific polytomous logistic likelihoods must be maximized jointly under the given restrictions.

Given that the subject-specific paths on the graph in Figure 2 are not observed, we regard the unavailable path information as missing and estimate the γ_{mk} using the EM algorithm. That is, for a Markov RMM process, the observed data is $(R, L_{(R)})$ and the complete data is $(R, L_{(R)}, Ph)$ where Ph is the path of length N taken by the subject on Figure 2 prior to quitting and N is the (random) number of components of X observed on that subject that is Ph is an ordering of the N observed components of X. Then there are $N!$ possible orderings (paths) consistent with a subject's observed data. In this setting, as shown in Appendix I, the EM algorithm is particularly straightforward. Given the current parameter estimates, the E-step assigns subject-specific weights $\hat{\omega}_\ell, \ell = 1, \ldots, N!$ corresponding to the estimated conditional probability given the subject's observed data $(R, L_{(R)})$ of each of the $N!$ compatible orderings. The M-step provides an updated estimate of γ_{mk} by maximizing a weighted polytomous logistic likelihood in which (i) each of the $\sum_{i=1}^n N_i!$ subject-specific compatible paths are treated as independent observations, and (ii) if the ℓth path for a subject, $\ell = 1, \ldots, N!$, corresponds to an ordering in which the variables in X^{mk} have been observed at stage m (that is, X^{mk} is on the ℓth path), the path contributes as an independent observation to a weighted polytomous likelihood for γ_{mk} with weight $\hat{\omega}_\ell$, and contributes a weight zero otherwise.

At convergence of the EM algorithm, the estimates of γ_{mk} will still depend on the starting values chosen for these parameters. However, the estimates of the missingness probabilities $\pi(r, \ell_{(r)})$ will, typically, be maximum likelihood and not depend on starting values. As discussed in Appendix II, this reflects the fact that the $\pi(r, \ell_{(r)})$ are, but the γ_{mk} may not be, identifiable.

As an example, consider a subject with $X = (X_1, X_2, X_3)$ fully observed so $N = 3$. In Figure 2, there are $N! = 6$ possible orderings of observation (that is, paths) by which the data (X_1, X_2, X_3) could have arisen. We compute estimated weights $\hat{\omega}_1, \ldots, \hat{\omega}_6$ corresponding to the various paths in the E-step. For example, suppose we let $\hat{\omega}_2$ correspond to the ordering (X_2, X_1, X_3). Then $\hat{w}_2 = \hat{w}_2^* / \sum_{j=1}^6 \hat{w}_j^*$ where, for example, $\hat{w}_2^* = p_2(Y) p_1(X_2, Y) p_3(X_1, X_2, Y)$ is the unconditional probability of the ordering (X_2, X_1, X_3). Note that the paths associated with the ordering (X_2, X_1, X_3) and (X_2, X_3, X_1) both pass through $X^{2k} = \{X_2\}$ at stage 2. In the M-step of the algorithm, each path through $X^{2k} = \{X_2\}$ is treated as an 'independent observation' in its contribution to the weighted likelihood for γ_{2k} with its own path-specific weight. For example, if both $X^{2k} = \{X_2\}$ and the group $\{X_1, X_2\}$ are on the ℓth path for a subject, the subject's path-specific contribution to the likelihood for γ_{2k} is the conditional probability $p_1(X_2, Y; \gamma_{2k})$ of observing variable X_1 next given the observed values of X_2 and Y. Note the multinomial outcome associated with the different paths in the likelihood for γ_{2k} depends on the path. Specifically, the path (X_2, X_3, X_1) is associated with the outcome 'observe variable X_3 next,' and the path (X_2, X_1, X_3) is associated with the outcome 'observe variable X_1 next.'

6.1. The Simulated EM Algorithm

When the dimension M of X is large, the EM algorithm becomes computationally unwieldy since, for a subject with complete data, there are $M!$ different orderings of the observations. Thus, at each iteration, the E-step must compute $M!$ conditional probabilities for the subject.

In this setting, we can still construct $n^{\frac{1}{2}}$-consistent estimators of the probabilities $\pi\left(r,\ell_{(t)};\gamma\right)$ of our Markov RMM model with a much diminished computational burden using the simulated (Sim) EM algorithm.[5] The Sim EM algorithm is a specific type of multiple imputation procedure.

The Sim EM algorithm replaces the E-step of the EM algorithm by an imputation step (described in Appendix III) in which V particular paths (that is, orderings) compatible with the subjects' observed data are imputed, thus filling in the missing path data V times. For example, for a subject with $R = (1,0,1,1)$ and (X_1,X_3,Y) observed, each imputation $v, v = 1,\dots,V$, imputes either the path $(X_1,X_3,quit)$ or the path $(X_3,X_1,quit)$. The M-step of the Sim EM algorithm maximizes the polytomous logistic likelihood for the parameters γ_{mk} as if we had $V \times n$ independent subjects whose paths were completely observed. That is, each of the $V \times n$ "pseudo-subjects" contributes a weight 1 or 0 to the likelihood for γ_{mk}, depending on whether their path passes through the variables X^{mk} at stage m.

As discussed in Appendix III, under regularity conditions, the estimates of $\pi\left(r,\ell_{(r)}\right)$ based on the Sim EM will be $n^{\frac{1}{2}}$-consistent even if the number of imputations V is as few as 1. A s the number of imputations $V \to \infty$, these estimates will become asymptotically equivalent to the maximum likelihood estimates obtained using the EM algorithm and thus asymptotically normal and efficient.[6,14] However, for a fixed finite number V of imputations, the Sim EM estimates of $\pi\left(r,\ell_{(r)}\right)$ are not guaranteed to be asymptotically normal and unbiased (although $n^{\frac{1}{2}}$-consistent) because there are non-identifiable parameters γ_{mk} in the model.

Even when the dimension M of X is small (as in our example), it is quite a bit more difficult to program the EM algorithm than the Sim EM algorithm. As a consequence, the analyses reported in this paper are based on the Sim EM algorithm.

7. DATA ANALYSIS

7.1. A Worked Example

Data on 529 second breast cancer cases and 529 controls (drawn from women whose initial breast cancer was not followed by the development of a second cancer) were obtained from the conference organizers. Details of the data collection and other substantive considerations are described in References 7 and 8. In this section, we shall analyse the association of radiation with second breast cancer while adjusting for the potential confounding variables of family history and the weight/height^2 by fitting the logistic model (4) where Y is the second breast cancer (case-control) indicator, $X = (X_0,X_1,X_2,X_3)'$, $\alpha' = (\alpha_0,\alpha_1,\alpha_2,\alpha_3)'$. Here $X_0 \equiv 1$; $X_1 = 1$ if a subject received radiation therapy, $X_1 = 0$ otherwise; $X_2 = 1$ if a subject had a family history of breast cancer, $X_2 = 0$ otherwise; X_3 is an ordinal variable taking the values $1,2,3$, or 4 as $10^4 \times$ weight/height^2 was $<20, 20–25, 25–30, >30$. The outcome Y is available on all subjects. However, X_1 was missing on 3 per cent of the subjects, X_2 on 49 per cent, and X_3 on 43 per cent. Let $R_j = 1$ if X_j was observed, and $R_j = 0$ otherwise. Table II gives the joint distribution of R_j's. Reading from the last row of Table II, we see that only 33.6 per cent of the subjects had complete covariate data. Furthermore, the missingness is non-monotone.

Table III gives estimates of the regression coefficients, $\alpha_1,\alpha_2,\alpha_3$ using two different estimation methods. Row 1 is based on logistic regression restricted to complete cases. The complete case estimator of row 1 will be biased when the missingness process depends on both the covariates X and the outcome Y.

The estimates in rows 2 and 3 were obtained by estimating the parameter α of the logistic regression model (4) using the weighted estimating equation (5) where $\hat{\gamma}$ was obtained by fitting the Markov RMM polytomous logistic regression model (8) by the Sim EM algorithm with $V = 5$

Table II.

R_1	R_2	R_3	%
0	0	0	1·2
1	0	0	27·4
0	1	0	0·9
1	1	0	14·8
0	0	1	0·2
1	0	1	21·3
0	1	1	0·38
1	1	1	33·6

Table III.

Method of estimation	$\hat{\alpha}_1$ (*Rad*)	$\hat{\alpha}_2$ (*FH$_x$*)	$\hat{\alpha}_3$ (*Wt/Ht*)
Complete case	0·24	0·43	0·31
RMM ($V = 5$)	0·19	0·44	0·31
RMM ($V = 200$)	0·19(0·28)	0·44(0·29)	0·32(0·13)

and $V = 200$ imputations. The standard errors in row 3 were calculated as described in Section 8 under the assumption that the Sim EM estimator with $V = 200$ has variance essentially equal to that of the EM estimator.

7.2. Testing for representation by an RMM process

In this section, we assume that the complete data L is discrete taking values in a set $\mathbf{L}$. We wish to test the hypothesis that the observed data distribution $f(r, \ell_{(r)})$ could have been generated by an RMM process. Let $\tilde{f}(r, \ell_{(r)})$ be the empirical distribution of the observed data. For example, if as in Section 2, $L = (X_1, X_2, X_3, Y)'$, $\tilde{f}\{r = (1, 1, 0, 1), \ell_{(r)} = (1, 0, 0)\}$ is the proportion of the n study subjects with $X_1 = 1$, $X_2 = 0$, and $Y = 0$ observed, but X_3 missing.

Suppose that $\tilde{f}(\mathbf{1}, \ell) \neq 0$ for all $\ell \in \mathbf{L}$. Then by the empirical analogue of theorem 2.1, there exists a unique $\tilde{f}^{\Delta}(r \mid \ell) \equiv \tilde{\pi}^{\Delta}(r, \ell_{(r)})$ and $\tilde{f}^{\Delta}(\ell)$ such that $\tilde{f}(r, \ell_{(r)})$ is the marginal distribution of $(R, L_{(R)})$ under the joint law $\tilde{f}^{\Delta}(\ell) \tilde{f}^{\Delta}(r \mid \ell)$. If $\tilde{f}^{\Delta}(r \mid \ell)$ can be represented by an RMM process, then there is no data evidence contradicting the hypothesis that the law $f^{\Delta}(r \mid \ell)$ of theorem 2.1 is representable by an RMM process. Theorem 7.1 below gives necessary and sufficient conditions for $\tilde{f}^{\Delta}(r \mid \ell)$ to be represented by an RMM process. If, using theorem 7.1, we conclude $\tilde{f}^{\Delta}(r \mid \ell)$ cannot be represented by an RMM process, it nonetheless remains possible that $f^{\Delta}(r \mid \ell)$ can be so represented but, due to sampling variability, its empirical counterpart $\tilde{f}^{\Delta}(r \mid \ell)$ cannot. In such a case, we would like to construct a test of the null hypothesis that $f^{\Delta}(r \mid \ell)$ can be represented. We discuss a possible test below.

Theorem 7.1. Given $\tilde{f}(\mathbf{1}, \ell) \neq 0$ for all $\ell \in \mathbf{L}$, $\tilde{f}^{\Delta}(r \mid \ell)$ can be represented by an RMM process if and only if $\tilde{f}(r, \ell_{(r)})$ is the marginal distribution of $(R, L_{(R)})$ corresponding to the joint law $\hat{f}(r \mid \ell) \hat{f}(\ell)$ where $\hat{f}(r \mid \ell) \equiv \hat{\pi}(r, \ell_{(r)})$ is the non-parametric maximum likelihood estimator (NPMLE) of $f(r \mid \ell)$ under the sole restriction that $f(r \mid \ell)$ is represented by an RMM process and $\hat{f}(\ell) \equiv \tilde{f}(\mathbf{1}, \ell) / \hat{f}(\mathbf{1} \mid \ell)$.

Proof. If $\tilde{f}^{\Delta}(r \mid \ell)$ can be represented, then $\hat{f}(r \mid \ell) = \tilde{f}^{\Delta}(r \mid \ell)$ since $\tilde{f}^{\Delta}(r \mid \ell)$ is the NPMLE of $f(r \mid \ell)$ under the MAR assumption (1). Further, by uniqueness of $\tilde{f}^{\Delta}(\ell)$, $\tilde{f}^{\Delta}(\ell)$ must equal $\hat{f}(\ell)$. Since $\hat{f}(r \mid \ell)$ is an MAR process, the converse holds by the empirical analogue of theorem 2.1.

Example: We can use theorem 7.1 to show that $\tilde{f}^{\Delta}(r \mid \ell)$ is not representable by an RMM process for the breast cancer data. To simplify the calculations we redefined X_3 to be a dichotomous indicator variable representing whether the $10^4 \times$ height/weight^2 ratio exceeds 25, so that all variables are now dichotomous. To obtain $\hat{f}(r \mid \ell)$, we can fit using the EM algorithm the Markov RMM linear logistic model in which $p_j(X^{mk}, Y; \gamma_{mk})$ is saturated; that is, it is given by equation (8) expanded so that all orders of interaction between the variables X_p themselves and Y are included. In practice, since we did not have the EM algorithm programmed, we used the Sim EM algorithm with $V = 500$ to approximate the EM algorithm. We first calculated $\hat{f}(\ell) = \tilde{f}(\mathbf{1}, \ell) / \tilde{f}(\mathbf{1} \mid \ell)$. We then calculated the marginal distribution $\hat{f}(r, \ell_{(r)})$ of $(R, L_{(R)})$ based on the joint law $\hat{f}(r \mid \ell) \hat{f}(\ell)$ and noted that, for several values of $(r, \ell_{(r)})$, $\hat{f}(r, \ell_{(r)})$ differed from $\tilde{f}(r, \ell_{(r)})$ in the second decimal place. This implies that $\tilde{f}^{\Delta}(r \mid \ell)$ is not representable by a Markov RMM process and, thus, by any RMM process. However, we remain concerned that the difference between $\hat{f}(r, \ell_{(r)})$ and $\tilde{f}(r, \ell_{(r)})$ may have been entirely due to the fact that we used the Sim EM rather than the EM algorithm and so only obtained an approximation to the non-parametric maximum likelihood estimator $\hat{f}(r \mid \ell)$.

When $\hat{f}(r, \ell_{(r)}) \neq \tilde{f}(r, \ell_{(r)})$, the NPMLE of $f(r, \ell_{(r)})$ under the sole restriction that $f^{\Delta}(r \mid \ell)$ is representable by an RMM process is the marginal distribution of $(R, L_{(R)})$ corresponding to the joint law $\tilde{f}^{\Delta}(\ell) \hat{f}(r \mid \ell)$. $\tilde{f}^{\Delta}(\ell)$ is the unique law $f(\ell)$ maximizing the log likelihood $\sum_i \sum_r I(R_i = r) \log f(L_{i(r)})$ and can be calculated using the EM algorithm.

When $\hat{f}(r, \ell_{(r)}) \neq \tilde{f}(r, \ell_{(r)})$, we suggest the following heuristic bootstrap procedure to assign a p-value to the hypothesis that $f^{\Delta}(r \mid \ell)$ of theorem 2.1 can be represented by an RMM process.

Step 1: Compute $T^2 \equiv \sum_{(r, \ell_{(r)})} \left[\hat{f}(r, \ell_{(r)}) - \tilde{f}(r, \ell_{(r)}) \right]^2$.

Step 2: Simulate 1000 data sets of n independent observations from the joint law $\tilde{f}^{\Delta}(\ell)$ $\hat{f}(r \mid \ell)$. Let $\tilde{f}_u(r, \ell_{(r)})$, $\hat{f}_u(r, \ell_{(r)})$ and T_u^2 be computed as above, except from data set u, $u = 1, \dots, 1000$. Report the p-value as the fraction of the T_u^2 that exceed T^2.

Simulation and theoretical investigation of this statistic remain to be done. However, results of Gill and Robins.[2] guarantee that there will exist MAR processes that are not RMM representable for which an α-level test based on the T^2 statistic will be consistent (that is, have power converging to 1 with increasing sample size).

8. VARIANCE ESTIMATION

In the appendix we sketch a proof of the following:

Theorem 8.1. Suppose (i) the full data model (4) for the conditional distribution of Y given X is correctly specified, (ii) a Markov RMM polytomous logistic model with parameters γ_{mk} is correctly specified, and (iii) $\hat{\gamma}$ is fit using either the EM or the Sim EM algorithm with $V \to \infty$ as $n \to \infty$. Then, under regularity conditions, $n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0)$ is asymptotically normal with mean zero

and variance that can be consistently estimated by $\hat{I}^{-1} \hat{\Omega}_{\mathrm{EM}} \hat{I}^{-1'}$ *where* $\hat{I} = n^{-1} \sum_i \partial U_i(\hat{\alpha}, \hat{\gamma}) / \partial \alpha$ *and* $\hat{\Omega}_{\mathrm{EM}}$ *is defined below.*

Definition of $\hat{\Omega}_{\mathrm{EM}}$**:** Let γ be a stacked vector of the γ_{mk}. We suppose we have implemented the Sim EM algorithm using a large number of imputations V. Let $S_{\gamma v}(\gamma)$ be a vector of the dimension of γ denoting the contribution of a subject's vth path to the polytomous logistic likelihood score for $\gamma, v = 1, \ldots, V$. Let $S_{\gamma v}^{mk}(\gamma)$ denote the subvector of $S_{\gamma v}(\gamma)$ corresponding to the parameters γ_{mk}. In Appendix I, we show $S_{\gamma v}^{mk}(\gamma) = 0$ unless the group X^{mk} lies on the vth path. If X^{mk} lies on the vth path, $S_{\gamma v}^{mk}(\gamma)$ is the polytomous logistic likelihood score; $\partial \log p_j(X^{mk}, Y; \gamma_{mk}) / \partial \gamma_{mk}$ if variable X_j is newly observed at stage $m + 1$ on path v and $\partial \log p_-(X^{mk}, Y; \gamma_{mk}) / \partial \gamma_{mk}$ if path v terminates at X^{mk} (that is, no further variables are observed past stage m).

Let $\bar{S}_\gamma(\gamma) = V^{-1} \sum_{v=1}^V S_{\gamma v}(\gamma)$ be a subject's average score contribution. Then $\tilde{Q}(\alpha, \gamma) = n^{-1} \sum_i U_i(\alpha, \gamma) \bar{S}_{\gamma i}(\gamma)'$ is an empirical estimate of $Q(\alpha, \gamma) \equiv E[U(\alpha, \gamma) \bar{S}_\gamma(\gamma)']$. A symmetrized non-negative definite estimate of $v(\gamma) \equiv E[S_{\gamma v}(\gamma) S_{\gamma v^*}'(\gamma)]$ is

$$\tilde{v}(\gamma) \equiv \sum_{v^*=1}^V \sum_{v=1}^V I(v \neq v^*) \left\{ n^{-1} \sum_i S_{\gamma v i}(\gamma) S_{\gamma v^* i}(\gamma)' + S_{\gamma v^* i}(\gamma) S_{\gamma v i}(\gamma)' \right\} / \{2V(V-1)\}.$$

Then $\hat{\Omega}_{\mathrm{EM}} = n^{-1} \sum_i U_i(\hat{\alpha}, \hat{\gamma})^{\otimes 2} - \tilde{Q}(\hat{\alpha}, \hat{\gamma}) v^\dagger(\gamma) \tilde{Q}(\hat{\alpha}, \hat{\gamma})'$ where $v^\dagger(\gamma)$ is the Moore–Penrose generalized inverse of $\tilde{v}(\gamma)$ and $b^{\otimes 2} \equiv bb'$.

9. DISCUSSION

We have argued that the MAR processes that do not have an RMM representation are implausible. Hence, in practical applications with non-monotone missing data, we suggest an analyst perform a test, such as that proposed in Section 7.2, of the hypothesis that the missing process is RMM representable. If the test rejects, analysis of the data based on the MAR assumption should either be avoided altogether or be viewed as only a rough approximation to a more appropriate analysis based on a non-ignorable model. In practice, the frequency of rejection will depend, in part, on the relative volumes (w.r.t. Lesbegue measure) of MAR laws that do and do not admit RMM representations. Calculation of these relative volumes is an open research problem.

Even if the test indicates that the data are compatible with having been generated by an RMM representable process, an analyst should still be reluctant to analyse that data under the assumption that missingness is MAR unless he or she holds the opinion, based on subject matter considerations, that the missingness process might be RMM representable.

To help analysts clarify their opinions, we consider whether some previously proposed non-monotone missing data processes are RMM representable.

Consider the factor analytic missingness process under which, conditional on an always unobserved latent factor U, L and its vector R of missing data indicators are independent, that is, $f(r \mid \ell, u) = f(r \mid u)$. The factor analytic missingness process is non-ignorable and thus not RMM representable since, upon marginalizing over the always unobserved U, the density $f(r \mid \ell)$ of R given L depends on the unobserved components $\ell_{(\bar{r})}$ of ℓ. Rubin[9] argued that, in the substantive context of election polls in Slovenia, a factor analytic missing process would be satisfactorily approximated by regarding it as MAR in the analysis. We would be less sanguine. Simulation studies might be useful in clarifying the accuracy of the approximation.

Robins et al.[10] and Mark and Gail[11] independently proposed an "observed past missingness process" under which the conditional probability the kth variable is observed depends only on

the observed past; formally $\mathrm{pr}\,(R_k = 1 \mid R_1, ..., R_{k-1}, L)$ depends on L only through the observed past $R_1 L_1, ..., R_{k-1} L_{k-1}$ for each k. An observed past missingness process is RMM representable. Indeed, the following generalization is representable. Define a permutation observed past missing process to be one where one of the $K!$ orderings (permutations) of the K variables is chosen at random according to some distribution. Then missing data is generated by an observed past missingness process based on the selected ordering. Finally, data $(R, L_{(R)})$ on the values of the observed variables are recorded but no record is kept of the chosen permutation. In fact, not only is every permutation observed past missing process representable by an RMM process but the converse is true; any RMM process is representable by a permutation observed past missingness process.

Appendix I: Derivation of the form of the EM Algorithm

Under our Markov RMM model, $\log \mathcal{L}(\gamma) \equiv \log f(Ph, R \mid L; \gamma) = \sum_{m=1}^{M} \sum_{k=1}^{\binom{M}{m-1}} \mathcal{L}_{mk}(\gamma_{mk})$, where $\mathcal{L}_{mk}(\gamma_{mk}) = \sum_{X_j \notin X^{mk}} I\left[\{X_j, X^{mk}\} \text{ lies on } Ph\right] \log p_j\left(X^{mk}, Y; \gamma_{mk}\right) +$ $I\left[Ph \text{ ends at } X^{mk}\right] p_-\left(X^{mk}, Y; \gamma_{mk}\right)$. Thus, with the current estimate of γ being γ^*, we obtain in the E-step[6] for each subject

$$E\left[\log \mathcal{L}(\gamma) \mid R, L_{(R)}; \gamma^*\right] = \sum_{m=1}^{M} \sum_{k=1}^{\binom{M}{m-1}} E\left[\log \mathcal{L}(\gamma_{mk}) \mid R, L_{(R)}; \gamma^*\right], \tag{9}$$

where

$$E\left[\log \mathcal{L}(\gamma_{mk}) \mid R, L_{(R)}; \gamma^*\right] =$$
$$\sum_{\ell=1}^{N!} \omega_\ell(\gamma^*) I\left[\{X_j, X^{mk}\} \text{ lies on } \ell^{\text{th}} \text{ path}\right] \log p_j\left(X^{mk}, Y; \gamma_{mk}\right) \tag{10}$$

where $\omega_\ell(\gamma^*) = \mathrm{pr}\left[\text{path } \ell \text{ taken} \mid R, L_{(R)}; \gamma^*\right]$ is called $\hat{\omega}_\ell$ in the text. In the M-step[6], we maximize over γ the quantity $\sum_{i=1}^{n} E\left[\log \mathcal{L}_i(\gamma) \mid R_i, L_{i(R_i)}; \gamma^*\right]$ with γ^* held fixed. When $p_j\left(X^{mk}, Y; \gamma_{mk}\right)$ is of polytomous logistic form, this is the weighted log-likelihood for a series of independent polytomous logistic regressions. $S_{\gamma v}^{mk}(\gamma_{mk}) = \partial \log \mathcal{L}_{mk}(\gamma_{mk}) / \partial \gamma_{mk}$ with Ph replaced by Ph_v.

Appendix II: Sketch of proof of theorem 8.1

Let $g_1(\gamma)$ be a function of γ such that $\pi(r, \ell_{(r)}; \gamma)$ depends on γ only through $g_1(\gamma)$ in the sense that $\pi(r, \ell_{(r)}; \gamma_1) = \pi(r, \ell_{(r)}; \gamma_2)$ for all $(r, \ell_{(r)})$ if and only if $g_1(\gamma_1) = g_1(\gamma_2)$. In general, $g_1(\gamma)$ will be a many to one map. Let $\phi = g(\gamma) \equiv \left(g_1(\gamma)' \, g_2(\gamma)'\right)'$ be a one to one reparameterization of γ. Write $\theta = g_1(\gamma), \psi = g_2(\gamma)$ so $\phi = (\theta', \psi')'$. Then we can rewrite $\pi(r, \ell_{(r)}; \gamma)$ and $U(\alpha, \gamma)$ as $\pi(r, \ell_{(r)}; \theta)$ and $U(\alpha, \theta)$. Note $f(Ph, R \mid L; \phi) = f\left(Ph \mid R, L_{(R)}; (\theta, \psi)\right) \pi(R, L_{(R)}; \theta)$. It follows from theorem 2.1 that θ is identified. We denote by θ_0 the true value of θ generating the data. In contrast, since only the data $(R, L_{(R)})$ is observed, ψ is not identified. Indeed, any value of ψ compatible with θ_0 could have generated the data. Both the EM and Sim EM algorithms are invariant under reparameterization in the sense that $\hat{\phi} = g(\hat{\gamma})$ for EM and Sim EM estimators $\hat{\gamma}$ and $\hat{\phi}$.

Since $\hat{\alpha}$ satisfies $0 = \sum_i U_i(\hat{\alpha}, \hat{\theta})$ and $E[U(\alpha_0, \theta_0)] = 0$ under the assumptions of the theorem, it follows that, under mild regularity conditions, by a Taylor series expansion,

$$n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0) = -I^{-1}\left\{n^{-\frac{1}{2}} \sum_i U_i + E\left[\partial U(\alpha_0, \theta_0)/\partial \theta'\right] n^{\frac{1}{2}}\left(\hat{\theta} - \theta_0\right) + o_p(1)\right\} \tag{11}$$

where $U_i = U_i(\alpha_0, \theta_0)$, $I = E\left[\partial U(\alpha_0, \theta_0)/\partial \alpha'\right]$. Hence, to obtain the asymptotic distribution of $n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0)$, we need to derive the asymptotic distribution of $n^{-\frac{1}{2}}(\hat{\theta} - \theta_0)$ under the EM algorithm.

Let $S_\phi^{\text{obs}}(\phi) \equiv \partial \log \pi\left(r; \ell_{(r)}; \theta\right)/\partial \phi \equiv \left(S_\theta^{\text{obs}}(\phi)', S_\psi^{\text{obs}}(\phi)'\right)' = \left(S_\theta^{\text{obs}}(\theta)', 0'\right)'$. Under general conditions, $\hat{\theta}$ is maximum likelihood. Hence, by standard likelihood theory,

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) = n^{-\frac{1}{2}} \sum_i E[S_\theta^{\text{obs}} S_\theta^{\text{obs}'}]^{-1} S_\theta^{\text{obs}} + o_p(1)$$

where $S_\theta^{\text{obs}} \equiv S_\theta^{\text{obs}}(\theta_0)$. Further, by the extended information equality[12], $E\left[\partial U(\alpha_0, \theta_0)/\partial \theta'\right] = -E[US_\theta^{\text{obs}'}]$. Hence, substituting into (11)

$$n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0) = -I^{-1} n^{-\frac{1}{2}} \sum_i \text{Resid}_i\left(U, S_\theta^{\text{obs}}\right) + o_p(1) \tag{12}$$

where $Resid\left(U, S_\theta^{\text{obs}}\right) \equiv U - E\left[US_\theta^{\text{obs}'}\right]\left\{E\left[S_\theta^{\text{obs}\otimes 2}\right]\right\}^{-1} S_\theta^{\text{obs}}$ is the residual from the population regression of U on S_θ^{obs}. Hence, $n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0)$ is asymptotically normal with asymptotic variance

$$I^{-1}\left\{\text{var}(U) - E\left[US_\theta^{\text{obs}'}\right] E\left[S_\theta^{\text{obs}\otimes 2}\right]^{-1} E\left[US_\theta^{\text{obs}'}\right]'\right\} I^{-1'}.$$

Thus it only remains to prove the consistency of the variance estimator $\hat{\Omega}_{\text{EM}}$. To do so, first note

$$E\left[US_\theta^{\text{obs}'}\right] E\left[S_\theta^{\text{obs}\otimes 2}\right]^{-1} E\left[US_\theta^{\text{obs}'}\right]' =$$

$$E\left[US_\phi^{\text{obs}'}\right] E\left[S_\phi^{\text{obs}\otimes 2}\right]^\dagger E\left[US_\phi^{\text{obs}'}\right]' \tag{13}$$

since the addition of the zeros of S_ψ^{obs} to S_θ^{obs} has no effect except that we need to use the generalized inverse $E\left[S_\phi^{\text{obs}\otimes 2}\right]^\dagger$ due to $E\left[S_\phi^{\text{obs}\otimes 2}\right]$ being singular. However, $S_\phi^{\text{obs}} = TS_\gamma^{\text{obs}}(\gamma_0)$ for a matrix T where $\gamma_0 = g^{-1}(\theta_0, \psi^*)$ and ψ^* is an arbitrary value of ψ. Upon substituting $TS_\gamma^{\text{obs}}(\gamma_0)$ into (13), by some matrix algebra, we find (13) is algebraically equal to $E\left[US_\gamma^{\text{obs}}\right] E\left[S_\gamma^{\text{obs}\otimes 2}\right]^\dagger E\left[US_\gamma^{\text{obs}'}\right]'$ where $S_\gamma^{\text{obs}} = S_\gamma^{\text{obs}}(\gamma_0)$. Hence the asymptotic variance of $n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0)$ is

$$I^{-1}\left\{\text{var}(U) - E\left[US_\gamma^{\text{obs}'}\right] E\left[S_\gamma^{\text{obs}\otimes 2}\right]^\dagger E\left[US_\gamma^{\text{obs}'}\right]'\right\} I^{-1'}. \tag{14}$$

A consistent variance estimator can be obtained by taking appropriate sample averages of $U(\hat{\alpha}, \hat{\gamma})$ and $\hat{S}_\gamma^{\text{obs}}(\hat{\gamma}) = \partial\left\{E\left[\log \mathscr{L}(\hat{\gamma}) \mid R, L_{(R)}; \hat{\gamma}\right]\right\}/\partial \gamma$ with $E\left[\log \mathscr{L}(\gamma) \mid R, L_{(R)}; \gamma\right]$ as defined in equations (9) and (10).[5,13] Note $\hat{\gamma} = g^{-1}(\hat{\theta}, \hat{\psi})$ converges to $\gamma_0 = g^{-1}(\theta_0, \psi_0)$ where ψ_0 depends on the starting value for γ in the EM algorithm. The alternative simulation estimator $\hat{\Omega}_{\text{EM}}$ of the asymptotic variance given in the text is derived below.

The Simulation Variance Estimator $\hat{\Omega}_{\text{EM}}$

Let $S_{\gamma v}(\gamma) = \partial \log f(Ph_v, R \mid L; \gamma)/\partial \gamma$. We shall need the following two lemmas. Expectations are with respect to laws indexed by γ_0 (equivalently, by $\phi_0 = (\theta_0, \psi_0)$). All scores are evaluated at γ_0 (that is, ϕ_0) unless shown otherwise.

Lemma A1: $E\left[US'_{\gamma v}\right] = E\left[US^{\mathrm{obs}'}_{\gamma}\right]$.

Lemma A2: $E\left[S_{\gamma v}S'_{\gamma v*}\right] = E\left[S^{\mathrm{obs}\otimes 2}_{\gamma}\right]$ if $v \neq v^*$.

Proof of lemma A1 : $E\left[US'_{\gamma v}\right] = E\left[US^{\mathrm{obs}'}_{\gamma}\right] + E\left[US^{\mathrm{mis}'}_{\gamma}\right]$ where $S^{\mathrm{mis}}_{\gamma}(\gamma) = \partial \log f\left[Ph_v \mid R, L_{(R)}; \gamma\right] / \partial \gamma$. But $E\left[US^{\mathrm{mis}'}_{\gamma}\right] = E\left\{UE\left[S^{\mathrm{mis}'}_{\gamma} \mid R, L_{(R)}\right]\right\} = 0$ since U is a function of $\left(R, L_{(R)}\right)$ and $S^{\mathrm{mis}}_{\gamma}$ is a conditional score.

Proof of lemma A2 : $E\left[S_{\gamma v}S'_{\gamma v*}\right] = E\left[S^{\otimes 2}_{\mathrm{obs}}\right] + E\left[S^{\mathrm{obs}}_{\gamma v}S^{\mathrm{mis}'}_{\gamma v*}\right] + E\left[S^{\mathrm{mis}}_{\gamma v}S^{\mathrm{obs}'}_{\gamma v*}\right] + E\left[S^{\mathrm{mis}}_{\gamma v}S^{\mathrm{mis}'}_{\gamma v*}\right]$. Now, since $S^{\mathrm{obs}}_{\gamma v}$ is a function of $\left(R, L_{(R)}\right)$, the second and third terms are zero by the argument of the Proof of Lemma A1. Finally, $E\left[S^{\mathrm{mis}}_{\gamma v}S^{\mathrm{mis}'}_{\gamma v*}\right] = E\{E\left[S^{\mathrm{mis}}_{\gamma v} \mid R, L_{(R)}\right] E[S^{\mathrm{mis}'}_{\gamma v*} \mid R, L_{(R)}]\} = 0$ since, given $\left(R, L_{(R)}\right)$, Ph_v and Ph_{v*} and thus $S^{\mathrm{mis}}_{\gamma v}$ and $S^{\mathrm{mis}}_{\gamma v*}$ are independent.

Consistency of $\hat{\Omega}_{\mathrm{EM}}$ now follows if we have drawn the Ph_v using the EM estimator at convergence $\hat{\gamma} = \left(\hat{\theta}, \hat{\psi}\right)$ since, by lemma A1 and A2, $\tilde{Q}(\hat{\alpha}, \hat{\gamma})$ will converge to $E\left[US^{\mathrm{obs}'}_{\gamma}\right]$ and $\tilde{v}(\hat{\gamma})$ will converge to $E\left[S^{\mathrm{obs}\otimes 2}_{\gamma}\right]$.

Appendix III: SIM EM Algorithm

The Imputation Step of the Sim EM Algorithm

For subjects with $N = 0$ (no variables observed) or $N = 1$ (a single variable observed), the path in Figure 2 is known, and we 'impute' the known path in each imputation $v, v = 1, \ldots, V$.

For a subject with $N > 1$, for each imputation v separately, we (i) draw $N - 1$ independent random numbers $u_1, \ldots, u_{N-1}$ from uniform distribution on $(0, 1)$ and then (ii) impute a single path recursively in reverse order as follows. Let $\tilde{p}_{(N+1)j}$ be the (current estimate of) the conditional probability, given the observed data $\left(R, L_{(R)}\right)$ that the variable newly observed at the subject's final stage $N + 1$ is variable X_j. For example, if (X_1, X_2, X_3, Y) is fully observed, so $N = 3$ and $N + 1 = 4$, $\tilde{p}_{41} = \hat{p}^{\dagger}_{41} / \sum_{j=1}^{3} \hat{p}^{\dagger}_{4j}$ where $\hat{p}^{\dagger}_{41} = \hat{\imath}(X_2, X_3, Y)\{\hat{p}_3(Y)\hat{p}_2(X_3, Y) + \hat{p}_2(Y)\hat{p}_3(X_2, Y)\}$ is the probability that the true but unobserved path (ordering) is either (X_2, X_3, X_1) or (X_3, X_2, X_1).

Given the $\tilde{p}_{(N+1)j}$, we randomly impute the variable X_j newly "observed" at stage $N + 1$ based on the random draw u_{N-1}.

Then recursively for $m = N, N - 1, \ldots, 3$, let $\tilde{p}_{mj}$ be the current conditional probability that X_j was the variable newly "observed" at stage m given Y and the total set X^{mk} of variables "observed" through this stage. Given the $\tilde{p}_{mj}$, we impute the variable X_j newly "observed" at stage m using the random number u_{m-2}. When the algorithm terminates at $m = 3$, we have imputed an entire path. This procedure is repeated V times so that V total paths are imputed. For example, suppose $N = 3$ and we previously imputed that X_2 was newly "observed" at stage $N + 1 = 4$. Then $X^{3k} = \{X_1, X_3\}$ are the variables through stage $m = 3$. Then $\tilde{p}_{33} = \hat{p}_1(Y)\hat{p}_3(X_1, Y) / \{\hat{p}_1(Y)\hat{p}_3(X_1, Y) + \hat{p}_3(Y)\hat{p}_1(X_3, Y)\}$ and $\tilde{p}_{31} = 1 - \tilde{p}_{33}$. Suppose we impute X_1 to be newly "observed" at stage 3 based on the random number $u_{3-2} = u_1$, and the probabilities $\tilde{p}_{33}$ and $\tilde{p}_{31}$. Then the entire imputed path is (X_3, X_1, X_2).

It is critical that in each iteration of the Sim EM algorithm the uniform random numbers are not redrawn. Otherwise, the algorithm will fail to converge.[5] That is, a total of $V(N - 1)$ random numbers are drawn for each subject with $N > 1$.

Sketch of properties of Sim EM Algorithm with non-identified parameters

Adopting the notation of Appendix II, let $S_\phi^{\mathrm{mis}}\left(Ph_v \mid R, L_{(R)}; \phi\right) = \partial \log f\left[Ph_v \mid R, L_{(R)}; \phi\right]/\partial \phi$. Write $Ph_v = ph(u_v, \phi^*)$ to emphasize the functional dependence of Ph_v on the vector u_v of $N-1$ independent uniform random draws and a parameter ϕ^*. Now define $Q_b(\psi, \theta) = \sum_i \sum_{v=1}^{V} S_b^{\mathrm{mis}}\left\{ph(u_{iv}; \psi, \theta) \mid R_i, L_{i(R_i)}; \psi, \theta\right\}$ for $b \in \{\phi, \psi, \theta\}$. Then, assuming convergence, the Sim EM estimator $\hat{\phi} = (\hat{\theta}', \hat{\psi}')'$ satisfies at convergence $Q_\phi(\hat{\psi}, \hat{\theta}) + V \sum_i (S_{\theta_i}^{\mathrm{obs}}(\hat{\theta}), 0)' = 0$. Let $\tilde{\psi}(\theta)$ be Sim estimate of ψ with θ fixed, so $\hat{\psi} = \hat{\psi}(\hat{\theta})$. If, as we shall argue below, $\hat{\psi}$ is $O_p(1)$ and $\hat{\theta}$ is $n^{\frac{1}{2}}$-consistent for θ_0, by a Taylor expansion around θ_0,

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) = -E\left[S_\theta^{\mathrm{obs}\otimes 2}\right]^{-1}\left\{n^{-\frac{1}{2}}V^{-1}Q_\theta(\hat{\psi}(\theta_0), \theta_0) + n^{-\frac{1}{2}}\sum_i S_{\theta_i}^{\mathrm{obs}} + o_p(1)\right\}$$

since $Q_\theta(\psi, \theta)$ has conditional mean zero given $\{R_i, L_{i(R_i)}; i = 1, \ldots, n\}$ for all (ψ, θ) and thus $E_{\theta_0}\left[Q_\phi(\psi, \theta)\right] = 0$.

In general, $\hat{\psi}$, since not identified, will not be guaranteed to converge to any random variable, that is, $|\hat{\psi} - c|$ will be $O_p(1)$ or greater for all vectors c. However, since $(nV)^{-\frac{1}{2}}Q_\theta(\psi, \theta)$ is $O_p(1)$ for all (ψ, θ) by the Central Limit Theorem, and, under regularity conditions, $Q_\theta(\psi, \theta)$ is Domker[17] as a stochastic process is (ψ, θ) we conclude that:

(a) If $V \to \infty$ as $n \to \infty$, $\hat{\theta}$ will be asymptotically equivalent to the MLE of θ. This reflects the fact that: (i) $n^{-\frac{1}{2}}V^{-1}Q_\theta(\tilde{\psi}\theta_0, \theta_0)$ will be $o_p(1)$ and thus (ii) $n^{\frac{1}{2}}(\hat{\theta} - \theta_0) = -E\left[S_\theta^{\mathrm{obs}\otimes 2}\right]^{-1} n^{-\frac{1}{2}}\sum_i S_{\theta_i}^{\mathrm{obs}} + o_p(1)$.

(b) If V is bounded as $n \to \infty$, then $n^{\frac{1}{2}}(\hat{\theta} - \theta_0)$ will be bounded in probability (that is, $\hat{\theta}$ will be $n^{\frac{1}{2}}$-consistent but, in general, will not have an asymptotic normal distribution). This reflects the fact that $n^{-\frac{1}{2}}V^{-1}Q_\theta(\tilde{\psi}\theta_0, \theta_0)$ will be $O_p(1)$, but its limiting distribution will be non-normal.

REFERENCES

1. Robins, J. M., Rotnitzky, A. and Greenland, S. 'Models for non-monotone missing data', Technical report, Department of Epidemiology, Harvard School of Public Health, 1994.
2. Gill, R., Robins, J. M. 'Missing at andom from an algorithmic viewpoint', *Proceedings of the First Seattle Symposium on Biostatistics: Survival Analysis*, ed. Lin, D. Y., Springer Verlag (1997).
3. Rubin, D. B. 'Inference and missing data' *Biometrika.* **63**, 581–592 (1976).
4. Horvitz, D. G. and Thompson, D. J. 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association*, **47**, 663–685 (1952).
5. Ruud, P. A. 'Extensions of estimation methods using the EM algorithm', *Journal of Econometrics*, **49**, 305–341 (1991).
6. Dempster, A. P., Laird, N. M., and Rubin, D. B. 'Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)' *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38 (1977).
7. Storm, H. H. *et al.* 'Adjuvant radiotherapy and risk of contralateral breast cancer' *Journal of the National Cancer Institute*, **84**, 1245–1250 (1992).
8. Vach, W. and Blettner, M. 'Logistic regression with incompletely observed categorical covariates: Investigating the sensitivity against violation of the missing at random assumption' *Statistics in Medicine*, **14**, 1315–1329 (1995).

 9. Rubin, D. B., Stern, H. S., and Vehovar, V. 'Handling don't know survey responses: The case of the Slovenian plebiscite' *Journal of the American Statistical Association*, **90**, 822–828 (1995).
10. Robins, J. M., Rotnitzky, A. and Zhao, L-P. 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data' *Journal of the American Statistical Association*, **90**, 106–121 (1995).
11. Mark, S. D. and Gail, M. H. 'A comparison of likelihood-based marginal estimating equation methods for analyzing repeated-ordered categorical responses with missing data: Application to an intervention trial of vitamin prophylaxis for esophageal dysplasia', *Statistics in Medicine*. **13**, 479–494 (1994).
12. Newey, W. K. 'Semiparametric efficiency bounds' *Journal of Applied Econometrics*, **5**, 99–135 (1990).
13. Meiljison, I. 'A fast improvement to the EM algorithm on its own terms' *Journal of the Royal Statistical Society*, *Series B*, **51**, 127–138 (1989).
14. Wei, C. G. and Tanner, M. A. 'A Monte Carlo implementation of the EM Algorithm and Poor Man's Data Augmentation Algorithm' *Journal of the American Statistical Association*, **85**, 699–704 (1990).
15. Gill, R., van der Laan, M. and Robins, J. M. 'Coarsening at random: characterizations, conjectures and counter-examples', Proceedings of The First Seattle Symposium on Biostatistics: Survival Analysis, ed. Lin, D. Y., Springer Verlag (1997).
16. Rotnitzky, A. and Robins. J. M. 'Analysis of semi-parametric progression models with non-ignorable non-response' *Statistics in Medicine*, **16**, 81–102 (1997).
17. Bickel, P. J., Klaasen, G. A. J., Ritov, Y. and Wellner, J. A. 'Efficient and adaptive inference for semi-parametric models' Baltimore, MD, Johns Hopkins University Press.