

Large-sample theory for parametric multiple imputation procedures

BY NAISYIN WANG

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
nwang@picard.tamu.edu

AND JAMES M. ROBINS

*Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue,
Boston, Massachusetts 02115, U.S.A.*
robins@epiun1.harvard.edu

SUMMARY

We consider the asymptotic behaviour of various parametric multiple imputation procedures which include but are not restricted to the ‘proper’ imputation procedures proposed by Rubin (1978). The asymptotic variance structure of the resulting estimators is provided. This result is used to compare the relative efficiencies of different imputation procedures. It also provides a basis to understand the behaviour of two Monte Carlo iterative estimators, stochastic EM (Celeux & Diebolt, 1985; Wei & Tanner, 1990) and simulated EM (Ruud, 1991). We further develop properties of these estimators when they stop at iteration K with imputation size m . An application to a measurement error problem is used to illustrate the results.

Some key words: Asymptotic distribution; EM algorithm; Loglikelihood score; Measurement error model; Missing data.

1. INTRODUCTION

In observational studies, data are often missing either by chance or design. A number of statistical procedures have been proposed that first use multiple imputation methods repeatedly to fill in the missing data and then use standard complete data methods to analyse the ‘completed’ data. Different multiple imputation methods have been proposed in the context of different statistical models: parametric (Ruud, 1991; Celeux & Diebolt, 1985), non- or semiparametric (Reilly, 1993), randomisation (Rubin, 1987, Ch. 4; Fay, 1996) and fully Bayesian models (Rubin, 1978, 1987, Ch. 3). They serve different purposes such as the ‘completion’ of public-use data tapes with missing values (Rubin, 1978, 1987, 1996; Fay, 1996; Meng, 1994), or to substitute for a computationally difficult or intractable expectation step in the EM algorithm (Celeux & Diebolt, 1985; Diebolt & Celeux, 1993; Diebolt & Ip, 1996; Tanner, 1993; Wei & Tanner, 1990; Ruud, 1991; McFadden & Ruud, 1994; Robins & Gill, 1997; Deltour, Richardson & Le Hesran, 1998).

The interrelationships between these multiple imputation methods have not been studied. Consequently, the performance of various methods as statistical estimation procedures has not been compared. It is the goal of this paper to provide an asymptotic theory for multiple imputation estimators of the parameter θ of a correctly specified regular para-

metric model. Generalisations of our results to semiparametric or misspecified parametric models and to settings, like those considered in Meng (1994), Rubin (1996) and Fay (1996), in which the parametric model used to produce imputations for public-use data tapes differs from the consumers' analysis model will be the subjects of later reports.

The multiple imputation method proposed by Rubin (1978, 1987), working in a Bayesian context, is one of the most popular among applied scientists who rely on simple complete data procedures for data analysis. Rubin's original goal was to impute m completed data sets for public usage in the context of public surveys in which a response rate of less than 60% for any variable was rare. However, because of its ease of implementation and increasing familiarity, Rubin's method is recently being used outside this context, for example as the primary analytical method for 'two-stage' studies in which response rates are as low as 0.1 to 1% (Greenland & Finkle, 1995). To reduce data storage costs, it may be desirable for the number of imputed datasets, m , to be as small as two or three.

Wei & Tanner (1990) obviated a full Bayesian formulation by suggesting that one imputed missing values from the conditional distribution of the missing data given the observed evaluated at the maximum likelihood estimate. Their method belongs to a class referred to as 'improper' by Rubin (1987, Ch. 3) in contrast to his class of 'proper' procedures. When one's goal is to create public-use data tapes, Rubin states a preference for proper imputation methods based on certain inferential concerns and the availability of a computationally convenient variance estimator that may remain useful even when the parametric model used to produce the imputations differs from the consumer's analysis model. However, when, as in this paper, one's concern is with estimation efficiency in a correctly specified parametric model, we show that the improper procedure of Wei & Tanner is always to be preferred to that of Rubin for finite m , because of its strictly smaller asymptotic variance. The asymptotic properties of these two types of non-iterative multiple imputation procedure and theoretical and numerical comparisons of their efficiencies are given in § 3. We find that the efficiency of the proper estimator relative to the improper may be as low as 60% when m is small and the response rate is low.

In this setting, the inefficiency of Rubin's procedure is even more striking when we turn our attention from point to interval estimation. Specifically, we derive a consistent estimator for the asymptotic variance of the 'improper' estimator and use it to construct a Wald-type interval estimator. In § 3, we show that, even when the efficiency of the proper point estimator relative to the improper is nearly 90%, nonetheless the median length of Rubin's interval estimator is a striking 1.9 times that of the Wald intervals. This poor performance arises because Rubin's variance estimator, although unbiased, is not consistent. As a result, t -intervals, which have a fairly large expected length, are used rather than the standard z -intervals in order to attain nominal coverage rates. Thus, the improper point and Wald interval estimates, though computationally more complicated, are clearly preferable in terms of efficiency when m is small and the fraction of missing data is at least moderate.

An entirely different use for multiple imputation is to substitute for an intractable E-step in the EM algorithm. Both the stochastic (Celeux & Diebolt, 1985) and simulated (Ruud, 1991) EM algorithms are 'iterative' imputation procedures. In general, for the simulated or stochastic EM estimator to be consistent, iteration must continue to convergence or stationarity; see § 4. This may require excessive computation time, especially when m is large of the convergence of the algorithms is mainly decided by the linear convergence rate of EM-like algorithms. However, in settings in which one has an inefficient but easily computed and consistent asymptotically linear initial estimate of θ , one may want to stop

the iteration when the relative efficiency reaches a desired level as each iterate is itself consistent asymptotically normal. For inference, one then only requires the asymptotic distribution of the k th iterate. Therefore, in § 4, we derive the asymptotic variances of the simulated and stochastic EM estimators when stopped after k iterations with m imputations per iteration. We show that the asymptotic variance of the stochastic EM estimator is less than that of the simulated EM and that, for the simulated EM estimator, it may be more efficient to stop after k iterations than to iterate to convergence.

Finally, we analyse a parametric measurement error problem with a validation sample and a dichotomous outcome using the stochastic and simulated EM. This model is a natural one for our approach since an inefficient complete-case estimator based on the validation sample is easily computed, the E-step of the EM algorithm is computationally burdensome, but it is easy to draw from the conditional law of the missing data given the observed data by rejection sampling.

2. THE MODEL

We shall study the following statistical model. The complete data $Y = (Y_1, \dots, Y_p)'$, of dimension $p \times 1$, is randomly drawn from a population whose density is $f(Y; \theta_0)$, a regular parametric family, $\{f(Y; \theta); \theta \in \Theta \subset R^q\}$, with respect to a dominating measure μ , where θ_0 is an unknown parameter to be estimated. Let R_k be the indicator of whether or not the k th component of Y was observed and let $R = (R_1, \dots, R_p)'$. We also denote the observed and unobserved components of Y by Y_R and $Y_{\bar{R}}$, respectively. Rubin refers to Y_R and $Y_{\bar{R}}$ as Y_{obs} and Y_{mis} . Throughout, we assume that the data are missing at random; that is, the probability that $R = r$ given Y does not depend on the unobserved component $Y_{\bar{R}}$ of Y . We observe n independent and identically distributed realisations $Z = \{R^i, Y_R^i; i = 1, \dots, n\}$ of (R, Y_R) . To avoid extraneous issues, we assume

- (i) that the complete-data maximum likelihood estimator, $\hat{\theta}_c$, and the observed-data maximum likelihood estimator, $\hat{\theta}_{\text{MLE}}$, are, respectively, the unique solutions to the complete-data score equation $\sum S_i(\theta) = 0$ and the observed-data score equation $\sum S_i^{\text{obs}}(\theta) = 0$, where $S(\theta) = S(Y; \theta) = \partial \log f(Y; \theta) / \partial \theta$ is the loglikelihood score if the data were completely observed and $S^{\text{obs}}(\theta) = E_{\theta} \{S(\theta) | Y_R\}$ is the score function of the observed-data loglikelihood; and
- (ii) $\hat{\theta}_c$ and $\hat{\theta}_{\text{MLE}}$ are consistent asymptotically linear estimators of θ_0 with influence functions $I_c^{-1} S(\theta_0)$ and $I_{\text{obs}}^{-1} S^{\text{obs}}(\theta_0)$, where $I_c = E_{\theta_0} \{S(\theta_0)^{\otimes 2}\}$ and $I_{\text{obs}} = E_{\theta_0} \{S^{\text{obs}}(\theta_0)^{\otimes 2}\}$ are the full- and observed-data information matrices and $A^{\otimes 2} = AA'$.

An estimator $\hat{\theta}$ of θ_0 is consistent asymptotically linear with influence function D if $n^{\frac{1}{2}}(\hat{\theta} - \theta_0) = n^{-\frac{1}{2}} \sum D_i + o_p(1)$, where D has a zero mean and a finite covariance matrix and $o_p(1)$ denotes a random variable converging to zero in probability. Consequently, $n^{\frac{1}{2}}(\hat{\theta} - \theta_0)$ is asymptotically normal with mean zero and variance $E[D^{\otimes 2}]$. For example, $\hat{\theta}_c$ and $\hat{\theta}_{\text{MLE}}$ have asymptotic variances I_c^{-1} and I_{obs}^{-1} , respectively.

3. ASYMPTOTIC PROPERTIES OF NON-ITERATIVE MULTIPLE IMPUTATION ESTIMATORS

3.1. Definition of the estimators

We shall study two types of non-iterative multiple imputation estimator, type A and type B. In Rubin's nomenclature, type A estimators use a proper and type B estimators use an improper imputation method. For $i = 1, \dots, n$, let $Y_{\bar{R}}^{ij}(\hat{\theta}_j)$ ($j = 1, \dots, m$) be m

imputed copies of subject i 's missing data; each $Y_{R}^{ij}(\tilde{\theta}_j)$ is drawn independently from $f(Y_{R}|Y_{R}^i; \tilde{\theta}_j)$. For type B, $\tilde{\theta}_j = \hat{\theta}_p$ is a preliminary consistent asymptotically linear estimator of θ_0 computed from the observed data Z . For type A, $\tilde{\theta}_j$ is a single independent draw from the posterior density $f(\theta|Z)$ of θ under a Bayesian model. Let $S_{ij}(\theta, \tilde{\theta}_j) = S[\{Y_{R}^i, Y_{R}^{ij}(\tilde{\theta}_j)\}; \theta]$ be the completed data score contributed by subject i in the j th completed dataset. Then the j th completed dataset maximum likelihood estimator of θ_0 , $\hat{\theta}_j \equiv \hat{\theta}_j(\tilde{\theta}_j)$, solves $n^{-1} \sum_i S_{ij}(\theta, \tilde{\theta}_j) = 0$. Rubin suggests a final inference based on $\bar{\theta} = \sum \hat{\theta}_j/m$. As an alternative to solving m separate score equations to obtain $\hat{\theta}_1, \dots, \hat{\theta}_m$ and then averaging, one could use the estimator $\hat{\theta}$ that solves the single 'clustered data' estimating equation $0 = m^{-1} \sum \sum S_{ij}(\theta, \tilde{\theta}_j)$ (Fay, 1996). The following lemma, which follows from simple algebraic arguments and Slutsky's theorem (Billingsley, 1968, § 1.5), implies that the two estimators have the same asymptotic distributions. We shall therefore treat the estimators $\hat{\theta}$ and $\bar{\theta}$ interchangeably in the sequel.

LEMMA 1. *Estimators $\hat{\theta}$ and $\bar{\theta}$ are asymptotically equivalent, that is, $n^{1/2}(\hat{\theta} - \bar{\theta})$ converges to zero in probability.*

3.2. Asymptotic distribution of the type B estimator

Write $\hat{\theta}_B$ to denote a type B estimator $\hat{\theta}$ based on a preliminary asymptotically linear estimator $\hat{\theta}_p$ with asymptotic variance B . The following theorem, proved in the Appendix, gives the asymptotic distribution of $\hat{\theta}_B$. Let $I_{\text{mis}} \equiv I_c - I_{\text{obs}}$ be the missing information matrix. Then $J = I_{\text{mis}} I_c^{-1}$ is the 'fraction' of missing information matrix (Rubin, 1987, p. ■■■). Let $\Delta(B) = B - I_{\text{obs}}^{-1}$ be the nonnegative definite difference between the asymptotic variance matrix of $\hat{\theta}_p$ and $\hat{\theta}_{\text{MLE}}$.

THEOREM 1. *Under the regularity conditions in the Appendix, $n^{1/2} \{\hat{\theta}_B - \theta_0\}$ is asymptotically normal with mean zero and variance*

$$V_B \equiv I_{\text{obs}}^{-1} + J' \Delta(B) J + m^{-1} I_c^{-1} J. \quad (1)$$

COROLLARY 1. *If $\hat{\theta}_p = \hat{\theta}_{\text{MLE}}$, the asymptotic variance of $\hat{\theta}_B$ is $V_B = I_{\text{obs}}^{-1} + m^{-1} I_c^{-1} J$.*

Theorem 1 and its corollary show that the asymptotic variance I_{obs}^{-1} of $\hat{\theta}_{\text{MLE}}$ differs from that of $\hat{\theta}_B$ by two positive semidefinite terms. The third term, $m^{-1} I_c^{-1} J$, in (1) is attributable to the additional variability resulting from the finite number of imputations m . This term, which does not depend on the initial estimator $\hat{\theta}_p$, increases with the fraction of missing information J and goes to zero as $m \rightarrow \infty$. The second term, $J' \Delta(B) J$, is attributable to the inefficiency of $\hat{\theta}_p$, and does not depend on the number of imputations m . Since, when $\hat{\theta}_p$ is inefficient, $B - \{I_{\text{obs}}^{-1} + J' \Delta(B) J\}$ is positive definite, $\hat{\theta}_B$ will always be more efficient than $\hat{\theta}_p$ when $m = \infty$. In fact, with $m = \infty$, $\hat{\theta}_B$ is exactly a one-step update of $\hat{\theta}_p$ using the EM algorithm. When m is finite and $\hat{\theta}_p$ is inefficient, $\hat{\theta}_B$ may or may not be more efficient than $\hat{\theta}_p$ depending on the relative sizes of the terms in (1).

3.3. Asymptotic distribution of type A estimators

Assume the standard regularity conditions hold for the model and the prior which guarantee that the posterior distribution of θ given the observed data Z is asymptotically normal with mean $\hat{\theta}_{\text{MLE}}$ and variance matrix $\{-\sum \partial S_i^{\text{obs}}(\hat{\theta}_{\text{MLE}})/\partial \theta'\}^{-1}$ almost surely on Z . The validity of z and t inference procedures proposed by Rubin (1987, Ch. 3) require this assumption. The following theorem then characterises the asymptotic distribution of the type A multiple imputation estimator $\hat{\theta}_A$.

THEOREM 2. Under the regularity conditions in the Appendix, $n^{\frac{1}{2}}(\hat{\theta}_A - \theta_0)$ is asymptotically normal with mean zero and variance

$$V_A = I_{\text{obs}}^{-1} + m^{-1} J_c^{-1} J + m^{-1} J' I_{\text{obs}}^{-1} J. \quad (2)$$

The proof of Theorem 2 is outlined in the Appendix. It follows that the inflation, $m^{-1} J' I_{\text{obs}}^{-1} J$, of the type A asymptotic variance over that of a type B estimator with the preliminary estimate $\hat{\theta}_P = \hat{\theta}_{\text{MLE}}$, goes to zero as $m \rightarrow \infty$ and increases with the fraction of missing information. This suggests that, when m is small and the fraction of missing data large, the type B, improper estimator could be significantly more efficient than the type A, proper estimator. We use Example (2.1) of Rubin (1987) to illustrate this result.

Example 1. Suppose Y is $N(\mu, \sigma^2)$ when $\theta = (\mu, \sigma^2)'$. Let $\pi = \text{pr}(R = 1|Y) = \text{pr}(R = 1)$ be the probability that Y is observed. Then $\hat{\theta}'_{\text{MLE}} = (\hat{\mu}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}}^2)$ is the sample mean and variance of Y calculated from the N_{obs} of the n subjects with Y observed. The type B estimator with $\hat{\theta}_P = \hat{\theta}_{\text{MLE}}$ imputes independent replications for each subject with missing data from a $N(\hat{\mu}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}}^2)$ distribution. For the type A estimator, Rubin shows that, with standard non-informative priors, the j th imputed value of Y^{ij} for a subject i with missing data is obtained as follows:

- (i) draw a random variable v_{1j} , from a χ^2 distribution with $N_{\text{obs}} - 1$ degrees of freedom, and let $\hat{\sigma}_j^2 = \hat{\sigma}_{\text{MLE}}^2 (N_{\text{obs}} - 1) / v_{1j}$;
- (ii) draw $\tilde{\mu}_j$ from a $N(\hat{\mu}_{\text{MLE}}, N_{\text{obs}}^{-1} \hat{\sigma}_j^2)$ distribution;
- (iii) draw Y^{ij} from a $N(\tilde{\mu}_j, \hat{\sigma}_j^2)$ distribution.

For various choices of π and m , we evaluated the relative efficiency of the type A estimator compared to the type B. The results are summarised in Fig. 1. The relative efficiency drops to nearly 60% when $m=2$ and $\pi=0.1$, but increases rapidly when either π or m is increased.

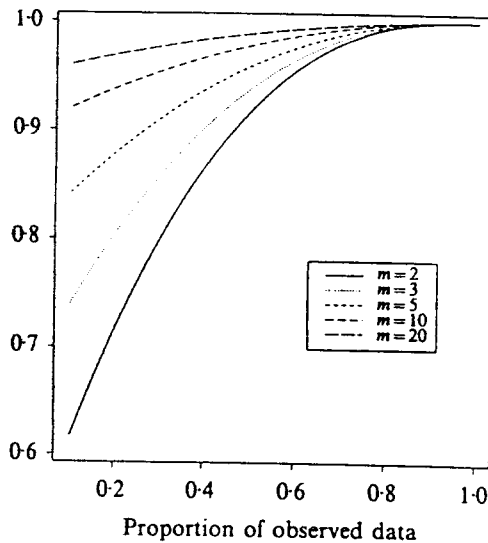


Fig. 1. Example 1. The relative efficiency of the type A estimator with respect to the corresponding type B estimator versus the proportion of observed data, π . Five curves, from bottom to top, correspond to imputation size $m = 2, 3, 5, 10$ and 20 , respectively.

3.4. Variance estimation

The asymptotic variances V_B and V_A of $\hat{\theta}_B$ and $\hat{\theta}_A$ depend on I_{obs} , I_c and the asymptotic variance B of $\hat{\theta}_P$. Usually, as in the measurement error example in § 4.3, a consistent estimator \hat{B} of B can be easily computed. It therefore remains to find consistent estimators of I_c and I_{obs} . A consistent estimator for I_c is $\hat{I}_c = m^{-1} \sum \hat{I}_{c,j}$, where $\hat{I}_{c,j} = -n^{-1} \sum \partial S_{ij}(\theta, \tilde{\theta}) / \partial \theta' |_{\theta = \hat{\theta}}$ is the usual observed information for the j th completed dataset given by off-the-shelf software packages. A consistent estimator for I_{obs} is motivated by the following lemma by Robins & Gill (1997), which states that the expected outer product of the completed data scores for two different completed datasets equals I_{obs} .

LEMMA 2. For $j \neq j^*$, $E\{S_{ij}(\theta_0, \theta_0) S'_{ij^*}(\theta_0, \theta_0)\} = E\{S_{\text{obs}}^{\otimes 2}(\theta_0)\} \equiv I_{\text{obs}}$.

A symmetrised positive definite consistent estimator of I_{obs} is

$$\hat{I}_{\text{obs}} = \frac{1}{2nm(m-1)} \sum_i \sum_{j \neq j^*} \{S_{ij}(\hat{\theta}, \tilde{\theta}_j) S'_{ij^*}(\hat{\theta}, \tilde{\theta}_{j^*})' + S_{ij^*}(\hat{\theta}, \tilde{\theta}_{j^*}) S'_{ij}(\hat{\theta}, \tilde{\theta}_j)'\}.$$

Let \hat{V}_B and \hat{V}_A be V_B and V_A in (1) and (2), with \hat{I}_c , \hat{I}_{obs} , $\hat{J} = I - \hat{I}_{\text{obs}} \hat{I}_c^{-1}$ and \hat{B} replacing I_c , I_{obs} , J and B , where I denotes the $q \times q$ identity matrix. Then \hat{V}_B and \hat{V}_A are consistent for V_B and V_A . Furthermore, by Slutsky's theorem, $n^{1/2} V_t^{-1/2}(\hat{\theta}_t - \theta_0)$, t in $\{A, B\}$, are asymptotically standard normal. Thus, Wald-type inferences can be directly applied.

3.5. Comparison with inference based on Rubin's variance estimators

Rubin (1987, Ch. 3, 4) proposed to use \tilde{V}_A , which does not require explicit estimation of I_{obs} , to estimate V_A , where

$$\tilde{V}_A = \hat{I}_c^{-1} + (1 + m^{-1}) \tilde{Q}_A \quad \text{and} \quad \tilde{Q}_A = n(m-1)^{-1} \sum (\hat{\theta}_{Aj} - \bar{\theta}_A)^{\otimes 2}.$$

Recall that $\hat{\theta}_{Aj}$ is the estimate based on the j th completed dataset using the type A imputation. In the above formula for \tilde{Q}_A , the empirical between-imputation variance of $\hat{\theta}_{Aj}$ is multiplied by the sample size n to make \tilde{Q}_A , the empirical between-imputation variance of $\hat{\theta}_{Aj}$ is multiplied by the sample size n to make \tilde{Q}_A an $O_p(1)$ random variable, where a random variable is $O_p(1)$ if it is bounded in probability.

As in Rubin (1987, Ch. 3, 4), we concentrate on the cases where $\text{var}\{n^{1/2}(\hat{\theta}_{Aj} - \theta_0) | Z\}$ is finite with probability one; when this condition fails, \tilde{V}_A might need to be robustified in some suitable fashion to control the influence of the heavy tails of $\hat{\theta}_{Aj}$. In §§ 4.1 and 4.2 below, we will use the following identities, proved in the Appendix,

$$V_A = I_c^{-1} + (1 + m^{-1})(I_{\text{obs}}^{-1} - I_c^{-1}), \quad (3)$$

$$E\{E(\tilde{Q}_A | Z)\} = I_{\text{obs}}^{-1} - I_c^{-1} + o(1), \quad (4)$$

to study the properties of \tilde{V}_A in the m -infinite and m -finite settings. Throughout §§ 4.1 and 4.2 references to $\hat{\theta}_B$ or V_B refer to the case in which $\hat{\theta}_P = \hat{\theta}_{\text{MLE}}$. Also, for any two matrices M_A and M_B , $M_A \geq M_B$ indicates that $M_A - M_B$ is positive semidefinite, whereas the strict inequality implies that $M_A - M_B$ is positive definite.

Case I (infinite m). If both m and $n \rightarrow \infty$, then (i) $\hat{\theta}_A$ and $\hat{\theta}_B$ will have the same limiting distribution, and (ii), by (3) and (4), \tilde{V}_A will consistently estimate $V_A = V_B = I_{\text{obs}}^{-1}$, since \tilde{Q}_A will be consistent for $I_{\text{obs}}^{-1} - I_c^{-1}$. Thus, by Slutsky's theorem, the Wald intervals using \tilde{V}_A will cover θ_0 at their nominal confidence level for large samples. However, as noted by Rubin (1987), such intervals with $\hat{\theta}_{Bj}$ substituted for $\hat{\theta}_{Aj}$ will under-cover. Define \tilde{V}_B and

\bar{Q}_B by analogy with the type A estimators; it is easy to see that \bar{Q}_B converges to $I_c^{-1}I_{\text{mis}}I_c^{-1}$, which is smaller than or equal to $I_{\text{obs}}^{-1} - I_c^{-1}$. This result explains why, from a frequentist perspective, Rubin (1987, Ch. 4) suggested using a type A, proper, rather than a type B, improper estimator even though $\hat{\theta}_A$ and $\hat{\theta}_B$ have the same limiting distribution. Note that, when m goes to infinity, both \hat{V}_A and \hat{V}_B of § 3.4 converge in probability to the same and correct limit. The comparison between two types of variance estimator becomes much more important when we consider the finite m case where $V_A \neq V_B$.

Case II (finite m). When m is finite, as noted by Rubin (1987), \bar{V}_A has a nondegenerate limiting distribution and thus does not converge to the constant $I_{\text{obs}}^{-1} - I_c^{-1}$. Nonetheless, the asymptotic means of \bar{Q}_A and \bar{V}_A equal $I_{\text{obs}}^{-1} - I_c^{-1}$ and V_A . Hence \bar{V}_A remains a sensible estimator of V_A . Since the asymptotic normality for $n^{\frac{1}{2}}(\bar{V}_A^{-\frac{1}{2}}(\hat{\theta}_A - \theta_0))$ no longer holds, Rubin suggests replacing a z -critical value by a t -critical value with an approximated degree of freedom (Rubin, 1987, p. 77). However, these t intervals will cover θ_0 at their nominal rate at the cost of a rather large expected confidence interval length due to the heavy t tails. This unfortunate trade-off can be bypassed by replacing \bar{V}_A by \hat{V}_A since it is consistent for V_A . An even better strategy is to use the type B, improper, Wald interval based on the estimator $\hat{\theta}_B$ and the estimated variance \hat{V}_B , which, for finite m , will (i) have a shorter expected length than the corresponding A intervals, since $V_B < V_A$, and (ii) still cover θ_0 at the nominal confidence level in large samples. This point is illustrated in the following simulation study.

Example 1 (cont.). In the setting of Example 1, we conducted a small simulation study with sample size $n = 50$, the number of imputations $m = 3$ and the response rate, π , equal to 0.9, 0.5 and 0.25, respectively. One thousand datasets were generated for each of the three scenarios. In Table 1, we report the Monte Carlo means and variances of $\hat{\theta}_A$ and $\hat{\theta}_B$, with the preliminary estimator $\hat{\theta}_P$ being the observed-data maximum likelihood estimator, that is, the sample mean of the observed data. The relative efficiency of $\hat{\theta}_A$ compared to $\hat{\theta}_B$ is never less than $0.86 = 0.8784/1.093$ since π was not very small. We also report the actual coverage rates of nominal 95% confidence intervals. For the type B estimator $\hat{\theta}_B$, we report the Wald z -interval based on \hat{V}_B . For $\hat{\theta}_A$, we report (i) the z -interval based on \bar{V}_A , (ii) the t -interval based on \bar{V}_A , and (iii) the Wald z -interval based on \hat{V}_A . We also report the Monte Carlo median of the ratios of the lengths of the type A intervals to those of the type B intervals. As predicted by the theory, when π is small, the z -interval based on \bar{V}_A under-covers and the t -interval has median length 1.87 times that of the type B z -interval and $1.8 = 1.87/1.069$ times that of the type A z -interval based on \hat{V}_A .

In practice when, as in Example 1, an efficient and easily computed estimator of θ is available, there is no reason to use a multiple imputation estimator. Rather, the purpose of Example 1 was to compare, in a simple tractable setting, the performance of alternative imputation procedures. Fitting parametric models by multiple imputation methods will be of practical importance for those models with a computationally difficult or intractable expectation step in the EM algorithm; see § 4 for examples.

4. ITERATIVE MULTIPLE IMPUTATION ESTIMATORS

4.1. *The stochastic EM and simulated EM algorithms*

Iterative versions of the type B estimator $\hat{\theta}_B$ have been proposed to substitute for a computationally difficult or intractable E-step in the EM algorithm. Both the stochastic EM algorithm (Celeux & Diebolt, 1985; Tanner, 1993) and the simulated EM algorithm (Ruud,

Table 1. *Example 1 (cont.). Results based on 1000 simulations, $m = 3$, summary statistics of estimated μ for a normal mean problem with the true $\mu = 0$. Entries reading from left to right in each row are sample mean $\times 10$, sample variance $\times 10$, coverage probability for 95% confidence intervals and median ratio of the length of each confidence interval over that of type B z-interval. Results for three different type A intervals are reported, (i) z-interval using \tilde{V}_A , (ii) t-interval using \tilde{V}_A and (iii) z-interval using \hat{V}_A ; first and second entries inside parentheses are for (ii) and (iii), respectively*

	Sample mean	Sample var.	Cover prob.	Med. length ratio
			$\pi = 0.9$	
Type B	0.1571	0.2207	0.961	1.00
Type A	0.1713	0.2250	0.933 (0.950, 0.958)	0.929 (0.992, 1.002)
			$\pi = 0.5$	
Type B	0.3411	0.4160	0.965	1.00
Type A	0.2472	0.4654	0.887 (0.943, 0.966)	0.776 (1.516, 1.033)
			$\pi = 0.25$	
Type B	0.3552	0.8784	0.978	1.00
Type A	-0.3206	1.0930	0.889 (0.980, 0.973)	0.852 (1.871, 1.069)

1991) start from an initial $\hat{\theta}_p$ and compute $\hat{\theta}_B^{(1)} = \hat{\theta}_B$ and then iterate by regarding the current estimate $\hat{\theta}_B^{(k)}$ as $\hat{\theta}_p$ and updating to $\hat{\theta}_B^{(k+1)} = \hat{\theta}_B$. In the simulated EM algorithm, the same pseudo-random numbers drawn in the first iteration are reused to draw the imputations in subsequent iterations. In the stochastic EM algorithm, an independent set of pseudo-random numbers is used in each iteration. As a consequence, the stochastic EM algorithm is computationally more intensive than the simulated EM algorithm, especially if it is computationally demanding to draw from $f\{Y_R^i | Y_R^i; \hat{\theta}_B^{(k-1)}\}$. Ruud (1991) did not require the starting value $\hat{\theta}_p$ to be a consistent, asymptotically linear estimator of θ_0 . Consequently, to ensure consistency he required the simulated EM algorithm to be iterated until convergence. McFadden & Ruud (1994) prove that, under regularity conditions, as $K \rightarrow \infty$, the iterative simulated EM estimates, $\hat{\theta}_{\text{sim},K} = \hat{\theta}_B^{(K)}$ converge to a consistent, asymptotically linear limit, $\hat{\theta}_{\text{sim}}$, with the variance

$$V_{\text{min}} = I_{\text{obs}}^{-1} + m^{-1} I_{\text{obs}}^{-1} I_{\text{mis}}^{-1} I_{\text{obs}}^{-1}. \quad (5)$$

Although the iterates $\hat{\theta}_{\text{sto},k}$ of the stochastic EM algorithm do not converge point-wise for finite m , nonetheless they do converge to a stationary distribution under regularity conditions (Diebolt & Celeux, 1993). These authors further show that, as $K \rightarrow \infty$, $\bar{\theta}_{\text{sto},K}$, the average of the first K iterates after the algorithm converges, goes to a consistent, asymptotically linear limit $\bar{\theta}_{\text{sto}}$ which is efficient for θ_0 . However, the computation time necessary to reach convergence or stationarity could be excessive for both algorithms, especially when m and the fraction of missing data are large.

4.2. Properties of $\hat{\theta}_{\text{sto},K}$ and $\hat{\theta}_{\text{sim},K}$ for finite m and K

If, as in the measurement error example of § 4.3, one has an initial inefficient consistent asymptotically linear estimator of θ_0 , then each iterate is itself consistent asymptotically linear so one needs not iterate until convergence or stationarity. Note that, even though these estimators may not be efficient especially when m is small, one can easily obtain the

estimated relative efficiency at each iteration. If, as in the example in § 4.3, after several iterations the asymptotic relative efficiency of the iterate is already near 1, then one may choose to stop the iteration at that point. The following theorems provide the asymptotic distribution of $\hat{\theta}_{\text{sto},K}$ and $\hat{\theta}_{\text{sim},K}$ when $\hat{\theta}_p$ is a consistent asymptotically linear estimate of θ_0 with asymptotic variance B .

THEOREM 3. *Under the regularity conditions in the Appendix, $n^{\frac{1}{2}}(\hat{\theta}_{\text{sim},K} - \theta_0)$ ($K \geq 1$) is asymptotically normal with mean 0 and variance*

$$V_{\text{sim},K} = I_{\text{obs}}^{-1} + \Sigma_p^{(K)} + \Sigma_{\text{imp,sim}}^{(K)}$$

where $\Sigma_p^{(K)} = (J^K)' \Delta(B) J^K$ and

$$\Sigma_{\text{imp,sim}}^{(K)} = m^{-1} I_c^{-1} \left(\sum_{k=1}^K J^{K-k} \right) I_{\text{mis}} \left(\sum_{k=1}^K J^{K-k} \right)' I_c^{-1}. \quad (6)$$

THEOREM 4. *Under regularity conditions, $n^{\frac{1}{2}}(\hat{\theta}_{\text{sto},K} - \theta_0)$ ($K \geq 1$) is asymptotically normal with mean 0 and asymptotic variance $V_{\text{sto},K} = I_{\text{obs}}^{-1} + \Sigma_p^{(K)} + \Sigma_{\text{imp,sto}}^{(K)}$, where*

$$\Sigma_{\text{imp,sto}}^{(K)} = \sum_{k=1}^K m_k^{-1} I_c^{-1} J^{K-k} I_{\text{mis}} (J^{K-k})' I_c^{-1}, \quad (7)$$

and where m_k ($k = 1, \dots, K$) is the number of imputations used in the k th iteration.

Both theorems can be obtained through simple algebraic derivations following the asymptotic expansions for $\hat{\theta}_{\text{sto},K}$ and $\hat{\theta}_{\text{sim},K}$ given in (A5) and (A6), respectively. In the Appendix we prove the following corollary.

COROLLARY 2. *Suppose for the stochastic EM algorithm $m_k = m$ for $k = 1, \dots, K$. Then $V_{\text{sim},K} - V_{\text{sto},K}$ is positive semidefinite. Further, as $K \rightarrow \infty$,*

- (i) $V_{\text{sim},K} \rightarrow V_{\text{sim}}$ given by (5), and
- (ii) $V_{\text{sto},K} \rightarrow V_{\text{sto}} \equiv I_{\text{obs}}^{-1} + m^{-1} I_c^{-1} I_{\text{mis}} I_c^{-1} (I - J^2)^{-1}$.

The two theorems and Corollary 2 above provide insight about the behaviour of $\hat{\theta}_{\text{sim},K}$ and $\hat{\theta}_{\text{sto},K}$. We summarise them in the following remarks. To simplify the notation in the presentation, unless otherwise we consider $m_k \equiv m$ in $\hat{\theta}_{\text{sto},K}$.

Remark 1. The matrices $\Sigma_{\text{imp}}^{(K)}$ and $\Sigma_p^{(K)}$ correspond to the extra variation caused by the imputations and by the inefficient preliminary estimator respectively. We prove in Lemma A1 in the Appendix that the matrix norm $\|J^K\|$ goes to zero as K increases. As a consequence, as indicated in Corollary 2, the effect of the initial inefficiency in $\hat{\theta}_p$ is eliminated as $K \rightarrow \infty$. On the other hand, $\Sigma_{\text{imp,sim}}^{(K)}$ and $\Sigma_{\text{imp,sto}}^{(K)}$ both increase with K and decrease with m .

Remark 2. By Corollary 2, it is obvious that V_{sim} and V_{sto} need not be smaller than the asymptotic variance B of the initial estimator $\hat{\theta}_p$. An extreme example is when $\hat{\theta}_p = \hat{\theta}_{\text{MLE}}$ and $B = I_{\text{obs}}^{-1}$. However, it can be easily shown that, for the stochastic EM algorithm, if, for the given imputation size, m , and the initial variance B , there is an improvement in efficiency after the first iteration, then there is an improvement in each iteration; that is, if $V_{\text{sto},1} - B$ is non-positive definite then so is $V_{\text{sto},k} - V_{\text{sto},k-1}$. This result does not hold for the simulated EM algorithm, however, where the optimal estimates in the class may not be attained in the limit as $K \rightarrow \infty$.

Remark 3. Define $\bar{\theta}_{\text{sto},K}$ to be the average of the first K iterates in a stochastic EM. We

obtain the asymptotic expression of $n^{\frac{1}{2}}(\bar{\theta}_{\text{sto},K} - \theta_0)$ in (A7) when we start with a consistent asymptotic linear $\hat{\theta}_p$. As pointed out in the Appendix, the asymptotic variance of $n^{\frac{1}{2}}(\bar{\theta}_{\text{sto},K} - \theta_0) \rightarrow I_{\text{obs}}^{-1}$ as n and $K \rightarrow \infty$. Equivalent results have been obtained earlier by Diebolt & Celeux (1993) for $m = 1$. For finite K , there is no guarantee that $\bar{\theta}_{\text{sto},K}$ is more efficient than $\hat{\theta}_{\text{sto},K}$. For both the stochastic and simulated EM algorithms, even greater efficiency can, in principle, be obtained by taking 'optimal' linear combinations of the first K iterates rather than using the K th iterate or the unweighted average.

Remark 4. Tanner (1993, p. ■■■) suggested starting a stochastic EM algorithm with a small number of imputations, and then increasing the imputation size at a later stage. Our Theorem 4 supports this suggestion by noting that, in $\Sigma_{\text{imp,sto}}^{(K)}$ the contribution of the k th iteration is relatively small for small k since $\|J^{K-k}\|$ is small. Therefore, for a fixed number of iterations, to improve the efficiency without performing extra imputations one should allocate more imputations to the later iterations. How to allocate the imputation resources efficiently is currently under investigation.

4.3. A measurement error example

To demonstrate properties of the stochastic and simulated EM, we study a simple logistic normal measurement error example, in which the true covariate X is $N(\mu_x, \sigma_x^2)$ distributed; the surrogate W equals $X + U$ with U distributed as $N(0, \sigma_u^2)$, independent of X ; the dichotomous response D follows a logistic model with $E(D|X) = H(\beta_0 + \beta_1 X)$ and $H(v) = \{1 + \exp(-v)\}^{-1}$; D and W are always observed. Subjects are randomly selected into a validation sample with probability π . The variable X is only observed in the validation sample. For more details about measurement error problems with this data structure, see Carroll, Ruppert & Stefanski (1995, Ch. 13). The parameter $\theta = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2)'$ is $(0, 1, 0, 1, 0.5)$, implying a reliability ratio of 66.7%. By using the results in § 4.2, we calculated the asymptotic relative efficiency of $\hat{\beta}_1^{(K)}$ with respect to the observed-data maximum likelihood estimator when the initial estimate was the maximum likelihood estimator calculated from the completely observed data in the validation sample. The results are shown in Fig. 2. Note that, if π is a known function of the always observed data, $Z = (D, W)$, and it is bounded away from zero, the Horvitz-Thompson estimate (Horvitz & Thompson, 1952) calculated from the validation sample data could be used as the initial estimate. In Fig. 2, we plot the asymptotic relative efficiency of $\hat{\beta}_1^{(K)}$ versus the number of iterations, K . The three curves in each plot, from bottom to top, respectively, correspond to the stochastic EM with $m = 5$, solid curve, the simulated EM with $m = 10$, dotted curve, and the stochastic EM with $m = 10$, dashed curve. Figures 2(a), (c) and (d) correspond to the cases where $\pi = 0.25, 0.5$ and 0.1 , respectively. For this example, our calculations indicate the following: (i) most of the improvement in efficiency is gained in the early iterations; (ii) even with a small number of imputations, $m = 5$ or 10 , and a relatively small validation proportion, $\pi = 0.1$, the asymptotic relative efficiencies of both estimator are still quite good, that is, above 90%; (iii) although not quite visible in Fig. 2, the maximum asymptotic relative efficiency for the simulated EM occurred from the third to the fifth iterations, but nonetheless there is little difference between these maxima and the asymptotic relative efficiencies at the later iterations. We also conducted a small simulation study for $n = 250$, $m = 5$ and 10 , $\pi = 0.25$, based on 1000 iterations. Estimated asymptotic relative efficiencies, with the calculated variances now replaced by the simulation sample variances, are presented in Fig. 2(b). The simulation result is close to what is suggested by the theory, except that the estimated asymptotic relative efficiency of the

simulated EM is slightly lower than predicted by our asymptotic theory. Using the completely observed cases only results in a preliminary estimate with 39% estimated asymptotic relative efficiency. The improvement resulting from even a small number of iterations is obvious.

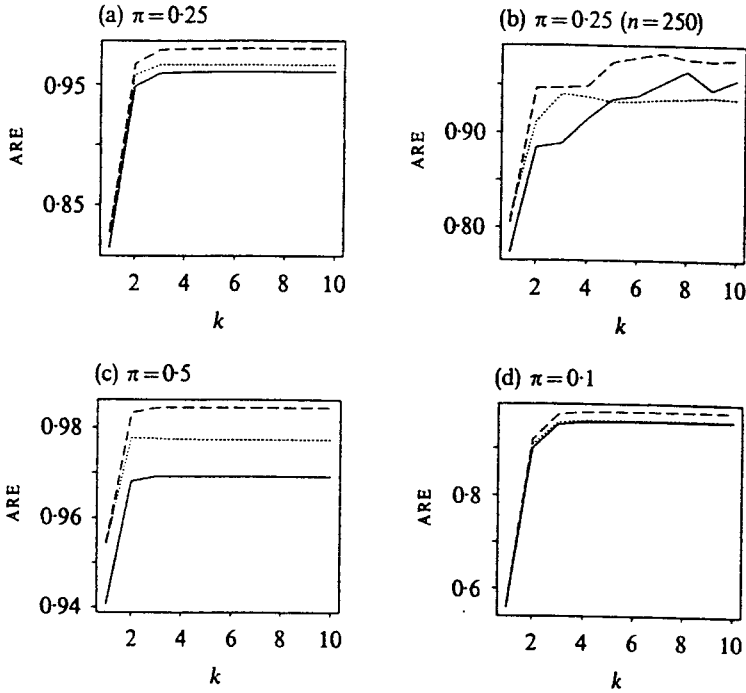


Fig. 2. Measurement error example. Asymptotic relative efficiencies, ARE, of $\hat{\beta}_1^{(k)}$ with respect to the observed-data maximum likelihood estimator versus the number of iterations k . The solid curves and the dashed curves in all plots correspond to the stochastic EM estimates with $m=5$ and 10, respectively; the dotted curves correspond to the simulated EM estimates with $m=10$. (a) $\pi=0.25$; (b) as in (a), but replacing the variances in the asymptotic relative efficiency with the sample variances of the estimates in a simulation study, with $n=250$; (c) and (d), as in (a) but with $\pi=0.5$ and 0.1, respectively.

ACKNOWLEDGEMENT

This research was supported by grants from the National Science Foundation, National Institutes of Health and the Texas Advanced Research program. We thank the editor and two referees, whose comments greatly improved the presentation of the paper.

APPENDIX

Assumptions and proofs

We consider a regular parametric family $\{f(Y; \theta): \theta \in \Theta\}$, where Θ is in a finite dimensional Euclidean space. Besides the regularity assumptions we mentioned in the text, we further assume that Assumptions (S1) and (S2) hold for θ in a neighbourhood of θ_0 throughout the proofs.

Assumption (S1). The partial derivative $\partial \log f(Y_R | Y_R, \theta) / \partial \theta'$ exists and is bounded in L^2 .

Assumption (S2). Let $\lambda(\theta, \eta)$ be $E_\theta[E\{S(Y_R, Y_{\bar{R}}, \theta)|Y_R, \eta\}]$, where $E(\cdot|Y_R, \eta)$ is the conditional mean given the observed data when the conditional distribution of $Y_{\bar{R}}$ is $f(\cdot|Y_R, \eta)$, and let

$$\mathcal{Z}_{n,\theta}(\tau, \eta) = n^{-\frac{1}{2}} \left| \sum S(\theta, \eta) - \sum S(\theta, \tau) - \lambda(\theta, \eta) + \lambda(\theta, \tau) \right|,$$

where $S(\theta, \eta)$ is $S\{Y_R, Y_{\bar{R}}(\eta); \theta\}$ defined in § 3. We assume that there exists a positive d such that, for θ, η and τ in a neighbourhood of θ_0 , $\sup_{|\eta-\tau|<d} \mathcal{Z}_{n,\theta}(\tau, \eta) \rightarrow 0$ uniformly in θ as n goes to ∞ . We also assume that $(\partial/\partial\eta)\lambda(\theta, \eta)$ exists.

Note that Assumption (S2) provides a continuity condition which is equivalent to what is assumed in § 4 of Huber (1967). We will now provide sketches of the following proofs.

Proof of Theorem 1. Define $S_i(\theta, \tilde{\theta}) = m^{-1} \sum_j S_{ij}(\theta, \tilde{\theta})$. Then $\hat{\theta}_B$ solves $n^{-\frac{1}{2}} \sum S_i(\theta, \hat{\theta}_B) = 0$. Let ψ be the influence function of $\hat{\theta}_B$, that is, $n^{\frac{1}{2}}(\hat{\theta}_B - \theta_0) = \sum_i n^{-\frac{1}{2}} \psi(Y_{\bar{R}}^i, \theta_0) + o_p(1)$. The standard M -estimator arguments lead to

$$n^{\frac{1}{2}}(\hat{\theta}_B - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n [I_c^{-1} \{S_i(\theta_0, \theta_0) + \lambda_2(\theta_0, \theta_0) \psi(Y_{\bar{R}}^i, \theta_0)\}] + o_p(1), \quad (\text{A1})$$

where by Assumption (S2) and similar derivations in Huber (1967), λ_2 is the partial derivative of λ with respect to its second argument. Define $S^{\text{mis}}(\theta, \eta) = S(\theta, \eta) - S^{\text{obs}}(\theta)$. By (2.4.1) of Meng & Rubin (1991), we obtain $S^{\text{mis}}(\theta, \theta) = (\partial/\partial\theta) \log f(Y_{\bar{R}}|Y_R, \theta)$. Therefore, we can write $\lambda_2(\theta_0, \theta_0)$ as

$$\int S(\theta_0, \theta_0) \{(\partial/\partial\eta) f(y_{\bar{R}}|Y_R, \eta) / f(y_{\bar{R}}|Y_R, \eta)\} dF(y_{\bar{R}}|Y_R, \eta)|_{\eta=\theta_0},$$

which, upon dropping the obvious arguments, is $E\{S(S^{\text{mis}})'|Y_R\}$. The fact that $E\{S^{\text{obs}}(S^{\text{mis}})'\} = E[E\{S^{\text{obs}}(S - S^{\text{obs}})'|Y_R\}] = 0$, with some simple derivations implies that $\lambda_2(\theta_0, \theta_0) = I_{\text{mis}}$. Write $S_i = S_i^{\text{obs}} + S_i^{\text{mis}}$ and define ρ_i to be $\psi(Y_{\bar{R}}^i, \theta_0) - I_{\text{obs}}^{-1} S_i^{\text{obs}}$. The influence function of $\hat{\theta}_B$ in (A1) can be written as $I_c^{-1} (I + I_{\text{mis}} I_{\text{obs}}^{-1}) S_i^{\text{obs}} + I_c^{-1} S_i^{\text{mis}} + J' \rho_i$; I is the identity matrix. The fact that S^{obs} , S^{mis} and ρ are mutually orthogonal, and the equality

$$I + I_{\text{mis}} I_{\text{obs}}^{-1} = I_c I_{\text{obs}}^{-1}, \quad (\text{A2})$$

imply the resulting variance in (1). \square

Proof of Theorem 2. By similar derivations to those leading to Theorem 1, we obtain

$$n^{\frac{1}{2}}(\hat{\theta}_A - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n I_{\text{obs}}^{-1} S_i^{\text{obs}} + m^{-1} \sum_{j=1}^m \left(n^{-\frac{1}{2}} \sum_{i=1}^n I_c^{-1} S_{ij}^{\text{mis}} \right) + m^{-1} \sum_{j=1}^m (n^{\frac{1}{2}} J' V_j) + o_p(1), \quad (\text{A3})$$

where $S_{ij}^{\text{mis}} = S_{ij} - S_{ij}^{\text{obs}}$; $V_j = \tilde{\theta}_j - \hat{\theta}_{\text{MLS}}$; and $\tilde{\theta}_j$, as defined in § 3, are independent draws from the posterior density of θ given the observed Z . Under the standard regularity conditions in § 3.2, given Z , $m^{-1} \sum (n^{\frac{1}{2}} J' V_j)$ converges, almost surely on Z , to an asymptotic normal with mean zero and variance $m^{-1} J' I_{\text{obs}}^{-1} J$. The sum of the first two terms in (A3) converges unconditionally to an asymptotic normal with mean zero and variance $I_{\text{obs}}^{-1} + m^{-1} I_c^{-1} J$. Theorem 2 thus follows from Lemma 1 of Schenker & Welsh (1988), which implies that the asymptotic distribution of $n^{\frac{1}{2}}(\hat{\theta}_A - \theta_0)$ is the convolution of the two normal distributions above. \square

Proof of (3) and (4). To prove (3), it is sufficient to show that $I_{\text{obs}}^{-1} - I_c^{-1} = I_c^{-1} J + J' I_{\text{obs}}^{-1} J$. The right-hand side of the equation equals $I_c^{-1} (I + I_{\text{mis}} I_{\text{obs}}^{-1}) I_{\text{mis}} I_c^{-1}$, which by (A2) is $I_{\text{obs}}^{-1} (I_c - I_{\text{obs}}) I_c^{-1}$; we thus obtain (3). To prove (4), by the law of large numbers it is sufficient to show that, when $m \rightarrow \infty$, \hat{Q}_A converges to $I_c^{-1} J + J' I_{\text{obs}}^{-1} J$. Write \hat{Q}_A as

$$(m-1)^{-1} \sum_j [n^{\frac{1}{2}} \{(\hat{\theta}_{Aj} - \theta_0) - (\hat{\theta}_A - \theta_0)\}]^{\otimes 2}.$$

Straightforward derivations further show that

$$\hat{Q}_A = (m-1) \sum \{n^{\frac{1}{2}} J' (V_j - \bar{V})\}^{\otimes 2} + \{n(m-1)\}^{-1} \sum \sum \{I_c^{-1} (S_{ij}^{\text{mis}} - \bar{S}_i^{\text{mis}})\}^{\otimes 2} + o_p(1),$$

where V_j is defined at (A3); \bar{V} is the average of the mV_j 's and \bar{S}_i^{mis} is defined analogously. When both n and m go to infinity, it is easy to show that the first term above converges to $J'I_{\text{obs}}^{-1}J$ while the second term converges to $I_c^{-1}J$. \square

Influence function for $\hat{\theta}_{\text{sto},K}$ and $\hat{\theta}_{\text{sim},K}$. Note that, with $J^0 = I$,

$$\sum_{k=1}^K J^{k-1} + I_c(J^K)'I_{\text{obs}}^{-1} \equiv I_c I_{\text{obs}}^{-1}, \quad (\text{A4})$$

which can be proved easily by induction on K . By iterative substitutions into (A1) and by applying (A4), we have

$$n^{\frac{1}{2}}(\hat{\theta}_{\text{sto},K} - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n I_{\text{obs}}^{-1} S_i^{\text{obs}} + n^{-\frac{1}{2}} \sum_{i=1}^n \left(\sum_{k=1}^K m_k^{-1} I_c^{-1} J^{K-k} S_i^{\text{mis},(k)} \right) + n^{-\frac{1}{2}} \sum_{i=1}^n (J^K)' \rho_i + o_p(1), \quad (\text{A5})$$

where $S_i^{\text{mis},(k)}$ is the S_i^{mis} component of the k th iterate, and ρ_i is defined in the proof of Theorem 1. Similarly,

$$n^{\frac{1}{2}}(\hat{\theta}_{\text{sim},K} - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n I_{\text{obs}}^{-1} S_i^{\text{obs}} + n^{-\frac{1}{2}} \sum_{i=1}^n m^{-1} I_c^{-1} \left(\sum_{k=1}^K J^{K-k} \right) S_i^{\text{mis},(1)} + n^{-\frac{1}{2}} \sum_{i=1}^n (J^K)' \rho_i + o_p(1). \quad (\text{A6})$$

Proof of Corollary 2. To show that $V_{\text{sim},K} - V_{\text{sto},K}$ is positive semidefinite, we only need to show that $\Sigma_{\text{imp},\text{sim}}^{(K)} - \Sigma_{\text{imp},\text{sto}}^{(K)}$ is positive semidefinite, which, by using the fact that $J^{k_1} I_{\text{mis}} (J^{k_2})'$ is positive semidefinite for any two nonnegative integers k_1 and k_2 can be easily obtained by induction. Note that $J^{k_1} I_{\text{mis}} (J^{k_2})'$ equals $MI_c^{-1}M'$ for some matrix M when $k_1 + k_2$ is odd, and equals $MI_{\text{mis}}M'$ when $k_1 + k_2$ is even, which implies that it is always positive semidefinite.

To show (i) and (ii) in Corollary 2, we need the following lemma, which we state first and prove afterwards.

LEMMA A1. *Consider the norm of a matrix to be the largest absolute value of all elements in the matrix. Then, when k goes to ∞ , $\|J^k\| \rightarrow 0$.*

Proof. Since I_c and I_{mis} are real symmetric matrices and I_c is positive definite, there exists a nonsingular matrix R such that $I_{\text{mis}} = R'\Lambda R$ and $I_c = R'R$, where Λ is a diagonal matrix with i th diagonal element λ_i (Rao, 1985, p. 41). Since $I_{\text{obs}} = I_c - I_{\text{mis}}$ is positive definite, it implies that the λ_i 's are all less than 1. The result follows since $J^K = (I_{\text{mis}} I_c^{-1})^K = R'(\Lambda)^K (R^{-1})'$. \square

By Lemma A1, $\|\Sigma_p^{(K)}\|$ goes to zero. To find the limit of $V_{\text{sim},K}$ and $V_{\text{sto},K}$, we only need to find the limit of $\Sigma_{\text{imp},\text{sim}}^{(K)}$ and $\Sigma_{\text{imp},\text{sto}}^{(K)}$. The results claimed in Corollary 2 can be easily obtained by noting that $I - J = I_{\text{obs}} I_c^{-1}$, and that

$$\left(\sum_{k=1}^K J^{K-k} \right) (I - J) = I - J^K, \quad \sum_{k=1}^K \{I_c^{-1} J^{K-k} I_{\text{mis}} (J^{K-k})' I_c^{-1}\} (I - J^2) = I_c^{-1} I_{\text{mis}} I_c^{-1} - I_c^{-1} K^{2K+1}.$$

Influence function for $\bar{\theta}_{\text{sto},K}$. By (A5), we have

$$\begin{aligned} n^{\frac{1}{2}}(\bar{\theta}_{\text{sto},K} - \theta_0) &= n^{-\frac{1}{2}} \sum_{i=1}^n I_{\text{obs}}^{-1} S_i^{\text{obs}} + n^{-\frac{1}{2}} \sum_{i=1}^n (mK)^{-1} I_c^{-1} \left\{ \sum_{k=1}^K \left(\sum_{l=0}^{K-k} J^l \right) S_i^{\text{mis},(k)} \right\} \\ &\quad + n^{-\frac{1}{2}} \sum_{i=1}^n \left\{ K^{-1} \sum_{k=1}^K (J^k)' \right\} \rho_i + o_p(1). \end{aligned} \quad (\text{A7})$$

Straightforward calculations imply that, when n and $K \rightarrow \infty$, the variances of the second and third terms in (A7) go to 0, that is the asymptotic variance of $n^{\frac{1}{2}}(\bar{\theta}_{\text{sto},K} - \theta_0) \rightarrow I_{\text{obs}}^{-1}$.

REFERENCES

- BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. New York: Wiley.
- CARROLL, R. J., RUPPERT, D. & STEFANSKI, L. A. (1995). *Measurement Error in Non-linear Models*. London: Chapman and Hall.
- CELEUX, G. & DIEBOLT, J. (1985). The EM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statist. Quart.* 2, 73–82.
- DELTOUR, I., RICHARDSON, S. & LE HESRAN, J. L. (1998). Stochastic algorithms for Markov models estimation with intermittent missing data. *Biometrics* ???.
- DIEBOLT, J. & CELEUX, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Commun. Statist. B* 9, 599–613.
- DIEBOLT, J. & IP, E. H. S. (1996). Stochastic EM: method and application. In *Markov Chain Monte Carlo in Practice*, Ed. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, pp. 259–68, New York: Springer-Verlag.
- FAY, R. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Am. Statist. Assoc.* 91, 490–8.
- GREENLAND, S. & FINKLE, W. D. (1995). A critical look at basic methods for handling missing covariates in epidemiologic regression analyses. *Am. J. Epidem.* 142, 1255–64.
- HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* 47, 663–85.
- HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, 1, pp. 221–33.
- McFADDEN, D. & RUDD, P. A. (1994). Estimation by simulation. *Rev. Econ. Statist.* 76, 591–608.
- MENG, X. L. (1994). Multiple imputation inferences with uncongenial sources of input. *Statist. Sci.* 9, 538–58.
- MENG, X. L. & RUBIN, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Assoc.* 86, 899–909.
- RAO, C. R. (1985). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.
- REILLY, M. (1993). Data analysis using hot-deck multiple imputation. *Statistician* 42, 307–13.
- ROBINS, J. M. & GILL, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statist. Med.* 16, 39–56.
- RUBIN, D. B. (1978). Multiple imputation in sample surveys—a phenomenological Bayesian approach to nonresponse. In *Proc. Survey Res. Meth. Sect., Am. Statist. Assoc.* pp. 20–34. Washington, DC: American Statistical Association.
- RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- RUBIN, D. B. (1996). Multiple imputation after 18 years. *J. Am. Statist. Assoc.* 91, 473–90.
- RUUD, P. A. (1991). Extensions of estimation methods using the EM algorithm. *J. Economet.* 49, 305–41.
- SCHENKER, N. & WELSH, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* 16, 1550–66.
- TANNER, M. A. (1993). *Tools for Statistical Inference*. New York: Springer-Verlag.
- WEI, G. C. G. & TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Assoc.* 85, 699–704.

[Received March 1997. Revised January 1998]