The Imputation Maximization Algorithm

Søren Feodor Nielsen
Department of Theoretical Statistics
University of Copenhagen

Abstract

An appealing and much used way of handling missing or incomplete data is to impute complete data, i.e. to fill out the holes in the data material. The main advantage of imputation methods is that they lead to complete data sets, and thus to simpler estimation problems. Obviously, the imputations must be chosen with care. This paper introduces a method, the Imputation Maximization algorithm, of constructing imputations for general incomplete data problems by iterating two simple steps: an imputation step and a maximization step. Large sample results are discussed.

Keywords: Incomplete observations, imputation, simulation, EM algorithm

1 Introduction

An appealing and much used way of handling missing or incomplete data is to impute complete data, i.e. to fill out the holes in the data material; a basic reference on imputation is Little and Rubin (1987). Imputing missing data leads to complete data sets and thus to simpler estimation problems. In order to obtain reasonable estimates, the imputations should be chosen with care.

The simplest example of an imputation procedure is mean imputation, where unobserved data is replaced by simple averages of observed data. This method gives reasonable estimates if the missing data is missing completely at random (cf Little and Rubin (1987)). It is however easy to see that the average of the imputed and observed data is just the average of the observed data; in other words the imputation is superfluous.

A major drawback of imputation methods is in fact that if "correct" values can be imputed, the missing data problem is so simple that better methods exist. In the mean imputation example, an average over observed data gives the same result. Conversely many problems do not facilitate simple "correct" imputations. Another problem with imputation methods is that special expressions for the variance of estimators are needed. In the mean imputation example given above, the naive choice of variance –the complete data variance– will be too small.

This paper introduces a method, the Imputation Maximization algorithm, of constructing valid imputations for general incomplete data problems by iterating two simple steps.

Let $X_1, X_2, \ldots, X_n$ be iid random variables from a distribution indexed by an unknown parameter θ. Suppose that $Y_i = Y(X_i)$ is observed rather than X_i. Throughout the paper X denotes a generic random variable from the unknown distribution and Y the corresponding incomplete observation; the double meaning of Y should not cause any confusion.

The IM-algorithm takes the following form: From an arbitrary starting value $\tilde{\theta}_n(0)$ a sequence $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is formed by going through an I-step (imputation) and an M-step (maximization):

I-step : Given a value of $\tilde{\theta}_n(k)$ impute values $\widetilde{X}_i$ from the conditional distribution of X given $Y = y_i$ under $\tilde{\theta}_n(k)$, i.e. simulate $\widetilde{X}_i \sim \mathcal{L}_{\tilde{\theta}_n(k)}(X|Y = y_i)$.

M-step : Maximize the resulting complete data log-likelihood, $\sum_{i=1}^{n} \log f_\theta(\widetilde{X}_i)$, and let the maximizer be the next value, $\tilde{\theta}_n(k+1)$.

As the maximization in the M-step is a complete data problem, it is often easy to solve either explicitly or iteratively using standard algorithms such as Newton-Raphson or Scoring.

It is clear that by imputing new independent $\widetilde{X}_i$-s in each step, the sequence of maximizers $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is a time-homogeneous Markov chain, when the observed data is fixed (i.e. conditioned upon). In Subsection 3.1, ergodicity of the chain is discussed. If ergodic, the algorithm will converge in the sense that as k, the number of iterations, tends to infinity, $\tilde{\theta}_n(k)$ converges in distribution to a random variable, $\tilde{\theta}_n$, where $\tilde{\theta}_n$ is distributed according to the stationary distribution of the Markov chain.

We will show that under mild assumptions this sequence of estimates, $(\tilde{\theta}_n)_{n\in\mathbb{N}}$, is asymptotically normal as n, the number of observations, tends to infinity. Similar results have been shown for the special case of finite mixtures (and conjectured to hold more generally) by Celeux and Diebolt (1993).

2 Preliminaries

In this section necessary results about Markov chains (Subsection 2.1), some notation is introduced (Subsection 2.2), and assumptions to be used in showing the asymptotic results are given (Subsection 2.3). The results in Subsection 2.1 are adapted from Ethier and Kurtz (1986, chapter 4).

2.1 Markov chains

Let P_n be transition probabilities for a Markov chain and let μ_n be the corresponding stationary initial distributions, which are assumed to exist. The following assumptions (referred to as assumption **C**) will be used:

C1 $P_n(x,\cdot) \xrightarrow{w} P(x,\cdot)$ uniformly over compacta, i.e. for all compact sets, $K \subseteq S$, $\sup_{x\in K} |\int f(y)P_n(x,dy) - \int f(y)P(x,dy)| \to 0$ as $n \to \infty$ for all $f \in \mathcal{C}_b(S)$.

C2 $x \to \int f(y)P(x,dy)$ is continuous for all $f \in \mathcal{C}_b(S)$.

Assumption **C** implies that $\int f(y)P_n(\cdot,dy)$ converges to $\int f(y)P(\cdot,dy)$ continuously, i.e. $\int f(y)P_n(x_n,dy) \to \int f(y)P(x,dy)$ when $x_n \to x$.

Conversely, continuous convergence of $\int f(y)P_n(\cdot,dy)$ to $\int f(y)P(\cdot,dy)$ implies **C1** and **C2**. See for instance Roussas (1972, p. 132). Thus assumption **C** may be replaced by

C* $\int f(y)P_n(\cdot,dy)$ converges continuously to $\int f(y)P(\cdot,dy)$ for each $f \in \mathcal{C}_b(S)$, where P is a transition probability.

Proposition 1 *Suppose that assumption* **C** *holds. If a subsequence of* $(\mu_n)_{n\in\mathbb{N}}$ *converges weakly to a probability* μ*, then* μ *is a stationary initial distribution for the Markov chain with transition kernel* P*.*

Proof:
Suppose $(\mu_{n'})_{n'}$ is a convergent subsequence with limit μ. Then

$$\begin{aligned}
&|\iint f(y)P(x,dy)d\mu(x) - \int f d\mu| \\
&\leq |\iint f(y)P(x,dy)d\mu(x) - \iint f(y)P(x,dy)d\mu_{n'}(x)| \\
&\quad + |\int \left(\int f(y)P(x,dy) - \int f(y)P_{n'}(x,dy)\right) d\mu_{n'}(x)| \\
&\quad + |\int f d\mu_{n'} - \int f d\mu|
\end{aligned}$$

Here the first and the third terms vanish as $n' \to \infty$. Let $\varepsilon > 0$ be arbitrary. According to Prohorov's Theorem, a compact set K can be chosen such that $\inf_{n'} \mu_{n'}(K) > 1 - \varepsilon/2$. Thus

$$\begin{aligned}
&|\int \left(\int f(y)P(x,dy) - \int f(y)P_{n'}(x,dy)\right) d\mu_{n'}(x)| \\
&\leq \frac{\varepsilon}{2} + \sup_{x\in K} |\int f(y)P(x,dy) - \int f(y)P_{n'}(x,dy)|
\end{aligned}$$

where the second term can be made arbitrarily small (smaller than $\varepsilon/2$, say) by assumption **C1**. □

From Proposition 1 follows:

Corollary 1 *Suppose assumption* **C** *holds. If μ is the unique stationary distribution corresponding to P and $(\mu_n)_{n\in\mathbb{N}}$ is tight, then $\mu_n \xrightarrow{w} \mu$.*

Proof:
Tightness implies that any subsequence of $(\mu_n)_{n\in\mathbb{N}}$ has a convergent sub-subsequence, $(\mu_{n'})$. According to Proposition 1, $\mu_{n'} \xrightarrow{w} \mu$, and the result follows from Theorem 2.3 in Billingsley (1968). □

In order to apply Corollary 1, a criterion for tightness is necessary:

Proposition 2 *Suppose that for each $n \in \mathbb{N}$ $(Z_k^n)_{k\in\mathbb{N}_0}$ is an ergodic Markov chain on a finite dimensional Euclidean state space with initial distribution ν_n and transition kernel P_n. Let μ_n be the corresponding stationary initial distribution.*
If there exists functions $\varphi_n : S \to [0;\infty[$, $\psi_n : S \to \mathbb{R}$, and $\psi : S \to]-\infty; C]$ for some $C > 0$, such that:

(i) $\int \varphi_n d\nu_n$ is finite for all $n \in \mathbb{N}$.

(ii) $\psi_n \leq \psi$ for all $n \in \mathbb{N}$.

(iii) $-\psi$ is a norm-like function.

(iv) $E(\varphi_n(Z_l^n)) \leq E(\varphi_n(Z_{l-1}^n)) + E(\psi_n(Z_{l-1}^n))$ for all $n, l \in \mathbb{N}$.

then $(\mu_n)_{n\in\mathbb{N}}$ is tight.

Remark:
$-\psi$ is norm-like (see Meyn and Tweedie, 1995, p. 214) if $\overline{\{z : -\psi(z) \leq m\}}$ is compact for each $m > 0$.

Proof:
Letting $K_m = \overline{\{z : \psi(z) \geq -m\}}$, we get $\psi_n \leq C \cdot 1_{K_m} - m \cdot 1_{K_m^c} = (C+m) \cdot 1_{K_m} - m$ by (ii). Now

$$\begin{aligned} 0 \;&\leq\; E(\varphi_n(Z_l^n)) \leq E(\varphi_n(Z_0^n)) + E(\sum_{k=0}^{l-1} \psi_n(Z_k^n)) \\ &\leq\; \int \varphi_n d\nu_n + (C+m) E(\sum_{k=0}^{l-1} 1_{K_m(Z_k^n)}) - ml \end{aligned}$$

so that

$$\frac{m}{C+m} - \frac{1}{l} \int \varphi_n d\nu_n \frac{1}{C+m} \leq E(\frac{1}{l} \sum_{k=0}^{l-1} 1_{K_m(Z_k^n)})$$

When $l \to \infty$ the right hand side converges to $\mu_n(K_m)$ and the second term on the left hand side vanishes. As m may be chosen arbitrarily large and K_m is compact, $(\mu_n)_{n\in\mathbb{N}}$ is tight. □

Remark:
A sufficient condition for condition (iv) of Proposition 2 is that $\varphi_n(Z_l^n) - \sum_{k=0}^{l-1} \psi_n(Z_k^n)$ is a super martingale (wrt some suitable filtration).

2.2 Notation

Let f_θ be the density of X wrt some dominating probability measure, μ. The resulting probability measure is denoted P_θ, and expectations with respect to this probability is denoted E_θ. Let $s_x(\theta)$ be the corresponding score function and $V(\theta) = E_\theta(s_X(\theta)^{\otimes 2})$.

The conditional density of X given $Y = y$ wrt some probability measure ν_y is denoted $k_\theta(x|y)$ and the corresponding score function $s_{x|y}(\theta)$. Let $I_y(\theta) = E_\theta(s_{X|y}(\theta)^{\otimes 2}|Y = y)$.

The density of Y is denoted h_θ and the corresponding score function is $s_y(\theta)$. Put $I(\theta) = E_\theta(s_Y(\theta)^{\otimes 2})$.

Notice that $s_y(\theta) = s_x(\theta) - s_{x|y}(\theta)$ and that $I(\theta) = V(\theta) - E_\theta I_Y(\theta)$, when $E_\theta(s_{X|y}(\theta)|Y = y) = 0$.

Let $F(\theta) = E_\theta I_Y(\theta)V(\theta)^{-1}$ be the expected fraction of missing information, and let θ_0 denote the true unknown value of $\theta \in \Theta \subseteq \mathbb{R}^d$.

As in Section 1 we shall use $\widetilde{X}_i$ for imputed values. The distribution of the imputed values is denoted $\tilde{P}_\theta$. This is of course just the conditional distribution of the unobserved X_is given the observed values of $Y_i = y_i$ under P_θ. Generally the imputed values are not from the same distribution as the observed Y_is (i.e. the correct distribution, P_{θ_0}), and the introduced notation should help to keep the distinction between imputed and observed variables clear. Notice that the $\tilde{P}_\theta$-notation is short-hand in the sense that the dependence upon the observed y_is is suppressed.

2.3 Assumptions

The necessary assumptions can be divided roughly into three groups corresponding to which model they relate to.

On the model for the observed data, $Y_1, Y_2, \ldots, Y_n$, it is assumed that there is a solution, $\hat{\theta}_n$ to the likelihood equation, $\frac{1}{n}\sum_{i=1}^n s_{y_i}(\theta) = 0$, such that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathrm{N}(0, I(\theta_0)^{-1})$.

It is also assumed that $\hat{\theta}_n$ converges to θ almost surely. This assumption is not necessary but it will simplify the proofs of Subsection 3.2. The strong consistency may be replaced by taking almost surely convergent subsequences of arbitrary subsequences and applying Theorem 2.3 in Billingsley (1968).

The assumption on the observed data model may be difficult to verify in practice, but it will typically follow if the complete data model and the missing data model are sufficiently smooth. It does not seem to be an unreasonable assumption since the IM algorithm attempts to mimic maximum likelihood estimation. Hence, we would not expect it to have better properties than maximum likelihood.

On the model for the complete data, $X_1, X_2, \ldots, X_n$, it is assumed that (for n sufficiently large) there is (with probability 1) a unique maximum likelihood estimator.

We must also assume that if $\theta_n \to \theta_0$ then for almost all y-sequences

$$\sqrt{n}(\tilde{\theta}_n(1) - \theta_n) = V(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\widetilde{X}_i}(\theta_n) + o_{\tilde{P}_{\theta_n}}(1) \tag{1}$$

where $\widetilde{X}_i \sim \mathcal{L}_{\theta_n}(X|Y = y_i)$ and $\tilde{\theta}_n(1)$ is the complete data maximum likelihood estimator

based on the imputed data. This condition may be interpreted as an assumption of the complete data MLE based on approximately correct imputations is approximately efficient. This is close to assuming local efficiency of the complete data MLE.

It is difficult to give simple yet general sufficient conditions for (1), but it should not be difficult to verify in actual applications. For instance, it is easily seen to hold if the complete data model is a full exponential family. If the complete data model is smooth (in the sense of Lehman (1983), say), then we need to show that

$$-\frac{1}{n}\sum_{i=1}^{n} D_\theta s_{\widetilde{X}_i}(\theta_n) \xrightarrow{\tilde{P}_{\theta_n}} V(\theta_0)$$

for almost every y-sequence when $\theta_n \to \theta_0$ and either that the integrable majorizer of the third derivative can be chosen to depend on y only or that a law of large numbers (again given y) applies to this majorizer.

Finally, some assumptions are necessary on the model for the missing data, i.e. the conditional distribution of X given $Y = y$. This is assumed to be regular (in the sense of Bickel et al. (1993, Section 2.1)). Let

$$\theta \to D_\theta k_\theta^{\frac{1}{2}}(\cdot|y) = \dot{k}_\theta^{\frac{1}{2}}(\cdot|y)$$

denote the ($\mathrm{L}(\nu_y)^2$-)derivative of the root density,

$$\theta \to \sqrt{k_\theta(\cdot|y)} = k_\theta^{\frac{1}{2}}(\cdot|y).$$

Assume that for almost every y-sequence and every compact $C \subseteq \Theta$

(U) $\sup_{\theta\in C} |\frac{1}{n}\sum_{i=1}^n I_{y_i}(\theta) - E_{\theta_0} I_Y(\theta)| \to 0$ as $n \to \infty$.

(D) $\forall h \in \mathbb{R}^d : \sup_{\theta\in C} \sup_{u\in[0;1/\sqrt{n}]} \frac{1}{n}\sum_{i=1}^n \int \left(h^t \dot{k}_{\theta+uh}^{\frac{1}{2}}(x|y_i) - h^t \dot{k}_\theta^{\frac{1}{2}}(x|y_i)\right)^2 d\nu_{y_i}(x) \to 0$ as $n \to \infty$

(L) $\forall \varepsilon > 0 : \sup_{\theta\in C} \frac{1}{n}\sum_{i=1}^n E_\theta\left((h^t s_{X_i|y_i}(\theta))^2 1_{\{|h^t s_{X_i|y_i}(\theta)|\geq\varepsilon/\sqrt{n}\}}\right) \to 0$ as $n \to \infty$

Then (by a straightforward extension of Theorem II.6.2 in Ibragimov and Has'minskii (1981))

$$\begin{aligned}\sum_{i=1}^n \log k_{\theta+\frac{1}{\sqrt{n}}h}(X_i|y_i) - \sum_{i=1}^n \log k_\theta(X_i|y_i) \\ = h^t \frac{1}{\sqrt{n}}\sum_{i=1}^n s_{X_i|y_i}(\theta) - \frac{1}{2}h^t E_{\theta_0} I_Y(\theta) h + R_n(\theta, h)\end{aligned} \tag{2}$$

where

$$\sup_{|h|\leq M}\sup_{\theta\in C} P_\theta\{|R_n(\theta,h)| > \varepsilon\} \to 0$$

for every $\varepsilon, M > 0$ and every compact set $C \subseteq \Theta$, and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{X_i|y_i}(\theta_n) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, \frac{1}{2}E_{\theta_0} I_Y(\theta_0)\right)$$

for every sequence $(\theta_n)_{n\in\mathbb{N}}$ converging to θ_0.

Assumption (L) may be verified in practice by showing that a Lyapounov-type condition holds. Assumptions (U) and (D) may be shown using empirical process techniques or from further smoothness.

Finally, the distributions of the conditional model, $(k_\theta(\cdot|y)\cdot\nu_y)_{\theta\in\Theta}$, must be mutually equivalent.

Unlike the assumptions on the observed data model, the other assumptions will generally be easy to check since the complete data may to some degree be chosen by the data analyser.

The assumptions on the complete data model and the model for the missing data may contradict each other; in order to get a sufficiently smooth complete data likelihood, the conditional model may turn out to be complicated and vice versa. So the assumptions represent a trade-off.

3 Results

In Subsection 3.1 the convergence of the IM algorithm (for a fixed sample size n) is discussed. The properties of the Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ are discussed, and sufficient criteria for ergodicity are given.

In Subsection 3.2 large sample results for the sequence of estimators $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ are given. The main result is Theorem 1, where the limiting distribution of $\sqrt{n}(\tilde{\theta}_n-\theta_0)$ is identified. A criteria for tightness of the sequence $\sqrt{n}(\tilde{\theta}_n-\theta_0)$ is given.

Both subsections are technical, and especially Subsection 3.2 –apart from Theorem 1– may be skipped by the less technically inclined reader.

3.1 Convergence

It is clear that by imputing new independent $\widetilde{X}_i$-s in each step, the sequence of maximizers $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is a time-homogeneous Markov chain. Also the imputed x-values make up a Markov chain; this is denoted $(\widetilde{X}(k))_{k\in\mathbb{N}_0}$ (suppressing the dependence on n).

Remark:
In the proofs of this subsection the dependence on n will be suppressed in the notation. We may –and will– assume that supp ν_y = supp $k_\theta(\cdot|y)\cdot\nu_y$ for all $\theta\in\Theta$ since the conditional distributions are mutually equivalent.

Lemma 1 *The Markov chain $(\widetilde{X}(k))_{k\in\mathbb{N}_0}$ is irreducible and aperiodic.*

Proof:
As $\tilde{P}\{\widetilde{X}(1)\in B|\widetilde{X}(0)=x\}=\tilde{P}\{\widetilde{X}(1)\in B|\tilde{\theta}(0)=\theta\}=\int_B k_\theta(x'|y)d\nu_y(x')>0$ for any x –here θ is the complete data MLE corresponding to the observation x– and any measurable set B such that $\nu_y(B)>0$, the chain is ν_y-irreducible.

The aperiodicity follows from the following observation. Suppose that the chain is periodic with period at least two. Then there exists two disjoint sets, D_1 and D_2 such that $\tilde{P}\{\widetilde{X}(1)\in D_2|\widetilde{X}(0)=x\}=1$ for all $x\in D_1$ (see Meyn & Tweedie (1993, Theorem 5.4.4)). Consequently, $\int_{D_2}k_\theta(x|y)d\nu_y(x)=1$ for all values of θ, and D_1 must be a null-set for all θ. In other words, $\tilde{P}\{\widetilde{X}(1)\in D_1|\widetilde{X}(0)=x\}=0$ for all x, contradicting Theorem 5.4.4 in Meyn & Tweedie (1993). □

The Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ inherits some of the properties of the chain $(\widetilde{X}(k))_{k\in\mathbb{N}_0}$.

Corollary 2 *The Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is irreducible and aperiodic. If the Markov chain $(\widetilde{X}(k))_{k\in\mathbb{N}_0}$ is ergodic, then so is $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$. In particular, if the imputations have a finite sample space, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is (uniformly) ergodic.*

The regularity assumption has the following consequence:

Lemma 2 *The Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ has the weak Feller property, and compact sets are small.*

Proof:
The assumed regularity of the model implies that for all measurable sets B we get $\int_B k_{\theta'}(x|y)d\nu_y(x) \to \int_B k_\theta(x|y)d\nu_y(x)$ when $\theta' \to \theta$. In particular, for an open set $O \subseteq \Theta$, $\tilde{P}\{\tilde{\theta}(1) \in O|\tilde{\theta}(0) \in \theta'\} \to \tilde{P}\{\tilde{\theta}(1) \in O|\tilde{\theta}(0) \in \theta\}$ as $\theta' \to \theta$. Thus the Markov chain has the weak Feller property.

If $K \subseteq \Theta$ is compact, then there is for each measurable B a $\theta' \in K$ such that $\inf_{\theta\in K} \int_B k_\theta(x|y)d\nu_y(x) = \int_B k_{\theta'}(x|y)d\nu_y(x)$ since the map $\theta \to \int_B k_\theta(x|y)d\nu_y(x)$ is continuous. Thus K is small. □

It is worth noticing that typically $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ has a smaller state space than Θ if the imputed x-values are discrete or the observed y-values restrict the sample space. For instance, if the sample space of the imputations is finite, then so is the actual state space of the Markov chain. The actual state space is the set of possible complete data maximum likelihood estimates based on imputed data; or more precisely the image of the support of the conditional distribution of $(X_i)_{i=1,\ldots,n}$ given $(Y_i)_{i=1,\ldots,n} = (y_i)_{i=1,\ldots,n}$ under the mapping that transforms complete data to the corresponding complete data MLE. Let $\widetilde{\Theta}$ denote the actual state space; the dependence on $y_1, y_2, \ldots, y_n$ is suppressed.

Since compact sets are small, we get:

Corollary 3 *If $\widetilde{\Theta}$ is compact, then $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, is (uniformly) ergodic.*

When $\widetilde{\Theta}$ is not compact, a drift criteria is needed in order to show ergodicity.

Let M denote the EM algorithm. Recall from general theory of the EM algorithm (e.g. Dempster, Laird & Rubin (1977)) that $\hat{\theta}_n$ is a fixed point of M. Let $\lambda(\theta)$ be the largest eigenvalue of $D_\theta M(\theta) = V(M(\theta))^{-1} \cdot \frac{1}{n}\sum_{i=1}^n I_{y_i}(\theta)$. Since the model is regular $V(\theta) - \frac{1}{n}\sum_{i=1}^n I_{y_i}(\theta)$ is positive definite and $0 \leq \lambda(\hat{\theta}_n) < 1$. Continuity (also a consequence of regularity) ensures that $\lambda < 1$ in a neighbourhood of $\hat{\theta}_n$.

Proposition 3 *Suppose that $\lambda(\theta) \leq \lambda < 1$ for all $\theta \in \widetilde{\Theta}$.*
If there exists an $\varepsilon \in [0; 1-\lambda[$ such that

$$\tilde{E}\left(\|\tilde{\theta}_n(1) - M(\tilde{\theta}_n(0))\|_2|\tilde{\theta}_n(0)\right) - \varepsilon\|\tilde{\theta}_n(0) - \hat{\theta}_n\|_2 \leq K1_C(\tilde{\theta}_n(0))$$

almost surely for some compact set, $C \subseteq \widetilde{\Theta}$, and some constant $K < \infty$, then the Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is (geometrically) ergodic.

Proof:
Since

$$\begin{aligned}\|\tilde{\theta}(1) - \hat{\theta}_n\|_2 &\leq \|\tilde{\theta}(1) - M(\tilde{\theta}(0))\|_2 + \|M(\tilde{\theta}(0)) - \hat{\theta}_n\|_2 \\ &\leq \|\tilde{\theta}(1) - M(\tilde{\theta}(0))\|_2 + \lambda\|\tilde{\theta}(0) - \hat{\theta}_n\|_2\end{aligned}$$

we get

$$\tilde{E}\left(1+\|\tilde{\theta}_n(1)-\hat{\theta}_n\|_2|\tilde{\theta}_n(0)\right) \leq (\lambda+\varepsilon)\cdot\left(1+\|\tilde{\theta}_n(0)-\hat{\theta}_n\|_2\right)+K1_C(\tilde{\theta}_n(0))-\lambda-(1-\varepsilon)$$

guaranteeing ergodicity (see Meyn & Tweedie (1993), Theorem 15.0.1). □

Remark:
In practice, the inequality in Proposition 3 may be verified by showing that $\tilde{E}(\|\tilde{\theta}_n(1)-M(\tilde{\theta}_n(0))\|_*|\tilde{\theta}_n(0)) = o(\|\tilde{\theta}_n(0)-\hat{\theta}_n\|_2)$ for some norm, $\|\cdot\|_*$. The boundedness typically follows from continuity.

For exponential family models –with θ the expectation of the canonical statistic– $M(\tilde{\theta}_n(0)) = \tilde{E}(\tilde{\theta}_n(1)|\tilde{\theta}_n(0))$. It thus suffices to show that

$$\sqrt{\text{trace}\left(\widetilde{Var}(\tilde{\theta}_n(1)|\tilde{\theta}_n(0)\right)}-\varepsilon\|\tilde{\theta}_n(0)-\hat{\theta}_n\|_2 \leq K1_C(\tilde{\theta}_n(0))$$

3.2 Asymptotic normality

In this subsection asymptotic normality of the estimators $\sqrt{n}(\tilde{\theta}_n-\theta_0)$ is discussed. The aim is to apply Corollary 1 on the sequence $\sqrt{n}(\tilde{\theta}_n-\hat{\theta}_n)$ conditional on the observed y-sequence.

In order to do this, we have to look at how the transition probabilities from a fixed point of the sample space behave as n tends to infinity. Here the "fixed points" of the Markov chains, $(\sqrt{n}(\tilde{\theta}_n(k)-\hat{\theta}_n))_{k\in\mathbb{N}_0}$ has the form $h=\sqrt{n}(\theta_n-\hat{\theta}_n)$. We will verify assumption **C*** in the following lemma. Thus we need to look at a convergent sequence, $h_n=\sqrt{n}(\theta_n-\hat{\theta}_n)$, of points in the sample space, i.e. θ-values of the type $\theta_n=\hat{\theta}_n+\frac{1}{\sqrt{n}}h+o(\frac{1}{\sqrt{n}})$, and show continuous convergence of the transition probabilities.

Lemma 3 *Let $\widetilde{X}_i\sim\mathcal{L}_{\theta_n}(X|Y=y_i)$, where $\theta_n=\hat{\theta}_n+\frac{1}{\sqrt{n}}h+o(\frac{1}{\sqrt{n}})=\tilde{\theta}_n(0)$. For almost all y-sequences and conditional on y:*

$$\sqrt{n}(\tilde{\theta}_n(1)-\hat{\theta}_n)\xrightarrow{\mathcal{D}}N\left(F(\theta_0)^th, V(\theta_0)^{-1}E_{\theta_0}I_Y(\theta_0)V(\theta_0)^{-1}\right)$$

where $\tilde{\theta}_n(1)$ is the complete data maximum likelihood estimator based on imputed data, and $F(\theta_0)=E_{\theta_0}(I_Y(\theta_0))V(\theta_0)^{-1}$ is the expected fraction of missing information.

Proof:
Observe that from (2) the log-likelihood based on the imputations of θ_n to $\hat{\theta}_n$ is

$$l_n(\theta_n;\hat{\theta}_n)=h^t\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\widetilde{X}_i|y_i}(\hat{\theta}_n)-\frac{1}{2}h^tE_{\theta_0}I_Y(\theta_0)h+o_{\tilde{P}_{\hat{\theta}_n}}(1)$$

From (1) follows that when $\widetilde{X}_i\sim\mathcal{L}_{\hat{\theta}_n}(X|Y=y_i)$

$$\begin{aligned}\sqrt{n}(\tilde{\theta}_n(1)-\hat{\theta}_n) &= V(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\widetilde{X}_i}(\hat{\theta}_n)+o_{\tilde{P}_{\hat{\theta}_n}}(1)\\ &= V(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\widetilde{X}_i|y_i}(\hat{\theta}_n)+o_{\tilde{P}_{\hat{\theta}_n}}(1)\end{aligned}$$

since $\sum_{i=1}^n s_{y_i}(\hat{\theta}_n)=0$.

Therefore, under $\hat{\theta}_n$, i.e. when $\widetilde{X_i} \sim \mathcal{L}_{\theta_n}(X|Y=y_i)$,

$$\begin{pmatrix} l_n(\theta_n;\hat{\theta}_n) \\ \sqrt{n}(\tilde{\theta}_n(1)-\hat{\theta}_n) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathrm{N}\left(\begin{pmatrix} -\frac{1}{2}h^t E_{\theta_0}(I_Y(\theta_0))h \\ 0 \end{pmatrix}, \begin{pmatrix} h^t E_{\theta_0} I_Y(\theta_0) h & h^t F(\theta_0) \\ F(\theta_0)^t h & V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1} \end{pmatrix}\right)$$

according to (2). LeCam's third Lemma implies that under θ_n

$$\sqrt{n}(\tilde{\theta}_n(1)-\hat{\theta}_n) \xrightarrow{\mathcal{D}} N\left(F(\theta_0)^t h, V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}\right) \tag{3}$$

as we wished to show. □

Lemma 3 shows that the transition probabilities of the $(\sqrt{n}(\tilde{\theta}_n(k)-\hat{\theta}_n))_{k\in\mathbb{N}_0}$ converges continuously to the transition probabilities of a multivariate Gaussian AR(1)-process. From this the main theorem follows.

Theorem 1 *Suppose $\sqrt{n}(\tilde{\theta}_n-\hat{\theta}_n)$ is tight.*
Then (unconditionally) $\sqrt{n}(\tilde{\theta}_n-\theta_0) \xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1}\left(2I-(I+F(\theta_0))^{-1}\right)\right)$,
where $F(\theta_0) = E_{\theta_0}(I_Y(\theta_0))V(\theta_0)^{-1}$ is the expected fraction of missing information.

Proof:
From Lemma 3 and Corollary 1 follows that the limiting stationary distribution of $\sqrt{n}(\tilde{\theta}_n-\hat{\theta}_n)$ given (almost every) y-sequence is the one corresponding to the Gaussian AR(1)-process with parameter $F(\theta_0)^t = V(\theta_0)^{-1}E_{\theta_0}I_Y(\theta_0)$ and innovation variance $V(\theta_0)^{-1}E_{\theta_0}I_Y(\theta_0)V(\theta_0)^{-1}$. This distribution is normal with expectation 0 and variance given by

$$\begin{aligned}
&\sum_{k=0}^{\infty} F(\theta_0)^{tk} V(\theta_0)^{-1} E_{\theta_0}(I_Y(\theta_0)) V(\theta_0)^{-1} F(\theta_0)^k \\
= \; & F(\theta_0)^t \sum_{k=0}^{\infty} F(\theta_0)^{tk} V(\theta_0)^{-1} F(\theta_0)^k = F(\theta_0)^t V(\theta_0)^{-1} \sum_{k=0}^{\infty} F(\theta_0)^{2k} \\
= \; & V(\theta_0)^{-1} F(\theta_0) \left(I - F(\theta_0)^2\right)^{-1} = V(\theta_0)\left(I-F(\theta_0)\right)^{-1} F(\theta_0)\left(I+F(\theta_0)\right)^{-1} \\
= \; & I(\theta_0)^{-1} F(\theta_0)\left(I+F(\theta_0)\right)^{-1} = I(\theta_0)^{-1}\left(I-(I+F(\theta_0))^{-1}\right)
\end{aligned}$$

From Lemma 1 in Schenker and Welsh (1987) follows that (unconditionally) $\sqrt{n}(\tilde{\theta}_n-\theta_0) \xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1}\left(2I-(I+F(\theta_0))^{-1}\right)\right)$. □

Remark:
The factor $(2I-(I+F(\theta_0))^{-1})$ in the expression of the variance represents the increase in variance due to the simulations. As the fraction of missing information has eigenvalues in $]0;1[$, we see that the simulations increase the eigenvalues of the variance by a factor between 1 and 3/2.

We can rewrite the asymptotic variance to give

$$V(\theta_0)^{-1}\left(I-F(\theta_0)\right)^{-1}\left(2I-(I+F(\theta_0))^{-1}\right). \tag{4}$$

Here $V(\theta_0)^{-1}$ is the complete data variance and $(I-F(\theta_0))^{-1}(2I-(I+F(\theta_0))^{-1})$ is the correction factor needed due to the imputations.

Tightness still remains to be shown. The following corollary (to Proposition 3) gives a sufficient condition for tightness.

Corollary 4 *If the assumptions of Proposition 3 holds with* $||\tilde{\theta}_n(k) - \theta_n||_2$ *replaced by* $||\sqrt{n}(\tilde{\theta}_n(k) - \theta_n)||_2$, *then* $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$ *is tight.*

Proof:
The result follows from Proposition 2 and the proof of Proposition 3. □

4 Concluding remarks

We conclude this paper by a discussion of the connections to some other algorithms (Subsections 4.1 and 4.2) and to multiple imputation ideas (Subsection 4.3). Also the estimation of the asymptotic variance (Subsection 4.4) and the necessity of the assumptions (Subsection 4.5) is discussed.

4.1 The Monte Carlo EM algorithm

The I-step estimates the expectation from the E-step in the EM algorithm by (an average of) one simulated value. Thus the I-step can be thought of as a very poor Monte Carlo integration, and the IM algorithm may be seen as a simple version of the MCEM algorithm.

The resemblance can be strengthened by improving the Monte Carlo approximation, i.e. by imputing for each observation y_i, several possible values of the missing data, $\widetilde{X}_{i,1}, \ldots, \widetilde{X}_{i,m}$, say, and maximizing the average of the m complete data log-likelihoods. Again, this leads to a complete data log-likelihood, $\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\log f_\theta(\widetilde{X}_{i,j})$, and maximization should be straightforward.

In the results of Subsection 3.2, the multiple imputations will correspond to replacing the innovation variance $V(\theta_0)^{-1}E_{\theta_0}(I_Y(\theta_0))V(\theta_0)^{-1}$ by $\frac{1}{m}V(\theta_0)^{-1}E_{\theta_0}(I_Y(\theta_0))V(\theta_0)^{-1}$ but keeping the parameter $F(\theta_0)^t$. Hence, the (unconditional) asymptotic variance of $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ becomes

$$I(\theta_0)^{-1} + \frac{1}{m}I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right)$$

Obviously, the additional variance due to imputations –the second term– decreases as m increases.

Underlying the theory of MCEM is the thought that m should be so large that the simulation variance can be ignored. Typically this is formulated as the requirement that $m/n \to \infty$. The result mentioned above is a *fixed* m result and may be used to give a finite simulation approximation to the distribution of the resulting estimator, when m is small compared to n.

4.2 The Stochastic EM algorithm

The Stochastic EM algorithm (see Diebolt and Ip (1995) and references therein) differs from the IM algorithm in the resulting estimate. The estimator in the Stochastic EM algorithm is the mean, $\tilde{E}\tilde{\theta}_n$, of the stationary distribution. For this to make sense the mean has to exist, and the bias, $\tilde{E}\tilde{\theta}_n - \hat{\theta}_n$, has to be of order $o(1/\sqrt{n})$ for this to be a reasonable estimator. It is better than the IM estimator in the sense that the simulation variance is eliminated. In general, however, the mean cannot be calculated but must be

estimated from the Markov chain. This will re-introduce a part of the simulation variance.

We will now extend the results of Section 3 to give asymptotic results for the simple estimator of the mean of $\tilde{\theta}_n$.

Suppose that the IM algorithm works in the sense that the results discussed in Section 3 are applicable. Choose a sequence of integers, $(k_n)_{n\in\mathbb{N}}$, such that

$$\sqrt{n}(\tilde{\theta}_n(k_n) - \hat{\theta}_n) \xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right)\right)$$

This is possible because of the convergence in total variation of $\tilde{\theta}_n(k)$ to $\tilde{\theta}_n$ as $k \to \infty$ implied by ergodicity. We now get

Proposition 4 *Let m be a fixed integer and assume ergodicity and tightness as in Theorem 1.*

(i) $(\sqrt{n}(\tilde{\theta}_n(k_n + j) - \hat{\theta}_n))_{j=0,\ldots,m-1}$ *converges in distribution as $n \to \infty$ to a sequence of the same length of the stationary Gaussian AR(1)-proces with parameter $F(\theta_0)^t$ and innovation variance $V(\theta_0)^{-1}E_{\theta_0}(I_Y(\theta_0))V(\theta_0)^{-1}$.*

(ii) $\sqrt{n}(\frac{1}{m}\sum_{j=0}^{m-1}\tilde{\theta}_n(k_n + j) - \hat{\theta}_n)$ *converges in distribution as $n \to \infty$ to the normal distribution with mean 0 and variance given by*

$$\begin{aligned}
&\frac{1}{m}I(\theta_0)^{-1}\left(2I - (I + F(\theta_0))^{-1}\right) \\
&+\frac{1}{m^2}\sum_{1\le i<j\le m}\Big[\left(F(\theta_0)^t\right)^{j-i} I(\theta_0)^{-1}\left(2I - (I + F(\theta_0))^{-1}\right) \\
&\qquad +I(\theta_0)^{-1}\left(2I - (I + F(\theta_0))^{-1}\right)\left(F(\theta_0))^t\right)^{j-i}\Big]
\end{aligned}$$

Proof:
The first statement of the proposition is easily shown by induction on m using assumption **C** and techniques as in the proof of Proposition 1 (see Ethier and Kurtz (1986, Chapter 4, Lemma 8.1) for details).

The second statement follows trivially. □

Proposition 4 gives asymptotic results for the naive estimator of the mean of the stationary distribution. It is worth noticing that the mean does not have to exist for this result to be valid, and that this result is a *fixed* m result.

4.3 Multiple imputations

If the imputed data is of independent interest, it may be useful to make m sets of imputations in order to be able to assess the variability due to the simulations in the I-step. This can be done in two simple ways.

Obviously, the approach of subsection 4.1 immediately gives us a set of m imputations. These imputations are independent given the observed y_is and $\tilde{\theta}$.

From the m pseudo-complete data sets, we can find m estimators of θ, $\tilde{\theta}_{n1}, \ldots, \tilde{\theta}_{nm}$, say. A natural way of combining these estimators would be to average. Using a standard argument given in detail in the appendix, we see that

$$\sqrt{n}(\frac{1}{m}\sum_{i=1}^{m}\tilde{\theta}_{ni} - \theta_0) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, I(\theta_0)^{-1} + \frac{1}{m}I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right)\right)$$

Notice that the asymptotic variance is the same as the one we get from the approach discussed in subsection 4.1. In other words, the multiple imputation IM-estimator given in subsection 4.1 is asymptotically equivalent to the simple multiple imputation estimator given here.

Alternatively, the imputations corresponding to the θ-values $(\tilde{\theta}_n(k_n+j))_{j=0,\ldots,m-1}$ from Subsection 4.2 can be used. These imputations are correlated given the observed y_is.

Asymptotic results for the average of the m estimators obtained from this type of multiple imputation was discussed in Subsection 4.2.

It should be noted that the imputations obtained from the IM algorithm are not proper in the sense of Rubin (1987).

4.4 Asymptotic variance

The asymptotic variance of $\hat{\theta}_n$ (see equation (4)) can be estimated consistently from consistent estimates of $V(\theta_0)$ and $E_{\theta_0}(I_Y(\theta_0))$.

If the complete data information is continuous, then $V(\tilde{\theta}_n)$ is consistent for $V(\theta_0)$, and by assumption (U) $\frac{1}{n}\sum_{i=1}^n I_{Y_i}(\tilde{\theta}_n)$ is consistent for $E_{\theta_0}(I_Y(\theta_0))$.

With further assumptions, $\frac{1}{n}\sum_{i=1}^n s_{\widetilde{X}_i|Y_i}(\tilde{\theta}_n)^{\otimes 2}$ may be a consistent estimator of $E_{\theta_0}(I_Y(\theta_0))$, when $\widetilde{X}_i \sim \mathcal{L}_{\tilde{\theta}_n}(X|Y=y_i)$. This will be the case if

$$|\frac{1}{n}\sum_{i=1}^n s_{\widetilde{X}_i|y_i}(\theta)^{\otimes 2} - \frac{1}{n}\sum_{i=1}^n I_{y_i}(\theta)| \xrightarrow{\tilde{P}_\theta} 0$$

uniformly in a neighbourhood of θ_0 for almost every y-sequence.

The additional variance due to the simulations is an increasing function of the fraction of missing information. Thus the increase in variance can to some extent be controlled by a sensible choice of missing data.

The results in Subsections 4.1 and 4.2 may be used to decrease the variance further. It may be noted that the approach outlined in Subsection 4.1 increases the amount of simulation far more than the approach discussed in Subsection 4.2. In the latter, the decrease in variance is clearly affected by the fraction of missing information –the more information missing, the smaller the gain of averaging–, whereas the former is less affected by the fraction of missing information, as the decrease is roughly $1/m$.

4.5 Relaxing the assumptions

The main assumption is the implicitly assumed feasibility of the two steps –the I- and the M-step– of the algorithm. This assumption may be relaxed to some extent.

The M-step may be replaced by a calculation of a suitable $\sqrt{n}$-consistent estimate (in a sense similar to (1)) and then one iteration of the method of Scoring, leading to an estimator fulfilling (1) (see Bickel et al. (1993, p. 44)).

The I-step is typically the time-consuming part of the algorithm. Thus it is of interest to perform the simulations directly and fast. In particular, running a Gibbs sampler (or another Markov chain iteration scheme) in order to perform the I-step is time-consuming. Furthermore, it should be considered if this amount of simulation might not be put to better use. For instance, the Gibbs sampler could instead be used to perform the Monte Carlo integration in the MCEM algorithm.

If simulating from the exact conditional distributions is infeasible, one may consider simulating from another distribution. This should work as long as the conclusion of Lemma 3 hold partially (the asymptotic parameters in Lemma 3 may change as long as we get the AR(1)-structure). The results in Subsection 4.1 can –when the log-likelihood is linear in the imputations– be seen as an example of how the correct conditional distributions may be replaced by a convolution of m correct conditional distributions. This does not generally simplify the simulations, though. More work is needed on this aspect of the algorithm.

The assumption of mutual equivalence of the distributions in the missing data model is rather essential. Lack of equivalence may easily create absorbing states for the Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, which will typically bias the estimates severely. Intuitively, this problem may be avoided by either restarting the Markov chain (according to some fixed distribution) when an absorbing state is reached or restricting the M-step so that absorbing states are excluded. The first approach was essentially applied by Celeux and Diebolt (1993), who gave asymptotic results for the special case of mixing proportions, where absorption is a problem.

Parameters unidentified from the observed data typically lead to either lack of equivalence or random walk like behaviour. Though lack of equivalence generally leads to serious problems (in the sense of absorption), random walk like behaviour may turn out not to be a problem. One would hope to get a limiting Markov chain, where the components corresponding to identified parameters has a normal stationary distribution. This conjecture requires more work.

The asymptotic normality and the rate of convergence of the observed data maximum likelihood estimator seems essential for asymptotic normality of the IM-estimator.

Finally, it should be noted that the proof of asymptotic normality in Celeux and Diebolt (1993) is very different in nature to the proof given in this paper, relying on moments rather than smoothness. This indicates that that the asymptotic results discussed in this paper may be proved under different –not necessarily weaker– assumptions than the smoothness assumptions applied here.

Also the Euclidean parameter may be replaced by a more general parameter. Tanner and Wong (1987) applies an algorithm, which is essentially the IM algorithm, to estimate a hazard function non-parametrically based on group survival data with some promising simulation results. The algorithm of Tanner and Wong (1987) can be seen as a frequentist Data Augmentation algorithm (cf Wei and Tanner (1991)).

4.6 Conclusion

The Imputation Maximization algorithm is a general tool for handling all sorts of missing or incomplete data. This being said, it is also clear that in many cases simpler –and often better– alternatives exist. These include the EM algorithm and simulating the entire likelihood surface.

In the EM algorithm simulation is avoided and convergence (if any) is deterministic. The IM algorithm converges stochastically and like other Markov chain methods it may be difficult to ascertain convergence. However, in many cases the E-step of the EM algorithm is not feasible, and the IM algorithm may prove a useful alternative. Also the IM algorithm has the advantages of not getting stuck in local maxima, the M-step may be simpler, and a pseudo-complete data set, which may be of independent interest, is

created.

Simulating the entire likelihood surface gives more information on the problem at hand than the point estimate given by the IM algorithm, but it will often require more simulations than the IM algorithm. Especially if the dimension of θ is large this approach may be infeasible.

In conclusion, the Imputation Maximization algorithm is a useful tool for imputation and estimation in complicated incomplete data problems.

References

BICKEL, P.J, KLAASSEN, C.A.J., RITOV, Y., AND WELLNER, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* John Hopkins University Press, Maryland.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

CELEUX, G. AND DIEBOLT, J. (1993). Asymptotic Properties of a Stochastic EM Algorithm for Estimating Mixing Proportions. *Commun. Statist.-Stochastic Models* **9**, 599-613.

DEMPSTER A.P., LAIRD N.M., AND RUBIN D.B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion). *J. Roy. Statist. Soc., B* **39**, 1-38.

DIEBOLT, J. AND IP, E.H.S. (1996). Stochastic EM: method and application. In: *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson, D.J. Spiegelhalter, eds.) Chapman & Hall, London.

ETHIER, S.N. AND KURTZ T.G. (1986). *Markov Processes. Characterization and Convergence.* Wiley, New York.

IBRAGIMOV, I.A. AND HAS'MINSKII, R.Z. (1981). *Statistical Estimation. Asymptotic Theory.* Springer, New York.

LITTLE, R.J.A AND RUBIN, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

MEYN, S.P. AND TWEEDIE, R.L. (1993). *Markov Chains and Stochastic Stability.* Springer, New York.

ROUSSAS, G.G. (1972). *Contiguity of Probability Measures.* Cambridge University Press, London.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Survey Sampling.* Wiley, New York.

SCHENKER, N. AND WELSH, A.H. (1987). Asymptotic Results for Multiple Imputation. *Ann. Statist.* **16**, 1550-1566.

TANNER, M.A. AND WONG, W. H. (1987). An Application of Imputation to an Estimation Problem in Grouped Lifetime Analysis. *Technometrics* **29**, 23-32.

WEI, G.C.G. AND TANNER, M.A. (1990) A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm. *J. Am. Statist. Assoc.* **85**, 699-704.

A Proof of the result in Subsection 4.3

Let $\tilde{\theta}_n$ be distributed according to the the stationary distribution when m imputations are performed as discussed in Subsection 4.1.

Then, given the y-sequence,

$$\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, \frac{1}{m} I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right)\right)$$

One more iteration of the IM algorithm gives us m complete data set as discussed in Subsection 4.3. From these m data sets, m estimators of θ are calculated. These estimators, $\tilde{\theta}_{n1}, \ldots, \tilde{\theta}_{nm}$, say, are iid given $\tilde{\theta}_n$ and the y-sequence.

Now choose an almost sure representation, $\tilde{\theta}'_n$, of the $\tilde{\theta}_n$-sequence such that $\tilde{\theta}_n$ and $\tilde{\theta}'_n$ are identically distributed given y for each n and

$$\sqrt{n}(\tilde{\theta}'_n - \hat{\theta}_n) \xrightarrow{a.s.} Z$$

given y, where Z has the normal distribution given above. The $\tilde{\theta}'_n$-sequence and the limit Z is defined on some measurable space $(\Omega, \mathcal{F})$.

Using $\tilde{\theta}'_n$ instead of $\tilde{\theta}_n$ for the imputation procedure, we get new estimators $\tilde{\theta}'_{n1}, \ldots, \tilde{\theta}'_{nm}$, say, which are iid given $\tilde{\theta}'_n$ and the y-sequence. Also, given y,

$$\left(\tilde{\theta}_{n1}, \ldots, \tilde{\theta}_{nm}\right) \overset{\mathcal{D}}{=} \left(\tilde{\theta}'_{n1}, \ldots, \tilde{\theta}'_{nm}\right)$$

since obviously these two random vectors are identically distributed given $\tilde{\theta}_n$ and $\tilde{\theta}'_n$, respectively, and given y.

Due to the almost sure representation, we can apply Lemma 3 to each of the new estimators, $\tilde{\theta}'_{nj}$. This tells us that, conditionally on y and for almost every $\omega \in \Omega$,

$$\sqrt{n}(\tilde{\theta}'_{nj} - \hat{\theta}_n) \xrightarrow{\mathcal{D}} N\left(F(\theta_0)^t Z(\omega), V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}\right)$$

leading to

$$\sqrt{n}(\frac{1}{m}\sum_{j=1}^{m} \tilde{\theta}'_{nj} - \hat{\theta}_n) - F(\theta_0)^t \sqrt{n}(\tilde{\theta}'_n(\omega) - \hat{\theta}_n) \xrightarrow{\mathcal{D}} N\left(0, \frac{1}{m} V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}\right)$$

conditionally on y for almost every $\omega \in \Omega$. Hence conditionally on y only, we find, using Lemma 1 in Schenker and Welsh (1987), that

$$\begin{aligned}
\sqrt{n}(\frac{1}{m}\sum_{j=1}^{m} \tilde{\theta}'_{nj} - \hat{\theta}_n) &= \sqrt{n}(\frac{1}{m}\sum_{j=1}^{m} \tilde{\theta}'_{nj} - \hat{\theta}_n) - F(\theta_0)^t \sqrt{n}(\tilde{\theta}'_n - \hat{\theta}_n) + F(\theta_0)^t \sqrt{n}(\tilde{\theta}'_n - \hat{\theta}_n) \\
&\xrightarrow{\mathcal{D}} N\left(0, F(\theta_0)^t \frac{1}{m} I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right) F(\theta_0) + \frac{1}{m} V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}\right) \\
&= N\left(0, \frac{1}{m} I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right)\right)
\end{aligned}$$

and consequently

$$\begin{aligned}
\sqrt{n}(\frac{1}{m}\sum_{j=1}^{m} \tilde{\theta}_{nj} - \theta_0) &\overset{\mathcal{D}}{=} \sqrt{n}(\frac{1}{m}\sum_{j=1}^{m} \tilde{\theta}'_{nj} - \theta_0) = \sqrt{n}(\frac{1}{m}\sum_{j=1}^{m} \tilde{\theta}'_{nj} - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0) \\
&\xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1} + \frac{1}{m} I(\theta_0)^{-1}\left(I - (I + F(\theta_0))^{-1}\right)\right)
\end{aligned}$$

Preprints 1996

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR
OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS,
UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK.
TELEPHONE 45 35 32 08 99, FAX 45 35 32 07 72.

No. 1 Lando, David: Modelling Bonds and Derivatives with Default Risk.

No. 2 Johansen, Søren: Statistical analysis of some non-stationary time series.

No. 3 Jørgensen, C., Kongsted, H.C. and Rahbek, A.: Trend-Stationarity in the I(2) Cointegration Model.

No. 4 Paruolo, Paolo and Rahbek, Anders C.: Weak Exogeneity in I(2) Systems.

No. 5 Christensen, Peter Ove, Lando, David and Miltersen, Kristian R.: State-Dependent Realignments in Target Zone Currency Regimes.

Preprints 1997

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR
OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS,
UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK.
TELEPHONE 45 35 32 08 99, FAX 45 35 32 07 72.

No. 1 Nielsen, Søren Feodor: Coarsening at Random.
No. 2 Nielsen, Søren Feodor: The Imputation Maximization Algorithm.