On simulated EM algorithms

Søren Feodor Nielsen
Department of Theoretical Statistics
University of Copenhagen

Abstract

In a paper on the EM algorithm, Ruud (1991) suggests a simulated EM algorithm. The idea is to replace the E-step by a simulated E-step, i.e. to replace the expectation by an estimate obtained by simulation.

The simulations can –at least in principle– be done in two ways. Either new independent random variables are drawn in each iteration, or the same uniforms are reused in each iteration.

In this paper the properties of these two versions of the simulated EM algorithm are discussed and compared.

Keywords: Simulation, EM algorithm

1 Simulated EM algorithms

The EM algorithm (cf Dempster, Laird, and Rubin (1978)) has two steps, an E-step and an M-step. The E-step is the calculation of the conditional expectation of the complete data log-likelihood given the observed data. The M-step is a maximization of this expression. These two steps are then iterated. It is well-known that each iteration of the algorithm increases the observed data log-likelihood, and though the general convergence theory (cf Wu (1983)) is rather vague, the algorithm often works well in practice. Ruud (1991) gives an overview of the general theory and some applications to econometrics.

However in some cases the E-step of the algorithm is not practically feasible, because the conditional expectation cannot be calculated. This happens when the expectation is a large sum or when the expectation corresponds to a high-dimensional integral without a closed form expression. In this case, Ruud (1991) suggests to replace the expectation by an estimate obtained by simulation.

Let $X_1, X_2, \ldots, X_n$ be iid random variables with density f_θ. Instead of observing X_i suppose we observe $Y_i = Y(X_i)$. Then the E-step of the EM-algorithm is the calculation of

$$\theta \to Q(\theta|\theta') = \frac{1}{n}\sum_{i=1}^{n} E_{\theta'}(\log f_\theta(X_i)|Y_i) \tag{1}$$

This is then in the M-step maximized over θ and the maximizer is then used as a new value of θ' in the following iteration. One iteration of the EM algorithm thus corresponds to calculating the conditional expectation (1) and maximizing it as a function of θ.

We denote the EM update, i.e. the θ-value given by one iteration of the EM algorithm starting in θ', by $M(\theta')$. The maximum likelihood estimator, $\widehat{\theta}_n$, is a fixed point of M. Hence, the EM algorithm finds solutions to $M(\theta) = \theta$ by the method of successive substitution; from a starting value, θ_1, of θ, $\theta_2 = M(\theta_1)$ is calculated and used to calculate $\theta_3 = M(\theta_2)$ etc.

In the simulated EM algorithm the expectation (1) is replaced by an estimate in the following way: Let $\widetilde{X}_{ij}$ for $j = 1, \ldots, m$ be a random variable drawn from the conditional distribution of X_i given Y_i under the distribution with parameter θ'. Then

$$\tilde{Q}(\theta|\theta') = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(\widetilde{X}_{ij}) \tag{2}$$

is an unbiased estimate of Q. We can pretend that $\widetilde{X}_{ij} = F^{-1}_{\theta'}(U_{ij}|Y_i)$ where the U_{ij}s are independent uniform random variables, independent of the Y_is, and $F^{-1}_{\theta'}(\cdot|Y_i)$ is the appropriate conditional distribution function.

It is clear that we get two different simulated EM algorithms according to whether we draw new independent random variables in each iteration or we re-use the uniforms, U_{ij}, in each iteration.

Drawing new uniforms in each iteration, the sequence of θ-values obtained from the algorithm, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}}$, is a Markov chain, and the estimator, $\tilde{\theta}_n$ is a random variable drawn from the limiting distribution of the chain. This version has been discussed in detail by Nielsen (1997) under the name of the Imputation Maximization algorithm and is closely connected to the Stochastic EM algorithm suggested by Celeux and Diebolt (1985); see Diebolt and Ip (1996) for a review.

Reusing the uniforms, we estimate the function $\theta \to M(\theta)$ once and for all and search for fixed points by the method of successive substitutions. Essentially this corresponds to

estimating

$$\theta' \to D_\theta Q(\theta|\theta')_{|\theta=\theta'}$$

by

$$G_n(\theta') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} D_\theta \log f_\theta(F_{\theta'}^{-1}(U_{ij}|Y_i))_{|\theta=\theta'}$$

and finding the root by the method of successive substitutions. Notice that $G_n(\theta')$ is an unbiased estimate of $D_\theta Q(\theta|\theta')_{|\theta=\theta'}$ if differentiation and expectation are interchangeable.

It should be clear that at least for moderate values of m the versions differ significantly. The purpose of this paper is to compare these two simulated EM algorithms.

To keep the distinction between these two versions of the algorithm clear we will use the name IM algorithm for the version, where new independent random variables are drawn in each iteration, and call the second version, where the uniforms are reused, the SEM algorithm as done by McFadden and Ruud (1984). One should note that the supplementary EM algorithm (cf Meng and Rubin (1991)) is also referred to as the SEM algorithm, and so is the Stochastic EM algorithm (cf Celeux and Diebolt (1985)). Thus SEM is not a perfectly chosen acronym, but it should not cause any confusion in the context of this paper.

2 Asymptotic results

We begin this section by introducing some notation. All score functions to be defined below are assumed to have expectation zero and finite variance.

Let f_θ be the density of X. Let $s_x(\theta)$ be the corresponding score function and put $V(\theta) = E_\theta(s_X(\theta)^{\otimes 2})$. Let θ_0 denote the true unknown value of θ.

The score function corresponding to the conditional distribution of X given $Y = y$ is denoted $s_{x|y}(\theta)$. Let $I_y(\theta) = E_\theta(s_{X|y}(\theta)^{\otimes 2}|Y = y)$.

Let $s_y(\theta)$ be the score function corresponding to the distribution of Y and put $I(\theta) = E_\theta(s_Y(\theta)^{\otimes 2})$.

Notice that $s_y(\theta) = s_x(\theta) - s_{x|y}(\theta)$ and that $I(\theta) = V(\theta) - E_\theta I_Y(\theta)$. Finally, put $F(\theta) = E_\theta I_Y(\theta)V(\theta)^{-1}$, which can be interpreted as the expected fraction of missing information (cf Dempster et al. (1977)).

Asymptotic results for the IM algorithm were discussed by Nielsen (1997, Theorem 1). He proved the following result under general regularity conditions:

Theorem 1 (Asymptotic results for the IM algorithm) *If the Markov chain is ergodic and tight, then* $\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma_m(\theta_0))$ *where*

$$\begin{aligned}\Sigma_m(\theta_0) &= I(\theta_0)^{-1} + \frac{1}{m}\sum_{k=0}^{\infty} F(\theta_0)^{tk} V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1} F(\theta_0)^k \\ &= I(\theta_0)^{-1} + \frac{1}{m} V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1} \left(I - F(\theta_0)^2\right)^{-1} \qquad (3)\end{aligned}$$

where I is the identity matrix.

See Nielsen (1997) for a precise statement of the regularity conditions assumed and for a proof of (3).

In order to obtain large sample results for the SEM version, we apply Corollary 3.2 and Theorem 3.3 in Pakes and Pollard (1989). McFadden and Ruud (1994) give similar results under stronger assumptions.

We note that

$$\begin{aligned} G(\theta) &= E_{\theta_0}(G_n(\theta)) = E_{\theta_0}\left(D_\theta \log f_\theta(\widetilde{X}_{ij})\right) = E_{\theta_0}E_{\theta_0}\left(D_\theta \log f_\theta(\widetilde{X}_{ij})|Y_i\right) \\ &= E_{\theta_0}E_\theta(s_X(\theta)|Y) = E_{\theta_0}\left(s_\theta(Y) + E_\theta(s_{X|Y}(\theta)|Y)\right) \\ &= E_{\theta_0}(s_Y(\theta)) \end{aligned}$$

which is differentiable under usual regularity conditions with

$$D_\theta G(\theta)_{|\theta=\theta_0} = I(\theta_0) \tag{4}$$

Furthermore, by the central limit theorem

$$\sqrt{n}G_n(\theta_0) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, I(\theta_0) + \frac{1}{m}E_{\theta_0}I_Y(\theta_0)\right) \tag{5}$$

By assumption $G(\theta_0) = 0$ and we will assume that there is a neighbourhood of θ_0 such that θ_0 is the only root of G inside this neighbourhood.

From the law of large numbers, we know that $G_n(\theta) \xrightarrow{P} G(\theta)$ for all θ. This needs to be extended to uniform convergence to obtain consistency. Also $\sqrt{n}(G_n(\theta) - G(\theta))$ is asymptotically normal by the central limit theorem; we need this to be close to $\sqrt{n}G_n(\theta_0)$ when θ is close to θ_0. To be precise we need these two assumptions:

(A1) $\sup_{\theta\in C} \frac{\|G_n(\theta) - G(\theta)\|}{1 + \|G_n(\theta)\| + \|G(\theta)\|} \xrightarrow{P} 0$, for a compact neighbourhood, C, of θ_0.

(A2) $\sup_{\|\theta-\theta_0\|<\delta_n} \frac{\sqrt{n}\|G_n(\theta) - G(\theta) - G_n(\theta_0)\|}{1 + \sqrt{n}\|G_n(\theta)\| + \sqrt{n}\|G(\theta)\|} \xrightarrow{P} 0$, for any $\delta_n \to 0$.

Stronger assumptions are obtained if the numerators are ignored. A sufficient condition for both assumptions is given in this lemma:

Lemma 1 *Suppose that in a neighbourhood of θ_0 for some $\alpha > 0$*

$$|\log f_\theta(F_\theta^{-1}(U|Y)) - \log f_{\theta'}(F_{\theta'}^{-1}(U|Y))| \leq \psi(U,Y)\|\theta - \theta'\|^\alpha$$

(i) If $E_{\theta_0}|\psi(U,Y)| < \infty$ then (A1) holds.

(ii) If $E_{\theta_0}\psi(U,Y)^2 < \infty$ then (A2) holds.

Proof:
See Lemma 2.13, Lemma 2.8 and Lemma 2.17 in Pakes and Pollard (1989). □

We say that $\tilde{\theta}_n$ is an asymptotic local minimum of $\|G_n(\theta)\|$ if for some open set $W \subseteq \Theta$ $\|G_n(\tilde{\theta}_n)\| \leq \inf_{\theta\in W}\|G_n(\theta)\| + o_P(n^{-1/2})$. From Corollary 3.2 and Theorem 3.3 in Pakes and Pollard (1989) we get the following result:

Theorem 2 (Asymptotic results for the SEM algorithm) *Under assumption (A1), there is an asymptotic local minimum, $\tilde{\theta}_n$, of $\|G_n(\theta)\|$ such that $\tilde{\theta}_n \xrightarrow{P} \theta_0$.*

Under assumption (A2), if $\tilde{\theta}_n$ is consistent, then
$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1} + \frac{1}{m}I(\theta_0)^{-1}E_{\theta_0}I_Y(\theta_0)I(\theta_0)^{-1}\right)$

3 A comparison

It is clear that both the SEM and the IM version of the simulated EM algorithm approximates the EM algorithm as m tends to infinity. Both versions lead to estimators with asymptotic variances tending to $I(\theta_0)^{-1}$ as $m \to \infty$. For finite values of m, however, the asymptotic variances differ.

Recalling that $F(\theta_0) = E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$ and $V(\theta_0) = I(\theta_0) + E_{\theta_0} I_Y(\theta_0)$ we find (with the usual ordering of positive definite matrices) that

$$
\begin{aligned}
& I(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) I(\theta_0)^{-1} \\
& \quad > V(\theta_0)^{-1} E_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1} \left(I - F(\theta_0)^2\right)^{-1} \\
& \Updownarrow \\
& \left(I - F(\theta_0)^2\right) V(\theta_0) E_{\theta_0} I_Y(\theta_0)^{-1} V(\theta_0) \\
& \quad > I(\theta_0) E_{\theta_0} I_Y(\theta_0)^{-1} I(\theta_0) \\
& \Updownarrow \\
& V(\theta_0) E_{\theta_0} I_Y(\theta_0)^{-1} V(\theta_0) - E_{\theta_0} I_Y(\theta_0) > I(\theta_0) E_{\theta_0} I_Y(\theta_0)^{-1} I(\theta_0) \\
& \Updownarrow \\
& E_{\theta_0} I_Y(\theta_0) + I(\theta_0) E_{\theta_0} I_Y(\theta_0)^{-1} I(\theta_0) + 2I(\theta_0) - E_{\theta_0} I_Y(\theta_0) \\
& \quad > I(\theta_0) E_{\theta_0} I_Y(\theta_0)^{-1} I(\theta_0) \\
& \Updownarrow \\
& 2I(\theta_0) > 0
\end{aligned}
$$

Hence, the IM version leads to asymptotically better estimates –in the sense of smaller variance– than the SEM version.

This result may seem counterintuitive as new random noise is added in each step of the IM algorithm. The following example suggests an explanation.

Example 1 *Let $X_1, X_2, \ldots, X_n$ be iid bivariate random variables, normally distributed with common expectation θ and known variance matrix Σ. Suppose only the first coordinate, X_{1i}, of each $X_i = (X_{1i}, X_{2i})^t$ is observed. The observed data MLE, $\hat{\theta}_n$, is just the average of the X_{1i}s.*

The simulated EM algorithm (with $m = 1$) corresponds to simulating X_{2i} from the conditional distribution of X_{2i} given X_{1i} and then averaging all the Xs, the observed as well as the simulated. This leads to

$$\sqrt{n}\left(\tilde{\theta}_n(k+1) - \hat{\theta}_n\right) = \frac{1-\rho}{2}\sqrt{n}\left(\tilde{\theta}_n(k) - \hat{\theta}_n\right) + \varepsilon_k \tag{6}$$

where ρ is the correlation coefficient of X_i and $\varepsilon_k \sim N(0, \sigma^2)$; σ^2 is given from the known matrix Σ but will not be specified here.

In the SEM case, $\varepsilon_k = \varepsilon_1$ for all k and as $k \to \infty$

$$\sqrt{n}\left(\tilde{\theta}_n(k+1) - \hat{\theta}_n\right) \longrightarrow \frac{2}{1+\rho}\varepsilon_1 \sim N\left(0, \frac{4}{(1+\rho)^2} \cdot \sigma^2\right) \tag{7}$$

whereas in the IM case (6) defines an AR(1) process, so that

$$\sqrt{n}\left(\tilde{\theta}_n(k+1) - \hat{\theta}_n\right) \xrightarrow{\mathcal{D}} N\left(0, \left(1 - \left(\frac{1-\rho}{2}\right)^2\right)^{-1} \cdot \sigma^2\right) \tag{8}$$

when $k \to \infty$.
The (asymptotic) distribution of $\sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0)$ in the SEM and the IM case is thus normal with expectation 0 and variance given by the variance of the MLE plus the variance specified above (in (7) and (8), respectively).
It is not difficult to show directly that the variance of the IM estimator is smaller than the variance of the SEM estimator. Instead, looking at (6), we observe that

$$Var\left(\sqrt{n}\left(\tilde{\theta}_n(k+1) - \hat{\theta}_n\right)\right) = \frac{(1-\rho)^2}{4} Var\left(\sqrt{n}\left(\tilde{\theta}_n(k) - \hat{\theta}_n\right)\right) + \sigma^2 + 2Cov\left(\varepsilon_k, \sqrt{n}\left(\tilde{\theta}_n(k) - \hat{\theta}_n\right)\right)\frac{1-\rho}{2}$$

For the IM version the covariance term is 0, whereas it is positive for the SEM version. Thus, we see that the variance of the SEM estimator is larger exactly because "the uniforms are reused".

Furthermore, the IM estimator can easily be improved –in the sense of reducing the variance– by averaging over the last 5-20, say, iterations. Since the iterations constitute a Markov chain, averaging will decrease the variance (see Nielsen (1997) for details). Obviously the only way to improve the SEM estimator is to increase m, i.e. to increase the simulation burden.

The SEM algorithm is a deterministic algorithm searching for a root of the (random) function $G(\theta)$. Hence, if it converges, it converges deterministically. Little is known about the convergence of the SEM algorithm, though. There is no apparent reason to believe that $G_n(\theta)$ only has one root, and the SEM algorithm stops when one is found. There may be no root at all as in the example below.

Example 2 *Let $X_1, X_2, \ldots, X_n$ be iid normally distributed with expectation 0 and variance θ. Observe X_i if $X_i \geq 0$. The SEM algorithm (with $m = 1$) corresponds to iterating*

$$\tilde{\theta}_n(k+1) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 \cdot 1_{\{X_i \geq 0\}} + \tilde{\theta}_n(k) \cdot \varepsilon_i^2 \cdot 1_{\{X_i < 0\}}\right)$$

where ε_i are iid random variables from the standard normal distribution given that they are negative.
If $\sum_{i=1}^{n} \varepsilon_i^2 \cdot 1_{\{X_i<0\}} \geq n$ (which happens with positive probability), then $\tilde{\theta}_n(k) \to \infty$. If $\sum_{i=1}^{n} \varepsilon_i^2 \cdot 1_{\{X_i<0\}} < n$ then $\tilde{\theta}_n(k)$ converges to $\sum_{i=1}^{n} X_i^2 \cdot 1_{\{X_i \geq 0\}}/(n - \sum_{i=1}^{n} \varepsilon_i^2 \cdot 1_{\{X_i<0\}})$ which may be arbitrarily far from the maximum likelihood estimator; the MLE is obtained when $\sum_{i=1}^{n} \varepsilon_i^2 \cdot 1_{\{X_i<0\}} = \sum_{i=1}^{n} 1_{\{X_i<0\}}$.

Using a contraction principle (cf Letac (1986)), it is not difficult to show that the Markov chain of the IM algorithm is ergodic so that the IM algorithm converges.

Assumptions (A1) and (A2) (in the SEM case) and tightness (in the IM case) can be shown to hold, but the details will not be given here. Consequently, the asymptotic results of Section 2 hold.

The Markov chain of the IM algorithm is irreducible under weak assumptions, and the IM algorithm does therefore not get stuck. If it converges, i.e. if the Markov chain is ergodic, the IM algorithm converges stochastically in the sense that the sequence of distributions of the θ-values obtained from the iterations converge in total variation. Thus, whereas it is a relatively simple task to see if the SEM algorithm has converged, the convergence of

the IM algorithm is much more awkward to ascertain.

It is not clear which algorithm converge faster. As indicated above, convergence of the IM algorithm is difficult to ascertain and this will typically lead to iterations after the algorithm has converged. The method of successive substitutions, which is applied in the SEM as well as in the EM algorithm, is typically slow, as is well-known in the case of the EM algorithm.

Finally, a word on the simulations. Using acceptance-rejection sampling or Markov chain Monte Carlo techniques, we can generally simulate the $\widetilde{X}_{ij}$s. However, neither technique will give us simulations suitable for the SEM version of the simulated EM algorithm, as neither method "re-uses uniforms" even if we start each iteration with the same random seed. Thus, unless we can simulate the $\widetilde{X}_{ij}$s in a non-iterative manner, the SEM algorithm is not really implementable. This problem does not occur for the IM algorithm, though it should be noted that if a MCMC simulation of the $\widetilde{X}_{ij}$s is necessary, then –for instance– a Monte Carlo EM algorithm can be performed with the same amount of simulation, and this may well be preferable.

In conclusion, the IM algorithm has some serious theoretical advantages compared to the SEM algorithm. The asymptotic variance is smaller and the algorithm does not get stuck. On the other hand, convergence is more difficult to detect in the IM algorithm than in the SEM algorithm, and we conjecture that this may make the latter algorithm faster. Notice however that the SEM algorithm is not always implementable.

In order to apply either algorithm, some assumptions must be checked. For the SEM algorithm we need (A1) and (A2). This can typically be shown from smoothness of the simulations, if the simulations are indeed smooth. If not, more general empirical process techniques can be applied (cf Pakes and Pollard (1989)). In the IM case, it is ergodicity and tightness that must be shown, typically by verifying a drift criterion; see Nielsen (1997).

4 Monte Carlo experiment

We conclude this paper with a small simulation study. McFadden and Ruud (1994) consider the following trivariate tobit model.

Let

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta t_1 \\ \beta t_2 \\ \beta t_3 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \rho^2 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

We observe $Y_i = X_i 1_{\{X_i > 0\}}$. The observed covariates, t_i, are simulated as iid normal variates with mean 1 and variance 2. We simulate 50 replications of X with $\beta = 1$, $\rho = 0.7$ and $\sigma^2 = 1.51$. The parameter σ^2 is the conditional variance of X_1 given (X_2, X_3). This parameterization is chosen because it makes the parameters variation independent and, consequently, the results easier to interpret.

To investigate the behaviour of the two versions of the simulated EM algorithm we have obtained 1000 estimates from each algorithm. The IM algorithm was run for 16,000 iterations and the last 1000 iterations were used. The SEM version was run 1000 times. We do this for $m = 1, 5, 10$.

The complete data MLE does not have a closed form and the M-step must be performed

numerically. We have used the method of scoring with a fixed tolerance.

The simulated E-step involves drawing values of (X_1, X_2) given that they are both negative. This cannot be done in a non-iterative manner, and as mentioned above this means that the SEM algorithm is not really implementable. As McFadden and Ruud (1994), we have run a Gibbs sampler for 10 cycles to do this simulation in the SEM version. In the IM version, this simulation is done by acceptance-rejection sampling.

The distribution of the simulation estimators, $\tilde{\theta} - \theta_0$, obtained from the SEM and the IM algorithms can be decomposed into a contribution from the simulations and a contribution from the data, $\tilde{\theta} - \hat{\theta}$ and $\hat{\theta} - \theta_0$, respectively. The latter –the distribution of the MLE– is a function of the data only and gives no information on how well the two simulation estimators behave. The simulation estimators, on the other hand, attempts to approximate the maximum likelihood estimator (cf. Example 1), since the simulated EM algorithm approximates the EM algorithm. Therefore, it is the distribution of $\tilde{\theta} - \hat{\theta}$ that is of interest when evaluating the performance of the simulation estimators. Hence, we look at the empirical distribution of the difference between the simulation estimators, $\tilde{\theta}$, obtained from the SEM and the IM algorithms and the maximum likelihood estimator, $\hat{\theta}$.

We give summary statistics for the estimators of each parameter (Tables 1–3), and histograms and QQ-plots of the distribution of $\tilde{\theta} - \hat{\theta}$ for each parameter and each value of m are shown. The two histograms in each figure –one for each version of the simulated EM algorithm– have the same axes and are thus directly comparable. The displayed part of the abscissae of the histograms all contain 0. The MLE based on the observed data is shown in the tables for completeness.

Looking first at the results for β (Table 1 and Figures 1–3) we notice that the variance of the IM estimator is larger than the variance of the SEM estimator. This is probably due to the small sample size, as the results of Section 2 indicate that the asymptotic variance is smaller for the IM estimator.

The bias of the SEM estimator is however large compared to the bias of the IM estimator, and the IM estimator clearly performs better in this example in spite of the larger variance. The fact that the bias of the IM estimator in the case $m = 5$ is larger than when $m = 1$ or $m = 10$ is surprising but also occurs in additional simulations (not shown here).

The QQ-plots suggests that the estimators are reasonably close to normally distributed. The tails of the distributions may be a bit too heavy, and the IM estimator seems to be slightly closer to normal than the SEM estimator.

$\beta : (\hat{\beta} = 1.042)$		Mean	StDev	25%	Median	75%
m=1:	IM	-0.0117	0.0196	-0.0249	-0.0115	0.0013
	SEM	-0.0798	0.0122	-0.0878	-0.0808	-0.0724
m=5:	IM	-0.0635	0.0104	-0.0702	-0.0637	-0.0565
	SEM	-0.0798	0.0055	-0.0838	-0.0799	-0.0762
m=10:	IM	-0.0118	0.0065	-0.0162	-0.0117	-0.0074
	SEM	-0.0797	0.0038	-0.0822	-0.0798	-0.0772

Table 1: Simulation results for $\tilde{\beta} - \hat{\beta}$.
StDev is the standard deviation, 25% and 75% are the lower and upper quartiles

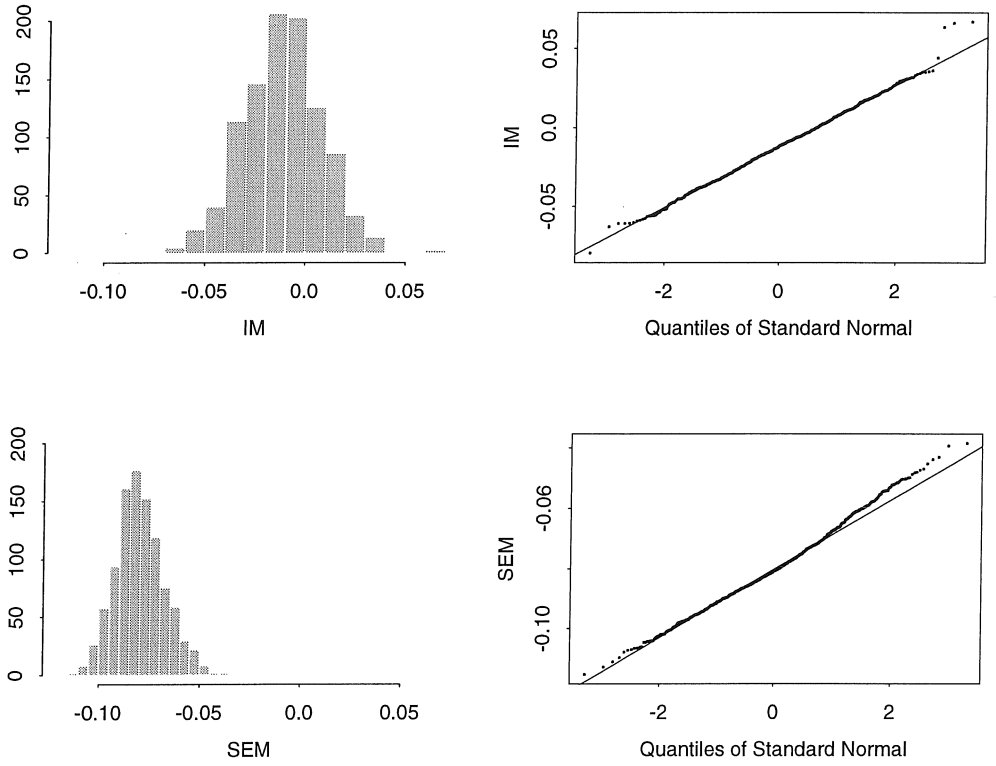

Figure 1: Histograms and QQ-plots for $\tilde{\beta} - \hat{\beta}$, $m = 1$

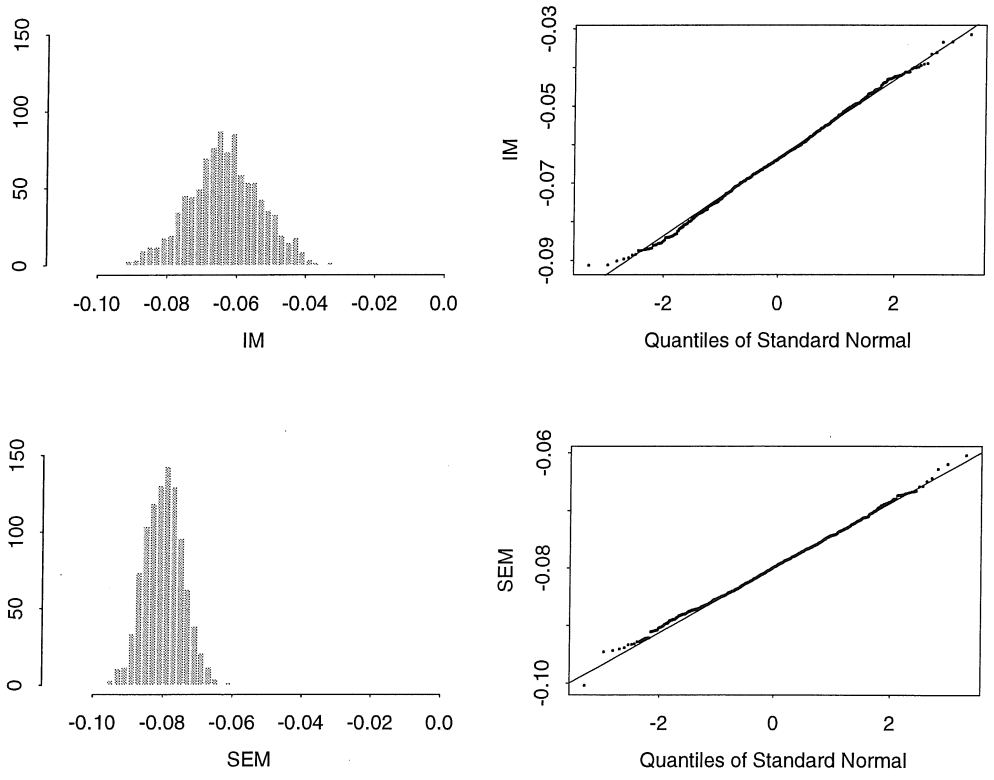

Figure 2: Histograms and QQ-plots for $\tilde{\beta} - \hat{\beta}$, $m = 5$

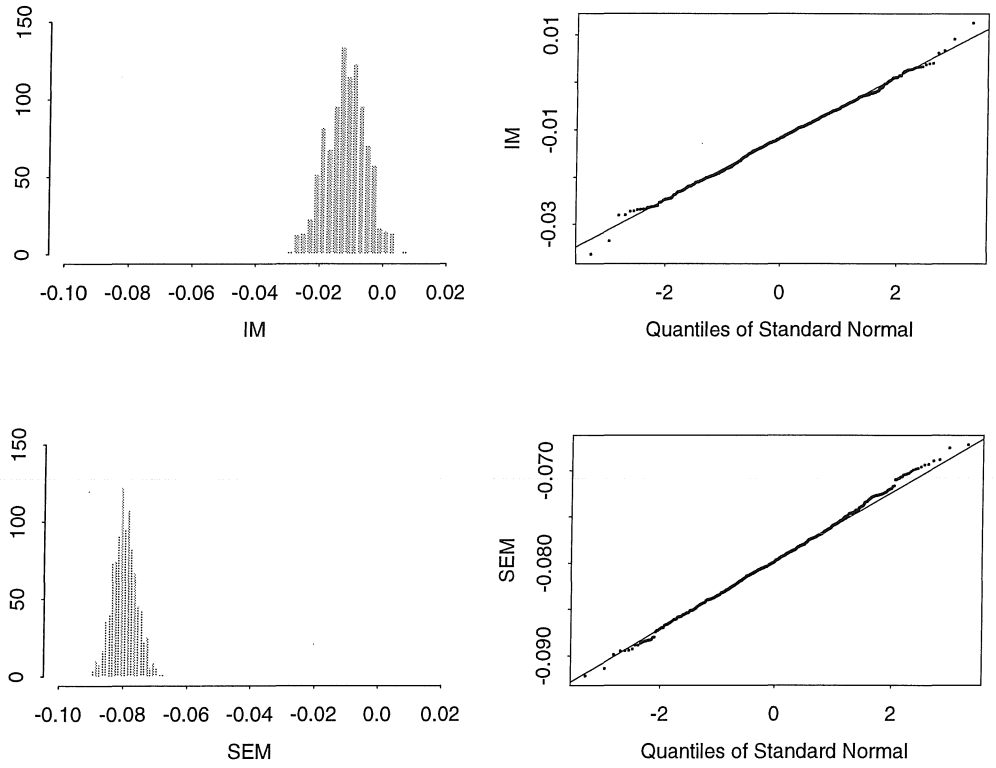

Figure 3: Histograms and QQ-plots for $\tilde{\beta} - \hat{\beta}$, $m = 10$

For ρ we see that the bias is significantly larger for the SEM estimator than for the IM estimator, but the variances are now roughly equal. Again the IM estimator is superior. The Figures 4–6 suggest that the estimators are not too far from normally distributed.

We may note that the SEM estimator is less biased than the IM estimator when looking at the empirical distribution of $\tilde{\rho} - \rho_0$ rather than $\tilde{\rho} - \hat{\rho}$. This is due to the large positive bias in the MLE and is thus an effect of the data and not the simulations. Consequently, we would not expect the SEM estimator to perform better than the IM estimator on all data sets, though it *may* perform better when $\hat{\rho}$ is positively biased. In particular, in large samples, where we would expect the bias of the MLE to be negligible, we would expect the IM estimator to be superior. Incidentally, this is the only parameter, θ, for which $\tilde{\theta} - \theta_0$ is more biased in the IM case.

$\rho : (\hat{\rho} = 1.101)$		Mean	StDev	25%	Median	75%
m=1:	IM	0.0386	0.0977	-0.0261	0.0417	0.1061
	SEM	-0.2122	0.0917	-0.2733	-0.2118	-0.1527
m=5:	IM	0.0276	0.0408	-0.0003	0.0269	0.0553
	SEM	-0.2179	0.0393	-0.2444	-0.2168	-0.1908
m=10:	IM	0.0372	0.0318	0.0148	0.0371	0.0595
	SEM	-0.2185	0.0277	-0.2361	-0.2180	-0.1996

Table 2: Simulation results for $\tilde{\rho} - \hat{\rho}$.
StDev is the standard deviation,
25% and 75% are the lower and upper quartiles

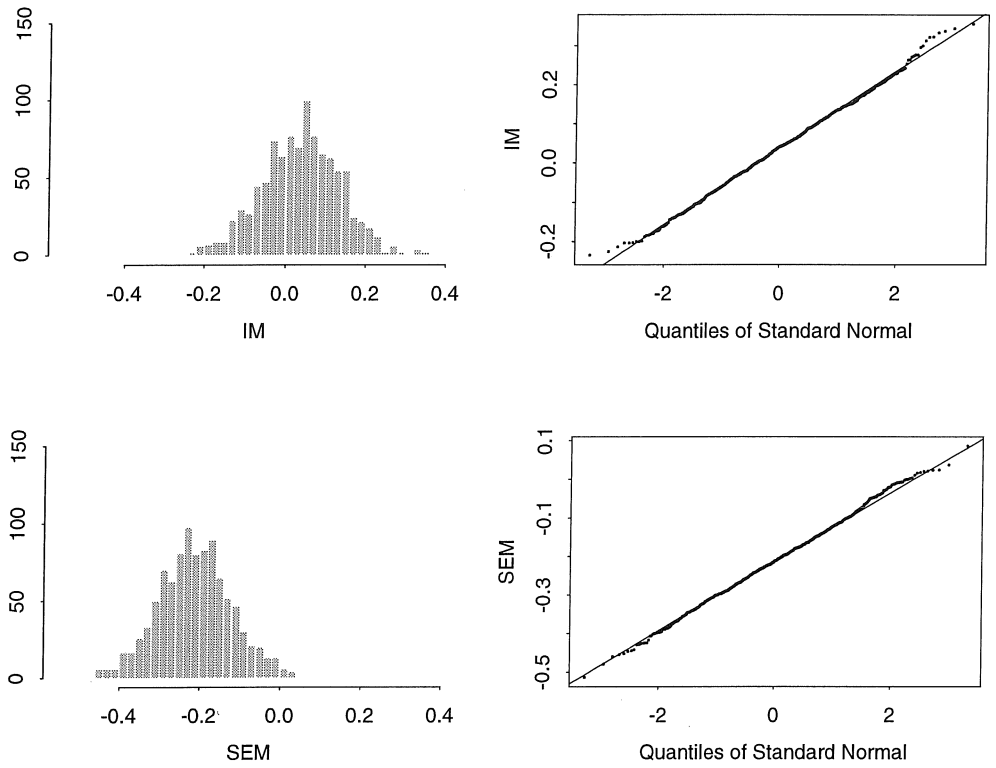

Figure 4: Histograms and QQ-plots for $\tilde{\rho} - \hat{\rho}$, $m = 1$

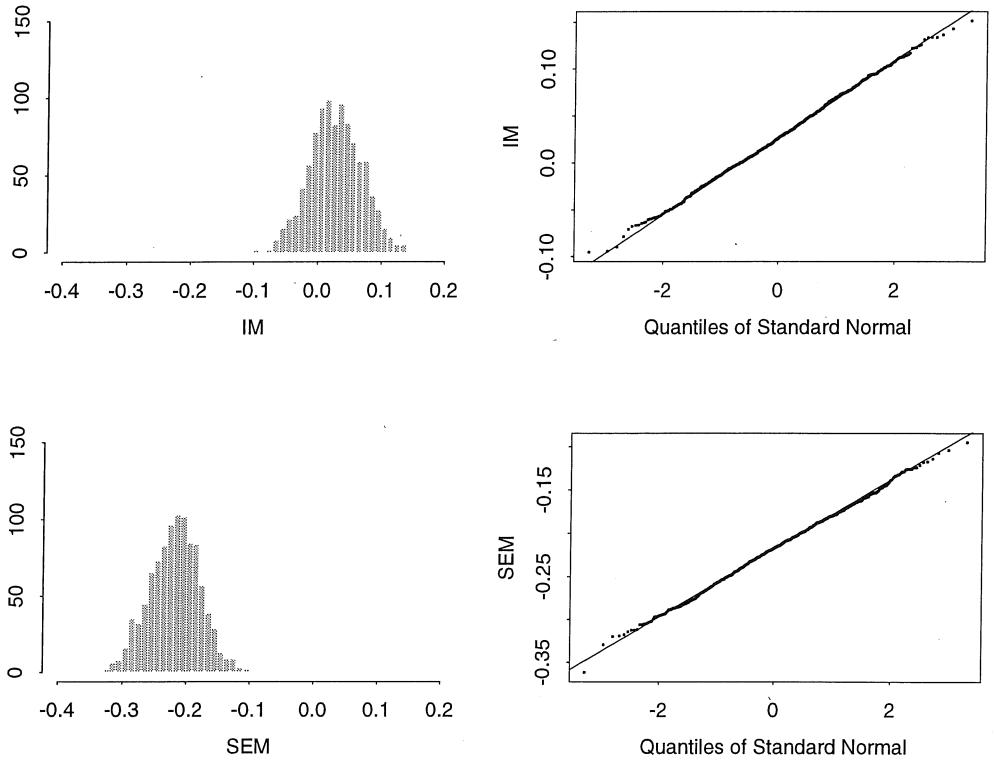

Figure 5: Histograms and QQ-plots for $\tilde{\rho} - \hat{\rho}$, $m = 5$

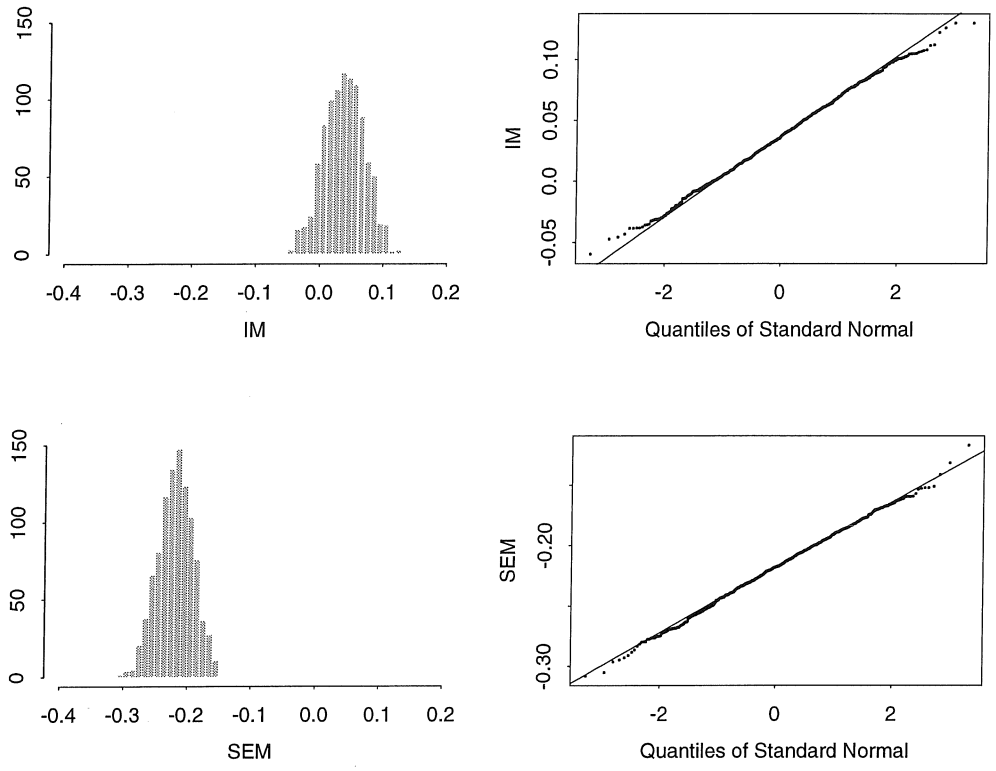

Figure 6: Histograms and QQ-plots for $\tilde{\rho} - \hat{\rho}$, $m = 10$

When estimating σ^2, the IM estimator again performs better; the bias is larger for the SEM estimator, and the variances are here smaller for the IM estimator.

Figures 7–9 shows that the distribution of $\tilde{\sigma}^2 - \hat{\sigma}^2$ is far from normal, especially for small values of m, but this is what we would expect. Ignoring the multivariate nature of the data and the fact that σ^2 is not the only parameter, the problem of estimating σ^2 is very similar to Example 2. Thus, we would expect the SEM estimator to be distributed approximately as $a/(50 - \chi^2_{12m})$ for some constant a, since only 12 X_1s are censored. In particular, $\tilde{\sigma}^2 - \hat{\sigma}^2$ is not approximately normal for small values of m in the SEM case. The distribution of the IM estimator is more difficult to describe but since the conditional distribution of the next step of the Markov chain in Example 2 given the past is an affine transformation of a χ^2_{12m}-distribution, we would not expect $\tilde{\sigma}^2 - \hat{\sigma}^2$ to be approximately normal for small values of m in the IM case, either.

$\sigma^2 : (\hat{\sigma}^2 = 1.538)$		Mean	StDev	25%	Median	75%
m=1:	IM	-0.1198	0.1594	-0.2287	-0.1347	-0.0291
	SEM	0.2712	0.1927	0.1390	0.2526	0.3787
m=5:	IM	-0.1703	0.0623	-0.2112	-0.1750	-0.1324
	SEM	0.2792	0.0836	0.2227	0.2772	0.3318
m=10:	IM	-0.1052	0.0523	-0.1424	-0.1066	-0.0712
	SEM	0.2809	0.0584	0.2402	0.2799	0.3216

Table 3: Simulation results for $\tilde{\sigma}^2 - \hat{\sigma}^2$.
StDev is the standard deviation,
25% and 75% are the lower and upper quartiles

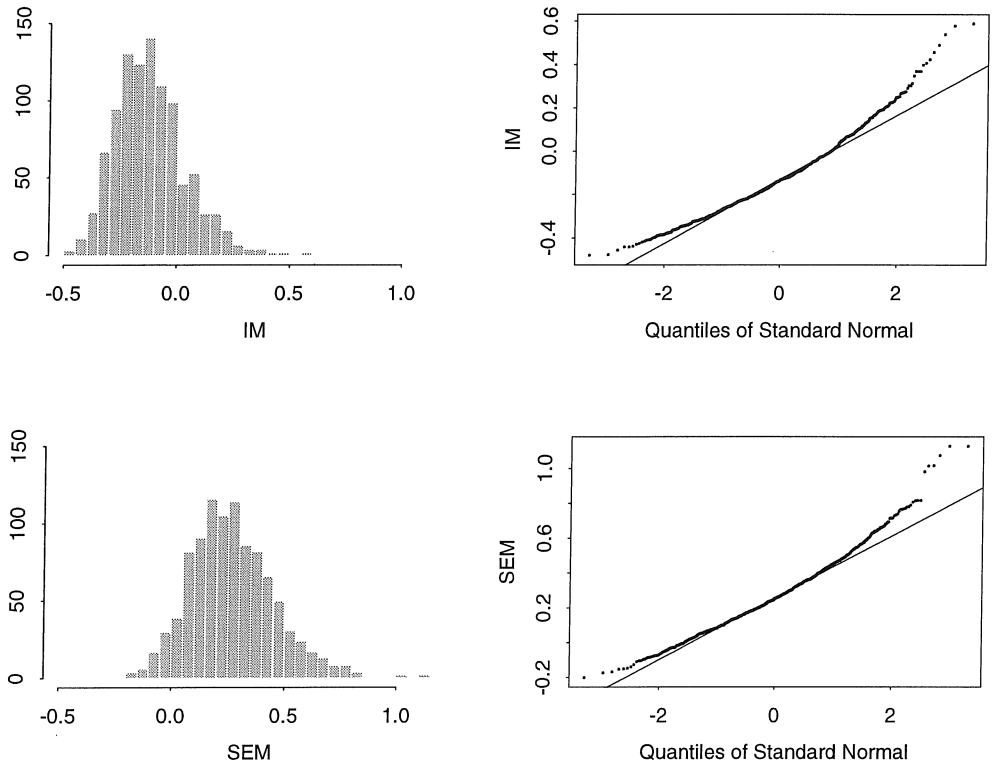

Figure 7: Histograms and QQ-plots for $\tilde{\sigma}^2 - \hat{\sigma}^2$, $m = 1$

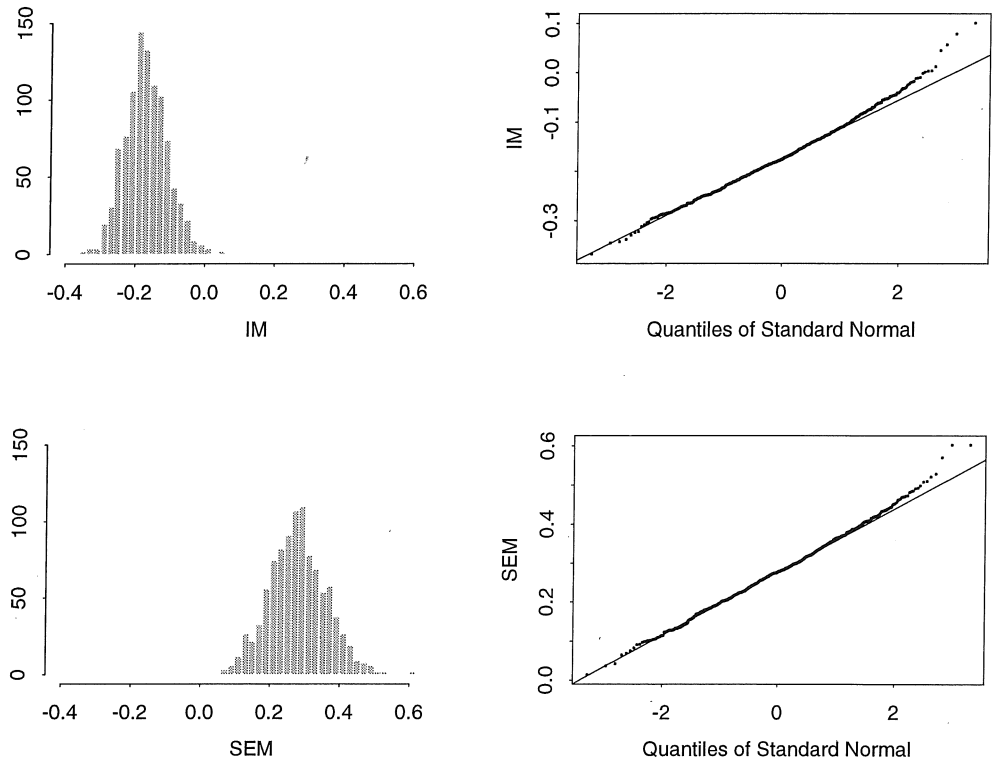

Figure 8: Histograms and QQ-plots for $\tilde{\sigma}^2 - \hat{\sigma}^2$, $m = 5$

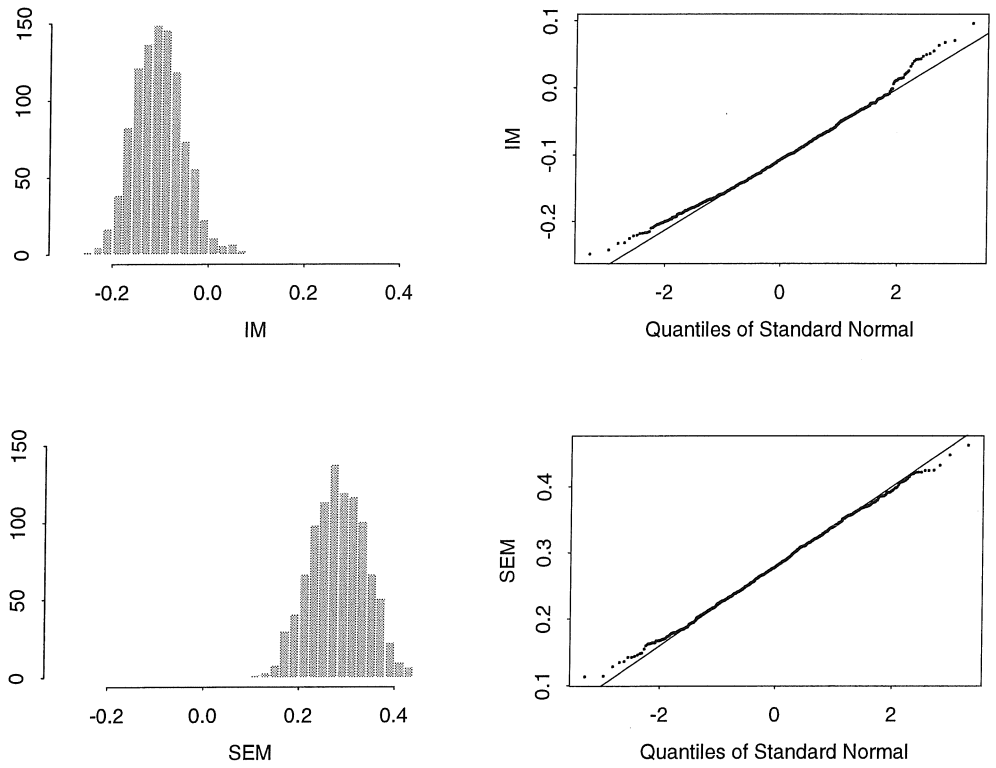

Figure 9: Histograms and QQ-plots for $\tilde{\sigma}^2 - \hat{\sigma}^2$, $m = 10$

For all parameters, the standard deviations decrease as m increases. The decrease is roughly $1/\sqrt{5} \approx 0.44$ when m increases from 1 to 5 and $1/\sqrt{2} \approx 0.71$ when going from $m = 5$ to $m = 10$, as we would expect from the asymptotic results. This fact and the distribution of $\tilde{\sigma}^2 - \hat{\sigma}^2$ indicates that as m increases the empirical distribution of the simulation estimators minus the MLEs gets closer to normality. This does not mean that the asymptotic results of Section 2 hold. For instance, the distribution of $\tilde{\sigma}^2 - \sigma_0^2$ does not approximate a normal distribution when m increases, since the distribution of $\hat{\sigma}^2 - \sigma_0^2$ is unaffected by m, and this distribution is probably not normal with the sample size in this example.

The bias of the SEM estimator is unaffected by the choice of m, whereas the mean of the IM estimator is always lower for $m = 5$. This leads to larger bias when $m = 5$ in the estimators of β and $\hat{\sigma}^2$, but smaller in the positively biased estimator of ρ. As mentioned previously this shows up in additional simulations and is thus not "accidental". Whether it is an effect of the observations or a more general phenomenon remains to be seen.

The components of the IM estimator appears to be independent whereas the SEM estimators of ρ and σ^2 are negatively correlated; the correlation coefficient is approximately -0.5, independent of m. Hence, some care should be exercised when interpreting the results for ρ and σ^2.

In conclusion, the IM algorithm performs better than the SEM algorithm in this example, even though the sample size is to small for the asymptotic results to hold. The SEM algorithm is faster than the IM algorithm in this example, but the number of iterations used in the IM version is probably a lot larger than what is necessary for convergence. We might hope to improve the SEM algorithm by increasing m but as we

have seen the main problem with the SEM estimator is bias rather than variance, and the bias does not seem to decrease for moderate values of m. One could also hope to improve the SEM algorithm in this example by allowing more iterations in the Gibbs sampler used for simulating (X_1, X_2) given that they are both negative. However, since only 7 observations have both X_1 and X_2 censored, the effect of this will probably not be very large.

Acknowledgment

I am grateful to Richard Gill for calling my attention to the subject of simulated EM algorithms.

References

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.

CELEUX, G. AND DIEBOLT, J. (1986). The SEM Algorithm: A Probabilistic Teacher Algorithm Derived From the EM Algorithm for the Mixture Problem. *CmpStQ* **2**, 73-82.

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.

LETAC, G. (1986). A contraction principle for certain Markov chains and its applications. *Contemporary Mathematics* **50**, 263-273.

MCFADDEN, D. AND RUUD P.A. (1994). Estimation By Simulation. *Review of Economics and Statistics* **76**, 591-608.

MENG, X.-L. AND RUBIN, D.B. (1991). Using EM to Obtain Asymptotic Variance-covariance Matrices: The SEM Algorithm. *JASA* **86**, 899-909.

NIELSEN, S.F. (1997). *The Imputation Maximization Algorithm.* Manuscript.

PAKES, A. AND POLLARD, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica* **57**, 1027-1057.

RUUD, P.A. (1991). Extensions of Estimation Methods Using the EM Algorithm. *J. Econometrics* **49**, 305-341.

WU, C.F.J. (1983). On the Convergence Properties of the EM Algorithm . *Ann Statist* **11**, 95-103.

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.
No. 3 Nielsen, Søren Feodor: On Simulated EM Algorithms.