NÚMERO 351

RICARDO SMITH RAMIREZ

A Monte Carlo EM Algorithm to Estimate Structural Equation Systems with Unobserved Information

DICIEMBRE 2005

www.cide.edu

Abstract

A Monte Carlo Expectation-Maximization algorithm for solving structural models with latent structures is formulated. It combines a Gibbs sampler to impute the unobserved information in the E-step, a sequential maximization procedure in the M-step and a stochastic version of Louis' method to estimate the Information matrix. I show that such an algorithm has a number of advantages with respect to traditional methods. First, it does not require integrating the unobserved information out from the likelihood function, which reduces the estimation time dramatically and permits to solve problems involving more than three latent variables. Second, it reduces the estimation of the vector of slopes to the calculation of a GLS estimator and numerical optimization is required only to estimate the elements in the disturbance covariance matrix. Third, it can accommodate potentially any linear-in-parameters equation system including cross-sectional models with latent variables, panel data models and stochastic frontier models. Finally, the estimation of the standard errors by Louis' method circumvents the limitations associated to the estimation of numerical Hessians by finite-difference methods.

Resumen

En este paper se formula un algoritmo Monte Carlo EM para estimar modelos de ecuaciones estructurales con variables latentes. El algoritmo combina un simulador de Gibbs en la etapa E para simular la información no observada, un proceso de maximización secuencial en la etapa M, y una versión estocástica del método de Louis para estimar la matriz de información. El algoritmo tiene varias ventajas con respecto a los métodos tradicionales de estimación. Primero, no requiere del cálculo de integrales en la función de verosimilitud, lo cual reduce dramáticamente el tiempo de estimación y permite solucionar problemas que involucran más de tres variables latentes. Segundo, reduce la estimación del vector de pendientes al cálculo de estimador de MCG, y optimización numérica sólo es requerida para estimar los elementos en la matriz de covarianzas de las perturbaciones. Puede se usado para estimar potencialmente cualquier sistema con ecuaciones lineales en parametros, incluyendo modelos de corte transversal con variables latentes, modelos de panel, y modelos de frontera estocástica. Finalmente, la estimación de los errores estándar por el método de Louis evita los problemas asociados al cálculo del Hessiano mediante técnicas numéricas tradicionales.

Introduction

Models involving equation systems with latent structures are abundant in applied economics literature. Examples include sample selectivity models, switching regression models, multivariate and nested tobit models, multivariate and multinomial probit models, and panel data models with random effects. The use of latent variables gives some level of independence from the limitations of the observed data to the applied econometrician. Completely or partially unobserved variables can be added to the empirical model in order to get a better representation of the phenomenon under study. However, the use of latent variables does not come without costs. An important issue is that both the number and the dimensionality of the integral terms in the likelihood function increase with the number of latent variables considered. High dimensional integration slows the estimation down and it can even make the estimation unfeasible in presence of integrals of dimension greater than three. A second issue arises from the inability of many conventional optimization algorithms to identify the parameters of these models even though conditions for formal identification are satisfied. "Fragile" identification, as it called by Keane (1992), tends to happen when the objective function shows little variation in a wide range of parameter values around the maximum, which prevents convergence of gradient-based algorithms. Finally, the selection of starting values in order to initiate the optimization routine is a frequent problem in maximum likelihood estimation of models with latent structures. Consequently, the development of algorithms with low sensitivity to the selection of starting values (i.e. with a larger approximation area) is always welcome.

The traditional approaches for estimating 2-equation systems that involve latent variables have been maximum likelihood and 2-step estimation methods (Heckman, 1979; Maddala, 1983). The first one is more desirable because it produces consistent and efficient estimates; however, it is prone to "fragile" identification and starting value problems. The second one is robust, but it is not efficient. For equation systems involving three or more latent variables, maximum likelihood estimation by numerical integration is often too costly computationally or even unfeasible since quadrature methods for high dimensional integrals are still in development. This so-called "curse of dimensionality" has, however, been partially overcome in the last years by the use of probability simulators (Börsch-Supan and Hajivassiliou, 1993; Geweke *et al.*, 1994) and Monte Carlo and Quasi-Monte Carlo integration methods (Sobol, 1998). Still, the focus of these approaches is only making the integration of the likelihood function feasible, which implies that the problems of "fragile" identification and starting values remain.

Instead of placing the attention on calculating the integrals in the likelihood function, the approach presented in this article focuses on the latent continuum that generates the observed information. By combining a Monte Carlo Expectation-Maximization (MCEM) algorithm with a sequential conditional maximization procedure, I show that the estimation of any system of linear (in parameters) equations with latent variables can be seen as equivalent to estimating a system of linear (in parameters) equations with fully observed information recursively. The use of a MCEM circumvents the integration problem by imputing the unobserved information using Gibbs sampling (Casella and George, 1992). Since the use of the Gibbs sampler in the Expectation step permits "restoring" the continuum, the Maximization step does not differ much from maximizing the likelihood function of a standard linear equation system. Complementarily, the use of sequential maximization steps permit to concentrate the optimization effort on those parameters that are frequently the hardest to identify, i.e. the elements of the disturbance covariance matrix. Finally, the MCEM framework confines the estimates to the parameter space at every iteration of the algorithm and reduces dependency on starting values. This study generalizes the procedure developed by Natarajan *et al*. (2000) for multinomial probit models. Its main contribution is the implementation of a robust algorithm that exploits explicitly the structural similarity between models that have been traditionally estimated by rather *ad-hoc* methods.

The remaining of this article is organized in the following way. The next section discusses the meaning of unobserved information in the context of this article. The second section presents the MCEM algorithm and exemplifies how it works by estimating a 3-equation problem. The third section solves the same problem as in the previous section by numerical integration. The outputs of both approaches are compared. The fourth section generalizes the algorithm to cover system of structural equations with latent variables. The fifth and last section gives final remarks.

1.- Unobserved information

For purposes of this article I consider two ways that unobserved information might enter in econometric estimations: by the existence of missing data, and by the presence of latent variables. Missing data are observations that the researcher failed to collect for all the individuals in the sample. Missing information can originate by multiple ways such as inability to sample the same unit along different years when constructing a panel data set, or unwillingness of the respondent to answer specific questions in a survey.

By latent variable I mean a continuous variable that is not observed fully; nonetheless, part of the information contained in the variable is available to the econometrician.

This observed counterpart originates a new variable, whose type (e.g. dichotomous, polytomous, limited-dependent) will depend on how it relates to the underlying latent variable.

It must be kept in mind that those variables containing missing data, although less structured, can also be seen and understood as another kind of latent variable. As a consequence, the methods that I am about to present in order to estimate models involving latent variables can be used to deal with missing information as well.

2.- The Monte Carlo Expectation- Maximization (MCEM) algorithm

In their presentation before the Royal Statistical Society, Dempster et al. (1977) introduced the Expectation-Maximization (EM) algorithm as an iterative procedure to compute maximum likelihood estimates "... when the observations can be viewed as incomplete data."

The way the notion of "incomplete data" is introduced above is indeed very general and it is this flexibility in the idea of incomplete data what is responsible of a good deal of EM algorithm's broad applicability. To give a flavor of how the algorithm works consider the following many-to-one mapping:

$$z \in Z \rightarrow y = y(z) \in Y$$

The information z in Z is not observed directly but through its observed realization y in Y.

In words, z is only know to lie in $Z(y)$, the subset of Z determined by the equation $y = y(z)$, where y is the observed (measurable) data.

Let the complete data be written as $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$, where $\boldsymbol{z}$ is the unobserved information. Then the log-likelihood function of the observed information can be written as,

$$\ell(\theta \mid \boldsymbol{y}) = \ln L(\theta \mid \boldsymbol{y}) = \ln \int_{Z(y)} L(\theta \mid \boldsymbol{x})\, d\boldsymbol{z} \qquad (1)$$

As previously discussed, the integrals present in (1) can make the maximization of $\ell(\theta \mid \boldsymbol{y})$ cumbersome or even impossible to solve by standard optimization methods. Instead of trying to solve (1) directly, the EM algorithm

focuses on the complete-information log-likelihood $\ell^c(\theta|\boldsymbol{x})$ and maximizes $E\left[\ell^c(\theta|\boldsymbol{x})\right]$ by executing iteratively two steps. The first one is the so-called Expectation step or E-step, which at iteration *m*+1 computes $Q\left(\theta|\theta^{(m)},\boldsymbol{y}\right)=E\left[\ell^c(\theta|\boldsymbol{x})\right]$, where $E\left[\ell^c(\theta|\boldsymbol{x})\right]$ is the expectation of the complete-information log-likelihood conditional on the observed information and provided that the conditional density $f\left(\boldsymbol{x}|\boldsymbol{y},\theta^{(m)}\right)$ is known. The E-step is followed by the Maximization step or M-step, which maximizes $Q\left(\theta|\theta^{(m)},\boldsymbol{y}\right)$ to find $\theta^{(m+1)}$. Then the procedure is repeated until convergence is attained. Often, however, this deterministic version of the EM algorithm has also to deal with hefty integrals in the calculation of the expectations in the E-step.

The stochastic version of the EM algorithm presented here avoids troublesome computations in the E-step by imputing the unobserved information conditional on what is observed and on distribution assumptions. In this approach the term $Q\left(\theta|\theta^{(m)},\boldsymbol{y}\right)$ is approximated by the mean $\frac{1}{K}\sum_{k=1}^{K}Q\left(\theta,\boldsymbol{z}^{(k)}|\boldsymbol{y}\right)$, where the $\boldsymbol{z}^{(k)}$ are random samples from $f\left(\boldsymbol{x}|\theta^{(m)},\boldsymbol{y}\right)$ (Wei and Tanner, 1990). No integrals need to be estimated in this procedure. Once the unobserved information is imputed, the latent continuum is made "visible" and the estimation can be carried out as we were solving a standard system of linear equations.

2.1.- Implementing the Monte Carlo EM algorithm

I illustrate the use of the MCEM algorithm by solving the following 3-equation system,

$$\begin{aligned} y_{1i}^{*} &= x_{1i}\beta_1+\varepsilon_{1i} \\ y_{2i}^{*} &= \gamma_2 y_{1i}+x_{2i}\beta_2+\varepsilon_{2i} \\ y_{3i}^{*} &= \gamma_3 y_{1i}+x_{3i}\beta_3+\varepsilon_{3i} \end{aligned} \qquad (2)$$

where y_{1i} is dichotomous, and y_{2i} and y_{3i} are censored from below at zero, i.e.

$$y_{1i}=\begin{cases} 1 \text{ if } y_{1i}^{*}>0 \\ 0 \text{ if } y_{1i}^{*}\le 0 \end{cases} \qquad y_{2i}=\begin{cases} y_{2i}^{*} \text{ if } y_{2i}^{*}>0 \\ 0 \text{ if } y_{2i}^{*}\le 0 \end{cases} \qquad y_{3i}=\begin{cases} y_{3i}^{*} \text{ if } y_{3i}^{*}>0 \\ 0 \text{ if } y_{3i}^{*}\le 0 \end{cases}$$

Equation system (2) contains only the observed counterparts of the latent variables on the right-hand side of the equations. It is clear that more general cases should consider both latent and observed endogenous regressors. The discussion of those cases will be delayed until Section 4. In the meantime the use of simpler models like (2) is more suitable to introduce the Monte Carlo EM algorithm. This will allow us to concentrate on methodological aspects and not get distracted by complications in the model structure.
The disturbance terms in (2) are assumed to have a trivariate normal distribution $N(0,\Sigma)$ with covariance matrix

$$\Sigma=\begin{pmatrix}1 & \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_1\varepsilon_3}\\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma^2_{\varepsilon_2} & \sigma_{\varepsilon_2\varepsilon_3}\\ \sigma_{\varepsilon_1\varepsilon_3} & \sigma_{\varepsilon_2\varepsilon_3} & \sigma^2_{\varepsilon_3}\end{pmatrix}=\begin{pmatrix}1 & \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_2} & \rho_{\varepsilon_1\varepsilon_3}\sigma_{\varepsilon_3}\\ \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_2} & \sigma^2_{\varepsilon_2} & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3}\\ \rho_{\varepsilon_1\varepsilon_3}\sigma_{\varepsilon_3} & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} & \sigma^2_{\varepsilon_3}\end{pmatrix} \qquad (3)$$

where $\sigma^2_{\varepsilon_1}=1$ is the usual normalization to ensure identification of the coefficients in an equation with a dichotomous dependent variable and $\rho_{\varepsilon_k\varepsilon_l}$ is the correlation coefficient between ε_k and ε_l $(k,l=1,2,3)$. I applied the method on data from a survey administered to Maryland farmers in 1998 in order to evaluate a conservation cost-sharing program. For a detailed description of the survey see Lichtenberg and Smith-Ramirez (2004).

There are two forms of the structural model for the system in (2) depending on the value of the observed counterpart of y^*_{1i}, i.e. e.

$$\begin{aligned} &y_{1i}=1 && y_{1i}=0\\ &y^*_{1i}=x_{1i}\beta_1+\varepsilon_{1i} && y^*_{1i}=x_{1i}\beta_1+\varepsilon_{1i}\\ &y^*_{2i}=\gamma_2+x_{2i}\beta_2+\varepsilon_{2i} && y^*_{2i}=x_{2i}\beta_2+\varepsilon_{2i}\\ &y^*_{3i}=\gamma_3+x_{3i}\beta_3+\varepsilon_{3i} && y^*_{3i}=x_{3i}\beta_3+\varepsilon_{3i}\end{aligned} \qquad (4)$$

According to (4) the parameters γ_2 and γ_3 only represent shifts in the intercepts of the second and third equations when $y_{1i}=1$. Thus, under the normality assumption, the complete data likelihood function can be written as,

$$L(\boldsymbol{\theta},\Sigma\,|\,\boldsymbol{x})=\prod_i f(y_{1i},y_{2i},y_{3i})=\prod_i\left[\frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}}\exp\left(-\frac{\boldsymbol{\varepsilon}_i'\Sigma^{-1}\boldsymbol{\varepsilon}_i}{2}\right)\right]$$

Where $\boldsymbol{\theta} = (\beta_1, \gamma_2, \beta_2, \gamma_3, \beta_3)$, $\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} = \begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i} - X_{2i}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i} - X_{3i}\beta_3 \end{pmatrix}$

Correspondingly, the complete information log-likelihood function and its expectation are,

$$\ell^c(\boldsymbol{\theta}, \Sigma \mid \boldsymbol{x}) = -\frac{3N}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_i \mathrm{tr}\left(\Sigma^{-1}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right)$$

$$E\left[\ell^c(\boldsymbol{\theta}, \Sigma \mid \boldsymbol{x})\right] = -\frac{3N}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_i E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right]\right) \qquad (5)$$

Where N is the total number of observations and the expectation operator indicates expectation conditional on observed information and distributional assumptions. The E-step is straightforward from equation (5) and, at iteration $m+1$, requires the calculation of,

$$Q_i\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right) = E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right] = E\left[\begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i} - X_{2j}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i} - X_{3j}\beta_3 \end{pmatrix}\begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i} - X_{2j}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i} - X_{3j}\beta_3 \end{pmatrix}' \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right] \qquad (6)$$

$$= \sigma_i^{2(m)} + \begin{pmatrix} \mu_{y_{1i}^*}^{(m)} - X_{1i}\beta_1 \\ \mu_{y_{2i}^*}^{(m)} - \gamma_2 y_{1i} - X_{2i}\beta_2 \\ \mu_{y_{3i}^*}^{(m)} - \gamma_3 y_{1i} - X_{3i}\beta_3 \end{pmatrix}\begin{pmatrix} \mu_{y_{1i}^*}^{(m)} - X_{1i}\beta_1 \\ \mu_{y_{2i}^*}^{(m)} - \gamma_2 y_{1i} - X_{2i}\beta_2 \\ \mu_{y_{3i}^*}^{(m)} - \gamma_3 y_{1i} - X_{3i}\beta_3 \end{pmatrix}'$$

where $\sigma_i^{2(m)} = \mathrm{Cov}\left(y_{1i}^*, y_{2i}^*, y_{3i}^* \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right) = \begin{pmatrix} \sigma_{y_{1i}^*}^{2(m)} & \sigma_{y_{1i}^* y_{2i}^*}^{(m)} & \sigma_{y_{1i}^* y_{3i}^*}^{(m)} \\ \sigma_{y_{1i}^* y_{2i}^*}^{(m)} & \sigma_{y_{2i}^*}^{2(m)} & \sigma_{y_{2i}^* y_{3i}^*}^{(m)} \\ \sigma_{y_{1i}^* y_{3i}^*}^{(m)} & \sigma_{y_{2i}^* y_{3i}^*}^{(m)} & \sigma_{y_{3i}^*}^{2(m)} \end{pmatrix}$ (7)

and $\begin{pmatrix} \mu_{y_{1i}^*}^{(m)} \\ \mu_{y_{2i}^*}^{(m)} \\ \mu_{y_{3i}^*}^{(m)} \end{pmatrix} = \begin{pmatrix} E\left[y_{1i}^* \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right] \\ E\left[y_{21i}^* \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right] \\ E\left[y_{3i}^* \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right] \end{pmatrix}$ (8)

The covariance matrix $\sigma_i^{2(m)}$ in (7) and the vector of means in (8) can be estimated by Gibbs sampling (Casella and George, 1992) from the joint distribution of $\left(y_{1i}^*, y_{2i}^*, y_{3i}^*\right)$ conditional on parameters $\left(\boldsymbol{\theta}^{(m)}, \Sigma^{(m)}\right)$ and the observed information $\boldsymbol{y}$. It is useful determining first the distribution of

$\left(y_{1i}^*, y_{2i}^*, y_{3i}^*\right)$. After recalling that γ_2 and γ_3 are only structural shifts in the second and third equations of (4) and given the distribution of the disturbances in (3), the distribution of $\left(y_{1i}^*, y_{2i}^*, y_{3i}^*\right)$ at iteration m is $N\left(\mu_i^{(m)}, \Sigma^{(m)}\right)$, where,

$$\mu_i^{(m)} = \begin{pmatrix} X_{1i}\beta_1^{(m)} \\ \gamma_2^{(m)} y_{1i} + X_{2i}\beta_2^{(m)} \\ \gamma_3^{(m)} y_{1i} + X_{3i}\beta_3^{(m)} \end{pmatrix} \quad \text{and} \quad \Sigma^{(m)} = \begin{pmatrix} 1 & \sigma_{\varepsilon_1\varepsilon_2}^{(m)} & \sigma_{\varepsilon_1\varepsilon_3}^{(m)} \\ \sigma_{\varepsilon_1\varepsilon_2}^{(m)} & \sigma_{\varepsilon_2}^{(m)} & \sigma_{\varepsilon_2\varepsilon_3}^{(m)} \\ \sigma_{\varepsilon_1\varepsilon_3}^{(m)} & \sigma_{\varepsilon_2\varepsilon_3}^{(m)} & \sigma_{\varepsilon_3}^{(m)} \end{pmatrix} \qquad (9)$$

2.2.- The Gibbs sampler

The moments in (7) and (8) could be easily calculated if the marginal densities (conditional on parameters and observed information) of y_{1i}^*, y_{2i}^*, and y_{3i}^* were known. However, obtaining those marginal densities may require solving hefty integrals. Instead of tackling the problem by integration, the Gibbs sampler provides a way to generate samples from the marginal distributions without requiring analytical expressions for the densities. The moments of interest can then be estimated from the simulated samples.

The implementation of the Gibbs sampler is straightforward using the definitions in (9). Before proceeding, let consider the following notation,

$$\boldsymbol{y}_{i|-j}^* = \begin{pmatrix} y_{1i}^* \\ \vdots \\ y_{j-1i}^* \\ y_{j+1i}^* \\ \vdots \\ y_{ki}^* \end{pmatrix} \qquad \boldsymbol{X}_{i|-j} = \begin{pmatrix} X_{1i} \\ \vdots \\ X_{j-1i} \\ X_{j+1i} \\ \vdots \\ X_{ki} \end{pmatrix} \qquad \boldsymbol{\gamma}_{-j}^{(m)} = \begin{pmatrix} \gamma_1^{(m)} \\ \vdots \\ \gamma_{j-1}^{(m)} \\ \gamma_{j+1}^{(m)} \\ \vdots \\ \gamma_k^{(m)} \end{pmatrix} \qquad \boldsymbol{\beta}_{-j}^{(m)} = \begin{pmatrix} \beta_1^{(m)} \\ \vdots \\ \beta_{j-1}^{(m)} \\ \beta_{j+1}^{(m)} \\ \vdots \\ \beta_k^{(m)} \end{pmatrix}$$

where $j = 1, \ldots, k$ and k is the number of equations in the system to estimate (equal to 3 in our example).

The implementation of the sampler begins with determining the distribution of each y_{ji}^* conditional on the value of the rest of the dependent variables $\boldsymbol{y}_{i|-j}^*$. It is well known that, under the normality assumption, this

conditional distribution is univariate normal. Thus, means $\mu_{ji|i(-j)}$ and variances $\sigma^2_{j|-j}$ at the $m+1$ iteration can be estimated by,

$$\begin{aligned}\mu^{(m)}_{ji|i(-j)} &= E\left(y^*_{ji}\middle|\boldsymbol{y}^*_{i|-j},\boldsymbol{\theta}^{(m)},\Sigma^{(m)}\right)\\ &= X_{ji}\beta^{(m)}_j + \operatorname{cov}\left(y^*_{ji}\middle|\boldsymbol{y}^*_{i|-j},\Sigma^{(m)}\right)\left[\operatorname{cov}\left(\boldsymbol{y}^*_{i|-j}\middle|\Sigma^{(m)}\right)\right]^{-1}\left(\boldsymbol{y}^*_{i|-j}-\boldsymbol{\gamma}^{(m)}_{-j}-\boldsymbol{X}_{i|-j}\boldsymbol{\beta}^{(m)}_{-j}\right)\end{aligned} \tag{10}$$

$$\begin{aligned}\sigma^{2(m)}_{j|-j} &= \operatorname{var}\left(y^*_{ji}\middle|\boldsymbol{y}^*_{i|-j},\boldsymbol{\theta}^{(m)},\Sigma^{(m)}\right)\\ &= \operatorname{var}\left(y^*_{ji}\middle|\Sigma^{(m)}\right)-\operatorname{cov}\left(y^*_{ji}\middle|\boldsymbol{y}^*_{i|-j},\Sigma^{(m)}\right)\left[\operatorname{cov}\left(\boldsymbol{y}^*_{i|-j}\middle|\Sigma^{(m)}\right)\right]^{-1}\operatorname{cov}\left(y^*_{ji}\middle|\boldsymbol{y}^*_{i|-j},\Sigma^{(m)}\right)'\end{aligned} \tag{11}$$

The next step is to sample iteratively from these conditional distributions in order to simulate a sample for the unobserved values of each y^*_{ji}. These samples will in turn allow estimating the values in (7) and (8). Since the simulations for $\boldsymbol{y}^*_i$ must be done conditional on its corresponding observed information $\boldsymbol{y}_i$, the implementation procedure depends on the structure imposed by $\boldsymbol{y}_i$ on $\boldsymbol{y}^*_i$.

The observed counterpart of y^*_{1i} in the first equation in (2) is dichotomous with y^*_{1i} being positive if y_{1i} equals one and non-positive if y_{1i} equals zero. Accordingly, it is necessary to simulate y^*_{1i} from a normal distribution with mean $\mu^{(m)}_{1i|i(-1)}$ and variance $\sigma^{2(m)}_{1i|-1}$ truncated below at zero if y_{1i} equals one and truncated above at zero if y_{1i} equals zero.

Variables y^*_{2i} and y^*_{3i} are both observed when having positive values. Consequently, it is only necessary to simulate them when $y_{2i}=0$ and $y_{3i}=0$, respectively. Thus, these variables must be simulated from normal distributions with means $\mu^{(m)}_{ji|i(-j)}$ and variances $\sigma^{2(m)}_{ji|-j}$ truncated above at zero when y_{ji} $(j=2,3)$ equals zero. When $y_{ji}>0$ set $y^*_{ji}=y_{ji}$.

Sampling from a truncated normal distribution can be easily accomplished by using the inverse distribution method. As an example, assume $y^* \sim N\left(\mu,\sigma^2\right)$ and y is limited to be in the interval $[l,u]$. Then according to Devroye (1986, p39), a random draw from the truncated normal distribution of y is given by,

$$y=\mu+\sigma\Phi^{-1}\left(P_l+U\left(P_u-P_l\right)\right) \tag{12}$$

Where $P_l = \Phi\left(\frac{l-\mu}{\sigma}\right)$, $P_u = \Phi\left(\frac{u-\mu}{\sigma}\right)$ and U is a random draw from the standard uniform distribution. MATLAB's pseudo random generator (Moler, 1995) was used to simulate the sampling from U in this study. As an alternatively to the use of pseudo random numbers in the sampling, the use of randomized low-discrepancy sequences or quasi-random numbers has been proposed in order to reduce Monte Carlo noise and speed up convergence (Liao, 1998; Jank, 2004).

A complete set of starting vectors $\boldsymbol{y}_i^*$ is necessary to begin the Gibbs sampler. In this study y_{ji}^* was set equal to zero $\forall i, j$ when the observed variable was dichotomous and equal to y_{ji} when censored. The simulation was then repeated iteratively until completing a sequence $\boldsymbol{y}_i^{*(1)}, \ldots, \boldsymbol{y}_i^{*(K^{(m)})}$, where $K^{(m)}$ is a number large enough to ensure convergence. Following Wei and Tanner (1990), it is more efficient to begin with a small $K^{(1)}$ and progressively increase $K^{(m)}$ as m increases. A simple linear rate of increment was used here. Then eliminate a number k_{burn} of simulations from the beginning of the sequence. The remaining observations in the sequence are used to estimate $\sigma_{ji}^{2(m)}$ and $\mu_{ji}^{(m)}$ in (7) and (8) according to,

$$\sigma_i^{2(m)} = \operatorname{cov}\left(y_{1i}^*, y_{2i}^*, y_{3i}^* \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right) \approx \begin{pmatrix} \hat{\sigma}^{2(m)}_{y_{1i}^*} & \hat{\sigma}^{(m)}_{y_{1i}^* y_{2i}^*} & \hat{\sigma}^{(m)}_{y_{1i}^* y_{3i}^*} \\ \hat{\sigma}^{(m)}_{y_{1i}^* y_{2i}^*} & \hat{\sigma}^{2(m)}_{y_{2i}^*} & \hat{\sigma}^{(m)}_{y_{2i}^* y_{3i}^*} \\ \hat{\sigma}^{(m)}_{y_{1i}^* y_{3i}^*} & \hat{\sigma}^{(m)}_{y_{2i}^* y_{3i}^*} & \hat{\sigma}^{2(m)}_{y_{3i}^*} \end{pmatrix}$$

Where $\hat{\sigma}^{(m)}_{y_{ri}^* y_{si}^*} = \frac{1}{K^{(m)} - k_{burn} - 1} \sum_{k=k_{burn}+1}^{K^{(m)}} \left(y_{ri}^{*(k)} - \overline{y}_{ri}^*\right)\left(y_{si}^{*(k)} - \overline{y}_{si}^*\right)$

$$\mu^{(m)}_{y_{ji}^*} = E\left[y_{ji}^* \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right] \approx \overline{y}_{ji}^* = \frac{1}{K^{(m)} - k_{burn}} \sum_{k=k_{burn}+1}^{K^{(m)}} y_{ji}^{*(k)}$$

Notice that when y_{ji}^* is fully observed (i.e. when $y_{ji}^* = y_{ji}$) then $\hat{\sigma}^{(m)}_{y_{ri}^* y_{si}^*} = 0$ and $\mu^{(m)}_{y_{ji}^*} = y_{ij} \quad \forall\, m, r, s$.

2.3.- Maximization Step

After obtaining $\sigma_{ji}^{2(m)}$ and $\mu_{ji}^{(m)}$ we are ready to move to the Maximization step. From (5) and (6) we maximize,

$$E\left[\ell^{c}\left(\boldsymbol{\theta},\Sigma\,|\,\boldsymbol{\theta}^{(m)},\Sigma^{(m)},\boldsymbol{y}\right)\right]=-\frac{3N}{2}\ln\left(2\pi\right)-\frac{N}{2}\ln\left|\Sigma\right|-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i}Q_{i}\left(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(m)},\Sigma^{(m)},\boldsymbol{y}\right)\right) \quad (13)$$

Note that the use of the Gibbs sampler has permitted us to circumvent the estimation of the high dimensional integrals present in (21). Also notice that, except for the covariance matrices $\sigma_{i}^{2(m)}$ present in the terms, the expression in (13) is the log-likelihood function of a system of linear equations, where the unobserved information has been replaced by its expected values. Thus, in a certain sense the latent continuum has been restored. Similarly to Meg and Rubin (1993) and Natarajan *et al*. (2000), I use two conditional maximization steps in order to maximize the expression in (13) with respect to $\boldsymbol{\theta}$ and the elements in Σ. The first maximization step maximizes (13) with respect to $\boldsymbol{\theta}$ conditional on $\Sigma^{(m)}$ to produce $\boldsymbol{\theta}^{(m+1)}$. This is followed by a maximization on the elements of Σ conditional on the recently updated $\boldsymbol{\theta}^{(m+1)}$ in order to obtain $\Sigma^{(m+1)}$.

It is clear from (6) that the maximizer in the first conditional maximization is the generalized least square estimator,

$$\boldsymbol{\theta}^{(m+1)}=\left[\tilde{X}_{d}^{'}\left(\Sigma^{(m)}\otimes I_{N}\right)^{-1}\tilde{X}_{d}\right]^{-1}\tilde{X}_{d}^{'}\left(\Sigma^{(m)}\otimes I_{N}\right)^{-1}\mu_{y^{*}}^{(m)} \quad (14)$$

where,

$\tilde{X}_{d}=\begin{bmatrix}X_{1} & 0 & 0\\ 0 & \tilde{X}_{2} & 0\\ 0 & 0 & \tilde{X}_{3}\end{bmatrix}$, $\tilde{X}_{2}$ and $\tilde{X}_{3}$ are the matrices $[y_{1i}\vdots X_{2}]$ and $[y_{1i}\vdots X_{3}]$ respectively, I_{N} is the identity matrix of dimension N and $\mu_{y^{*}}^{(m)}$ is a column vector of dimension Nk constructed by stacking the elements $\mu_{y_{ji}^{*}}^{(m)}$ from (8) in the following way,

$$\mu_{y^{*}}^{(m)}=\left(\mu_{y_{11}^{*}}^{(m)},\mu_{y_{12}^{*}}^{(m)},...,\mu_{y_{1N}^{*}}^{(m)},\mu_{y_{21}^{*}}^{(m)},...,\mu_{y_{2N}^{*}}^{(m)},\mu_{y_{31}^{*}}^{(m)},...,\mu_{y_{3N}^{*}}^{(m)}\right)^{'} \quad (15)$$

After plugging (14) in (13), $\Sigma^{(m+1)}$ is obtained by maximizing,

$$E\left[\ell^c\left(\Sigma \mid \boldsymbol{\theta}^{(m+1)}, \Sigma^{(m)}, \boldsymbol{y}\right)\right] = -\frac{3N}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_i Q_i\left(\boldsymbol{\theta}^{(m+1)} \mid \boldsymbol{\theta}^{(m)}, \Sigma^{(m)}, \boldsymbol{y}\right)\right) \quad (16)$$

with respect to the $3(3-1)/2+2=5$ different elements in Σ. The maximization of (16) can be easily accomplished with the routine FMINUNC in Matlab. Unlike the log-likelihood function in (21), the function in (16) is simple enough to obtain an analytical expression for its gradient. This is useful since no time need to be spent in a numerical estimation of the gradient by the optimization routine. Contrasting with the time required to maximize a function like (21), the calculation of (14) and the maximization of (16) consume almost no time (less than 0.2 seconds when using the routines implemented here).

2.4.- Convergence issues and stopping rules

Literature discussing convergence of the MCEM is scarce and it suggests that MCEM convergence relies mainly on properties of the deterministic EM algorithm and the Gibbs sampler. Convergence of the EM algorithm is discussed by Dempster *et al*. (1977), Boyles (1983) and Wu (1983) and convergence properties of the Gibbs sampler are studied in Geman and Geman (1984) and Casella and George (1992). In one of these studies, Wu (1983) clarifies a common misconception about the superior properties of the EM algorithm to converge to a global maximum. He shows that, like other maximization methods, the EM algorithm converges monotonically to some stationary point of a bounded log-likelihood function; however there is no guarantee that point is the global maximum. Consequently, the EM algorithm is susceptible to starting value problems and may converge to a local maximum, a saddle point, or even may not converge to a unique optimum, getting trapped in a connected set of local maxima (e.g. a plateau in the objective function) instead.

In the same study cited above Wu shows that if the log-likelihood function is well behaved and it has only one stationary point (a maximum) then the EM sequence will converge to the unique maximizer. This property has two immediate implications for unimodal and differentiable log-likelihood functions. First, the EM algorithm has a greater approximation area to the global maximum or, in other words, its estimates are less sensitive to starting values than other optimization techniques. Second and directly related to the

first implication, the EM estimates are confined to lie in the parameter space at every iteration. Consequently, problems with estimates of the disturbance covariance matrix like negative variances or correlation coefficients with absolute values greater than one (which are frequent under Newton and Quasi-Newton techniques) do not happen when using the EM algorithm.

Different sets of starting values were used in this work to reduce the possibility of missing the global maximum. For a quite broad array of starting values, the MCEM algorithm implemented as above always converged to the same maximizer and always kept the estimates in the interior of the parameter space.

Some closely related issues must be discussed before finishing the implementation of the MCEM algorithm. They are the criteria to use in order to determine the size of the Gibbs sample $K^{(m)}$ and to determine when convergence has been attained. Wei and Tanner (1990) have indicated that it is inefficient to begin with large Gibbs samples since MCEM estimates are likely to be far from the true maximizer during the first iterations. Rather it is more reasonable to begin with small samples and make $K^{(m)}$ an increasing function of m in order to reduce the Monte Carlo error as the algorithm approaches the maximizer. However, there is not a single criterion about the way $K^{(m)}$ must be increased at every iteration.

Some approaches consider separately the issues of determining the optimal size of the Gibbs simulation and monitoring convergence. Thus, McCulloch (1997) considers rather abrupt increments in the size of the Gibbs sample every time that m had achieved certain arbitrary values, while McCulloch (1994) uses a linear rate of increment. Convergence monitoring in these works is accomplished by plotting the expected log-likelihood versus iteration number and the algorithm is stopped manually when the process is observed to stabilize (Wei and Tanner, 1990; Natarajan *et al.*, 2000).

More elaborate approaches consider evaluating the Monte Carlo error at iteration m and use that estimation both to determine $K^{(m+1)}$ and to evaluate convergence. These methods can be classified either as likelihood-distance-based or as parameter-distance-based depending on whether they focus on likelihood differences $\left|E\left[\ell^c\left(\vartheta^{(j)}\right)\right]-E\left[\ell^c\left(\vartheta^{(j-1)}\right)\right]\right|$ or parameter differences $\left|\vartheta^{(j)}-\vartheta^{(j-1)}\right|$, where $\vartheta^{(j)}$ is the estimation of the parameter vector at iteration j. The idea is that if parameter or likelihood differences show variation no greater than the Monte Carlo error then the estimation has been saturated by random variation and the simulation size must be increased in the next iteration. In a complementarily way, a stopping rule can be implemented by establishing a target level for the Monte Carlo error. Examples of the likelihood-distance-based approach can be found in Chan and Ledolter (1995) and Eickhoff *et al*. (2004). Probably the sounder parameter-

distance-based approach belongs to Booth and Hobert (1999). They use a Taylor series approximation to construct a confidence ellipsoid around $\vartheta^{(m+1)}$, where the length of the ellipsoid axis on every dimension of the parameter space is a measure of the MC error on the respective dimension. Thus, if the estimate $\vartheta^{(m)}$ is contained in the ellipsoid, the current estimate $\vartheta^{(m+1)}$ is swamped in MC error and $K^{(m+1)}$ must be increased.

This study uses a linear rate of increment for the size of the Gibbs sample and a stopping ruled based both on likelihood and parameter distances. The idea is simply to automate the plotting method of Wei and Tanner (1990) by introducing the following criteria:

$$\sum_{j=M-J}^{M}\left|\frac{E\left[\ell^{c}\left(\vartheta^{(j)}\right)\right]-E\left[\ell^{c}\left(\vartheta^{(j-1)}\right)\right]}{E\left[\ell^{c}\left(\vartheta^{(j-1)}\right)\right]}\right|<10^{-5} \qquad (17)$$

$$\max_{k}\left[\sum_{j=M-J}^{M}\left|\frac{\vartheta_{k}^{(j)}-\vartheta_{k}^{(j-1)}}{\vartheta_{k}^{(j-1)}}\right|\right]<10^{-3}$$

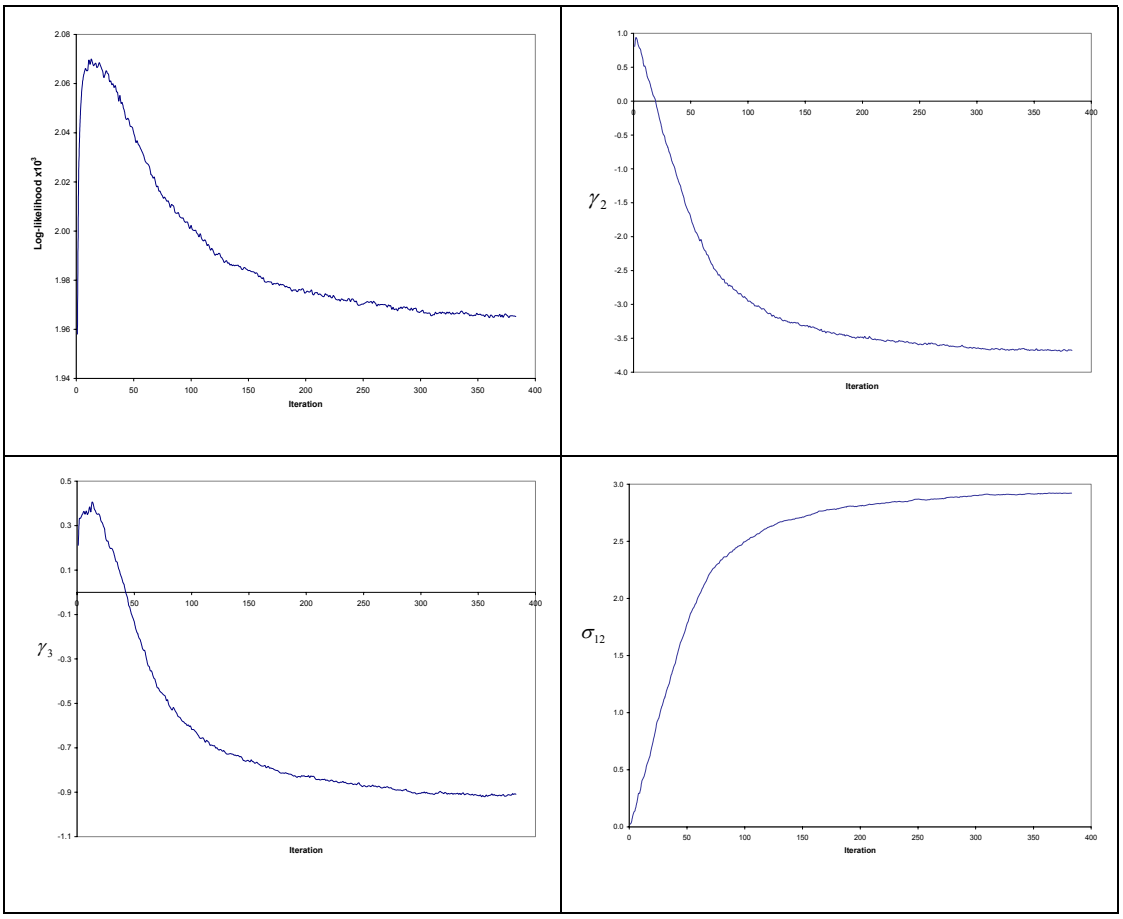
where $\vartheta_{k}^{(j)}$ is the estimate of the k component of the parameter vector at iteration *j*, *M* is the current number of iterations, and *J* is a researcher choice. In this example *J* was set equal to $0.25\times M$. The algorithm was stopped only when both criteria were satisfied for at least ten consecutive iterations. This last requirement was introduced to avoid false convergence due to the tendency of the MCEM algorithm to stall temporarily before reaching the maximizer. The criteria in (17) are simple to implement and, somewhat, they are more stringent than any of those presented in the articles mentioned above and may increase unnecessarily the number of iterations required for convergence. However, given the speed of today's computer, the computational cost is not very high.

The iteration paths for the likelihood function and selected parameters are presented in Figure 1. OLS estimates were used as starting values for the parameters in θ and an identity matrix was used for the covariance matrix of the disturbances. The routine converged after 393 iterations when using the stopping criteria described above. The Gibbs sampler was started with 300 simulations and increased by 15 simulations at every iteration of the EM algorithm, i.e. $K^{(m)}=300+15(m-1)$. The number of dismissed simulations, k_{burn}, was kept constant at 150. The algorithm converged after 2.5 hours[1].

Figure 1.

[1] On an AMD Athlon XP-M 2000+, 512 MB RAM, Windows XP, Matlab 6.5.

Iteration paths of the expected log-likelihood and selected parameter estimates.



2.5.- Estimation of the Information matrix

The asymptotic standard errors of the estimates are not among the outputs of the EM algorithm and, typically, additional code needs to be appended to the algorithm in order to estimate them. Louis's identity (Louis, 1982) was used in this article to obtain a Monte Carlo estimation of the information matrix (Guo and Thompson, 1992, Ibrahim *et. al.*, 2001). A description of how the approach works follows. Let the complete information likelihood function be $L^c(\boldsymbol{\theta};\boldsymbol{x})$, where $\boldsymbol{\theta}$ is the full set of parameters to estimate. Then, the observed log-likelihood can be written as,

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \ln L(\boldsymbol{\theta};\boldsymbol{y}) = \ln L^c(\boldsymbol{\theta};\boldsymbol{x}) - \ln \frac{L^c(\boldsymbol{\theta};\boldsymbol{x})}{L(\boldsymbol{\theta};\boldsymbol{y})} = \ell^c(\boldsymbol{\theta};\boldsymbol{x}) - \ell^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y}) \qquad (18)$$

where $\ell^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y}) = \ln \frac{L^c(\boldsymbol{\theta};\boldsymbol{x})}{L(\boldsymbol{\theta};\boldsymbol{y})}$ is the logarithm of the complete information likelihood function conditional on the observed information. After taking second derivatives on both sides of (18) we have,

$$\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \frac{\partial^2 \ell^c(\boldsymbol{\theta};\boldsymbol{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 \ell^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'},$$

which can be written in terms of information matrices in order to apply the "missing information principle" (Orchard and Woodbury, 1972),

$$I(\boldsymbol{\theta};\boldsymbol{y}) = I^c(\boldsymbol{\theta};\boldsymbol{x}) - I^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y}) \qquad (19)$$

where $I^c(\boldsymbol{\theta};\boldsymbol{x}) = -E\left[H^c(\boldsymbol{\theta};\boldsymbol{x})\right]$ is the complete information matrix, $H^c(\boldsymbol{\theta};\boldsymbol{x}) = \frac{\partial^2 \ell^c(\boldsymbol{\theta};\boldsymbol{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ is the complete information Hessian, and $I^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y})$ can be viewed as the missing information matrix. Louis (1982) showed that this last matrix could be written as,

$$I^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y}) = -E\left[\frac{\partial^2 \ell^m(\boldsymbol{\theta};\boldsymbol{x}\mid\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \mathrm{Var}\left[S^c(\boldsymbol{\theta};\boldsymbol{x})\right] = E\left[S^c(\boldsymbol{\theta};\boldsymbol{x})S^c(\boldsymbol{\theta};\boldsymbol{x})'\right] - E\left[S^c(\boldsymbol{\theta};\boldsymbol{x})\right]E\left[S^c(\boldsymbol{\theta};\boldsymbol{x})'\right] \qquad (20)$$

where $S^c(\boldsymbol{\theta};\boldsymbol{x}) = \frac{\partial \ell^c(\boldsymbol{\theta};\boldsymbol{x})}{\partial\boldsymbol{\theta}}$ is the complete information score vector. All the expectations are taken with respect to the distribution $f\left(\boldsymbol{x}\mid\boldsymbol{y},\boldsymbol{\theta}^{EM}\right)$, where $\boldsymbol{\theta}^{EM}$ is the final MCEM estimation of $\boldsymbol{\theta}$. The evaluation of all the expectations involved commonly prevents the estimation of the observed information matrix in (20) by direct calculation. Monte Carlo estimates of the expected complete information Hessian and score can be used to circumvent the problem and estimate the terms in the right hand side of (20). To sum up, the procedure implemented in this study is:

Step 1. Use the Gibbs sampler described above to simulate a sequence $\boldsymbol{y}_i^{*(1)},\ldots,\boldsymbol{y}_i^{*(R+r_{burn})}$ while holding $\boldsymbol{\theta}=\boldsymbol{\theta}^{EM}$. Eliminate a number r_{burn} of simulations from the beginning of the sequence.

Step 2. Use the remaining simulations to estimate the expectation of the complete and missing information matrices by using:

$$I^c\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}\right)=-\sum_{i=1}^{N}E\left[H_i^c\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}_i\right)\right]\cong-\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R}H_i^{c(r)}\left(\boldsymbol{\theta}^{EM};\boldsymbol{y}_i^{*(r)}\mid\boldsymbol{y}_i\right)$$

$$I^m\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}\mid\boldsymbol{y}\right)=\sum_{i=1}^{N}\left\{E\left[S_i^c\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}_i\right)S_i^c\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}_i\right)'\right]-E\left[S_i^c\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}_i\right)\right]E\left[S_i^c\left(\boldsymbol{\theta}^{EM};\boldsymbol{x}_i\right)'\right]\right\}$$

$$\cong\sum_{i=1}^{N}\left\{\frac{1}{R}\sum_{r=1}^{R}S_i^{c(r)}\left(\boldsymbol{\theta}^{EM},\boldsymbol{y}_i^{*(r)}\mid\boldsymbol{y}_i\right)S_i^{c(r)}\left(\boldsymbol{\theta}^{EM},\boldsymbol{y}_i^{*(r)}\mid\boldsymbol{y}_i\right)'-\frac{1}{R}\sum_{r=1}^{R}S_i^{c(r)}\left(\boldsymbol{\theta}^{EM};\boldsymbol{y}_i^{*(r)}\mid\boldsymbol{y}_i\right)\frac{1}{R}\sum_{r=1}^{R}S_i^{c(r)}\left(\boldsymbol{\theta}^{EM};\boldsymbol{y}_i^{*(r)}\mid\boldsymbol{y}_i\right)'\right\}$$

Expressions for the contributions from each observation to the Hessian and score are standard results from the theory of the multivariate normal distribution. Finally, plug the Monte Carlo estimates of $I^c\left(\theta;\boldsymbol{x}\right)$ and $I^m\left(\theta;\boldsymbol{x}\mid\boldsymbol{y}\right)$ in (20) and take the inverse of the resulting estimate of $I\left(\theta;\boldsymbol{y}\right)$ to get the asymptotic covariance matrix of θ^{EM}.

A sample with $R=3300$ and $r_{burn}=300$ was used in this study. Results of the Monte Carlo EM estimation of equation system (2) on the Lichtenberg and Smith-Ramirez (2004), data are presented in Table 1.

3.- The numerical integration approach and comparison between the two approaches.

In this section I solve by numerical integration the same model and data used to illustrate the MCEM algorithm. The performances of the two approaches are then compared.

The general form of the observed-information likelihood function for the equation system (2) is,

$$L=\prod_{\substack{y_{1i}=0\\y_{2i}=0\\y_{3i}=0}}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{0}f\left(y_{1i}^{*},y_{2i}^{*},y_{3i}^{*}\right)dy_{1i}^{*}dy_{2i}^{*}dy_{3i}^{*}\cdot\prod_{\substack{y_{1i}=1\\y_{2i}=0\\y_{3i}=0}}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{0}^{\infty}f\left(y_{1i}^{*},y_{2i}^{*},y_{3i}^{*}\right)dy_{1i}^{*}dy_{2i}^{*}dy_{3i}^{*}\cdot\prod_{j=2,3}\prod_{\substack{y_{1i}=0\\y_{ji}=0}}\int_{-\infty}^{0}\int_{-\infty}^{0}f\left(y_{1i},y_{2i},y_{3i}\right)dy_{1i}dy_{ji}\cdot$$
$$\prod_{j=2,3}\prod_{\substack{y_{1i}=1\\y_{ji}=0}}\int_{-\infty}^{0}\int_{0}^{\infty}f\left(y_{1i}^{*},y_{2i},y_{3i}\right)dy_{1i}^{*}dy_{ji}\cdot\prod_{y_{1i}=0}\int_{-\infty}^{0}f\left(y_{1i}^{*},y_{2i},y_{3i}\right)dy_{1i}^{*}\cdot\prod_{y_{1i}=1}\int_{0}^{\infty}f\left(y_{1i}^{*},y_{2i},y_{3i}\right)dy_{1i}^{*}$$

However, since all possible combinations of values for the dependent variables do not exist in the data set used, the observed information likelihood can be reduced to (Lichtenberg and Smith-Ramirez, 2004),

$$L(\boldsymbol{\theta},\Sigma\,|\,\boldsymbol{y},\boldsymbol{X}) = \prod_{\substack{y_{1i}=0\\y_{2i}=0\\y_{3i}=0}} \int_{-\infty}^{-X_{3i}\beta_3}\int_{-\infty}^{-X_{2i}\beta_2}\int_{-\infty}^{-X_{1i}\beta_1} \phi_3\left(\varepsilon_{1i},\varepsilon_{2i},\varepsilon_{3i}\right) d\varepsilon_{1i}d\varepsilon_{2i}d\varepsilon_{3i} \times$$

$$\prod_{\substack{y_{1i}=0\\y_{2i}>0\\y_{3i}>0}} \int_{-\infty}^{-X_{1i}\beta_1} \phi_3\left(\varepsilon_{1i},\varepsilon_{2i},\varepsilon_{3i}\right) d\varepsilon_i \times \prod_{\substack{y_{1i}=1\\y_{2i}>0\\y_{3i}>0}} \int_{-X_{1i}\beta_1}^{\infty} \phi_3\left(\varepsilon_{1i},\varepsilon_{2i},\varepsilon_{3i}\right) d\eta_i =$$

$$\prod_{\substack{y_{1i}=0\\y_{2i}=0\\y_{3i}=0}} \int_{-\infty}^{-X_{3i}\beta_3}\int_{-\infty}^{-X_{2i}\beta_2}\int_{-\infty}^{-X_{1i}\beta_1} \phi_3\left(\varepsilon_{1i},\varepsilon_{2i},\varepsilon_{3i}\right) d\varepsilon_{1i}d\varepsilon_{2i}d\varepsilon_{3i} \times$$

$$\prod_{\substack{y_{1i}=0\\y_{2i}>0\\y_{3i}>0}} \phi_2\left(\varepsilon_{2i},\varepsilon_{3i}\right) \int_{-\infty}^{-X_{1i}\beta_1} \phi_{1|2,3}\left(\varepsilon_{1i}\,|\,\varepsilon_{2i},\varepsilon_{3i}\right) d\varepsilon_i \times \prod_{\substack{y_{1i}=1\\y_{2i}>0\\y_{3i}>0}} \phi_2\left(\varepsilon_{2i},\varepsilon_{3i}\right) \int_{-X_{1i}\beta_1}^{\infty} \phi_{1|2,3}\left(\varepsilon_{1i}\,|\,\varepsilon_{2i},\varepsilon_{3i}\right) d\varepsilon_i$$

where $\boldsymbol{\theta} = \left(\beta_1,\gamma_2,\beta_2,\gamma_3,\beta_3\right)$, $\phi_m\left(\boldsymbol{\varepsilon}\right) = \left(2\pi\right)^{-m/2}\left|\Sigma\right|^{-1/2}\exp\left(-\frac{\boldsymbol{\varepsilon}'\Sigma^{-1}\boldsymbol{\varepsilon}}{2}\right)$ is the m-dimensional normal pdf and $\phi_{j|k,l}\left(\varepsilon_{ji}\,|\,\varepsilon_{ki},\varepsilon_{li}\right)$ is the normal pdf of ε_{ji} conditional on $\left(\varepsilon_{ki},\varepsilon_{li}\right)$. After a little algebra the log-likelihood of the observed data can be written as

$$\ell\left(\boldsymbol{\theta},\Sigma\,|\,\boldsymbol{y}\right) = \sum_{\substack{y_{2i}>0\\y_{3i}>0}} \ln\phi_2\left(\frac{\varepsilon_{2i}}{\sigma_{\varepsilon_2}},\frac{\varepsilon_{3i}}{\sigma_{\varepsilon_3}},\rho_{\varepsilon_2\varepsilon_3}\right) +$$

$$\sum_{\substack{y_{1i}=0\\y_{2i}=0\\y_{3i}=0}} \ln\Phi_3\left(-\frac{X_{1i}\beta_1}{1},-\frac{X_{2i}\beta_1}{\sigma_{\varepsilon_2}},-\frac{X_{3i}\beta_2}{\sigma_{\varepsilon_3}},\rho_{\varepsilon_1\varepsilon_2},\rho_{\varepsilon_1\varepsilon_3},\rho_{\varepsilon_2\varepsilon_3}\right) +$$

$$\sum_{\substack{y_{1i}=0\\y_{2i}>0\\y_{3i}>0}} \ln\Phi_1\left(-\frac{X_{1i}\beta_1 + \frac{\rho_{\varepsilon_1\varepsilon_2}-\rho_{\varepsilon_1\varepsilon_3}\rho_{\varepsilon_2\varepsilon_3}}{1-\rho^2_{\varepsilon_2\varepsilon_3}}\frac{\varepsilon_{2i}}{\sigma_{\varepsilon_2}} + \frac{\rho_{\varepsilon_1\varepsilon_3}-\rho_{\varepsilon_1\varepsilon_2}\rho_{\varepsilon_2\varepsilon_3}}{1-\rho^2_{\varepsilon_2\varepsilon_3}}\frac{\varepsilon_{3i}}{\sigma_{\varepsilon_3}}}{\sqrt{\left(1-\rho^2_{\varepsilon_1\varepsilon_2}-\rho^2_{\varepsilon_1\varepsilon_3}-\rho^2_{\varepsilon_2\varepsilon_3}+2\rho_{\varepsilon_1\varepsilon_2}\rho_{\varepsilon_1\varepsilon_3}\rho_{\varepsilon_2\varepsilon_3}\right)/\left(1-\rho^2_{\varepsilon_2\varepsilon_3}\right)}}\right) + \tag{21}$$

$$\sum_{\substack{y_{1i}=1\\y_{2i}>0\\y_{3i}>0}} \ln\Phi_1\left(\frac{X_{1i}\beta_1 + \frac{\rho_{\varepsilon_1\varepsilon_2}-\rho_{\varepsilon_1\varepsilon_3}\rho_{\varepsilon_2\varepsilon_3}}{1-\rho^2_{\varepsilon_2\varepsilon_3}}\frac{\varepsilon_{2i}}{\sigma_{\varepsilon_2}} + \frac{\rho_{\varepsilon_1\varepsilon_3}-\rho_{\varepsilon_1\varepsilon_2}\rho_{\varepsilon_2\varepsilon_3}}{1-\rho^2_{\varepsilon_2\varepsilon_3}}\frac{\varepsilon_{3i}}{\sigma_{\varepsilon_3}}}{\sqrt{\left(1-\rho^2_{\varepsilon_1\varepsilon_2}-\rho^2_{\varepsilon_1\varepsilon_3}-\rho^2_{\varepsilon_2\varepsilon_3}+2\rho_{\varepsilon_1\varepsilon_2}\rho_{\varepsilon_1\varepsilon_3}\rho_{\varepsilon_2\varepsilon_3}\right)/\left(1-\rho^2_{\varepsilon_2\varepsilon_3}\right)}}\right)$$

where $\Phi_m\left(\cdot\right)$ is the m-dimensional standard normal cdf.

In this study the log-likelihood function in (21) was maximized using the routine FMINUNC in Matlab. I programmed the 3-dimensional standard normal

cdf according to the methodology proposed by Steck (1958), which allows reducing the 3-dimensional integral to functions involving only 1-dimensional integrals of exponential functions and the univariate normal cdf. The information matrix was calculated from a finite-difference estimation of the Hessian of the objective function.

Starting values in the approximation area of the maximum were very hard to find. A first attempt using OLS estimates failed to converge. A second approach attempted to estimate the equations in (2). by pairs and then use a combination of the resulting estimates as starting values for the 3-equation system. Matlab routines using OLS estimates as starting values were written to estimate these smaller equation systems. However, although convergence in the system constituted by the second and third equations was easily accomplished, neither the routine for the system constituted by the first and second equations nor the one for the system constituted by the first and third equations converged. The routines either ceased to improve in the search for the optimum or the correlation between the disturbances escaped the parameter space.

Finally, a grid search was implemented. To decide the dimension of the grid, estimation attempts were made by fixing one, two, three and four parameters in Σ. Convergence was attained only after one of the two variances and all the correlation coefficients were fixed. Thus, a four-dimensional grid search was implemented on those parameters. Ten equally spaced points from –0.9 to 0.9 were chosen for the correlation coefficients and four equally spaced points from 2 to 8 were taken for σ_{ε_2}, which generated a 4,000-point grid. The time required solving a grid of this size easily becomes unaffordable when the objective function involves high dimensional integrals as in (21); however, the actual number of grid points that need to be used can be drastically reduced by two ways. First, it must be noticed that Σ must be kept positive definite at every moment during the estimation. Many points in the grid described above do not satisfy that requirement and they must be eliminated from the search. Second, since we are interested in finding a neighborhood of the global maximum only, by careful monitoring of the search it is possible to exclude large sets of grid points surrounding low values of the objective function. Notice that this last approach is advisable only when the objective function behaves smoothly. A real risk of missing the global maximum exists otherwise. By proceeding this way the grid search used in this study required less than 350 points to locate a point in the approximation area of the global maximum. However, despite of the significant reduction in the size of the grid, it took about 60 hours[2] to

[2] On an AMD Athlon XP-M 2000+, 512 MB RAM, Windows XP, Matlab 6.5.

solve the grid search and make the final estimation to obtain the maximum maximorum.. Results[3] of the final estimation are presented in Table 1.

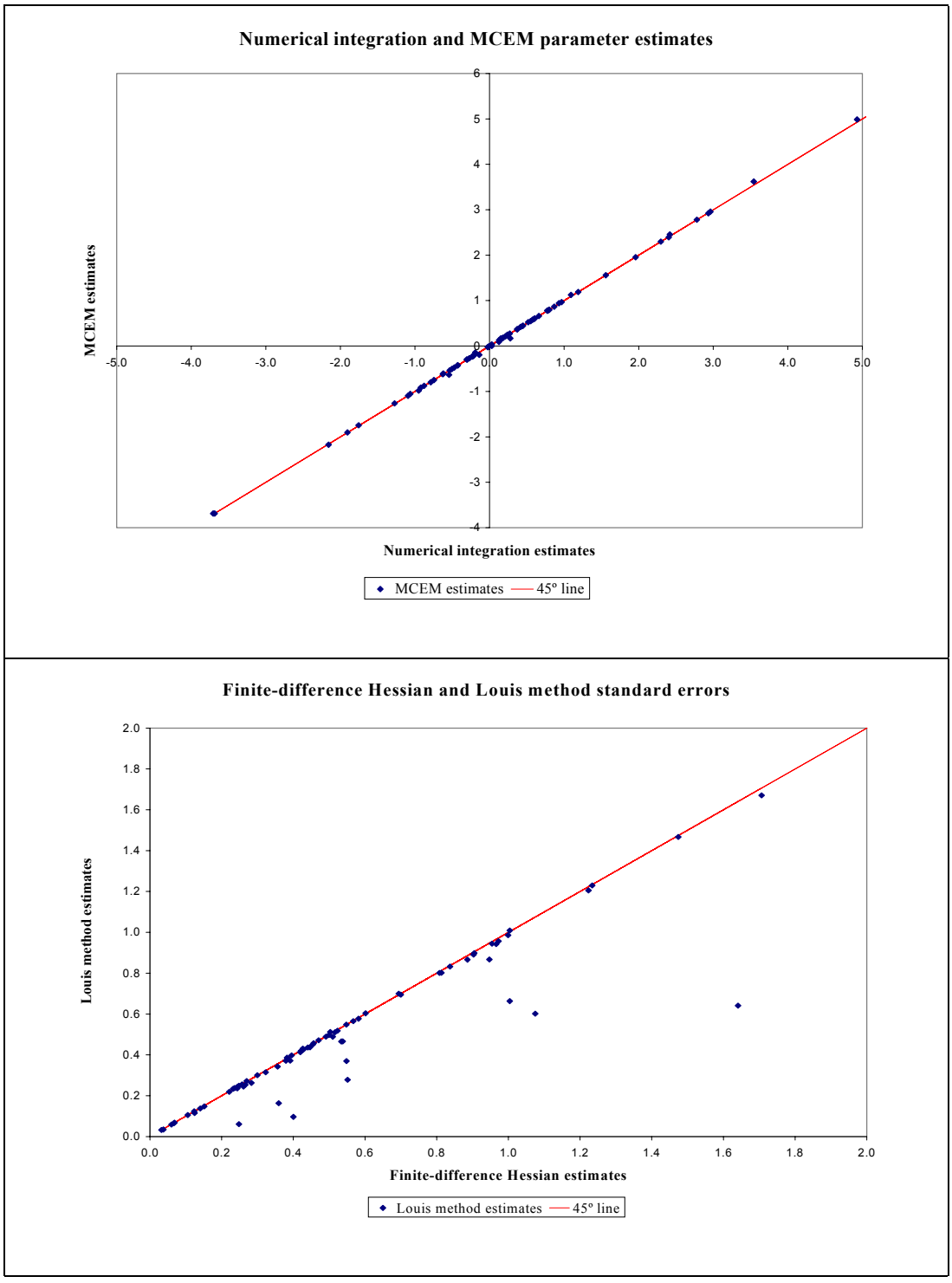
[3] I have not included the variable names since these results are presented only for comparison purposes. Details can be found in Lichtenberg and Smith-Ramirez (2004).

Table 1
ML estimates obtained by Numerical Integration and Monte Carlo EM algorithm

Equation 1					Equation 2					Equation 3				
Parameter	Numerical Integration		MCEM		Parameter	Numerical Integration		MCEM		Parameter	Numerical Integration		MCEM	
	Estim.	St. err.	Estim.	St. err.		Estim.	St. err.	Estim.	St. err.		Estim.	St. err.	Estimate	St. err.
					γ_2	-3.7074	1.0037	-3.6860	0.6633	γ_3	-0.9229	0.5486	-0.9163	0.3701
$\beta_{1,1}$	-0.7881	0.8856	-0.7987	0.8661	$\beta_{2,1}$	5.6232	1.7064	5.6173	1.6708	$\beta_{3,1}$	2.4189	0.8131	2.4572	0.8024
$\beta_{1,2}$	-0.2282	0.1515	-0.2259	0.1481	$\beta_{2,2}$	-0.8770	0.2572	-0.8756	0.2557	$\beta_{3,2}$	-0.2127	0.1234	-0.2175	0.1237
$\beta_{1,3}$	0.2126	0.1246	0.2108	0.1156	$\beta_{2,3}$	0.5992	0.2212	0.5981	0.2189	$\beta_{3,3}$	0.2409	0.1055	0.2474	0.1055
$\beta_{1,4}$	0.5194	0.2436	0.5245	0.2364	$\beta_{2,4}$	0.7933	0.5164	0.7917	0.5109	$\beta_{3,4}$	0.1260	0.2447	0.0919	0.2452
$\beta_{1,5}$	0.1809	0.0598	0.1807	0.0591	$\beta_{2,5}$	0.1665	0.1404	0.1656	0.1376	$\beta_{3,5}$	-0.0077	0.0670	-0.0045	0.0660
$\beta_{1,6}$	-0.0254	0.0374	-0.0255	0.0349	$\beta_{2,6}$	0.1776	0.0683	0.1777	0.0682	$\beta_{3,6}$	0.1279	0.0322	0.1279	0.0326
$\beta_{1,7}$	-0.6267	0.4196	-0.6193	0.4132	$\beta_{2,7}$	-0.3023	0.9046	-0.2999	0.8983	$\beta_{3,7}$	0.2795	0.4265	0.1697	0.4310
$\beta_{1,8}$	0.6602	0.2999	0.6603	0.3009	$\beta_{2,8}$	2.9621	0.9029	2.9580	0.8907	$\beta_{3,8}$	0.1504	0.4280	0.1656	0.4259
$\beta_{1,9}$	0.2370	0.4209	0.2368	0.4196	$\beta_{2,9}$	0.9667	0.9991	0.9667	0.9861	$\beta_{3,9}$	0.0294	0.4704	0.0008	0.4715
$\beta_{1,10}$	-0.5050	0.4405	-0.5042	0.4364	$\beta_{2,10}$	-0.4288	0.9658	-0.4272	0.9424	$\beta_{3,10}$	-0.9509	0.4544	-0.9809	0.4512
$\beta_{1,11}$	-0.4257	0.4467	-0.4211	0.4372	$\beta_{2,11}$	1.1889	0.9724	1.1912	0.9567	$\beta_{3,11}$	-0.1385	0.4567	-0.1933	0.4570
$\beta_{1,12}$	0.7764	0.5478	0.7765	0.5484	$\beta_{2,12}$	1.9579	1.4739	1.9549	1.4671	$\beta_{3,12}$	1.0929	0.6942	1.1256	0.6993

$\beta_{1,13}$	1.5578	0.6997	1.5608	0.6946	$\beta_{2,13}$	2.4029	2.0006	2.3937	1.9660	$\beta_{3,13}$	0.3721	0.9544	0.3632	0.9447
$\beta_{1,14}$	0.8016	0.3230	0.7989	0.3147	$\beta_{2,14}$	2.7804	0.5673	2.7797	0.5651	$\beta_{3,14}$	0.9317	0.2699	0.9430	0.2720
$\beta_{1,15}$	-0.3057	0.3915	-0.2967	0.3720	$\beta_{2,15}$	0.8671	0.8077	0.8654	0.8013	$\beta_{3,15}$	-0.5440	0.3825	-0.6334	0.3864
$\beta_{1,16}$	-0.5323	0.3793	-0.5337	0.3714	$\beta_{2,16}$	2.2971	0.8371	2.2984	0.8327	$\beta_{3,16}$	0.5500	0.3952	0.5474	0.3974
$\beta_{1,17}$	-0.2323	0.2829	-0.2327	0.2629	$\beta_{2,17}$	0.6061	0.4912	0.6062	0.4891	$\beta_{3,17}$	0.0311	0.2321	0.0364	0.2338
$\beta_{1,18}$	-0.7459	0.2661	-0.7483	0.2534	$\beta_{2,18}$	-0.4733	0.5233	-0.4723	0.5186	$\beta_{3,18}$	-1.0625	0.2484	-1.0542	0.2485
$\beta_{1,19}$	0.3718	0.2607	0.3725	0.2449	$\beta_{2,19}$	0.4493	0.4998	0.4481	0.4973	$\beta_{3,19}$	-0.6189	0.2362	-0.6046	0.2378
$\beta_{1,20}$	0.4153	0.5098	0.4219	0.4890	$\beta_{2,20}$	0.4398	1.2337	0.4407	1.2298	$\beta_{3,20}$	-1.2735	0.6017	-1.2638	0.6035
$\beta_{1,21}$	-2.1581	0.9470	-2.1740	0.8672	$\beta_{2,21}$	0.5753	1.2234	0.5797	1.2053	$\beta_{3,21}$	-0.1927	0.5817	-0.1420	0.5771
$\beta_{1,22}$	-1.0930	1.0749	-1.0957	0.6018	$\beta_{2,22}$	-3.6848	1.0036	-3.6882	1.0091	$\beta_{3,22}$	-1.7546	0.5032	-1.7449	0.5119
$\beta_{1,23}$	-1.9051	0.5381	-1.9013	0.4659										
$\beta_{1,24}$	0.2723	0.3561	0.2740	0.3431										
$\beta_{1,25}$	-0.2717	0.5340	-0.2715	0.4657										
σ_{12}	2.9339	0.4003	2.9217	0.0966										
σ_{13}	0.7756	0.2481	0.7823	0.0609	σ_{23}	4.9290	0.5513	4.9887	0.2782					
σ_{22}	16.0218	1.6406	16.001	0.6411	σ_{33}	3.5432	0.3591	3.6201	0.1633					

Figure 2
Comparison between numerical integration and MCEM estimates.
Flexibles.



Graphical comparisons between the estimates obtained by numerical integration and those obtained by the MCEM algorithm are depicted in Figure 2. The remarkable match in the parameter estimates proves that both approaches converged to the same maximizer. The main difference, of course, is the robustness of the MCEM to the election of starting values, which allowed achieving the solution in less than one-twentieth of the time needed when using numerical integration. Certainly, the use of QMC integration or the use of probability simulators may help to reduce the estimation time when using the numerical integration approach. Yet, these methods only provide an alternative to estimate the integral terms in the likelihood function. They do not help neither with problems of "practical" identification nor with the starting value problem.

The matching for the standard errors in table 1, however, is not that close. The lower graph in Figure 2 shows that, in general, the standard errors obtained by using a finite-difference Hessian are larger than those produced by Louis' method. In an attempt to reduce the disparity, the simulation size used to estimate the information matrix in the Monte Carlo approach was enlarged from $R = 3300$ to $R = 5300$. However, no significant change was observed in the estimates.

In order to determine if the origin of the mismatch was in the Hessian estimated by finite differences, the Hessian estimation was repeated several times reducing iteratively the size of the perturbation size[4]. This approach reduced the mismatch, which suggests that the origin of the problem is in the numerical estimation of the Hessian. The size of the perturbation, however, cannot be reduced arbitrarily. The numerical integration approach used to estimate the likelihood function in (21) may be free of Monte Carlo error but it has inaccuracies originated in the numerical integration procedure. This fact sets a lower bound for the size of the perturbation that we can use to estimate the Hessian: it cannot be smaller than the estimation error of the likelihood function. The accuracy of the numerical integrals can, certainly, be increased; however, it is well known that the computational costs of proceeding that way increment exponentially. Figure 2 presents the standard errors for a percentage perturbation equal to 10^{-4}.

The limitations of the numerical Hessian approach contrast with the advantages of the stochastic version of the Louis' method. Louis' method is easy to implement and we can use it to obtain standard errors with any needed accuracy simply by increasing the number of simulations. Doing this is relatively inexpensive since the score and the Hessian of the expectation in (5) exists in closed form and their calculation involves only matrix algebra.

[4] By perturbation size I mean the magnitude of the finite difference used to calculate the numerical derivatives

4.- Implementing the Monte Carlo EM algorithm for models with latent endogenous regressors

This section extends the MCEM algorithm presented in the last section in order to include systems of structural equations, i.e. cases where latent variables show up on both sides of the equations. Only for the purpose of illustrating the flexibility of the method, consider again the equation system (2) but now with y_{1i}^* instead of y_{1i} at the right-hand side of the second and third equations, and y_{1i} defined as a polytomous variable,

$$y_{1i} = \begin{cases} b_1 & \text{if } a_0 < y_{1i}^* \le a_1 \\ b_2 & \text{if } a_1 < y_{1i}^* \le a_2 \\ \vdots & \\ b_k & \text{if } a_{k-1} < y_{1i}^* \le a_k \end{cases} \qquad (22)$$

The structural model is now,

$$\begin{aligned} y_{1i}^* &= X_{1i}\beta_1 + \varepsilon_{1i} \\ y_{2i}^* &= \gamma_2 y_{1i}^* + X_{2i}\beta_2 + \varepsilon_{2i} \\ y_{3i}^* &= \gamma_3 y_{1i}^* + X_{3i}\beta_3 + \varepsilon_{3i} \end{aligned} \qquad (23)$$

Thus, under the normality assumption, the complete data likelihood function can be written as,

$$L(\boldsymbol{\beta}, \Gamma, \Sigma \mid \boldsymbol{x}) = \prod_i f\left(y_{1i}^*, y_{2i}^*, y_{3i}^*\right) = \prod_i \left[\frac{1}{(2\pi)^{3/2} |\Gamma| |\Sigma|^{1/2}} \exp\left(-\frac{\varepsilon_i' \Sigma^{-1} \varepsilon_i}{2} \right) \right]$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$, $\Gamma = \begin{bmatrix} 1 & -\gamma_2 & -\gamma_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\varepsilon_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} = \begin{pmatrix} y_{1i}^* - X_{1i}\beta_1 \\ y_{2i}^* - \gamma_2 y_{1i}^* - X_{2i}\beta_2 \\ y_{3i}^* - \gamma_3 y_{1i}^* - X_{3i}\beta_3 \end{pmatrix}$, and Σ is defined as in (3).

Correspondingly, the complete information log-likelihood function and its expectation are:

$$\ell^c(\boldsymbol{\beta}, \Gamma, \Sigma \mid \boldsymbol{x}) = -\frac{3N}{2}\ln(2\pi) - N\ln|\Gamma| - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_i \operatorname{tr}\left(\Sigma^{-1}\varepsilon_i\varepsilon_i'\right) \qquad (24)$$

$$E\left[\ell^c\left(\boldsymbol{\beta},\Gamma,\Sigma\mid \boldsymbol{x}\right)\right]=-\frac{3N}{2}\ln\left(2\pi\right)-N\ln\left|\Gamma\right|-\frac{N}{2}\ln\left|\Sigma\right|-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_i E\left[\varepsilon_i\varepsilon_i^{'}\right]\right) \qquad (25)$$

Where N is the total number of observations and the expectation operator indicates expectation conditional on observed information and distribution assumptions. The E-step at iteration $m+1$ requires the calculation of,

$$\begin{aligned} Q_i\left(\boldsymbol{\beta},\Gamma,\Sigma\mid\boldsymbol{\beta}^{(m)},\Gamma^{(m)},\Sigma^{(m)},\boldsymbol{y}\right) &= E\left[\varepsilon_i\varepsilon_i^{'}\mid\boldsymbol{\beta}^{(m)},\Gamma^{(m)},\Sigma^{(m)},\boldsymbol{y}\right] \\ &= E\left[\begin{pmatrix} y_{1i}^*-X_{1i}\beta_1 \\ y_{2i}^*-\gamma_2 y_{1i}^*-X_{2j}\beta_2 \\ y_{3i}^*-\gamma_3 y_{1i}^*-X_{3j}\beta_3\end{pmatrix}\begin{pmatrix} y_{1i}^*-X_{1i}\beta_1 \\ y_{2i}^*-\gamma_2 y_{1i}^*-X_{2j}\beta_2 \\ y_{3i}^*-\gamma_3 y_{1i}^*-X_{3j}\beta_3\end{pmatrix}^{'}\mid\boldsymbol{\beta}^{(m)},\Gamma^{(m)},\Sigma^{(m)},\boldsymbol{y}\right] \qquad (26) \\ &= \Gamma^{'}\sigma_i^{2(m)}\Gamma \;+\; \begin{pmatrix} \mu_{y_{1i}^*}^{(m)}-X_{1i}\beta_1 \\ \mu_{y_{2i}^*}^{(m)}-\gamma_2\mu_{y_{1i}^*}^{(m)}-X_{2i}\beta_2 \\ \mu_{y_{3i}^*}^{(m)}-\gamma_3\mu_{y_{1i}^*}^{(m)}-X_{3i}\beta_3\end{pmatrix}\begin{pmatrix} \mu_{y_{1i}^*}^{(m)}-X_{1i}\beta_1 \\ \mu_{y_{2i}^*}^{(m)}-\gamma_2\mu_{y_{1i}^*}^{(m)}-X_{2i}\beta_2 \\ \mu_{y_{3i}^*}^{(m)}-\gamma_3\mu_{y_{1i}^*}^{(m)}-X_{3i}\beta_3\end{pmatrix}^{'} \end{aligned}$$

Where $\sigma_i^{2(m)}$ and $\mu_{y_{ji}^*}^{(m)}$ are defined as in expressions (7) and (8). Some small modifications must be introduced to the Gibbs sampler implemented in section 3.2 in order to estimate these moments. The conditional mean $\mu_{ji|i(-j)}^{(m)}$ must be now estimated according to,

$$\mu_{ji|i(-j)}^{(m)}=X_{ji}\beta_j^{(m)}+\mathrm{cov}\left(y_{ji}^*\middle|\boldsymbol{y}_{i|-j}^*,\Sigma^{(m)}\right)\left[\mathrm{cov}\left(\boldsymbol{y}_{i|-j}^*\middle|\Sigma^{(m)}\right)\right]^{-1}\left(\boldsymbol{y}_{i|-j}^*-\boldsymbol{y}_{i|-j}^*\boldsymbol{\gamma}_{-j}^{(m)}-\boldsymbol{X}_{i|-j}\boldsymbol{\beta}_{-j}^{(m)}\right)$$

where, $$\boldsymbol{\gamma}_{-j}^{(m)}=\begin{pmatrix}\gamma_1^{(m)}\\ \vdots \\ \gamma_{j-1}^{(m)}\\ \gamma_{j+1}^{(m)}\\ \vdots \\ \gamma_k^{(m)}\end{pmatrix}$$

To construct a sample conditional on the observed information as defined in (22) proceed as follow. For every $a_\kappa < y_{1i} \le a_{\kappa+1}$ simulate y_{1i}^* from a normal distribution with mean $\mu_{1i|i(-1)}^{(m)}$ and variance $\sigma_{1i|-1}^{2(m)}$ truncated below at a_κ and truncated above at $a_{\kappa+1}$. Do the same for every $\kappa=0,1,2,\ldots k$, where k is the number of intervals defined by the polytomous variable.

Simulations for the unobserved values of y_{2i}^{*} and y_{3i}^{*} are obtained in the same way as in Section 2.2.

The maximization step does not differ significantly from the case analyzed previously except by the presence of Γ in the log-likelihood function, which motivates a slight change in the arguments of the conditional maximization steps. The objective function is,

$$-\frac{3N}{2}\ln(2\pi)-N\ln|\Gamma|-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i}Q_i\left(\boldsymbol{\beta},\Gamma,\Sigma\,|\,\boldsymbol{\beta}^{(m)},\Gamma^{(m)},\Sigma^{(m)},\boldsymbol{y}\right)\right) \qquad (27)$$

The first conditional maximization updates $\boldsymbol{\beta}$ conditional on the elements in Γ and Σ. From (26) the estimate of $\boldsymbol{\beta}$ can still be written as a generalized least squares estimator:

$$\boldsymbol{\beta}^{(m+1)}=\left[X_d'\left(\Sigma^{(m)}\otimes I_N\right)^{-1}X_d\right]^{-1}X_d'\left(\Sigma^{(m)}\otimes I_N\right)^{-1}\hat{\mu}_{y^*}^{(m)} \qquad (28)$$

where,

$X_d=\begin{bmatrix}X_1 & 0 & 0\\ 0 & X_2 & 0\\ 0 & 0 & X_3\end{bmatrix}$, $\hat{\mu}_{y^*}^{(m)}=\left(\Gamma'^{(m)}\otimes I_N\right)\mu_{y^*}^{(m)}$, I_N is the identity matrix of dimension N and $\mu_{y^*}^{(m)}$ is the column vector defined in (15).

The second conditional maximization updates Γ and Σ conditional on the updated estimate of $\boldsymbol{\beta}$. Numerical techniques must be used in this step to maximize,

$$-\frac{3N}{2}\ln(2\pi)-N\ln|\Gamma|-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i}Q_i\left(\boldsymbol{\beta}^{(m+1)},\Gamma\,|\,\boldsymbol{\beta}^{(m)},\Gamma^{(m)},\Sigma^{(m)},\boldsymbol{y}\right)\right)$$

with respect to the elements in Γ and Σ and obtain estimates for $\Gamma^{(m+1)}$ and $\Sigma^{(m+1)}$. Notice that the second term in the objective function vanishes in this particular example as Γ is triangular.

It is evident that the combination of the Expectation and Maximization steps as described here can be extended to admit systems with a larger number of linear equations involving any type of latent variables. Only small adjustments to the Gibbs sampler in order to take into account the types of latent variables involved are necessary.

Conclusions

This article has presented a MCEM algorithm suitable for estimating systems of simultaneous equations and structural models that contain latent variables. The applicability of the model is independent of whether the latent variables appear in the model as dependent variables or as endogenous regressors. The general formulation presented in Section 4 permits that the algorithm can be applied to solve a variety of models with latent structures from a one-equation tobit to a n-equation multinomial probit. Only small adjustments in the Gibbs sampler are necessary to shift from one model to another in order to internalize the type of latent variables involved and the nature of the unobserved information in the different models.

The MCEM algorithm as formulated in this article has a number of advantages over more traditional methods. First, it does not require integrating the unobserved information out from the likelihood function. This characteristic reduces the estimation time dramatically as no numerical integration is needed and, similarly to methods based on probability simulators, permits to solve problems involving more than three latent variables. Second, it reduces the estimation of the vector of slopes to the calculation of a GLS estimator and numerical optimization is required only to estimate the elements in the disturbance covariance matrix. Since the GLS estimator and the gradient and Hessian of the objective function for the estimation of the disturbance covariance matrix have closed forms, almost no time is consumed in the Maximization step and it is easier to keep the whole set of parameters in the parameter space. This property of the MCEM reduces substantially the problems of "fragile" identification and selection of starting values, which are serious limitations of traditional approaches. Third, it can accommodate potentially any linear-in-parameters equation system. This is valid not only for the cross-sectional models mentioned above but also for panel data models and stochastic frontier models, where the random effects and efficiency terms can be treated as one more latent variable. Finally, the estimation of standard errors by the Louis method circumvents the limitations associated to the estimation of numerical Hessians by finite-difference methods, which is the standard in traditional procedures. The accuracy of the estimates of the standard errors can be improved easily by increasing the number of simulations of a closed form of the Information matrix, which is much less expensive than reducing the perturbation size in the numerical Hessian approach.

Bibliografía

Booth, J. and Hobert, J. (1999), "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm", J. R. Statist. Soc. B. 61 Part 1, pp. 265-285.

Börsch-Supan, A. and V. Hajivassiliou (1993), "Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models", J. of Econometrics 58, pp. 347-368.

Boyles, R. (1983), "On the convergence of the EM algorithm", J. R. Statist. Soc. B, 45(1), pp. 47-50.

Casella, G. and George, E. (1992), "Explaining the Gibbs Sampler", The American Statistician. 46(3), pp. 167-174.

Chan, K. And Ledolter, J. (1995), "Monte Carlo EM estimation for time series models involving counts", J. Am. Statist. Assoc., 90(429), pp. 242-252.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), *Maximum Likelihood Estimation from incomplete observations*, J. Roy. Statist. Soc. B 39, pp:1-38.

Devroye, Luc. (1986), *Non-uniform random variate generation*, Springer-Verlag. 843 pps.

Dickens, W. and Lang, K. (1985), "A Test of Dual Labor Market Theory", The American Economic Review 75(4), pp. 792-805.

Eickhoff, J. Zhu, J. Amemiya, Y. (2004), "On the simulation size and the convergence of the Monte Carlo EM algorithm via likelihood-based distances", Statistics & Probability Letters 67, pp. 161-171.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, pp. 721-741.

Genz, A. (2004), "Numerical computation of rectangular bivariate and trivariate normal and t probabilities", To appear in *Statistics and Computing*.

Geweke, J.; M. Keane, and D. Runkle. (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model", Review of Economics and Statistics, 76, pp. 609-632.

Halton, J.H. (1960), "On the eficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals", Numer. Math. 2,pp. 84-90.

Heckman, J. (1976), "Simultaneous equation models with continuos and discrete endogenous variables and structural shifts", In S.M. Goldfeld and R. E. Quandt (eds.), *Studies in non-linear estimation*, Cambridge. 278 pps.

Ibrahim, J.G., Chen, M., and Lipsitz, S.R. (2001), *Missing Responses in Generalized Linear Mixed Models when the Missing Data Mechanism is Nonignorable*, Biometrika, 88(2), pp. 551-564.

Jank, W. (2004), "Quasi-Monte Carlo Sampling to improve the Efficiency of Monte Carlo EM", Forthcoming at Computational Statistics and Data Analysis.

Keane, Michael P. (1992), "A Note on Identification in the Multinomial Probit Model", Journal of Business and Economic Statistics 10, pp. 193-200.

Kendall, M. and and Stuart, A. (1978), "The advanced Theory of statistics", McMillan, New York.

Lichtenberg, E. and Smith-Ramirez, R. (2004), *Cost Sharing Transaction Costs and Conservation*, Mimeo, Agr. and Res. Ec. Dep., University of Maryland, College Park.

Louis, T.A. (1982), *Finding the Observed Information Matrix when using the EM Algorithm*, J. Roy. Statist. Soc. B 44, pp. 226-233.

Maddala, G. (1983), "Limited-dependent and qualitative variables in econometrics", Econometric Society Monographs. 401 pps.

McCulloch, C. (1994), "Maximum likelihood variance components estimation for binary data", J. Am. Statist. Assoc. 89, pp. 330-335.

McCulloch, C. (1997), "Maximum likelihood algorithms for generalized linear mixed models", J. Am. Statist. Assoc. 92, pp. 162-170.

McFadden, D. (1989), "A Method of Simulated Moments for estimation of Discrete Response Models without Numerical Integration", Econometrica 57, pp. 995-1026.

Meng, X. And Rubin, D. (1993), "Maximum likelihood estimation via the ECM algorithm: a general framework", Biometrika 80, pp. 267-278.

Moler, C. (1995), "Random thoughts. 10435 years is a very long time", Matlab News and Notes, Fall 12-13.

Natarajan, Ranjini; McCulloch, Charles E. and Nicholas Kiefer (2000), "A Monte Carlo EM Method for estimating Multinomial Probit Models", Computational Statistics and Data Analysis 34, pp. 33-50.

Orchard, T. and Woodbury, M.A. (1972), *A Missing Information Principle: Theory and Applications*, In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability Vol. 1. Berkeley, California, University of California Press, pp. 697-715.

Sobol, I.M. (1998), "On quasi-Monte Carlo integration", Math. Comput. Simulation 47, pp. 103-112.

Sowden, R. And Ashford, J. (1969), "Computation of the bi-variate normal integral", Applied Statistics, 18(2), pp. 169-180.

Steck, G.P. (1958), "A Table for Computing Trivariate Normal Probabilities", Ann. Math. Statisc. 29, pp. 780-800.

Wei, C. and Tanner, M. (1990), "A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation Algorithms", J. Am. Statist. Assoc., 85(411), pp. 699-704.

Wu, C.F.J. (1983), "On the convergence properties of the EM algorithm", The Annals of Statistics. 11(1), pp. 95-108.