Journal of Computational and Applied Mathematics 366 (2020) 112378

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Multilevel simultaneous equation model: A novel specification and estimation approach

Rocío Hernández-Sanjaime *, Martín González, Jose J. López-Espín

Center of Operations Research, Miguel Hernández University, Elche, Alicante, Spain

ARTICLE INFO

Article history:
Received 25 February 2019
Received in revised form 31 July 2019

Keywords:
Multilevel simultaneous equation model
Maximum likelihood estimation
Matrix normal distribution
Simultaneous equation model
Multilevel model

ABSTRACT

Conventional simultaneous equation models assume that the error terms are serially independent. In some situations, data may present hierarchical or grouped structure and this assumption may be invalid. A new multivariate model referred as to Multilevel Simultaneous Equation Model (MSEM) is developed under this motivation. The maximum likelihood estimation of the parameters of an MSEM is considered. A matrix-valued distribution, namely, the matrix normal distribution, is introduced to incorporate an among-row and an among-column covariance matrix structure in the specification of the model. In the absence of an analytical solution of the system of likelihood equations, a general-purpose optimization solver is employed to obtain the maximum likelihood estimators. In a first approach to the solution of the problem, the adequacy of the matrix normal distribution is evaluated empirically in the case in which the double covariance structure is known. Using simulated data under the model assumptions, the performance of the maximum likelihood estimator (MLE) is assessed with regard to other conventional alternatives such as two-stage least squares estimator (2SLS).

1. Introduction

The limitations of classic statistical models to accurately reproduce the complexity of problems in which data are hierarchically structured or there is endogeneity between variables make the use of new methodological techniques necessary. Multilevel models and simultaneous equation models (SEM) have been developed for the statistical analysis of hierarchy and simultaneity, respectively. Nevertheless, models combining endogeneity and hierarchically structured data open a new line of research. The literature handling this mixture of factors is scarce and limited to recursive models.

The present paper addresses the estimation of the parameters of a Multilevel Simultaneous Equation Model (MSEM), that is, a SEM in which observed data are clustered into independent groups. An among-row and among-column covariance matrix structure is considered in order to take into account data correlation within groups. The matrix normal distribution allows to incorporate this specific patterned matrix in the estimation process and seems to be appropriate for this purpose. Further details of this distribution will be described in Section 3. Alternative matrix non-normal distributions, for example the matrix Student-t distribution, have also been assumed in recent studies to estimate parameters incorporating variability among individuals [1].

Previously, matrix normal distribution has been applied to the analysis of multivariate repeated measurements [2]. In this context, one can encompass an m-variate response observed on n occasions, either m variables measured at n time points for one subject or m variables measured for n subjects that belong to the same group, yielding in both situations

* Corresponding author.
E-mail addresses: rocio.hernandezs@umh.es (R. Hernández-Sanjaime), martin.gonzaleze@umh.es (M. González), jlopez@umh.es (J.J. López-Espín).

https://doi.org/10.1016/j.cam.2019.112378
0377-0427/

an $n \times m$ observation matrix X. It should be noted that univariate repeated measurements and growth curves (also known as latent trajectory models where the repeated measurements are viewed as outcomes that depend on some metric of time (e.g. age, day or wave of measurement)) correspond to the $m = 1$ case. In these types of model analysing change, one variable is observed on n occasions and the degenerate matrix normal distribution is applied [3].

In the SEM estimation framework, it is usual to assume that errors are generated by a multivariate procedure with uncorrelated observations. In general, the errors have been supposed to follow a multivariate normal distribution [4]. Multilevel models allow dealing with grouped data but have been scarcely developed for multivariate response [5]. The aim of this paper is to merge multivariate response models with simultaneity and clustered data. Given the double covariance matrix structure, it is possible to bring these two relevant situations together: multivariate response as in a SEM and grouped correlated observations as in multilevel models.

Simultaneous equation models have been traditionally used in Econometrics, the best-known examples are the Klein's Model [6] or the macroeconomic IS-LM models [7]. However, their use has also been recently extended to other fields such as health sciences for modelling complex phenotypes [8] or even transport research for modelling the air traffic in the New York area [9]. Multilevel models have been widely implemented in cross-sectional studies from social and biomedical sciences in which units are naturally grouped at different levels (e.g. students in schools, voters in districts, etc.) or in longitudinal data such as clinical trials, when the same individual or unit response is repeatedly measured at several time points [10–12].

Applications of multilevel simultaneous equation modelling can be mainly found in studies analysing resources allocation in health or educational systems [13,14]. Nonetheless, the approach adopted basically consists of a multilevel model in which the endogeneity of some of the variables is adjusted including a second equation that creates a recursive simultaneous equation model. The extension proposed in this paper would not be confined to recursive models and aims to expand such practical situations.

Under the matrix normal distribution assumption, a random sample of independent and identically distributed (i.i.d.) groups provides the basis to derive the joint density of the new model. The parameters estimation is carried out via the maximum likelihood method. In the absence of a closed solution, a data sample is simulated and the maximum likelihood estimator is examined calling the R optimization function *nlm* from the stats package [15].

The rest of the article is organized as follows: Section 2 includes a brief overview of the statistical models employed and their most relevant characteristics. In Section 3, the MSEM is defined and its estimation via the maximum likelihood method is introduced. The simulation experiment proposed for solving the model in absence of analytic solutions is described in Section 4. This section also summarizes the numerical results obtained by using simulated data. Finally, main conclusions are listed in Section 5.

2. Statistical models

2.1. Simultaneous equation models (SEM)

Simultaneous equation model [16] consists of a system of linear regression equations that reflects the presence of jointly endogenous variables, i.e. the simultaneity between the set of variables of the model. Unlike single-equation models in which a dependent variable is a function of a set of independent variables, a SEM is a multi-equation model in which the dependent variable can appear as an explanatory variable in other equations. Formally, the structural form of the model

$$Y = YA + XB + U \tag{2.1.1}$$

where $Y = [y^1, \ldots, y^m]$ is a $N \times m$ matrix of N observations of m endogenous variables, $X = [x^1, \ldots, x^k]$ is a $N \times k$ matrix of N observations of k non-random predetermined variables which contains both exogenous and lagged endogenous variables, and $U = [u^1, \ldots, u^m]$ is a $N \times m$ matrix of the structural disturbances of the system. The matrices A $(m \times m)$ and B $(k \times m)$ are the endogenous and exogenous unknown coefficient matrices, respectively.[1]

The error terms $u_{t\cdot}$ $(t = 1, \ldots, N)$ are assumed to be serially independent random vectors normally distributed with 0 mean vector and covariance matrix Σ. Thus, the errors may be *contemporaneously* correlated but are *intertemporally* uncorrelated. That is, the rows of U, denoted $u_{t\cdot}$, have the properties:

$$u'_{t\cdot} \sim N(0, \Sigma), \quad E(u'_{t\cdot}, u_{t'\cdot}) = \delta_{tt'}\Sigma \qquad t, t' = 1, 2, \ldots, N \tag{2.1.2}$$

$\delta_{tt'}$ being the Kronecker delta and Σ a positive definite matrix.

Extensions to non-normal errors are possible [see [17] and [18]] but not considered in this work.

In addition, it is assumed that error terms are uncorrelated with the predetermined variables of the system, and there is no linear dependence among the predetermined variables so that the model has a unique interpretation in terms of its unknown parameters:

$$E(X'U) = 0 \quad \text{and} \quad rank(X) = k \tag{2.1.3}$$

[1] By convention, $a_{ii} = 0$, $i = 1, 2, \ldots, m$.

Finally, the coefficient matrix $(I - A)$ is assumed to be non-singular and the reduced form of the system becomes

$$Y = X\Pi + V \quad \text{where} \quad \Pi = B(I - A)^{-1} \quad \text{and} \quad V = U(I - A)^{-1} \tag{2.1.4}$$

The rows of V, $v_{t\cdot}$, are independent identically distributed (i.i.d.) random vectors with 0 mean vector and covariance matrix Ω :

$$\begin{aligned} &v'_{t\cdot} \sim N(0, \Omega), \quad v_{t\cdot} = u_{t\cdot}(I - A)^{-1} \quad \text{and} \quad \Omega = ((I - A)')^{-1}\Sigma(I - A)^{-1} \\ &Cov(v_{t\cdot}{}', v_{t'\cdot}) = \delta_{tt'}\Omega \qquad t, t' = 1, 2, \ldots, N \end{aligned} \tag{2.1.5}$$

The estimation of the parameters of the model can be tackled in two different ways: using limited information methods (single-equation methods) or full information techniques (system methods). The first approach that includes estimators such as indirect least squares (ILS), two-stage least squares (2SLS) or limited information maximum likelihood (LIML) treats each equation of the system in isolation. System methods such us three-stage least squares (3SLS) or full information maximum likelihood (FIML) estimate all the unknown parameters of the system simultaneously [19].

2.2. Multilevel models

Multilevel models, also called hierarchical linear models or linear mixed models among other denominations, are statistical techniques suitable for handling data that have a hierarchical, nested or clustered structure [5]. The existence of such dependent data structures implies that members of the same group share a set of features that derives in an intraclass correlation. Group effects describe how strongly units in the same group tend to resemble and influence each other. The statistical problems of ignoring these relationships may render invalid statistical conclusions [20].

For simplicity's sake, we will consider the 2-level model hereafter. More levels in the model and complex hierarchical structures shall be consulted in [5]. Model specification for multilevel models can be formulated in two different but equivalent approaches. One is based on a single equation that involves both fixed and random effects [21] while the other approach explicitly specifies the model in two levels with two different equations [22]. In this paper, we adopt the former representation expressed as follows:

$$y_i = X_i\beta + Z_i\mu_i + \varepsilon_i \qquad i = 1, \ldots, l \tag{2.2.1}$$

where y_i represents the n_i-response vector for the ith group of n_i individuals in cross-sectional data whereas it represents the n_i repeated measurements of the ith subject in longitudinal data, X_i is the $n_i \times p$ design matrix of the fixed effects, β is the p-vector of the fixed effects coefficients to be estimated, Z_i is the $n_i \times q$ design matrix of the random effects, μ_i is the q-vector of random effects for the ith group and ε_i is the n_i-vector of residuals [23]. It should be noted that X_i combines both level-1 and level-2 explanatory variables and $Z_i's$ columns are a subset of $X_i's$ $(q \leq p)$ incorporating random effects μ_i to y_i. That is, any component of β can be allowed to vary randomly by simply including the corresponding columns of X_i in Z_i [24].

The following distributional assumptions are made:

$$\begin{aligned} &\mu_i \sim N(0, D) \\ &\varepsilon_i \sim N(0, R_i) \\ &\mu_1, \ldots, \mu_l, \varepsilon_1, \ldots, \varepsilon_l \quad \text{independent} \end{aligned} \tag{2.2.2}$$

where μ_i reflects how the subset of regression coefficients for group i deviates from those of the population and ε_i comprises the residuals not explained by fixed or random effects. D and R_i are the covariance matrices of the multivariate normal distributions and l is the total number of groups [25]. The between variance component D is the same for all groups while R_i may vary across units.

Formally, the introduction of random effects helps to distinguish the conditional (group-specific) mean $E(y_i|\mu_i)$ and marginal (population-average) mean $E(y_i)$ as well as group-specific covariance $Cov(y_i|\mu_i)$ and population-average covariance $Cov(y_i)$:

$$\begin{aligned} E(y_i|\mu_i) &= X_i\beta + Z_i\mu_i \\ E(y_i) &= X_i\beta \\ Cov(y_i|\mu_i) &= R_i \\ Cov(y_i) &= Z_iDZ_i{}' + R_i \end{aligned} \tag{2.2.3}$$

Therefore, for each group

$$y_i \sim N(X_i\beta, Z_iDZ_i{}' + R_i) \tag{2.2.4}$$

This model structure allows units of the same group to be positively correlated, i.e. to account for intra-subject variability, and each group to diverge from the population allowing for inter-subject variability [25].

Finally, let consider the general model by stacking up all the groups, y_i, into a single column vector

$$y = X\beta + Z\mu + \varepsilon \tag{2.2.5}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_l \end{bmatrix} \quad Z = \begin{bmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_l \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_l \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_l \end{bmatrix}$$

Hence,

$$\begin{aligned} &y \sim N(X\beta, V) \quad \text{with} \quad V = ZGZ' + R \\ &G = diag(D, D, \dots, D) \quad R = diag(R_1, R_2, \dots, R_l) \end{aligned} \tag{2.2.6}$$

Standard estimation methods in multilevel models are maximum likelihood (ML) [26] and restricted maximum likelihood (REML) [27]. The aim is to calculate the fixed effects coefficients β as well as the variance components involved in V. The estimation of the fixed effects given the variance components is straightforward. Unfortunately, solutions to the variance components are not easy to handle computationally.

Maximum likelihood estimation methods require the maximization of the likelihood function which involves solving nonlinear equations. Historically, obtaining the estimators was a challenging computationally task. Nowadays, most statistical software have integrated routines for linear mixed models estimation in their packages: HLM [22], MLwiN [5] or nlme [23,28].

3. Multilevel simultaneous equation model

3.1. Definition of the Multilevel Simultaneous Equation Model (MESM)

Consider again a simultaneous equation model specified as in (2.1.1), but with observed data clustered into l independent groups

$$Y_j = Y_j A + X_j B + E_j \qquad j = 1, \dots, l \quad \text{independent groups} \tag{3.1.1}$$

Bearing in mind that ignoring groupings may invalidate many of the traditional statistical techniques, model assumptions of a SEM shall be modified. The error terms are no longer generated by a multivariate procedure with *intertemporally* uncorrelated observations. Therefore, distributional assumptions need to be reformulated.

A first approach to deal with this problem is to consider a double covariance matrix structure. The incorporation of the among-row and among-column covariance matrices allows specifying a covariance matrix for the variables and a covariance matrix for the group autocorrelation. This separable variance–covariance structure will provide the error distribution and will lead to more efficient inference.

Prior to introducing the MLE for the model proposed, the matrix normal distribution must be presented. Let X be an $n \times m$ random matrix and M, U, Σ $n \times m$, $n \times n$, $m \times m$ matrices, respectively, with U and Σ non-negative definite. Matrix M will represent the mean of the distribution whereas U and Σ the temporal autocorrelation and contemporaneous covariance matrices, respectively. By definition [29], X follows a matrix normal distribution with parameters M, U and Σ, denoted by $X \sim N_{n,m}(M, U, \Sigma)$, if X has the moment-generating function:

$$M_X(T) = exp\left\{ tr(M'T) + \frac{1}{2} tr(T'UTV) \right\} \quad \text{with } T \text{ an } n \times m \text{ matrix} \tag{3.1.2}$$

An equivalent definition involving the Kronecker product $\otimes$ and the vec operator is specified as:

$$X \sim N_{n,m}(M, U, \Sigma) \quad \text{if} \quad \text{vec}(X) \sim N_{np}(\text{vec}(M), \Sigma \otimes U) \tag{3.1.3}$$

Being U and Σ positive definite matrices, the distribution of X is said to be regular if X has the probability density function

$$f_X(X) = c^{-1} exp\left[-\frac{1}{2} tr\{U^{-1}(X - M)\Sigma^{-1}(X - M)^T\} \right] \tag{3.1.4}$$

with $c = (2\pi)^{nm/2} |U|^{m/2} |\Sigma|^{n/2}$

For model (3.1.1), the condition that each group has the same number of units will be imposed, so that the matrix U is common to all groups. The notation $n_1 = n_2 = \cdots = n_l = n$ will be used hereafter.

Consider again a SEM with clustered data and applying the normal matrix distribution exposed above, for each group it results

$$Y_j = Y_j A + X_j B + E_j \qquad E_j \sim N_{n,m}(0, U, \Sigma) \tag{3.1.5}$$

with 0, U and Σ an $n \times m$, $n \times n$, $m \times m$ matrix, respectively.

And applying some basic properties:

$$Y_j = X_jB(I-A)^{-1} + E_j(I-A)^{-1} \quad \text{and} \quad W_j = Y_j - X_jB(I-A)^{-1} \tag{3.1.6}$$

we have that,

$$W_j \sim N(0, U, ((I-A)^{-1})^T \Sigma (I-A)^{-1}) \tag{3.1.7}$$

3.2. The MLE for a MESM

Under the normality assumption, a random sampling of l groups provides $n \times m$ i.i.d. matrices $E_1, \ldots, E_l$, from which the parameters estimators can be derived using maximum likelihood methods by formulating the appropriate function.

In view of the error distribution (3.1.6) and replacing W_j by the observable quantities $Y_j - X_jB(I-A)^{-1}$, the form of the joint likelihood function is stated by

$$\begin{aligned} f(W_1, \ldots, W_l) &= \prod_{j=1}^{l} f_j(W_j) = (2\pi)^{-\frac{nml}{2}} |U|^{-\frac{ml}{2}} |((I-A)^{-1})^T \Sigma (I-A)^{-1}|^{-\frac{nl}{2}} \\ exp &\left\{ -\frac{1}{2} \sum_{j=1}^{l} tr\left(U^{-1}(Y_j - X_jB(I-A)^{-1})(I-A)\Sigma^{-1}(I-A)^T(Y_j - X_jB(I-A)^{-1})^T \right) \right\} \end{aligned} \tag{3.2.1}$$

The logarithm of the likelihood function, $L = log\, f(W_1, \ldots, W_l)$, is given by

$$\begin{aligned} L = &-\frac{nml}{2} ln(2\pi) - \frac{ml}{2} ln|U| - \frac{nl}{2} ln|((I-A)^{-1})^T \Sigma (I-A)^{-1}| \\ &- \frac{1}{2} \sum_{j=1}^{l} tr\left(U^{-1}(Y_j - X_jB(I-A)^{-1})(I-A)\Sigma^{-1}(I-A)^T(Y_j - X_jB(I-A)^{-1})^T \right) \end{aligned} \tag{3.2.2}$$

The application of matrix derivatives [30–32] provides the system of likelihood equations:

$$\begin{aligned} \frac{\partial L}{\partial U} = &-mlU^{-1} + \frac{ml}{2} diag(U^{-1}) + \sum_{j=1}^{l} \left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1} \right) \\ &-\frac{1}{2} \sum_{j=1}^{l} diag\left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1} \right) = 0 \end{aligned} \tag{3.2.3}$$

$$\begin{aligned} \frac{\partial L}{\partial \Sigma} = &-nl\Sigma^{-1} + \frac{nl}{2} diag(\Sigma^{-1}) + \sum_{j=1}^{l} \left(\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1} \right) \\ &-\frac{1}{2} \sum_{j=1}^{l} diag\left(\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1} \right) = 0 \end{aligned} \tag{3.2.4}$$

$$\frac{\partial L}{\partial B} = \sum_{j=1}^{l} (X_j^T U^{-1} Y_j)(I-A)\Sigma^{-1} - \sum_{j=1}^{l} (X_j^T U^{-1} X_j) B \Sigma^{-1} = 0 \tag{3.2.5}$$

$$\frac{\partial L}{\partial (I-A)} = nl((I-A)^{-1})^T - \sum_{j=1}^{l} (Y_j^T U^{-1} Y_j (I-A)\Sigma^{-1} - Y_j^T U^{-1} X_j B \Sigma^{-1}) = 0 \tag{3.2.6}$$

Let $\hat{U}$, $\hat{\Sigma}$, $\hat{A}$ and $\hat{B}$ denote the maximum likelihood estimators of U, Σ, A, B, respectively. If we isolate some of the parameters above, it results from (3.2.5) and (3.2.6) that

$$\hat{B} = \left[\sum_{j=1}^{l} X_j^T \hat{U}^{-1} X_j \right]^{-1} \left[\sum_{j=1}^{l} X_j^T \hat{U}^{-1} Y_j \right] (I - \hat{A}) \tag{3.2.7}$$

$$\hat{\Sigma} = \frac{1}{nl} (I-\hat{A})^T \left\{ -\sum_{j=1}^{l} Y_j^T \hat{U}^{-1} X_j \hat{B} (I-\hat{A})^{-1} + \sum_{j=1}^{l} Y_j^T \hat{U}^{-1} Y_j \right\} (I - \hat{A}) \tag{3.2.8}$$

Replacing (3.2.7) and (3.2.8) in (3.2.3):

$$
\begin{aligned}
&-ml\hat{U}^{-1}+\frac{ml}{2}diag(\hat{U}^{-1})\\
&+\sum_{j=1}^{l}\left(\hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)\right.\\
&\left.\times\ V^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\hat{U}^{-1}\right)\\
&-\frac{1}{2}\sum_{j=1}^{l}diag\left(\hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)\right.\\
&\left.\times\ V^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\hat{U}^{-1}\right)=0
\end{aligned}
\tag{3.2.9}
$$

where

$$
V=\frac{1}{nl}\left\{-\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]+\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}Y_j\right\}
$$

Replacing (3.2.7) and (3.2.8) in (3.2.4):

$$
\begin{aligned}
&-nl\hat{\Sigma}^{-1}+\frac{nl}{2}diag(\hat{\Sigma}^{-1})\\
&+\sum_{j=1}^{l}\left(\hat{\Sigma}^{-1}(I-\hat{A})^T\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\right.\\
&\left.\times\ \hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)(I-\hat{A})\hat{\Sigma}^{-1}\right)\\
&-\frac{1}{2}\sum_{j=1}^{l}diag\left(\hat{\Sigma}^{-1}(I-\hat{A})^T\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\right.\\
&\left.\times\ \hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)(I-\hat{A})\hat{\Sigma}^{-1}\right)=0
\end{aligned}
\tag{3.2.10}
$$

where

$$
\hat{\Sigma}=\frac{1}{nl}(I-\hat{A})^T\left\{-\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]+\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}Y_j\right\}(I-\hat{A})
$$

By replacing (3.2.7) and (3.2.8) in (3.2.3) and also in (3.2.4) the four equation system is reduced to a two equation system that depends on U and $(I-A)$. System (3.2.9)–(3.2.10) has not a closed analytic solution and the estimation of the two matrices of parameters U and $(I-A)$ needs to be solved iteratively by designing a two-stage algorithm. Once these two matrices have been estimated, the pair $\hat{B}$ and $\hat{\Sigma}$ can be obtained by substitution in (3.2.7) and (3.2.8).

4. Numerical results

By definition, the MLE is the global maximum of the (log)-likelihood function. The standard way to proceed to obtain this estimator implies solving the system of likelihood equations described in Section 3 by setting each derivative equal

Table 1
Mean Euclidean distances $||\hat{A}-A||_{2,s}$ and $||\hat{B}-B||_{2,s}$ between estimate $\hat{A}$ and parameter A and between estimate $\hat{B}$ and parameter B, over $s = 10$ simulation runs. Mean fitness value and percentage of runs MLE improves 2SLS fitness score. $U = (u_{ij}) \in [-5, 5]$.

Size			2SLS		MLE_{nlm}		Fitness		%
m	k	l	$\|\|\hat{A}-A\|\|$	$\|\|\hat{B}-B\|\|$	$\|\|\hat{A}-A\|\|$	$\|\|\hat{B}-B\|\|$	2SLS	MLE_{nlm}	Improvement
2	3	5	$1.55_{1.73}$	$1.80_{1.92}$	$1.50_{1.62}$	$1.67_{1.74}$	−737.53	−267.43	100%
2	3	10	$2.22_{2.49}$	$3.50_{4.75}$	$1.42_{0.94}$	$1.91_{2.06}$	−5019.17	−760.49	80%
2	3	25	$0.98_{1.46}$	$1.82_{2.84}$	$0.94_{1.39}$	$1.67_{2.37}$	−1207.75	−296.99	100%
2	3	50	$1.02_{1.42}$	$1.29_{2.09}$	$1.04_{1.42}$	$1.39_{2.09}$	−50 338.95	−672.09	90%
8	12	5	$6.41_{1.50}$	$10.89_{3.19}$	$6.40_{1.46}$	$10.82_{3.15}$	−474 794	−31 080.5	90%
8	12	10	$8.62_{8.29}$	$13.03_{12.04}$	$8.63_{8.03}$	$12.99_{12.02}$	−932 483	−435 080	100%
8	12	25	$7.68_{3.90}$	$11.07_{4.67}$	$7.56_{3.79}$	$11.08_{4.69}$	−4 814 615.7	−603 524.25	100%
8	12	50	$4.31_{3.24}$	$5.70_{3.00}$	$4.32_{3.23}$	$5.69_{2.99}$	−1 182 940	−1 028 962.6	100%
10	15	5	$9.09_{3.95}$	$16.60_{8.00}$	$9.12_{4.03}$	$16.54_{7.96}$	−950 988.07	−80 876.12	100%
10	15	10	$6.71_{2.36}$	$9.20_{2.75}$	$6.68_{2.39}$	$9.18_{2.75}$	−3 281 584.5	−755 413.76	100%
10	15	25	$5.16_{1.96}$	$7.21_{2.73}$	$5.17_{1.95}$	$7.19_{2.73}$	−684 998 313	−424 518 142	100%
10	15	50	$5.03_{2.68}$	$6.32_{2.56}$	$5.03_{2.69}$	$6.31_{2.57}$	−55 048 498	−19 943 667	100%
15	20	5	$15.21_{2.35}$	$26.65_{7.54}$	$15.20_{2.35}$	$26.64_{7.54}$	−32 944 034	−13 236 945	100%
15	20	10	$13.13_{3.23}$	$20.00_{7.75}$	$13.13_{3.23}$	$20.00_{7.75}$	−3 953 024	−1 599 824.4	100%
15	20	25	$11.60_{3.17}$	$15.45_{4.77}$	$11.60_{3.17}$	$15.43_{4.77}$	−12 677 072	−1 599 824.4	100%
15	20	50	$9.91_{1.98}$	$12.27_{3.44}$	$9.91_{1.98}$	$12.27_{3.43}$	−1 394 508.6	−707 533.98	100%

to zero. Instead, the scheme here suggested on finding the MLE is to use a generic optimization solver based on numerical methods. The idea, in this paper, is simply to obtain a first approach to the MLE by setting up starting parameter values for the log-likelihood function and computing the *nlm* optimization solver included in the statistical software R.

Since the maximization of the log-likelihood function is a nonlinear problem, calculations for obtaining the MLE of the model proposed are cumbersome and numerical procedures are often sensitive to initial values. At this point, two situations will be distinguished: (1) estimation of coefficient matrices A and B for known covariance matrices U and Σ and (2) estimation of A and B with an unknown covariance structure.

In this paper, we focus on the estimation in MSEMs with known covariance matrices U and Σ. In the absence of *a priori* information, the choice of $\hat{A}_0 = A_{2SLS}$ and $\hat{B}_0 = B_{2SLS}$ will generally constitute a suitable initial solution for $\hat{A}$ and $\hat{B}$ although it postulates intertemporally uncorrelated observations.

The experiment aims to compare 2SLS algorithm and the optimization function *nlm* for different sizes of an MSEM in order to determine the most efficient method in each case. These two techniques differ in nature, 2SLS is based on least squares and thus minimizes the sum of squared residuals while the *nlm* function is applied to obtain the maximum of the likelihood function. In a SEM, 2SLS and limited information maximum likelihood estimators are asymptotically equivalent [19]. We seek to analyse whether the MLE for the new model proposed obtains better estimates than the 2SLS estimator in presence of serial dependence.

Experiments have been executed in a parallel NUMA node with 4 Intel hexa-core Nehalem-EX EC E7530, with 24 cores, at 1.87 GHz and 32 GB of RAM. All tests were carried out with a C code, including the call of the optimization function nlm of R. Namely, the R statistical package used is GNU R version 3.5.2.

Four different values for endogenous and exogenous variables were considered and $l = 5$, 10, 25, 50. Whatever the problem size, the number of observations in each group is $n = 5$. Tables 1 and 2 show the experiment outcomes for the same among-column covariance matrix Σ, but two different among-row covariance matrices U of the error terms distribution. In both cases, error disturbances E_j in Eq. (3.1.5) were generated by using the property stated in [29]: If $Z_j = (z_{st})$ denotes an $n \times m$ random matrix with z_{st} $(s = 1, \ldots, n; t = 1, \ldots, m)$ i.i.d. $N(0, 1)$, then

$$E_j = U^{1/2} Z_j \Sigma^{1/2} \sim N_{n,m}(0, U, \Sigma) \qquad j = 1, \ldots, l$$

On the basis of the simulation results, the fitness value of the likelihood function provided by the optimization solver is the same as the likelihood value given by the 2SLS estimator or enhances it in all cases. Tables 1 and 2 show the percentage of simulation runs in which the fitness value of likelihood function calculated by the optimization solver purely outperforms the 2SLS fitness likelihood value and also the mean fitness value in each case. As a measure of the dispersion of the estimator around the parameter, the mean Euclidean distance between estimate and parameter is evaluated.

For example in Table 1, for a problem size $m = 8$, $k = 12$, $l = 5$ and $n = 5$ the optimization solver score outperforms fitness 2SLS value 90% of the simulation runs and 10% of the times both techniques obtain the same likelihood value. The mean fitness value illustrates this improvement being −474 794 for 2SLS and −31 080.5 for the maximum likelihood method. Moreover, the mean Euclidean norm of the coefficient matrices is closer to the parameters using the maximum likelihood estimator. For the endogenous variables, the distance between estimate and parameter over $s = 10$ simulation runs is 6.41 with a standard deviation of 1.50 with 2SLS and 6.40 with a standard deviation of 1.46 with the MLE. The same situation is repeated for the exogenous variables.

From the results, one can gather that the MLE tends to outperform the 2SLS fitness score for small values of U, that is when the serial dependence is not very strong, as shown in Table 1. Nevertheless, the greater the U values are, the more

Table 2
Mean Euclidean distances $||\hat{A} - A||_{2,s}$ and $||\hat{B} - B||_{2,s}$ between estimate $\hat{A}$ and parameter A and between estimate $\hat{B}$ and parameter B, over $s = 10$ simulation runs. Mean fitness value and percentage of runs MLE improves 2SLS fitness score. $U = (u_{ij}) \in [-500, 500]$.

Size			2SLS		MLE_{nlm}		Fitness		%
m	k	l	$\|\|\hat{A} - A_0\|\|$	$\|\|\hat{B} - B_0\|\|$	$\|\|\hat{A} - A_0\|\|$	$\|\|\hat{B} - B_0\|\|$	2SLS	MLE_{nlm}	Improvement
2	3	5	$1.60_{1.46}$	$4.10_{4.72}$	$1.63_{1.48}$	$4.20_{4.64}$	−4355.14	−446.46	60%
2	3	10	$1.40_{1.61}$	$2.06_{1.85}$	$1.31_{1.51}$	$1.90_{1.79}$	−19 732.70	−786.22	70%
2	3	25	$0.70_{0.94}$	$1.69_{2.66}$	$0.69_{0.94}$	$1.67_{2.68}$	−2659.69	−1661.99	40%
2	3	50	$0.90_{0.95}$	$2.56_{5.38}$	$0.98_{1.01}$	$2.63_{5.37}$	−6870.14	−5600.67	20%
8	12	5	$8.27_{3.09}$	$16.86_{8.73}$	$8.16_{3.07}$	$16.77_{8.60}$	−1.18E+10	−90 849 708	100%
8	12	10	$7.02_{4.05}$	$10.78_{5.81}$	$6.95_{4.03}$	$10.75_{5.75}$	−2 409 591 512	−12 709 197	80%
8	12	25	$4.89_{2.01}$	$6.70_{2.42}$	$4.96_{2.08}$	$6.72_{2.45}$	−4 461 250.85	−761 458.57	100%
8	12	50	$3.17_{1.25}$	$4.27_{1.75}$	$3.20_{1.25}$	$4.28_{1.75}$	−6 406 963.37	−501 370.69	80%
10	15	5	$9.32_{1.90}$	$16.05_{3.04}$	$9.31_{1.90}$	$16.03_{3.05}$	−12 121 082	−146 919.62	90%
10	15	10	$11.13_{4.35}$	$16.89_{8.39}$	$11.09_{4.31}$	$16.89_{8.38}$	−180 931 364	−18 311 430	90%
10	15	25	$8.78_{4.85}$	$13.80_{7.67}$	$8.77_{4.85}$	$13.77_{7.69}$	−6 033 087.5	−10 361 773.7	100%
10	15	50	$5.46_{2.22}$	$7.42_{1.89}$	$5.45_{2.22}$	$7.40_{1.88}$	−20 594 215	−10 283 613	100%
15	20	5	$15.26_{2.57}$	$33.78_{17.26}$	$15.24_{2.60}$	$33.75_{17.23}$	−825 513.65	−236 981.94	100%
15	20	10	$14.65_{3.27}$	$24.08_{8.00}$	$14.65_{3.28}$	$24.07_{7.99}$	−793 610.27	−455 702.25	100%
15	20	25	$11.38_{3.01}$	$16.63_{5.64}$	$11.37_{3.02}$	$16.60_{5.63}$	$-1.17E+11$	−994 995 734	80%
15	20	50	$9.33_{2.18}$	$11.57_{2.42}$	$9.33_{2.19}$	$11.56_{2.42}$	−3 518 530.8	−2 077 883.1	60%

cases the MLE obtains the same 2SLS fitness score, as shown in Table 2. One of the reasons is attributed to increasing difficulties in the calculations needed for the implementation of the optimization solver.

Expectedly, according to the tables above, dispersion decreases when the sample size l increases, so $\hat{A}$ and $\hat{B}$ are consistent estimators of A and B, respectively. In each problem, there is little difference in the mean Euclidean distance between estimate and parameter for the 2SLS algorithm and the ML method. However, in general, the MLE shows lower values of dispersion. In both tables, the exogenous coefficient matrices show the largest values in the mean Euclidean norm.

5. Conclusions

The introduction of a double variance structure in a SEM in which the assumption of intertemporally uncorrelated error terms is violated lays the basis for the development of a modified model that we referred to as MSEM. The maximum likelihood (ML) estimation of an MSEM has been set out. In the absence of an analytical solution of the system of likelihood equations, the estimation of an MSEM has been carried out using a general-purpose optimization solver with simulated data under the assumption of known variance–covariance matrices.

In a first approach, selecting the 2SLS estimates of the coefficient matrices as starting values for the optimization solver has empirically proved that the obtained estimates are closer to the parameter than those calculated when the serial dependence of the errors is ignored. However, limitations in the optimization method integrated in the solver currently used to find the maximum of the likelihood function might underperform MLE. For this reason, other alternatives need to be explored and other general-purpose optimization solvers not based on gradient methods should be examined.

The estimation of the variance matrices is not straightforward and in our humble opinion, we consider that it requires a deep and separate study. Thus, the development of a complete methodology for estimating the parameters of a MSEM, including the variance components of the model in the case in which these parameters are unknown must be incorporated as future work. Moreover, it is interesting to include the development of restricted maximum likelihood method (REML) for MSEM and to compare estimates of variance components with maximum likelihood results. Finally, extensions of MSEM to a matrix non-Gaussian distribution of errors must be considered as further work.

Acknowledgements

This research was partially supported by a grant from the Ministerio de Economía y Competitividad of Spain (TIN2016-8056-R) and a predoctoral contract from the Generalitat Valenciana and the European Social Fund (ACIF/2018/219) to R.H. The authors gratefully acknowledge the computer resources and assistance provided by the Scientific Computing and Parallel Programming Group of the University of Murcia for the simulation study.

References

[1] J.E. Contreras-Reyes, F.O.L. Quintero, R. Wiff, Bayesian modeling of individual growth variability using back-calculation: Application to pink cusk-eel (genypterus blacodes) off Chile, Ecol. Model. 385 (2018) 145–153.
[2] P. Dutilleul, The mle algorithm for the matrix normal distribution, J. Stat. Comput. Simul. 64 (2) (1999) 105–123.

[3] M. Byakagaba, Apport de la matrice normale aux modèles d'analyse de la variance et des mesures rèpètèes (Unpublished Doctoral Dissertation), Université catholique de Louvain, Louvain-la-Neuve, Belgium, 1987, Faculty of Science.
[4] P.C. Phillips, Exact small sample theory in the simultaneous equations model, Handb. Econom. 1 (1983) 449–516.
[5] H. Goldstein, Multilevel Statistical Models, Vol. 922, John Wiley & Sons, 2011.
[6] L.R. Klein, Economic Fluctuations in the United States, Wiley, 1950, pp. 1921–1941.
[7] R. Dornbusch, S. Fischer, Macroeconomics, third ed., McGraw-Hill, New York, 1984.
[8] T.M. King, Using simultaneous equation modeling for defining complex phenotypes, in: BMC Genetics, Vol. 4, BioMed Central, 2003, p. S10.
[9] I. Lu, J. Peixoto, W. Taam, A simultaneous equation model for air traffic in the new york area, in: Air Transport Research Society World ConferenceAir Transportation Research SocietyAir Transport Research Society, 2003.
[10] S.A. Graham-Bermann, L. Miller-Graff, Community-based intervention for women exposed to intimate partner violence: A randomized control trial., J. Family Psychol. 29 (4) (2015) 537.
[11] J.H. Shin, Application of repeated-measures analysis of variance and hierarchical linear model in nursing research, Nurs. Res. 58 (3) (2009) 211–217.
[12] I.L. Simone, A. Ceccarelli, C. Tortorella, A. Bellacosa, F. Pellegrini, I. Plasmati, M.F. De Caro, M. Lopez, F. Girolamo, P. Livrea, Influence of interferon beta treatment on quality of life in multiple sclerosis patients, Health Qual. Life Outcomes 4 (1) (2006) 96.
[13] F. Steele, A. Vignoles, A. Jenkins, The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach, J. R. Stat. Soc. A 170 (3) (2007) 801–824.
[14] R. Blundell, F. Windmeijer, Cluster effects and simultaneity in multilevel models, Health Econ. 6 (4) (1997) 439–443.
[15] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2012. URL: http://www.R-project.org/.
[16] J.A. Hausman, Specification and estimation of simultaneous equation models, Handb. Econom. 1 (1983) 391–448.
[17] P. Phillips, Finite sample theory and the distributions of alternative estimators of the marginal propensity to consume, Rev. Econom. Stud. 47 (1) (1980) 183–224.
[18] J. Knight, Non-Normality of disturbances and the k-class structural estimator (Unpublished manuscript), Univ. of New South Wales, 1981.
[19] P.J. Dhrymes, Econometrics: Statistical Foundations and Applications, Springer Science & Business Media, 2012.
[20] M. Aitkin, D. Anderson, J. Hinde, Statistical modelling of data on teaching styles (with discussion), J. R. Stat. Soc. A (1981) 419–461.
[21] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics (1982) 963–974.
[22] S.W. Raudenbush, A.S. Bryk, Hierarchical Linear Models: Applications and Data Analysis Methods, Vol. 1, Sage, 2002.
[23] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, et al., Linear and nonlinear mixed effects models, R package version 3 (2014).
[24] J.L. Bernal-Rusiel, D.N. Greve, M. Reuter, B. Fischl, M.R. Sabuncu, A.D.N. Initiative, et al., Statistical analysis of longitudinal neuroimage data with linear mixed effects models, Neuroimage 66 (2013) 249–260.
[25] X. Zhang, A Tutorial on Restricted Maximum Likelihood Estimation in Linear Regression and Linear Mixed-Effects Model, 2015.
[26] H.O. Hartley, J.N. Rao, Maximum-likelihood estimation for the mixed analysis of variance model, Biometrika 54 (1–2) (1967) 93–108.
[27] D.A. Harville, Maximum likelihood approaches to variance component estimation and to related problems, J. Amer. Statist. Assoc. 72 (358) (1977) 320–338.
[28] J.C. Pinheiro, D.M. Bates, Mixed-effects models in s and s-plus, 2011.
[29] S.F. Arnold, The Theory of Linear Models and Multivariate Analysis, Wiley, New York, 1981.
[30] D.A. Harville, Matrix Algebra from a Statistician's Perspective, Vol. 1, Springer, 1997.
[31] P.S. Dwyer, Some applications of matrix derivatives in multivariate analysis, J. Amer. Statist. Assoc. 62 (318) (1967) 607–625.
[32] K.B. Petersen, M.S. Pedersen, The matrix cookbook, Tech. Univ. Denmark 7 (15) (2008) 510.