GME ESTIMATION OF SPATIAL STRUCTURAL EQUATIONS MODELS

Rosa Bernardini Papalia,
Università di Bologna, Dipartimento di Scienze Statistiche 'Paolo Fortunati'
rossella.bernardini@unibo.it

Enrico Ciavolino,
Università del Salento, Dipartimento di Filosofia e Scienze Sociali
enrico.ciavolino@ateneo.unile.it

Abstract: *The objective of this paper is to develop a GME formulation for the class of spatial structural equations models (S-SEM) into a panel data framework. In this respect, two innovatory aspects are introduced: (i) the formalization of the GME estimation approach of SEM to allow for spatial heterogeneity and spatial dependence (spatially sampled data); (ii) the extension of the methodology panel data.*

1. Introduction

The focus of this paper is on Structural Equation Models (SEM) approach to spatial models. Panel data are considered by assuming that the spatial interaction across spatial units comes from the effect of an endogenous lag process.

A Generalized Maximum Entropy (GME) estimation approach for the Spatial SEM here introduced has been suggested as a suitable solution to endogeneity and collinearity problems implied by the presence of spatial dependence. This model offers a very flexible tool for modeling multivariate spatial data and answering research questions about latent factors underlying spatial samples data.

The paper is organized as follows. In section 2, we present the basic notation and the S-SEM model specification for panel data. Section 3 illustrates the proposed GME formulation. An application focused on the analysis of regional unemployment rates in Europe is presented in section 4.

2. Spatial Structural Equations Model (S-SEM) framework

2.1 Basic S-SEM specification for panel data

The Structural Equation Model (Bollen, 1989, Duncan, 1975, Jöreskog, 1973) is defined by two main parts: the first part - the *Structural Model* -refers to the relationships among the latent variables [1], while the second part - the *Measurement Model* - represents the relationships between manifest and latent, endogenous [2] and exogenous [3] variables, respectively. We introduce a general model specification for the SEM, where *m* endogenous and *l* exogenous latent variables are specified by the vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, while **x** and **y**, denote the vectors of the *r* and *q* manifest endogenous and exogenous variables, respectively.

The **B** matrix specifies the structural relationships among the endogenous latent variables while $\boldsymbol{\Gamma}$ is a matrix of coefficients of the exogenous latent variables on the endogenous ones.

The coefficient matrices $\boldsymbol{\Lambda}^{y}$ and $\boldsymbol{\Lambda}^{x}$ measure the relationships between the manifest and latent endogenous and exogenous variables, respectively.

The vectors of errors $\boldsymbol{\tau}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$, are the structural and the measurement errors vectors, respectively. The matrix $\boldsymbol{\Phi}$ represents the co-variance between the latent variables ξ, while $\boldsymbol{\Psi}$ is that one between the error term $\boldsymbol{\tau}$. Finally, $\boldsymbol{\Theta}^{\varepsilon}$ and $\boldsymbol{\Theta}^{\delta}$, are the measurement error covariance matrices between the error terms $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$**,** respectively.

$$\boldsymbol{\eta}_{m,1(t)} = \boldsymbol{\beta}_{m,m(t)} \cdot \boldsymbol{\eta}_{m,1(t)} + \boldsymbol{\Gamma}_{m,l(t)} \cdot \boldsymbol{\xi}_{l,1(t)} + \boldsymbol{\zeta}_{m,1(t)} \qquad [1]$$

$$\mathbf{y}_{r,1(t)} = \boldsymbol{\Lambda}^{y}_{r,m(t)} \cdot \boldsymbol{\eta}_{m,1(t)} + \boldsymbol{\varepsilon}_{r,1(t)} \qquad [2]$$

$$\mathbf{x}_{q,1(t)} = \mathbf{\Lambda}^{x}_{q,l(t)} \cdot \mathbf{\xi}_{l,1(t)} + \mathbf{\delta}_{q,1(t)} \qquad [3]$$

Moreover the basic model specification has some assumptions: The variables are centered, E(**η**)= E(ξ)= E(**τ**)= 0; E(**y**)= E(**ε**)= 0; E(**x**)= E(**δ**)= 0; The independent variables and the error terms are uncorrelated, E(ξ **τ**')= 0; E(**η** **ε**')=0; E(ξ **δ**')= 0; E(**η** **δ**')= 0; E(ξ **ε**')= 0; The error terms are uncorrelated, E(**τ** **ε**')= 0; E(**τ** **δ**')= 0; E(**ε** **δ**')= 0; The matrix **B** is not singular

Since we are considering a panel of *N* units within *T* periods, the model we introduce refers to (*N*·*T*) independent replications (observations) typically originated from a sample of randomly drawn subjects. The model reports the classic SEM notation (Jöreskog, 1973) with the addition of the index *t* relative to the *T* time periods and indicating that the manifest variables are referred to a *N*·*T* by 1 vector of observations, N observations for the Countries and T for time periods.

2.2 Introduction of spatial unobserved heterogeneity and spatial dependence among spatial units:

Panel data from individual observations give us the advantage of introducing fixed effects in the measurement model to allow of *spatial unobserved heterogeneity among* spatial units (N locations) (Bernardini Papalia, 2006). In this respect we proceed by including an individual specific "dummy variable" to capture unobserved heterogeneity for each unit *i* (i=1….N).

For the *spatial dependence*, we focus on one of the widely used approach (a so-called spatial LAG model) where the spatial correlation pertains to the dependent variable.

In this context, it is assumed interdependence of latent exogenous variables across areas. This assumption may be formalized by including a set of exogenous spatial lag variables into the measurement model which represent the relationship between the manifest and latent exogenous variables. In doing this, a spatial weights matrix W of non-stochastic time constant weights has to be specified. This is a (N×N) matrix in which the rows and columns correspond to the cross-sectional observations.

An element w_{ij} of the matrix expresses the prior strength of the interaction between location *i* (in the row of the matrix) and location *j* (column). This can be interpreted as the presence and strength of a link between nodes (observations) in a network representation that matches the spatial weights structure. In most application, the choice is driven by geographic criteria, such as contiguity (sharing a common border) or distance, including nearest neighbor distance (Anselin 1988; Lesage and Pace 2004).

More specifically, using the stacked equation [3], the set of latent exogenous variables ξ is enlarged to include: (i) *Spatial Lag variables* [4], that is the first-order contiguity spatially lagged dependent variable, here considered as exogenous and so defined with *X* instead of *Y*; (ii) The *fixed effects* that are country and year dummies as reported in equations [5] and [6], respectively; (iii) and the set of *q* exogenous variables $X_{NT,1}$.

$$Spatial - Lag = (\mathbf{I}_{T,T} \otimes \mathbf{W}_{N,N}) \cdot \mathbf{x}_{NT,1} \qquad [4]$$

$$Dummy - Times = (\mathbf{I}_{T,T} \otimes \mathbf{1}_{N,1}) \qquad [5]$$

$$Dummy - Space = (\mathbf{I}_{N,N} \otimes \mathbf{1}_{T,1}) \qquad [6]$$

The equation [3] can be reformulated considering also the spatial lag variable and both the fixed effects, obtaining:

$$\mathbf{X}^{*}_{NT,q+1+N+T} = \left[\mathbf{X}_{NT,q} \,|\, (\mathbf{I}_{T,T} \otimes \mathbf{W}_{N,N}) \cdot \mathbf{x}_{NT,1} \,|\, (\mathbf{I}_{N,N} \otimes \mathbf{1}_{T,1}) \,|\, (\mathbf{I}_{T,T} \otimes \mathbf{1}_{N,1}) \right] \qquad [7]$$

Then, the associated $\mathbf{\Lambda}^{x}$ matrix which specifies the regression coefficients of the observed variables on the latent exogenous variables, is defined as $\mathbf{\Lambda}^{x}$= [ρ | **α** | **τ**], including: the *spatial autoregressive* *parameter* (ρ), the vector of *fixed effects*, relative to time and spatial effects defined as $\boldsymbol{\alpha}=[\alpha_t \mid \alpha_s]$, and the set of coefficients relative to the exogenous variables ($\boldsymbol{\tau}$).

In the estimation of a S-SEM, it is then essential to deal with the problem of endogeneity of the spatial lag term given by the correlation between latent endogenous and exogenous variables and as a consequence the correlation between exogenous observed variables and errors. Our proposal here is to introduce a GME formulation which produces consistent parameter estimates in presence of collinearity and endogeneity of some explanatory variables.

3. The GME formulation

The objective here is to recover the unknown parameters of the fixed effects spatial panel SEM as specified in the previous section, with minimal distributional assumptions. In this respect and also with the aim of deal with the problems due to the endogeneity of the spatial lag and fixed effects components, it is suggest a Generalized Maximum Entropy (GME) estimation approach (Golan, 1996). Following the Information-theoretic GME idea based on the (Shannon, 1948) entropy principle, since the entropy function is defined on proper probability distributions, we first proceed by reformulating all parameters and noise components of the model [1-3] as set of proper probabilities (*P* indicates a probability distribution for parameters and *G* a probability distribution for errors) defined on some support spaces (*Z* support space for parameters and *V* support space for random errors).

3.1 Basic GME: Specification of SEM

The GME approach for the SEM considers the *Re- Parameterization* of the unknown parameters and the disturbance terms, as a convex combination *of expected value of a discrete random variable*. The coefficient matrices, $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}^{y}$, $\boldsymbol{\Lambda}^{x}$, and the co-variance matrices, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}^{\varepsilon}$, $\boldsymbol{\Theta}^{\delta}$, are all Re-Parameterized as expected values of discrete random variable with *M* fixed points for the coefficients and *J* for the errors. The Re-Parameterized coefficients, reported without the index *t* relative to the *T* time periods, are so defined: $\mathbf{B}_{(m,m)}=\mathbf{Z}_{(m,m\cdot M)}\cdot\mathbf{P}^{\mathrm{B}}_{(m\cdot M,m)}$; $\boldsymbol{\Gamma}_{(m,l)}=\mathbf{Z}_{(m,l\cdot M)}\cdot\mathbf{P}^{\Gamma}_{(l\cdot M,l)}$; $\boldsymbol{\Lambda}^{y}_{(r,m)}=\mathbf{Z}_{(r,m\cdot M)}\cdot\mathbf{P}^{\Lambda^{y}}_{(m\cdot M,m)}$; $\boldsymbol{\Lambda}^{x}_{(q,l)}=\mathbf{Z}_{(q,l\cdot M)}\cdot\mathbf{P}^{\Lambda^{x}}_{(l\cdot M,l)}$; $\boldsymbol{\zeta}_{(m,1)}=\mathbf{V}_{(m,j\cdot J)}\cdot\mathbf{G}^{\zeta}_{(j\cdot J,1)}$; $\boldsymbol{\varepsilon}_{(r,1)}=\mathbf{V}_{(r,j\cdot J)}\cdot\mathbf{G}^{\varepsilon}_{(j\cdot J,1)}$; $\boldsymbol{\delta}_{(q,1)}=\mathbf{V}_{(q,j\cdot J)}\cdot\mathbf{G}^{\delta}_{(j\cdot J,1)}$.

The S-SEM model [1-7] can be re-formulated in a unique formulae in function of the re-parameterized coefficients:

$$\mathbf{Y}=\psi\left(\mathbf{P}^{\mathrm{B}},\mathbf{P}^{\Gamma},\mathbf{P}^{\Lambda^{y}},\mathbf{P}^{\Lambda^{x}},\mathbf{G}^{\zeta},\mathbf{G}^{\varepsilon},\mathbf{G}^{\delta}\right)=\left(\mathbf{Z}\cdot\mathbf{P}^{\Lambda^{y}}\right)\cdot\left[\mathbf{I}-\left(\mathbf{Z}\cdot\mathbf{P}^{\mathrm{B}}\right)\right]^{-1}\cdot$$
$$\left\{\left(\mathbf{Z}\cdot\mathbf{P}^{\Gamma}\right)\left(\mathbf{Z}\cdot\mathbf{P}^{\Lambda^{x}}\right)^{-1}\left[\mathbf{X}-\left(\mathbf{V}\cdot\mathbf{G}^{\delta}\right)\right]+\left(\mathbf{V}\cdot\mathbf{G}^{\zeta}\right)\right\}+\left(\mathbf{V}\cdot\mathbf{G}^{\varepsilon}\right) \qquad [8]$$

Given the re-parameterization and the re-formulation, the GME system can be expressed as a constrained non-linear programming problem. The coefficients and the error terms are estimated by recovering the probability distribution of the discrete random variables set. The vectors $\mathbf{p^{B}}$=*vec*($\mathbf{P^{B}}$), $\mathbf{p}^{\boldsymbol{\Gamma}}$=*vec*($\mathbf{P}^{\boldsymbol{\Gamma}}$), $\mathbf{p}^{\boldsymbol{\Lambda}\mathbf{y}}$=*vec*($\mathbf{P}^{\boldsymbol{\Lambda}\mathbf{y}}$), $\mathbf{p}^{\boldsymbol{\Lambda}\mathbf{x}}$=*vec*($\mathbf{P}^{\boldsymbol{\Lambda}\mathbf{x}}$), are obtained by using the *vec* operator of the matrices $\mathbf{P^{B}}$, $\mathbf{P}^{\boldsymbol{\Gamma}}$, $\mathbf{P}^{\boldsymbol{\Lambda}\mathbf{y}}$, $\mathbf{P}^{\boldsymbol{\Lambda}\mathbf{x}}$. The vectors: $\mathbf{p^{B}}$, $\mathbf{p}^{\boldsymbol{\Gamma}}$, $\mathbf{p}^{\boldsymbol{\Lambda}\mathbf{y}}$, $\mathbf{p}^{\boldsymbol{\Lambda}\mathbf{x}}$, $\mathbf{G}^{\boldsymbol{\zeta}}$, $\mathbf{G}^{\boldsymbol{\varepsilon}}$, $\mathbf{G}^{\boldsymbol{\delta}}$,_are calculated by the maximization of the following entropy function:

$$H(\mathbf{p}^{\mathrm{B}},\mathbf{p}^{\Gamma},\mathbf{p}^{\Lambda^{y}},\mathbf{p}^{\Lambda^{x}},\mathbf{G}^{\zeta},\mathbf{G}^{\varepsilon},\mathbf{G}^{\delta})=$$
$$-\mathbf{p}^{\mathrm{B}'}\cdot\ln\mathbf{p}^{\mathrm{B}}-\mathbf{p}^{\Gamma'}\cdot\ln\mathbf{p}^{\Gamma}-\mathbf{p}^{\Lambda^{y}\,'}\cdot\ln\mathbf{p}^{\Lambda^{y}}-\mathbf{p}^{\Lambda^{x}\,'}\cdot\ln\mathbf{p}^{\Lambda^{x}}-\mathbf{G}^{\zeta'}\cdot\ln\mathbf{G}^{\zeta}-\mathbf{G}^{\varepsilon'}\cdot\ln\mathbf{G}^{\varepsilon}-\mathbf{G}^{\delta'}\cdot\ln\mathbf{G}^{\delta} \qquad [9]$$

subjected to the *consistency* and *normalization constraints* (the sum of each coefficients and the error terms probability vector have to be equal to 1).

4. The Case Study

In order to show the S-SEM application to a real data set, an analysis aiming at investigating the significance of spatial effects for regional unemployment disparities in OECD is presented. The data, relative to a panel of OECD countries over the period 1998-2006, refer to the main macro economy variables relevant in explaining unemployment differences across local labour markets such as: the Gross Domestic Product, the Inflation rate, the Wages, the Innovation, the Human Capital. The effects of the labour and market deregulation variables on the unemployment rates is also analyzed. Finally, the role of these variables is relied on static and dynamic specifications for regional unemployment rates.

Bibliography

Al-Nasser, A. D., (2003), *Customer Satisfaction Measurement Models: Generalized Maximum Entropy Approach*, Pak Journal of Statistics, 19(2): 213–226.

Anselin L. (1988). *Spatial Econometrics: Methods and models*. Dordrecht, the Netherlands, Kluwer.

Anselin L. (2001). Spatial Econometrics. In *A companion to theoretical econometrics*, ed. B.H.

Bernardini Papalia (2006) Modeling mixed spatial processes and spatio-temporal dynamics in Information theoretic frameworks, (2006), COMPSTAT 2006. In A. Rizzi, M. Vichi (Eds), Physica Verlag Heidelberg, New York, pp. 1483-1491.

Bollen, K. A. (1989). Structural Equations with Latent Variables. New York: John Wiley and Sons.

Ciavolino E., Dahlgaard J.J. (2007), ECSI - Customer Satisfaction Modelling and Analysis: a case study, in press on Total Quality Management & Business Excellence, Volume 18, Issue 5, 545 – 554.

Duncan, O.D., (1975), *Introduction to Structural Equation Models*. New York: Academic Press.

Fornell C., Bookstein F. L., (1982), Two structural equation models: LISREL and PLS Applied to Consumer Exit-Voice Theory, Journal of Marketing Research, Vol 19, n. 4, pp 440-452.

Golan A, Judge G, Miller D (1996). Maximum entropy econometrics: robust estimation with limited data. Wiley, New York.

Jöreskog, K. G., (1973), *A general Method of Estimating a Linear Structural Equation system*. In Goldberg, S. A. & Duncan, D. O (Eds). Structural Equation Models in the Social Sciences. New York: Seminar Press. 85-112.

Lesage J.P., Pace R.K. (2004) *Spatial and Spatiotemporal Econometrics* - Advances in Econometrics, Vol. 18. Elsevier Ltd.

Shannon, C. E., (1948), *A mathematical Theory of Communications*, Bell System Technical Journal, 27: 379–423.