TESTING CLUSTERS OF LOCATIONS IN SPATIAL DYNAMIC PANEL DATA MODELS

Feo G. [1], Giordano F.[1], Niglio M.[1], Milito S.[1] and Parrella M.L.[(*)1]

[1] Department of Economics and Statistics, University of Salerno, (e-mail[(*)]: mparrella@unisa.it)

ABSTRACT: The *SDPD* (Spatial Dynamic Panel Data) models have been proposed in the socio-econometric literature to analyze spatio-temporal data. In this paper we consider a particular version of such models, where the set of spatial units is assumed to be partitioned into clusters and the parameters of the model are assumed to be constant within clusters and not constant across clusters. We propose a mutiple testing procedure that helps to choose the best model for a dataset by testing a given partition of clusters assumed under the null hypothesis.

KEYWORDS: spatial dynamic panel data models, model selection, spatial clustering.

1 Introduction

Let us consider a multivariate stationary process $\{\mathbf{y}_t, t = 1, 2, \ldots\}$ of dimension p, where the vector $\mathbf{y}_t$ collects the observations at time t from p different locations (=*spatial units*). In this framework, the dependence between the p time series is usually due to spatial correlation.

The following model, in equation (1), belongs to the so called *SDPD* class of models, proposed in the socio-econometric literature (see Lee & Yu, 2010, Dou *et al.*, 2016 and references therein)

$$\begin{aligned} \mathbf{y}_t &= D(\lambda_0)\mathbf{W}\mathbf{y}_t + D(\lambda_1)\mathbf{y}_{t-1} + D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1} + D(\beta_1)\mathbf{x}_t^{(1)} + \ldots \qquad (1) \\ &\quad \ldots + D(\beta_k)\mathbf{x}_t^{(k)} + \mathbf{c} + \varepsilon_t. \end{aligned}$$

A typical feature of these models is the presence of the *spatial matrix*, denoted by $\mathbf{W}$, a known weight matrix with zero main diagonal, reflecting the physical distances between spatial units. It is used to deal with spatial correlation.

The parameters of the model are collected in the diagonal matrices $D(\lambda_j)$ and $D(\beta_l)$, with $j = 0, 1, 2$ and $l = 1, \ldots, k$, where the vectors $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jp})'$ and $\beta_l = (\beta_{l1}, \ldots, \beta_{lp})'$ assure that each location has its own parameter (*i.e.*, the model is *spatially heterogeneous*). Model (1) is characterized by the sum of several components: *a)* a *spatial component*, $D(\lambda_0)\mathbf{W}\mathbf{y}_t$, for spatial correlation; *b)* a *dynamic component*, $D(\lambda_1)\mathbf{y}_{t-1}$, for serial correlation; *c)* a *spatial–dynamic component*, $D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1}$, for the interactions between spatial and serial correlation; *d)* the component $D(\beta_l)\mathbf{x}_t^{(l)}$, for the effects of some covariates on the time series data $\mathbf{y}_t$ (the vector $\mathbf{x}_t^{(l)}$ collects the data observed at time t on the p locations and for a given covariate l, with $l = 1,\ldots,k$). Finally, $\mathbf{c}$ contains the fixed effects while $\varepsilon_t \sim i.i.d.$ with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \Sigma_\varepsilon$.

It is important to note that the number of parameters in model (1) is equal to $(4+k)p$ and may explode, since the number of locations p is allowed to increase to infinity asymptotically with the time series length. Many variants of *SDPD* models can be formulated starting from model (1) and considering some restrictions on the parameters. First of all, not all the components *a)-d)* are always active in the model. For example, in the well-known *SAR* model, only the parameters of the *spatial component a)* are active, while other parameters are zero. Moreover, sometimes the vectors λ_j and β_l may have constant parameters (*spatial homogeneity*, as Lee & Yu, 2010 and references therein), other times they are not constant (*spatial heterogeneity*, as in Dou *et al.*, 2016).

In this paper we consider a hybrid *SDPD* model, a cross between homogeneous and heterogeneous spatial models. By imagining that the spatial units can be subdivided into clusters, we assume that the model has parameters that are homogeneous within clusters and heterogeneous between clusters. This model has not yet been considered in the spatial econometric literature, as far as we know, and will be referred to as the *clusterized SDPD* model. It can be estimated by adapting the estimation procedure proposed in Dou *et al.*, 2016. But in order to estimate this model consistently and efficiently, one has to know the clustering structure (how many clusters there are and which locations are included in each cluster). The aim here is to propose a testing procedure which allows to test if a given partition of clusters assumed under H_0 can be accepted, so that one can use this information to estimate the *clusterized SDPD* model. The proposed testing procedure is briefly described in the following section.

2 The multiple testing procedure in a nutshell

Giordano *et al.*, 2023 propose a strategy to test a specific version of *SDPD* model for a given spatio-temporal dataset. The idea underlying their method is based on comparing two setups: *A)* the general version of the spatial model, shown in equation (1) and assumed under the alternative hypothesis (unrestricted model); *B)* a nested model, assumed under the null (restricted model).

Here we extend the procedure in Giordano *et al.*, 2023 to the case of a *clusterized SDPD* model. Denote with S the number of clusters assumed under H_0 and let G_s, $s = 1, \ldots, S$, be a partition of $\{1, \ldots, p\}$ with p_s the number of units in the s-th cluster, G_s. So, it is $\sum_{s=1}^{S} p_s = p$. The testing procedure is based on the following test-statistics

$$\widehat{\delta}_{ji} = \widehat{\theta}_{ji}^{(u)} - \widehat{\theta}_{js}^{(r)} \qquad j = 1, \ldots, 3+k; i \in G_s; s = 1, \ldots, S; \tag{2}$$

where $\widehat{\theta}_{ji}^{(u)}$ is the *unrestricted estimator* of the j-th parameter in the vector $\theta_i = (\lambda_{0i}, \lambda_{1i}, \lambda_{2i}, \beta_{1i}, \ldots, \beta_{ki})'$, while $\widehat{\theta}_{js}^{(r)}$ is the restricted estimator, derived under the null hypothesis as

$$\widehat{\theta}_{js}^{(r)} = \frac{1}{p_s} \sum_{i \in G_s} \widehat{\theta}_{ji}^{(u)}, \tag{3}$$

that is the average of the unrestricted estimated values for the spatial units in the s-th cluster. These estimators are described in details in Giordano *et al.*, 2023. When the true *SDPD* model is the one assumed under H_0 (i.e., the assumed clustering partition is correct), the two estimators $\widehat{\theta}_{ji}^{(u)}$ and $\widehat{\theta}_{js}^{(r)}$ are expected to produce similar results (in mean) and the statistics $\widehat{\delta}_{ji}$ are expected to be centered around zero. A graphical evidence is given in Figure 1, where we simulated 200 replications of a *clusterized SDPD* model with $p = 10$ locations (on the x-axis) and $S = 4$ clusters. The clusters are shown by colours, but note that we assume only 3 clusters under the null hypothesis (more specifically, H_0 is true for the first two clusters while the last two clusters are erroneously assumed to be one). The boxplots summarize the unrestricted $\widehat{\theta}_{ji}^{(u)}$ (on the left) and restricted $\widehat{\theta}_{js}^{(r)}$ (in the center) estimations of the parameters. On the right, the values of the test-statistics $\widehat{\delta}_{ji}$, for each location. As evident from the figure, the test-statistics correctly deviate from the null hypothesis for the last two clusters. Note that the procedure is organized as a mutiple test (one test for each location), where a Bonferroni-type correction is used to calibrate the global size (details are reported in Giordano *et al.*, 2023).

In the simulation study we have further considered different values of dimension $p = (10, 50, 100)$ and sample size $T = (100, 500, 1000)$. Other settings are fixed as in Giordano *et al.*, 2023. The results are consistent in terms of False Positive Rate (*i.e.*, the average proportion of locations for which we wrongly reject H_0; note that it is not equivalent to the global size) and False Negative Rate (the average proportion of locations for which we wrongly accept H_0), as reported in the following table for the parameter λ_{i1}.

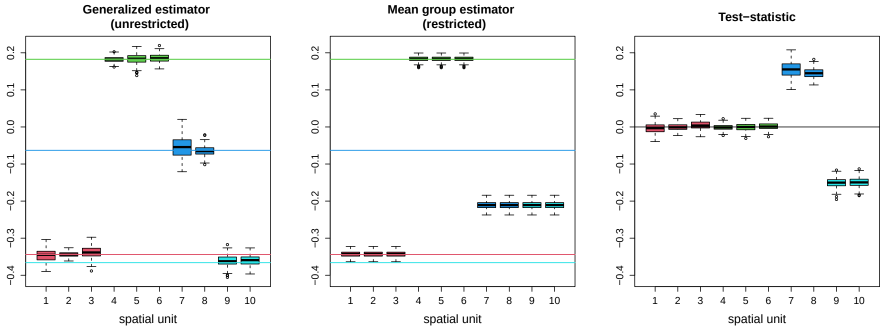

Figure 1. *For a* clusterized SDPD *model with 10 locations (x-axis), the boxplots summarize the unrestricted (left) and restricted (center) estimations of the parameters. On the right, the values of the test-statistics for each location. There are 4 clusters (=colours) in the true model, but we assume only 3 clusters under the null hypothesis (so, H_0 is true for the first two clusters while it is false for the last two).*

	False Positive Rate			False Negative Rate		
$T =$	100	500	1000	100	500	1000
$p = 10$	0	0	0	0.53	0.27	0.15
50	0	0	0	0.09	0.06	0.06
100	0	0	0	0.12	0.06	0.04

There are many real cases where one can apply our testing procedure. For example, one may consider spatial data observed in a country and may want to test if the *SDPD* model is homogeneous within counties and heterogeneous between counties. In such a case, the clusters are the counties and the units in each cluster are perfectly identified under H_0. Our procedure allows to test if the assumed *clusterized SDPD* model is a good model for the dataset at hand.

References

DOU, B., PARRELLA, M.L., & YAO, Q. 2016. Generalized Yule-Walker Estimation for Spatio-Temporal Models with Unknown Diagonal Coefficients. *Journal of Econometrics*, **194**, 369–382.

GIORDANO, F., NIGLIO, M., & PARRELLA, M.L. 2023. Model structure identification in spatial dynamic panel data models. *Submitted*.

LEE, L.F., & YU, J. 2010. Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, **154**, 165–185.