Computational Statistics and Data Analysis 71 (2014) 914–933

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Maximum likelihood estimation of spatially and serially correlated panels with random effects[☆]

Giovanni Millo[*]

Generali Research and Development, Assicurazioni Generali S.p.A., Via Machiavelli 3 34131 Trieste, Italy

ARTICLE INFO

Article history:
Received 1 July 2012
Received in revised form 2 July 2013
Accepted 13 July 2013
Available online 26 July 2013

Keywords:
Spatial panel
Maximum likelihood
Serial correlation
R

ABSTRACT

An estimation framework and a user-friendly software implementation are described for maximum likelihood estimation of panel data models with random effects, a spatially lagged dependent variable and spatially and serially correlated errors. This specification extends static panel data models in the direction of serial error correlation, allowing richer modelling possibilities and more thorough diagnostic assessments. The estimation routines extend the functionalities of the **splm** package for spatial panel econometrics in the open source **R** system for statistical computing.

1. Introduction

The econometric literature has recently considered panel regression models with spatially autocorrelated outcomes or disturbances and random or fixed individual effects. After the pioneering works of Anselin (1988) and Case (1991), the more recent methodological contributions by Elhorst (2003) and Baltagi et al. (2003) and the first comprehensive treatments of the subject in Anselin et al. (2008) and Elhorst (2009) have helped the diffusion of spatial panel methods in applied practice, although hindered by the lack of a user-friendly software (see Millo and Piras, 2012). Baltagi et al. (2012) have discussed the bias from neglecting spatial structure in panels from a forecasting perspective. Meanwhile, Baltagi et al. (2007b) have extended the spatial panel framework to considering serial correlation in the remainder errors. Notable alternative approaches to space–time dependence are: the dynamic spatial panel framework of Lee and Yu (2010a); and the space–time correlation model of Elhorst (2008), where spatial and serial correlation are modelled jointly. Lee and Yu (2012), in a recent comprehensive treatment, are the first to analyse a very general specification including spatial lags, spatially and serially correlated errors together with individual effects (their paper appeared when this one was already under revision. We thank an anonymous referee for bringing it to our attention.) As they observe, "[i]n empirical applications with spatial panel data, it seems that investigators tend to limit their focus on some spatial structures and ignore others, and in addition, no serial correlation is considered" (Lee and Yu, 2012, p. 1370). They also document through Monte Carlo simulation the biases due to neglecting serial correlation or some parts of the spatial structure, and recommend a general to specific strategy.

We describe the implementation in the **R** system for statistical computing (R Development Core Team, 2012) of maximum likelihood estimation of panel models with spatial lags, spatial errors, individual random effects that are/are not spatially autocorrelated and a temporally autocorrelated remainder error term. In economic terms, this specification

[☆] **R** code for replicating all examples appears as Inline Supplementary Code in the electronic version of the paper. The **R** package **splm**, version 1-1.0 or higher, is required.

[*] Tel.: +39 040671184; fax: +39 040671160.
E-mail addresses: giovanni_millo@generali.com, the_sculler@hotmail.com.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.024

can accommodate spatial spillovers from the outcome variable at neighbouring locations, spatial diffusion of idiosyncratic shocks, individual heterogeneity and time-persistence of idiosyncratic shocks. In terms of the previous literature, while building on the general estimation theory in Anselin (1988), this framework extends Case (1991) in adding serial error correlation, and Baltagi et al. (2007b) in adding a spatially lagged dependent variable. Moreover, unlike Baltagi et al. (2007b) whose primary goal is to derive Lagrange Multiplier diagnostic tests under a restricted specification, we take a general to specific approach aiming primarily at estimating the full model (for a review and comparison of the general to specific and specific to general strategies in applied spatial econometrics, see Florax et al., 2003; Mur and Angulo, 2009).

The common specifications (spatial lag and/or spatial error with or without random effects) have already been available for some time in the **R** package **splm** described in Millo and Piras (2012), although implementation details were still unpublished. Here we describe an extended framework allowing for serial correlation of the autoregressive type, along with all the other features. The corresponding software procedures extend the capabilities of the **splm** package in this direction.

One of the main goals has been to produce a complete set of estimators able to cope with all combinations of the dependence structures considered, while keeping software as easy to read and maintain as possible. To this end, the theoretical estimation framework has been geared towards modularity, so as to have software counterparts to theoretical objects such as spatial filters or covariances, combining them and "plugging them" in the likelihood as prescribed by the reference theory.

In the spirit of the **R** project, we have taken advantage of some peculiar features of the language. In particular abstraction of tasks into functions for easier readability and maintenance; functions as a data type to be passed on as arguments to other functions, in the spirit of correspondence between conceptual and software tools; and lastly, object-orientation, for example, in applying specialized algebraic methods to matrices with a peculiar structure for reliability and performance.

The main contribution of the paper is in combining existing results in the literature — the general estimation framework of Anselin (1988) and analytical derivations in Baltagi et al. (2007b) — into an operational framework for the estimation of a number of models for which, to our knowledge, no algorithms are currently available. Moreover, it describes a user-friendly, open-source implementation in the **R** language which will hopefully open up new modelling possibilities in the field of spatio-temporally correlated panels for a number of applied researchers.

Given the large number of specifications considered, some preliminary words on notation are in order. We consider either cross-sections of N data points (only in Section 4.1) or balanced panels of N data points observed over T time periods. Contrary to standard panel data practice, data are generally meant to be stacked by time, then by cross-section (so that the individual index is the "fastest" one) in order to simplify formulae especially as regards spatial filtering representations. While important for presentation clarity, this is nevertheless completely transparent for software users, who need only supply suitably indexed `data.frames` or `pdata.frames`. In the following, we will denote the composite error term in the standard linear regression model as u; ε will stand for the idiosyncratic error term, as opposed to the random effect μ, so that $u = \mu + \varepsilon$ throughout. Error covariance matrices will be denoted by Ω if unscaled, by Σ if scaled by the innovation's variance σ_e^2, so that $\Omega = \sigma_e^2 \Sigma$. Lastly, as we consider balanced panel datasets with N spatial units observed over T time periods, the dimension of spatial and covariance matrices involved will usually be $NT \times NT$. In some cases, though, we will consider $N \times N$ $(T \times T)$ submatrices pertaining to one cross-section (time period), denoting them, e.g., as Σ_N (Σ_T). The spatial weight matrix $W_N = W$ is assumed time invariant, as customary in the literature, and enters spatial panel models as $I_T \otimes W_N$ where $\otimes$ is the Kronecker product, dropping the index N when unambiguous. Software package names are in **bold**, commands and arguments are in `typewriter` font.

Estimation of all models is based on maximum likelihood methods; an assumption of normality is maintained throughout. An assessment of the appropriateness of the procedures presented here under non-normality of errors, in the spirit of Lee and Yu (2012), is left for future work.

The paper is organized as follows: the next section discusses the specification; then we review estimation theory, building on existing approaches to illustrate ours. In the subsequent section we address the practical aspects of estimation, from computational issues to the design of a user-friendly software package. A practical illustration and the conclusions follow.

2. Spatial panel models with error components

In this section we discuss the common specifications of random effects spatial panel models most frequent in the literature. At the end we introduce the general spatial autoregressive model with random effects and both spatially and temporally autoregressive errors. This last is the most general specification we consider here, and also the main subject of the paper.

Spatial panel data models capture spatial interactions across spatial units observed over time. There is an extensive literature both on static as well as dynamic models. Here we consider a general static panel model that includes a spatial lag of the dependent variable and spatial autoregressive disturbances:

$$y = \lambda(I_T \otimes W_N)y + X\beta + u$$

where y is an $NT \times 1$ vector of observations on the dependent variable, X is a $NT \times k$ matrix of observations on the non-stochastic exogenous regressors, I_T an identity matrix of dimension T, W_N is the $N \times N$ spatial weights matrix of known