`splm`: econometric analysis of spatial panel data

Giovanni Millo[1] Gianfranco Piras[2]

[1]Research Dept., Generali S.p.A. and DiSES, Univ. of Trieste
[2]REAL, UIUC

useR! Conference
Rennes, July 8th 2009

Università degli Studi di Trieste

Introduction

- `splm` = Spatial Panel Linear Models
- Georeferenced data and Spatial Econometrics
- Individual heterogeneity, unobserved effects and panel models

Outline

MOTIVATION

Motivation

Reasons for developing an R library for spatial panel data:

- Spatial econometrics has experienced an increasing interest in the last decade.
- Spatial panel data are probably one of the most promising but at the same time underdeveloped topics in spatial econometrics.
- Recently, a number of theoretical papers have appeared developing estimation procedures for different models. Among these: Anselin et al. (2008); Kapoor et al. (2007); Elhorst (2003); Lee and Yu (2008); Yu et al. (2008); Korniotis (2007); Mutl and Pfaffermayer (2008) .

Motivation

- Theoretical papers developing test procedures to discriminate among different specifications Baltagi et al. (2007); Baltagi et al. (2003).
- Although there exist libraries in R, Matlab, Stata, Phyton to estimate cross-sectional models, to the best of our knowledge, there is no other software available to estimate spatial panel data models than Elhorst's Matlab code and one (particular) Stata example on Prucha's web site

Why is R the ideal environment?

- (Faith)
- (Reason)
- ...more specifically, because of the availability of three libraries: `spdep`, `plm` and `Matrix`
  - `spdep`: extremely well known library for spatial analysis that contains spatial data infrastructure and estimation procedures for spatial cross sectional models
  - `plm`: recently developed panel data library performing the estimation of most non-spatial models, tests and data management
  - `Matrix`: library containing methods for sparse matrices (and much more). Turns out to be very relevant because of the particular properties of a spatial weights matrix.

.

The `splm` package

Taking advantage of these existing libraries we are working on the development of a library for spatial panel data models.

- From `spdep`: we are taking:
  - contiguity matrices' and spatial data management infrastructure
  - specialized object types like `nb` and `listw`
- From `plm`:
  - how to handle the double (space and time) dimension of the data
  - model object characteristics and printing methods
- From `Matrix`:
  - efficient specialized methods for sparse matrices

.

General ML Framework

General ML framework

Anselin (1988) outlines the general procedure for a model with a spatial lag and spatially autocorrelated (and possibly non-spherical) innovations:

$$\begin{aligned} y &= \lambda W_1 y + X\beta + u \\ u &= \rho W_2 u + \eta \end{aligned} \tag{1}$$

with $\eta \sim N(0, \Omega)$ and $\Omega \neq \sigma^2 \mathrm{I}$.
Special cases:

- $\lambda = 0$, **spatial error model**
- $\rho = 0$, **spatial lag model.**

General ML framework

Introducing the simplifying notation

$$A = I - \lambda W_1$$
$$B = I - \rho W_2$$

model (1) can be rewritten as:

$$\begin{aligned} Ay &= X\beta + u \\ Bu &= \eta. \end{aligned} \tag{2}$$

If there exists Ω such that $e = \Omega^{-\frac{1}{2}}\eta$ and $e \sim N(0, \sigma_e^2 I)$ and B is invertible, then

$$u = B^{-1}\Omega^{\frac{1}{2}}e$$

General ML framework

Model (1) can be written as

$$Ay = X\beta + B^{-1}\Omega^{\frac{1}{2}}e$$

or, equivalently,

$$\Omega^{-\frac{1}{2}}B(Ay - X\beta) = e$$

with e a “well-behaved" error term with zero mean and constant variance.
Unfortunately, the error term e is unobservable and therefore, to make the estimator operational, the likelihood function needs to be expressed in terms of y. This in turn requires calculating $J = det(\frac{\partial e}{\partial y}) = |\Omega^{-\frac{1}{2}}BA| = |\Omega^{-\frac{1}{2}}||B||A|$, the Jacobian of the transformation.

General ML framework

These determinants directly enter the log-likelihood, which becomes

$$logL = -\frac{N}{2} ln\pi - \frac{1}{2} ln|\Omega| + ln|B| + ln|A| - \frac{1}{2} e'e$$

Note that:

- The difference compared to the usual likelihood of the classic linear model is given by the **Jacobian** terms
- The likelihood is a function of β, λ, ρ and parameters in Ω.
- The overall errors covariance $B'\Omega B$ can in turn be scaled and written as $B'\Omega B = \sigma_e^2 \Sigma$.

Computational approach to the ML problem

A general description of the ML estimation problem consists of

- defining an efficient and parsimonious transformation of the model at hand to (unobservable) spherical errors, and the corresponding log-likelihood for the observables;
- translating the inverse and the determinant of the scaled covariance of the errors, Σ^{-1} and $|\Sigma|$, and the determinants $|A|$ and $|B|$ into computationally manageable objects;
- implementing the two-step optimization iterating between maximization of the concentrated log-likelihood and the closed-form GLS solution.

RE Models

General Random Effects Panel Model

Let us concentrate on the error model considering a panel model within a more specific, yet quite general setting, allowing for the following features of the composite error term (i.e., parameters describing Σ):

- random effects ($\phi = \sigma^2_\mu / \sigma^2_\epsilon$)
- spatial correlation in the idiosyncratic error term (λ)
- serial correlation in the idiosyncratic error term (ρ)

$$
\begin{aligned}
y &= X\beta + u \\
u &= (\iota_T \otimes \mu) + \epsilon \\
\epsilon &= \lambda(I_T \otimes W_2)\epsilon + \nu \\
\nu_t &= \rho \nu_{t-1} + e_t
\end{aligned}
$$

Particular cases of the general model

Error features can be combined, giving rise to the following possibilities:

	$\lambda, \rho \neq 0$	$\lambda \neq 0$	$\rho \neq 0$	$\lambda = \rho = 0$
$\sigma^2_\mu \neq 0$	SEMSRRE	**SEMRE**	SSRRE	RE
$\sigma^2_\mu = 0$	SEMSR	**SEM**	SSR	OLS

where SEMRE is the 'usual' random effects spatial error panel and SEM the standard spatial error model (here, pooling with $W = I_T \otimes w$)

The likelihoods involved give rise to computational issues that limit the application to data sets of certain dimensions.

FE Models

FE Models

A **fixed effect spatial lag** model can be written in stacked form as:

$$y = \lambda(I_T \otimes W_N)y + (\iota_T \otimes \alpha) + X\beta + \varepsilon \tag{3}$$

where λ is a spatial autoregressive coefficient.
On the other hand, a **fixed effects spatial error model** can be written as:

$$\begin{aligned} y &= (\iota_T \otimes \alpha) + X\beta + u \\ u &= \rho(I_T \otimes W_N)u + \varepsilon \end{aligned} \tag{4}$$

where ρ is a spatial autocorrelation coefficient.

FE Spatial Panel estimation

The estimation procedure for both models is based on maximum likelihood and can be summarized as follows.

1. Applying OLS to the demeaned model.
2. Plug the OLS residuals into the expression for the concentrated likelihood to obtain an initial estimate of ρ.
3. The initial estimate for ρ can be used to perform a spatial FGLS to obtain estimates of the regression coefficients, the error variance and a new set of estimated GLS residuals.
4. An iterative procedure may then be used in which the concentrated likelihood and the GLS estimators for, respectively, ρ and β are alternately estimated until convergence.

GM Approach

GM approach

Consider again the panel model written stacking the observations by cross-section:

$$y_N = X_N\beta + u_N \tag{5}$$

and

$$u_N = \rho(I_T \otimes W_N)u_N + \varepsilon_N \tag{6}$$

Kapoor et al. (2007) consider an error component structure for the whole innovation vector in (6), that is:

$$\varepsilon_N = (e_T \otimes I_N)\mu_N + \nu_N \tag{7}$$

i.e., the individual error components are spatially correlated as well (different economic interpretation).

GM RE Estimation Procedure

The estimation procedure can be summarized by the following steps:

1. Estimate the regression equation by OLS to obtain an estimate of the vector of residuals to employ in the GM procedure
2. Estimate ρ and the variance components σ_1^2 and σ_ν^2 using one of the three set of GM estimators proposed
3. Use the estimators of ρ, σ_ν^2 and σ_1^2 to define a corresponding feasible GLS estimator of β.

Spatial simultaneous equation model

The system of spatially interrelated cross sectional equations corresponding to N cross sectional units can be represented as (Kelejian and Prucha, 2004) :

$$\mathbf{Y} = \mathbf{YB} + \mathbf{XC} + \mathbf{Q}\lambda + \mathbf{U} \tag{8}$$

with $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$, $\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_m)$ and $\mathbf{q}_j = \mathbf{W}\mathbf{y}_j$, $j = 1, \ldots m$.
$\mathbf{y}_j$ is an $N \times 1$ vector of cross-sectional observations on the dependent variable in the jth equation, $\mathbf{x}_l$ is an $N \times 1$ vector of observations on the lth exogenous variable, $\mathbf{u}_j$ is the $N \times 1$ disturbance vector in the jth equation. Finally, $\mathbf{W}$ is an $N \times N$ weights matrix of known constants, and $\mathbf{B}$, $\mathbf{C}$ and λ are corresponding matrices of parameters.

Spatial simultaneous equation model

The model also allows for **spatial autocorrelation in the disturbances.** In particular, it is assumed that the disturbances are generated by the following spatial autoregressive process:

$$\mathbf{U} = \mathbf{M}\mathbf{R} + \mathbf{E} \tag{9}$$

where $\mathbf{E} = (\varepsilon_1, \ldots, \varepsilon_m)$, $\mathbf{R} = diag_{j=1}^{m}(\rho_j)$, $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_m)$ and $\mathbf{m}_j = \mathbf{W}\mathbf{u}_j$. ε_j and ρ_j denotes, respectively, the vector of innovations and the spatial autoregressive parameter in the jth equation.

TESTS

Baltagi et al. 2003; 2007

Consider the general random effect model with serial and cross-sectional correlation in the error term:

$$
\begin{aligned}
y &= X\beta + u \\
u &= (\imath_T \otimes \mu) + \epsilon \\
\epsilon &= \lambda(I_T \otimes W_2)\epsilon + \nu \\
\nu_t &= \rho\nu_{t-1} + e_t
\end{aligned}
$$

Baltagi et al. (2003) derive tests for the following hypothesis (for a model where ρ is zero):

- λ, μ (needs OLS estimates of $\hat{u}$)
- $\lambda|\mu$ (needs SEM estimates of $\hat{u}$)
- $\mu|\lambda$ (needs RE estimates of $\hat{u}$)

Baltagi et al. 2007

Baltagi et al. (2007) derive tests for the following hypothesis:

- $\lambda|\rho, \mu$ (needs SSRRE estimates of $\hat{u}$)
- $\rho|\lambda, \mu$ (needs SEMRE estimates of $\hat{u}$)
- $\mu|\lambda, \rho$ (needs SEMSR estimates of $\hat{u}$)

So a viable and computationally parsimonious strategy for the error model can be to test in the three directions by means of conditional LM tests and see whether one can estimate a simpler model than the general one.

DEMONSTRATIONS

Munnel (1990) data set and spatial weights matrix

Munnell's (1990) data on public capital productivity:

- 48 continental US states
- 17 years (1970-1987): but we consider the subset 1970-74
- binary proximity (contiguity) matrix

```
> data(Produc,package="Ecdat")
> Produc <- Produc[Produc$year %in% 1970:1974, ]
> fm <- log(gsp) ~ log(pcap) + log(pc) + log(emp) + unemp
> load("usaww.rda")
```

Is the coefficient on *pcap* in this production function significantly positive?

ML estimator of the Munnell (1990) model (Random Effects)

```
> system.time(re.mod<-spreml(fm,data=Produc,w=usaww,errors="re"))
   user  system elapsed
  4.773   0.789   5.547
> summary(re.mod)
Spatial panel random effects ML model

Call:
spreml(formula = fm, data = Produc, w = usaww, errors = "re")

Residuals:
    Min.   1st Qu.   Median  3rd Qu.     Max.
-0.22800 -0.07210 -0.00288  0.05930  0.36400

Error variance parameters:
    Estimate Std. Error t-value  Pr(>|t|)
phi  17.5954     3.6329  4.8433 1.277e-06 ***

Coefficients:
              Estimate Std. Error t-value  Pr(>|t|)
(Intercept)  1.6794519  0.1961403  8.5625 < 2.2e-16 ***
log(pcap)    0.0946963  0.0525863  1.8008  0.071738 .
log(pc)      0.3411079  0.0424213  8.0410 8.914e-16 ***
log(emp)     0.6271917  0.0377445 16.6168 < 2.2e-16 ***
unemp       -0.0100227  0.0026808 -3.7386  0.000185 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Munnell's model (Random Effects and Spatial Dependence)

```
> system.time(semre.mod<-spreml(fm,data=Produc,w=usaww,errors="semre"))
   user  system elapsed
  1.297   0.304   1.993
> summary(semre.mod)
Spatial panel random effects ML model

Call:
spreml(formula = fm, data = Produc, w = usaww, errors = "semre")

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.24300 -0.07350 -0.00701 -0.00119  0.06370  0.35500

Error variance parameters:
        Estimate Std. Error t-value  Pr(>|t|)
phi    21.828390   5.065227  4.3095 1.637e-05 ***
lambda  0.548171   0.064972  8.4370 < 2.2e-16 ***

Coefficients:
              Estimate Std. Error t-value  Pr(>|t|)
(Intercept)  1.7044885  0.1999468  8.5247 < 2.2e-16 ***
log(pcap)    0.0469036  0.0510134  0.9194  0.357867
log(pc)      0.3717287  0.0413017  9.0003 < 2.2e-16 ***
log(emp)     0.6416880  0.0377657 16.9913 < 2.2e-16 ***
unemp       -0.0067800  0.0026266 -2.5813  0.009843 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Munnell’s model (Complete Model)

```
> system.time(semsrre.mod<-spreml(fm,data=Produc,w=usaww,errors="semsrre"))
   user  system elapsed
 25.949   7.606  33.612
> summary(semsrre.mod)
Spatial panel random effects ML model

Call:
spreml(formula = fm, data = Produc, w = usaww, errors = "semsrre")

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.24500 -0.06770 -0.00149  0.00526  0.07420  0.36300

Error variance parameters:
       Estimate Std. Error t-value  Pr(>|t|)
phi    5.422708   1.287023  4.2134 2.516e-05 ***
rho    0.588131   0.073678  7.9824 1.435e-15 ***
lambda 0.669867   0.052941 12.6530 < 2.2e-16 ***

Coefficients:
              Estimate Std. Error t-value  Pr(>|t|)
(Intercept)  1.7263508  0.2341863  7.3717 1.685e-13 ***
log(pcap)    0.0202843  0.0567305  0.3576   0.72068
log(pc)      0.3851314  0.0474891  8.1099 5.067e-16 ***
log(emp)     0.6536615  0.0439349 14.8780 < 2.2e-16 ***
unemp       -0.0057097  0.0024116 -2.3676   0.01790 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Munnell's model (Error components GM estimator)

```
> system.time(spregm.mod<-spregm(fm,data=Produc,w=usaww))
   user  system elapsed
  0.073   0.001   0.107
> summary(spregm.mod)
Spatial panel random effects GM model

Call:
spregm(formula = fm, data = Produc, w = usaww)

Residuals:
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.088000 -0.012800 -0.000280  0.000122  0.013200  0.097200

Estimated spatial coefficient, variance components and theta:
             Estimate
rho        0.54467770
sigma^2_v  0.00059389
sigma^2_1  0.03442833
theta      0.86866112

Coefficients:
               Estimate Std. Error t-value  Pr(>|t|)
(Intercept)  1.3441929  0.2118448  6.3452 2.222e-10 ***
log(pcap)    0.0617207  0.0525437  1.1747   0.24013
log(pc)      0.4374493  0.0441520  9.9078 < 2.2e-16 ***
log(emp)     0.5740508  0.0437860 13.1104 < 2.2e-16 ***
unemp       -0.0073729  0.0031552 -2.3367   0.01945 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Baltagi et al. 2003

```
> system.time(bsk<-bsktest(lm(fm,data=Produc), test="LMJOINT",w=mat2listw(usaww),
index=Produc[,c(1,2)]))
   user  system elapsed
  0.131   0.002   0.137
> bsk

   Baltagi, Song and Koh LM-H one-sided joint test

data:  lm(formula = fm, data = Produc)
LM-H = 387.6268, p-value = 0.01
alternative hypothesis: Random Regional Effects and Spatial autocorrelation
```

Baltagi et al. 2003

```
> system.time(bsk<-bsktest(fm,data=Produc, w=mat2listw(usaww), test="CLMlambda"))
   user  system elapsed 
  3.629   0.762   4.371 
> bsk

   Baltagi, Song and Koh LM*-lambda conditional LM test (assuming sigma^2_mu >= 0)

data:  log(gsp) ~ log(pcap) + log(pc) + log(emp) + unemp 
LM*-lambda = 7.6657, p-value = 8.893e-15
alternative hypothesis: Spatial autocorrelation
```

CONCLUSIONS

Recap and Next Steps

The functionalities of the package allow to conduct specification search and model estimation according to current practice in a straightforward way. Most future developments will take place "under the hood".

Directions for future work:

- Complete Montecarlo checks
- Improve interface consistency
- Including different estimators and additional models that are already available in the literature
- Improve the efficiency of ML estimation
- ...complete packaging and put package on CRAN