ETH zürich

A Panel Data Approach for Spatial and Network Selection Models

Sophia Ding[1], Peter H. Egger[2]

[1]**ETH Zurich;** ETH Zurich, CEPR, CESifo

1 Introduction

Motivation:

- Common features of economic data: **spatial/network pattern** (cross-sectional interdependence), non-randomly missing observations (**sample selection/treatment selection**), panel-data
- No model exists to deal with all three simultaneously!
- Neglecting selection and/or spatial/network correlation results in **biased** coefficient estimates!

	Cross Section	Panel
Non-Spatial/ Non-Network	Heckman (1976, 1979)	Wooldridge (1995)
Spatial/ Network	McMillen (1995), Flores-Lagunes, Schnier (2012), Doğan,Taşpinar (2017)	**This paper!**

Example: Export-Wage Premium

- Empirical and theoretical evidence that **exporters pay higher wage/worker** than non-exporting firms (**treatment effect** of exporter status)
- Exporting decision as well as wage/worker depends on latent export profitability → **treatment ≠ random**
- Wages may have a **spatial pattern** due to local labor markets, commuting, etc. → Shocks to wages are **correlated across firms**!
- Profitability of exporting may have **network pattern** due to input/output linkages or industry affiliation → Shocks to profitability are **correlated across firms**!

This paper:

Develop **two-step approach** towards selection on unobservables akin to Heckman (1976, 1979) and Wooldridge (1995) but for **panel-data** with **spatial or network interdependencies** in both the selection and the outcome equation.

2 Econometric Model

Selection equation

- Fixed effects (Mundlak 1978, Wooldridge 1995)

$$\begin{aligned} y_{ti}^{A*} &= x_{ti}^{A\prime}\beta^A + e_{ti}^A, \qquad \forall \quad t=1,...,T; i=1,...,N, \\ \text{with} \quad e_{ti}^A &= \bar{x}_i^{A\prime}\delta^A + \underbrace{\mu_i^A + \varepsilon_{ti}^A}_{=\xi_{ti}^A} \\ y_{ti}^A &= 1[y_{ti}^{A*} > 0], \end{aligned}$$

- Panel SAR process (Kapoor, Kelejian, and Prucha, 2007)

$$\begin{aligned} e_{ti}^A &= \rho^A \sum_{j=1}^N w_{tij} e_{tj}^A + \bar{x}_i^{A\prime}\delta^A + \xi_{ti}^A \\ e_{ti}^A &= \sum_{j=1}^N r_{tij}^A \bar{x}_j^{A\prime}\delta^A + \underbrace{\sum_{j=1}^N r_{tij}^A \xi_{tj}^A}_{=u_{ti}^A}, \quad \text{using} \quad R_t^A = (I_N - \rho^A W_t)^{-1} = (r_{tij}^A) \end{aligned}$$

Selection equation restated: $y_{ti}^{A*} = x_{ti}^{A\prime}\beta^A + \sum_{j=1}^N r_{tij}^A \bar{x}_j^{A\prime}\delta^A + u_{ti}^A$

Outcome equation

- Fixed effects + panel SAR process
- Correct for selection bias by making use of joint normality assumption of spatial/network error components

Spatial/Network Sample Selection

$$E[y_{ti}^B | y_{ti}^A = 1, x^{A0}, x^B] = x_{ti}^{B\prime}\beta^B + \sum_{j=1}^N r_{tij}^B \bar{x}_j^{B\prime}\delta^B + E[u_{ti}^B | y_{ti}^A = 1, x^{A0}, x^B]$$

Spatial/Network Treatment Selection

$$E[y_{ti}^B | y_{ti}^A, x^{A0}, x^B] = \alpha y_{ti}^A + x_{ti}^{B\prime}\beta^B + \sum_{j=1}^N r_{tij}^B \bar{x}_j^{B\prime}\delta^B + E[u_{ti}^B | y_{ti}^A, x^{A0}, x^B]$$

- Spatially/network adjusted (generalized) Inverse Mills' Ratio (=Correction Function):

$$\begin{aligned} E[u_{ti}^B | y_{ti}^A = 1, x^{A0}, x^B] &= \frac{\sigma_{\xi^{BA}} \sum_{j=1}^N r_{tij}^B r_{tij}^A}{\sqrt{\sigma_{\xi^A}^2 \sum_j^N (r_{tij}^A)^2}} \frac{\phi(z_{ti})}{\Phi(z_{ti})} = \tau\psi_{ti}\lambda_{ti} \\ E[u_{ti}^B | y_{ti}^A, x^{A0}, x^B] &= \frac{\sigma_{\xi^{BA}} \sum_{j=1}^N r_{tij}^B r_{tij}^A}{\sqrt{\sigma_{\xi^A}^2 \sum_j^N (r_{tij}^A)^2}} \phi(z_{ti}) \frac{y_{ti}^A - \Phi(z_{ti})}{\Phi(z_{ti})[1-\Phi(z_{ti})]} = \tau\psi_{ti}\lambda_{ti}^g \end{aligned}$$

3 Estimation Strategy (Outline)

Step 1: Estimate selection equation using **Pooled Bayesian Spatial/ Network Error Probit** model to obtain $\hat{\theta}_A = \{\hat{\tilde{\beta}}^A, \hat{\tilde{\delta}}^B, \hat{\rho}^A\}$, where $\tilde{\beta}^A = \frac{\beta^A}{\sigma_{\xi^A}}$, $\tilde{\delta}^A = \frac{\delta^A}{\sigma_{\xi^A}}$.

Step 2: Use estimated parameters to construct spatially/network adjusted (generalized) Inverse Mills' Ratio.

Step 3: Add estimated spatially/network adjusted (generalized) Inverse Mills' Ratio in outcome equation and estimate using **Pooled Non-linear Least Squares** to obtain $\hat{\theta}^B = \{\hat{\beta}^B, \hat{\delta}^B, \hat{\tau}, \hat{\rho}^B\}$, or $\hat{\theta}^B = \{\hat{\alpha}, \hat{\beta}^B, \hat{\delta}^B, \hat{\tau}, \hat{\rho}^B\}$.

4 Monte Carlo Evidence (Results)

Case 1: Medium Spatial/Network Correlation

			$\tilde{\beta}_1^A$	$\tilde{\beta}_2^A$	$\tilde{\delta}_1^A$	$\tilde{\delta}_2^A$	ρ^A	β_1^B	δ_1^B	τ	ρ^B
		True	0.707	0.707	0.707	0.707	0.5	1	3	0.707	0.5
N=250	SNSS	Mean	0.732	0.735	0.736	0.734	0.449	0.999	3.011	0.699	0.487
		Bias	0.025	0.028	0.029	0.027	-0.051	-0.001	0.011	-0.009	-0.013
		RMSE	0.091	0.084	0.198	0.185	0.131	0.095	0.192	0.173	0.109
	WPS	Mean	0.653	0.670	0.681	0.708		0.964	3.038	0.679	
Ignore spatial/ network correlation		Bias	-0.054	-0.037	-0.026	0.001		-0.036	0.038	-0.028	
		RMSE	0.096	0.081	0.140	0.134		0.102	0.192	0.189	
	NLLS	Mean						0.987	2.988		0.421
Ignore sample selection		Bias						-0.013	-0.012		-0.079
		RMSE						0.095	0.189		0.149
N=500	SNSS	Mean	0.716	0.719	0.721	0.711	0.482	0.999	3.001	0.702	0.497
		Bias	0.009	0.012	0.014	0.004	-0.018	-0.001	0.001	-0.005	-0.003
		RMSE	0.055	0.060	0.134	0.140	0.074	0.061	0.141	0.127	0.061
	WPS	Mean	0.660	0.657	0.686	0.671		0.982	3.222	0.808	
Ignore spatial/ network correlation		Bias	-0.047	-0.050	-0.022	-0.036		-0.018	0.222	0.101	
		RMSE	0.069	0.074	0.098	0.109		0.064	0.264	0.170	
	NLLS	Mean						0.990	2.987		0.508
Ignore sample selection		Bias						-0.010	-0.013		0.008
		RMSE						0.061	0.141		0.064

Case 2: No Spatial/Network Correlation

			$\tilde{\beta}_1^A$	$\tilde{\beta}_2^A$	$\tilde{\delta}_1^A$	$\tilde{\delta}_2^A$	ρ^A	α	β_1^B	δ_1^B	τ	ρ^B
		True	0.707	0.707	0.707	0.707	0	1	1	3	0.707	0
N=250	SNTS	Mean	0.739	0.740	0.743	0.744	-0.114	1.005	1.000	3.006	0.697	-0.009
		Bias	0.032	0.033	0.036	0.037	-0.114	0.005	0.000	0.006	-0.010	-0.009
		RMSE	0.092	0.085	0.193	0.177	0.245	0.112	0.051	0.132	0.106	0.137
	WPS	Mean	0.715	0.716	0.719	0.718		1.000	0.999	3.009	0.706	
Ignore spatial/ network correlation		Bias	0.008	0.009	0.011	0.011		0.000	-0.001	0.009	-0.001	
		RMSE	0.082	0.075	0.143	0.140		0.112	0.050	0.130	0.106	
	NLLS	Mean						1.350	0.942	2.973		0.067
Ignore sample selection		Bias						0.350	-0.058	-0.027		0.067
		RMSE						0.362	0.076	0.130		0.148
N=500	SNTS	Mean	0.721	0.723	0.730	0.718	-0.066	1.003	0.999	3.002	0.701	-0.006
		Bias	0.014	0.016	0.023	0.011	-0.066	0.003	-0.001	0.002	-0.006	-0.006
		RMSE	0.055	0.059	0.133	0.133	0.176	0.086	0.033	0.099	0.076	0.087
	WPS	Mean	0.709	0.711	0.714	0.706		1.001	0.999	3.002	0.705	
Ignore spatial/ network correlation		Bias	0.002	0.004	0.007	-0.001		0.001	-0.001	0.002	-0.002	
		RMSE	0.052	0.056	0.102	0.102		0.086	0.033	0.096	0.076	
	NLLS	Mean						1.358	0.942	2.935		-0.016
Ignore sample selection		Bias						0.358	-0.058	-0.065		-0.016
		RMSE						0.365	0.066	0.114		0.093

References

Doğan, O., Taşpinar, S., 2017. Bayesian Inference in Spatial Sample Selection Models. Oxford Bulleting of Economics and Statistics 80, 90-121.

Flores-Lagunes, A., Schnier, K.E., 2012. Estimation of Sample Selection Models with Spatial Dependence. Journal of Applied Econometrics 27, 173-204.

Heckman, J., 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. The Annals of Economic and Social Measurement 5, 475-492.

Heckman, J., 1979. Sample Selection Bias as a Specification Error. Econometrica 47 (1), 153-61.

Kapoor, M, Kelejian, H.H. Prucha, I.R., 2007. Panel Data Models with Spatially Correlated Error Components. Journal of Econometrics 140, 97-130.

Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69-85.

McMillen, D.P., 1995. Selection Models in Spatial Econometric Models. Journal of Regional Science 35 (3), 417-436.

Wooldridge, J., 1995. Selection corrections for panel data models under conditional mean independence assumptions. Journal of Econometrics 68 (1), 115-132.