Mixture Distributions with Spatial Dependency

Gary Cornwall[1] Olivier Parent[2]

[1]Department of Economics
University of Cincinnati

[2]Department of Economics
University of Cincinnati

SRSA Conference, 2016

Outline

Spatial Dependency and Heterogeneity

Why is it important?

Spatial Dependence

- My outcomes are influenced by those of my neighbors.
- Spatial Autoregressive Model (SAR) - Global and Simultaneous
- Creates a homogenizing effect as the dependence gets large.

Spatial Heterogeneity

- Spatial process for which the mean varies from cluster to cluster.
- Creates areas of higher spatial intensity.

Spatial Heterogeneity vs Spatial Mixtures

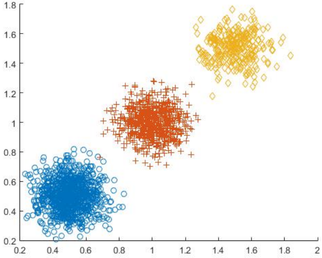

Figure: Spatial Heterogeneity

- Mean varies across **clusters**.
- A form of fixed effects can be implemented by allowing intercept to vary.
- Approach requires delineation of data into smaller clusters which may or may not be feasible.

Spatial Heterogeneity vs Spatial Mixtures

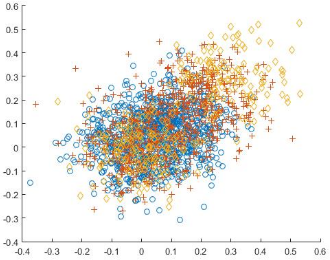

Figure: Spatial Dependence/Mixture

- Mean may or may not vary across **groups** $(g = 1, 2, \ldots, G)$
- Fixed effects may be spurious.
- SAR model overstates spatial dependency.
- Delineation into small **clusters** is nearly impossible with any sort of accuracy

What is the big deal?

- Specification is important, the two are not modeled with the same techniques.
- Work by Anselin and Arribas-Bel (2013) shows that results can be spurious if specification is picked incorrectly.
- The Spatial-Mix (S-M) Model nests both the spatial heterogeneity and dependence as special cases.

The Spatial Autoregressive Mixture Model

Model the heterogeneity as a finite mixture of distributions.

$$p(y_i|x_i,\beta,\Sigma,\pi) = \sum_{g=1}^{G} \pi_g N(y_i|x_i\beta_g,\sigma_g^2), \quad \sum_{g=1}^{G} \pi_g = 1 \tag{1}$$

Model the spatial dependency in a SAR form

$$\begin{aligned} y &= \rho W y + x\beta + \epsilon \\ \epsilon &\sim N(0,\sigma^2 I_N) \end{aligned} \tag{2}$$

The SAR Mixture DGP
Reduced Form

$$Y = (I - \rho W)^{-1}\tilde{X}\beta + (I - \rho W)^{-1}\tilde{\epsilon} \tag{3}$$

Where,

$$\begin{aligned}
\tilde{X} &= (\iota_g \otimes x) \odot (z \otimes \iota_K) \\
\tilde{\epsilon} &= (z\Sigma^{1/2} \odot \epsilon) \\
x &\sim N(5,9)^* \\
\epsilon &\sim N(0,1)^*
\end{aligned}$$

- $\iota_{G/K}$ are sizing vectors.
- Σ is the diagonal of a $(G \times G)$ variance-covariance matrix.
- z is a $(N \times G)$ indicator matrix.

* indicate values used in simulations

The Spatial Autoregressive Mixture Model

Given the latent component indicator the joint density of the data has the following representation:

$$p(y|x,z,\psi,\beta,\Sigma,\pi) = |I_N - \tilde{\psi}W|^{-1} \prod_{g=1}^{G} \Bigg\{ (2\pi\sigma_g^2)^{N_g/2} \exp\left[-\frac{1}{2\sigma_g^2} \sum_{i\in I_g} \left(y_i - \tilde{\psi}_i \sum_{j\in\delta_i} w_{ij} y_j - x_i\beta_g \right)^2 \right] \Bigg\}, \quad (4)$$

With a full joint density of:

$$\begin{aligned} p(y,z|x,\beta,\Sigma,\psi,\pi) &= p(y|z,x,\beta,\Sigma,\psi,\pi)p(z|x,\beta,\Sigma,\psi,\pi) \quad (5) \\ &= \prod_{g=1}^{G} \prod_{i\in I_g} N(\tilde{y}_i|x_i\beta_g,\sigma_g^2)\pi_g^{N_g} \end{aligned}$$

Unknown Parameters

- β - $(N \times KG)$ vector of coefficients/parameters.
- Σ - $(G \times G)$ variance-covariance matrix.
- Ω - $(N \times G)$ identification probability matrix.
- Z - $(N \times G)$ group indicator matrix.

Observed Information

- Y - $(N - dimensional)$ vector of outcomes.
- X - $(N \times K)$ matrix of information.
- W - $(N \times N)$ row normalized weight matrix.

Simulation Architecture

- Data is generated using outlined process.
- Gibbs sampler is employed to estimate parameters (5000 iterations).
- Process was repeated 100 times for each DGP.
- DGPs were varied based on signal to noise ratio, values of ρ, locational, and group parameters.

Simulation Results $\rho = 0$

Table: Estimation Results $\rho = 0$

Parameter	True Values	S-M	L95	U95	O-M	L95	U95		SAR	L95	U95
β_{11}	-1.100	-1.105	-1.148	-1.060	-1.099	-1.141	-1.055	β_1	-0.430	-0.592	-0.265
β_{12}	1.400	1.394	1.349	1.436	1.400	1.356	1.442	β_2	1.312	1.162	1.461
β_{21}	2.500	2.482	2.408	2.551	2.489	2.415	2.558	-	-	-	-
β_{22}	1.900	1.869	1.787	1.935	1.875	1.795	1.941	-	-	-	-
β_{31}	-0.800	-0.739	-0.884	-0.561	-0.736	-0.880	-0.557	-	-	-	-
β_{32}	-2.100	-2.079	-2.229	-1.912	-2.076	-2.225	-1.908	-	-	-	-
ρ	0.000	0.012	0.001	0.023	-	-	-	ρ	0.136	0.071	0.200
σ_1^2	3.000	7.686	3.226	14.962	7.676	3.212	14.965	σ^2	128.035	117.440	139.334
σ_2^2	2.000	4.703	1.914	6.901	4.696	1.899	6.891	-	-	-	-
σ_3^2	1.000	13.357	4.592	34.490	13.360	4.583	34.541	-	-	-	-
π_1	0.500	0.499	0.475	0.523	0.499	0.475	0.523	-	-	-	-
π_2	0.350	0.350	0.328	0.372	0.350	0.329	0.372	-	-	-	-
π_3	0.150	0.151	0.134	0.168	0.151	0.134	0.168	-	-	-	-

Simulation Results $\rho = .4$

Table: Estimation Results $\rho = .4$

Parameter	True Values	S-M	L95	U95	O-M	L95	U95		SAR	L95	U95
β_{11}	-1.100	-1.111	-1.154	-1.066	-0.838	-0.933	-0.743	β_1	-0.454	-0.615	-0.290
β_{12}	1.400	1.383	1.338	1.425	1.785	1.688	1.881	β_2	1.279	1.127	1.430
β_{21}	2.500	2.469	2.394	2.539	2.927	2.773	3.074	-	-	-	-
β_{22}	1.900	1.856	1.774	1.923	2.200	2.037	2.350	-	-	-	-
β_{31}	-0.800	-0.744	-0.889	-0.566	-0.294	-0.640	0.074	-	-	-	-
β_{32}	-2.100	-2.079	-2.228	-1.911	-1.931	-2.254	-1.586	-	-	-	-
ρ	0.400	0.417	0.409	0.427	-	-	-	ρ	0.512	0.464	0.559
σ_1^2	3.000	7.656	3.230	14.880	39.499	25.578	57.182	σ^2	124.795	114.395	135.906
σ_2^2	2.000	4.694	1.946	6.888	21.232	16.144	27.538	-	-	-	-
σ_3^2	1.000	13.377	4.603	34.357	71.101	43.555	109.165	-	-	-	-
π_1	0.500	0.499	0.475	0.523	0.479	0.451	0.508	-	-	-	-
π_2	0.350	0.350	0.328	0.372	0.377	0.351	0.402	-	-	-	-
π_3	0.150	0.151	0.134	0.169	0.145	0.125	0.166	-	-	-	-

Simulation Results $\rho = .8$

Table: Estimation Results $\rho = .8$

Parameter	True Values	S-M	L95	U95	O-M	L95	U95		SAR	L95	U95
β_{11}	-1.100	-1.119	-1.163	-1.074	2.073	0.657	2.377	β_1	-0.538	-0.699	-0.374
β_{12}	1.400	1.369	1.324	1.413	2.698	2.227	3.628	β_2	1.182	1.026	1.337
β_{21}	2.500	2.455	2.378	2.527	3.371	3.524	4.760	-	-	-	-
β_{22}	1.900	1.846	1.764	1.913	2.891	2.660	4.178	-	-	-	-
β_{31}	-0.800	-0.753	-0.896	-0.578	2.265	1.667	3.043	-	-	-	-
β_{32}	-2.100	-2.079	-2.226	-1.914	1.887	-0.578	1.838	-	-	-	-
ρ	0.800	0.810	0.805	0.815	-	-	-	ρ	0.867	0.846	0.888
σ_1^2	3.000	7.700	3.248	14.875	783.200	611.200	976.700	σ^2	121.371	111.180	132.278
σ_2^2	2.000	4.695	1.977	6.876	559.600	473.800	653.300	-	-	-	-
σ_3^2	1.000	13.313	4.634	34.168	1008.100	809.200	1282.200	-	-	-	-
π_1	0.500	0.498	0.475	0.522	0.431	0.417	0.639	-	-	-	-
π_2	0.350	0.349	0.327	0.371	0.304	0.196	0.381	-	-	-	-
π_3	0.150	0.153	0.136	0.170	0.264	0.130	0.260	-	-	-	-

The Model

A simple hedonic pricing model for homes in the greater Cincinnati area:

$$
\begin{aligned}
ln(Price) = \alpha + \beta_1 DistCBD + \beta_2 Sqftland + \beta_3 Sqftbuild \\
+ \beta_4 Numbdr + \beta_5 Age + \beta_6 Styheight + \beta_7 Age65 \\
+ \beta_8 White + \beta_9 HSdegree + \epsilon
\end{aligned}
$$

The Data

Table: Data Statistics

Variable	Mean	S.D.	Min	Max
Price	127,385.142	167,194.08	3,500	3,500,000.00
DistCBD	10.02	0.54	7.52	10.92
Sqftland (in log)	8.69	0.75	3.69	12.98
Sqftbuilt (in log)	6.90	0.31	3.69	8.77
Numbdrooms	3.07	1.09	0.00	20.00
Age	86.23	30.15	0.00	198.00
Styheight	1.70	0.51	1.00	4.00
Age65	11.51	4.68	0.00	28.10
White	59.25	28.58	3.23	100.00
HSdegree	84.69	10.30	58.50	100.00

The Data

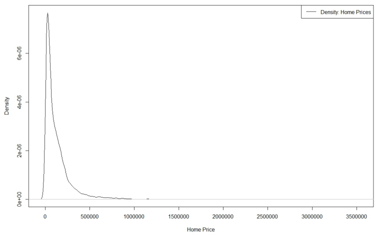

Figure: Density: Home Price

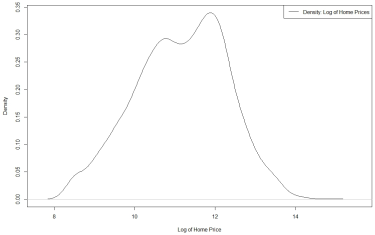

Figure: Density: Log of Home Price

Figure: Probabilities (ω_{i1}) belong to 1st group

Results

Table: Estimation Results

Parameter	SAR Model				3-component mixture model											
	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95
Constant	0.122	0.265	-0.304	0.561	1.865	0.354	1.285	2.445	2.950	0.677	1.839	4.045	11.318	2.392	7.700	15.532
DistCBD	**-0.150**	0.018	-0.179	-0.120	**-0.289**	0.024	-0.329	-0.249	**-0.169**	0.043	-0.241	-0.099	**-0.533**	0.180	-0.844	-0.252
Sqftland	0.070	0.012	0.050	0.090	0.090	0.019	0.060	0.121	0.028	0.046	-0.044	0.107	0.004	0.107	-0.135	0.212
Sqfrbuild	0.477	0.029	0.427	0.525	0.731	0.040	0.665	0.796	0.516	0.075	0.390	0.637	0.329	0.215	-0.025	0.679
Numbdrooms	0.005	0.009	-0.010	0.020	0.014	0.011	-0.004	0.032	-0.012	0.016	-0.039	0.015	0.153	0.078	0.030	0.285
Age	-0.005	0.000	-0.005	-0.004	-0.002	0.000	-0.003	-0.002	-0.014	0.001	-0.016	-0.013	0.003	0.002	-0.001	0.006
Styheight	0.288	0.018	0.258	0.316	0.437	0.024	0.398	0.477	0.353	0.040	0.289	0.418	0.428	0.201	0.074	0.730
Age65	0.004	0.002	0.001	0.007	0.012	0.002	0.008	0.015	0.008	0.004	0.002	0.014	0.017	0.015	-0.007	0.043
White	0.003	0.001	0.002	0.003	0.004	0.001	0.003	0.005	0.008	0.001	0.007	0.009	0.001	0.003	-0.004	0.005
HSdegree	0.027	0.001	0.025	0.029	0.044	0.002	0.041	0.047	0.044	0.002	0.040	0.048	-0.003	0.009	-0.019	0.010
σ^2	0.266	0.006	0.257	0.276	0.172	0.011	0.154	0.191	0.456	0.022	0.420	0.492	0.684	0.181	0.456	1.045
π_g					0.520	0.021	0.485	0.555	0.418	0.023	0.381	0.456	0.062	0.009	0.049	0.077
ρ	0.543	0.010	0.525	0.559	0.182	0.016	0.156	0.207								

Results

Table: Estimation Results

Parameter	SAR Model				3-component mixture model											
	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95
Constant	0.122	0.265	-0.304	0.561	1.865	0.354	1.285	2.445	2.950	0.677	1.839	4.045	11.318	2.392	7.700	15.532
DistCBD	**-0.150**	0.018	-0.179	-0.120	**-0.289**	0.024	-0.329	-0.249	**-0.169**	0.043	-0.241	-0.099	**-0.533**	0.180	-0.844	-0.252
Sqftland	0.070	0.012	0.050	0.090	0.090	0.019	0.060	0.121	0.028	0.046	-0.044	0.107	0.004	0.107	-0.135	0.212
Sqfrbuild	0.477	0.029	0.427	0.525	0.731	0.040	0.665	0.796	0.516	0.075	0.390	0.637	0.329	0.215	-0.025	0.679
Numbdrooms	0.005	0.009	-0.010	0.020	0.014	0.011	-0.004	0.032	-0.012	0.016	-0.039	0.015	0.153	0.078	0.030	0.285
Age	-0.005	0.000	-0.005	-0.004	-0.002	0.000	-0.003	-0.002	-0.014	0.001	-0.016	-0.013	0.003	0.002	-0.001	0.006
Styheight	0.288	0.018	0.258	0.316	0.437	0.024	0.398	0.477	0.353	0.040	0.289	0.418	0.428	0.201	0.074	0.730
Age65	0.004	0.002	0.001	0.007	0.012	0.002	0.008	0.015	0.008	0.004	0.002	0.014	0.017	0.015	-0.007	0.043
White	0.003	0.001	0.002	0.003	0.004	0.001	0.003	0.005	0.008	0.001	0.007	0.009	0.001	0.003	-0.004	0.005
HSdegree	0.027	0.001	0.025	0.029	0.044	0.002	0.041	0.047	0.044	0.002	0.040	0.048	-0.003	0.009	-0.019	0.010
σ^2	0.266	0.006	0.257	0.276	0.172	0.011	0.154	0.191	0.456	0.022	0.420	0.492	0.684	0.181	0.456	1.045
π_g					0.520	0.021	0.485	0.555	0.418	0.023	0.381	0.456	0.062	0.009	0.049	0.077
ρ	0.543	0.010	0.525	0.559	0.182	0.016	0.156	0.207								

Results

Table: Estimation Results

Parameter	SAR Model Mean	s.d.	0.05	0.95	3-component mixture model Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95	Mean	s.d.	0.05	0.95
Constant	0.122	0.265	-0.304	0.561	1.865	0.354	1.285	2.445	2.950	0.677	1.839	4.045	11.318	2.392	7.700	15.532
DistCBD	**-0.150**	0.018	-0.179	-0.120	**-0.289**	0.024	-0.329	-0.249	**-0.169**	0.043	-0.241	-0.099	**-0.533**	0.180	-0.844	-0.252
Sqftland	0.070	0.012	0.050	0.090	0.090	0.019	0.060	0.121	0.028	0.046	-0.044	0.107	0.004	0.107	-0.135	0.212
Sqfrbuild	0.477	0.029	0.427	0.525	0.731	0.040	0.665	0.796	0.516	0.075	0.390	0.637	0.329	0.215	-0.025	0.679
Numbdrooms	0.005	0.009	-0.010	0.020	0.014	0.011	-0.004	0.032	-0.012	0.016	-0.039	0.015	0.153	0.078	0.030	0.285
Age	-0.005	0.000	-0.005	-0.004	-0.002	0.000	-0.003	-0.002	-0.014	0.001	-0.016	-0.013	0.003	0.002	-0.001	0.006
Styheight	0.288	0.018	0.258	0.316	0.437	0.024	0.398	0.477	0.353	0.040	0.289	0.418	0.428	0.201	0.074	0.730
Age65	0.004	0.002	0.001	0.007	0.012	0.002	0.008	0.015	0.008	0.004	0.002	0.014	0.017	0.015	-0.007	0.043
White	0.003	0.001	0.002	0.003	0.004	0.001	0.003	0.005	0.008	0.001	0.007	0.009	0.001	0.003	-0.004	0.005
HSdegree	0.027	0.001	0.025	0.029	0.044	0.002	0.041	0.047	0.044	0.002	0.040	0.048	-0.003	0.009	-0.019	0.010
σ^2	0.266	0.006	0.257	0.276	0.172	0.011	0.154	0.191	0.456	0.022	0.420	0.492	0.684	0.181	0.456	1.045
π_g					0.520	0.021	0.485	0.555	0.418	0.023	0.381	0.456	0.062	0.009	0.049	0.077
ρ	0.543	0.010	0.525	0.559	0.182	0.016	0.156	0.207								

What have we learned?

Conclusions

- The S-M model provides a robust modeling framework that nests both heterogeneity and spatial dependence.
  - Heterogeneous coefficients and error terms are special cases nested within this model as well.
- Failing to account for heterogeneity across the sample population can lead to biased estimates of spatial dependence.
- Failing to account for spatial dependence leads to biased coefficients (as expected).

Where do we go from here?

Future Research

- Direct and indirect effects still need to be worked out for this model.
- The SAR is limited in its empirical application so the S-M model needs to be extended to include other spatial specifications.
- The model presented is merely a cross-sectional application and a time series approach still needs to be developed.

Thank you!