A Two Stage Approach to Spatiotemporal Analysis with Strong and Weak Cross-Sectional Dependence

Natalia Bailey[⋆], Sean Holly[†] and M. Hashem Pesaran[‡]

[⋆]Queen Mary, University of London
[†]Cambridge University

[‡]University of Southern California, and Trinity College, Cambridge

20th International Panel Data Conference

July 9-10, 2014

Cross Section Dependence - An Overview

- Growing literature on econometric methods for modelling and measuring cross section dependence in panel data.
- Researchers in many fields have turned to network theory, spatial and factor models to obtain a better understanding of the extent and nature of such cross dependencies.
- Some issues reflect:
  - Testing for the presence of cross-sectional dependence
  - Measuring the degree of cross-sectional dependence
  - Modelling cross-sectional dependence, and
  - Carrying out counter-factual exercises under alternative network formations or market inter-connections

This Paper's Contribution

- Assume a panel data set that contains both strong and weak cross dependence.

- In this paper we propose a two-stage estimation and inference strategy to spatiotemporal analysis:

  - **First step**: tests of cross-sectional dependence are applied to ascertain if the cross-sectional dependence is weak. If the null of weak cross-sectional dependence is rejected, the implied strong cross-sectional dependence is modelled by means of a (hierarchical) factor model.

  - **Second step**: residuals from such factor models are used to estimate possible connections amongst pairs of cross section units, and ultimately to model the remaining weak cross dependencies, making use of either extant techniques from spatial econometrics or more general spatial model specifications.

Modelling Cross-sectional Dependence

- Currently, there are two main approaches to modelling CD in large panels: spatial processes and factor structures.
- Spatial processes were pioneered by Whittle (1954) and developed further in econometrics by Anselin (1988), Kelejian and Prucha (1999), and Lee (2002), amongst others.
- Factor models were introduced by Hotelling (1933) and applied in economics first by Stone (1947, JRSS). They have been applied extensively:
  - in finance (Chamberlain and Rothschild 1983; Connor and Korajczyk, 1993; Stock and Watson, 1998; Kapetanios and Pesaran, 2007), and
  - in macroeconomics (Forni and Reichlin, 1998; Stock and Watson, 2002).

A Summary Statistic of the Degree of Cross-sectional Dependence

- The degree of cross-sectional dependence among N units, $\mathbf{x}_{\circ t} = (x_{1t}, ..., x_{Nt})'$, can be summarised conveniently by their average cross-correlation (excluding the diagonal elements),

$$\bar{\rho}_N = \frac{\boldsymbol{\tau}'\mathbf{R}\boldsymbol{\tau} - N}{N(N-1)} = \frac{\boldsymbol{\tau}'\mathbf{R}\boldsymbol{\tau}}{N(N-1)} - \frac{1}{N-1}, \tag{1}$$

  where $\boldsymbol{\tau}$ is an $N \times 1$ vector of ones and $\mathbf{R}$ is their corresponding the correlation matrix.

- Noting that $(\boldsymbol{\tau}'\boldsymbol{\tau})\,\lambda_{\min}(\mathbf{R}) \leq \boldsymbol{\tau}'\mathbf{R}\boldsymbol{\tau} \leq (\boldsymbol{\tau}'\boldsymbol{\tau})\,\lambda_{\max}(\mathbf{R})$, then

$$\frac{\lambda_{\min}(\mathbf{R}) - 1}{(N-1)} \leq \bar{\rho}_N \leq \frac{\lambda_{\max}(\mathbf{R}) - 1}{(N-1)}.$$

Spatial Econometric Models

- Consider the first-order spatial autoregressive, SAR(1), model

$$\mathbf{x}_{\circ t} = \psi \mathbf{W} \mathbf{x}_{\circ t} + \boldsymbol{\Sigma}_u^{1/2} \tilde{\mathbf{u}}_{\circ t},$$

  where $\mathbf{x}_{\circ t} = (x_{1t}, ..., x_{Nt})'$, $\tilde{\mathbf{u}}_{\circ t} = (\tilde{u}_{1t}, ..., \tilde{u}_{Nt})' \backsim (\mathbf{0}, \mathbf{I}_N)$, $\boldsymbol{\Sigma}_u$ is an $N \times N$ diagonal matrix with $\sigma^2_{u_i} < K < \infty$ on its i^{th} diagonal, and $\mathbf{W}$ is the spatial weight matrix.

- In the spatial literature, $\mathbf{W}$ is assumed to have non-negative elements and is typically row-standardized: $\|\mathbf{W}\|_\infty = 1$.

- Assuming that $(\mathbf{I}_N - \psi \mathbf{W})$ is invertible, then we have

$$\mathbf{x}_{\circ t} = (\mathbf{I}_N - \psi \mathbf{W})^{-1} \boldsymbol{\Sigma}_u^{1/2} \tilde{\mathbf{u}}_{\circ t} = \mathbf{G} \tilde{\mathbf{u}}_{\circ t}, \tag{2}$$

- If $\|\mathbf{W}\|_\infty = 1$, then $|\psi| < 1$ ensures that $|\psi| \|\mathbf{W}\|_\infty < 1$, and

$$\|\mathbf{G}\|_\infty < K < \infty; \;\; \|\mathbf{G}\|_1 < K < \infty.$$

Spatial Econometric Models

- Then the covariance matrix of (2), $\mathbf{\Sigma} = \mathbf{G}\mathbf{G}'$, will also be row (column) bounded:

$$\|\mathbf{\Sigma}\|_1 = \left\|\mathbf{G}\mathbf{G}'\right\|_1 \leq \|\mathbf{G}\|_1 \left\|\mathbf{G}'\right\|_1 = \|\mathbf{G}\|_1 \|\mathbf{G}\|_\infty < K < \infty.$$

- Similarly, assuming that $var(x_{it}) = \sigma_i^2 > 0$, for the correlation matrix of (2), $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$, where $\mathbf{D} = diag(\ \sigma_i^2,\ i = 1, 2, ..., N)$ we have

$$\|\mathbf{R}\|_1 = \left\|\mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}\right\|_1 \leq \frac{1}{\min_i(\sigma_i^2)} \|\mathbf{\Sigma}\|_1 < K < \infty. \qquad (3)$$

- Note that $\lambda_{\max}(\mathbf{R}) \leq \|\mathbf{R}\|_1$ controls for the degree of cross-sectional dependence.

- For weakly cross correlated processes, where $\lambda_{\max}(\mathbf{R})$ is bounded we have $\bar{\rho}_N \longrightarrow 0$, as $N \longrightarrow \infty$, and standard spatial econometric models cannot deal with cases where $\bar{\rho}_N$ differs from zero even for sufficiently large N.

The Factor Model

- Suppose instead that $\mathbf{x}_{\circ t}$ are generated according to the following factor model

$$\mathbf{x}_{\circ t} = \mathbf{\Gamma f}_t + \mathbf{\Omega}^{1/2}\tilde{\boldsymbol{\varepsilon}}_{\circ t}, \tag{4}$$

where $\mathbf{f}_t = (f_{1t}, f_{2t}, ..., f_{\ell t})'$ is the $\ell \times 1$ vector of unobserved common factors (ℓ being fixed) with $E(\mathbf{f}_t) = \mathbf{0}$, $\mathbf{\Sigma}_{ff} = Cov(\mathbf{f}_t) = \mathbf{I}_\ell$, and $\mathbf{\Gamma}$ is the $N \times \ell$ matrix of the factor loadings $\gamma_i = (\gamma_{i1}, \gamma_{i2}, ..., \gamma_{i\ell})$, $\tilde{\boldsymbol{\varepsilon}}_{\circ t} = (\tilde{\varepsilon}_{1t}, ..., \tilde{\varepsilon}_{Nt})'$, with $\tilde{\varepsilon}_{it} \sim IID(0, 1)$, $i = 1, ..., N$.

- Also, $\mathbf{\Omega} = Diag\left(\omega_i^2,\ i = 1, ..., N\right)$ and $\varepsilon_{it} = \omega_i \tilde{\varepsilon}_{it}$ so that $\varepsilon_{it} \sim IID(0, \omega_i^2)$, $i = 1, ..., N$.

- Again, for the degree of cross-sectional dependence of $\mathbf{x}_{\circ t}$ we look at the correlation matrix of (4), $\mathbf{R}$.

Degree of Cross-sectional Dependence

- For strongly cross correlated processes, such as the factor model, $\rho_{ij} = Corr(x_{it}, x_{jt}) = \delta_i \delta_j'$, for $i \neq j$, where

$$\delta_i = \frac{\gamma_i}{\sqrt{\omega_i^2 + \gamma_i \gamma_i'}}, \tag{5}$$

are the scaled factor loadings and $\delta_i = (\delta_{i1}, \delta_{i2}, ..., \delta_{i\ell})$. Then,

$$\bar{\rho}_N = \left(\frac{N}{N-1} \right) \left(\bar{\delta}_N' \bar{\delta}_N - \frac{\sum_{i=1}^N \delta_i \delta_i'}{N^2} \right), \tag{6}$$

where $\bar{\delta}_N = N^{-1} \sum_{i=1}^N \delta_i$ - Pesaran (2013, Econometrics Review - *forthcoming*).

Degree of Cross-sectional Dependence

- For the scaled loadings of the j^{th} factor we have:

$$\begin{aligned}\bar{\delta}_{j,N} &= \frac{1}{N}\left(\sum_{i=1}^{M_j}\delta_{ij}+\sum_{i=M_j+1}^{N}\delta_{ij}\right)=\frac{M_j}{N}\left(M_j^{-1}\sum_{i=1}^{M_j}\delta_{ij}\right)\\ &= N^{\alpha_j-1}\mu_j=O(N^{\alpha_j-1}),\end{aligned}$$

  where $\mu_j=\left(M_j^{-1}\sum_{i=1}^{M_j}\delta_{ij}\right)\neq 0$ and $M_j=N^{\alpha_j}$ are the number of non-zero factor loadings; $N^{-2}\sum_{i=1}^{N}\delta_{ij}^2=O(N^{\alpha_j-2})$.

- Exponent $\alpha_j\in[0,1]$ measures the degree of CSD due to the j^{th} factor where $\alpha_j=\frac{\ln(M_j)}{\ln(N)}$. The overall degree of CSD is given by $\alpha=\max_j(\alpha_j)$- Bailey, Kapetanios and Pesaran (2014).
- Using (6) we have

$$\bar{\rho}_N=O(N^{2\alpha-2}).$$

A Test for Weak Cross-Sectional Dependence

- Denote the pair-wise correlations of (i, j) units by

$$\hat{\rho}_{ij} = \hat{\rho}_{ji} = \frac{\sum_{t=1}^{T} \left(x_{it} - \bar{x}_i\right)\left(x_{jt} - \bar{x}_j\right)}{\left(\sum_{t=1}^{T} \left(x_{it} - \bar{x}_i\right)^2\right)^{1/2} \left(\sum_{t=1}^{T} \left(x_{jt} - \bar{x}_j\right)^2\right)^{1/2}}, \tag{7}$$

where $\bar{x}_i = N^{-1} \sum_{i=1}^{N} x_{it}$. The CD statistic is defined by

$$\begin{aligned} CD &= \left[\frac{TN(N-1)}{2}\right]^{1/2} \widehat{\bar{\rho}}_N, \\ \widehat{\bar{\rho}}_N &= \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \hat{\rho}_{ij}. \end{aligned} \tag{8}$$

- Pesaran (2013) shows that $CD \rightarrow N(0,1)$, under the null hypothesis that cross-sectional exponent of $\mathbf{x}_{\circ t}$, $t = 1, ..., T$, is $\alpha < (2 - \epsilon)/4$, as $N \rightarrow \infty$, such that $T = \kappa N^{\epsilon}$, for some $0 \leq \epsilon \leq 1$, and a finite $\kappa > 0$.

Evaluating the Degree of Cross-sectional Dependence

Suppose observations $\mathbf{x}_{\circ t} = (x_{1t}, ..., x_{Nt})'$, $t = 1, 2, ..., T$ are available and the aim is to model how x_{it} and x_{jt} are dependent, across all i and j, with N and T relatively large.

1. Apply the cross section dependence test developed in Pesaran (2013) to $\mathbf{x}_{\circ t}$, $t = 1, ..., T$, to find out if the observations are cross-sectionally weakly or strongly dependent.
    1. Only proceed to spatial modelling if the null of weak cross dependence is not rejected.
    2. If the null of weak dependence is rejected, model the (semi-) strong dependence by use of a (hierarchical) factor model, and check that the residuals from (4), denoted by $\hat{\boldsymbol{\varepsilon}}_{\circ t} = (\hat{\varepsilon}_{1t}, ..., \hat{\varepsilon}_{Nt})'$, are weakly cross-correlated (by applying the CD test to $\hat{\boldsymbol{\varepsilon}}_{\circ t}$, $t = 1, ..., T$).
2. Apply spatial or network modelling techniques to $\hat{\boldsymbol{\varepsilon}}_{\circ t}$ and/or identify local connections for the spatial weights matrix $\mathbf{W}$.

Choice of the Spatial Weight Matrix - using exogenous information

- Typically, the sparse spatial weight matrix, $\mathbf{W}$, is constructed using information brought in exogenously, such as geodesic, demographic or economic information, not contained in the data set under consideration, $\mathbf{x}_{\circ t}$, $t = 1, ..., T$.

- In economic applications, economic measures, such as commuting times, trade and migratory flows across geographical areas have been used. For example, trade weights are used in constructing cross-sectional averages used in GVAR modelling (Pesaran et al. 2004, JBES).

- Such measures are often preferable over the geodesic measures - since they are closer to the decisions that underlie the observations, x_{it}, and they allow for possible time variations in the weight matrix.

Correlation-based Specification of Spatial Weight Matrix

- In practice, it is often difficult to obtain appropriate measures of economic distance for the analysis of interdependencies.
- It is, therefore, desirable to see if **W** can be constructed without recourse to such exogenous information.
- In applications where the time dimension is reasonably large (around 50-80), it is possible to identify the non-zero elements of **W** with those elements of $\hat{\rho}_{ij}$ (marginal correlations) that are different from zero at a suitable significance level.
- But since there are a large number of such statistical tests, multiple testing procedures that control the overall size of the tests must be applied.

Multiple Testing Literature

- The multiple testing problem arises when we are faced with a number of (possibly) dependent tests and our aim is to control the overall size of the test.

- Suppose we are interested in a family of null hypotheses, $H_{01}, H_{02}, ..., H_{0m}$ and we are provided with corresponding test statistics, $Z_{1T}, Z_{2T}, ..., Z_{mT}$, with separate rejection rules given by (using a two sided alternative)

$$\Pr\left(|Z_{iT}| > CV_{iT} \,|H_{0i}\right) \leq p_{iT},$$

where CV_{iT} is some suitably chosen critical value of the test. p_{iT} is the observed p value for H_{0i}.

Family-wise Error Rate and Bonferroni's Bound

- Consider now the family-wise error rate (FWER) defined by

$$FWER_T = \Pr\left[\cup_{i=1}^{m}\left(|Z_{iT}| > CV_{iT}\,|H_{0i}\right)\right],$$

and suppose that we wish to control $FWER_T$ to lie below a pre-determined value, p.

- Bonferroni provides a general solution, which holds for all possible degrees of dependence across the separate tests. By Boole's inequality we have

$$\begin{aligned} FWER_T &= \Pr\left[\cup_{i=1}^{m}\left(|Z_{iT}| > CV_{iT}\,|H_{0i}\right)\right] \\ &\leq \sum_{i=1}^{m}\Pr\left(|Z_{iT}| > CV_{iT}\,|H_{0i}\right) \leq \sum_{i=1}^{m} p_{iT} \end{aligned}$$

- Hence to achieve $FWER_T \leq p$, it is sufficient if we set $p_{iT} \leq p/m$.

Holm's Procedure

- Bonferroni's procedure can be quite conservative, particularly when the tests are highly correlated.
- A step-down procedure is proposed by Holm (1979) which is more powerful than the Bonferroni's procedure, without imposing any further restrictions on the degree to which the underlying tests depend on each other.
- Abstract from the T subscript and order the p-values of the tests so that

$$p_{(1)} \leq p_{(2)} \leq .... \leq p_{(m)}.$$

  These are associated with the null hypotheses, $H_{(01)}, H_{(02)}, ..., H_{(0m)}$, respectively.
- Holm procedure rejects $H_{(01)}$ is $p_{(1)} \leq p/m$, rejects $H_{(01)}$ and $H_{(02)}$ if $p_{(2)} \leq p/(m-1)$, rejects $H_{(01)}, H_{(02)}$ and $H_{(03)}$ if $p_{(3)} \leq p/(m-2)$, and so on.

Holm Procedure on Full Correlation Matrix

- Under the null i and j are unconnected, and $\hat{\rho}_{ij}$ is approximately distributed as $N\left(0, T^{-1}\right)$.

- Thus, the p-values of the individual tests are (approximately) given by $p_{ij} = 2\left[1 - \Phi\left(\sqrt{T}\left|\hat{\rho}_{ij}\right|\right)\right]$ for $i = 1, 2, ..., N-1, j = i+1, ..., N$, with the total number of tests being carried out given by $m = N(N-1)/2$.

- We order these p-values in an ascending manner, or equivalently order $\left|\hat{\rho}_{ij}\right|$ in a descending manner. Denote the largest value of $\left|\hat{\rho}_{ij}\right|$ over all $i \neq j$, by $\left|\hat{\rho}_{(1)}\right|$, the second largest value by $\left|\hat{\rho}_{(2)}\right|$, and so on, to obtain the ordered sequence $\left|\hat{\rho}_{(s)}\right|$, for $s = 1, 2, ..., m$.

Correlation-based Connection Matrix

- Then the (i,j) pair associated with $\left|\hat{\rho}_{(s)}\right|$ are connected if $\left|\hat{\rho}_{(s)}\right| > T^{-1/2}\Phi^{-1}\left(1-\frac{p/2}{m-s+1}\right)$, otherwise disconnected, for $s = 1, 2, ..., m$.

- p is the pre-specified overall size of the test (set to 5% in the empirical application), and $\Phi^{-1}(.)$ is the inverse of the standard normal distribution function.

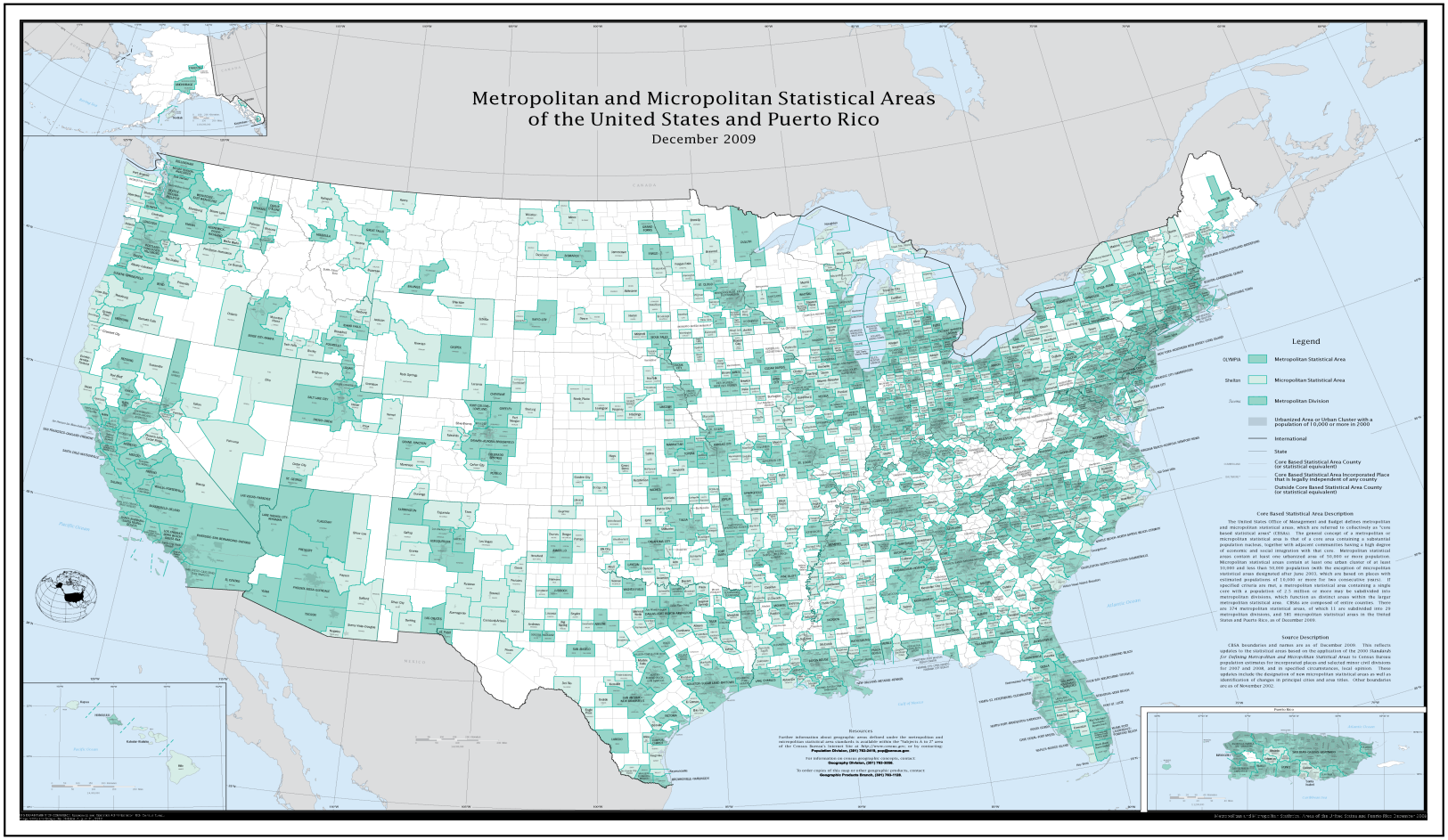
- The resultant connection matrix will be denoted by $\hat{\mathbf{W}} = \left(\hat{w}_{ij}\right)$, where $\hat{w}_{ij} = 1$ if the (i,j) pair are connected according to the Holm procedure, otherwise $\hat{w}_{ij} = 0$.

- Connections can also be classified as positive ($\hat{w}_{ij}^{+}$) if $\hat{\rho}_{ij} > 0$, and negative ($\hat{w}_{ij}^{-}$) if $\hat{\rho}_{ij} < 0$.

Application to US house prices

- We consider the existence of clusters of similar behaviour in US house prices at a Metropolitan Statistical Area (MSA) level.
- An MSA is defined by a core area with a large population concentration, together with adjacent areas that have a high degree of economic and social integration with that core through commuting and transport links.
- We consider a total of 363 MSAs, with the District of Columbia treated as a single MSA. We exclude 3 MSAs located in Alaska and Hawaii.
- The time dimension under consideration covers the period 1975Q1 to 2010Q4, with $T = 144$.

Metropolitan Statistical Areas (MSAs) in Medium Green



Spatiotemporal analysis of US real house price changes

- We denote house prices in the i^{th} MSA, located in state s, in quarter t by P_{ist}, for $i = 1, 2, ..., N_s$, $s = 1, 2, ..., S$, and $t = 1, 2, ..., T$ where $T = 144$ quarters (covering the period 1975Q1-2010Q4), $S = 49$ (comprised of 48 States and District of Columbia), and $\sum_{s=1}^{S} N_s = N = 363$.
- We then compute real house prices as:

$$p_{ist} = \ln\left(\frac{P_{ist}}{CPI_{st}}\right), \text{ for } i = 1, 2, ..., N_s;\ s = 1, ..., S;\ t = 1, 2, ..., T,$$

  where CPI_{st} is the Consumer Price Index for state s.

- We then regress $p_{ist} - p_{is,t-1}$, the rate of change of real house prices, on an intercept and three quarterly seasonal dummies to obtain the seasonally adjusted series π_{ist}, as residuals.

Evidence of Strong or Semi-strong Cross-sectional Dependence in Real House Price Changes

- As noted above, the **first step** in the analysis is to check for the presence of strong cross-sectional dependence in the house price data.
- Ignoring the State within which a particular MSA is located, we computed the CD statistic for the seasonally adjusted real house price changes, π_{it}, for $i = 1, 2..., N = 363$, and $t = 2, ..., 144$, and obtained $\mathbf{CD} = \mathbf{640.46}$ (as compared to the 5% critical value of 1.96).
- The estimate of the exponent of cross-sectional dependence amounted to $\mathring{\boldsymbol{\alpha}}_{\hat{\pi}} = \mathbf{0.989}$ (0.03).
- The CD test clearly rejects the null of weak cross-sectional dependence.
- It is clearly inappropriate to apply our method of estimating spatial matrix $\mathbf{W}$ and spatial modelling using π_{it}, without eliminating the effects of the strong dependence.

Modelling Strong Dependence of Real House Price Changes

- The effects of strong cross-sectional dependence in real house prices changes can be modelled by using observed (national/regional income and interest rates), or unobserved common factors (using principal components).
- Alternatively, as argued in Pesaran (2006), we could use cross-sectional averages at the national and regional levels. We also considered using State level averages, but there were only a few MSAs in some States.
- The analysis can be conducted using the Principal Components approach either applied to the full data set or at a regional level.

A Hierarchical Spatiotemporal Model of HP Changes

- More specifically, we run the following regressions

$$\boldsymbol{\pi}_t = \mathbf{a} + \mathbf{B}\mathbf{R}_N\boldsymbol{\pi}_t + \boldsymbol{\Gamma}\mathbf{P}_N\boldsymbol{\pi}_t + \boldsymbol{\xi}_t, \tag{9}$$

  where $\boldsymbol{\pi}_t$ is an $N \times 1$ vector of (real) house price changes, partitioned by regions (from 1 to R). The same holds for the intercepts $\mathbf{a}$.

- $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are an $N \times N$ diagonal matrices with ordered elements in line with those of $\boldsymbol{\pi}_t$
- $\mathbf{R}_N$ and $\mathbf{P}_N$ are $N \times N$ projection matrices. $\mathbf{R}_N\boldsymbol{\pi}_t$ give the **regional** means and $\mathbf{P}_N\boldsymbol{\pi}_t$ the **national** mean of house price changes. For $\boldsymbol{\tau}_N$ and $\boldsymbol{\tau}_{N_r}$ an $N \times 1$ and $N_r \times 1$ vector of ones,

$$\mathbf{P}_N = \boldsymbol{\tau}_N(\boldsymbol{\tau}_N'\boldsymbol{\tau}_N)^{-1}\boldsymbol{\tau}_N' \text{ and } \mathbf{R}_N = Diag\left(\mathbf{P}_{N_r},\ r = 1, ..., R\right),$$

  where $\mathbf{P}_{N_r} = \boldsymbol{\tau}_{N_r}(\boldsymbol{\tau}_{N_r}'\boldsymbol{\tau}_{N_r})^{-1}\boldsymbol{\tau}_{N_r}'$.

- R is assumed fixed, and for each r, $N_r/N \to K > 0$, $N \to \infty$. We identified a total of $R = 8$ regions in the US.

De-factoring House Price Changes using Cross Section Averages

- The **second** stage of our analysis is based on de-factored real house price changes, given by residuals from (9), namely

$$\hat{\boldsymbol{\xi}}_t = \left(\mathbf{I}_N - \hat{\mathbf{B}}\mathbf{R}_N - \hat{\boldsymbol{\Gamma}}\mathbf{P}_N\right)\boldsymbol{\pi}_t - \hat{\mathbf{a}}, \text{ for } t = 2, ..., T.$$

- Application of the CD test to these residuals resulted in the much reduced **CD** statistic of $\mathbf{-6.05}$ (as compared to 640.46 when the test was applied to price changes without de-factoring).
- We then computed the exponent of cross-sectional dependence of $\hat{\boldsymbol{\xi}}_t = \left(\hat{\xi}_{it}\right)$, for $i = 1, ..., N$ which amounted to $\mathring{\alpha}_{\hat{\xi}} = \mathbf{0.637}$ (0.03).
- The simple hierarchical de-factoring procedure has managed to eliminate almost all of the strong cross-sectional dependence that had existed in the house price changes.
- What remains could be due to the local dependencies that need to be modelled using spatial techniques.

Estimating Spatial Connections

- We apply Holm's multiple testing to the $N(N-1)/2$ pair-wise correlation coefficients, $\hat{\rho}_{\hat{\xi},ij} = \hat{\sigma}_{\hat{\xi},ij} / \sqrt{\hat{\sigma}_{\hat{\xi},ii}\hat{\sigma}_{\hat{\xi},jj}}$,

  $$\hat{\sigma}_{\hat{\xi},ij} = T^{-1}\sum_{t=1}^{T}\hat{\xi}_{it}\hat{\xi}_{jt}, \text{ for } i = 1,2,...,N-1, j = i+1,...,N.$$

- We denote the resultant connection matrix by $\hat{\mathbf{W}}_{cs} = \left(\hat{w}_{cs_{ij}}\right)$. Here *cs* stands for multiple testing applied to residuals extracted from de-factoring using the cross-sectional averages approach.

- We obtain 1.08% of non-zero elements for $\hat{\mathbf{W}}_{cs}$ but the pattern of these is of more interest.

- It is best to view the non-zero elements of $\hat{\mathbf{W}}_{cs}$ as connections rather than as neighbours (in a physical sense).

- According to $\hat{\mathbf{W}}_{cs}$, the connections extend well beyond geographical boundaries.

Positive and Negative Connections

- Next, we separate out the **positive** and **negative** connections by creating network matrices $\hat{\mathbf{W}}_{cs}^{+} = \left(\hat{w}_{cs_{ij}}^{+} \right)$ and $\hat{\mathbf{W}}_{cs}^{-} = \left(\hat{w}_{cs_{ij}}^{-} \right)$ having elements

$$\hat{w}_{cs_{ij}}^{+} = \hat{w}_{cs_{ij}} I\left(\hat{\rho}_{\hat{\xi},ij} > 0 \right)$$

and

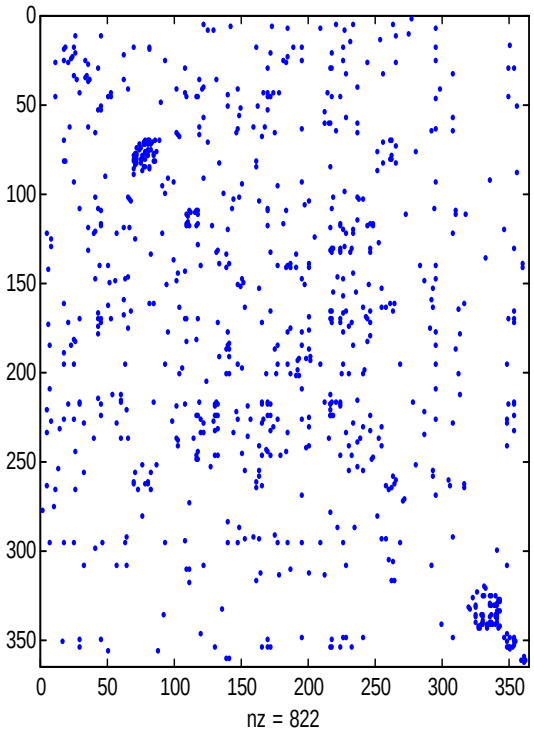
$$\hat{w}_{cs_{ij}}^{-} = \hat{w}_{cs_{ij}} I\left(\hat{\rho}_{\hat{\xi},ij} < 0 \right).$$

- Note that $\hat{\mathbf{W}}_{cs} = \hat{\mathbf{W}}_{cs}^{+} + \hat{\mathbf{W}}_{cs}^{-}$.

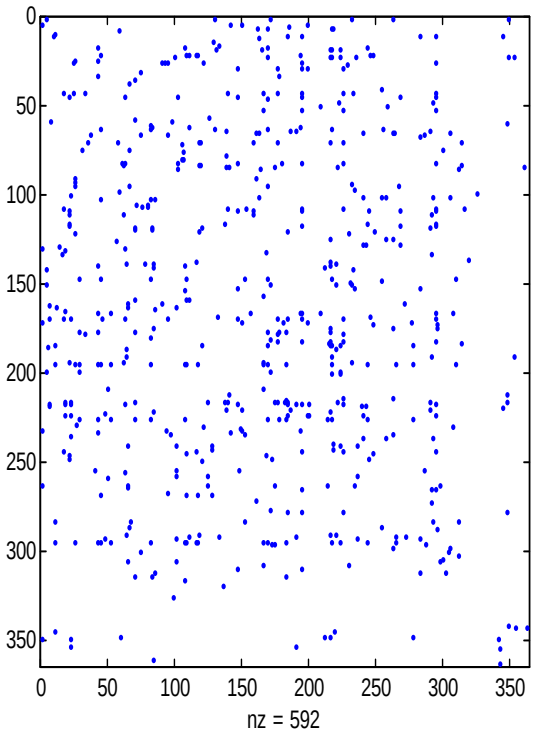
- It is evident that geographical proximity is not the only factor driving spatial connections between MSAs. There are significant correlations (positive or negative) well away from the diagonal.

Spatial Weights matrices: Distance- and Correlation-based Connections

$\hat{\mathbf{W}}_{cs}^{+}$



$\hat{\mathbf{W}}_{cs}^{-}$



De-Factoring House Price Changes using regional PC's

- Alternatively we can repeat estimation of regression (9) where the national and regional cross-sectional averages are now replaced by the strongest PC of the full data set and the strongest two PCs in each of the 8 regions identified:

$$\begin{aligned} \pi_{irt} &= a_{ir} + \boldsymbol{\beta}_{ir}' \hat{\mathbf{f}}_{rt} + \gamma_{ir} \hat{f}_{gt} + \xi_{irt}, \\ i &= 1, 2, ..., N_r; \; r = 1, 2, ..., R; \; t = 2, ..., T, \end{aligned} \tag{10}$$

  where $\hat{\mathbf{f}}_{rt}$ is an $\ell_r \times 1$ vector of regional principal components of house price changes for $r = 1, ..., R$ and $\boldsymbol{\beta}_{ir} = \left(\beta_{i1}, \beta_{i2}, ..., \beta_{i\ell_r}\right)'$ is the associated vector of factor loadings.

- $\hat{f}_{gt}$ is the 'global' or 'national' principal component with coefficients γ_{ir} representing the factor loadings.

- Note that Bai and Ng (2002) test gave little guidance as to the number of PCs we should use.

- We repeated the analysis by de-factoring US house price changes using these Principal Components (PC)

Closeness of correlation-based W_{cs}^{+} and W_{pc}^{+}

- The CD statistic of the de-factored residuals from regressions (10) amounted to a value of 3.320. This is slightly higher than the value recorded using the CSA but again markedly reduced from the value of the statistic attached to the house price data in their original form.
- The corresponding exponent of cross-sectional dependence was $\mathring{\alpha}_{pc} = 0.773\ (0.03)$. This α estimate resides in the region of $\alpha = 3/4$ which in line with $\bar{\rho}_N \longrightarrow 0$ at a rate of $\frac{1}{\sqrt{N}}$ compared with $\bar{\rho}_N$ tending to zero at the faster rate of $\frac{1}{N}$ when α is recorded around $1/2$.
- As before we constructed $\hat{\mathbf{W}}_{pc}^{+}$ and $\hat{\mathbf{W}}_{pc}^{-}$ network matrices.
- These are similar to those produced using Cross Section Averages approach.

A Heterogeneous Spatio-temporal Model of US HP Changes

- We are now in a position to illustrate the utility of separate identification of positive and negative connections for the spatial analysis of house price changes.
- The de-factored house price changes, $\hat{\tilde{\xi}}_{it}$, can be modelled using the following spatiotemporal model

$$\hat{\tilde{\xi}}_{it} = a_{i\xi} + \sum_{j=1}^{h_{\lambda i}} \lambda_{ij} \hat{\tilde{\xi}}_{i,t-j} + \sum_{j=0}^{h_{\psi i}} \psi_{ij} \hat{\tilde{\xi}}^{*}_{i,t-j} + \zeta_{it}, \tag{11}$$

for $i = 1, 2, ..., N$, $t = 2, ..., T$, where

$$\begin{aligned} \hat{\tilde{\xi}}^{*}_{it} &= \frac{\mathbf{w}_i \hat{\tilde{\boldsymbol{\xi}}}_t}{\mathbf{w}_i \boldsymbol{\tau}_N}, \text{ if } \mathbf{w}_i \boldsymbol{\tau}_N > 0, \\ &= 0 \text{ if } \mathbf{w}_i \boldsymbol{\tau}_N = 0, \end{aligned}$$

and $\mathbf{w}_i$ denotes the i^{th} row of the $N \times N$ spatial matrix $\mathbf{W}$, which we take as given.

A Heterogeneous Spatio-temporal Model of US HP Changes

- Writing the above model in matrix notation we have

$$\hat{\boldsymbol{\xi}}_t = \mathbf{a}_\xi + \sum_{j=1}^{h_\lambda} \boldsymbol{\Lambda}_j \hat{\boldsymbol{\xi}}_{t-j} + \sum_{j=0}^{h_\psi} \boldsymbol{\Psi}_j \mathbf{W} \hat{\boldsymbol{\xi}}_{t-j} + \boldsymbol{\zeta}_t, \tag{12}$$

  where $h_\lambda = \max(h_{\lambda 1}, h_{\lambda 2}, ..., h_{\lambda N})$, $h_\psi = \max(h_{\psi 1}, h_{\psi 2}, ..., h_{\psi N})$, $\boldsymbol{\Lambda}_j$ and $\boldsymbol{\Psi}_j$ are $N \times N$ diagonal matrices with λ_{ij} and ψ_{ij} over i as their diagonal elements, and $\boldsymbol{\zeta}_t = (\zeta_{1t}, \zeta_{2t}, ..., \zeta_{Nt})'$.
- This model provides a generalisation of the spatiotemporal models analysed in the literature.
- The slope coefficients, λ_{ij} and ψ_{ij}, and the error variances, $\sigma^2_{\zeta_i} = var(\zeta_{it})$ are allowed to differ across i - Aquaro, Bailey and Pesaran (2013).

Model accommodating positive and negative connections

- We accommodate negative and positive connections, (here set h_λ, h_ψ^+ and h_ψ^- equal to unity), so that (12) becomes

$$\hat{\boldsymbol{\xi}}_t = \begin{cases} \mathbf{a}_\xi + \boldsymbol{\Lambda}_1 \hat{\boldsymbol{\xi}}_{t-1} + \boldsymbol{\Psi}_0^+ \tilde{\mathbf{W}}_{cs}^+ \hat{\boldsymbol{\xi}}_t + \boldsymbol{\Psi}_0^- \tilde{\mathbf{W}}_{cs}^- \hat{\boldsymbol{\xi}}_t \\ \quad + \boldsymbol{\Psi}_1^+ \tilde{\mathbf{W}}_{cs}^+ \hat{\boldsymbol{\xi}}_{t-1} + \boldsymbol{\Psi}_1^- \tilde{\mathbf{W}}_{cs}^- \hat{\boldsymbol{\xi}}_{t-1} + \boldsymbol{\zeta}_t \end{cases}, \quad t = 3, ..., 144. \tag{13}$$

  where $\tilde{\mathbf{W}}_{cs}^+$ and $\tilde{\mathbf{W}}_{cs}^-$ are the $N \times N$ scaled (row-standardised when applicable) versions of $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$.

- Here $\boldsymbol{\Lambda}_1 = diag\left(\boldsymbol{\lambda}_1\right)$, $\boldsymbol{\Psi}_0^+ = diag\left(\boldsymbol{\psi}_0^+\right)$, $\boldsymbol{\Psi}_0^- = diag\left(\boldsymbol{\psi}_0^-\right)$, $\boldsymbol{\Psi}_1^+ = diag\left(\boldsymbol{\psi}_1^+\right)$, and $\boldsymbol{\Psi}_1^- = diag\left(\boldsymbol{\psi}_1^-\right)$.

- Also, $\boldsymbol{\lambda}_1$, $\boldsymbol{\psi}_0^+$, $\boldsymbol{\psi}_0^-$, $\boldsymbol{\psi}_1^+$ and $\boldsymbol{\psi}_1^-$ are $N \times 1$ vectors of parameters for the $N = 363$ MSAs.

- Finally, for quasi maximum likelihood (QML) estimation of the parameters we assume that $\zeta_{it} \sim IIDN\left(0, \sigma_{\zeta_i}^2\right)$, for $i = 1, ..., N$.

QML Summary Estimates

Table: Quasi-ML estimates of spatiotemporal model (19)

Applied to residual house price changes of 363 MSAs in the United States

	λ_1	ψ_0^+	ψ_0^-	ψ_1^+	ψ_1^-	σ_ζ
	Computed over non-zero parameter coefficients					
Median	0.3986	0.3124	-0.2493	-0.0430	0.0608	1.2416
Mean Group Estimates	0.3921	0.3454	-0.2763	-0.0398	0.0706	1.3056
	(0.0086)	(0.0168)	(0.0209)	(0.0147)	(0.0156)	(0.0181)
% significant (at 5% level)	89.8%	64.8%	61.9%	28.1%	26.4%	-
Number of non-zero coef.	363	253	197	253	197	363

[1] *Restricted parameter coefficients are set to zero.* $\hat{\psi}_{i0}^+ = 0$ *and* $\hat{\psi}_{i1}^+ = 0$ *if MSA* i *has no positive connections;* $\hat{\psi}_{i0}^- = 0$ *and* $\hat{\psi}_{i1}^- = 0$ *if MSA* i *has no negative connections;* $\hat{\psi}_{i0}^+ = 0$, $\hat{\psi}_{i1}^+ = 0$, $\hat{\psi}_{i0}^- = 0$ *and* $\hat{\psi}_{i1}^- = 0$ *if MSA* i *has no positive or negative connections, for* $i = 1, ..., 363$.

[2] *MGE standard errors are in brackets.*

Concluding Remarks

- This paper provides a general approach to spatiotemporal modelling in the case of large spatial data sets observed over a relatively long time period.

- The paper:
  - Highlights the importance of distinguishing between weak and strong cross sectional dependence in modelling of spatial effects.
  - Proposes a correlation-based measure of spatial weights matrix.
  - Implements a heterogeneous spatiotemporal model specification.
  - Provides a detailed application to the analysis of house price diffusion across 363 MSAs in the US over the 1975-2010 period.

The starting point was data on the Latitude-Longitude of zip codes cross referenced to MSAs.
Missing LL coordinates were coded manually from Google searches.
Haversine formula for calculating the geodesic distance between a pair of latitude/longitude coordinates:

$$a = sin^2(\Delta lat/2) + cos(lat1).cos(lat2).sin^2(\Delta long/2)$$

$$c = 2.atan2(\sqrt{a}, \sqrt{1-a})$$

$$d = R.c$$

where R is the radius of the earth in miles and d is the distance.
$\Delta lat = lat2 - lat1$, $\Delta long = long2 - long1$.

Data Sources

Monthly data for house prices from January 1975 to December 2010 taken from Freddy Mac.
These data are available at http://www.freddiemac.com/finance/cmhpi.
The quarterly figures are the arithmetic average of the monthly figures.

State level consumer prices are taken from the Bureau of Labor Statistics.

Distribution of MSAs by Connections

Table: Distribution of MSAs by connections across 8 regions in the US

Region\No. of MSAs	$N_{+/-}$	N_{-}	N_{+}	N_0	$\sum_{row}$
New England	9	1	1	4	15
Mid East	17	2	9	8	36
South East	63	10	25	16	114
Great Lakes	28	8	13	12	61
Plains	16	5	8	3	32
South West	14	3	7	14	38
Rocky Mountains	7	3	3	9	22
Far West	9	2	24	10	45
$\sum_{col}$	163	34	90	76	363
	Proportional to total no. of MSAs per region				
New England	60.0%	6.7%	6.7%	26.7%	100.0%
Mid East	47.2%	5.6%	25.0%	22.2%	100.0%
South East	55.3%	8.8%	21.9%	14.0%	100.0%
Great Lakes	45.9%	13.1%	21.3%	19.7%	100.0%
Plains	50.0%	15.6%	25.0%	9.4%	100.0%
South West	36.8%	7.9%	18.4%	36.8%	100.0%
Rocky Mountains	31.8%	13.6%	13.6%	40.9%	100.0%
Far West	20.0%	4.4%	53.3%	22.2%	100.0%

$N_{+/-}$ denotes the number of MSAs with both positive and negative connections; N_{-} the no. of MSAs with only negative connections; N_{+} the no. of MSAs with only positive connections; finally N_0 the no. of MSAs with no connections.