

A Two Stage Approach to Spatio-Temporal Analysis with Strong and Weak Cross-Sectional Dependence*

Natalia Bailey

Queen Mary, University of London

Sean Holly

University of Cambridge

M. Hashem Pesaran

University of Southern California and Trinity College, Cambridge

Revised July 2014

Abstract

An understanding of the spatial dimension of economic and social activity requires methods that can separate out the relationship between spatial units that is due to the effect of common factors from that which is purely spatial even in an abstract sense. The same applies to the empirical analysis of networks in general. We use cross unit averages to extract common factors (viewed as a source of strong cross-sectional dependence) and compare the results with the principal components approach widely used in the literature. We then apply multiple testing procedures to the de-factored observations in order to determine significant bilateral correlations (signifying connections) between spatial units and compare this to an approach that just uses distance to determine units that are neighbours. We apply these methods to real house price changes at the level of Metropolitan Statistical Areas in the USA, and estimate a heterogeneous spatio-temporal model for the de-factored real house price changes and obtain significant evidence of spatial connections, both positive and negative.

Keywords: Spatial and factor dependence, spatio-temporal models, positive and negative connections, house price changes.

JEL Classification: C21, C23

*The authors would like to thank Alex Chudik, George Kapetanios, Ron Smith, Vanessa Smith, Jushan Bai (guest editor), two anonymous referees and participants at the following conferences for valuable comments and suggestions: 2012 SEA conference, Salvador, Brazil, 2013 conference on Cross-Sectional Dependence in Panel Data Models, Cambridge, UK, 2013 Econometrics of Social Interaction Symposium, York, UK, 2013 SEA conference, Washington DC, US, 1st IAAE Annual Conference, June 2014, Queen Mary University of London, UK, and 20th International Panel Data Conference, July 2014, Tokyo, Japan. The authors also wish to acknowledge financial support under ESRC Grant ES/I031626/1.

1 Introduction

The nature and degree of spatial dependence in economic, geographical, epidemiological and ecological systems has long been the focus of intensive study. Geographers regard the fundamental question in economic geography to be what explains the uneven pattern of economic activity in space. Indeed the New Economic Geography starting with Krugman (1991) addresses exactly this question. But where we have a data rich environment with observations on many spatial units over many time periods there may be obstacles to understanding these uneven patterns in spatial data because of complex dependencies between spatial units that reflect both local (clustering) and common factors. Recent developments in spatial econometrics have generated a growing literature on methods for modelling and measuring spatial or cross-sectional dependence in data sets with a panel structure where there are observations over time (T) and over space (N). This in turn has identified a number of central research questions. What is the source of dependencies in space? To what extent are the observed dependencies between different spatial units due to common factors - for example, aggregate shocks - that affect different units rather than being the result of local interactions that generate spatial spill-over effects? Is the implementation of estimation procedures of panels that implicitly assume spatially correlated units justifiable when the degree of their cross dependence has not been established? Do existing methods of identifying neighbouring relationships fully reflect the spatial structure of the underlying data studied?

To answer these and other related questions two basic approaches have been developed in the literature, namely spatial and common factor models. 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 first applied in economics by Stone (1947), and have been developed further to deal with data sets where the cross section and time series dimensions are both relatively large, by Forni and Lippi (2001), Forni, Hallin, Lippi, and Reichlin (2000), and Stock and Watson (1998). These methods have been applied extensively in finance and macroeconomics, notably by Chamberlain and Rothschild (1983), Connor and Korajczyk (1993), Forni and Reichlin (1998), Stock and Watson (2002a,b), and Kapetanios and Pesaran (2007).

Factors can be represented by cross-sectional averages at regional and/or national levels - Pesaran (2006), or can be estimated by Principal Components (PCs). The number of principal components can be determined, for example, using the various information criteria proposed by Bai and Ng (2002), amongst others. Estimation of panels with spatially correlated errors include the use of parametric methods based on maximum likelihood - Lee (2004), Yu, de Jong and Lee (2008), Lee and Yu (2010), or the GMM approach proposed by Kelejian and Prucha (1999, 2010), Kapoor, Kelejian and Prucha (2007), and Lin and Lee (2010). Furthermore, non-parametric methods using spatial HAC estimators have been applied by Conley (1999), Kelejian and Prucha (2007), and Bester, Conley and Hansen (2011). Chudik and Pesaran (2014) provide a review of recent literature on estimation and inference in large panel data models with cross-sectional dependence.

The factor and spatial econometric approaches tend to complement one another, with the factor approach more suited to modelling strong cross-sectional dependence, whilst the spatial approach generally requires the spatial dependence to be weak. See, for ex-

ample, Chudik, Pesaran and Tosetti (2011). This presents a challenge as most panel data sets are subject to a combination of strong and weak cross dependencies, and a methodology that is capable of identifying and dealing with both forms of cross dependence is needed. This paper proposes a two-stage estimation and inference strategy, whereby in the 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 factor model. Residuals from such factor models, referred to as de-factored observations, are then used to estimate possible connections amongst pairs of cross section units, and ultimately to model the remaining weak cross dependencies, making use of extant techniques from spatial econometrics or generalisations thereof.

In our application of spatial econometric techniques, in addition to using standard spatial weights matrices based on contiguity and geodesic distance, we also consider the use of pair-wise correlations of the de-factored observations to identify if a given pair of cross section units is connected by testing whether the associated pair-wise correlation is non-zero. To avoid the multiple testing problem that such an approach entails we employ Bonferroni (1935) and Holm (1979) procedures discussed in Bailey, Pesaran and Smith (2014). Finally, following Aquaro, Bailey and Pesaran (2014), we consider a generalisation of the traditional spatial autoregressive model for large panel data sets that allows the spatial parameters to vary over the cross section units.

The paper also provides a detailed application of the proposed two-step methodology to the analysis of real house price changes across different Metropolitan Statistical Areas (MSAs) in the US. We consider de-factoring of the observations using Principal Components applied to the full data set as well as by regions, and compare the results to de-factoring using simple national and regional cross-sectional averages. The de-factored observations are then used to estimate the patterns of connections across MSAs by the Holm multiple testing procedure applied to the $N(N - 1)/2$ pair-wise correlations, distinguishing between positively related $N \times N$ connection (weights) matrix, $\hat{\mathbf{W}}^+$, and the negatively related connection matrix, $\hat{\mathbf{W}}^-$. These positive and negative connection matrices are then compared to geodesic based spatial matrices, \mathbf{W}_d , (d being the selected distance measure between different MSAs) and their closeness examined by means of contingency tables. A heterogeneous spatio-temporal model of de-factored house price changes is then estimated by the quasi maximum likelihood (QML). The results confirm important dynamics in the de-factored house price changes as well as statistically significant positive and negative spill-over effects across the MSAs, with the positive effects being more prevalent.

The rest of the paper is organised as follows: Section 2 motivates and describes the first stage of the proposed two-step spatio-temporal modelling strategy. Section 3 focuses on the second stage of the proposed approach and suggests first a correlation based method for approximating network connections using de-factored observations from the first stage, and second introduces a heterogeneous version of the traditional spatial econometrics model. Section 4 presents the empirical application to the US real house price changes, with some concluding remarks provided in Section 5. Data specifications and sources are relegated to the Appendices.

Notation: The largest and the smallest eigenvalues of the $N \times N$ matrix $\mathbf{A} = (a_{ij})$, are denoted by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$ is the spectral (or operator) norm of \mathbf{A} , $\|\mathbf{A}\|_1 = \max_{1 \leq j \leq N} \left\{ \sum_{i=1}^N |a_{ij}| \right\}$ is its maximum absolute column

sum norm, and $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq N} \left\{ \sum_{j=1}^N |a_{ij}| \right\}$ is its maximum absolute row sum norm. K is used to represent a finite generic positive constant.

2 Cross-sectional dependence (CSD) in panels

Cross-sectional dependence may be due to common effects coming from aggregate factors which are pervasive by nature, and/or can result from spatial interactions between units which are localised in character. In order to build a strategy for distinguishing between the two types of cross dependence it is useful first to analyse their properties with the help of two established and widely used econometric models: spatial and factor models.

2.1 Spatial dependence - a form of weak CSD

The standard spatial econometric model with a homogeneous spatial autoregressive coefficient, ψ , can be written as

$$x_{it} = \psi x_{it}^* + u_{it}, \text{ for } i = 1, 2, \dots, N, \quad (1)$$

where $x_{it}^* = \mathbf{w}_i' \mathbf{x}_{ot}$, $\mathbf{x}_{ot} = (x_{1t}, x_{2t}, \dots, x_{Nt})'$, $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{iN})'$ is the $N \times 1$ vector of fixed weights attached to the neighbours of the i^{th} unit, and u_{it} is the idiosyncratic component of x_{it} which is assumed to be serially and cross-sectionally independently distributed, with zero means and variances $var(u_{it}) = \sigma_{u_i}^2, 0 < \sigma_{u_i}^2 < K < \infty$. Writing the above N equations in matrix notation we have

$$\mathbf{x}_{ot} = \psi \mathbf{W} \mathbf{x}_{ot} + \mathbf{u}_{ot}, \quad (2)$$

where $\mathbf{u}_{ot} = (u_{1t}, u_{2t}, \dots, u_{Nt})'$, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)'$, and $var(\mathbf{u}_{ot}) = \mathbf{\Sigma}_u = diag(\sigma_{u_i}^2, i = 1, 2, \dots, N)$. This model is commonly known as a first-order spatial autoregression, or SAR(1) for short.

The $N \times N$ weight matrix, \mathbf{W} , is typically sparse, with its non-zero elements set *a priori*, using physical or economic distance. In practice \mathbf{W} is row-standardised so that $\mathbf{W}\mathbf{\tau} = \mathbf{\tau}$, where $\mathbf{\tau}$ is a $N \times 1$ vector of ones. A review of spatial econometrics literature is provided by Anselin (2001), and more recently by LeSage and Pace (2010). Many different choices of \mathbf{W} are considered in the literature. Physical and economic distances have been used (Conley and Dupor (2003), Conley and Topa (2003), Pesaran, Schuermann and Weiner (2004)). In some instances trade flows might be relevant, whilst in the case of inter-industry dependencies input-output matrices might provide the appropriate ‘spatial’ metric - Holly and Petrella (2012). Alternatively, there may be dependencies between geographical areas that reflect cultural similarity, and migration or commuting relationships.¹ Using one of these distance metrics the \mathbf{W} matrix is then constructed *a priori* and equation (2) is estimated directly.²

Irrespective of the distance measure used, spatial dependence relates to spill-over effects that are not pervasive in nature. In other words it conforms to the notion of cross-sectional weak dependence (CWD) as defined in Chudik, Pesaran and Tosetti (2011).

¹Interactions in social networks can also be ‘spatial’ in an abstract sense. For example Bhattacharjee and Holly (2013) explore interactions among members of a committee using a spatial analogy.

²The regional science literature has long been aware of the potential problems with the prior specification of the \mathbf{W} matrix. For recent contributions see Corrado and Fingleton (2012).

To see why, consider the SAR(1) model defined in (2). Assuming that $(\mathbf{I}_N - \psi \mathbf{W})$ is invertible, we have

$$\mathbf{x}_{ot} = \mathbf{G} \mathbf{u}_{ot}, \quad (3)$$

where $\mathbf{G} = (\mathbf{I}_N - \psi \mathbf{W})^{-1}$. It is now easily seen that $Var(\mathbf{x}_{ot}) = \mathbf{\Sigma} = \mathbf{G} \mathbf{\Sigma}_u \mathbf{G}'$, and we have

$$\begin{aligned} \|\mathbf{\Sigma}\|_1 &= \|\mathbf{G} \mathbf{\Sigma}_u \mathbf{G}'\|_1 \leq \|\mathbf{\Sigma}_u\| \|\mathbf{G}\|_1 \|\mathbf{G}'\|_1 \leq \sup_i (\sigma_{u_i}^2) \|\mathbf{G}\|_1 \|\mathbf{G}\|_\infty \\ &\leq K \|\mathbf{G}\|_1 \|\mathbf{G}\|_\infty. \end{aligned} \quad (4)$$

Suppose that $|\psi| \|\mathbf{W}\|_\infty < 1$, then

$$\begin{aligned} \|\mathbf{G}\|_\infty &= \|\mathbf{I}_N + \psi \mathbf{W} + \psi^2 \mathbf{W}^2 + \dots\|_\infty \\ &\leq (1 + |\psi| \|\mathbf{W}\|_\infty + |\psi|^2 \|\mathbf{W}\|_\infty^2 + \dots) = \frac{1}{1 - |\psi| \|\mathbf{W}\|_\infty} < K. \end{aligned}$$

Similarly, $\|\mathbf{G}\|_1 < K < \infty$, if it is further assumed that $|\psi| \|\mathbf{W}\|_1 < 1$. Therefore, \mathbf{G} has bounded row and column sum matrix norms if $|\psi| < \min(1/\|\mathbf{W}\|_1, 1/\|\mathbf{W}\|_\infty)$, and under the same condition, using (4), it follows that $\mathbf{\Sigma}$ will also be row (column) bounded.

Denoting the correlation matrix of \mathbf{x}_{ot} by $\mathbf{R} = (\rho_{ij})$ and assuming that $var(x_{it}) = \sigma_{x_i}^2 > 0$ is bounded away from zero, then $\mathbf{R} = \mathbf{D}^{-1/2} \mathbf{\Sigma} \mathbf{D}^{-1/2}$, where $\mathbf{D} = diag(\sigma_{x_i}^2, i = 1, 2, \dots, N)$. Hence

$$\|\mathbf{R}\|_1 = \|\mathbf{D}^{-1/2} \mathbf{\Sigma} \mathbf{D}^{-1/2}\|_1 \leq \frac{1}{\min_i (\sigma_{x_i}^2)} \|\mathbf{\Sigma}\|_1 < K. \quad (5)$$

Also, $\lambda_{\max}(\mathbf{R}) \leq \|\mathbf{R}\|_1 < K$. Therefore, $\lambda_{\max}(\mathbf{R})$ controls the degree of cross-sectional dependence and is bounded in the case of SAR(1) models so long as $|\psi| < \min(1/\|\mathbf{W}\|_1, 1/\|\mathbf{W}\|_\infty)$, a condition generally assumed in the spatial econometrics literature. This result readily extends to higher order SAR models.

2.2 The factor model - a form of strong CSD

At other end of the spectrum we consider the factor model as a form of strong cross-sectional dependence. To this end we draw from the analysis in Pesaran (2014) which tests for weak cross-sectional dependence. Suppose that \mathbf{x}_{ot} is generated according to the following m -factor model

$$\mathbf{x}_{ot} = \mathbf{\Gamma} \mathbf{f}_t + \boldsymbol{\varepsilon}_{ot}, \quad (6)$$

where $\mathbf{f}_t = (f_{1t}, f_{2t}, \dots, f_{mt})'$ is the $m \times 1$ vector of unobserved common factors (m being fixed) with $E(\mathbf{f}_t) = \mathbf{0}$, $\mathbf{\Sigma}_{ff} = Cov(\mathbf{f}_t) = \mathbf{I}_m$, and $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_N)'$ is the $N \times m$ matrix of the factor loadings $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \dots, \gamma_{im})'$, for $i = 1, 2, \dots, N$, $\boldsymbol{\varepsilon}_{ot} = (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{Nt})'$ are idiosyncratic errors that are cross-sectionally and serially independent, with zero means and constant variances, ω_i^2 , $0 < \omega_i^2 < K$, $i = 1, 2, \dots, N$. As before, the degree of cross-sectional dependence of \mathbf{x}_{ot} is governed by the largest eigenvalue of the correlation matrix of \mathbf{x}_{ot} , $\mathbf{R} = (\rho_{ij})$. It is easily seen that under the above factor model

$$\rho_{ij} = \sum_{\ell=1}^m \delta_{i\ell} \delta_{j\ell}, \text{ for } i \neq j, \quad (7)$$

where $\delta_{i\ell} = \gamma_{i\ell} / \sqrt{\omega_i^2 + \gamma_i' \gamma_i}$, and $\rho_{ii} = 1$. Also, since ω_i^2 is a fixed non-zero constant, then the scaling of $\gamma_{i\ell}$ by $\sqrt{\omega_i^2 + \gamma_i' \gamma_i}$ ensures that $\gamma_{i\ell}$ and $\delta_{i\ell}$ are of the same order of magnitude. For example, $\delta_{i\ell} = 0$ if $\gamma_{i\ell} = 0$, and $\delta_{i\ell} \neq 0$ if $\gamma_{i\ell} \neq 0$, and *vice versa*.

Consider now the effects of the ℓ^{th} factor, $f_{\ell t}$, on the i^{th} unit, x_{it} , as measured by $\delta_{i\ell}$, and suppose that these factor loadings take non-zero values for M_ℓ out of the N cross section units. Then, following Bailey, Kapetanios and Pesaran (2014 - BKP), the degree of cross-sectional dependence due to the ℓ^{th} factor can be measured by $\alpha_\ell = \ln(M_\ell) / \ln(N)$, and the overall degree of cross-sectional dependence by $\alpha = \max_\ell(\alpha_\ell)$. They define α as the exponent of N that gives the maximum number of x_{it} units, $M = \max_\ell(M_\ell)$, that are pair-wise correlated. The remaining $N - M$ units are either uncorrelated or their correlations tend to zero at a sufficiently fast rate. BKP refer to α as the exponent of cross-sectional dependence and takes any value in the range 0 to 1, with 1 indicating the highest degree of cross-sectional dependence.

The exponent of cross-sectional dependence of x_{it} can be equivalently defined in terms of the scaled factor loadings, δ_i . Without loss of generality, suppose that only the first M_ℓ elements of $\delta_{i\ell}$ over i are non-zero, and note that³

$$\begin{aligned} \bar{\delta}_{\ell N} &= N^{-1} \left(\sum_{i=1}^{M_\ell} \delta_{i\ell} + \sum_{i=M_\ell+1}^N \delta_{i\ell} \right) \\ &= \left(\frac{M_\ell}{N} \right) \left(M_\ell^{-1} \sum_{i=1}^{M_\ell} \delta_{i\ell} \right) = N^{\alpha_\ell-1} \mu_\ell = O(N^{\alpha_\ell-1}), \end{aligned} \quad (8)$$

where $\mu_\ell = M_\ell^{-1} \sum_{i=1}^{M_\ell} \delta_{i\ell} \neq 0$, and $\alpha = \max_\ell(\alpha_\ell)$.

2.3 Average pair-wise correlations as a measure of cross-sectional dependence

As a statistical measure of cross-sectional dependence, $\lambda_{\max}(\mathbf{R})$, can be difficult to analyse, especially for temporally and cross-sectionally dependent data sets. Instead, we can summarise the degree of cross-sectional dependence among the N units conveniently by their average pair-wise correlations, defined by

$$\bar{\rho}_N = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \rho_{ij},$$

which can be written equivalently in terms of the correlation matrix, \mathbf{R} , as

$$\bar{\rho}_N = \frac{\boldsymbol{\tau}' \mathbf{R} \boldsymbol{\tau} - N}{N(N-1)} = \frac{\boldsymbol{\tau}' \mathbf{R} \boldsymbol{\tau}}{N(N-1)} - \frac{1}{N-1}, \quad (9)$$

where $\boldsymbol{\tau}$ is an $N \times 1$ vector of ones. In general, noting that $(\boldsymbol{\tau}' \boldsymbol{\tau}) \lambda_{\min}(\mathbf{R}) \leq \boldsymbol{\tau}' \mathbf{R} \boldsymbol{\tau} \leq (\boldsymbol{\tau}' \boldsymbol{\tau}) \lambda_{\max}(\mathbf{R})$, we have

$$\frac{\lambda_{\min}(\mathbf{R})}{(N-1)} \leq \frac{\boldsymbol{\tau}' \mathbf{R} \boldsymbol{\tau}}{N(N-1)} \leq \frac{\lambda_{\max}(\mathbf{R})}{(N-1)},$$

³The main results in Pesaran (2014) and Bailey, Kapetanios and Pesaran (2014) remain valid even if $\sum_{i=M_\ell+1}^N \delta_{i\ell} = O(1)$. But for expositional simplicity we assume that $\sum_{i=M_\ell+1}^N \delta_{i\ell} = 0$.

and

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

Therefore, in the case of weakly cross correlated processes, such as the spatial autoregressive models, where $\lambda_{\max}(\mathbf{R})$ is bounded in N , $\bar{\rho}_N \rightarrow 0$, as $N \rightarrow \infty$. It is also clear that standard spatial econometric models cannot deal with cases where $\bar{\rho}_N$ differs from zero for sufficiently large N .⁴

In the case of the factor model given by (6), we have,

$$\bar{\rho}_N = \left(\frac{N}{N - 1} \right) \left(\bar{\boldsymbol{\delta}}_N' \bar{\boldsymbol{\delta}}_N - \frac{\sum_{i=1}^N \boldsymbol{\delta}_i' \boldsymbol{\delta}_i}{N^2} \right), \quad (10)$$

where $\bar{\boldsymbol{\delta}}_N = N^{-1} \sum_{i=1}^N \boldsymbol{\delta}_i = (\bar{\delta}_{1N}, \bar{\delta}_{2N}, \dots, \bar{\delta}_{mN})'$, and $\bar{\delta}_{\ell N}$ is given by (8). Further, noting that m is fixed and using (8) we have

$$\bar{\boldsymbol{\delta}}_N' \bar{\boldsymbol{\delta}}_N = \sum_{\ell=1}^m \bar{\delta}_{\ell N}^2 = \sum_{\ell=1}^m O(N^{2\alpha_{\ell}-2}) = O(N^{2\alpha-2}),$$

where as before, $\alpha = \max_{\ell}(\alpha_{\ell})$. Similarly

$$N^{-2} \sum_{i=1}^N \boldsymbol{\delta}_i' \boldsymbol{\delta}_i = \sum_{\ell=1}^m \left(N^{-2} \sum_{i=1}^N \delta_{i\ell}^2 \right) = \sum_{\ell=1}^m O(N^{\alpha_{\ell}-2}) = O(N^{\alpha-2}).$$

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

The values of α in the range $[0, 1/2)$ correspond to different degrees of weak cross-sectional dependence. For these values of α , $\bar{\rho}_N$ tends to zero very fast, at orders that range from N^{-2} to N^{-1} . The values of α in the range $[1/2, 3/4)$ represent moderate degrees of cross-sectional dependence. In this case, $\bar{\rho}_N$ tends to zero at rates ranging from N^{-1} to $N^{-1/2}$. $\bar{\rho}_N$ converges to a non-zero value only if $\alpha = 1$, although for values of α in the range of $[3/4, 1)$, cross-sectional dependence is still quite strong with $\bar{\rho}_N$ tending to zero rather slowly.

2.4 Degrees of cross-sectional dependence and the two-step approach

Since in many applications cross-sectional dependence could be due to common factors as well as spatial or network dependence, it is important that both sources of cross-sectional dependence are taken into account. Mistaking factor dependence, as in (6), for spatial dependence can lead to spurious inference as to the pervasiveness and the degree of the cross-sectional dependence. In consequence, identifying the strength of such dependence is of special significance.

⁴In cases where the degree of cross-sectional dependence is relatively high, one would expect $\lambda_{\max}(\mathbf{R})$ associated with the correlation matrix of the spatial model to be relatively large when \mathbf{W} is row-standardized and ψ is close to unity. Using simulations we can confirm that in such cases $\lambda_{\max}(\mathbf{R})$ rises with N but at a slower rate, such that $\alpha_N = \ln(\lambda_{\max}(\mathbf{R})) / \ln(N)$ tends to a value which is below $1/2$. Also, as is expected, for each N , α_N rises with $|\psi|$ - see Bailey, Kapetanios and Pesaran (2014) for details regarding the specification of α_N .

Suppose observations $\mathbf{x}_{\circ t} = (x_{1t}, x_{2t}, \dots, x_{Nt})'$, $t = 1, 2, \dots, T$, are available and the aim is to model the cross dependence between x_{it} and x_{jt} across $i, j = 1, 2, \dots, N$, with N and T relatively large. A first step requires one to evaluate the strength of the cross-sectional correlation in $\mathbf{x}_{\circ t}$. The application of spatial methods should only be considered if the cross-sectional exponent of the observations, α , is sufficiently small, and particularly not close to unity. Regarding temporal dependence, this can be modelled through common factors or unit-specific dynamics using autoregressive distributed lag models or GVAR specifications (Pesaran, Schuermann and Weiner (2004), and Dees, di Mauro, Pesaran and Smith (2007)).

A two-step procedure suggests itself:

Step 1: Apply the cross-sectional dependence (CD) test developed in Pesaran (2004, 2014) to $\mathbf{x}_{\circ t}$, $t = 1, 2, \dots, T$.

- (a) Only proceed to spatial modelling if the null of weak cross dependence is not rejected.
- (b) If the null of weak dependence is rejected, model the (semi-) strong dependence implied by the test outcome using factor models and check that the residuals from (6), denoted by $\hat{\mathbf{e}}_{\circ t} = (\hat{e}_{1t}, \dots, \hat{e}_{Nt})'$ and referred to as de-factored observations, are weakly cross-correlated (by applying the CD test now to $\hat{\mathbf{e}}_{\circ t}$, $t = 1, 2, \dots, T$).

Step 2: Supposing that the analysis in the first step yields de-factored observations that are reasonably weakly cross-correlated, proceed to spatial or network modelling of the de-factored observations. The residuals $\hat{\mathbf{e}}_{\circ t}$ can also be used to estimate the strength of connections over the cross section units.

In order to test for weak dependence, denote the sample estimates of the pair-wise correlations of the (i, j) units of $\mathbf{x}_{\circ t}$, $t = 1, 2, \dots, T$, by

$$\hat{\rho}_{ij} = \hat{\rho}_{ji} = \frac{\sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)}{\left(\sum_{t=1}^T (x_{it} - \bar{x}_i)^2\right)^{1/2} \left(\sum_{t=1}^T (x_{jt} - \bar{x}_j)^2\right)^{1/2}}, \quad (11)$$

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

$$CD = \left[\frac{TN(N-1)}{2} \right]^{1/2} \hat{\rho}_N, \quad (12)$$

where

$$\hat{\rho}_N = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \hat{\rho}_{ij}. \quad (13)$$

Pesaran (2014) shows that $CD \rightarrow_d N(0, 1)$, under the null hypothesis that the cross-sectional exponent of $\mathbf{x}_{\circ t}$, $t = 1, 2, \dots, T$, 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 $\kappa > 0$.

If the null hypothesis of weak dependence is rejected in step 1 of the above procedure, then according to BKP the exponent of cross-sectional dependence, α , can be estimated.

There are different ways of estimating this exponent if $1/2 < \alpha \leq 1$. We refer to Bailey, Kapetanios and Pesaran (2014) for details.

For de-factoring of the observations common factors can be estimated either by the Principal Components Analysis (PCA), or by cross-sectional averages. Both approaches can be applied to the data set as a whole, or to sub-groups of the data set separately. The grouping can in turn be based on some *a priori* criteria, or could be estimated using discriminant or some other suitable statistical analysis. In the empirical application we use the regional classifications of the MSAs in the USA, but it would also be interesting to apply the group membership procedure recently developed for factor models by Bai and Ando (2014).

The success of de-factoring could be checked by the application of Pesaran's CD test to the de-factored observations. The spatial modelling stage can then begin once we are satisfied that the de-factored observations do not exhibit strong forms of cross-sectional dependence. In the case where the spatial weight matrix is given *a priori*, it might be possible to combine the two steps in one meta approach that simultaneously deals with factor and spatial dependence. It is not clear that such a meta approach would also be possible if the spatial weights, w_{ij} , are to be endogenously determined.

3 Spatial econometrics models revisited

The de-factored observations can now be modelled using spatial econometric techniques that are available in the literature. However, in our application we consider two extensions of this literature, namely endogenising the choice of the spatial weights matrix, \mathbf{W} , and allowing the spatial parameter, ψ , in the SAR model, (1), to be heterogeneous across i .

3.1 Correlation-based specification of spatial weights matrices

Typically \mathbf{W} is constructed using geodesic, demographic or economic information brought in exogenously, and not contained in the data set under consideration, here $\{x_{it}, i = 1, 2, \dots, N; t = 1, 2, \dots, T\}$. Geographical contiguity can be used as in Holly, Pesaran and Yamagata (2011a,b). In economic applications, economic measures, such as commuting times, trade and migratory flows across geographical areas have been used. For example, in GVAR modelling trade weights are used in the construction of link matrices that relate individual economies to their trading partners in the global economy - Pesaran, Schuermann and Weiner (2004). Such measures are often preferable over geodesic measures - since they are closer to the decisions that underlie the observations, x_{it} , and they allow also for possible time variations in the weighting matrix which of course is not possible if we use only physical distance measures in the construction of \mathbf{W} .

In practice, however, suitable economic distance might not be available in many applications, and it is desirable to see if \mathbf{W} can be constructed without recourse to such exogenous information. In applications where the time dimension is reasonably large (around 60-80), it is possible to identify the non-zero elements of \mathbf{W} with those elements of $\hat{\rho}_{ij}$, as expressed in (11), that are different from zero at a suitable significance level. There is a related literature that addresses the issue of identification of neighbours by non-zero elements of an assumed sparse covariance matrix or its inverse. The inverse covariance matrix is used in Markov networks defined as a graphical model that represents variables as nodes and 'conditional' dependencies (partial correlations) between variables

as (undirected) edges. Estimation then amounts to setting elements of the *inverse* covariance matrix to zero - Dempster (1972).⁵ A number of estimation approaches have followed, using lasso type penalties applied to the estimation of the inverse covariance matrix directly. The main problem with this approach is that once de-factoring has taken place the interpretation of the resulting inverse covariance matrix is ambiguous. Furthermore, finding a good estimate of the inverse covariance matrix especially when $N > T$ can be challenging. The second approach uses the so-called covariance graph at its focus which is the corresponding graphical model for ‘marginal’ dependencies (marginal correlations). Methods of estimating covariance matrices with zero elements include contributions by Chaudhuri, Drton and Richardson (2007), Khare and Rajaratnam (2011), Butte, Tamayo, Slonim, Golub and Kohane (2000), and Rothman, Bickel, Levina and Zhu (2008, 2010).⁶

In this paper we follow the second approach and employ the multiple testing procedure recently developed in Bailey, Pesaran and Smith (2014, BPS) which tests the statistical significance of the pair-wise correlations of the de-factored observations, ρ_{ij} . As shown below, it is simple to implement, is invariant to the ordering of the underlying units, and consistently estimates the true positive rate and the false positive rate of zeros and ones of the underlying \mathbf{W} matrix.

The multiple testing problem arises when we are faced with a number of (possibly) dependent tests and our aim is to control for the size of the overall test. Suppose we are interested in a family of null hypotheses, $H_{01}, H_{02}, \dots, H_{0n}$ and we are provided with corresponding test statistics, $Z_{1T}, Z_{2T}, \dots, Z_{nT}$, with separate rejection rules given by (using a two sided alternative)

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

where CV_{iT} is some suitably chosen critical value of the test, and p_{iT} is the observed p value for H_{0i} . Consider now the family-wise error rate (FWER) defined by

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

and suppose that we wish to control $FWER_T$ to lie below a pre-determined value, p . Bonferroni (1935, 1936) provides a general solution, which holds for all possible degrees of dependencies across the separate tests. By Boole’s inequality we have

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

Hence, to achieve $FWER_T \leq p$, it is sufficient to set $p_{iT} \leq p/n$. However, Bonferroni’s procedure can be quite conservative, particularly when the tests are highly correlated. This means that the procedure does not reject as often as it should and therefore lacks power. A step-down procedure is proposed by Holm (1979) which is more powerful than

⁵A recent economic application of this approach is provided by Barigozzi and Brownlees (2013).

⁶For further contributions in this area see Meinshausen and Bühlmann (2006), Peng, Wang, Zhou and Zhu (2009), and Bien and Tibshirani (2011), among others.

Bonferroni's procedure, without imposing any further restrictions on the degree to which the underlying tests depend on each other.

If we abstract from the T subscript and order the p -values of the tests, so that

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)},$$

are associated with the null hypotheses, $H_{(01)}, H_{(02)}, \dots, H_{(0n)}$, respectively, Holm's procedure rejects $H_{(01)}$ if $p_{(1)} \leq p/n$, rejects $H_{(01)}$ and $H_{(02)}$ if $p_{(2)} \leq p/(n-1)$, rejects $H_{(01)}, H_{(02)}$ and $H_{(03)}$ if $p_{(3)} \leq p/(n-2)$, and so on.⁷

In our application, we apply multiple testing procedures to distinct non-diagonal elements of the sample estimate of $\mathbf{R} = (\rho_{ij})$, namely $\hat{\mathbf{R}} = (\hat{\rho}_{ij})$, where $\hat{\rho}_{ij}$ is the correlation of the de-factored price changes between i and j MSAs. BPS show that the application of the Bonferroni procedure to $\hat{\mathbf{R}}$ yields a regularised version which converges to \mathbf{R} at the rate of $\sqrt{m_N N/T}$ under the Frobenius norm, where m_N is bounded in N , and represents the number of non-zero off-diagonal elements in each row of \mathbf{R} . More importantly for the present application, BPS establish that the zeros of $\mathbf{W} = (w_{ij})$ are consistently estimated by

$$\hat{w}_{ij} = I \left(|\hat{\rho}_{ij}| > \frac{C_p(N)}{\sqrt{T}} \right),$$

where $C_p(N) = \Phi^{-1} \left(1 - \frac{p}{N(N-1)} \right)$, p is the pre-specified overall size of the test (which we set to 5% in the empirical application), and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative standard normal distribution. More specifically, consider the true positive rate (TPR) and the false positive rate (FPR) of ones/zeros in the \mathbf{W} matrix as defined by

$$TPR = \frac{\sum_{i \neq j} I(\hat{w}_{ij} \neq 0, \text{ and } \rho_{ij} \neq 0)}{\sum_{i \neq j} I(\rho_{ij} \neq 0)} \quad (14)$$

$$FPR = \frac{\sum_{i \neq j} I(\hat{w}_{ij} \neq 0, \text{ and } \rho_{ij} = 0)}{\sum_{i \neq j} I(\rho_{ij} = 0)}. \quad (15)$$

BPS show that under certain plausible regularity conditions $TPR \rightarrow 1$ and $FPR \rightarrow 0$ as N and $T \rightarrow \infty$ with probability one, so long as $\rho_{\min} = \min_{i,j}(\rho_{ij}) > C_p(N)/\sqrt{T}$.

Similar results also hold if the Holm procedure is applied. Let $n = N(N-1)/2$ and order the p -values of n individual tests in an ascending manner, which is equivalent to ordering $|\hat{\rho}_{ij}|$ in a descending manner. Denote the largest value of $|\hat{\rho}_{ij}|$ over all $i \neq j$, by $|\hat{\rho}_{(1)}|$, the second largest value by $|\hat{\rho}_{(2)}|$, and so on, to obtain the ordered sequence $|\hat{\rho}_{(s)}|$, for $s = 1, 2, \dots, n$. Then the (i, j) pair associated with $|\hat{\rho}_{(s)}|$ are connected (i.e. $w_{ij} \neq 0$) if $|\hat{\rho}_{(s)}| > T^{-1/2} \Phi^{-1} \left(1 - \frac{p/2}{m-s+1} \right)$, otherwise disconnected (i.e. $w_{ij} = 0$), for $s = 1, 2, \dots, n$. The resultant connection matrix will be denoted by $\hat{\mathbf{W}} = (\hat{w}_{ij})$, 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 $\hat{\rho}_{ij} < 0$.

⁷Other multiple testing procedures can also be considered such as Hochberg (1988), Hommel (1988) and Hommel (1989) among others - we thank an anonymous referee for drawing our attention to these contributions to the multiple testing literature. Efron (2010) also provides a recent review. But most of these procedures tend to place undue prior restrictions on the dependence of the underlying test statistics while the Holm method is not subject to such restrictions.

3.2 A heterogeneous spatio-temporal model

Almost all spatial econometric models estimated in the literature assume that the spatial parameters do not vary across the units. For example, in the case of the SAR(1) model defined by (1), the parameter ψ is restricted to be the same across $i = 1, 2, \dots, N$. Such parameter homogeneity assumptions are unavoidable when T is very small, but need not be imposed in the case of large panels where T is sufficiently large. The evidence of parameter heterogeneity in panel data models is quite prevalent particularly in the case of cross county or country data sets. In such cases and when T is sufficiently large, reducing the spatial effects into a single parameter appears rather restrictive. Instead, Aquaro, Bailey and Pesaran (2014, ABP) allow the spatial effects to differ across the units, and derive the conditions needed for identification and consistent estimation under parameter heterogeneity.⁸ ABP consider the following heterogeneous version of (1)

$$x_{it} = \psi_i x_{it}^* + u_{it}, \text{ for } i = 1, 2, \dots, N; t = 1, 2, \dots, T,$$

where $x_{it}^* = \mathbf{w}_i' \mathbf{x}_{ot}$, \mathbf{w}_i' denotes the i^{th} row of the $N \times N$ row-standardised spatial matrix, \mathbf{W} , which is taken as given. In the spatial econometrics literature it is assumed that all units have at least one neighbour (or connection), which ensures that $\mathbf{w}_i' \boldsymbol{\tau} = 1$ for all i , when \mathbf{W} is row-standardised. But as we shall see, particularly when using correlation-based weights, it is possible for some units not to have any connections. In such cases $x_{it}^* = 0$, and the associated coefficient, ψ_i , is unidentified, and to resolve the identification problem, and without loss of generality, we set $\psi_i = 0$. Notice that since for such units $x_{it}^* = 0$, the choice of ψ_i will not impact the results.

In matrix notation we have

$$\mathbf{x}_{ot} = \boldsymbol{\Psi} \mathbf{W} \mathbf{x}_{ot} + \mathbf{u}_{ot}, \quad t = 1, 2, \dots, T, \quad (16)$$

where $\boldsymbol{\Psi} = \text{diag}(\boldsymbol{\psi})$, and $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_N)'$. An extension of (16), that incorporates richer temporal and spatial dynamics and accommodates negative as well as positive connections, is given by

$$\mathbf{x}_{ot} = \sum_{j=1}^{h_\lambda} \boldsymbol{\Lambda}_j \mathbf{x}_{ot-j} + \sum_{j=0}^{h_\psi^+} \boldsymbol{\Psi}_j^+ \mathbf{W}^+ \mathbf{x}_{ot-j} + \sum_{j=0}^{h_\psi^-} \boldsymbol{\Psi}_j^- \mathbf{W}^- \mathbf{x}_{ot-j} + \mathbf{u}_{ot}, \quad (17)$$

where $h_\lambda = \max(h_{\lambda 1}, h_{\lambda 2}, \dots, h_{\lambda N})$, $h_\psi^+ = \max(h_{\psi 1}^+, h_{\psi 2}^+, \dots, h_{\psi N}^+)$, $h_\psi^- = \max(h_{\psi 1}^-, h_{\psi 2}^-, \dots, h_{\psi N}^-)$, $\boldsymbol{\Lambda}_j$, $\boldsymbol{\Psi}_j^+$ and $\boldsymbol{\Psi}_j^-$ are $N \times N$ diagonal matrices with λ_{ij} , ψ_{ij}^+ and ψ_{ij}^- over i as their diagonal elements. Also, \mathbf{W}^+ and \mathbf{W}^- are $N \times N$ network matrices for positive and negative connections, respectively such that $\mathbf{W} = \mathbf{W}^+ + \mathbf{W}^-$. In this model the slope coefficients, λ_{ij} , ψ_{ij}^+ and ψ_{ij}^- , and the error variances, $\sigma_{u_i}^2 = \text{var}(u_{it})$ are allowed to differ across i . In what follows for expositional simplicity we set $h_\lambda = h_\psi^+ = h_\psi^- = 1$.

For consistent estimation of the parameters ABP propose a QML procedure, and suggest using the following concentrated log-likelihood function to simplify the computations:

$$\ell(\boldsymbol{\psi}_0^+, \boldsymbol{\psi}_0^-) \propto T \ln |I_N - \boldsymbol{\Psi}_0^+ \mathbf{W}^+ - \boldsymbol{\Psi}_0^- \mathbf{W}^-| - \frac{T}{2} \sum_{i=1}^N \ln \left(\frac{1}{T} \tilde{\mathbf{x}}_i' \mathbf{M}_i \tilde{\mathbf{x}}_i \right), \quad (18)$$

⁸The literature on panel data models with heterogeneous slopes is reviewed in Hsiao and Pesaran (2008).

where $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \psi_{i0}^+ \mathbf{x}_i^+ - \psi_{i0}^- \mathbf{x}_i^-$, $\mathbf{M}_i = \mathbf{I}_T - \mathbf{Z}_i (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i'$, $\mathbf{Z}_i = (\mathbf{x}_{i,-1}, \mathbf{x}_{i,-1}^+, \mathbf{x}_{i,-1}^-)$, $\boldsymbol{\psi}_0^+ = (\psi_{10}^+, \psi_{20}^+, \dots, \psi_{N0}^+)'$ and $\boldsymbol{\psi}_0^- = (\psi_{10}^-, \psi_{20}^-, \dots, \psi_{N0}^-)'$. The parameters of the lagged variables, $\boldsymbol{\lambda}_1, \boldsymbol{\psi}_1^+$ and $\boldsymbol{\psi}_1^-$, can be estimated by least squares applied to the equations for individual units conditional on ψ_{i0}^+ and ψ_{i0}^- . Inference on individual coefficients are then carried out using second cross derivatives of the full likelihood function of (17), with respect to $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \dots, \boldsymbol{\theta}'_N)'$, where $\boldsymbol{\theta}_i = (\psi_{i0}^+, \psi_{i0}^-, \psi_{i1}^+, \psi_{i1}^-, \lambda_{i1}, \sigma_{u_i}^2)'$. The variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{ML}$ is computed as

$$\hat{\Sigma}_{\boldsymbol{\theta}_{ML}} = \left[-\frac{1}{T} \frac{\partial^2 \ell(\hat{\boldsymbol{\theta}}_{ML})}{\partial \hat{\boldsymbol{\theta}}_{ML} \partial \hat{\boldsymbol{\theta}}_{ML}'} \right]^{-1}. \quad (19)$$

Further details regarding the econometric analysis of this model are provided in Aquaro, Bailey and Pesaran (2014).

4 Application: US house prices

The two-step procedure developed in this paper can be applied to different types of panel data sets so long as the time series dimension of the panel is reasonably large such that reliable estimates of pair-wise correlations, ρ_{ij} , can be obtained. There are many such panels, covering regions or countries, that can be considered. Regional data in the United States have been studied by many including Cromwell (1992), Pollakowski and Ray (1997), Carlino and DeFina (1998, 2004), Carlino and Sill (2001), Del Negro (2002), Owyang, Piger and Wall (2005), and Partridge and Rickman (2005). The cross country data sets used in global modelling provide another example.

Here we opt to study house price changes at the level of Metropolitan Statistical Areas in the US.⁹ Metropolitan Statistical Areas (MSAs) are geographic entities delineated by the Office of Management and Budget and are used by Federal statistical agencies when collecting, tabulating, and publishing Federal statistics for spatial units in the USA. The 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. They range in size, as measured by population in 2008, from the smallest - Carson City - with a population of 55,000, to New York and its environs with a population of 18.97 million. Moreover, there can be considerable distances between MSAs. The pair-wise average distance is 1,156 miles, though of course this is exaggerated by the relative sparseness of the distribution of MSAs in the Midwest. Indeed, by comparison, the study of regional house prices in the UK by Holly, Pesaran and Yamagata (2011b) deals with distances of a much smaller magnitude. Distance is, therefore, likely to be an important factor for the spatial distribution of house prices, though size could play a role as well.

⁹There already exists a large literature on the spatial dimension of house price changes, partly because of the availability of spatially disaggregated data, but also because of the role that housing plays in household wealth and in the transmission of monetary policy shocks, and more recently as a conduit for transmission of global shocks. Recent contributions include Rapach and Strauss (2007, 2009), Kadiyala and Bhattacharya (2009), Gupta and Das (2010), Gupta, Kabundi and Miller (2011a,b), Kuethe and Pede (2011), and Gupta and Miller (2012). These papers apply a number of models to house price data without prior assessment of their degree of cross-sectional dependence.

Our choice of house prices is also motivated by the role that housing plays in spatial equilibrium models (Glaeser, Gyourko and Saiz (2008), Glaeser and Gottlieb (2009)). The standard approach in urban and regional economics is to assume a spatial equilibrium. At the margin firms and households have to be indifferent between alternative locations. Firms employ labour up to the point at which the wage is equal to marginal product; construction companies supply housing up to the point at which marginal cost is equal to marginal product. Finally, households have to be indifferent about where they are located, taking into account wages, the price of houses and the local availability of amenities (proximity to schools, sea, mountains, temperature, etc.). The combination of the labour supply curve, the supply curve for housing and the labour demand determines simultaneously the population of say a locality, wages and the price of housing. Idiosyncratic differences in space in terms of productivity, particular characteristics of an area and the construction sector determine differences across space in population density, household incomes and the house prices. There are a number of equilibrating processes at work. Households tend to move across geographical areas in response to differences in wages, house prices and area characteristics. There can also be agglomeration effects due to economies of scale in relation to size and population density of cities. But, it should be clear that such equilibrating tendencies are likely to operate fully only in the long run, over a number of years rather than quarters. It takes time for households to relocate in response to changing economic circumstances. It also takes time for construction companies to increase the supply of housing. Any model of house price diffusion across MSAs must also adequately deal with dynamics, both within and across MSAs.

We consider 363 MSAs in total, excluding three MSAs located in Alaska and Hawaii,¹⁰ over the period 1975Q1 to 2010Q4 ($T = 144$). We denote the level of house prices in MSA i , located in State s , in quarter t , by P_{ist} , for $i = 1, \dots, N_s$, $s = 1, \dots, S$, and $t = 1, \dots, T$, where $\sum_{s=1}^S N_s = N = 363$, $S = 49$ (comprised of 48 contiguous States and the district of Columbia), and $T = 144$ quarters. Then we compute real house prices as:

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

where CPI_{st} is the Consumer Price Index of State s in quarter t . Details on the sources of these data can be found in Appendix I. We ordered the MSAs by State, starting at the East Coast and moving towards the West Coast, following the list provided in Table A of Appendix II. Finally, we obtain seasonally adjusted changes in real house prices, π_{ist} , as residuals from regressing $p_{ist} - p_{is,t-1}$, the seasonally unadjusted rate of change in real house prices, on an intercept and three quarterly seasonal dummies. Before modelling the spatial dimension of the price changes, π_{ist} , we first need to examine the extent to which these price changes are cross-sectionally strongly correlated and then de-factor such effects if necessary. (see sub-section 2.4)

4.1 Cross-sectional dependence in house price changes and de-factoring of observations

To examine the degree of cross-sectional dependence in house price changes, we computed the CD statistic of Pesaran (2004, 2014) for price changes, π_{ist} , without any de-factoring.

¹⁰Note that the District of Columbia is treated as a single MSA.

(See (12) and (13)). We obtained $CD_\pi = 640.46$ ($\hat{\rho}_\pi = 0.209$) as compared to a critical value of 1.96 at the 5% significance level. The test outcome is statistically highly significant and suggests a high degree of cross-sectional dependence in house price changes, which could be due to common national and regional effects. Applying the method proposed by BKP we calculate the exponent of cross-sectional dependence (standard error in parenthesis) for house price changes and obtain $\hat{\alpha}_\pi = 0.989$ (0.03), which is very close to unity and suggests that house price changes are strongly correlated across MSAs. Clearly, it would be inappropriate to apply standard spatial modelling techniques directly to π_{ist} , as noted earlier since the maintained assumption of spatial econometric models is weak cross-sectional dependence as shown in sub-section 2.1.

4.1.1 De-factoring using cross-sectional averages

The strong cross-sectional dependence in house price changes can be modelled using observed (national/regional income, unemployment and interest rates), or unobserved common factors (using principal components). Alternatively, one can use cross-sectional averages at the national and regional levels, which corresponds to assuming the existence of a national factor together with a number of region-specific factors.¹¹ The correspondence of cross-sectional averages and the national and regional factors can be established using arguments similar to the ones advanced in Pesaran (2006). In what follows we employ the Bureau of Economic Analysis classification and consider a total of $R = 8$ regions in the US containing an average of approximately 45 MSAs each. These are: (i) New England, (ii) Mid East, (iii) South East, (iv) Great Lakes, (v) Plains, (vi) South West, (vii) Rocky Mountains, and (viii) Far West. See Table A of Appendix II for more details.

Accordingly, let π_{irt} denote the rate of change of real house prices (after seasonal adjustments) in the i^{th} MSA located in region $r = 1, 2, \dots, R$, at time t , and consider the following hierarchical model

$$\begin{aligned}\pi_{irt} &= a_{ir} + \beta_{ir}\bar{\pi}_{rt} + \gamma_{ir}\bar{\pi}_t + \xi_{irt}, \\ i &= 1, 2, \dots, N_r; r = 1, 2, \dots, R; t = 2, 3, \dots, T,\end{aligned}\tag{20}$$

where, $\bar{\pi}_{rt} = N_r^{-1} \sum_{i=1}^{N_r} \pi_{irt}$, and $\bar{\pi}_t = N^{-1} \sum_{r=1}^R \sum_{i=1}^{N_r} \pi_{irt}$, with $N = \sum_{r=1}^R N_r$. Write the above model more compactly as

$$\boldsymbol{\pi}_t = \mathbf{a} + \mathbf{B}\mathbf{Q}_N\boldsymbol{\pi}_t + \mathbf{\Gamma}\mathbf{P}_N\boldsymbol{\pi}_t + \boldsymbol{\xi}_t,\tag{21}$$

where $\boldsymbol{\pi}_t$ is an $N \times 1$ vector of house price changes partitioned by regions, namely

$$\boldsymbol{\pi}_t = (\pi_{11t}, \pi_{21t}, \dots, \pi_{N_1t}; \pi_{12t}, \pi_{22t}, \dots, \pi_{N_2t}; \dots; \pi_{1Rt}, \pi_{2Rt}, \dots, \pi_{N_Rt})'.$$

Similarly

$$\mathbf{a} = (a_{11}, a_{21}, \dots, a_{N_11}; a_{12}, a_{22}, \dots, a_{N_22}; \dots; a_{1R}, a_{2R}, \dots, a_{N_RR})'.$$

\mathbf{B} and $\mathbf{\Gamma}$ are $N \times N$ diagonal matrices with their ordered elements given by

$$\beta_{11}, \beta_{21}, \dots, \beta_{N_11}; \beta_{12}, \beta_{22}, \dots, \beta_{N_22}; \dots; \beta_{1R}, \beta_{2R}, \dots, \beta_{N_RR},$$

¹¹We also considered using State level averages, but there were only a few MSAs in some States.

and

$$\gamma_{11}, \gamma_{21}, \dots, \gamma_{N_1 1}; \gamma_{12}, \gamma_{22}, \dots, \gamma_{N_2 2}; \dots; \gamma_{1R}, \gamma_{2R}, \dots, \gamma_{N_R R},$$

respectively. Finally, \mathbf{Q}_N and \mathbf{P}_N are $N \times N$ projection matrices such that $\mathbf{Q}_N \boldsymbol{\pi}_t$ give the regional means and $\mathbf{P}_N \boldsymbol{\pi}_t$ the national mean of the local feature. More specifically, let $\boldsymbol{\tau}_{N_r}$ be an $N_r \times 1$ vector of ones, and $\boldsymbol{\tau}_N$ an $N \times 1$ vector of ones, then

$$\mathbf{P}_N = \boldsymbol{\tau}_N (\boldsymbol{\tau}_N' \boldsymbol{\tau}_N)^{-1} \boldsymbol{\tau}_N',$$

and

$$\mathbf{Q}_N = \begin{pmatrix} \mathbf{P}_{N_1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{N_2} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{P}_{N_{R-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{P}_{N_R} \end{pmatrix},$$

where $\mathbf{P}_{N_r} = \boldsymbol{\tau}_{N_r} (\boldsymbol{\tau}_{N_r}' \boldsymbol{\tau}_{N_r})^{-1} \boldsymbol{\tau}_{N_r}'$. It is assumed that R is fixed, and for each r , N_r/N tends to a non-zero constant as $N \rightarrow \infty$. $\mathbf{P}_{N_r} \boldsymbol{\pi}_t$, for $r = 1, 2, \dots, R$, and $\mathbf{P}_N \boldsymbol{\pi}_t$ can be viewed as regional and national factors that are consistently estimated by simple averages. They also represent the strong form of cross-sectional dependence in the real house price changes across MSAs.

The de-factored real house price changes are then given by residuals from (21), namely

$$\hat{\boldsymbol{\xi}}_t = \boldsymbol{\pi}_t - \hat{\mathbf{a}} - \hat{\mathbf{B}} \mathbf{Q}_N \boldsymbol{\pi}_t - \hat{\mathbf{\Gamma}} \mathbf{P}_N \boldsymbol{\pi}_t, \quad t = 2, \dots, T. \quad (22)$$

To check if de-factoring has been successful (step 1(b) in sub-section 2.4), we apply the CD test to the residuals, $\hat{\boldsymbol{\xi}}_t$. The resulting CD statistic is much reduced, falling from 640.45 to -6.05 , giving a very small estimate for the average pair-wise correlations, $\hat{\rho}_{\hat{\boldsymbol{\xi}}} = -0.002$. Therefore, the simple hierarchical de-factoring procedure has managed to eliminate almost all of the strong cross-sectional dependence that had existed in house price changes, and what remains could be due to the local dependencies that need to be modelled using spatial techniques. Also, the estimate of the exponent of cross-sectional dependence, α , which stood at $\hat{\alpha}_{\pi} = 0.989$ (0.03) is now reduced to $\hat{\alpha}_{\hat{\boldsymbol{\xi}}} = 0.637$ (0.03) which is close to the borderline value of $1/2$, representing weak cross-sectional dependence.

4.1.2 De-factoring using Principal Components

For comparison we repeat the de-factoring analysis by applying the method of principal components developed for large panels in Bai (2003) to price changes, to begin with without a regional classification. Consider the following m -factor model for π_{it}

$$\pi_{it} = a_i + \boldsymbol{\gamma}_i' \mathbf{f}_t + \xi_{f,it}, \quad i = 1, 2, \dots, N; \quad t = 2, 3, \dots, T, \quad (23)$$

where \mathbf{f}_t is an $m \times 1$ vector of unobserved factors. For the time being suppose that m is known. Then the factors can be estimated by the first m principal components (PC) of real house price changes, and the de-factored observations can be obtained as residuals from regressions of π_{it} on the m largest PCs. To select m we applied the six information criteria (IC) proposed in Bai and Ng (2002), specifying 8 as the maximum number of factors to match the number of cross-sectional averages used in the hierarchical specification of the previous sub-section. All six IC ended up selecting 8 as the optimal

number of factors. We increased the maximum number of factors in the procedure, but still ended up selecting the maximum as the optimal. In view of the failure of the IC to lead to any meaningful outcome, we decided to conduct the de-factoring analysis using $m = 2, 3, \dots, 8$ principal components. We then computed CD statistics for the de-factored residuals for all 7 choices of m , and obtained the values of 53.39, 10.21, 2.73, 3.27, 2.31, -1.96 and -4.42 respectively, for $m = 2, 3, \dots, 8$. The corresponding estimates of the exponent of cross-sectional dependence were $\hat{\alpha}_{2pc} = 0.932$ (0.04), $\hat{\alpha}_{3pc} = 0.799$ (0.04), $\hat{\alpha}_{4pc} = 0.793$ (0.03), $\hat{\alpha}_{5pc} = 0.785$ (0.03), $\hat{\alpha}_{6pc} = 0.831$ (0.02), $\hat{\alpha}_{7pc} = 0.718$ (0.02), and $\hat{\alpha}_{8pc} = 0.622$ (0.02), respectively. It is evident that as more factors are added the strength of cross-sectional dependence of the resulting residuals diminishes progressively.

However since the testing procedure for the determination of m appears inconclusive, and following a suggestion by an anonymous referee, we considered a hierarchical PCA, similar to the specification in (21), where the national factor is represented by the strongest principal component extracted from the full data set and the regional factor(s) are expressed as the strongest principal component(s) extracted from each of the eight regions separately. Such a hierarchical factor model can be written as

$$\begin{aligned}\pi_{irt} &= a_{ir} + \beta'_{ir} \mathbf{f}_{rt} + \gamma_{ir} f_{gt} + \xi_{irt}, \\ i &= 1, 2, \dots, N_r; \quad r = 1, 2, \dots, R; \quad t = 2, 3, \dots, T,\end{aligned}\tag{24}$$

where \mathbf{f}_{rt} is an $m_r \times 1$ vector of regional factors for $r = 1, 2, \dots, R$, and f_{gt} is the ‘global’ or ‘national’ factor. $\beta_{ir} = (\beta_{i1}, \beta_{i2}, \dots, \beta_{im_r})'$ and γ_{ir} are associated factor loadings. Again to select the number of PCs, m_r , in each region we applied the six information criteria of Bai and Ng (2002). The results using as the maximum number of factors of 4 and 8 in their procedure, are shown in Table 1 below:

Table 1: Number of factors selected at national and regional levels using the Bai and Ng information criteria

	National	New Engl.	Mid East	South East	Gr. Lakes	Plains	South West	Rock. Mount.	Far West
Maximum number of factors assumed = 8									
PC_1	8	8	7	8	7	7	6	7	8
PC_2	8	8	6	8	7	6	5	7	7
PC_3	8	8	8	8	8	8	8	8	8
IC_1	8	8	4	8	7	4	2	2	8
IC_2	8	8	4	6	3	2	2	2	4
IC_3	8	8	8	8	8	6	5	7	8
Maximum number of factors assumed = 4									
PC_1	4	4	4	4	4	4	4	4	4
PC_2	4	4	4	4	4	4	3	3	4
PC_3	4	4	4	4	4	4	4	4	4
IC_1	4	4	4	4	4	4	2	2	4
IC_2	4	4	4	4	3	2	2	2	4
IC_3	4	4	4	4	4	4	4	4	4

Notes: The Bai and Ng (2002) information criteria are set out in what follows: $PC_1(m_r) = V_{m_r} + m_r \hat{\sigma}^2 \frac{(T+N_r)}{(TN_r)} \ln \left(\frac{TN_r}{(T+N_r)} \right)$, $PC_2(m_r) = V_{m_r} + m_r \hat{\sigma}^2 \frac{(T+N_r)}{(TN_r)} \ln C_{NT}^2$, $PC_3(m_r) = V_{m_r} + m_r \hat{\sigma}^2 \left(\frac{\ln C_{N_r T}^2}{C_{N_r T}^2} \right)$, $IC_1(m_r) = \ln(V_{m_r}) + m_r \left(\frac{(T+N_r)}{TN_r} \right) \ln \left(\frac{TN_r}{(T+N_r)} \right)$, $IC_2(m_r) = \ln(V_{m_r}) + m_r \left(\frac{(T+N_r)}{TN_r} \right) \ln C_{N_r T}^2$ and $IC_3(m_r) = \ln(V_{m_r}) + m_r \left(\frac{\ln C_{N_r T}^2}{C_{N_r T}^2} \right)$. Here $\hat{\sigma}^2$ is a consistent estimate of $(N_r T)^{-1} \sum_{i=1}^N \sum_{t=1}^T E(\xi_{irt}^2)$ from $\pi_{irt} = \beta'_{ir} \mathbf{f}_{rt} + \xi_{irt}$, $i = 1, 2, \dots, N_r; r = 1, 2, \dots, R; t = 2, \dots, T$, and $C_{N_r T}^2 = T \wedge N_r := \min(T, N_r)$. Also, $V_{m_r} = (N_r T)^{-1} \sum_{i=1}^{N_r} \sum_{t=1}^T \left(\pi_{irt} - \sum_{j=1}^{m_r} \hat{\beta}_{j,i} \hat{f}_{j,t} \right)^2$, for $r = 1, \dots, R$.

In the global and eight regional cases all six IC selected more than one PC though some variation in the number of PCs is evident across regions. Again given the lack of clear guidance as to the number of principal components to use in (24), and in order to strike

a balance between purging the house price changes of common effects and still leaving a sufficient degree of spatial dependence in the de-factored observations we decided to opt for one PC as the ‘national’ factor and $m_r = 2$ for all regions considered. As before we computed the CD statistic of the de-factored residuals from regressions (24) which amounted to a value of 3.320 ($\hat{\rho}_{\xi_{pc}} = 0.001$), which is comparable to the results obtained using the national and regional cross-sectional averages. The estimate of the exponent of cross-sectional dependence in this case turned out to be $\hat{\alpha}_{pc} = 0.773$ (0.03), which is somewhat larger than the estimate $\hat{\alpha}_{\xi} = 0.637$ (0.03) obtained using the cross-sectional averages as factors.

It is clear that both de-factoring methods (cross-sectional averages and PC) do reasonably well in purging price changes of common effects, although the hierarchical version of the PC seems preferable to the standard application of PC to house price changes without any regional classification.

4.2 Estimation of spatial connections

Having computed de-factored price changes, $\hat{\xi}_{it}$ - from (22), and $\hat{\xi}_{it,pc}$ - from (24), we are now in a position to apply the methodology developed in sub-section 3.1 to estimate the matrix of connections, $\hat{\mathbf{W}}$, using pair-wise correlations of de-factored price changes.

4.2.1 Spatial weights matrices based on distance

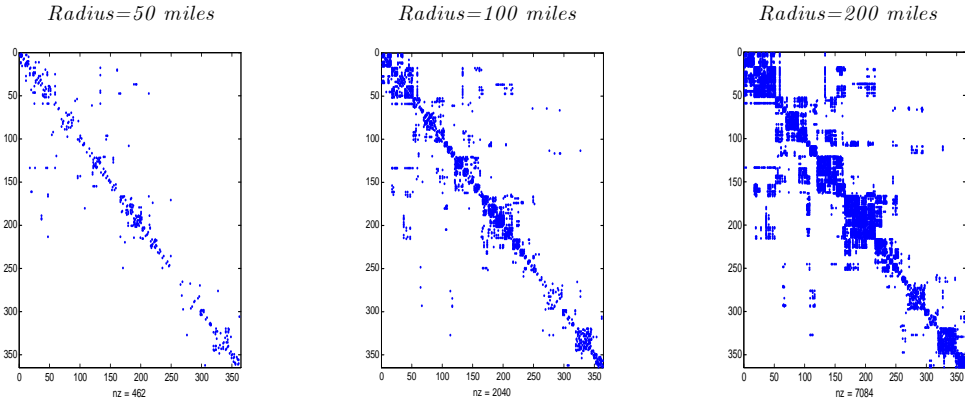
We start our analysis with a standard specification of \mathbf{W} based on contiguity measures, which we also use as a benchmark to examine the estimates of \mathbf{W} that are based on pair-wise correlations, $\hat{\rho}_{ij}$. As noted above, MSAs are deliberately defined as one or more large cities with their core having a substantial influence over the surrounding region, with a high degree of economic and social integration through commuting or transport links. Hence, it can be argued that geographical distance can play an important role in the determination of connections across different MSAs. We consider the same 363 MSAs as before. We denote a weights matrix based on physical distance by \mathbf{W}_d , and make use of data for geodesic distance (d) by applying the Haversine formula to data on the Latitude-Longitude of zip codes, cross referenced to each of the $N = 363$ MSAs.¹² We regard \mathbf{W}_d (of dimension $N \times N$) to be symmetric. We identify as neighbours for each MSA, i ($i = 1, \dots, N$), all MSAs that lie within a radius of d miles. This pattern translates into a value of 1 for elements (i, j) and (j, i) of \mathbf{W}_d if MSA i is a neighbour (falling within the given radius) of MSA j , or a value of 0 otherwise. Diagonal entries (i, i) take a value of 0, indicating that MSA i cannot be a neighbour of itself. Also under this specification all non-zero elements of \mathbf{W}_d are viewed as representing a positive connection, which should be contrasted with connections that are based on economic factors that could lead to negative as well as positive connections.

We study three cases: (i) MSAs within a radius of $d = 50$ miles, (ii) MSAs within a radius of $d = 100$ miles, and (iii) MSAs within a radius of $d = 200$ miles. These give rise to three \mathbf{W}_d matrices, namely (i) \mathbf{W}_{50m} , (ii) \mathbf{W}_{100m} , and (iii) \mathbf{W}_{200m} , which are sparse by nature, but of different degree depending on the cut-off point set by the radius, d . We compare the degree of sparseness of \mathbf{W}_{50m} , \mathbf{W}_{100m} , and \mathbf{W}_{200m} in terms of their percentage of non-zero elements (excluding the diagonal elements). This percentage is

¹²See Appendix III for details of this formula.

0.35% for \mathbf{W}_{50m} , 1.55% for \mathbf{W}_{100m} , and 5.39% for \mathbf{W}_{200m} . As expected, the number of non-zero elements increases when the radius within which MSAs are considered to be neighbours rises. Figure 1 displays all three \mathbf{W}_d matrices. In this figure we have ordered the MSAs by States starting at the East Coast as before and following the list provided in Table A of Appendix II, from top to bottom and from left to right. The sparseness of the \mathbf{W}_d matrices is captured by white areas in the graph when the relevant entries are equal to zero. As is to be expected there is considerable clustering along the diagonal, but because we are using a line to depict a plane, sometimes an MSA may lie at the edge of a State (or region) and fall within the radius of another State or region. Clearly, as the radius is increased from 50 to 200 miles the degree of leaching increases.

Figure 1: Spatial weights matrices specified by distance



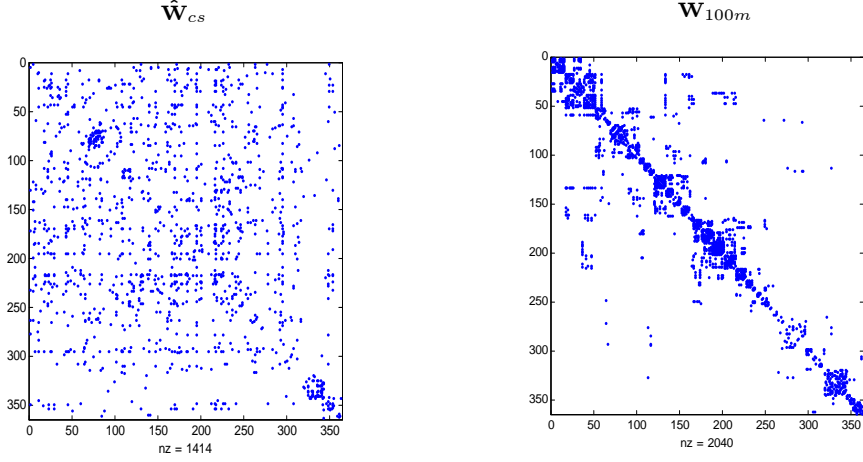
4.2.2 Spatial weights matrices based on pair-wise correlations

Next we make use of the de-factored price changes, $\hat{\xi}_{it}$ and $\hat{\xi}_{it,pc}$, to estimate the matrix of connections. Focussing on $\hat{\xi}_{it}$, first we obtain the sample correlation matrix of $\hat{\xi}_t = (\hat{\xi}_{it})$, $\hat{\mathbf{R}}_{\xi} = (\hat{\rho}_{\xi,ij})$ from the residuals of regression (21), where $\hat{\rho}_{\xi,ij} = \hat{\rho}_{\xi,ji} = \hat{\sigma}_{\xi,ij} / \sqrt{\hat{\sigma}_{\xi,ii} \hat{\sigma}_{\xi,jj}}$, and $\hat{\sigma}_{\xi,ij} = T^{-1} \sum_{t=1}^T \hat{\xi}_{it} \hat{\xi}_{jt}$. Next, we apply Holm's multiple testing to the $N(N-1)/2$ pair-wise correlation coefficients, $\hat{\rho}_{\xi,ij}$, for $i = 1, 2, \dots, N-1, j = i+1, \dots, N$, as described in sub-section 3.1. We denote the resultant connection matrix by $\hat{\mathbf{W}}_{cs} = (\hat{w}_{cs,ij})$. Here *cs* stands for multiple testing applied to residuals extracted from de-factoring using the cross-sectional averages approach.

As in sub-section 4.2.1, measuring the degree of sparseness of $\hat{\mathbf{W}}_{cs}$ by the percentage of its non-zero elements we obtain the figure of 1.08% which is comparable to the 1.55% we obtained for \mathbf{W}_{100m} , although as can be seen from Figure 2 the pattern of sparseness of the two matrices, \mathbf{W}_{100m} and $\hat{\mathbf{W}}_{cs}$, are quite different. In fact 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, though distinct clusters are evident especially in the West Coast and parts of the East Coast regions.¹³ Divisions of the connections into the East, the Middle and West of the country are also visible.

¹³Recall that in these graphs MSAs are ordered by State, moving from East to West, from top to bottom and left to right.

Figure 2: Spatial weights matrix using multiple testing and \mathbf{W}_{100m}
De-factoring using cross-sectional averages



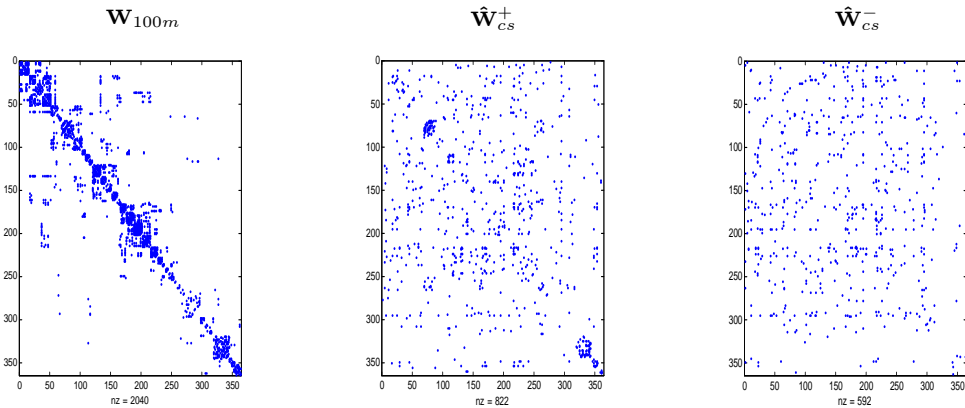
4.2.3 Positive and negative connections

Unlike the distance based \mathbf{W}_d weights matrices, the connections identified by the correlation based approach can be readily distinguished into positive and negative ones. This can be done by associating positive connections with statistically significant evidence of a positive correlation, and the negative connections with the evidence of statistically significant pair-wise negative correlations. Accordingly, we can now define the positively and negatively connected weights matrices, $\hat{\mathbf{W}}_{cs}^+ = (\hat{w}_{ij,cs}^+)$ and $\hat{\mathbf{W}}_{cs}^- = (\hat{w}_{ij,cs}^-)$, respectively, by

$$\hat{w}_{ij,cs}^+ = \hat{w}_{ij,cs} I\left(\hat{\rho}_{\xi,ij} > 0\right), \text{ and } \hat{w}_{ij,cs}^- = \hat{w}_{ij,cs} I\left(\hat{\rho}_{\xi,ij} \leq 0\right).$$

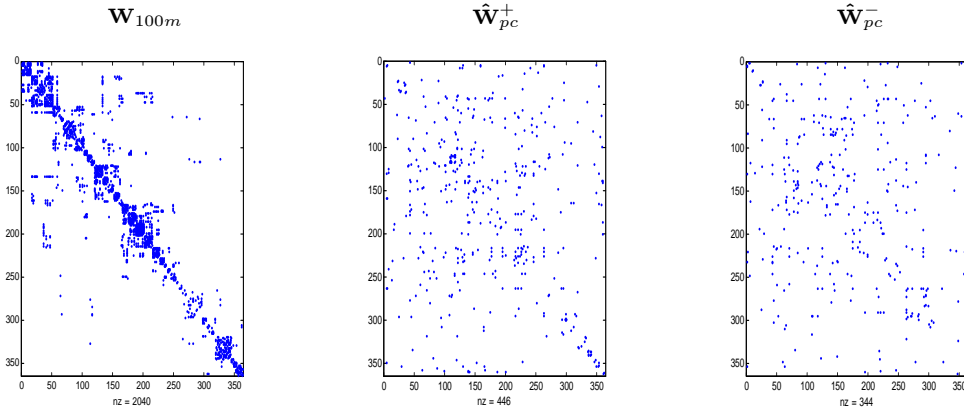
Clearly, $\hat{\mathbf{W}}_{cs} = \hat{\mathbf{W}}_{cs}^+ + \hat{\mathbf{W}}_{cs}^-$. Comparing these with the distance based weights matrix, \mathbf{W}_{100m} , in Figure 3, at first glance we notice that $\hat{\mathbf{W}}_{cs}^+$ is more closely related to \mathbf{W}_{100m} than is $\hat{\mathbf{W}}_{cs}^-$. Further, 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, with a number of clusters suggesting connections at considerable distances.

Figure 3: Spatial weights matrices - distance and correlation-based connections (CS)



We also applied the multiple testing procedure to the de-factored house price changes using the hierarchical factor model, (24), and obtained the weights matrix $\hat{\mathbf{W}}_{pc}$. The degree of sparseness of this matrix, as measured by the percentage of their non-zero elements, is 0.60%. Comparing $\hat{\mathbf{W}}_{pc}$ with the distance-based matrices \mathbf{W}_{50m} , \mathbf{W}_{100m} , and \mathbf{W}_{200m} , it appears that using the hierarchical principal components approach the degree of sparseness of $\hat{\mathbf{W}}_{pc}$ is mid-way between \mathbf{W}_{50m} and \mathbf{W}_{100m} . We also constructed $\hat{\mathbf{W}}_{pc}^+$ and $\hat{\mathbf{W}}_{pc}^-$ in line with the procedure described earlier. $\hat{\mathbf{W}}_{pc}^+$, $\hat{\mathbf{W}}_{pc}^-$ and \mathbf{W}_{100m} are plotted in Figure 4 below. As with the cross-sectional averages approach, the PC method suggests that positively correlated connections match more closely the distance-based connections than do the negatively correlated connections.

Figure 4: Spatial weights matrices - distance and correlation-based connections (PCs)



4.2.4 Statistical associations of different connection weights matrices

We assess the closeness of the correlation-based estimates, $\hat{\mathbf{W}}^+$ and $\hat{\mathbf{W}}^-$ (using either *cs* or *pc* regressions for de-factoring) with the distance-based weight matrix, \mathbf{W}_d , more formally by quantifying the statistical association of the two types of weights matrices. The analysis is complicated by the fact that these matrices are by nature sparse, and hence the probability of a zero realisation in both adjacency matrices $\hat{\mathbf{W}}^+$ (or $\hat{\mathbf{W}}^-$) and \mathbf{W}_d is higher than obtaining a common entry of 1.¹⁴ Given the symmetry of the weights matrices in our application, we focus on the upper triangular elements. We create contingency tables from these upper-triangular elements of the form

$$\begin{pmatrix} n_{11} & n_{10} \\ n_{01} & n_{00} \end{pmatrix},$$

where:

- n_{11} equals the number of times $\hat{\mathbf{W}}^+$ (or $\hat{\mathbf{W}}^-$) displays entry of 1 when \mathbf{W}_d displays 1.
- n_{00} equals the number of times $\hat{\mathbf{W}}^+$ (or $\hat{\mathbf{W}}^-$) displays entry of 0 when \mathbf{W}_d displays 0.

¹⁴For more details regarding testing the dependence among multicategory variables see Pesaran and Timmermann (2009).

- n_{01} equals the number of times $\hat{\mathbf{W}}^+$ (or $\hat{\mathbf{W}}^-$) displays entry of 0 when \mathbf{W}_d displays 1.
- n_{10} equals the number of times $\hat{\mathbf{W}}^+$ (or $\hat{\mathbf{W}}^-$) displays entry of 1 when \mathbf{W}_d displays 0.

Then, $n_{11} + n_{00} + n_{01} + n_{10} = N(N - 1)/2$, and Pearson's chi-squared statistic - Pearson (1900) - is given by

$$\chi^2 = \frac{1}{2}N(N - 1) \left[\sum_{i,j=0}^1 \frac{n_{ij}^2}{(n_{i.} + n_{.j})} - 1 \right].$$

We set the significance level at 5%. We compare $\hat{\mathbf{W}}^+$ (or $\hat{\mathbf{W}}^-$) with the three versions of \mathbf{W}_d , namely \mathbf{W}_{50m} , \mathbf{W}_{100m} , and \mathbf{W}_{200m} . For brevity of exposition we present the contingency tables for $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$ versus \mathbf{W}_{100} only:

Table 2: Contingency tables - $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$ versus \mathbf{W}_{100m} spatial weights matrices

	\mathbf{W}_{100m}			
		1	0	\sum_{rows}
$\hat{\mathbf{W}}_{cs}^+$	1	54	357	411
	0	966	64326	65292
	\sum_{cols}	1020	64683	65703

	\mathbf{W}_{100m}			
		1	0	\sum_{rows}
$\hat{\mathbf{W}}_{cs}^-$	1	8	288	296
	0	1012	64395	65407
	\sum_{cols}	1020	64683	65703

It is clear that $\hat{\mathbf{W}}_{cs}^+$ has more elements in common with \mathbf{W}_d than does $\hat{\mathbf{W}}_{cs}^-$. The $\chi_{5\%}^2$ statistics for $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$ versus \mathbf{W}_{50m} , \mathbf{W}_{100m} , and \mathbf{W}_{200m} respectively are shown in Table 3 below (to be compared with a critical value of 3.84):

Table 3: Pearson's $\chi_{5\%}^2$ test statistics
 $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$ versus \mathbf{W}_d , $d = 50, 100, 200m$

	\mathbf{W}_{50m}	\mathbf{W}_{100m}	\mathbf{W}_{200m}
$\hat{\mathbf{W}}_{cs}^+$	267.24	363.27	298.41
$\hat{\mathbf{W}}_{cs}^-$	0.89	2.57	4.30

The chi squared test statistics are highly significant especially when $\hat{\mathbf{W}}_{cs}^+$ is considered. Elements of $\hat{\mathbf{W}}_{cs}^+$ are much more closely associated with the spatial weights, \mathbf{W}_d , than the elements of $\hat{\mathbf{W}}_{cs}^-$. The association between $\hat{\mathbf{W}}_{cs}^+$ and \mathbf{W}_d is the largest when $d = 100$.

Finally, we repeat these comparisons with weights based on de-factored price changes using the hierarchical PCs, and obtain similar results. See Tables 3 and 4 where $\hat{\mathbf{W}}_{pc}^+$ and $\hat{\mathbf{W}}_{pc}^-$ are compared with \mathbf{W}_d , $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$. The association between $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{pc}^+$ is particularly high, and gives a chi-squared statistic of 9673.7 (compared with a critical value of 3.84).

Table 4: Pearson's $\chi^2_{5\%}$ test statistics

$\hat{\mathbf{W}}_{pc}^+$ and $\hat{\mathbf{W}}_{pc}^-$ versus \mathbf{W}_d , $d = 50, 100, 200m$

	\mathbf{W}_{50m}	\mathbf{W}_{100m}	\mathbf{W}_{200m}
$\hat{\mathbf{W}}_{pc}^+$	86.70	149.55	84.66
$\hat{\mathbf{W}}_{pc}^-$	0.61	0.17	3.75

Table 5: Contingency tables - $\hat{\mathbf{W}}_{cs}^{+/-}$ versus $\hat{\mathbf{W}}_{pc}^{+/-}$ spatial weights matrices

$\hat{\mathbf{W}}_{pc}^+$					$\hat{\mathbf{W}}_{pc}^-$				
		1	0	\sum_{rows}			1	0	\sum_{rows}
$\hat{\mathbf{W}}_{cs}^+$	1	117	106	223	$\hat{\mathbf{W}}_{cs}^-$	1	71	101	172
	0	294	65186	65480		0	225	65306	65531
	\sum_{cols}	411	65292	65703		\sum_{cols}	296	65407	65703

4.3 A heterogeneous spatio-temporal model of US house price changes

Finally, we are in a position to illustrate the utility of the separate identification of positive and negative connections for the spatial analysis of house price changes. We model the de-factored house price changes, $\hat{\xi}_{it}$, using the spatio-temporal model given by (17), with \mathbf{W}^+ and \mathbf{W}^- replaced by their estimates $\hat{\mathbf{W}}^+$ and $\hat{\mathbf{W}}^-$ using de-factored residuals. The estimation results are similar for both approaches to de-factoring, and to save space we focus on the de-factored observations and \mathbf{W} estimates using cross-sectional averages as factors (see (20)). We also restrict the lag orders, h_λ , h_ψ^+ and h_ψ^- , to unity (as they seem adequate for capturing the temporal dynamics), and row-standardise the weight matrices to identify the spatial effects across the MSAs. This yields

$$\hat{\xi}_t = \mathbf{a}_\xi + \mathbf{\Lambda}_1 \hat{\xi}_{t-1} + \mathbf{\Psi}_0^+ \tilde{\mathbf{W}}_{cs}^+ \hat{\xi}_t + \mathbf{\Psi}_0^- \tilde{\mathbf{W}}_{cs}^- \hat{\xi}_t + \mathbf{\Psi}_1^+ \tilde{\mathbf{W}}_{cs}^+ \hat{\xi}_{t-1} + \mathbf{\Psi}_1^- \tilde{\mathbf{W}}_{cs}^- \hat{\xi}_{t-1} + \zeta_t, \quad (25)$$

where $\tilde{\mathbf{W}}_{cs}^+$ and $\tilde{\mathbf{W}}_{cs}^-$ are the scaled (row-standardised when applicable) versions of $\hat{\mathbf{W}}_{cs}^+$ and $\hat{\mathbf{W}}_{cs}^-$, and $\mathbf{a}_\xi = (\alpha_{1\xi}, \alpha_{2\xi}, \dots, \alpha_{N\xi})'$ is the $N \times 1$ vector of intercepts. $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, $\mathbf{\Psi}_0^+ = \text{diag}(\boldsymbol{\psi}_0^+)$, $\mathbf{\Psi}_0^- = \text{diag}(\boldsymbol{\psi}_0^-)$, $\mathbf{\Psi}_1^+ = \text{diag}(\boldsymbol{\psi}_1^+)$, and $\mathbf{\Psi}_1^- = \text{diag}(\boldsymbol{\psi}_1^-)$, where $\boldsymbol{\lambda}_1 = (\lambda_{11}, \lambda_{21}, \dots, \lambda_{N1})'$, $\boldsymbol{\psi}_s^+ = (\psi_{1s}^+, \psi_{2s}^+, \dots, \psi_{Ns}^+)$, $\boldsymbol{\psi}_s^- = (\psi_{1s}^-, \psi_{2s}^-, \dots, \psi_{Ns}^-)$, for $s = 0$ and 1 , and $\zeta_t = (\zeta_{1t}, \zeta_{2t}, \dots, \zeta_{Nt})'$ are the error terms. For quasi-maximum likelihood (QML) estimation of the parameters we assume that $\zeta_{it} \sim \text{IIDN}(0, \sigma_{\zeta_i}^2)$, for $i = 1, 2, \dots, N$. Notice that we continue to have $\tilde{\mathbf{W}}_{cs}^+ + \tilde{\mathbf{W}}_{cs}^- = \tilde{\mathbf{W}}_{cs}$ where $\tilde{\mathbf{W}}_{cs}$ is the row-standardized version of $\hat{\mathbf{W}}_{cs}$, and (25) reduces to the more familiar specification

$$\hat{\xi}_t = \mathbf{a}_\xi + \mathbf{\Lambda}_1 \hat{\xi}_{t-1} + \mathbf{\Psi}_0 \tilde{\mathbf{W}}_{cs} \hat{\xi}_t + \mathbf{\Psi}_1 \tilde{\mathbf{W}}_{cs} \hat{\xi}_{t-1} + \zeta_t, \quad (26)$$

under the restrictions $\mathbf{\Psi}_j^+ = \mathbf{\Psi}_j^- = \mathbf{\Psi}_j$, for all $j = 0$ and 1 , a restriction that can be tested within our framework.

The model specification in (25) allows for a high degree of heterogeneity in dynamics and spatial dependence across the 363 MSAs. By comparison, the traditional spatial model assumes that the dynamic and spatial coefficients are homogeneous, namely $\lambda_{1i} = \lambda_1$, $\psi_{is}^+ = \psi_s^+$, $\psi_{is}^- = \psi_s^-$, for $s = 0$ and 1, and for all i . The only forms of heterogeneity allowed in the literature are intercept and error variance heterogeneity. Furthermore, the existing spatial literature assumes that all units in \mathbf{W}_d have at least one (positive) neighbour - see Kelejian and Prucha (1999, 2010) among others. This need not always hold. When applying the multiple testing procedure to the de-factored observations we find a relatively small number of units, $N_0 = 76$ (out of 363), that are not connected to the remaining MSAs. There are also a number of MSAs with only negative connections, $N_- = 34$, and a number with only positive connections, $N_+ = 90$, with the remaining $N_{+/-} = 163$ MSAs having both positive and negative connections, so that $N = N_{+/-} + N_- + N_+ + N_0 = 363$. The distribution of connections by the eight regions are given in Table 6.

Table 6: 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
<i>Proportional to total no. of MSAs per region</i>					
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.

It is clear that, for the most part, MSAs in all regions have both positive and negative connections. Also, more MSAs have exclusively positive connections than only negative connections across all regions, the most polarised regions being the South East and the Far West. On the other hand, a more balanced distribution of MSAs across N_- and N_+ can be seen for the Plains, South West and Rocky Mountains regions, with the latter two also having a proportionately larger number of MSAs with no connections at all.

To estimate ψ_{i0}^+ and ψ_{i0}^- , for $i = 1, 2, \dots, N$, we use the concentrated log-likelihood function (18), in which $\tilde{\mathbf{x}}_i$ is replaced by $\hat{\tilde{\mathbf{x}}}_i$, and \mathbf{Z}_i is augmented by a $T \times 1$ vector of ones, $\boldsymbol{\tau}_T$. The intercepts, $a_{i\xi}$, and the parameters of the lagged values, $\lambda_{i1}, \psi_{i1}^+$ and

ψ_{i1}^- can be estimated via least squares applied to the equations for individual MSAs conditional on the ML estimates of ψ_{i0}^+ and ψ_{i0}^- , for $i = 1, 2, \dots, N$. We also note that for the N_0 units with no connections, we set $\psi_{i0}^+ = \psi_{i0}^- = \psi_{i1}^+ = \psi_{i1}^- = 0$, and estimate the remaining parameters, intercepts and λ_{i1} by least squares. For MSAs with only negative or positive connections we impose the restriction that the corresponding ψ_{i0}^+ , ψ_{i1}^+ , ψ_{i1}^- , and ψ_{i0}^- coefficients are set to zero. This restriction is needed for identification purposes due to the simultaneity problem that arises in this case. Clearly these coefficients can be set to other values as well, such as the average of each coefficient within the region they belong to or even to the national average of each coefficient.

For the purpose of inference, the variance-covariance matrix of the estimates is computed as in (19), again using the full log-likelihood with respect to the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \dots, \boldsymbol{\theta}'_N)'$, where $\boldsymbol{\theta}_i = (\psi_{i0}^+, \psi_{i0}^-, \psi_{i1}^+, \psi_{i1}^-, \alpha_{i\xi}, \lambda_{i1}, \sigma_{\zeta_i}^2)'$. For further details, see Aquaro, Bailey and Pesaran (2014).

QML estimates for individual MSAs are available on request. In what follows we give median and mean estimates, and the proportion of MSAs with statistically significant parameters (at the 5% level). The results are summarized in Table 7, and give the median and the mean group (MG) estimates of $\hat{\lambda}_{i1}$, $\hat{\psi}_{i0}^+$, $\hat{\psi}_{i0}^-$, $\hat{\psi}_{i1}^+$, $\hat{\psi}_{i1}^-$ and $\hat{\sigma}_{\zeta_i}$. The standard errors of the MG estimates are given in parenthesis.¹⁵

A number of general conclusions readily emerge from an examination of the results in Table 7. The mean and median estimates are very close suggesting that the estimates across the MSAs are approximately symmetrically distributed. All the mean estimates are statistically significant at the 5% level, with the mean lagged spatial effects ($\hat{\psi}_{1,MGE}^+$ and $\hat{\psi}_{1,MGE}^-$) being less precisely estimated as compared to the other mean effects, namely $\hat{\lambda}_{1,MGE}$, $\hat{\psi}_{0,MGE}^+$ and $\hat{\psi}_{0,MGE}^-$. The size of the mean temporal effect, $\hat{\lambda}_{1,MGE}$, at 0.392 (0.009) is reasonably large considering that de-factoring is likely to have removed some of the common dynamics in house price changes. With regard to the cross section dynamics, contemporaneous positive spill-over effects have a larger magnitude than their equivalent negative effects with the MG estimates of ψ_{i0}^+ and ψ_{i0}^- given by $\hat{\psi}_{0,MGE}^+ = 0.345$ (0.017) and $\hat{\psi}_{0,MGE}^- = -0.2763$ (0.021), respectively. The estimates in both cases are correctly signed and clearly reject the hypothesis that $\boldsymbol{\Psi}_j^+ = \boldsymbol{\Psi}_j^-$, for $j = 0$ (and for $j = 1$ as discussed below). Also, the MG estimates of the lagged spatial effects, $\hat{\psi}_{1,MGE}^+ = -0.040$ (0.015) and $\hat{\psi}_{1,MGE}^- = 0.071$ (0.016), are opposite in sign to the respective contemporaneous effects, suggesting a certain degree of reversal of the effects. We also note that the mean spatial effects of positive connections at 0.345 is somewhat higher than the mean effects of negative connections at 0.276 (in absolute terms).

With regard to the statistical significance of the estimates for individual MSAs (abstracting from multiple testing issues) we note that $\hat{\lambda}_{i1}$ is statistically significant in 90% of the MSAs, whilst the contemporaneous spatial effects is significant in 65% of MSAs with positive connections, and significant in 62% of MSAs with negative connections. In

¹⁵The MG estimator is defined as the simple average of the estimates across the MSAs with non-zero coefficients. For example, the MGE of $E(\psi_{i0}^+) = \psi_0^+$, is given by $\hat{\psi}_{0,MGE}^+ = (1/N_+^*) \sum_{i=1}^{N_+^*} \hat{\psi}_{i0}^+$, where N_+^* denotes the number of MSAs with positive connections ($N_+^* = N_{+/-} + N_+$), and $\hat{\psi}_{i0}^+$ is the QMLE of ψ_{i0}^+ . The non-parametric estimator of the variance of $\hat{\psi}_{0,MGE}^+$ is given by: $\widehat{Var}(\hat{\psi}_{0,MGE}^+) = \frac{1}{N_+^*(N_+^*-1)} \sum_{i=1}^{N_+^*} (\hat{\psi}_{i0}^+ - \hat{\psi}_{0,MGE}^+)^2$. For more details see Pesaran and Smith (1995).

contrast, the lagged spatial effects turned out to be much weaker, with only 28 present of positive connections and 26% of negative connections being statistically significant. Overall, the estimates suggest there exists a reasonably rich temporal and cross-sectional dependence in US house price changes even after stripping them of strong, pervasive national and regional factors.

Table 7: Quasi-ML estimates of spatio-temporal model (25)
Applied to de-factored house price changes given by (21)

	λ_1	ψ_0^+	ψ_0^-	ψ_1^+	ψ_1^-	σ_ζ
	<i>Computed over non-zero parameter coefficients</i>					
Median	0.3986	0.3124	-0.2493	-0.0430	0.0608	1.2416
Mean Group Estimates	0.3921 (0.0086)	0.3454 (0.0168)	-0.2763 (0.0209)	-0.0398 (0.0147)	0.0706 (0.0156)	1.3056 (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

¹Restricted 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 $i = 1, 2, \dots, 363$.

²MGE standard errors are in brackets.

To give an idea of the extent of parameter heterogeneity across MSAs, in Table 8 we provide median and mean estimates by regions. Interestingly enough the regional differences are not very pronounced, particularly if we focus on the more precisely estimated parameters. The regional estimates of λ_1 range from the low value of 0.316 (0.021) for the Great Lakes to a high value of 0.458 (0.025) for the Rocky Mountains. The regional differences in the mean estimates of ψ_0^+ are even slightly lower and range from 0.264 (0.082) in the South West to 0.374 (0.042) for the Plains. In contrast, the estimates of the negative connections, ψ_0^- are less precisely estimated and range from -0.078 (0.099) for New England to -0.370 (0.053) in the South West. But one should consider such comparisons with care since in the case of some regions the number of non-zero estimates was quite small. Nevertheless, one of our main conclusions that positive and negative connections have opposite effects seems to be robust to the regional disaggregation. The estimates of ψ_0^+ and ψ_0^- are respectively positive and negative across all regions. The results in Table 8 also support our conclusion that lagged spatial effects are generally not that important and tend to be statistically insignificant in a number of regions. But once again we need to bear in mind that some of the regional estimates are based on a rather small number of non-zero estimates.

Finally, to assess the importance of de-factoring of house price changes we also estimated the connection matrices $\hat{\mathbf{W}}^+$ and $\hat{\mathbf{W}}^-$ without de-factoring, using the hierarchical factor model (20). Not surprisingly we found $\hat{\mathbf{W}}^+$ to be much denser as compared to the estimates obtained based on de-factored price changes, and $\hat{\mathbf{W}}^-$ to be less dense. The many more connections that we are finding when using price changes without de-factoring reflect the presence of common factors rather than genuine spatial effects. In line with this result we also find an estimate of spatial effects, ψ_0^+ , which is very close to unity when we use a estimated $\hat{\mathbf{W}}^+$ based on non-defactored price changes. Details of these results are available upon request.

Table 8: Quasi-ML estimates of spatio-temporal model (25) summarised by region
Applied to de-factored house price changes given by (21)

	<i>Computed over non-zero parameter coefficients</i>					
	λ_1	ψ_0^+	ψ_0^-	ψ_1^+	ψ_1^-	σ_ζ
<i>New England</i>						
Median	0.4064	0.2762	-0.0843	-0.0514	0.0209	1.1684
Mean Group Estimates	0.3944	0.3563	-0.0781	-0.0050	-0.0412	1.2704
	(0.0303)	(0.0996)	(0.0991)	(0.0430)	(0.0784)	(0.0966)
% significant (5% level)	86.7%	60.0%	50.0%	0.0%	30.0%	-
Number of non-zero coef.	15	10	10	10	10	10
<i>Mid East</i>						
Median	0.4278	0.3439	-0.1904	-0.0096	0.0625	1.3977
Mean Group Estimates	0.3990	0.3603	-0.1938	-0.0755	0.1129	1.4368
	(0.0319)	(0.0465)	(0.1163)	(0.0487)	(0.0747)	(0.0634)
% significant (5% level)	91.7%	65.4%	68.4%	30.8%	26.3%	-
Number of non-zero coef.	36	26	19	26	19	36
<i>South East</i>						
Median	0.4013	0.3242	-0.2686	-0.0538	0.0847	1.2384
Mean Group Estimates	0.4001	0.3563	-0.3062	-0.0596	0.0977	1.3469
	(0.0162)	(0.0262)	(0.0326)	(0.0234)	(0.0242)	(0.0427)
% significant (5% level)	90.4%	64.8%	61.6%	27.3%	31.5%	-
Number of non-zero coef.	114	88	73	88	73	114
<i>Great Lakes</i>						
Median	0.3176	0.2660	-0.2227	0.0149	0.0361	1.2492
Mean Group Estimates	0.3160	0.3304	-0.2846	0.0229	0.0407	1.3142
	(0.0209)	(0.0463)	(0.0383)	(0.0435)	(0.0351)	(0.0392)
% significant (5% level)	78.7%	63.4%	50.0%	31.7%	13.9%	-
Number of non-zero coef.	61	41	36	41	36	61
<i>Plains</i>						
Median	0.3808	0.3015	-0.2491	-0.1573	0.0597	1.1128
Mean Group Estimates	0.3751	0.3744	-0.2409	-0.1280	0.0825	1.1254
	(0.0243)	(0.0427)	(0.0540)	(0.0290)	(0.0421)	(0.0324)
% significant (5% level)	93.8%	75.0%	57.1%	29.2%	28.6%	-
Number of non-zero coef.	32	24	21	24	21	32
<i>South West</i>						
Median	0.3935	0.2944	-0.3053	-0.1023	0.0077	1.2877
Mean Group Estimates	0.4024	0.2642	-0.3695	-0.0576	0.0377	1.3385
	(0.0209)	(0.0823)	(0.0525)	(0.0630)	(0.0560)	(0.0301)
% significant (5% level)	94.7%	57.1%	82.4%	38.1%	23.5%	-
Number of non-zero coef.	38	21	17	21	17	38
<i>Rocky Mountains</i>						
Median	0.4435	0.3486	-0.2756	0.0155	0.1396	1.1618
Mean Group Estimates	0.4581	0.3177	-0.3086	0.0083	0.1033	1.2096
	(0.0253)	(0.0667)	(0.0542)	(0.0430)	(0.0557)	(0.0409)
% significant (5% level)	100.0%	70.0%	80.0%	10.0%	40.0%	-
Number of non-zero coef.	22	10	10	10	10	22
<i>Far West</i>						
Median	0.4672	0.3667	-0.2438	0.0330	0.0480	1.2158
Mean Group Estimates	0.4400	0.3591	-0.2673	0.0137	0.0155	1.2437
	(0.0237)	(0.0488)	(0.0898)	(0.0428)	(0.0293)	(0.0336)
% significant (5% level)	91.1%	63.6%	63.6%	30.3%	18.2%	-
Number of non-zero coef.	45	33	11	33	11	45

¹Restricted 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 $i = 1, 2, \dots, 363$.

²MGE standard errors are in brackets below their respective Mean Group Estimates.

5 Conclusions

An understanding of the spatial dimension of economic and social activity requires methods that can separate out the relationship between spatial units that is due to the effect of common factors from that which is purely spatial. We are able to distinguish between cross-sectional strong dependence and weak or spatial dependence. Strong dependence in turn suggests that there are common factors. We have proposed the use of cross unit averages to extract common factors and contrast this to a principal components approach widely used in the literature. We then use multiple testing to determine significant bilateral correlations (signifying connections) between spatial units and compare this to an approach that just uses distance to determine units that are neighbours. In a very data rich environment with observations on many spatial units over long periods of time a way of filtering the data to uncover spatial connections is crucial. We have applied these methods to real house price changes in the US at the level of the Metropolitan Statistical Area. Although there is considerable overlap between neighbours determined by distance and those by multiple testing, there is also considerable correlation between MSAs across the United States that suggests that other forces are at work.

We also find that our analysis of connections based on pair-wise correlations of de-factored house price changes clearly points to the existence of negative as well as positive connections. This feature is absent if we base the spatial analysis exclusively on contiguity. It is common in the literature to think of spatial relationships as involving spillover from one area to another with the (implicit) assumption that the spillover effects are positive. But this need not be the case. Migration across space could raise/lower wages or house prices in one locality and lower/raise them into another locality.

Furthermore, we verify that basing the spatial analysis on house price changes without de-factoring ignores the possibility that there may be common national and regional factors and failing to condition on the common factors may bias the inferences that can be drawn. Our analysis strips out such common effects and allows us to focus on estimation of spillover effects (positive or negative) which is of primary interest in spatial analysis. Although proximity measured by distance is a useful metric for constructing the weights matrix, our analysis suggests that correlation analysis, once applied to de-factored price changes with appropriate application of multiple testing techniques can lead to important new insights as to the nature of spatial connections.

Appendices

Appendix I: Data sources

Monthly data for US house prices from January 1975 to December 2010 are taken from Freddie Mac. These data are available at:

<http://www.freddiemac.com/finance/fmhpi>

The quarterly figures are arithmetic averages of monthly figures.

Annual CPI data at State level are obtained from the Bureau of Labor Statistics:

<http://www.bls.gov/cpi/>

The quarterly figures are interpolated using the interpolation technique described in the appendix of GVAR toolbox 1.1 user guide.

The annual population data at MSA level are obtained from the Regional Economic Information System, Bureau of Economic Analysis, U.S. Department of Commerce:

<http://www.bea.gov/regional/docs/footnotes.cfm?tablename=CA1-3>

Appendix II: Geographical classification of the United States

Table A provides the geographical breakdown of the eight regions used in our analysis of US house price changes which is based on the Bureau of Economic Analysis classification (<http://www.bea.gov/regional/docs/regions.cfm>). Each region covers an average of 6 States, each of which contains 45 Metropolitan Statistical Areas on average. The classifications are shown in Table A together with the number of MSAs included in each State. Details of the MSAs by region are available upon request.

Appendix III: Calculation of distance

The original data used were Latitude-Longitude of zip codes, cross referenced with each of the 366 Metropolitan Statistical Areas (MSAs). Any missing Latitude-Longitude coordinates were coded manually from Google searches. The geodesic distance between a pair of latitude/longitude coordinates was then calculated using the Haversine formula:

$$\begin{aligned} a &= \sin^2 \left(\frac{\Delta lat}{2} \right) + \cos(lat1) \cos(lat2) \sin^2 \left(\frac{\Delta long}{2} \right), \\ c &= 2a \tan 2 \left(\sqrt{a}, \sqrt{1-a} \right), \\ d &= Rc, \end{aligned}$$

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

Table A: Geographical Classification of Regions, States and MSAs in the United States

Regions	States	No of MSAs	Regions	States	No of MSAs	Regions	States	No of MSAs
New England	Connecticut (CT)	4	Great Lakes	Illinois (IL)	9	South West	Arizona (AZ)	6
	Maine (ME)	3		Indiana (IN)	13		New Mexico (NM)	4
	Massachusetts (MA)	5		Michigan (MI)	14		Oklahoma (OK)	3
	New Hampshire (NH)	1		Ohio (OH)	13		Texas (TX)	25
	Rhode Island (RI)	1		Wisconsin (WI)	12			
	Vermont (VT)	1						
Mid-East	Delaware (DE)	1	Plains	Iowa (IA)	8	Rocky Mountains	Colorado (CO)	7
	District of Columbia (DC)	1		Kansas (KS)	4		Idaho (ID)	5
	Maryland (MD)	4		Minnesota (MN)	5		Montana (MT)	3
	New Jersey (NJ)	4		Missouri (MO)	8		Utah (UT)	5
	New York (NY)	12		Nebraska (NE)	2		Wyoming (WY)	2
	Pennsylvania (PA)	14		North Dakota (ND)	3			
				South Dakota (SD)	2			
	Alabama (AL)	11				Far West	California (CA)	26
	Arkansas (AR)	6					Nevada (NV)	3
	Florida (FL)	20					Oregon (OR)	6
South East	Georgia (GA)	14					Washington (WA)	10
	Kentucky (KY)	5	Far West			Far West		
	Louisiana (LA)	8						
	Mississippi (MS)	4						
	North Carolina (NC)	14						
	South Carolina (SC)	8						
	Tennessee (TN)	10						
	Virginia (VA)	9						
	West Virginia (WV)	5						

This geographical classification of the United States follows the Bureau of Economic Analysis specifications.

References

- [1] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Boston: Kluwer Academic Publishers.
- [2] Anselin, L. (2001). *Spatial Econometrics*. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics*. Blackwell, Oxford.
- [3] Aquaro, M., Bailey, N. and Pesaran, M. H. (2014). Quasi-Maximum Likelihood Estimation of Spatial Models with Heterogenous Coefficients. Under preparation.
- [4] Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71, 1, 135–171.
- [5] Bai, J. and Ando, (2014). Panel Data Models with Grouped Factor Structure under Unknown Group Membership. MPRA Paper No. 52782, January 2014 05:46 UTC.
- [6] Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70, 191–221.
- [7] Bailey, N., Kapetanios, G. and Pesaran, H. M. (2014). Exponent of Cross-Sectional Dependence: Estimation and Inference. Jan 2012, revised Apr 2014. Cambridge Working Paper 1206.
- [8] Bailey, N., Pesaran, H. M. and Smith, L.V. (2014). A Multiple Testing Approach to the Regularisation of Sample Correlation Matrices. CESifo Working Paper no. 4834.
- [9] Barigozzi, M. and Brownlees, C. (2013). NETS: Network Estimation for Time Series. SSRN Working Paper. Available at SSRN: <http://ssrn.com/abstract=2249909> or <http://dx.doi.org/10.2139/ssrn.2249909>.
- [10] Bester, C. A., Conley, T. G. and Hansen, C. B. (2011). Inference with Dependent Data Using Cluster Covariance Estimators. *Journal of Econometrics*, 165, 2, 137–151.
- [11] Bhattachajee, A. and Holly, S. (2013). Understanding Interactions in Social Networks and Committees. *Spatial Economic Analysis*, 8, 1, 23–53.
- [12] Bien, J. and Tibshirani, R. J. (2011). Sparse Estimation of the Covariance Matrix. *Biometrika*, 98, 4, 807–820.
- [13] Bonferroni, C. (1935). Il Calcolo delle Assicurazioni su Gruppi di Teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, Rome, Italy, 13–60.
- [14] Bonferroni, C. (1936). Teoria Statistica delle Classi e Calcolo delle Probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62.
- [15] Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. and Kohane, I. S. (2000). Discovering Functional Relationships between RNA Expression and Chemotherapeutic Susceptibility using Relevance Networks. *Proceedings of the National Academy of Sciences of the U.S.A.*, 97, 12182–12186.

- [16] Carlino, G. A. and DeFina, R. H. (1998). The Differential Regional Effects of Monetary Policy. *Review of Economics and Statistics*, 80, 4, 572–587.
- [17] Carlino, G. A. and DeFina, R. H. (2004). How Strong Is Co-movement in Employment over the Business Cycle? Evidence from State/Sector Data. *Journal of Urban Economics*, 55, 2, 298–315.
- [18] Carlino, G. and Sill, K. (2001). Regional Income Fluctuations: Common Trends and Common Cycles. *Review of Economics and Statistics*, 83, 3, 446–456.
- [19] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets. *Econometrica*, 51, 1305–1324.
- [20] Chaudhuri, S., Drton, M. and Richardson, T. S. (2007). Estimation of a Covariance Matrix with Zeros. *Biometrika*, 94, 199–216.
- [21] Chudik, A., Pesaran, M. H. and Tosetti, E. (2011). Weak and Strong Cross Section Dependence and Estimation of Large Panels. *The Econometrics Journal*, 14, C45–C90.
- [22] Chudik, A. and Pesaran, M. H. (2014). Large Panel Data Models with Cross-Sectional Dependence: A Survey. Center for Applied Financial Economics (CAFE) Research Paper no. 13.15. In B. H. Baltagi (Ed.), forthcoming in *The Oxford Handbook on Panel Data*. Oxford University Press.
- [23] Conley, T. G. (1999). GMM Estimation with Cross-Sectional Dependence. *Journal of Econometrics*, 92, 1–45.
- [24] Conley, T. G. and Dupor, B. (2003). A Spatial Analysis of Sectoral Complementarity. *Journal of Political Economy*, 111, 311–352.
- [25] Conley, T. G. and Topa, G. (2003). Identification of Local Interaction Models with Imperfect Location Data. *Journal of Applied Econometrics*, 18, 605–618.
- [26] Connor, G. and Korajczyk, R. (1993). A Test for the Number of Factors in an Approximate Factor Model. *The Journal of Finance*, XLVIII, 4, 1263–1291.
- [27] Corrado, L. and Fingleton, B. (2012). Where is the Economics in Spatial Econometrics? *Journal of Regional Science*, 52, 210–239.
- [28] Cromwell B. (1992). Does California Drive the West? An Econometric Investigation of Regional Spillovers. *Economic Review of the Federal Reserve Bank of San Francisco*, 2, 13–23.
- [29] Dees, S., di Mauro, F., Pesaran, M. H. and Smith, L. V. (2007). Exploring the International Linkages of the Euro Area: A Global VAR Analysis. *Journal of Applied Econometrics*, 22, 1–38.
- [30] Del Negro, M. (2002). Asymmetric Shocks Among U.S. States. *Journal of International Economics*, 56, 2, 273–297.
- [31] Dempster, A. P. (1972). Covariance Selection. *Biometrics*, 28, 157–175.

- [32] Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction, Cambridge.
- [33] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalised Dynamic Factor Model: Identification and Estimation. *Review of Economics and Statistics*, 82, 540–554.
- [34] Forni, M. and Lippi, M. (2001). The Generalised Dynamic Factor Model: Representation Theory. *Econometric Theory*, 17, 1113–1141.
- [35] Forni, M. and Reichlin, L. (1998). Let’s Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics. *Review of Economic Studies*, 65, 453–473.
- [36] Glaeser, E. L. and Gottlieb, J. D. (2009). The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States. *Journal of Economic Literature*, 47, 983–1028.
- [37] Glaeser, E. L., Gyourko, J. and Saiz, A. (2008). Housing Supply and Housing Bubbles. *Journal of Urban Economics*, 64, 2, 198–217.
- [38] Gupta, R. and Das, S. (2010). Predicting Downturns in the US Housing Market. *Journal of Real Estate Economics and Finance*, 41, 3, 294–319.
- [39] Gupta, R., Kabundi, A. and Miller, S. M. (2011a). Using Large Data Sets to Forecast Housing Prices: A Case Study of Twenty US States. *Journal of Housing Research*, 20, 2, 161–190.
- [40] Gupta, R., Kabundi, A. and Miller, S. M. (2011b). Forecasting the US Real House Price Index: Structural and Non-structural Models with and without Fundamentals. *Economic Modelling*, 28, 4, 2013–2021.
- [41] Gupta R. and Miller, S. M. (2012). The Time Series Properties on Housing Prices: A Case Study of the Southern California Market. *Journal of Real Estate Finance and Economics*, 44, 3, 2012.
- [42] Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*, 75, 800–802.
- [43] Holly, S., Pesaran, H. M. and Yamagata, T. (2011a). A Spatio-Temporal Model of House Prices in the US. *Journal of Econometrics*, 158, 160–173.
- [44] Holly, S., Pesaran, H. M. and Yamagata, T. (2011b). The Spatial and Temporal Diffusion of House Prices in the UK. *Journal of Urban Economics*, 69, 1, 2–23.
- [45] Holly, S. and Petrella, I. (2012). Factor Demand Linkages, Technology Shocks and the Business Cycle. *Review of Economics and Statistics*, 94, 4, 948–963.
- [46] Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6, 2, 65–70.
- [47] Hommel, G. (1988). A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni test. *Biometrika*, 75, 383–386.

- [48] Hommel, G. (1989). A Comparison of Two Modified Bonferroni procedures. *Biometrika*, 76, 624–625.
- [49] Hotelling, H. (1933). Analysis of a Complex of Statistical Variables with Principal Components. *Journal of Educational Psychology*, 24, 6, 417–441.
- [50] Hsiao, C. and Pesaran, M. H. (2008). Random Coefficient Models. Chapter 6 in L. Matyas and Sevestre (eds.), *The Econometrics of Panel Data*, Springer-Verlag, Berlin, 185–213.
- [51] Kadiyala, K.R. and Bhattacharya, R. (2009). Regional Housing Prices in the USA: An Empirical Investigation of Nonlinearity. *The Journal of Real Estate Finance and Economics*, 38, 4, 443–460.
- [52] Kapetanios, G. and Pesaran, M. H. (2007). Alternative Approaches to Estimation and Inference in Large Multifactor Panels: Small Sample Results with an Application to Modelling of Asset Return. In Garry Phillips and Elias Tzavalis (eds.), *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*, Cambridge University Press, Cambridge.
- [53] Kapoor, M., Kelejian, H. H. and Prucha, I. R. (2007). Panel Data Models with Spatially Correlated Error Components. *Journal of Econometrics*, 140, 97–130.
- [54] Khare, K. and Rajaratnam, B. (2011). Wishart Distributions for Decomposable Covariance Graph Models. *Annals of Statistics*, 30, 514–555.
- [55] Kelejian, H. H. and Prucha, I. R. (1999). A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review*, 40, 509–533.
- [56] Kelejian, H. H. and Prucha, I. R. (2007). HAC Estimation in a Spatial Framework. *Journal of Econometrics*, 140, 131–154.
- [57] Kelejian, H. H. and Prucha, I. R. (2010). Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances. *Journal of Econometrics*, 157, 53–67.
- [58] Krugman, P. R. (1991). Increasing Returns and Economic Geography. *Journal of Political Economy*, 99, 483–499.
- [59] Kuethe, T. and Pedde, V. (2011). Regional Housing Price Cycles: A Spatio-Temporal Analysis Using US State-Level Data. *Regional Studies*, 45, 5, 563–574.
- [60] Lee, L.-F. (2002). Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models. *Econometric Theory*, 18, 2, 252–277.
- [61] Lee, L.-F. (2004). Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica*, 72, 6, 1899–1925.
- [62] Lee, L.-F. and Yu, J. (2010). Estimation of Spatial Autoregressive Panel Data Models with Fixed Effects. *Journal of Econometrics*, 154, 165–185.

- [63] LeSage, J. and Pace R. K. (2010). Introduction to Spatial Econometrics, Chapman & Hall/CRC, Boca Raton, Florida.
- [64] Lin, X. and Lee, L.-F. (2010). GMM Estimation of Spatial Autoregressive Models with Unknown Heteroskedasticity. *Journal of Econometrics*, 157, 1, 34–52.
- [65] Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso. *Annals of Statistics*, 34, 3, 1436–1462.
- [66] Owyang, M. T., Piger, J. and Wall, H. J. (2005). Business Cycle Phases in U.S. States. *Review of Economics and Statistics*, 87, 4, 604–616.
- [67] Partridge, M. D. and Rickman, D. S. (2005). Regional Cyclical Asymmetries in an Optimal Currency Area: An Analysis Using US State Data. *Oxford Economic Papers*, July, 57, 3, 373–397.
- [68] Pearson, K. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *Philosophical Magazine Series*, 5, 50 (302), 157–175.
- [69] Peng, J., Wang, W., Zhou, N. and Zhu, J. (2009). Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association*, June, 104, 486, 735–746.
- [70] Pesaran, M. H. (2004). General Diagnostic Tests for Cross Section Dependence in Panels. CESifo Working Paper Series No. 1229. Available at SSRN: <http://ssrn.com/abstract=572504>.
- [71] Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica*, 74, 4, 967–1012.
- [72] Pesaran, M. H. (2014). Testing Weak Cross-Sectional Dependence in Large Panels. Forthcoming in *Econometric Reviews*.
- [73] Pesaran, M. H., Schuermann, T. and Weiner, S. M. (2004). Modeling Regional Interdependencies using a Global Error-Correcting Macroeconomic Model. *Journal of Business and Economics Statistics*, 22, 2, 129–162.
- [74] Pesaran, M. H. and Smith, R. P. (1995). Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. *Journal of Econometrics*, 68, 79–113.
- [75] Pesaran, M. H. and Timmermann, A. (2009). Testing Dependence Among Serially Correlated Multicategory Variables. *Journal of American Statistical Association*, 104, 485, 325–337.
- [76] Pollakowski H. O. and Ray T. S. (1997). Housing Price Diffusion Patterns at Different Aggregation Levels: An Examination of Housing Market Efficiency. *Journal of Housing Research*, 8, 1, 107–124.
- [77] Rapach, D. E. and Strauss, J. K. (2007). Forecasting Real Housing Price Growth in the Eighth District States. *Federal Reserve Bank of St. Louis. Regional Economic Development*, 3, 2, 33–42.

- [78] Rapach, D. E. and Strauss, J. K. (2009). Differences in Housing Price Forecast Ability Across U.S. States. *International Journal of Forecasting*, 25, 2, 351–372.
- [79] Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*, 2, 494–515.
- [80] Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2010). Sparse Multivariate Regression with Covariance Estimation. *Journal of Computational and Graphical Statistics*, 19, 4, 947–962.
- [81] Stock, J. H. and Watson, M. W. (1998). Diffusion Indexes. NBER Working Papers 6702, *National Bureau of Economic Research*, Inc.
- [82] Stock, J. H. and Watson, M. W. (2002a). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97, 460, 147–162.
- [83] Stock, J. H. and Watson, M. W. (2002b). Macroeconomics Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, 20, 2, 147–162.
- [84] Stone, R. (1947). On the Interdependence of Blocks of Transactions. *Journal of the Royal Statistical Society* (Supplement), 9, 1, 1–45.
- [85] Whittle, P. (1954). On Stationary Processes on the Plane. *Biometrika*, 41, 434–449.
- [86] Yu, J., de Jong, R. M. and Lee, L.-F. (2008). Quasi-Maximum Likelihood Estimators for Spatial Dynamic Panel Data with Fixed Effects when both N and T are Large. *Journal of Econometrics*, 146, 1, 118–134.