Modelling the Cross-section Dependence, the Spatial Heterogeneity and the Network Diffusion in the Multi-dimensional Dataset

Yongcheol Shin, University of York
Workshop at University of Bari

18-21 September, 2018 (Preliminary and incomplete)

1 Introduction

Cross-section dependence (CSD) seems pervasive in panels, since it seems rare that the covariance of the errors is zero. Recently, there has been much progress in characterising and modelling CSD. Phillips and Sul (2003) note that the consequences of ignoring CSD can be serious: pooling may provide little gain in efficiency over single equation estimation; estimates may be badly biased and tests for unit roots and cointegration may be misleading.

CSD has always been central in spatial econometrics (Baltagi, 2005) where there is a natural way to characterise dependence in terms of distance, but for most economic and social science problems there is no obvious distance measure. For instance, trade between countries reflects not just geographical distance, but transport costs, common language, policy and historical factors such as colonial links as well as the multilateral barriers (Anderson and van Wincoop, 2003). For large T, it is straightforward to test for cross-section dependence using the squared correlations between the residuals (e.g. Pesaran, 2015). See also Pesaran, Ullah and Yamagata (2007) for the survey of the various tests.

2 Overview on Cross-section Dependence (CSD)

We consider a generic panel data model advanced by Serlenga an Shin (2007):

$$y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \boldsymbol{\lambda}'\boldsymbol{z}_i + \boldsymbol{\pi}_i'\boldsymbol{s}_t + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{1}$$

where $\boldsymbol{x}_{it} = (x_{1,it}, ..., x_{k,it})'$ is a $k \times 1$ vector of variables that vary across individuals and over time periods, $\mathbf{s}_t = (s_{1t}, ..., s_{st})'$ is an $s \times 1$ vector of observed time-specific factors, $\mathbf{z}_i = (z_{1i}, ..., z_{gi})'$ is a $g \times 1$ vector of individual-specific variables, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_g)'$ and $\boldsymbol{\pi}_i = (\pi_{1i}, ..., \pi_{si})'$ are conformably defined column vectors of parameters.

To address the heterogeneous individual effects and common time effects, we consider the following one-way and two-way error components specifications:

$$\varepsilon_{it} = \alpha_i + u_{it} \tag{2}$$

$$\varepsilon_{it} = \alpha_i + \theta_t + u_{it} \tag{3}$$

However, in the presence of cross-section correlations among ε_{it}'s, the conventional panel data estimators such as FE, tend to become biased. To control for CSD, the most popular approach is to add heterogeneous factors:

$$\varepsilon_{it} = \alpha_i + \boldsymbol{\gamma}_i'\boldsymbol{f}_t + u_{it}, \tag{4}$$

where α_i is an unobserved individual effect (heterogeneity), $\boldsymbol{f}_t$ is the $c \times 1$ vector of unobserved common factors with heterogeneous parameter vector, $\boldsymbol{\gamma}_i = (\gamma_{1i}, ..., \gamma_{ci})'$, and u_{it} is a zero mean idiosyncratic uncorrelated random disturbance. Notice that both α_i and $\boldsymbol{f}_t$ might be correlated with explanatory variables $\mathbf{x}_{it}$ and $\mathbf{z}_i$.

The distinguishing feature of this model is that it allows for observed and unobserved time effects both of which are cross-sectionally correlated. Factors are expected to provide good proxy for any remaining complex time-varying patterns associated with multilateral resistance and globalisation trends, e.g. Mastromarco et al. (2016). Notice that the cross-section dependence in (1) is explicitly allowed through heterogeneous loadings, $\boldsymbol{\gamma}_i$, see Pesaran (2006) and Bai (2009).

2.1 Representations of CSD

There are various sources of CSD (neighborhood or network effects, the influence of a dominant unit or the influence of common unobserved factors). Factor models propose that the errors reflect a vector of unobserved common factors:

$$y_{it} = \boldsymbol{\delta}_i' \boldsymbol{z}_t + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + \varepsilon_{it} \text{ with } \varepsilon_{it} = \boldsymbol{\gamma}_i' \boldsymbol{f}_t + u_{it} \tag{5}$$

where y_{it} is a scalar dependent variable, $\boldsymbol{z}_t$ is a $k_z \times 1$ vector of variables that do not differ over units, e.g. intercept and trend, $\boldsymbol{x}_{it}$ is a $k_x \times 1$ vector of observed regressors which differ over units, $\boldsymbol{f}_t$ is an $r \times 1$ vector of unobserved factors, which may influence each unit differently and which may be correlated with the $\boldsymbol{x}_{it}$, and u_{it} is an unobserved disturbance with $E(u_{it}) = 0$, $E\left(u_{it}^2\right) = \sigma_i^2$, which is independently distributed across i and (possibly) t. The covariance of the errors, $\varepsilon_{it} = \boldsymbol{\gamma}_i' \boldsymbol{f}_t + u_{it}$ is determined by the factor loadings $\boldsymbol{\gamma}_i$. Notice that if $\boldsymbol{f}_t$ is correlated with $\boldsymbol{x}_{it}$, as is likely in many economic applications, then not allowing for CSD by omitting $\boldsymbol{f}_t$ causes the estimates of $\boldsymbol{\beta}_i$ to be biased.

Spatial models allow the $N \times 1$ vector of errors, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$ to follow:

$$\boldsymbol{\varepsilon}_t = \boldsymbol{W} \boldsymbol{u}_t$$

where $\boldsymbol{u}_t = (u_{1t}, ..., u_{Nt})'$ is cross-sectionally independent and $\boldsymbol{W}$ is a known (possibly time-varying) matrix, e.g. reflecting whether the units share a common border. This can be used to represent spatial autoregressive, moving average or error component models.

This approach assumes that the structure of cross section correlation is related to the location and distance among units on the basis of a pre-specified weight matrix. Hence, cross section correlation is represented by means of a spatial process, which explicitly relates each unit to its neighbors. A number of approaches for modeling spatial dependence has been suggested. The most popular ones are the Spatial Autoregressive (SAR), the Spatial Moving Average (SMA), and the Spatial Error Component (SEC) specifications. The spatial panel data model is estimated using the maximum likelihood (ML) or the generalized method of moments (GMM) techniques (Elhorst, 2011).

We consider a spatial panel data gravity (SARAR) model, which combines a spatial lagged variable and a spatial autoregressive error term:

$$y_{it} = \rho y_{it}^* + \boldsymbol{\beta}' \boldsymbol{x}_{it} + \boldsymbol{\gamma}' \mathbf{z}_i + \tilde{\alpha}_i + v_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{6}$$

$$v_{it} = \lambda v_{it}^* + u_{it} \tag{7}$$

where $y^*_{it} = \sum_{j \neq i}^{N} w_{ij} y_{jt}$ is the spatial lagged variable, and $v^*_{it} = \sum_{j \neq i}^{N} w_{ij} v_{jt}$ is the spatial autoregressive error term, w_{ij}'s are the spatial weight with the row-sum normalisation, $\sum_i w_{ij} = 1$ and $w_{ii} = 0$, and u_{it} is a zero mean idiosyncratic disturbance with constant variance. ρ is the spatial lag coefficient and λ the spatial error component coefficient. They capture the spatial spillover effects and measure the influence of the weighted average of neighboring observations on cross section units. Chudik *et al.* (2011) show that a particular form of a weak cross dependent process arises when pairwise correlations take non-zero values only across finite units that do not spread widely as the sample size rises. A similar case occurs in the spatial/netwrok processes, where the local dependency exists only among adjacent observations.

2.2 Weak and Strong CSD

Chudik et al. (2011) show that these factor models exhibit the strong form of cross-sectional dependence since the maximum eigenvalue of the covariance matrix for ε_{it} tends to infinity at rate N. On the other hand spatial econometric models display the weak form of cross-sectional dependence, which can be represented by an infinite number of weak factors and no idiosyncratic error.

With weak CSD, the dependences are local and decline with N. This could be the case with spatial correlations, where each cross-section unit is correlated with near neighbors but not others; with strong CSD the dependences influence all units. The distinction can be expressed in various ways.

Suppose the elements of y_t are stationary, e.g. growth rates, and the weighted average of the elements $\bar{y}_t = \sum_{i=1}^{N} y_{it}/N$, where the weights are 'granular', go to zero as $N \to \infty$. Then, with weak CSD the variance of $\bar{y}_t$ goes to zero as $N \to \infty$. If there is strong CSD, it does not, for instance there may be a global cycle in $\bar{y}_t$. If there is weak CSD, the influence of the factors, $\sum_{i=1}^{N} \gamma_i^2$ is bounded as $N \to \infty$, while if there is strong dependence, it goes to infinity with N. If there is weak dependence, all the eigenvalues of the covariance matrix of the errors are bounded as $N \to \infty$. If there is strong dependence, the largest eigenvalue goes to infinity with N. Bailey, Kapetanios and Pesaran (2016) characterise the strength of the dependence in terms of the exponent of CSD, defined as $\alpha = \ln(n)/\ln(N)$, where n is the number of units with non zero factor loadings. In the case of a strong factor, $\alpha = 1$ while $\alpha < 1/2$ indicates the weak factor. The values of $\frac{1}{2} \leq \alpha \leq \frac{3}{4}$ represent a moderate degree of CSD.[1]

CSD is central to all the issues discussed. For instance, there is a growing literature on testing for structural change in panels. However, the apparent structural change may result from having left out an unobserved global variable, $\boldsymbol{f}_t$. If $\boldsymbol{f}_t$ is omitted and the correlation between $\boldsymbol{f}_t$ and $\boldsymbol{x}_{it}$ changes, this will change the estimate of $\boldsymbol{\beta}_i$ giving the appearance of structural change. Similarly,

[1] Vega and Elhorst (2016) argue "the terminology of weak and strong cross-sectional dependence is to some extent misleading, because the terms strong and weak suggest that the former is more important than the latter, while we find that their contributions are almost equally as important. We therefore propose the descriptions common factors and spatial dependence to acknowledge the importance of both properties.

an omitted factor may give the impression of non-linearity.[2] Since unobserved factors play a major role in the treatment of CSD, we begin by discussing the estimation of such factors. The implications for estimation are different depending on whether $\boldsymbol{f}_t$ are merely regarded as nuisance parameters that we wish to control for in order to get better estimates of $\boldsymbol{\beta}$ or whether they are the parameters of interest: one wishes to estimate $\boldsymbol{f}_t$ as variables of economic interest in their own right.

3 Factor Models

The meaning of the 'factor' has a variety of different meanings in different areas. Here some observed variables x_{it}, $i = 1, 2, ..., N$ are determined by unobserved factors, f_{jt}, $j = 1, ..., r$:

$$x_{it} = \lambda_{i0} + \sum_{j=1}^{r} \lambda_{ij} f_{jt} + e_{it} \tag{8}$$

where λ_{ij} are called factor loadings, and the e_{it} are idiosyncratic effects. Usually r is much smaller than N so the variation in a large number of observed variables can be reduced to a few unobserved factors

3.1 Uses

Factor models are used in various applications:

- In economics the oldest is the decomposition of time series into unobserved factors labelled trend, cycle, seasonals etc.
- The observed series may be generated by some underlying unobserved factors and the objective is to measure them. This was developed extensively in psychometrics, where the x_{it} are answers to a variety of questions by a sample of people. The underlying factors are aspects of personality, e.g. neuroticism, openness, conscientiousness, agreeableness and extroversion. It has also been used in economics for unobserved variables like: development, natural rates, permanent components, core inflation, etc.
- Factor models can be used to measure the dimension of the independent variation in a set of data, e.g. how many factors are needed to account for most of the variation in x_{it}. For I(1) series these dimensions may be the stochastic trends.
- Factor models can be used to reduce the dimensionality of a set of possible explanatory variables in regression or forecasting models, i.e. replace the large number of x_{it} by a few f_{jt} which contain most of the information in the x_{it}. This may reduce omitted variable problems.

[2] Cerrato, de Peretti and Sarantis (2007) extend the Kapetanios, Shin and Snell (2003) test for a unit root against a non-linear ESTAR alternative to allow for cross-section dependence.

- **Factor models are used to model residual cross-section dependence in panel data models**.

- Factor models have been used to choose instruments for IV or GMM estimators when there is a large number of potential instruments.

One can distinguish two different types of problem. The Pesaran approach to the role of unobserved factors in panels is primarily motivated by the need to allow for "error" cross-sectional dependence. The aim is to estimate the (mean) coefficient of x_{it} allowing for error cross-sectional dependence and/or missing unobserved effects.

Alternatively, it might be relevant to view the unobserved factor as "missing" (omitted) common effects. An example is "technology" in the aggregate production function. In modelling error CSD the error of each cross section unit should have mean zero (otherwise the model suffers from omitted variables) and could be serially correlated. The errors could also be I(1). But if the aim is to test for cointegration between observables, y_{it} and x_{it} (which could also contain observed common effects such as oil prices), and if we maintain the possibility that the errors of the relationship between y_{it} and x_{it} can be I(1), then it is clear that y_{it} and x_{it} cannot cointegrate.

One could hypothesize that y_{it} and x_{it} and f_t are cointegrated where f_t is an "unobserved" factor. Even if such a possibility existed, it may not be relevant if the economic relation of interest is between y_{it} and x_{it}; e.g. in the case of PPP or UIP.

In the case where the common factor represents a missing variable, such as the technological variable in the production function, our main interest is in fact that y_{it} (log output per man hour), k_{it} (log capital per man hour) and f_t (global technology) are cointegrated. The role of f_t is not to model error CSD, but is an integral part of the model, which happens to be unobserved. One could try to obtain proxies for f_t - some assume that f_t is a linear trend with a stationary component while others assume that it is a latent variable and use HP-filter to measure it. **In the context of growth convergence one might estimate f_t by the cross sectional average of y_{it} over i.** But, there will be some degree of arbitrariness. For example, how can we establish that f_t is I(1) or trend stationary? Not knowing whether f_t is I(1) or trend stationary, how can we test that y_{it}, k_{it} and f_t are cointegrated.[3]

3.2 Estimation Methods

There are various ways to estimate factors:

- Univariate ($N = 1$) filters (e.g. the Hodrick-Prescott filter for trends).

- Multivariate ($N > 1$) filters such as the Kalman filter used to estimate unobserved-component models, see Canova (2007).

[3] In the case of testing for panel unit roots, the cross-sectionally augmented CADF test, CADF is a joint test of a unit root and a stationary f_t.

- Multivariate judgemental approaches, e.g. NBER cycle dating based on many series.
- Using a priori weighted averages of the variables.
- Deriving estimates from a model, e.g. Beveridge Nelson decompositions which treat the unobserved variable as the long-horizon forecast.
- **Principal component based methods**.

The relative attractiveness of these methods depends on the number of observed series, N, and the number of unobserved factors, r. The method emphasised is Principal Components (PC). This can be appropriate for large N and small r. Unobserved component models for small N tend to put more parametric structure on the factors whilst PCs do not. **The size of N and T are crucial**. There are some methods that work for small N, other methods that work for large N, but no obvious methods for the medium sized N that we have in practice.

Factor models have a long history. In the early days, it was not clear whether the errors in variables model (observed data generated by unobserved factors) or the errors in equation model was appropriate and both models were used. From the late 1940s the errors in equations model came to dominate. The basic approach to measuring unobserved variables by the PCs of a data matrix was developed by Hotelling (1933). Stone (1947) used this method to show that most of the variation in a large number of financial accounts series could be accounted for by three factors, which could be interpreted as trend, cycle and rate of change of the cycle. Factor analysis was extensively developed in psychometrics and played relatively little role in the development of econometrics, which focussed on the errors in equations model. There are some exceptions, such as factor interpretations of Friedman's permanent income.

However, there has been an explosion of papers on factor models. The original statistical theory was developed for cases where one dimension, say N was fixed and the other say T went to infinity. It is only recently that the theory for large panels, where both dimensions can go to infinity, has been developed, e.g. Bai (2003).

3.3 Calculating Principal Components

Static models Suppose that we have a $T \times N$ data matrix, $\boldsymbol{X}$ with element x_{it} for units $i = 1, ..., N$ and periods $t = 1, ..., T$. The direction in which you take the factors could also be reversed, i.e. treat $\boldsymbol{X}$ as an $N \times T$ matrix. We assume that the $T \times N$ data matrix $\boldsymbol{X}$ is generated by a smaller set of r unobserved factors stacked in the $T \times r$ matrix $\boldsymbol{F}$. In matrix notation,

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda} + \boldsymbol{E} \tag{9}$$

where $\boldsymbol{\Lambda}$ is an $r \times N$ matrix of factor loadings and $\boldsymbol{E}$ is a $T \times N$ matrix of idiosyncratic components. Units can differ in the weight that is given to each

of the factors. Strictly factor analysis involves making some distributional assumptions about e_{it} and applying ML to estimate factor loadings, but we use a different approach and estimate the factors as the PCs of the data matrix.

The PCs of $\boldsymbol{X}$ are the linear combinations of $\boldsymbol{X}$ that have maximal variance and are orthogonal to (uncorrelated with) each other. Often the $\boldsymbol{X}$ matrix is first standardised (subtracting the mean and dividing by the standard deviation), to remove the effect of units of measurement on the variance. $\boldsymbol{X}'\boldsymbol{X}$ is then the correlation matrix. To obtain the first PC we construct a $T\times 1$ vector $\boldsymbol{f}_1 = \boldsymbol{X}\boldsymbol{a}_1$ such that $\boldsymbol{f}_1'\boldsymbol{f}_1 = \boldsymbol{a}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_1$ is maximised. We need some normalisation, so use $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.[4] The problem is to choose $\boldsymbol{a}_1$ to maximise the variance of $\boldsymbol{f}_1$ subject to this constraint. The Langrangian is

$$\begin{aligned}
\mathcal{L} &= \boldsymbol{a}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_1 - \phi_1(\boldsymbol{a}_1'\boldsymbol{a}_1 - 1) \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_1} &= 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_1 - 2\phi_1\boldsymbol{a}_1 = 0 \\
\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_1 &= \phi_1\boldsymbol{a}_1
\end{aligned}$$

so $\boldsymbol{a}_1$ is the first eigenvector of $\boldsymbol{X}'\boldsymbol{X}$, (the one corresponding to the largest eigenvalue, ϕ_1) or the first eigenvector of the correlation matrix of $\boldsymbol{X}$ if the data are standardised. This gives us the weights we need for the first PC.

The second PC, $\boldsymbol{f}_2 = \boldsymbol{X}\boldsymbol{a}_2$ is the linear combination which has the second largest variance, subject to being uncorrelated with $\boldsymbol{a}_1$ i.e. $\boldsymbol{a}_2'\boldsymbol{a}_1 = 0$; so $\boldsymbol{a}_2$ is the second eigenvector. If $\boldsymbol{X}$ is of full rank, there are N distinct eigenvalues and associated eigenvectors and the number of PCs is N.

We can stack the results:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{A}$$

where $\boldsymbol{A}$ is the matrix of eigenvectors and $\boldsymbol{\Phi} = diag(\phi_1, ..., \phi_N)$ is the diagonal matrix of eigenvalues. We can also write this

$$\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A} = \boldsymbol{\Phi} \quad or \quad \boldsymbol{F}'\boldsymbol{F} = \boldsymbol{\Phi}$$

The eigenvalues can be used to calculate the proportion of the variation in $\boldsymbol{X}$ that each principal component explains: $\phi_i / \sum_{i=1}^{N}\phi_i$. If the data are standardised, then $\sum_{i=1}^{N}\phi_i = N$ is the total variance. Forming the PCs is a mathematical operation replacing the $T \times N$ matrix $\boldsymbol{X}$ by the $T \times N$ matrix $\boldsymbol{F}$.

We define the PCs as $\boldsymbol{F} = \boldsymbol{X}\boldsymbol{A}$, but we can also write $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{A}'$ defining $\boldsymbol{X}$ in terms of the PCs.[5] Usually, we want to reduce the number of PCs. To reduce the dimensionality, we can write:

$$\boldsymbol{X} = \boldsymbol{F}_1\boldsymbol{A}_1' + \boldsymbol{F}_2\boldsymbol{A}_2' \tag{10}$$

[4] We need to impose normalizations on the factors and factor loadings to pin down the rotational indeterminacy. This is due to the fact that $\boldsymbol{F}\boldsymbol{\Lambda} = \boldsymbol{F}\boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{\Lambda}$ for any $r \times r$ full-rank matrix $\boldsymbol{Q}$. Because an arbitrary $r \times r$ matrix has r^2 degrees of freedom, we need to impose at least r^2 restrictions (order condition) to remove the indeterminacy.

[5] Note that $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}_N$ and $\boldsymbol{A}' = \boldsymbol{A}^{-1}$.

where the $T \times r$ matrix $\boldsymbol{F}_1$ contains the $r < N$ largest PCs, the $r \times N$ matrix $\boldsymbol{A}_1'$ contains the first r eigenvectors corresponding to the largest eigenvalues. We treat $\boldsymbol{F}_1 = (f_{1t}, ..., f_{rt})$ as the common factors and $\boldsymbol{F}_2\boldsymbol{A}_2'$ as the idiosyncratic factors corresponding to the e_{it} in (9). While it is an abuse of this notation, we usually write $\boldsymbol{F}_1$ as $\boldsymbol{F}$ and $\boldsymbol{F}_2\boldsymbol{A}_2'$ as $\boldsymbol{E}$.

Dynamic models We write the factor model (9) in time series form:

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t \tag{11}$$

where $\boldsymbol{x}_t$ is an $N \times 1$ vector, $\boldsymbol{\Lambda}$ an $N \times r$ matrix of loadings, $\boldsymbol{f}_t$ an $r \times 1$ vector of factors and $\boldsymbol{e}_t$ an $N \times 1$ vector of errors. In using PCs to calculate the factors we have ignored all the information in the lagged values of $\boldsymbol{x}_t$. It may be that some lagged elements of x_{it-j} contain information that help predict x_{it}; e.g. factors influence the variables at different times. Standard PCs, which just extract the information from the covariance matrix, are often called static factor models, because they ignore the dynamic information and the idiosyncratic component, $\boldsymbol{e}_t$, may be serially correlated. There are also dynamic factor models which extract the PCs of the long-run covariance matrix or spectral density matrix, see Forni et al. (2000, 2003, 2005). The spectral density matrix is estimated using some weight function, like the Bartlett or Parzen windows, with some truncation lag.

The dynamic factor model gives different factors, say

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}^*\boldsymbol{f}_t^* + \boldsymbol{e}_t^* \tag{12}$$

where $\boldsymbol{f}_t^*$ is an $r^* \times 1$ vector. In practice, we can approximate the dynamic factors by using lagged values of the static factors,

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}(L)\boldsymbol{f}_t + \boldsymbol{e}_t^s \tag{13}$$

where $\boldsymbol{\Lambda}(L)$ is a pth order lag polynomial. This may be less efficient in the sense that $r < rp$: one can get the same degree of fit with fewer parameters using the dynamic factors than using current and lagged static factors. Determining whether the dynamics in $\boldsymbol{x}_t$ comes from an autoregression in $\boldsymbol{x}_t$, dynamics in $\boldsymbol{f}_t$ or serial correlation in $\boldsymbol{e}_t$ raises quite difficult issues of identification.

Dynamic PCs are two sided filters, taking account of future as well as past information, thus are less suitable for forecasting purposes. This problem does not arise with using current and lagged static factors. Forni et al. (2003) discuss one sided dynamic PCs which can be used for forecasting. Forecasting also includes 'nowcasting', where one has a series, say quarterly GDP, produced with a lag but various monthly series produced very quickly, such as industrial production and retail sales. PCs of the rapidly produced series are then used to provide a 'flash' estimate of current GDP.

3.4 Issues in Using PCs

How to choose r? How many factors to use depends on statistical criteria, the purpose of the exercise and the context. Traditional rules of thumb for

determining r included choosing the PCs that correspond to eigenvalues that are above average value or equivalently greater than unity for standardised data or graphing the eigenvalues and seeing where they drop off sharply. There are also various tests and information criteria. There has been work on information criteria when both N and T are large.

Kapetanios (2004) suggests to use the largest eigenvalue to choose the number of factors. Onatski (2009) suggests another function of the largest eigenvalues of the spectral density matrix at a specified frequency. The statistical properties of the various tests, information criteria and other methods of choosing r for economic data is still a matter of research.

We focus on two types of methods; one based on information criteria and the other based on the distribution of eigenvalues. Bai and Ng (2002) proposed a model selection procedure which can consistently estimate the number of factors when N and T converge to ∞. Let $\hat{\lambda}_i^k$ and $\hat{F}_t^k$ be the PC estimators assuming that the number of factors is k. We may treat the sum of squared residuals (divided by NT) as a function of k:

$$V(k) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \hat{\lambda}_i^{k\prime} \hat{F}_t^k \right)^2$$

Define the following loss function:

$$IC(k) = \ln(V(k)) + kg\left(N, T\right)$$

where the penalty function $g\left(N, T\right)$ satisfies two conditions: (i) $g\left(N, T\right) \to 0$, and (ii) $\min\left\{N^{1/2}, T^{/12}\right\} g\left(N, T\right) \to \infty$, as $N, T \to \infty$. Define the estimator for the number of factors as

$$\hat{k}_{IC} = \arg \min_{0<k<k_{\max}} IC(k);$$

where k_{max} is the upper bound. Then consistency can be established under standard conditions:

$$\hat{k}_{IC} \to r \quad as \quad N, T \to \infty;$$

Bai and Ng (2002) propose the following information criteria:

$$IC_1(k) = \ln(V(k)) + k \left(\frac{N+T}{NT} \right) \ln \left(\frac{NT}{N+T} \right)$$

$$IC_2(k) = \ln(V(k)) + k \left(\frac{N+T}{NT} \right) \ln \left(C_{NT}^2 \right)$$

$$IC_3(k) = \ln(V(k)) + k \left(\frac{\ln \left(C_{NT}^2 \right)}{C_{NT}^2} \right)$$

YC: def. C_{NT}^2??

Monte Carlo simulations show that all criteria perform well when both N and T are large. For the cases where either N or T is small, and if errors are

uncorrelated across units and time, the preferred criteria tend to be IC1 and IC2. Still, they may not work well when N or T are small, leading to too many factors being estimated, e.g. always choosing the maximum number allowed.

Some desirable features of the above method are worth mentioning. Firstly, the consistency is established without any restriction between N and T. Secondly, the results hold under heteroskedasticity in both the time and the cross-section dimensions, as well as under weak serial and cross-section correlation.

Kapetanios (2004) suggests to use the largest eigenvalue to choose the number of factors. Based on random matrix theory, Onatski (2009) established a test of k_0 factors against the alternative that the number of factors is between k_0 and k_1. The test statistic is given by

$$R = \max_{k_0 < k \leq k_1} \frac{\gamma_k - \gamma_{k+1}}{\gamma_{k+1} - \gamma_{k+2}}$$

where γ_k is the k-th largest eigenvalue of the sample spectral density of data at a given frequency. For macroeconomic data, the frequency could be chosen at the business cycle frequency. The basic idea is that under the null of k_0 factors, the first leading k_0 eigenvalues will be unbounded, while the remaining eigenvalues are all bounded. As a result, R will be bounded under the null, while explode under the alternative, making R asymptotically pivotal. The limiting distribution of R is derived under the assumption that T grows sufficiently faster than N, which turns out to be a function of the Tracy-Widom distribution. Ahn and Horenstein (2013) proposed two estimators, the Eigenvalue Ratio (ER) estimator and the Growth Ratio (GR) estimator, based on simple calculation of eigenvalues. The ER estimator is defined as maximizing the ratio of two adjacent eigenvalues in decreasing order. The intuition is similar to Onatski (2009, 2010).

The statistical properties of the various tests, information criteria and other methods of choosing r for economic data is still a matter of research. The choice of r will depend not just on statistical criteria but also the purpose of the exercise and the context.

YC: update and how to extend to the MD case?

Determining the number of dynamic factors The dynamic factor model considers the case in which lags of factors also directly affect x_{it}. The methods for static factor models can be extended to estimate the number of dynamic factors. Consider

$$x_{it} = \lambda_{i0}' f_t + \lambda_{i1}' f_{t-1} + ... + \lambda_{is}' f_{t-s} + e_{it} = \lambda_i(L)' f_t + e_{it} \tag{14}$$

where f_t is $q \times 1$ and $\lambda_i(L) = \lambda_{i0} + \lambda_{i1} L + ... + \lambda_{is} L^s$. While Forni et al. (2000, 2004, 2005) consider the case with $s \to \infty$, we focus on the case with a fixed s. Model (14) can be represented as a static factor model with $r = q(s+1)$ static factors:

$$x_{it} = \lambda_i' F_t + e_{it}$$

$$\lambda_i = \begin{bmatrix} \lambda_{i0} \\ \lambda_{i1} \\ \vdots \\ \lambda_{is} \end{bmatrix}; \; F_t = \begin{bmatrix} f_t \\ f_{t-1} \\ \vdots \\ f_{t-s} \end{bmatrix}$$

We refer to f_t as the dynamic factors and F_t as the static factors. Regarding the dynamic process for f_t, we may use a finite-order VAR:

$$\Phi(L) f_t = \varepsilon_t$$

where $\Phi(L) = I_q - \Phi_1 L - ... - \Phi_h L^h$. Then, we may form the $VAR(k)$ representation of the static factor, F_t, where $k = \max\{h, s+1\}$,

$$\Phi_F(L) F_t = u_t \text{ with } \; u_t = R\varepsilon_t$$

where $\Phi_F(L) = I_{q(s+1)} - \Phi_{F1} L - ... - \Phi_{Fk} L^k$, and the $q(s+1) \times q$ matrix R are given by $R = [I_q, 0, ..., 0]'$.

The spectrum of the static factors has rank q instead of $r = q(s+1)$. Given that

$$\Phi_F(L) F_t = R\varepsilon_t;$$

the spectrum of F at frequency ω is

$$S_F(\omega) = \Phi_F\left(e^{-i\omega}\right)^{-1} R S_\varepsilon(\omega) R' \Phi_F\left(e^{i\omega}\right)^{-1};$$

whose rank is q if $S_\varepsilon(\omega)$ has rank q for $|\omega| \leq \pi$. This implies $S_F(\omega)$ has only q nonzero eigenvalues. Hallin and Liska (2007) estimate the rank of this matrix to determine the number of dynamic factors. Onatski (2009) also considers estimating q using the sample estimates of $S_F(\omega)$.

Alternatively, we may first estimate a static factor model using Bai and Ng (2002) to obtain $\hat{F}_t$. Next, we may estimate a $VAR(p)$ for $\hat{F}_t$ to obtain the residuals $\hat{u}_t$. Let $\hat{\Sigma}_u = T^{-1} \sum \hat{u}_t \hat{u}_t'$. Note that the theoretical moments $E(u_t u_t')$ has rank q. We may estimate q using the information about the rank of $\hat{\Sigma}_u$.

Using a different approach, Stock & Watson (2005) considered a richer dynamics in error terms, and transformed the model such that the residual of the transformed model has a static factor representation with q factors. Bai and Ng (2002)'s IC can then be directly applied to estimate q.

Amengual and Watson (2007) derived the corresponding econometric theory for estimating q. They started from the static factor model,

$$X_t = \Lambda F_t + e_t,$$

and considered a $VAR(p)$ for F_t,

$$F_t = \sum_{i=1}^{p} \Phi_i F_{t-i} + \varepsilon_t; \quad \varepsilon_t = G\eta_t;$$

where G is $r \times q$ with full column rank and η_t is a sequence of shocks with mean 0 and variance I_q. The shock η_t is the dynamic factor shock, whose dimension is the number of dynamic factors. Let

$$Y_t = X_t - \sum_{i=1}^{p} \Lambda \Phi_i F_{t-i}$$

and $\Gamma = \Lambda G$, then Y_t has a static factor representation with q factors,

$$Y_t = \Gamma \eta_t + e_t :$$

If Y_t is observed, q can be directly estimated using Bai and Ng (2002)'s IC. In practice, Y_t needs to be estimated. Let $\hat{Y}_t = X_t - \sum_{i=1}^{p} \hat{\Lambda} \hat{\Phi}_i \hat{F}_{t-i}$ where $\hat{\Lambda}$ and $\hat{F}_t$ are PC estimators from X_t, and $\hat{\Phi}_i$ is obtained by $VAR(p)$ regression of $\hat{F}_t$.

How to choose N? One may have very large amounts of potential data available (e.g. thousands of time series on different economic, social, and demographic variables for different countries) and an issue is how many you should use in constructing the PCs. Information seems better so one should include as many as possible, but this may not be the case. Adding variables that are weakly dependent on the common factors will add very little information.

To calculate the PCs the weights on the series have to be estimated and adding more series adds more parameters to be estimated. This increases the noise due to parameter estimation error. If the series have little information on the factors of interest, they just add noise, worsening the estimation problem. The series may be determined by different factors, increasing the number of factors needed to explain the variance. They may also have outliers or idiosyncratic jumps and this will introduce a lot of variance which may be picked up by the estimated factors. Many of the disputes in the literature about the relevant number of factors reflect the range of series used to construct the PCs.

If the series are mainly different sort of price and output measures, two factors may be adequate; but if one adds financial series such as stock prices and interest rates, or foreign variables, more factors may be needed. One may be able to look at the factor loading matrix and see whether it has a block structure, certain factors loading on certain sets of variables. If this is the case one may want to split the data using different dataset to estimate different factors. But it may be difficult to determine the structure of the factor loading matrix.

N may be larger than the number of variables, if transformations of the variables (e.g. logarithms, powers, first differences, etc.) are also included. This trade-off between the size of the information set and the errors introduced by estimation may be a particular issue in forecasting, where parsimony tends to produce better forecasting models. Then using more data may not improve forecasts, e.g. Mitchell et al. (2005) and Elliott and Timmerman (2008). Notice that in forecasting we would need to update our estimates of F_t, and perhaps r the number of factors, as our sample size, T changes.

How to Identify and interpret factors? To interpret the factors requires just identifying restrictions. Suppose that we have obtained estimates:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda} + \boldsymbol{E}$$

For any non-singular $r \times r$ matrix, $\boldsymbol{Q}$, the new factors and loadings $(\boldsymbol{F}\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{\Lambda})$ are observationally equivalent to $\boldsymbol{F}\boldsymbol{\Lambda}$. The new loadings are $\boldsymbol{\Lambda}^* = \boldsymbol{Q}^{-1}\boldsymbol{\Lambda}$ and factors are $\boldsymbol{F}^* = \boldsymbol{F}\boldsymbol{Q}$. The r^2 just identifying restrictions used to calculate PCs are the unit length and orthogonality normalisations which come from treating it as an eigenvalue problem. Thus, the factors are only defined up to a non-singular transformation. A major problem in applications is to interpret the estimated PCs. Often in time-series the first PC has roughly equal weights and corresponds to the mean of the series. Looking at the factor loadings and the graphs of the PCs may help interpret them. The choice of $\boldsymbol{Q}$, just identifying restrictions, called 'rotations' in psychometrics, is an important part of traditional factor analysis. These are needed to provide some interpretation of the factors.[6]

The same identification issue arises in simple regression. For

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

is observationally equivalent to the reparameterisation:

$$\boldsymbol{y} = \left(\boldsymbol{X}\boldsymbol{Q}^{-1}\right)(\boldsymbol{Q}\boldsymbol{\beta}) + \boldsymbol{u} = \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{u}$$

For instance, $\boldsymbol{Z}$ could be the PCs, which have the advantage that they are orthogonal and so the estimates of the factor coefficients are invariant to which other factors are included. But there could be other $\boldsymbol{Q}$ matrices. To interpret the regression coefficients we need to choose a parameterisation, k^2 restrictions that specify $\boldsymbol{Q}$. We tend to take the parameterisation for granted in economics, so this is not usually called an identification problem.

For some purposes, e.g. forecasting, one may not need to identify the factors, but for other purposes their interpretation is crucial. It is quite often the case that one estimates the PCs and has no idea what they mean or measure.

Estimated or imposed weights? Factors are estimated as linear combinations of observed data series. Above it has been assumed that the weights in the linear combination should be estimated to optimise some criterion function, e.g. to maximise variance explained in the case of PCs. However, there are possible *a priori* weights, imposing the weights rather than estimating them. Examples are equal weights as in the mean or trade weights as in effective exchange rates. There is a bias-variance trade-off. The imposed weights are almost certainly biased, but have zero variance. The estimated weights may be unbiased but may have large variance because of estimation error. The imposed

[6] Rotations in psychometrics are as controversial as just-identifying restrictions in economics, so while many psychologists agree that there are five dimensions to personality, $r = 5$; how they are described differs widely.

weights may be better than the estimated weights in the sense of having smaller mean square error (bias squared plus variance). Forecast evaluation of regression models indicates that simple models with imposed coefficients tend to do very well. Measures constructed with imposed weights are usually also much easier to interpret.

The most obvious candidate for imposed weights is to use equal weights, a simple average (perhaps after having standardised the variable to have mean zero and variance one). In many cases the first PC seems to have roughly equal weights and thus behave like an average or sum.

Alternatively, economic theory may suggest suitable weights. For instance, effective exchange rates for country i (weighted averages of exchange rates with all other countries) use trade weights: exports plus imports of i with j as a share of total exports plus imports of country i). PCs might give a lot of weight to a set of countries which have very volatile exchange rates even though country i does not trade with them. Measures of core inflation give zero weights to the inflation rates of certain volatile components of total expenditure while a PC might give a high weight to those volatile components because they account for a lot of the variance. Monte Carlo evaluation of estimators that allow for CSD indicate the methods that use imposed weights, like the CCE estimator, often do much better than estimators that rely on estimating the number of PCs and their weights. See the comparison between the CCE estimator by Pesaran and the IPC estimator by Bai see also the studies by Westlund and coauthors.

Explanation using PCs Suppose the model of interest is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{15}$$

where $\boldsymbol{\beta}$ is an $N \times 1$ vector of parameters and we wish to reduce the dimension of $\boldsymbol{X}$. This could be because there are a very large number of candidate variables or because there is multicollinearity. Replacing $\boldsymbol{X}$ by all the PCs, $\boldsymbol{F} = \boldsymbol{X}\boldsymbol{A}$ is just a reparameterisation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{\beta} + \boldsymbol{u} = \boldsymbol{F}\boldsymbol{\delta} + \boldsymbol{u} \tag{16}$$

However, we could reduce the number of PCs by writing it:

$$\boldsymbol{y} = \boldsymbol{F}_1\boldsymbol{\delta}_1 + \boldsymbol{F}_2\boldsymbol{\delta}_2 + \boldsymbol{u} \tag{17}$$

where $\boldsymbol{F}_1$ are the $r < N$ PCs (corresponding to the r largest eigenvalues). Setting $\boldsymbol{\delta}_2 = \boldsymbol{A}_2'\boldsymbol{\beta} = 0$ to give

$$\boldsymbol{y} = \boldsymbol{F}_1\boldsymbol{\delta}_1 + \boldsymbol{v} \tag{18}$$

In this case the original coefficients could be recovered as $\boldsymbol{\beta} = \boldsymbol{A}_1\boldsymbol{\delta}_1$. The hypothesis, $\boldsymbol{\delta}_2 = 0$ is testable (as long as $N < T$). This has been suggested as a way of dealing with multicollinearity, or choosing a set of instruments.

There are some problems. First, it is quite possible that a PC which has a small eigenvalue and explains a very small part of the total variation of $\boldsymbol{X}$

may explain a large part of the variation of $\boldsymbol{y}$. The PCs are chosen on the basis of their ability to explain $\boldsymbol{X}$ not $\boldsymbol{y}$, but the regression is designed to explain $\boldsymbol{y}$. Secondly, unless $\boldsymbol{F}_1$ can be given an interpretation, e.g. as an unobserved variable, it is not clear whether the hypothesis, $\boldsymbol{\delta}_2 = \boldsymbol{A}_2'\boldsymbol{\beta} = 0$ has prior plausibility or what the interpretation of the estimated regression is. Thirdly, estimation error is being introduced by using $\boldsymbol{F}_1$ and these are generated regressors with implications for the estimation of the standard errors of $\boldsymbol{\delta}_1$. As a result, until recently with the Factor augmented VARS and ECMs discussed below, economists have tended not to use PCs as explanatory variables in regressions. Instead multicollinearity tended to be dealt with through the use of theoretical information, either explicitly through Bayesian estimators or implicitly by a priori weights e.g. through the construction of aggregates. Notice that we could include certain elements of $\boldsymbol{X}$ directly and have others summarised in factors.

3.5 Factor-Augmented Regressions

One of the popular applications of large factor model is the factor-augmented regressions. Bai and Ng (2006) develop the econometric theory for such factor-augmented regressions. Consider the following forecasting model for y_t:

$$y_{t+h} = \alpha' F_t + \beta' W_t + \varepsilon_{t+h}$$

where W_t is the vector of a small number of observables including lags of y_t, and F_t is unobservable. Suppose there is a large number of series x_{it}, $i = 1, ..., N$, $t = 1, ..., T$, which has a large factor representation as

$$x_{it} = \lambda_i' F_t + e_{it} :$$

When y_t is a scalar, these become the diffusion index forecasting model of Stock and Watson (2002b). Clearly, each x_{it} is a noisy predictor for y_{t+h}. Because F_t is latent, the conventional mean-squared optimal prediction of y_{t+h} is not feasible. Alternatively, consider the method of PCs to estimate $\hat{F}_t$, which is a consistent estimator for $H'F_t$ for some rotation matrix H. Then, regress y_{t+h} on $\hat{F}_t$ and W_t to obtain $\hat{\alpha}$ and $\hat{\beta}$. The feasible prediction for $\hat{y}_{T+h|T} = E(y_{T+h}|\Omega_T)$, where $\Omega_T = [F_T, W_T, F_{T-1}, W_{T-1}, ...]$ is given by

$$\hat{y}_{T+h|T} = \hat{\alpha}' \hat{F}_t + \hat{\beta}' W_t$$

Let $\delta = \left(\alpha' H^{-1}, \beta'\right)'$ and $\varepsilon_{T+h|T} = y_{T+h} - y_{T+h|T}$, Bai and Ng (2006) show that when N is large relative to T (i.e., $\sqrt{T}/N \to 0$), $\hat{\delta}$ will be $\sqrt{T}$-consistent and asymptotically normal. $\hat{y}_{T+h|T}$ and $\hat{\varepsilon}_{T+h|T}$ are $\min\left\{N^{1/2}, T^{1/2}\right\}$-consistent and asymptotically normal. Inference needs to take into account the estimated factors, except for the special case $T/N \to 0$. In particular, under standard assumptions for large approximate factor model, when $\sqrt{T}/N \to 0$, we have

$$\hat{\delta} - \delta \to_d (0, \Sigma_\delta)$$

Let $z_t = [F_t', W_t']'$, $\hat{z}_t = [\hat{F}_t', W_t']'$, and $\hat{\varepsilon}_{t+h} = y_{t+h} - \hat{y}_{t+h|t}$, a heteroskedasticity-consistent estimator for Σ_δ is given by

$$\Sigma_\delta = \left(\frac{1}{T}\sum_{t=1}^{T-h} \hat{z}_t \hat{z}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T-h} \hat{\varepsilon}_{t+h}^2 \hat{z}_t \hat{z}_t'\right) \left(\frac{1}{T}\sum_{t=1}^{T-h} \hat{z}_t \hat{z}_t'\right)^{-1}$$

If in addition, we assume $\sqrt{N}/T \to 0$, then

$$\frac{\hat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{var\left(\hat{y}_{T+h|T}\right)}} \to_d N(0,1)$$

where

$$var\left(\hat{y}_{T+h|T}\right) = \frac{1}{N}\hat{z}_T' Avar\left(\hat{\delta}\right)\hat{z}_T + \frac{1}{N}\hat{\alpha}' Avar\left(\hat{F}_T\right)\hat{\alpha}$$

A notable feature of the limiting distribution of the forecast is that the overall convergence rate is given by $\min\left\{N^{1/2}, T^{1/2}\right\}$. Given that

$$\hat{\varepsilon}_{T+h} = \hat{y}_{T+h|T} - y_{T+h} = \hat{y}_{T+h|T} - y_{T+h|T} + \varepsilon_{T+h}$$

if we further assume that ε_t is normal with variance σ_ε^2, then the forecasting error also becomes approximately normal

$$\hat{\varepsilon}_{T+h} \sim N\left(0, \sigma_\varepsilon^2 + var\left(\hat{y}_{T+h|T}\right)\right)$$

so that confidence intervals can be constructed for the forecasts.

3.6 Factor Augmented VAR (FAVAR)

The analysis of monetary policy often involves estimating a small VAR in some focus variables, e.g. output, inflation and interest rates. Then, the VAR is used to examine the effect of a monetary shock to interest rates on the time paths of the variables (impulse response functions). To identify the monetary shock involves making some short-run just identifying assumptions, e.g. a Choleski decomposition imposes a recursive causal ordering, in which some variables (e.g. output and inflation) are assumed to respond slowly, and others (e.g. interest rates) to respond fast. VARs plus identifying assumptions are often called structural VARs. Generalised Impulse Response Functions do not require any just identifying assumptions but cannot be given a structural interpretation.

Small VARs can give implausible impulse response functions, e.g. the "price puzzle", that a contractionary monetary shock was followed by a price increase rather than a price decrease as economic theory would predict. This was interpreted as reflecting misspecification errors, the exclusion of relevant conditioning information. One response was to add variables and use larger VARs, but this route rapidly runs out of degrees of freedom, since Central Bankers monitor hundreds of variables. The results are also sensitive to the choice of variables.

A central question of using VAR is how to identify structural shocks, which in turn depends on what variables to include in the VAR system. A small VAR

cannot fully capture the structural shocks. In the meantime, including more variables in the VAR system could be problematic due to either the degree of freedom problem or the variable selection problem. We now focus on another popular response, the factor-augmented vector autoregressions (FAVAR), originally proposed by Bernanke et al. (2005). The FAVAR assumes that a large number of economic variables are driven by a small VAR, which can include both latent and observed variables. The dimension of structural shocks can be estimated instead of being assumed to be known and fixed.

FAVAR is used to measure US monetary policy in Bernanke Boivin and Eliasz (2005, BBE); UK monetary policy in Lagana and Mountford (2005, LM); US and Eurozone monetary policy in Favero, Marcellini and Neglia (2005, FMN). The technical issues are discussed by Stock and Watson (2005, SW).

Consider an $M \times 1$ vector of observed focus variables $\boldsymbol{Y}_t$, a $K \times 1$ vector of unobserved factors $\boldsymbol{F}_t$ with a VAR structure:

$$\begin{pmatrix} \boldsymbol{F}_t \\ \boldsymbol{Y}_t \end{pmatrix} = \boldsymbol{A}(L) \begin{pmatrix} \boldsymbol{F}_{t-1} \\ \boldsymbol{Y}_{t-1} \end{pmatrix} + \boldsymbol{v}_t \tag{19}$$

where $\boldsymbol{A}(L)$ is a polynomial lag operator. The unobserved factors $\boldsymbol{F}_t$ are related to an $N \times 1$ vector $\boldsymbol{X}_t$, which contains a large number (BBE use $N = 120$; LM $N = 105$) of potentially relevant observed variables by

$$\boldsymbol{X}_t = \boldsymbol{\Lambda} \boldsymbol{F}_t + \boldsymbol{e}_t$$

where $\boldsymbol{F}$ are estimated as the PCs of $\boldsymbol{X}_t$, which may include $\boldsymbol{Y}_t$. There is an identification problem, since

$$\boldsymbol{X}_t = \boldsymbol{\Lambda} \boldsymbol{F}_t + \boldsymbol{e}_t = \boldsymbol{\Lambda} \boldsymbol{Q} \boldsymbol{Q}^{-1} \boldsymbol{F}_t + \boldsymbol{e}_t = \boldsymbol{\Lambda}^* \boldsymbol{F}_t^* + \boldsymbol{e}_t.$$

It is common to use an arbitrary statistical assumption to identify the loadings as eigenvectors, but other assumptions are possible. The standard practice is to difference the observable data so that they are stationary, $I(0)$. The factors are therefore stationary. Differencing loses levels information about the level relationships, but if one does not difference one has to take account of cointegration *etc.*

The argument is that (a) a small number of factors can account for a large proportion of the variance of $\boldsymbol{X}_t$ and thus reduce omitted variable bias in the VAR; (b) the factor structure for $\boldsymbol{X}_t$ allows one to calculate impulse response functions for all the elements of $\boldsymbol{X}_t$ in response to a (structural) shock in $\boldsymbol{Y}_t$ transmitted through $\boldsymbol{F}_t$; (c) the factors may be better measures of underlying theoretical variables such as economic activity than the observed proxies such as GDP or industrial production; (d) FAVARs may forecast better than standard VARs; (e) factor models can approximate infinite dimensional VARs, see Chudik and Pesaran (2011).

BBE conclude: "the results provide some support for the view that the "price puzzle" results from the exclusion of conditioning information. The conditioning information also leads to reasonable responses of monetary aggregates".

The simplest approach (the two step method) is to (i) estimate K PCs from the $\boldsymbol{X}$, (ii) estimate the VAR treating the PCs as measures of $\boldsymbol{F}_t$ variables along with the M observed focus variables $\boldsymbol{Y}_t$. The standard errors produced by the two-step estimates are subject to the generated regressor problem and thus can potentially lead to misleading inference. In large samples $\boldsymbol{F}_t$ can be treated as known, thus there is no generated regressor problem, but it is not clear how good this approximation is in practice.

Choosing M and K, the number of focus variables and the number of factors, raises difficult issues. SW for the US and LM for the UK argue for 7 factors, BBE argue for smaller numbers e.g. $M = 3$ and $K = 1$; or $M = 1$ and $K = 3$. They use monthly data with either output, inflation and the interest rate as focus variables and one factor or the interest rate as the only observed focus variable and 3 unobserved factors, their preferred specification. If a large number of factors are needed, it reduces the attraction of the procedure and may make interpretation of the factors more difficult. The procedure is sensitive to the choice of $\boldsymbol{X}_t$. Just making the set of variables large does not solve the problem, because there may be factors that are very important in explaining $\boldsymbol{X}_t$, but do not help in explaining $\boldsymbol{Y}_t$ and *vice versa*. BBE motivate the exercise with the standard 3 equation macro model with the unobserved factors being the natural level of output and supply shocks. However, they do not use this interpretation in the empirical work, just note the need to interpret the estimated factors more explicitly. Boivin and Giannoni (2006) use the theory putting the factor model in the context of a DSGE with imperfect measurement of the theoretical variables.

Bai et al. (2015) show that, under suitable identification conditions, inferential theory can be developed for such a two-step estimator, which differs from a standard large factor model. The second method involves a one-step likelihood approach, implemented by Gibbs sampling, which leads to joint estimation of both the latent factors and the impulse responses. The two methods can be complement of each other. A useful feature of the FAVAR is that the impulse response function of all variables to the fundamental shocks can be readily calculated.

4 The Factor-based Models of Cross Sectionally Dependent Panels

CSD has attracted considerable attention and a large number of estimators have been suggested. Currently, the market leader, according to Monte Carlo studies, appears to be CCE estimators. It is common to transform the data to make it stationary before calculating PCs by differencing. If one tries to measure a stationary unobservable, e.g. a global trade cycle, this is clearly sensible. It is equally not sensible if one is trying to measure a non-stationary unobservable, e.g. a global trend. Even in the stationary case it is important that transformations beyond differencing be considered, stationary transformations of levels

variables, such as interest rate spreads, may also contain valuable information. We describe main techniques for dealing with factor-based CSD.

SURE Suppose that the model is heterogeneous:

$$y_{it} = \boldsymbol{\delta}_i' \boldsymbol{z}_t + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + \boldsymbol{\gamma}_i' \boldsymbol{f}_t + u_{it},\ i = 1, ..., N,\ t = 1, ..., T \tag{20}$$

where y_{it} is a scalar dependent variable, $\boldsymbol{z}_t$ is a k_z vector of variables that do not differ over groups (intercept and trend), and $\boldsymbol{x}_{it}$ is a $k_x \times 1$ vector of observed regressors which differ over groups, $\boldsymbol{f}_t$ is an $r \times 1$ vector of unobserved factors and u_{it} is an unobserved disturbance with $E(u_{it}) = 0$, $E(u_{it}^2) = \sigma_i^2$ which is independently distributed across i and t. Estimating

$$y_{it} = \boldsymbol{\delta}_i' \boldsymbol{z}_t + \boldsymbol{b}_i' \boldsymbol{x}_{it} + v_{it},\ i = 1, ..., N,\ t = 1, ..., T \tag{21}$$

will give inconsistent estimates of $\boldsymbol{\beta}_i$ if $\boldsymbol{f}_t$ is correlated with $\boldsymbol{x}_{it}$ and inefficient estimates even if $\boldsymbol{f}_t$ is not correlated with $\boldsymbol{x}_{it}$. In the latter case if N is small, the equations can be estimated by SURE, but if N is large relative to T, SURE is not feasible, because the estimated covariance matrix cannot be inverted. Robertson and Symons (2007) suggest using the factor structure to obtain an invertible covariance matrix. Their estimator is quite complicated and will not be appropriate if the factors are correlated with the regressors.

Time effects/demeaning If $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, and there is a single factor which influences each group in the same way, i.e. $\gamma_i = \gamma$, then including time effects, a dummy variable for each period, i.e. the two way fixed effect estimator:

$$y_{it} = \theta_t + \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + u_{it}$$

will estimate $f_t' \gamma = \theta_t$. This can be implemented by using time-demeaned data, $\tilde{y}_{it} = y_{it} - \bar{y}_t$, where $\bar{y}_t = N^{-1} \sum_{i=1}^{N} y_{it}$ and similarly for $\tilde{\boldsymbol{x}}_{it}$. **Unlike SURE the factor does not have to be distributed independently of x_{it} for this to work.**

It is sometimes suggested (e.g. for unit root tests) that demeaned data be used even in the case of heterogeneous slopes. Suppose we have heterogeneous random parameters:

$$y_{it} = \theta_t + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + u_{it} \text{ with } \boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_i$$

Averaging over groups for each period we get:

$$\bar{y}_t = \theta_t + \boldsymbol{\beta}' \bar{\boldsymbol{x}}_t + \bar{u}_t + N^{-1} \sum_{i=1}^{N} \boldsymbol{\eta}_i' \boldsymbol{x}_{it}$$

Noting that

$$\boldsymbol{\beta}_i' \boldsymbol{x}_{it} - \boldsymbol{\beta}' \bar{\boldsymbol{x}}_t = \boldsymbol{\beta}_i' \tilde{\boldsymbol{x}}_{it} + \boldsymbol{\eta}_i' \bar{x}_t$$

Demeaning ($\tilde{y}_{it} = y_{it} - \bar{y}_t$) gives:

$$\tilde{y}_{it} = \boldsymbol{\beta}_i' \tilde{\boldsymbol{x}}_{it} + \tilde{u}_{it} + e_{it} \text{ with } e_{it} = \boldsymbol{\eta}_i' \bar{x}_t - N^{-1} \sum_{i=1}^{N} \boldsymbol{\eta}_i' \boldsymbol{x}_{it} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\eta}_i' \tilde{\boldsymbol{x}}_{it}$$

This removes the common factor θ_t, but has added new terms to the error reflecting the effect of slope heterogeneity. If $\boldsymbol{\eta}_i$ is independent of the regressors, e_t will have expected value zero and be independent of the regressors, so one can obtain large T consistent estimates of $\boldsymbol{\beta}_i$, but the variances will be larger. One can compare the fit of the panels using the original data y_{it} and the demeaned data $\tilde{y}_{it}$ to see which effect dominates, i.e. whether the reduction in variance from eliminating θ_t is greater or less than the increase in variance from adding e_{it}.

This model assumes that the factor has identical effects on each unit, **implying that they impose the same CSD across all cross-section units. Rather than demeaning, it is usually better to include the means directly.**

4.1 The Correlated Common Effect (CCE) Estimator

If one wishes to treat factors as nuisance parameters and remove the effect of CSD, a simple and effective procedure, for large N, T, is the CCE estimator of Pesaran (2006). Consider the panel data model with unobserved common factors:

$$y_{it} = \boldsymbol{\delta}_i' \boldsymbol{d}_t + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + \varepsilon_{it} \text{ with } \varepsilon_{it} = \boldsymbol{\gamma}_i' \boldsymbol{f}_t + u_{it} \tag{22}$$

where y_{it} is a scalar dependent variable, $\boldsymbol{d}_t$ is a $k_d \times 1$ vector of variables that do not differ over units, e.g. intercept and trend, $\boldsymbol{x}_{it}$ is a $k_x \times 1$ vector of observed regressors which differ over units, $\boldsymbol{f}_t$ is an $r \times 1$ vector of unobserved factors, which may influence each unit differently and which may be correlated with the $\boldsymbol{x}_{it}$, and u_{it} is an unobserved disturbance with $E(u_{it}) = 0$ and $E(u_{it}^2) = \sigma_i^2$, which is independently distributed across i and t. If $\boldsymbol{x}_{it}$ is correlated with $\boldsymbol{f}_t$ and $\boldsymbol{\gamma}_i$, then not allowing for CSD by omitting $\boldsymbol{f}_t$ causes the estimates of $\boldsymbol{\beta}_i$ to be biased and inconsistent.

Pesaran suggests to include the means of y_{it} and $\boldsymbol{x}_{it}$ as additional regressors, to remove the effect of the factors. The CCE procedure can handle multiple factors which are I(0) or I(1), which can be correlated with the regressors, and handles serial correlation in the errors. The consistency holds for any linear combination of the dependent variable and the regressors, not just the arithmetic mean, subject to the assumptions that the weights, w_i satisfy:

$$(i) : w_i = O\left(\frac{1}{N}\right);\ (ii) : \sum_{i=1}^{N} |w_i| < K;\ (iii) : \sum_{i=1}^{N} w_i \gamma_i \neq 0$$

These clearly hold for the mean:

$$w_i = \frac{1}{N};\ \sum_{i=1}^{N} |w_i| = 1;\ \ \sum_{i=1}^{N} w_i \gamma_i = N^{-1} \sum_{i=1}^{N} \gamma_i \neq 0$$

as long as the mean effect of the factor on the dependent variable is non-zero.

This involves adding the means of the dependent and independent variables to the regression, (22):

$$y_{it} = \boldsymbol{\delta}_i' \boldsymbol{d}_t + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + \boldsymbol{\pi}_{yi} \bar{y}_t + \boldsymbol{\pi}_{xi}' \bar{\boldsymbol{x}}_t + u_{it} \tag{23}$$

To see the motivation, assume a single factor and average (5) across units to give:

$$\bar{y}_t = \bar{\boldsymbol{\delta}}' \boldsymbol{z}_t + \bar{\boldsymbol{\beta}}_i' \bar{\boldsymbol{x}}_t + \bar{\gamma} f_t + \bar{u}_t + N^{-1} \sum \left(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}} \right)' \boldsymbol{x}_{it} \tag{24}$$

and thus

$$f_t = \frac{1}{\bar{\gamma}} \left\{ \bar{y}_t - \bar{\boldsymbol{\delta}}' \boldsymbol{z}_t - \bar{\boldsymbol{\beta}}_i' \bar{\boldsymbol{x}}_t - \bar{u}_t - N^{-1} \sum \left(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}} \right)' \boldsymbol{x}_{it} \right\} \tag{25}$$

so the $\bar{y}_t$ and $\bar{\boldsymbol{x}}_t$ provide a proxy for the unobserved factor. As the covariance between $\bar{y}_t$ and u_{it} goes to zero with N, so for large N there is no endogeneity problem. The CCE generalises to many factors and lagged dependent variables, **but requires that $\bar{\gamma}$ or the vector equivalent, is non-zero**.

Pesaran (2006) showed that β_i can be consistently estimated through the augmented OLS regression, (23) under the large N, T. Namely,

$$\hat{\boldsymbol{\beta}}_i = \left(\boldsymbol{X}_i' \boldsymbol{M}_D \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i' \boldsymbol{M}_D \boldsymbol{y}_i$$

where $\boldsymbol{y}_i$ is a $T \times \ 1$ vector of the dependent variable for the ith unit, $\boldsymbol{X}_i$ is a $T \times k_x$ vector of regressors, and $\boldsymbol{M}_D = \boldsymbol{I}_T - \boldsymbol{D} \left(\boldsymbol{D}' \boldsymbol{D} \right)^{-1} \boldsymbol{D}$ and $\boldsymbol{D}$ consists of observed common factor and cross sectional average of dependent and independent variables. We can use the mean group estimator:

$$\hat{\boldsymbol{\beta}}_{MG} = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\beta}}_i.$$

This assumes heterogeneous coefficients, but there are homogeneous versions. Alternatively, $\boldsymbol{\beta} = E(\boldsymbol{\beta}_i)$ can be obtained by a pooled estimate:

$$\hat{\beta}_P = \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{M}_D \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{M} D \boldsymbol{y}_i$$

A clear advantage of the CCE method is that the number of unobserved factors need not be estimated. In fact, the method is valid with a single or multiple unobserved factors and does not require the number of factors to be smaller than the number of observed cross-section averages. In addition, CCE is easy to compute as an outcome of OLS and no iteration is needed. Desirable small sample properties of CCE are also demonstrated.

There are sometimes economic reasons for adding averages, but in other cases the economic interpretation is not straightforward. In a variety of circumstances

estimating the factors by the means, seems to work better than estimating them directly by the PC estimator. This procedure determines the weights *a priori* rather than estimating them by PCs. Not estimating the weights seems to improve the performance of the procedure. Kapetanios, Pesaran and Yamagata (2011) show that this procedure is robust to a wide variety of data generation processes including unit roots.

Remark: Westerlund and Urbain (2013) show that Pesaran's estimator becomes inconsistent when the factor loadings in the y equation are correlated with the factor loadings in the x equation.

YC: update

4.2 Panel Data Models with Interactive Fixed Effects

Bai (2009) considers the following large N large T panel data model:

$$y_{it} = \boldsymbol{X}_{it}'\boldsymbol{\beta} + u_{it} \text{ with } u_{it} = \boldsymbol{\lambda}_i'\boldsymbol{F}_t + \varepsilon_{it} \tag{26}$$

where $\boldsymbol{X}_{it}$ is a $k \times 1$ vector of regressors and $\boldsymbol{f}_t$ is an $r \times 1$ vector of unobserved factors. The main difference is that it assumes homogeneous parameters. Bai interprets it as a generalisation of the additive two-way fixed effect model. We observe y_{it} and X_{it}, but do not observe λ_i, F_t, and ε_{it}. Such model nests conventional fixed effects models as special cases due to the following transformation:

$$y_{it} = \boldsymbol{X}_{it}'\boldsymbol{\beta} + \alpha_i + \xi_t + \varepsilon_{it} = \boldsymbol{X}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{F}_t + \varepsilon_{it}$$

where $\boldsymbol{\lambda}_i = (1, \alpha_i)'$ and $\boldsymbol{F}_t = (t, 1)'$. The interactive fixed effects allow a much richer form of unobserved heterogeneity. For example, F_t can represent a vector of macroeconomic common shocks and $\boldsymbol{\lambda}_i$ captures individual i's heterogeneous response to such shocks.

YC: update on macro and micro

Bai (2009) allows X_{it} to be correlated with λ_i, F_t or both. Under the large N, T, we may estimate the model by minimising a LS objective function:

$$\begin{aligned} SSR(\beta, F, \Lambda) &= \sum_{i=1}^{N} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{F}\boldsymbol{\lambda}_i)' (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{F}\boldsymbol{\lambda}_i) \\ s.t.\ \frac{\boldsymbol{F}'\boldsymbol{F}}{T} &= \boldsymbol{I}_r,\ \boldsymbol{\Lambda}'\boldsymbol{\Lambda} \text{ is diagonal} \end{aligned}$$

Although no closed-form solution is available, the estimators can be obtained by iterations.

1. Obtain some initial values $\boldsymbol{\beta}^{(0)}$, such as least squares estimators from regressing $\boldsymbol{Y}_i$ on $\boldsymbol{X}_i$.
2. Perform principal component analysis for the pseudo-data, $\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(0)}$ to obtain $\boldsymbol{F}^{(1)}$ and $\Lambda^{(1)}$.
3. Next, regress $\boldsymbol{Y}_i - \boldsymbol{F}^{(1)}\boldsymbol{\lambda}_i^{(1)}$ on $\boldsymbol{X}_i$ to obtain $\boldsymbol{\beta}^{(1)}$.

4. Iterate such steps until convergence is achieved.

Bai (2009) showed that the resulting estimator $\hat{\beta}$ is $\sqrt{NT}$-consistent, and the limiting distributions for $\hat{F}$ and $\hat{\Lambda}$ are the same as in Bai (2003) due to their slower convergence rates. The limiting distribution for $\hat{\beta}$ depends on specific assumptions on ε_{it} and on the ratio T/N. If $T/N \to 0$, then the limiting distribution of $\hat{\beta}$ will be centered around zero, given that $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for $t \neq s$, and $E(\varepsilon_{it}\varepsilon_{js}) = \sigma_{ij}$ for all i, j, t.

On the other hand, if N and T are comparable such that $T/N \to K > 0$, then the limiting distribution will not be centered around zero, which poses a challenge for inference. Bai (2009) provided a bias-corrected estimator for β, whose limiting distribution is centered around zero. The bias-corrected estimator allows for heteroskedasticity across both N and T. Let $\tilde{\beta}$ be the bias-corrected estimator, assume that $T/N^2 \to 0$ and $N/T^2 \to 0$, $E\left(\varepsilon_{it}^2\right) = \sigma_{it}^2$, and $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for $i \neq j$ and $t \neq s$, then

$$\sqrt{NT}\left(\tilde{\beta} - \beta\right) \to_d N(0, \Sigma_\beta)$$

where a consistent estimator for Σ_β is available. Notice that the issue of choosing r remains.

Ahn et al. (2001, 2013) studied the model (26) under large N but fixed T. They employ the GMM method that applies moments of zero correlation and homoskedasticity. Moon and Weidner (2014) consider the same model as (26)), but allow lagged dependent variable as regressors. They devise a quadratic approximation of the profile objective function to show the asymptotic theory for the least square estimators and test statistics. Moon and Weidner (2015) extend their study by allowing unknown number of factors. They show that the limiting distribution of the least square estimator is not affected by the number of factors used in the estimation, as long as this number is no smaller than the true number of factors. Lu and Su (2015) propose the adaptive group LASSO (least absolute shrinkage and selection operator), which can simultaneously select the regressors and determine the number of factors.

Heterogenous panel models with interactive effects are also studied by Ando and Bai (2014), where the number of regressors can be large and the regularisation method is used to select relevant regressors. Ando and Bai (2015) provide a formal test for homogenous coefficients. The ML estimation of (26) is studied by Bai and Li (2014). They consider the case in which X_{it} also follows a factor structure and is jointly modeled.

Lu and Su (2015) consider the problem of determining the number of factors and selecting the proper regressors in linear dynamic panel data models with interactive fixed effects. Based on the preliminary estimates of the slope parameters and factors a la Bai and Ng (2009) and Moon and Weidner (2014a), they propose a method for simultaneous selection of regressors and factors and estimation through the method of adaptive group Lasso. With probability approaching one, this method can correctly select all relevant regressors and factors and shrink the coefficients of irrelevant regressors and redundant factors to zero.

Further, the shrinkage estimators of the nonzero slope parameters exhibit some oracle property. Monte Carlo simulations demonstrate the superb finite-sample performance of the proposed method.

YC: update, Moon and Weidner (2015)???

5 The Spatial-based Models of CSD

In general, an important research issue is to model the spatial dependence, the spatial heterogeneity and nonlinearity, simultaneously.

5.1 The Spatial Autoregressive (SAR) Process

Some models for cross sectional data may capture spatial interactions across spatial units. Consider the first-order spatial autoregressive (SAR) process:

$$y_i = \lambda \boldsymbol{w}_{i,n} \boldsymbol{Y}_n + \varepsilon_i, \quad i = 1, ..., n, \tag{27}$$

where $\boldsymbol{Y}_n = (y_1, ..., y_n)'$ is an $n \times 1$ vector of dependent variable, $\boldsymbol{w}_{i,n}$ is a $1 \times n$ row vector of weights, and $\varepsilon_i \sim iid(0, \sigma^2)$. We write (389) in the matrix form:

$$\boldsymbol{Y}_n = \lambda \boldsymbol{W}_n \boldsymbol{Y}_n + \boldsymbol{E}_n \tag{28}$$

where $\boldsymbol{W}_n \boldsymbol{Y}_n$ is 'the spatial lag'. Under the assumption that $\boldsymbol{S}_n(\lambda) = \boldsymbol{I}_n - \lambda \boldsymbol{W}_n$ is nonsingular, we have:

$$\boldsymbol{Y}_n = \boldsymbol{S}_n(\lambda)^{-1} \boldsymbol{E}_n.$$

Also consider the regression model with SAR disturbance:

$$\boldsymbol{Y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{U}_n, \quad \boldsymbol{U}_n = \rho \boldsymbol{W}_n \boldsymbol{U}_n + \boldsymbol{E}_n \tag{29}$$

The disturbances in $\boldsymbol{U}_n$ are spatially-correlated. The variance matrix of $\boldsymbol{U}_n$ is $\sigma^2 \boldsymbol{S}_n(\rho)^{-1} \boldsymbol{S}_n(\rho)^{-1\prime}$. As the off-diagonal elements of $\boldsymbol{S}_n(\rho)^{-1} \boldsymbol{S}_n(\rho)^{-1\prime}$ may be nonzero, u_i's are cross-sectionally correlated across units.

Spatial autoregressive model with covariates This generalises SAR by incorporating exogenous variables $\boldsymbol{x}_i$:

$$\boldsymbol{Y}_n = \lambda \boldsymbol{W}_n \boldsymbol{Y}_n + \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{E}_n \tag{30}$$

where $\boldsymbol{E}_n \sim iid(0, \sigma^2 \boldsymbol{I}_n)$. This model has the feature of a simultaneous equation model and its reduced form is:

$$\boldsymbol{Y}_n = \boldsymbol{S}_n(\lambda)^{-1} \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{S}_n(\lambda)^{-1} \boldsymbol{E}_n.$$

Some Intuitions on Spatial Weights Matrix, $\boldsymbol{W}_n$ The jth element of $\boldsymbol{w}_{i,n}$, $w_{n,ij}$, represents the link (or distance) between the neighbor j and the spatial unit i. The diagonal of $\boldsymbol{W}_n$ is specified to be zero, i.e., $w_{n,ii} = 0$ for all i, because $\lambda \boldsymbol{w}_{i,n}$ represents the effect of other spatial units on the spatial unit i. It is a common practice to have $\boldsymbol{W}_n$ having a zero diagonal and being row-normalized such that the summation of elements in each row of $\boldsymbol{W}_n$ is unity. In some applications, the ith row $\boldsymbol{w}_{i,n}$ of $\boldsymbol{W}_n$ may be constructed as $\boldsymbol{w}_{i,n} = (d_{i1}, d_{i2}, ..., d_{in}) / \sum_{j=1}^{n} d_{ij}$, where $d_{ij} \geq 0$, represents a function of the spatial distance between i and j, in some space. The weighting operation may be interpreted as an average of neighboring values.

When neighbors are defined as adjacent ones for each unit, the correlation is local in the sense that correlations will be stronger for neighbors but weak for units far away. Suppose that $\|\rho \boldsymbol{W}_n\| \leq 1$ for matrix norm $\|.\|$, then

$$\boldsymbol{S}_n(\rho)^{-1} = \boldsymbol{I}_n + \sum_{i=1}^{\infty} \rho^i \boldsymbol{W}_n^i$$

Notice that

$$\left\| \sum_{i=m}^{\infty} \rho^i \boldsymbol{W}_n^i \right\| \leq |\rho \boldsymbol{W}_n|^i \left\| \boldsymbol{S}_n(\rho)^{-1} \right\|$$

If $\boldsymbol{W}_n$ is row-normalized, then

$$\left\| \sum_{i=m}^{\infty} \rho^i \boldsymbol{W}_n^i \right\|_{\infty} \leq \sum_{i=m}^{\infty} \left| \rho^i \right| = \frac{|\rho|^m}{1 - |\rho|}$$

will be small as m gets larger. $\boldsymbol{U}_n$ can be represented as

$$\boldsymbol{U}_n = \boldsymbol{E}_n + \rho \boldsymbol{W}_n \boldsymbol{E}_n + \rho^2 \boldsymbol{W}_n^2 \boldsymbol{E}_n + ...,$$

where $\rho \boldsymbol{W}_n$ may represent the influence of neighbors on each unit, $\rho^2 \boldsymbol{W}_n^2$ is the second layer neighborhood influence, *etc.* **In the social interactions literature, $W_n S_n(\rho)^{-1}$ is a vector of measures of centrality**, which summaries the position of each spatial unit in a network.

In conventional spatial cases, neighbor units are defined by only a few adjacent ones. However, there are cases where 'neighbors' may consist of many units. An example is a social interactions model, where 'neighbors' refer to individuals in a same group. The latter may be regarded as models with large group interactions. For models with a large number of interactions for each unit, the spatial weights matrix $\boldsymbol{W}_n$ will associate with the sample size. Suppose that there are R groups and there are m individuals in each group, with the sample size, $n = mR$. In a special network (e.g., friendship), one may assume that each individual in a group is given an equal weight. In that case,

$$\boldsymbol{W}_n = \boldsymbol{I}_R \otimes \boldsymbol{B}_m \text{ with } \boldsymbol{B}_m = \left(\boldsymbol{\ell}_m \boldsymbol{\ell}_m' - \boldsymbol{I}_m \right) / (m-1),$$

where $\boldsymbol{\ell}_m$ is the m-dimensional vector of ones. In general, the number of members in each district may be large but have different sizes. This model has many interesting applications in the social interactions.

Other generalizations We may combine the SAR equation with SAR disturbances:

$$\boldsymbol{Y}_n = \lambda \boldsymbol{W}_n \boldsymbol{Y}_n + \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{U}_n, \ \boldsymbol{U}_n = \rho \boldsymbol{M}_n \boldsymbol{U}_n + \boldsymbol{E}_n \tag{31}$$

where $\boldsymbol{W}_n$ and $\boldsymbol{M}_n$ are spatial weights matrices, which may not be identical.

Further extension of a SAR model may allow high-order spatial lags as in

$$\boldsymbol{Y}_n = \sum_{j=1}^{p} \lambda_j \boldsymbol{W}_{jn} \boldsymbol{Y}_n + \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{E}_n \tag{32}$$

where $\boldsymbol{W}_{jn}$'s are p distinct spatial weights matrices.

5.2 Estimation Methods

We consider the QML, the 2SLS (IV), and the generalized method of moments (GMM). The QML has usually good finite sample properties. However, the ML method is not computationally attractive for the higher spatial lags model, in which case the IV and GMM methods may be feasible, (Lee, 2007).

MLE For the SAR process in (386), we have the log-likelihood function:

$$\ln L_n\left(\lambda, \sigma^2\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln\left|\boldsymbol{S}_n\left(\lambda\right)\right| - \frac{1}{2\sigma^2}\boldsymbol{Y}_n'\boldsymbol{S}_n\left(\lambda\right)'\boldsymbol{S}_n\left(\lambda\right)\boldsymbol{Y}_n, \tag{33}$$

where $\boldsymbol{S}_n\left(\lambda\right) = \boldsymbol{I}_n - \lambda \boldsymbol{W}_n$. For the model with SAR disturbances, (388), the log likelihood function is

$$\begin{aligned} \ln L_n\left(\rho, \beta, \sigma^2\right) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln\left|\boldsymbol{S}_n\left(\rho\right)\right| \\ &\quad -\frac{1}{2\sigma^2}\left(\boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\right)'\boldsymbol{S}_n\left(\rho\right)'\boldsymbol{S}_n\left(\rho\right)\left(\boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\right) \end{aligned} \tag{34}$$

The log likelihood function for the SAR model with covariates in (30) is

$$\begin{aligned} \ln L_n\left(\lambda, \beta, \sigma^2\right) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln\left|\boldsymbol{S}_n\left(\lambda\right)\right| \\ &\quad -\frac{1}{2\sigma^2}\left(\boldsymbol{Y}_n\boldsymbol{S}_n\left(\lambda\right) - \boldsymbol{X}_n\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}_n\boldsymbol{S}_n\left(\lambda\right) - \boldsymbol{X}_n\boldsymbol{\beta}\right) \end{aligned} \tag{35}$$

The likelihood function involves the computation of the determinant of $\boldsymbol{S}_n\left(\lambda\right)$, which is a function of the unknown parameter λ, and may have a large dimension n. A computationally tractable method is due to Ord (1975), where $\boldsymbol{W}_n$ is a row-normalized weights matrix with

$$\boldsymbol{W}_n = \boldsymbol{D}_n \boldsymbol{W}_n^*,$$

where $\boldsymbol{W}_n^*$ is a symmetric matrix and $\boldsymbol{D}_n = diag\left\{\sum_{j=1}^n w_{n,ij}^*\right\}^{-1}$. Though $\boldsymbol{W}_n$ is not symmetric, the eigenvalues of $\boldsymbol{W}_n$ are still all real. This is because

$$\begin{aligned}
|\boldsymbol{W}_n - v\boldsymbol{I}_n| &= |\boldsymbol{D}_n\boldsymbol{W}_n^* - v\boldsymbol{I}_n| \\
&= \left|\boldsymbol{D}_n\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}\boldsymbol{D}_n^{-1/2} - v\boldsymbol{D}_n^{1/2}\boldsymbol{D}_n^{-1/2}\right| \\
&= \left|\boldsymbol{D}_n^{1/2}\right|\left|\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2} - v\boldsymbol{I}_n\right|\left|\boldsymbol{D}_n^{-1/2}\right| \\
&= \left|\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2} - v\boldsymbol{I}_n\right|
\end{aligned}$$

Thus, the eigenvalues of $\boldsymbol{W}_n$ are the same as those of $\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}$, which is a symmetric matrix. As the eigenvalues of a symmetric matrix are real, the eigenvalues of $\boldsymbol{W}_n$ are real. Let μ_i's be the eigenvalues of $\boldsymbol{W}_n$, which are the same as those of $\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}$. Let $\boldsymbol{\Gamma}$ be the orthogonal matrix such that

$$\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2} = \boldsymbol{\Gamma}diag\left\{\mu_i\right\}\boldsymbol{\Gamma}'$$

The above relations show that

$$\begin{aligned}
|\boldsymbol{I}_n - \lambda\boldsymbol{W}_n| &= |\boldsymbol{I}_n - \lambda\boldsymbol{D}_n\boldsymbol{W}_n^*| = \left|\boldsymbol{I}_n - \lambda\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}\right| \\
&= \left|\boldsymbol{D}_n^{1/2}\boldsymbol{D}_n^{-1/2} - \lambda\boldsymbol{D}_n^{1/2}\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}\boldsymbol{D}_n^{-1/2}\right| \\
&= \left|\boldsymbol{D}_n^{1/2}\right|\left|\boldsymbol{I}_n - \lambda\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}\right|\left|\boldsymbol{D}_n^{-1/2}\right| \\
&= \left|\boldsymbol{I}_n - \lambda\boldsymbol{D}_n^{1/2}\boldsymbol{W}_n^*\boldsymbol{D}_n^{1/2}\right| = \left|\boldsymbol{I}_n - \lambda\boldsymbol{\Gamma}diag\left\{\mu_i\right\}\boldsymbol{\Gamma}'\right| \\
&= |\boldsymbol{I}_n - \lambda diag\left\{\mu_i\right\}| = \Pi_{i=1}^n\left(1 - \lambda\mu_i\right)
\end{aligned}$$

Thus, $|\boldsymbol{I}_n - \lambda\boldsymbol{W}_n|$ can be easily updated during iterations within a maximization subroutine, as μ_i's need be computed only once.

Another tractable method is the characteristic polynomial. The determinant, $|\boldsymbol{W}_n - \mu\boldsymbol{I}_n|$ is a polynomial in μ and is called the characteristic polynomial of $\boldsymbol{W}_n$ (the zeros of the characteristic polynomial are the eigenvalues of $\boldsymbol{W}_n$). Thus,

$$|\boldsymbol{I}_n - \lambda\boldsymbol{W}_n| = a_n\lambda^n + ... + a_1\lambda_1 + a_0,$$

where the constant a depends only on $\boldsymbol{W}_n$. So a's can be computed once during the maximization algorithm.

2SLS estimation For the SAR model with covariates, (30), the spatial lag $\boldsymbol{W}_n\boldsymbol{Y}_n$ can be correlated with the disturbance, $\boldsymbol{E}_n$. So OLS may not be a consistent estimator. However, there is a class of spatial $\boldsymbol{W}_n$ (with large group interaction) that the OLS estimator can be consistent (Lee 2002).

To avoid the bias due to the correlation of $\boldsymbol{W}_n\boldsymbol{Y}_n$ with $\boldsymbol{E}_n$, Kelejian and Prucha (1998) suggested the use of instrumental variables. Let $\boldsymbol{Q}_n$ be a matrix of IVs. Denote $\boldsymbol{Z}_n = (\boldsymbol{W}_n\boldsymbol{Y}_n, \boldsymbol{X}_n)$ and $\boldsymbol{\theta} = \left(\lambda, \boldsymbol{\beta}'\right)'$, and rewrite (30) as

$$\boldsymbol{Y}_n = \boldsymbol{Z}_n\boldsymbol{\theta} + \boldsymbol{E}_n \tag{36}$$

The 2SLS estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{2SLS} = \left[\boldsymbol{Z}'_n \boldsymbol{Q}_n \left(\boldsymbol{Q}'_n \boldsymbol{Q}_n\right)^{-1} \boldsymbol{Q}'_n \boldsymbol{Z}_n\right] \left[\boldsymbol{Z}'_n \boldsymbol{Q}_n \left(\boldsymbol{Q}'_n \boldsymbol{Q}_n\right)^{-1} \boldsymbol{Q}'_n \boldsymbol{Y}_n\right]$$

The asymptotic distribution of $\hat{\boldsymbol{\theta}}_{2SLS}$ follows:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{2SLS} - \boldsymbol{\theta}\right) \rightarrow_d N\left(0, \sigma^2 \left(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}, \boldsymbol{X}_n\right)' \boldsymbol{Q}_n \left(\boldsymbol{Q}'_n \boldsymbol{Q}_n\right)^{-1} \boldsymbol{Q}'_n \left(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}, \boldsymbol{X}_n\right)\right)$$

where $\boldsymbol{G}_n = \boldsymbol{W}_n \boldsymbol{S}_n^{-1}$, under the assumption that the limiting matrix $\frac{1}{n}\left(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}, \boldsymbol{X}_n\right)$ has the full column rank $(k+1)$ where k is the number of columns of $\boldsymbol{X}_n$. Kelejian and Prucha (1998) suggest the use of linearly independent variables in $(\boldsymbol{X}_n, \boldsymbol{W}_n \boldsymbol{X}_n)$ for the construction of $\boldsymbol{Q}_n$.

By the Schwartz inequality, the optimum IV matrix is $(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}, \boldsymbol{X}_n)$. This 2SLS cannot be used for the estimation of the (pure) SAR process with $\boldsymbol{\beta} = 0$. When $\boldsymbol{\beta} = 0$, $\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta} = 0$. Hence, $(\boldsymbol{G}_n \boldsymbol{X}_n \boldsymbol{\beta}, \boldsymbol{X}_n) = (\boldsymbol{0}, \boldsymbol{X}_n)$ would have rank k but not full rank $(k+1)$. Intuitively, when $\boldsymbol{\beta} = 0$, $\boldsymbol{X}_n$ does not appear in the model and there is no other IV available.

Method of moments Kelejian and Prucha (1999) suggest an MOM estimation:

$$\min_{\theta} \boldsymbol{g}'_n \left(\boldsymbol{\theta}\right) \boldsymbol{g}_n \left(\boldsymbol{\theta}\right).$$

The moment equations are based on three moments:

$$E\left(\boldsymbol{E}'_n \boldsymbol{E}_n\right) = n\sigma^2; \quad E\left(\boldsymbol{E}'_n \boldsymbol{W}'_n \boldsymbol{W}_n \boldsymbol{E}_n\right) = \sigma^2 tr\left(\boldsymbol{W}'_n \boldsymbol{W}_n\right); \quad E\left(\boldsymbol{E}'_n \boldsymbol{W}_n \boldsymbol{E}_n\right) = 0$$

In this case we have:

$$g_n \left(\boldsymbol{\theta}\right) = \begin{pmatrix} \boldsymbol{Y}'_n \boldsymbol{S}_n \left(\lambda\right)' \boldsymbol{S}_n \left(\lambda\right) \boldsymbol{Y}_n - n\sigma^2 \\ \boldsymbol{Y}'_n \boldsymbol{S}_n \left(\lambda\right)' \boldsymbol{W}'_n \boldsymbol{W}_n \boldsymbol{S}_n \left(\lambda\right) \boldsymbol{Y}_n - \sigma^2 tr\left(\boldsymbol{W}'_n \boldsymbol{W}_n\right), \\ \boldsymbol{Y}'_n \boldsymbol{S}_n \left(\lambda\right)' \boldsymbol{W}_n \boldsymbol{S}_n \left(\lambda\right) \boldsymbol{Y}_n \end{pmatrix}$$

For the regression model with SAR disturbances, $\boldsymbol{Y}_n$ shall be replaced by least squares residuals.

GMM estimation For the SAR model with covariates, we can obtain other moment equations in addition to $\boldsymbol{Q}_n$. Let $\boldsymbol{Q}_n$ be an $n \times k_x$ IV matrix constructed as functions of $\boldsymbol{X}_n$ and $\boldsymbol{W}_n$. Let

$$\boldsymbol{\varepsilon}_n \left(\boldsymbol{\theta}\right) = \boldsymbol{S}_n \left(\lambda\right) \boldsymbol{Y}_n - \boldsymbol{X}_n \boldsymbol{\beta}$$

for any possible $\boldsymbol{\theta}$. The orthogonality conditions, $\boldsymbol{Q}'_n \boldsymbol{\varepsilon}_n \left(\boldsymbol{\theta}\right) = 0$ provide the $k_x \times 1$ vector of moment conditions.

Now, consider a finite number, say m, of $n \times n$ constant matrices, $\boldsymbol{P}_{1n}, ..., \boldsymbol{P}_{mn}$, each of which has a zero diagonal. Then, $\left(\boldsymbol{P}_{jn} \boldsymbol{\varepsilon}_n \left(\boldsymbol{\theta}\right)\right)' \boldsymbol{\varepsilon}_n \left(\boldsymbol{\theta}\right)$ can be used as the

moment functions in addition to $\boldsymbol{Q}_n'\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})$. Then, we have the following moment conditions vector:

$$\begin{aligned}\boldsymbol{g}_n(\boldsymbol{\theta}) &= (\boldsymbol{P}_{1n}\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}),...,\boldsymbol{P}_{mn}\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}),\boldsymbol{Q}_n)'\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \\ &= \begin{pmatrix} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})'\boldsymbol{P}_{1n}(\boldsymbol{\theta})\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})'\boldsymbol{P}_{mn}(\boldsymbol{\theta})\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \\ \boldsymbol{Q}_n'\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \end{pmatrix}\end{aligned}$$

Proposition. For any constant $n \times\ n$ matrix $\boldsymbol{P}_n$ with $tr(\boldsymbol{P}_n)=0$, $\boldsymbol{P}_n\boldsymbol{\varepsilon}_n$ is uncorrelated with $\boldsymbol{\varepsilon}_n$, i.e., $E\left((\boldsymbol{P}_n\boldsymbol{\varepsilon}_n)'\boldsymbol{\varepsilon}_n\right)=0$.

Proof:

$$E\left((\boldsymbol{P}_n\boldsymbol{\varepsilon}_n)'\boldsymbol{\varepsilon}_n\right)=E\left(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_n'\boldsymbol{\varepsilon}_n\right)=E\left(\boldsymbol{\varepsilon}_n'\boldsymbol{P}_n\boldsymbol{\varepsilon}_n\right)=\sigma^2 tr(\boldsymbol{P}_n)=0$$

This shows that $E(\boldsymbol{g}_n(\boldsymbol{\theta}_0))=0$. Thus, $\boldsymbol{g}_n(\boldsymbol{\theta})$ are valid moment equations for GMM. Intuitively, as

$$\boldsymbol{W}_n\boldsymbol{Y}_n=\boldsymbol{G}_n\boldsymbol{X}_n\boldsymbol{\beta}_0+\boldsymbol{G}_n\boldsymbol{\varepsilon}_n \text{ with } \boldsymbol{G}_n=\boldsymbol{W}_n\boldsymbol{S}_n^{-1} \text{ and } \boldsymbol{S}_n=\boldsymbol{S}_n(\lambda_0),$$

and $\boldsymbol{G}_n\boldsymbol{\varepsilon}_n$ is correlated with the disturbance $\boldsymbol{\varepsilon}_n$ in the model,

$$\boldsymbol{Y}_n=\lambda\boldsymbol{W}_n\boldsymbol{Y}_n+\boldsymbol{X}_n\boldsymbol{\beta}+\boldsymbol{\varepsilon}_n,$$

hence, any $\boldsymbol{P}_{jn}\boldsymbol{\varepsilon}_n$, which is uncorrelated with $\boldsymbol{\varepsilon}_n$, can be used as IV for $\boldsymbol{W}_n\boldsymbol{Y}_n$ as long as $\boldsymbol{P}_{jn}\boldsymbol{\varepsilon}_n$ and $\boldsymbol{G}_n\boldsymbol{\varepsilon}_n$ are correlated.

5.3 The Spatial Dynamic Panel Data (SDPD) Model

Dynamic panel data models consider not only heterogeneity but also state dependence that cannot be handled by cross-sectional or static panel data models. The most general case is the "time-space dynamic" model (Anselin, Le Gallo, and Jayet, 2008),[7] which is termed spatial dynamic panel data (SDPD) model in Yu, de Jong and Lee(2008):

$$\boldsymbol{Y}_{nt}=\lambda_0\boldsymbol{W}_n\boldsymbol{Y}_{nt}+\gamma_0\boldsymbol{Y}_{n,t-1}+\rho_0\boldsymbol{W}_n\boldsymbol{Y}_{n,t-1}+\boldsymbol{X}_{nt}\boldsymbol{\beta}_0+\boldsymbol{c}_{n_0}+\alpha_{t0}\boldsymbol{l}_n+\boldsymbol{V}_{nt}, \quad (37)$$

where c_{n0} is $n\times 1$ column vector of fixed effects and α_{t0}'s are time effects. γ_0 captures the pure dynamic effect and ρ_0 captures the spatial-time or diffusion effect. Due to the presence of fixed individual and time effects, $\boldsymbol{X}_{nt}$ will not include any time invariant or individual invariant regressors.

[7] Anselin, Le Gallo, and Jayet (2008) divide spatial dynamic models into four categories, namely, "pure space recursive" if only a spatial time lag is included; "time-space recursive" if an individual time lag and a spatial time lag are included; "time-space simultaneous" if an individual time lag and a contemporaneous SL term are specified; and "time-space dynamic" if all forms of lags are included. Korniotis (2010) studies a time-space recursive model with fixed effects, which is applied to the growth of consumption in each state in the United States to investigate habit formation. Su and Yang (2007) derive the quasi-maximum likelihood (QML) estimation of the above model under both fixed and random effects specifications.

Define

$$\boldsymbol{S}_n(\lambda) = \boldsymbol{I}_n - \lambda \boldsymbol{W}_n;\ \boldsymbol{S}_n \equiv \boldsymbol{S}_n(\lambda_0) = I_n - \lambda_0 \boldsymbol{W}_n.$$

Presuming that $\boldsymbol{S}_n$ is invertible and denoting $\boldsymbol{A}_n = \boldsymbol{S}_n^{-1}(\gamma_0 \boldsymbol{I}_n + \rho_0 \boldsymbol{W}_n)$, (37) can be rewritten as

$$\boldsymbol{Y}_{nt} = \boldsymbol{A}_n \boldsymbol{Y}_{n,t-1} + \boldsymbol{S}_n^{-1}\left(\boldsymbol{X}_{nt}\boldsymbol{\beta}_0 + \boldsymbol{c}_{n_0} + \alpha_{t0}\boldsymbol{l}_n + \boldsymbol{V}_{nt}\right), \tag{38}$$

We study the eigenvalues of $\boldsymbol{A}_n$ by focusing on the case with $\boldsymbol{W}_n$ being row-normalized. Let $\varpi_n = diag\{\varpi_{n1}, ..., \varpi_{nn}\}$ be the $n \times n$ diagonal eigenvalues matrix of $\boldsymbol{W}_n$ such that $\boldsymbol{W}_n = \Gamma_n \varpi_n \Gamma_n^{-1}$ where Γ_n is the eigenvector matrix. Because $\boldsymbol{A}_n = \boldsymbol{S}_n^{-1}(\gamma_0 I_n + \rho_0 \boldsymbol{W}_n)$, the eigenvalues matrix of $\boldsymbol{A}_n$ is $\boldsymbol{D}_n = (I_n - \lambda_0 \boldsymbol{W}_n)^{-1}(\gamma_0 I_n + \rho_0 \boldsymbol{W}_n)$ such that $\boldsymbol{A}_n = \Gamma_n \boldsymbol{D}_n \Gamma_n^{-1}$. As $\boldsymbol{W}_n$ is row-normalized, all the eigenvalues are less than or equal to 1 in absolute value. Denote m_n as the number of unit eigenvalues of $\boldsymbol{W}_n$ and let the first m_n eigenvalues of $\boldsymbol{W}_n$ be the unity. $\boldsymbol{D}_n$ can be decomposed into two parts, one corresponding to the unit eigenvalues of $\boldsymbol{W}_n$, and the other corresponding to the eigenvalues of $\boldsymbol{W}_n$ which are smaller than 1. Define $J_n = diag\left(l'_{m_n}, 0, ..., 0\right)$ with l_{m_n} being an $m_n \times 1$ vector of ones and $\tilde{D}_n = diag\{0, ..., 0, d_{n,m_n+1}, ..., d_{nn}\}$, where $|d_{ni}| < 1$, for $i = m_n + 1, ..., n$. We have

$$\boldsymbol{A}_n^h = \left(\frac{\gamma_0 + \rho_0}{1 - \lambda_0}\right)^h \Gamma_n J_n \Gamma_n^{-1} + B_n^h \text{ with } B_n = \Gamma_n \tilde{D}_n \Gamma_n^{-1}$$

Depending on the value of $\frac{\gamma_0+\rho_0}{1-\lambda_0}$, we may divide the process into four cases.

- Stable case when $\gamma_0 + \rho_0 + \lambda_0 < 1$ (and with some other restrictions on three parameters).
- Spatial cointegration case when $\gamma_0 + \rho_0 + \lambda_0 = 1$ but $\gamma_0 < 1$.
- Unit roots case when $\gamma_0 + \rho_0 + \lambda_0 = 1$ and $\gamma_0 = 1$.
- Explosive case when $\gamma_0 + \rho_0 + \lambda_0 > 1$.

For stability, or in terms of stationarity in the time series notion, there are more restrictions on the parameter space of $(\gamma_0, \rho_0, \lambda_0)$ in addition to $\gamma_0 + \rho_0 + \lambda_0 < 1$. Such a parameter region can be revealed from conditions such that all the eigenvalues d_{ni}'s of $\boldsymbol{A}_n$ are less than one in absolute value. The eigenvalues of $\boldsymbol{A}_n$ are $d_{ni} = \frac{\gamma_0 + \rho_0 \varpi_{ni}}{1 - \lambda_0 \varpi_{ni}}$, where ϖ_{ni}'s are eigenvalues of $\boldsymbol{W}_n$. By regarding d as a function of ϖ, we have:

$$\frac{\partial}{\partial \varpi}\left(\frac{\gamma_0 + \rho_0 \varpi_{ni}}{1 - \lambda_0 \varpi_{ni}}\right) = \frac{\rho_0 + \lambda_0 \gamma_0}{(1 - \lambda_0 \varpi)^2}$$

Thus, we have three different situations:

1. $\rho_0 + \lambda_0 \gamma_0 > 0$ if and only if d_{ni} has the same increasing order as ϖ_{ni}.

2. $\rho_0 + \lambda_0\gamma_0 = 0$, i.e., separable space-time filter, if and only if d_{ni} is a constant (under this case, $d_{ni} = \gamma_0$).

3. $\rho_0 + \lambda_0\gamma_0 < 0$ if and only if d_{ni} has the decreasing order of ϖ_{ni}.

(38) expresses the model in terms of a space-time multiplier (Anselin, Le Gallo, and Jayet 2008), which specifies how the joint determination of the dependent variables is a function of both spatial and time lags of explanatory variables and disturbances of all spatial units. This is useful for calculating marginal effects of changes of exogenous variables on outcomes over time and across spatial units. LeSage and Pace (2009) have introduced the concept of direct impact, total impact, and indirect impact. In a SAR model:

$$\boldsymbol{Y}_n = \alpha_0 l_n + \lambda_0 \boldsymbol{W}_n \boldsymbol{Y}_n + \sum_{k=1}^{k_x} \beta_{k0} X_{nk} + \varepsilon_n,$$

where $\boldsymbol{W}_n$ does not depend on X_n, the impact of a regressor X_{nk} on $\boldsymbol{Y}_n$ is

$$\frac{\partial Y_n}{\partial X'_{nk}} = \left(I_n - \lambda_0 \boldsymbol{W}_n\right)^{-1}_{k0} \beta_{k0} \quad \text{for the } k\text{th regressor.}$$

The average direct, average total and average indirect impacts are defined as

$$f_{k,direct}\left(\theta_0\right) \equiv \frac{1}{n} tr\left(\left(I_n - \lambda_0 \boldsymbol{W}_n\right)^{-1} \beta_{k0}\right),$$

$$f_{k,total}\left(\theta_0\right) \equiv \frac{1}{n} l'_n \left(\left(I_n - \lambda_0 \boldsymbol{W}_n\right)^{-1} \beta_{k0}\right) l_n,$$

$$f_{k,indirect}\left(\theta_0\right) \equiv f_{k,total}\left(\theta_0\right) - f_{k,direct}\left(\theta_0\right),$$

with l_n being an n-dimensional column of ones.

Debarsy, Ertur and LeSage (2012) extend such impact analyses to spatial dynamic panel models to investigate diffusion effects over time. Lee and Yu (2012b) extend the impact analysis to the case with time-varying spatial weights. Elhorst (2012) points out various restrictions on spatial dynamic panel models with marginal effects implied by specified models.

For an impact analysis for SDPD model, the change in exogenous variables will only influence the dependent variable of current period, but not the future ones. By changing the value of a regressor by the same amount across all spatial units in some consecutive time periods, say, from the time period t_1 to t_2. Thus, we have

$$\frac{\partial E\left(\boldsymbol{Y}_{nt}\right)}{\partial x} = \beta_0 \sum_{h=t-t_2}^{t-t_1} \boldsymbol{A}_n^h \boldsymbol{S}_n^{-1} l_n$$

where x is a regressor with its coefficient being β_0. Hence, by denoting $\theta = (\lambda, \gamma, \rho, \beta)$, the average total impact is

$$f_{t,total}\left(\theta_0\right) \equiv \frac{1}{n} l'_n \frac{\partial E\left(\boldsymbol{Y}_{nt}\right)}{\partial x} = \beta_0 \sum_{h=t-t_2}^{t-t_1} \frac{1}{n} \left[l'_n A_n^h S_n^{-1} l_n\right],$$

and similarly for other impacts. Debarsy, Ertur and LeSage (2012) study the case of $t_2 = t$ so that

$$f_{t,total}(\theta_0) \equiv \beta_0 \sum_{h=0}^{t-t_1} \frac{1}{n} \left[l_n' A_n^h S_n^{-1} l_n \right],$$

and the object of interest is how a permanent change in $X_{n,t1}$ will affect the future horizons (cumulatively) until t.

5.4 The Spatial Durbin Model

Elhorst (2012) proposes the following general spatial Durbin-type model:

$$\boldsymbol{Y}_t = \tau \boldsymbol{Y}_{t-1} + \delta \boldsymbol{W}\boldsymbol{Y}_t + \eta \boldsymbol{W}\boldsymbol{Y}_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\beta}_2 + \boldsymbol{X}_{t-1}\boldsymbol{\beta}_3 + \boldsymbol{W}\boldsymbol{X}_{t-1}\boldsymbol{\beta}_4 + \boldsymbol{Z}_t\boldsymbol{\theta} + \boldsymbol{v}_t \tag{39}$$

$$\boldsymbol{v}_t = \gamma \boldsymbol{v}_{t-1} + \rho \boldsymbol{W} v_t + \boldsymbol{\mu} + \lambda_t \boldsymbol{i}_N + \boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\mu} = \kappa \boldsymbol{W}\boldsymbol{\mu} + \boldsymbol{\xi}$$

where $\boldsymbol{Y}_t$ is an $N \times 1$ vector of the dependent variable for spatial unit ($i = 1, \ldots, N$) observed at time ($t = 1, \ldots, T$), $\boldsymbol{X}_t$ is an $N \times K$ matrix of exogenous regressors, and $\boldsymbol{Z}_t$ is an $N \times L$ matrix of endogenous regressors. The $N \times N$ matrix $\boldsymbol{W}$ is a nonnegative spatial-weight matrix with its diagonal elements being zero. τ, δ and η are the scalar parameters on $\boldsymbol{Y}_{t-1}$, $\boldsymbol{W}\boldsymbol{Y}_t$, and $\boldsymbol{W}\boldsymbol{Y}_{t-1}$. $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, and $\boldsymbol{\beta}_4$ are the $K \times 1$ vectors of the parameters on exogenous regressors and $\boldsymbol{\theta}$ is the $L \times 1$ vector of the parameters on endogenous regressors. $\boldsymbol{v}_t$ is the $N \times 1$ vector of error term, which may be allowed to be serially and spatially correlated. $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)'$ is the $N \times 1$ vector of the spatial-specific effects, and λ_t ($t = 1, \ldots T$) denotes time effects with $\boldsymbol{i}_N$ an $N \times 1$ vector of ones. We may allow the spatial-specific effects to be spatially autocorrelated. Finally, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$ and $\boldsymbol{\xi}$ are vectors of *iid* disturbances with zero mean and finite variance σ^2 and σ_ξ^2.

Stationarity conditions We impose the restrictions on the parameters and $\boldsymbol{W}$ to obtain the stationarity. The characteristic roots of $(I - \delta \boldsymbol{W})^{-1}(\tau I + \eta \boldsymbol{W})$ in (39) should lie within the unit circle as follows:

$$\begin{aligned}
\tau &< 1 - (\delta + \eta)\,\omega_{\max}, \quad if \quad \delta + \eta \geq 0 \\
\tau &< 1 - (\delta + \eta)\,\omega_{\min}, \quad if \quad \delta + \eta < 0 \\
-1 + (\delta - \eta)\,\omega_{\max} &< \tau, \quad if \quad \delta - \eta \geq 0 \\
-1 + (\delta - \eta)\,\omega_{\min} &< \tau, \quad if \quad \delta - \eta < 0
\end{aligned} \tag{40}$$

where $\omega_{\min}$ is the smallest (most negative) and $\omega_{\max}$ the largest real characteristic root of $\boldsymbol{W}$.

Remark: The restriction $\tau + \delta + \eta < 1$ in Lee and Yu (2010a) is not too restrictive. The stationarity conditions in (40) are more difficult to work with.

Estimation methods The spatial model has been estimated mainly by the iterative ML estimation procedure developed by Anselin (1988). Three estimation methods: (i) the bias-corrected quasi-maximum likelihood (QML) estimator, (ii) the IV or GMM estimator, and (iii) the Bayesian Markov Chain Monte Carlo (MCMC) approach.

Yu et al. (2008) construct a bias-corrected estimator for a dynamic model with $(\boldsymbol{Y}_{t-1}, \boldsymbol{WY}_t, \boldsymbol{WY}_{t-1})$ and spatial fixed effects. Lee and Yu (2010d) extend it to include time effects. They provide an asymptotic theory for bias-corrected LSDV (BCLSDV) estimator when both N and T tend to infinity, but T cannot be too small relative to N.

A few studies consider IV/GMM estimators, building on works by Arrelano and Bond (1991) and Blundell and Bond (1998). Elhorst (2010d) extends the FD-GMM to include endogenous interaction effects and finds that this estimator can still be severely biased. Lee and Yu (2010c) show that a 2SLS estimator, which is based on lagged values of $\boldsymbol{Y}_{t-1}$, $\boldsymbol{WY}_{t-1}$, $\boldsymbol{X}_t$ and $\boldsymbol{WY}_t$, is inconsistent due to too many moments; the dominant bias is caused by the endogeneity of $\boldsymbol{WY}_t$. They propose an optimal GMM estimator. Kukenova and Monteiro (2009) and Jacobs et al. (2009) consider a dynamic panel data model with $(\boldsymbol{Y}_{t-1}, \boldsymbol{WY}_t)$ and extend the system GMM to account for endogenous interaction effects. GMM can also be used to instrument endogenous explanatory variables (other than $\boldsymbol{Y}_{t-1}$ and $\boldsymbol{WY}_t$).

The dynamic spatial Durbin model Burridge (1981) recommends the first-order spatial autoregressive distributed lag model, in which $\boldsymbol{Y}$ is regressed on $\boldsymbol{WY}$ and $\boldsymbol{X}$ and $\boldsymbol{WX}$. This is known as the spatial Durbin model. The cost of ignoring spatial dependence in the dependent/independent variables is relatively high (biased) whilst ignoring spatial dependence in the disturbances will only cause a loss of efficiency (LeSage and Pace, 2009). The spatial Durbin model produces unbiased estimates, even if the DGP contains a spatial error.

If explanatory variables are endogenous, the best estimation method is the IV/GMM estimator (Fingleton and LeGallo, 2008).[8] Elhorst et al. (2010b) propose a dynamic spatial Durbin model:

$$\boldsymbol{Y}_t = \tau \boldsymbol{Y}_{t-1} + \delta \boldsymbol{WY}_t + \eta \boldsymbol{WY}_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{WX}_t\boldsymbol{\beta}_2 + \boldsymbol{v}_t \tag{41}$$

Rewriting (41) as

$$\boldsymbol{Y} = (\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} (\tau \boldsymbol{I} + \eta \boldsymbol{W}) \boldsymbol{Y}_{t-1} + (\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} (\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{WX}_t\boldsymbol{\beta}_2) + (\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} \boldsymbol{v}_t, \tag{42}$$

we can derive the partial derivatives of $\boldsymbol{Y}$ with respect to the kth explanatory variable of $\boldsymbol{X}$ at time t by

$$\left[\begin{array}{ccc} \frac{\partial Y}{\partial x_{1k}} & \cdots & \frac{\partial Y}{\partial x_{Nk}} \end{array} \right]_t = (\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} (\beta_{1k} \boldsymbol{I}_N + \beta_{2k} \boldsymbol{W}) \tag{43}$$

[8] The studies on growth and convergence typically regress economic growth on economic growth in neighboring economies, and the initial income level, the rates of saving, population growth, technological change, and depreciation in the own and in neighboring countries.

These denote the effect of a change of an explanatory variable in a spatial unit on the dependent variable of all other units in the short term. Similarly, the long-term effects can be:

$$\left[\begin{array}{ccc} \frac{\partial Y}{\partial x_{1k}} & \cdots & \frac{\partial Y}{\partial x_{Nk}} \end{array} \right] = \left[(1-\tau)\, \boldsymbol{I}_N - (\delta+\eta)\, \boldsymbol{W} \right]^{-1} (\beta_{1k} \boldsymbol{I}_N + \beta_{2k} \boldsymbol{W}) \quad (44)$$

(43) and (44) show that the short-term indirect effects do not occur if both $\delta = 0$ and $\beta_{2k} = 0$ while the long-term indirect effects do not occur if both $\delta = -\eta$ and $\beta_{2k} = 0$. By simulating the effects of shocks in $\boldsymbol{v}_t$, it is possible to find the path along which an economy moves to its long term equilibrium (De Groot and Elhorst 2010).

The dynamic spatial Durbin model can be used to determine direct effects and indirect (spatial spillover) effects in the short- and long-term. Anselin et al. (2008) criticise that this model might suffer from identification problems. By continuous substitution of $\boldsymbol{Y}_{t-1}$ up to $\boldsymbol{Y}_1$ in (42), we have:

$$\begin{aligned} \boldsymbol{Y} \;=\; & \left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-T} \left(\tau \boldsymbol{I}_N + \eta \boldsymbol{W}\right)^{T} \boldsymbol{Y}_{t-T} \\ & + \sum_{p=1}^{T} \left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-p} \left(\tau \boldsymbol{I}_N + \eta \boldsymbol{W}\right)^{p-1} \left(\boldsymbol{X}_{t-(p-1)} \boldsymbol{\beta}_1 + \boldsymbol{W} \boldsymbol{X}_{t-(p-1)} \boldsymbol{\beta}_2 + \boldsymbol{v}_{t-(p-1)}\right). \end{aligned} \quad (45)$$

This shows that two global spatial multiplier matrices, $\left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-p}$ and $\left(\tau \boldsymbol{I}_N + \eta \boldsymbol{W}\right)^{p-1}$, are at work at the same time in conjunction with one process that produces local spatial spillover effects, $\boldsymbol{W} \boldsymbol{X}_{t-(p-1)} \boldsymbol{\beta}_2$.[9] Second, more empirical research is needed to find out whether the short-term and long-term direct and indirect effects make sense.

To avoid identification problems, 4 restrictions are imposed. The first restriction is $\boldsymbol{\beta}_2 = 0$ in which the local indirect effects (spatial spillover) are zero. The indirect effects in relation to the direct effects become the same for every explanatory variable both in the short- and the long-term. For example, the ratio for the kth explanatory variable in the short term takes the form:

$$\frac{\left[\left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-1} \beta_{1k} \boldsymbol{I}_N\right]^{rsum}}{\left[\left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-1} \beta_{1k} \boldsymbol{I}_N\right]^{\bar{d}}} = \frac{\left[\left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-1}\right]^{rsum}}{\left[\left(\boldsymbol{I}_N - \delta \boldsymbol{W}\right)^{-1}\right]^{\bar{d}}}$$

where $\bar{d}$ is the operator that calculates the mean diagonal element of a matrix and $rsum$ is the operator that calculates the mean row sum of the non-diagonal elements.[10] This is independent of β_{1k} and thus the same for. A similar result in the long term.

[9] That is one process too much. One should examine whether the log-likelihood function is flat or almost flat. Hendry (1995) recommends to regress Y_t on Y_{t-1}, X_t and X_{t-1} as a generalization of the first-order autocorrelation model for time-series while Burridge (1981) recommends to regress Y_t on WY_t, X_t and WX_t as a generalization of the first-order spatial autocorrelation model for cross-section data. Elhorst (2001) suggests to regress Y_t on Y_{t-1}, WY_t, WY_{t-1}, X_t, WX_t, X_{t-1} and WX_{t-1}. This extension, however, worsens the identification problem.

[10] LeSage and Pace (2009) propose to report one direct effect measured by the average of the diagonal elements, and one indirect effect measured by the average of the row sums of the non-diagonal elements of that matrix.

The second restriction is $\delta = 0$ in which case $(\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} = \boldsymbol{I}_N$. Thus, the global short-term indirect (spatial spillover) effect is zero. This model is less suitable if the analysis focuses on spatial spillover effects in the short term.

The third restriction is $\eta = -\tau\delta$ (Parent and LeSage, 2011). The advantage is that the impact of a change in one of the explanatory variables on the dependent variable can be decomposed into a spatial effect and a time effect; the impact over space falls by $\delta \boldsymbol{W}$ for every higher-order neighbor, and over time by the factor τ for every period. The disadvantage is that the indirect (spatial spillover) effects in relation to the direct effects remain constant over time. The ratio of the kth explanatory variable takes the form:

$$\left[(\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} (\beta_{1k} \boldsymbol{I}_N + \beta_{2k} \boldsymbol{W})\right]^{rsum} / \left[(\boldsymbol{I}_N - \delta \boldsymbol{W})^{-1} (\beta_{1k} \boldsymbol{I}_N + \beta_{2k} \boldsymbol{W})\right]^{\bar{d}}$$

both in the short term and the long term.

The fourth restriction is $\eta = 0$. Although this model limits the flexibility of the ratio between indirect and direct effects, it seems to be the least restrictive model.

5.5 QMLE of Heterogeneous Spatial Models

Recently, Aquaro, Bailey and Pesaran (2015) extend the spatial autoregressive panel data model to the case where the spatial coefficients differ across the spatial units. They develop the QML estimator which is shown to be consistent and asymptotically normally distributed when both the time and cross section dimensions are large. Small sample properties of the proposed estimators are investigated by Monte Carlo simulations for Gaussian and non-Gaussian errors, and with spatial weight matrices of differing degree of sparseness, showing that the QML estimators have satisfactory small sample properties under certain sparsity conditions on the spatial weight matrix.

Consider the heterogeneous spatial autoregressive (HSAR) model:

$$y_{it} = \psi_i \sum_{j=1}^{N} w_{ij} y_{jt} + \varepsilon_{it},\ i = 1, ..., N; t = 1, ..., T \tag{46}$$

where $\boldsymbol{w}_i' \boldsymbol{y}_t = \sum_{j=1}^{N} w_{ij} y_{jt}$, $\boldsymbol{w}_i = (w_{i1}, ..., w_{iN})'$ with $w_{ii} = 0$, and $\boldsymbol{y}_t = (y_{1t}, ..., y_{Nt})'$. Here, $\boldsymbol{w}_i$ denotes an $N \times 1$ non-stochastic vector. Stacking the observations on individual units for each t, we have

$$(\boldsymbol{I}_N - \boldsymbol{\Psi} \boldsymbol{W})\, \boldsymbol{y}_t = \boldsymbol{\varepsilon}_t,\ t = 1, ..., T$$

where $\Psi = diag(\boldsymbol{\psi})$ with $\boldsymbol{\psi} = (\psi_1, ..., \psi_N)'$. The true value of ψ_i will be denoted by ψ_{i0}. Define:

$$\boldsymbol{S}(\boldsymbol{\psi}) = (\boldsymbol{I}_N - \boldsymbol{\Psi} \boldsymbol{W}), \quad \text{and } \boldsymbol{S}_0 = (\boldsymbol{I}_N - \boldsymbol{\Psi}_0 \boldsymbol{W}).$$

Assumption 1 The $N \times N$ spatial weight matrix, $\boldsymbol{W} = (w_{ij})$ is exactly sparse such that

$$h_N = \max_{i \leq N} \sum_{j \leq N} I(w_{ij} \neq 0);$$

is bounded in N, where $I(A)$ denotes the indicator function.

Assumption 2 ε_{it} are independently distributed over i and t, have zero means, and constant variances, $E\left(\varepsilon_{it}^2\right) = \sigma^2$.

Assumption 3 The $(N+1) \times 1$ parameter vector, $\boldsymbol{\theta} = \left(\boldsymbol{\psi}', \sigma^2\right)' \in \Theta$ is a subset of the $N+1$ dimensional Euclidean space, $\mathbb{R}^{N+1}$. Θ is a closed and bounded (compact) set and includes the true value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$ as an interior point.

Assumption 4 $\lambda_{\min}\left[\boldsymbol{S}'\left(\boldsymbol{\psi}\right)\boldsymbol{S}\left(\boldsymbol{\psi}\right)\right] > 0$ for all values of $\boldsymbol{\psi} \in \Theta$ and N.

Assumption 5 Let $\boldsymbol{y}_t = (y_{1t}, ..., y_{Nt})'$, and consider the sample covariance matrix of $\boldsymbol{y}_t$, $\hat{\boldsymbol{\Sigma}}_T = T^{-1}\sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t'$. For a given N we have (as $T \to \infty$)

$$\hat{\boldsymbol{\Sigma}}_T \to_p \boldsymbol{\Sigma}_0 \text{ uniformly in } \boldsymbol{\theta} = \left(\boldsymbol{\psi}', \sigma^2\right)',$$

where

$$\boldsymbol{\Sigma}_0 = \sigma_0^2 \left(\boldsymbol{I}_N - \boldsymbol{\Psi}_0 \boldsymbol{W}\right)^{-1} \left(\boldsymbol{I}_N - \boldsymbol{\Psi}_0 \boldsymbol{W}\right)^{-1\prime} = \sigma_0^2 \boldsymbol{S}_0^{-1} \boldsymbol{S}_0^{-1\prime}.$$

Under Assumption 1 and under the following condition:

$$\sup_i |\psi_i| < \max\left\{\frac{1}{\|\boldsymbol{W}\|_1}, \frac{1}{\|\boldsymbol{W}\|_\infty}\right\}$$

then $\boldsymbol{S}\left(\boldsymbol{\psi}\right) = \left(\boldsymbol{I}_N - \boldsymbol{\Psi}\boldsymbol{W}\right)$ is non-singular and (46) can be expressed as

$$\boldsymbol{y}_t = \boldsymbol{S}^{-1}\left(\boldsymbol{\psi}\right)\boldsymbol{\varepsilon}_t; \; t = 1, ..., T \tag{47}$$

Then, the joint density function of $y_{1t}, ..., y_{Nt}$ is given by

$$\prod_{t=1}^{T} \frac{\boldsymbol{S}\left(\boldsymbol{\psi}\right)}{\left|\sigma^2 \boldsymbol{I}_N\right|^{1/2} \left(2\pi\right)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}_t' \boldsymbol{S}'\left(\boldsymbol{\psi}\right)\boldsymbol{S}\left(\boldsymbol{\psi}\right)\boldsymbol{y}_t\right)$$

and the (quasi) log-likelihood function can be written as

$$\ell\left(\boldsymbol{\theta}\right) = -\frac{NT}{2}\ln 2\pi - \frac{NT}{2}\ln\sigma^2 + T\ln\left|\boldsymbol{S}\left(\boldsymbol{\psi}\right)\right| - \frac{1}{2\sigma^2}\sum_{t=1}^{T} \boldsymbol{y}_t' \boldsymbol{S}'\left(\boldsymbol{\psi}\right)\boldsymbol{S}\left(\boldsymbol{\psi}\right)\boldsymbol{y}_t$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\psi}', \sigma^2\right)'$. The last term of the LLF can be written conveniently as

$$\sum_{t=1}^{T} \mathbf{y}_t' \mathbf{S}\left(\boldsymbol{\psi}\right)' \mathbf{S}\left(\boldsymbol{\psi}\right)\mathbf{y}_t = T tr\left[\mathbf{S}\left(\boldsymbol{\psi}\right)' \mathbf{S}\left(\boldsymbol{\psi}\right)\hat{\Sigma}_T\right]$$

Hence,

$$\ell\left(\boldsymbol{\theta}\right) = -\frac{NT}{2}\ln 2\pi - \frac{NT}{2}\ln\sigma^2 + T\ln\left|\mathbf{S}\left(\boldsymbol{\psi}\right)\right| - \frac{T}{2\sigma^2}\left[\mathbf{S}\left(\boldsymbol{\psi}\right)' \mathbf{S}\left(\boldsymbol{\psi}\right)\hat{\Sigma}_T\right] \tag{48}$$

Proposition 1 (**The main identification**) Consider the HSAR model (46) and suppose that Assumptions 1 to 5 hold and the invertibility condition (47) is met. Then the true parameter values, σ_0^2 and ψ_{i0}, for $i = 1, 2, ..., N$, are identified if $\lambda_{\min}\left[\Lambda_N\left(\bar{\boldsymbol{\varphi}}\right)\right] > 0$, where

$$\boldsymbol{\Lambda}_N\left(\boldsymbol{\varphi}\right) = \begin{bmatrix} \left(\boldsymbol{A}_0 \odot \boldsymbol{A}_0'\right) + (1-\delta)\, diag\left(\boldsymbol{G}_0\boldsymbol{G}_0'\right) & diag\left(\boldsymbol{G}_0\right)\boldsymbol{\tau}_N - diag\left(\boldsymbol{G}_0\boldsymbol{G}_0'\boldsymbol{D}\right)\boldsymbol{\tau}_N \\ \left[diag\left(\boldsymbol{G}_0\right)\boldsymbol{\tau}_N - diag\left(\boldsymbol{G}_0\boldsymbol{G}_0'\boldsymbol{D}\right)\boldsymbol{\tau}_N\right]' & \frac{N}{2(1-\delta)^2} \end{bmatrix}$$

$\boldsymbol{\tau}_N$ is an $N \times 1$ vector of ones, $\odot$ is the Hadamard product matrix operator,

$$\boldsymbol{A}_0 = \boldsymbol{G}_0\left(\boldsymbol{I}_N - \boldsymbol{D}\boldsymbol{G}_0\right)^{-1} = \boldsymbol{W}\left(\boldsymbol{I}_N - \boldsymbol{\Psi}\boldsymbol{W}\right)^{-1}.$$

For large N, $\lambda_{\min}\left[\Lambda_N\left(\bar{\boldsymbol{\varphi}}\right)\right] > 0$ is necessary for identification but need not be sufficient if the aim is to identify all the spatial coefficients, ψ_i for $i = 1, 2, ..., N$. For local identification the condition simplifies to

$$\lambda_{\min}\left[\boldsymbol{H}_0\right] > \epsilon > 0, \ for\ all\ N,$$

$$\boldsymbol{H}_0 = \left[\left(\boldsymbol{G}_0 \odot \boldsymbol{G}_0'\right) + diag\left(\boldsymbol{G}_0\boldsymbol{G}_0'\right)\right] - \frac{2}{N} diag\left(\boldsymbol{G}_0\right)\boldsymbol{\tau}_N\boldsymbol{\tau}_N' diag\left(\boldsymbol{G}_0\right)$$

$$\boldsymbol{G}_0 = \boldsymbol{G}_0\left(\boldsymbol{\psi}_0\right) = \boldsymbol{W}\left(I_N - \boldsymbol{\Psi}_0\boldsymbol{W}\right)^{-1},$$

the ith element of $\text{diag}\left(\boldsymbol{G}_0\boldsymbol{G}_0'\right)$ is given by $g_{0i}'g_{0i}$. For large N it is also required that $\lambda_{\max}\left[\boldsymbol{H}_0\right] < K$ for all N.

Proposition 2 (**The main asymptotic result**) Consider the HSAR model (46) and suppose that: (a) Assumptions 1 to 5 hold, (b) the invertibility condition (47) is met, (c) the $N \times N$ information matrix

$$\mathbf{H}_{11,2} = \left(\mathbf{G}_0 \odot \mathbf{G}_0'\right) + diag\left(\mathbf{G}_0\mathbf{G}_0'\right) - \frac{2}{N} diag\left(\mathbf{G}_0\right)\boldsymbol{\tau}_N'\boldsymbol{\tau}_N diag\left(\mathbf{G}_0'\right)$$

is full rank, where $\mathbf{G}_0 = \boldsymbol{W}(\boldsymbol{I}_N - \Psi\boldsymbol{W})^{-1}$, $\boldsymbol{\Psi}_0 = diag(\boldsymbol{\psi}_0)$; $\boldsymbol{\psi}_0 = (\psi_{10}, ..., \psi_{N0})'$, the ith element of $diag\left(\boldsymbol{G}_0\boldsymbol{G}_0'\right)$ is given by $\boldsymbol{g}_{0i}'\boldsymbol{g}_{0i}$, and $\boldsymbol{W}$ is the spatial weight matrix, and (d) $\varepsilon_{it} \sim IIDN(0, \sigma_0^2)$. The MLE of $\boldsymbol{\psi}_0$ has the following asymptotic distribution as $T \to \infty$,

$$\sqrt{T}\left(\hat{\boldsymbol{\psi}}_T - \boldsymbol{\psi}_0\right) \to_d N\left(0, AsyVar\left(\hat{\boldsymbol{\psi}}_T\right)\right)$$

$$AsyVar\left(\hat{\boldsymbol{\psi}}_T\right) = \left[\left(\mathbf{G}_0 \odot \mathbf{G}_0'\right) + diag\left(\mathbf{G}_0\mathbf{G}_0'\right) - \frac{2}{N} diag\left(\mathbf{G}_0\right)\boldsymbol{\tau}_N'\boldsymbol{\tau}_N diag\left(\mathbf{G}_0'\right)\right]^{-1}$$

which does not depend on σ_0^2.

The HSAR model with heteroskedastic error variances and exogenous regressors The HSAR model (46) can be extended to include exogenous regressors as well as heteroskedastic errors:

$$y_{it} = \psi_i \sum_{j=1}^{N} w_{ij} y_{jt} + \boldsymbol{\beta}_i'\boldsymbol{x}_{it} + \varepsilon_{it} \tag{49}$$

where $\boldsymbol{w}_i'\boldsymbol{y}_t = \sum_{j=1}^{N} w_{ij}y_{jt}$, $\boldsymbol{w}_i = (w_{i1},...,w_{iN})'$ with $w_{ii} = 0$, and $\boldsymbol{y}_t = (y_{1t},...,y_{Nt})'$. Now, we also introduce a $k \times 1$ vector of exogenous regressors $\boldsymbol{x}_{it} = (x_{i1,t},...,x_{ik,t})'$ with parameters $\boldsymbol{\beta}_i = (\beta_{i1},...,\beta_{ik})'$. The above specification allows for the fixed effects by setting one of the regressors equal to unity.

We also allow ε_{it} to be cross-sectionally heteroskedastic, $Var(\varepsilon_{it}) = \sigma_i^2$ for $i = 1,...,N$. Stacking by individual units for each t, (49) becomes

$$\boldsymbol{y}_t = \boldsymbol{\Psi}\boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{B}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \tag{50}$$

where $\boldsymbol{\Psi} = diag(\boldsymbol{\psi})$, $\boldsymbol{\psi} = (\psi_1,...,\psi_N)'$, $\boldsymbol{B} = diag\left(\boldsymbol{\beta}_1',...,\boldsymbol{\beta}_N'\right)'$, and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t},...,\varepsilon_{Nt})'$. Then (50) can be written as

$$\mathbf{y}_t = \left(\mathbf{I}_N - \boldsymbol{\Psi}\mathbf{W}\right)^{-1}\left(\mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t\right)$$

The (quasi) LLF can be written as (assuming that the errors are Gaussian)

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= -\frac{NT}{2}\ln 2\pi - \frac{T}{2}\sum_{i=1}^{N}\ln\sigma_i^2 + T\ln\left|\mathbf{I}_N - \boldsymbol{\Psi}\mathbf{W}\right| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\left[\left(\mathbf{I}_N - \boldsymbol{\Psi}\mathbf{W}\right)\mathbf{y}_t - \mathbf{B}\mathbf{x}_t\right]'\boldsymbol{\Sigma}_\varepsilon\left[\left(\mathbf{I}_N - \boldsymbol{\Psi}\mathbf{W}\right)\mathbf{y}_t - \mathbf{B}\mathbf{x}_t\right]\end{aligned}$$

where $\boldsymbol{\Sigma}_\varepsilon = diag(\sigma_1^2,...,\sigma_N^2)$. Alternatively, it is often more convenient to write the above log-likelihood function as

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= -\frac{NT}{2}\ln 2\pi - \frac{T}{2}\sum_{i=1}^{N}\ln\sigma_i^2 + T\ln\left|\mathbf{I}_N - \boldsymbol{\Psi}\mathbf{W}\right| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\frac{\left(\boldsymbol{y}_i - \psi_i\boldsymbol{y}_i^* - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)'\left(\boldsymbol{y}_i - \psi_i\boldsymbol{y}_i^* - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)}{\sigma_i^2}\end{aligned}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\psi}, \boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_N', \sigma_1^2, ..., \sigma_N^2\right)'$ with $\boldsymbol{\beta}_i = (\beta_{i1},...,\beta_{iN})'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1},...,\boldsymbol{x}_{iT})'$ is the $T \times k$ matrix of regressors on the ith cross section unit with $\boldsymbol{x}_{it} = (x_{i1,t},...,x_{ik,t})'$, $\boldsymbol{y}_i = (y_{i1},...,y_{iT})'$, and $\boldsymbol{y}_i^* = (y_{i1}^*,...,y_{iT}^*)'$, with the elements $y_{it}^* = \boldsymbol{w}_i'\boldsymbol{y}_t = \sum_{j=1}^{N} w_{ij}y_{jt}$.

Assumption 6 The $N(k+2)\times 1$ parameter vector, $\boldsymbol{\theta} = \left(\boldsymbol{\psi}, \boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_N', \sigma_1^2, ..., \sigma_N^2\right)' \in \Theta$ is a sub-set of the $N(k+2)$ dimensional Euclidean space, $\mathbb{R}^{N(k+2)}$. Θ is a closed and bounded (compact) set and includes $\boldsymbol{\theta}_0$ as an interior point, and $\sup_i \|\boldsymbol{\beta}_i\|_1 < K$.

Assumption 7 ε_{it} are independently distributed over i and t; have zero means, and constant variances, $E(\varepsilon_{it}^2) = \sigma_i^2$.

Assumption 8 $\boldsymbol{x}_{it}$ are exogenous such that $E\left(\boldsymbol{x}_{it}\varepsilon_{jt}\right) = 0$ for all i and j,

$$T^{-1}\sum_{t=1}^{T}\boldsymbol{x}_{it}\varepsilon_{jt} \to_p 0, \text{ uniformly in } i \text{ and } j = 1,...,N.$$

The covariance matrices $E\left(\boldsymbol{x}_{it}\boldsymbol{x}_{jt}'\right)=\boldsymbol{\Sigma}_{ij}$, for all i and j, are time-invariant and finite, $\boldsymbol{\Sigma}_{ii}$ is non-singular, $T^{-1}\sum_{t=1}^{T}\boldsymbol{X}_i\boldsymbol{X}_j'\rightarrow_p\boldsymbol{\Sigma}_{ij}$;

$$\sup_i\lambda_{\max}\left(T^{-1}\boldsymbol{X}_i\boldsymbol{X}_i'\right)<K,\ \inf_i\lambda_{\min}\left(T^{-1}\boldsymbol{X}_i\boldsymbol{X}_i'\right)>0$$

Proposition 3 (**The main asymptotic result**) Consider the HSAR model (49) and suppose that: (a) Assumptions 1, 4, 5, and 6, 7, and 8 hold, (b) the invertibility condition (47) is met, (c) $\lambda_{min}\left(\tilde{\mathbf{H}}_{11,2}\right)>\epsilon>0$ for all N, where $\tilde{\mathbf{H}}_{11,2}$ is the $N\times N$ matrix

$$\begin{aligned}\tilde{\mathbf{H}}_{11,2} &= (\mathbf{G}_0\odot\mathbf{G}_0')+diag\left(-g_{0,ii}+\sum_{s=1,s\neq i}^{N}\frac{\sigma_s^2}{\sigma_i^2}g_{0,is}^2,\ i=1,...,N\right)\\ &\quad+diag\left[\frac{1}{\sigma_i^2}\sum_{r=1}^{N}\sum_{s=1}^{N}g_{0,is}g_{0,ir}\beta_r'\left(\Sigma_{rs}-\Sigma_{ri}\Sigma_{ii}^{-1}\Sigma_{it}\right)\beta_s,\ i=1,...,N\right]\end{aligned}$$

$\mathbf{G}_0=\mathbf{W}(\mathbf{I}_N-\Psi\mathbf{W})^{-1}$, $\boldsymbol{\Psi}_0=diag(\boldsymbol{\psi}_0)$; $\boldsymbol{\psi}_0=(\psi_{10},...,\psi_{N0})'$, the ith element of $diag\left(\boldsymbol{G}_0\boldsymbol{G}_0'\right)$ is given by $\boldsymbol{g}_{0i}'\boldsymbol{g}_{0i}$, and $\boldsymbol{W}$ is the spatial weight matrix, and (d) $\varepsilon_{it}\sim IIDN(0,\sigma_i^2)$. Then the MLE of $\boldsymbol{\psi}_0$ has the following asymptotic distribution as $T\rightarrow\infty$,

$$\sqrt{T}\left(\hat{\boldsymbol{\psi}}_T-\boldsymbol{\psi}_0\right)\rightarrow_d N\left(0,AsyVar\left(\hat{\boldsymbol{\psi}}_T\right)\right)$$

where $AsyVar\left(\hat{\boldsymbol{\psi}}_T\right)=\tilde{\mathbf{H}}_{11,2}^{-1}$.

5.5.1 BHP (2016) application to US housing prices

Almost all spatial econometric models assume that the spatial parameters do not vary across the units. Such parameter homogeneity is unavoidable when T is very small. The evidence of parameter heterogeneity in panel data models is quite prevalent particularly in the case of cross-county or country datasets. In such cases and when T is sufficiently large, reducing the spatial effects to a single parameter appears rather restrictive. ABP allow the spatial effects to differ across the units, and derive the conditions needed for identification and consistent estimation under parameter heterogeneity. Consider the following heterogeneous equation:

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

where $x_{it}^*=\mathbf{w}_i'\mathbf{x}_{\circ t}$, $\mathbf{w}_i'$ denotes the ith row of the $N\times N$ row-standardized spatial matrix, W. In the spatial econometrics literature it is assumed that all units have at least one neighbour, which ensures that $\mathbf{w}_i'\tau=1$ for all i. But 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, we set $\psi_i = 0$. In matrix notation we have

$$\mathbf{x}_{\circ t} = \mathbf{\Psi W x}_{\circ t} + \mathbf{u}_{\circ t};\ for\ t = 1, ..., T;$$

where $\mathbf{\Psi} = diag\,(\psi)$, $\psi = (\psi_1, ..., \psi_N)'$, and $\sigma_{ui}^2 = var\,(u_{it})$, $i = 1, ..., N$.

An extension that incorporates richer temporal and spatial dynamics and accommodates negative as well as positive connections is given by

$$\mathbf{x}_{\circ t} = \sum_{j=1}^{h_\lambda} \mathbf{\Lambda}_j \mathbf{x}_{\circ t-j} + \sum_{j=1}^{h_\psi^+} \mathbf{\Psi}_j^+ \mathbf{W}^+ \mathbf{x}_{\circ t-j} + \sum_{j=1}^{h_\psi^-} \mathbf{\Psi}_j^- \mathbf{W}^- \mathbf{x}_{\circ t-j} + \mathbf{u}_{\circ t}$$

where $h_\lambda = \max\,(h_{\lambda 1}, ..., h_{\lambda N})'$; $h_\psi^+ = \left(h_{\psi 1}^+, ..., h_{\psi N}^+\right)'$; $h_\psi^- = \left(h_{\psi 1}^-, ..., h_{\psi N}^-\right)'$; $\mathbf{\Lambda}_j$, $\mathbf{\Psi}_j^+$, $\mathbf{\Psi}_j^-$ 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 $\mathbf{W} = \mathbf{W}^+ + \mathbf{W}^-$. We set $h_\lambda = h_\psi^+ = h_\psi^- = 1$ for expositional simplicity.

ABP propose a QML procedure. The following concentrated log-likelihood function can be used:

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

where

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\psi}_{i0}^+ \mathbf{x}_i^+ - \boldsymbol{\psi}_{i0}^- \mathbf{x}_i^-$$

$$\mathbf{M}_i = \mathbf{I}_T - \mathbf{Z}_i \left(\mathbf{Z}_i' \mathbf{Z}_i\right)^{-1} \mathbf{Z}_i, \mathbf{Z}_i = \left(\mathbf{x}_{i,-1}, \mathbf{x}_{i,-1}^+, \mathbf{x}_{i,-1}^-\right)$$

$$\boldsymbol{\psi}_0^+ = \left(\psi_{10}^+, ..., \psi_{N0}^+\right)', \boldsymbol{\psi}_0^- = \left(\psi_{10}^-, ..., \psi_{N0}^-\right)',$$

The parameters of the lagged variables, λ_1, ψ_1^+ and ψ_1^-, can be estimated by least squares applied to the equations for individual units conditional on ψ_{i0}^+ and ψ_{i0}^-. For inference the analysis must be carried out with respect to the unconcentrated log-likelihood function in terms of $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_N'\right)'$, where $\boldsymbol{\theta}_i = \left(\psi_{i0}^+, \psi_{i0}^-, \psi_{i1}^+, \psi_{i1}^-, \lambda_{i1}, \sigma_{ui}^2\right)'$. The variance–covariance matrix of $\hat{\boldsymbol{\theta}}_{ML}$ is computed as

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

See BHP for more detailed applications.

5.6 The Spatiotemporal Autoregressive Distributed Lag (STARDL) Modelling

Motivations and Plan:

1. Issues: spatial dependence, **spatial heterogeneity,** nonlinearity and anything else;

2. Narrow focus:

   (a) Most general model in the spatial literature is SDPD or the dynamic spatial Durbin model; identification and estimation??

   (b) SP-VAR extension (Xie, 2015), still subject to the homogeneous coefficients.

   (c) ABP approach to HSAR, QML estimation of the heterogeneous coefficients, still not accommodating the dynamic spatial Durbin model;

   (d) Hence, we propose an alternative novel STARDL approach that can allow for consistent estimation of heterogeneous coefficients.

   (e) Then, we derive the generalised spatial model representation from which we develop the dynamic, the spatial and the total multipliers or IRFs...

We consider the STARDL model with the heterogeneous parameters:

$$y_{it} = \phi_i y_{it-1} + \boldsymbol{\pi}'_{i0}\boldsymbol{x}_{it} + \boldsymbol{\pi}'_{i1}\boldsymbol{x}_{i,t-1} + \phi^*_{i0} y^*_{it} + \phi^*_{i1} y^*_{it-1} + \boldsymbol{\pi}^{*\prime}_{i0}\boldsymbol{x}^*_{it} + \boldsymbol{\pi}^{*\prime}_{i1}\boldsymbol{x}^*_{i,t-1} + u_{it} \quad (51)$$

where y_{it} is the scalar dependent variable of the ith spatial unit at time t, $\boldsymbol{x}_{it} = \left(x^1_{it}, ..., x^K_{it}\right)'$ is a $K \times 1$ vector of exogenous regressors with a $K \times 1$ vector of parameters, $\boldsymbol{\pi}_{i0} = \left(\pi^1_{i0}, ..., \pi^K_{i0}\right)'$. Similarly for $\boldsymbol{x}_{i,t-1} = \left(x^1_{i,t-1}, ..., x^K_{i,t-1}\right)'$ and $\boldsymbol{\pi}_{i1} = \left(\pi^1_{i1}, ..., \pi^K_{i1}\right)'$.

The spatial variables, y^*_{it} and $\boldsymbol{x}^*_{it}$, are defined by

$$y^*_{it} = \sum_{j=1}^{N} w_{ij} y_{jt} = \boldsymbol{w}_i \boldsymbol{y}_t \quad \text{with} \quad \underset{N\times 1}{\boldsymbol{y}_t} = (y_{1t}, ..., y_{Nt})', \quad (52)$$

$$\underset{K\times 1}{\boldsymbol{x}^*_{it}} = \begin{pmatrix} x^{1*}_{it} \\ \vdots \\ x^{K*}_{it} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^N w_{ij} x^1_{jt} \\ \vdots \\ \sum_{j=1}^N w_{ij} x^K_{jt} \end{pmatrix} = (\boldsymbol{w}_i \otimes \boldsymbol{\iota}_K)\, \boldsymbol{x}_t; \; \underset{NK\times 1}{\boldsymbol{x}_t} = \begin{pmatrix} \boldsymbol{x}_{1t} \\ \vdots \\ \boldsymbol{x}_{Nt} \end{pmatrix} \quad (53)$$

where $\boldsymbol{w}_i = (w_{i1}, ..., w_{iN})$ is a $1 \times N$ vector of spatial weights determined *a priori* with $w_{ii} = 0$, and $\boldsymbol{\iota}_K$ is a $K \times 1$ vector of unity. Similarly,

$$y^*_{i,t-1} = \sum_{j=1}^{N} w_{ij} y_{j,t-1} = \boldsymbol{w}_i \boldsymbol{y}_{t-1} \text{ with } \underset{N\times 1}{\boldsymbol{y}_{t-1}} = (y_{1,t-1}, ..., y_{N,t-1})',$$

$$\boldsymbol{x}^*_{i,t-1} = \left(x^{1*}_{i,t-1}, ..., x^{K*}_{i,t-1}\right)' = (\boldsymbol{w}_i \otimes \boldsymbol{\iota}_K)\, \boldsymbol{x}_{t-1} \text{ with } \underset{NK\times 1}{\boldsymbol{x}_{t-1}} = \left(\boldsymbol{x}'_{1,t-1}, ..., \boldsymbol{x}'_{N,t-1}\right)'.$$

Then, $\boldsymbol{y}_t^* = (y_{1t}^*, ..., y_{Nt}^*)'$ and $\boldsymbol{x}_t^* = (\boldsymbol{x}_{1t}^{*\prime}, ..., \boldsymbol{x}_{Nt}^{*\prime})'$ can be expressed as

$$\underset{N\times 1}{\boldsymbol{y}_t^*} = \boldsymbol{W}\boldsymbol{y}_t \text{ and } \underset{NK\times 1}{\boldsymbol{x}_t^*} = (\boldsymbol{W}\otimes\boldsymbol{\iota}_K)\,\boldsymbol{x}_t \tag{54}$$

where $\boldsymbol{W}$ is the $N\times N$ matrix of the spatial weights given by

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NN} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix} \text{ with } w_{ii} = 0 \tag{55}$$

Stacking the individual STARDL(1,1) regressions, (51), we have the following generalised spatial representation:

$$\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{y}_{t-1} + \boldsymbol{\Pi}_0\boldsymbol{x}_t + \boldsymbol{\Pi}_1\boldsymbol{x}_{t-1} + \boldsymbol{\Phi}_0^*\boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{\Phi}_1^*\boldsymbol{W}\boldsymbol{y}_{t-1} + \boldsymbol{\Pi}_0^*(\boldsymbol{W}\otimes\boldsymbol{\iota}_K)\,\boldsymbol{x}_t + \boldsymbol{\Pi}_1^*(\boldsymbol{W}\otimes\boldsymbol{\iota}_K)\,\boldsymbol{x}_{t-1} + \boldsymbol{u}_t \tag{56}$$

where

$$\underset{N\times N}{\boldsymbol{\Phi}} = \begin{bmatrix} \phi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_N \end{bmatrix},\ \underset{N\times N}{\boldsymbol{\Phi}_h^*} = \begin{bmatrix} \phi_{1h}^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh}^* \end{bmatrix} \text{ for } h = 0, 1$$

$$\underset{N\times NK}{\boldsymbol{\Pi}_h} = \begin{bmatrix} \boldsymbol{\pi}_{1h}' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\pi}_{Nh}' \end{bmatrix},\ \underset{N\times NK}{\boldsymbol{\Pi}_h^*} = \begin{bmatrix} \boldsymbol{\pi}_{1h}'^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\pi}_{Nh}'^* \end{bmatrix} \text{ for } h = 0, 1$$

It is straightforward to develop the general STARDL(p,q) model with the heterogeneous parameters as follows:

$$y_{it} = \sum_{h=1}^{p}\phi_{ih}y_{i,t-h} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'\boldsymbol{x}_{i,t-h} + \sum_{h=0}^{p}\phi_{ih}^*y_{i,t-h}^* + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'^*\boldsymbol{x}_{i,t-h}^* + u_{it} \tag{57}$$

Suppose that the lag orders p and q are selected sufficiently large and assume that $\boldsymbol{x}_{it}$ are exogenous. In such a case u_{it}'s in (57) are free from serial correlations. Stacking the individual STARDL(p,q) regressions, (57), we have the following generalised spatial representation:

$$\boldsymbol{y}_t = \sum_{h=1}^{p}\boldsymbol{\Phi}_h\boldsymbol{y}_{t-h} + \sum_{h=0}^{q}\boldsymbol{\Pi}_h\boldsymbol{x}_{t-h} + \sum_{h=0}^{p}\boldsymbol{\Phi}_h^*\boldsymbol{W}\boldsymbol{y}_{t-h} + \sum_{h=0}^{q}\boldsymbol{\Pi}_h^*(\boldsymbol{W}\otimes\boldsymbol{\iota}_K)\,\boldsymbol{x}_{t-h} + \boldsymbol{u}_t \tag{58}$$

where

$$\underset{N\times N}{\boldsymbol{\Phi}_h} = \begin{bmatrix} \phi_{1h} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh} \end{bmatrix}, h = 1, ...p,\ \underset{N\times N}{\boldsymbol{\Phi}_h^*} = \begin{bmatrix} \phi_{1h}^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh}^* \end{bmatrix}, h = 0, 1, ..., p$$

$$\underset{N\times NK}{\boldsymbol{\Pi}_h} = \begin{bmatrix} \boldsymbol{\pi}'_{1h} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\pi}'_{Nh} \end{bmatrix}, \underset{N\times NK}{\boldsymbol{\Pi}^*_h} = \begin{bmatrix} \boldsymbol{\pi}'^*_{1h} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\pi}'^*_{Nh} \end{bmatrix}, h = 0, 1, ..., q.$$

Remark: **Spatial stability**: The eigenvalues of $\boldsymbol{\Phi}^*_0 \boldsymbol{W}$ lie inside the unit circle.

Remark: **Time stability**: We rewrite equation (58) as

$$\boldsymbol{y}_t = \sum_{h=1}^{p} \tilde{\boldsymbol{\Phi}}_h \boldsymbol{y}_{t-h} + \sum_{h=0}^{q} \tilde{\boldsymbol{\Pi}}_h \boldsymbol{x}_{t-h} + \tilde{\boldsymbol{u}}_t, \tag{59}$$

where $\tilde{\boldsymbol{\Phi}}_h = \left(\boldsymbol{I}_N - \boldsymbol{\Phi}^*_0 \boldsymbol{W}\right)^{-1} \left(\boldsymbol{\Phi}_h + \boldsymbol{\Phi}^*_h \boldsymbol{W}\right)$, $\tilde{\boldsymbol{\Pi}}_h = \left(\boldsymbol{I}_N - \boldsymbol{\Phi}^*_0 \boldsymbol{W}\right)^{-1} \left[\boldsymbol{\Pi}_h + \boldsymbol{\Pi}^*_h \left(\boldsymbol{W} \otimes \boldsymbol{\iota}_K\right)\right]$, and $\tilde{\boldsymbol{u}}_t = \left(\boldsymbol{I}_N - \boldsymbol{\Phi}^*_0 \boldsymbol{W}\right)^{-1} \boldsymbol{u}_t$. The roots of the $N \times N$ matrix polynomial $\tilde{\boldsymbol{\Phi}}(z) = \boldsymbol{I}_N - \sum_{h=1}^{p} \tilde{\boldsymbol{\Phi}}_h z$ lie outside the unit circle.

5.6.1 Estimation and Inference

To deal with the endogeneity of y^*_{it} in (57) we apply the control function approach.[11] Consider the following CF DGP for y^*_{it}:

$$y^*_{it} = \boldsymbol{\varphi}'_i \boldsymbol{z}_{it} + v_{it} \quad \text{with } E\left(\boldsymbol{z}'_{it} v_{it}\right) = \boldsymbol{0} \tag{60}$$

where $\boldsymbol{z}_{it}$ be the $L \times 1$ vector of exogenous variables:

$$\boldsymbol{z}_{it} = \left(\boldsymbol{z}^{1\prime}_{it}, \boldsymbol{z}^{2\prime}_{it}\right)$$

where $\boldsymbol{z}^1_{it}$ be the $L_1 \times 1$ vector of exogenous variables included in (57) and $\boldsymbol{z}^2_{it}$ be the $L_2 \times 1$ vector of exogenous variables excluded.

We assume that the error terms u_{it} and v_{it} have a joint distribution:

$$\begin{pmatrix} u_{it} \\ v_{it} \end{pmatrix} \sim iid\left(0, \Sigma_{uv}\right), \quad \Sigma_{uv} = \begin{bmatrix} \sigma^2_u & \sigma_{uv} \\ \sigma_{uv} & \sigma^2_v \end{bmatrix} \tag{61}$$

Furthermore, $E\left(u_{it}|v_{it}\right) = \rho v_{it}$ and $E\left(u_{it}|v_{it}\right) = \sigma^2_e$. The endogeneity of y^*_{it} comes from the correlation between u_{it} and v_{it}. Using (169), we can construct[12]

$$u_{it} = \rho v_{it} + e_{it} \tag{62}$$

[11] Most linear models are estimated using IV methods –two stage least squares (2SLS). An alternative, the control function (CF) approach, relies on the same kind of identification conditions. However, in models with nonlinearities or random coefficients, the form of exogeneity is stronger and more restrictions are imposed on the reduced forms.

[12] In the special case where $(u_{it}, v_{it})'$ has a jointly normal distribution, then

$$u_{it}|v_{it} \sim N\left(\frac{\sigma_{uv}}{\sigma^2_v} v_{it}, \frac{\sigma^2_{uv}}{\sigma^2_v}\right)$$

and e_{it} is independent of v_{it}.

where $\rho = E(v_{it}u_{it})/E(v_{it}^2)$. By construction, $E(\boldsymbol{z}_{it}'e_{it}) = \mathbf{0}$ and $E(v_{it}e_{it}) = 0$. Replacing u_{it} by (62), we obtain the following transformation of (57):

$$y_{it} = \sum_{h=1}^{p}\phi_{ih}y_{i,t-h} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'\boldsymbol{x}_{i,t-h} + \sum_{h=0}^{p}\phi_{ih}^{*}y_{i,t-h}^{*} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'^{*}\boldsymbol{x}_{i,t-h}^{*} + \rho v_{it} + e_{it} \quad (63)$$

where v_{it} is the additional control variable, rendering the new error terms, e_{it} uncorrelated with y_{it}^{*} as well as with v_{it} and other regressors in (63). In practice, we use the two-step procedure: (i) obtain the reduced form residuals, $\hat{v}_{it} = y_{it}^{*} - \hat{\boldsymbol{\varphi}}_i'\boldsymbol{z}_{it}$ from (60) and (ii) then run the following regression:

$$y_{it} = \sum_{h=1}^{p}\phi_{ih}y_{i,t-h} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'\boldsymbol{x}_{i,t-h} + \sum_{h=0}^{p}\phi_{ih}^{*}y_{i,t-h}^{*} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'^{*}\boldsymbol{x}_{i,t-h}^{*} + \rho\hat{v}_{it} + e_{it}^{*} \quad (64)$$

where $e_{it}^{*} = e_{it}^{*} + \rho(\hat{\boldsymbol{\varphi}}_i - \boldsymbol{\varphi}_i)'\boldsymbol{z}_{it}$ depends on the sampling error in $\hat{\boldsymbol{\varphi}}_i$ unless $\rho = 0$ (exogeneity test). Then, the OLS estimator from (64) will be consistent for all the parameters.

Remark: The selection of exogenous IVs, $\boldsymbol{z}_{it}$. We prefer to construct them internally. Potentially, under the assumption of exogeneity of $\boldsymbol{x}_{it}$, $\boldsymbol{x}_{it}$ and $\boldsymbol{x}_{it}^{*}$ would be good candidates. Then, why not $\boldsymbol{x}_{jt}$ and $\boldsymbol{x}_{jt}^{*}$, $j(\neq i) = 1, ..., N$? For the current application on the relationship between inflation and output gap, we may suggest to use the cross-section average of y_{it-1}, $\bar{y}_{t-1}$ as the just-identified IV for y_{it}^{*}. y_{it-1} is unocrrelated with u_{it} whereas $\bar{y}_{t-1}$ (measuring the sort of global inflation) should be highly correlated with y_{it}^{*} (measuring the individual country-specific global inflation).

5.6.2 The Spatiotemporal Dynamic Multipliers

It is straightforward to derive (dynamic) multipliers associated with unit changes in x_t, y_t^{*} and x_t^{*} on y_t. Rewrite the STARDL(p, q) model, (57) as[13]

$$\phi_i(L)\,y_{it} = \phi_i^{*}(L)\,y_{it}^{*} + \boldsymbol{\pi}_i(L)\,\boldsymbol{x}_{it} + \boldsymbol{\pi}_i^{*}(L)\,\boldsymbol{x}_{it}^{*} + u_{it} \quad (65)$$

where

$$\phi_i(L) = 1 - \sum_{h=1}^{p}\phi_{ih}L^h;\ \phi_i^{*}(L) = 1 - \sum_{h=0}^{p}\phi_{ih}^{*}L^h;\ \boldsymbol{\pi}_i(L) = \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'L^h;\ \boldsymbol{\pi}_i^{*}(L) = \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'^{*}L^h.$$

Premultiplying (65) by the inverse of $\phi_i(L)$, we obtain:

$$y_{it} = \tilde{\phi}_i^{*}(L)\,y_{it}^{*} + \tilde{\boldsymbol{\pi}}_i(L)\,\boldsymbol{x}_{it} + \tilde{\boldsymbol{\pi}}_i^{*}(L)\,\boldsymbol{x}_{it}^{*} + \tilde{u}_{it} \quad (66)$$

where $\tilde{\phi}_i^{*}(L)\left(=\sum_{j=0}^{\infty}\tilde{\phi}_{ij}^{*}L^j\right) = [\phi_i(L)]^{-1}\phi_i^{*}(L)$, $\tilde{\boldsymbol{\pi}}_i(L)\left(=\sum_{j=0}^{\infty}\tilde{\boldsymbol{\pi}}_{ij}'L^j\right) = [\phi_i(L)]^{-1}\boldsymbol{\pi}_i(L)$, $\tilde{\boldsymbol{\pi}}_i^{*}(L)\left(=\sum_{j=0}^{\infty}\tilde{\boldsymbol{\pi}}_{ij}^{*\prime}L^j\right) = [\phi_i(L)]^{-1}\boldsymbol{\pi}_i^{*}(L)$ and $\tilde{u}_{it} = [\phi_i(L)]^{-1}u_{it}$.

[13] To construct the dynamic multipliers, we should use the structural parameters in (57) which are consistently estimated by the CF-augmented regression, (64).

The dynamic multipliers, $\tilde{\phi}^*_{ij}$, $\tilde{\boldsymbol{\pi}}'_{ij}$ and $\tilde{\boldsymbol{\pi}}^{*\prime}_{ij}$ for $j = 0, 1, ...$, can be evaluated using the following recursive relationships:

$$\tilde{\phi}^*_{ij} = \phi_{i1}\tilde{\phi}^*_{i,j-1} + \phi_{i2}\tilde{\phi}^*_{i,j-2} + \cdots + \phi_{i,j-1}\tilde{\phi}^*_{i1} + \phi_{ij}\tilde{\phi}^*_{i0} + \phi^*_{ij},\ j = 1, 2, ... \quad (67)$$

where $\phi_{ij} = 0$ for $j < 1$ and $\tilde{\phi}^*_{i0} = \phi^*_{i0},\ \tilde{\phi}^*_{ij} = 0$ for $j < 0$ by construction,

$$\tilde{\boldsymbol{\pi}}'_{ij} = \phi_{i1}\tilde{\boldsymbol{\pi}}'_{i,j-1} + \phi_{i2}\tilde{\boldsymbol{\pi}}'_{i,j-2} + \cdots + \phi_{i,j-1}\tilde{\boldsymbol{\pi}}'_{i,1} + \phi_{ij}\tilde{\boldsymbol{\pi}}'_{i0} + \boldsymbol{\pi}'_{ij},\ j = 1, 2, ... \quad (68)$$

where $\tilde{\boldsymbol{\pi}}'_{i0} = \boldsymbol{\pi}'_{i0},\ \tilde{\boldsymbol{\pi}}'_{ij} = 0$ for $j < 0$, and

$$\tilde{\boldsymbol{\pi}}^{*\prime}_{ij} = \phi_{i1}\tilde{\boldsymbol{\pi}}^{*\prime}_{i,j-1} + \phi_{i2}\tilde{\boldsymbol{\pi}}^{*\prime}_{i,j-2} + \cdots + \phi_{i,j-1}\tilde{\boldsymbol{\pi}}^{*\prime}_{i,1} + \phi_{ij}\tilde{\boldsymbol{\pi}}^{*\prime}_{i0} + \boldsymbol{\pi}'_{ij},\ j = 1, 2, ... \quad (69)$$

where $\tilde{\boldsymbol{\pi}}^{*\prime}_{i0} = \boldsymbol{\pi}^{*\prime}_{i0},\ \tilde{\boldsymbol{\pi}}^{*\prime}_{ij} = 0$ for $j < 0$.

Define the dynamic multiplier effects as

$$\frac{\partial y_{i,t+h}}{\partial y^*_{it}},\ \frac{\partial y_{i,t+h}}{\partial \boldsymbol{x}'_{it}} = \left[\frac{\partial y_{i,t+h}}{\partial x^1_{it}}, ..., \frac{\partial y_{i,t+h}}{\partial x^K_{it}}\right]_{1\times K},\ \frac{\partial y_{i,t+h}}{\partial \boldsymbol{x}^{*\prime}_{it}} = \left[\frac{\partial y_{i,t+h}}{\partial x^{*1}_{it}}, ..., \frac{\partial y_{i,t+h}}{\partial x^{*K}_{it}}\right]_{K\times 1}$$

The cumulative dynamic multiplier effects of y^*_{it}, $\boldsymbol{x}_{it}$ and $\boldsymbol{x}_{it}$ on $y_{i,t+h}$ for $h = 0, ..., H$, can be evaluated as follows:

$$m_{y_i}(y^*_i, H) = \sum_{h=0}^{H} \tilde{\phi}^*_{ih},\ \boldsymbol{m}_{y_i}(\boldsymbol{x}_i, H) = \sum_{h=0}^{H} \tilde{\boldsymbol{\pi}}'_{ih},\ \boldsymbol{m}_{y_i}(\boldsymbol{x}^*_i, H) = \sum_{h=0}^{H} \tilde{\boldsymbol{\pi}}^{*\prime}_{ih}, H = 0, 1, ...$$

By construction, as $H \to \infty$,

$$m_{y_i}(y^*_i, H) \to \beta_{yi};\ \boldsymbol{m}_{y_i}(\boldsymbol{x}_i, H) \to \boldsymbol{\beta}'_{xi};\ \boldsymbol{m}_{y_i}(\boldsymbol{x}^*_i, H) \to \boldsymbol{\beta}^{*\prime}_{xi}$$

where β_{yi}, $\boldsymbol{\beta}_{xi}$ and $\boldsymbol{\beta}^*_{xi}$ are the associated long-run coefficients.

Remark: An important feature of the SARDL model is to capture three different forms of dynamic adjustment from initial equilibrium to the new equilibrium following an economic perturbation with respect to domestic conditions, overseas conditions and the overseas policy decisions.

- An investigation of the key parameters in (57) enables us to categorise the group of countries, say countries that focus on domestic conditions only (e.g. the US), and those that pay attention to both domestic and overseas conditions. The small open economy may be likely to depend more on overseas conditions and *vice versa*. Also differently in the short- and the long-run.

- We may apply the mean group estimation of the key parameters and the dynamic multipliers to see the overall mean patterns on a global scale.

We now develop spatial-dynamic multipliers more generally in terms of the spatial system representation (58). We rewrite (58) as

$$\boldsymbol{\Phi}(L)\,\boldsymbol{y}_t = \boldsymbol{\Phi}^*(L)\,\boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{\Pi}(L)\,\boldsymbol{x}_t + \boldsymbol{\Pi}^*(L)\,(\boldsymbol{W}\otimes\boldsymbol{\iota}_K)\,\boldsymbol{x}_t + \boldsymbol{u}_t \quad (70)$$

where

$$\boldsymbol{\Phi}(L) = \boldsymbol{I}_N - \sum_{h=1}^{p} \boldsymbol{\Phi}_h L^h; \boldsymbol{\Phi}^*(L) = \sum_{h=0}^{p} \boldsymbol{\Phi}_h^* L^h; \boldsymbol{\Pi}(L) = \sum_{h=0}^{q} \boldsymbol{\Pi}_h L^h; \boldsymbol{\Pi}^*(L) = \sum_{h=0}^{q} \boldsymbol{\Pi}_h^* (\boldsymbol{W} \otimes \boldsymbol{\iota}_K) L^h$$

Premultiplying (70) by the inverse of $\boldsymbol{\Phi}(L)$, we obtain:

$$\boldsymbol{y}_t = \tilde{\boldsymbol{\Phi}}^*(L) \boldsymbol{W} \boldsymbol{y}_t + \tilde{\boldsymbol{\Pi}}(L) \boldsymbol{x}_t + \tilde{\boldsymbol{\Pi}}^*(L) (\boldsymbol{W} \otimes \boldsymbol{\iota}_K) \boldsymbol{x}_t + \tilde{\boldsymbol{u}}_t \tag{71}$$

where $\tilde{\boldsymbol{\Phi}}^*(L) \left(= \sum_{h=0}^{\infty} \tilde{\boldsymbol{\Phi}}_h^* L^h\right) = [\boldsymbol{\Phi}(L)]^{-1} \boldsymbol{\Phi}^*(L)$, $\tilde{\boldsymbol{\Pi}}(L) \left(= \sum_{h=0}^{\infty} \tilde{\boldsymbol{\Pi}}_h L^h\right) = [\boldsymbol{\Phi}(L)]^{-1} \boldsymbol{\Pi}(L)$, $\tilde{\boldsymbol{\Pi}}^*(L) \left(= \sum_{h=0}^{\infty} \tilde{\boldsymbol{\Pi}}_h^* L^h\right) = [\boldsymbol{\Phi}(L)]^{-1} \boldsymbol{\Pi}^*(L)$, and $\tilde{\boldsymbol{u}}_t = [\boldsymbol{\Phi}(L)]^{-1} \boldsymbol{u}_t$. The dynamic multipliers, $\tilde{\boldsymbol{\Phi}}_j^*$, $\tilde{\boldsymbol{\Pi}}_j$ and $\tilde{\boldsymbol{\Pi}}_j^*$ for $j = 0, 1, ...$, can be evaluated using the following recursive relationships:

$$\tilde{\boldsymbol{\Phi}}_j^* = \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Phi}}_{j-1}^* + \boldsymbol{\Phi}_2 \tilde{\boldsymbol{\Phi}}_{j-2}^* + \cdots + \boldsymbol{\Phi}_{j-1} \tilde{\boldsymbol{\Phi}}_1^* + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Phi}}_0^* + \boldsymbol{\Phi}_j^*, \ j = 1, 2, ... \tag{72}$$

where $\boldsymbol{\Phi}_j = 0$ for $j < 1$ and $\tilde{\boldsymbol{\Phi}}_0^* = \boldsymbol{\Phi}_0^*$, $\tilde{\boldsymbol{\Phi}}_j^* = 0$ for $j < 0$,

$$\tilde{\boldsymbol{\Pi}}_j = \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Pi}}_{j-1} + \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Pi}}_{j-2} + \cdots + \boldsymbol{\Phi}_{j-1} \tilde{\boldsymbol{\Pi}}_1 + \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Pi}}_0 + \boldsymbol{\Pi}_j, \ j = 1, 2, ... \tag{73}$$

where $\tilde{\boldsymbol{\Pi}}_0 = \boldsymbol{\Pi}_0$, $\tilde{\boldsymbol{\Pi}}_j = 0$ for $j < 0$, and

$$\tilde{\boldsymbol{\Pi}}_j^* = \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Pi}}_{j-1}^* + \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Pi}}_{j-2}^* + \cdots + \boldsymbol{\Phi}_{j-1} \tilde{\boldsymbol{\Pi}}_1^* + \boldsymbol{\Phi}_1 \tilde{\boldsymbol{\Pi}}_0^* + \boldsymbol{\Pi}_j^*, \ j = 1, 2, ... \tag{74}$$

where $\tilde{\boldsymbol{\Pi}}_0^* = \boldsymbol{\Pi}_0^*$, $\tilde{\boldsymbol{\Pi}}_j^* = 0$ for $j < 0$. The matrices of the cumulative dynamic multiplier effects can be evaluated as follows:

$$\boldsymbol{m}_{y^*}(H) = \sum_{h=0}^{H} \frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{y}_t^{*\prime}} = \sum_{h=0}^{H} \tilde{\boldsymbol{\Phi}}_h^*,$$

$$\boldsymbol{m}_x(H) = \sum_{h=0}^{H} \frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{x}_t'} = \sum_{h=0}^{H} \tilde{\boldsymbol{\Pi}}_h, \boldsymbol{m}_{x^*}(H) = \sum_{h=0}^{H} \frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{x}_t^{*\prime}} = \sum_{h=0}^{H} \tilde{\boldsymbol{\Pi}}_h^*,$$

Remark: $\boldsymbol{m}_{y^*}(H)$, $\boldsymbol{m}_x(H)$ and $\boldsymbol{m}_{x^*}(H)$ capture the dynamic multiplier effects with respect to three different types of regressors, $\boldsymbol{y}_t^*$, $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^*$, respectively. But, they are block-diagonal by construction because $\tilde{\boldsymbol{\Phi}}_h^*$, $\tilde{\boldsymbol{\Pi}}_h$ and $\tilde{\boldsymbol{\Pi}}_h^*$ are block-diagonal.

5.6.3 Diffusion Multipliers

We rewrite (59) as

$$\tilde{\boldsymbol{\Phi}}(L) \boldsymbol{y}_t = \tilde{\boldsymbol{\Pi}}(L) \boldsymbol{x}_t + \tilde{\boldsymbol{u}}_t, \tag{75}$$

where $\tilde{\boldsymbol{u}}_t = (\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W})^{-1} \boldsymbol{u}_t$,

$$\tilde{\boldsymbol{\Phi}}(L) = \boldsymbol{I}_N - \sum_{j=1}^{p} \tilde{\boldsymbol{\Phi}}_j L^j \text{ with } \tilde{\boldsymbol{\Phi}}_j = (\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W})^{-1} \left(\boldsymbol{\Phi}_j + \boldsymbol{\Phi}_j^* \boldsymbol{W}\right)$$

$$\tilde{\boldsymbol{\Pi}}(L)=\sum_{j=0}^{q}\tilde{\boldsymbol{\Pi}}_j L^j \text{ with } \tilde{\boldsymbol{\Pi}}_j=\left(\boldsymbol{I}_N-\boldsymbol{\Phi}_0^*\boldsymbol{W}\right)^{-1}\left[\boldsymbol{\Pi}_h+\boldsymbol{\Pi}_h^*\left(\boldsymbol{W}\otimes\boldsymbol{\iota}_K\right)\right]$$

Premultiplying (75) by the inverse of $\tilde{\boldsymbol{\Phi}}(L)$, we obtain:

$$\boldsymbol{y}_t=\boldsymbol{B}(L)\,\boldsymbol{x}_t+\left[\tilde{\boldsymbol{\Phi}}(L)\right]^{-1}\tilde{\boldsymbol{u}}_t,\ \boldsymbol{B}(L)\left(=\sum_{j=0}^{\infty}\boldsymbol{B}_j L^j\right)=\left[\tilde{\boldsymbol{\Phi}}(L)\right]^{-1}\tilde{\boldsymbol{\Pi}}(L) \quad (76)$$

The diffusion multipliers, $\boldsymbol{B}_j$ for $j=0,1,...$, can be evaluated as follows:

Algebra:

$$\tilde{\boldsymbol{\Phi}}(L)\,\boldsymbol{B}(L)=\tilde{\boldsymbol{\Pi}}(L)$$

$$\left(\boldsymbol{I}_N-\sum_{j=1}^{p}\tilde{\boldsymbol{\Phi}}_j L^j\right)\left(\sum_{j=0}^{\infty}\boldsymbol{B}_j L^j\right)=\sum_{j=0}^{q}\tilde{\boldsymbol{\Pi}}_j L^j$$

$$\begin{aligned}&\left(\boldsymbol{I}_N-\tilde{\boldsymbol{\Phi}}_1 L-\tilde{\boldsymbol{\Phi}}_2 L^2-\tilde{\boldsymbol{\Phi}}_3 L^3-\cdots\right)\left(\boldsymbol{B}_0+\boldsymbol{B}_1 L+\boldsymbol{B}_2 L^2+\boldsymbol{B}_3 L^3+\cdots\right)\\ =\ &\boldsymbol{B}_0+\left(-\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_0+\boldsymbol{B}_1\right)L+\left(-\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_1-\tilde{\boldsymbol{\Phi}}_2\boldsymbol{B}_0+\boldsymbol{B}_2\right)L^2\\ &+\left(-\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_2-\tilde{\boldsymbol{\Phi}}_2\boldsymbol{B}_1-\tilde{\boldsymbol{\Phi}}_3\boldsymbol{B}_0+\boldsymbol{B}_3\right)L^3+\cdots\end{aligned}$$

Therefore,

$$\begin{aligned}\boldsymbol{B}_0&=\tilde{\boldsymbol{\Pi}}_0\\ \boldsymbol{B}_1&=\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_0+\tilde{\boldsymbol{\Pi}}_1\\ \boldsymbol{B}_2&=\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_1+\tilde{\boldsymbol{\Phi}}_2\boldsymbol{B}_0+\tilde{\boldsymbol{\Pi}}_2\\ \boldsymbol{B}_3&=\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_2+\tilde{\boldsymbol{\Phi}}_2\boldsymbol{B}_1+\tilde{\boldsymbol{\Phi}}_3\boldsymbol{B}_0+\tilde{\boldsymbol{\Pi}}_3\\ &\cdots\end{aligned}$$

$$\boldsymbol{B}_j=\tilde{\boldsymbol{\Phi}}_1\boldsymbol{B}_{j-1}+\tilde{\boldsymbol{\Phi}}_2\boldsymbol{B}_{j-2}+\cdots+\tilde{\boldsymbol{\Phi}}_{j-1}\boldsymbol{B}_1+\tilde{\boldsymbol{\Phi}}_j\boldsymbol{B}_0+\tilde{\boldsymbol{\Pi}}_j,\ j=1,2,... \quad (77)$$

where $\boldsymbol{B}_0=\tilde{\boldsymbol{\Pi}}_0$ and $\boldsymbol{B}_j=0$ for $j<0$ by construction.

Define the dynamic multiplier effects as

$$\frac{\partial\boldsymbol{y}_{t+h}}{\partial\boldsymbol{x}_t'}=\begin{bmatrix}\frac{\partial y_{1,t+h}}{\partial x_{1t}^1} & \cdots & \frac{\partial y_{1,t+h}}{\partial x_{1t}^k} & \cdots & \frac{\partial y_{1,t+h}}{\partial x_{Nt}^1} & \cdots & \frac{\partial y_{1,t+h}}{\partial x_{Nt}^k}\\ \frac{\partial y_{2,t+h}}{\partial x_{1t}^1} & \cdots & \frac{\partial y_{2,t+h}}{\partial x_{1t}^k} & \cdots & \frac{\partial y_{2,t+h}}{\partial x_{Nt}^1} & \cdots & \frac{\partial y_{2,t+h}}{\partial x_{Nt}^1}\\ \vdots & & \vdots & & \vdots & & \vdots\\ \frac{\partial y_{N,t+h}}{\partial x_{1t}^1} & \cdots & \frac{\partial y_{N,t+h}}{\partial x_{1t}^k} & \cdots & \frac{\partial y_{N,t+h}}{\partial x_{Nt}^1} & \cdots & \frac{\partial y_{N,t+h}}{\partial x_{Nt}^k}\end{bmatrix}_{N\times NK} \quad (78)$$

Then, the matrices of the cumulative diffusion multiplier effects can be evaluated as follows:

$$\boldsymbol{m}_x(H)=\sum_{h=0}^{H}\frac{\partial\boldsymbol{y}_{t+h}}{\partial\boldsymbol{x}_t'}=\sum_{h=0}^{H}\boldsymbol{B}_h,\ H=0,1,2,...$$

The cumulative diffusion multiplier effects of x_{jt}^{ℓ}on $y_{i,t+h}$ are given by the $(i,(j-1)k+\ell)$th element of the $N \times NK$ matrix, $\boldsymbol{m}_x(H)$.

Consider the special case of the single regressor, x_t with $k=1$. Then, (78) is simplified to

$$\frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{x}_t'} = \begin{bmatrix} \frac{\partial y_{1,t+h}}{\partial x_{1t}} & \cdots & \frac{\partial y_{1,t+h}}{\partial x_{Nt}} \\ & \ddots & \\ \frac{\partial y_{N,t+h}}{\partial x_{1t}} & \cdots & \frac{\partial y_{N,t+h}}{\partial x_{Nt}} \end{bmatrix}_{N \times N}, \; h = 0, 1, ... \tag{79}$$

and similarly for $\boldsymbol{m}_x(H)$.

Remark: In the case of homogeneous spatial panel models, LeSage and Pace [2009] propose using the average of the main diagonal elements of the $N \times N$ matrix as a summary measure of the own-partial derivatives that they label a direct (own-region) effect. The direct effect for region i includes some feedback loop effects that arise as a result of impacts passing through neighboring regions j and back to region i. LeSage and Pace also propose an average of the (cumulative) off diagonal elements over all rows (observations) to produce a summary that corresponds to the cross-partial derivative or indirect (other-region) effect associated with changes in the rth explanatory variable. Debarsy et al. (2012) extend this cross-sectional reasoning to the case of dynamic space–time panel data. This allows us to compute own- and cross-partial derivatives that trace the effects (own-region and other-region) through time and space. Space–time dynamic models produce a situation where a change in the ith observation of the rth explanatory variable at time t will produce contemporaneous and future responses in all regions' dependent variables y_{it+T} as well as other-region future responses y_{jt+T}. This is due to the presence of an individual time lag (time dependence), a spatial lag (spatial dependence) and a cross-product term reflecting the space–time diffusion. The main diagonal elements of the $N \times N$ matrix sums for time horizon T represent (cumulative) own-region impacts that arise from both time and spatial dependence. The sum of off-diagonal elements reflects both spillovers measuring contemporaneous cross-partial derivatives and diffusion measuring cross-partial derivatives that involve different time periods. We note that it is not possible to separate out the time dependence from spillover and diffusion effects. This implies that except from the contemporaneous effects that represent pure spatial effects, future time horizons contain both time and space diffusion effects, which cannot be distinguished from each other. Indeed this approach is somewhat similar to Diebold-Yilmaz aggregate measures.

Remark: In the case with heterogeneous spatial coefficients, LeSage and Chin (2016) propose use of the N diagonal elements in (79) to produce observation-level direct effects estimates for each of the N regions. As estimates of region specific (observation-level) indirect spill-in and spill-out effects (similar to from and to-effects or in-degree or out-degree in network approach), they propose use of the sum of off-diagonal elements in each row and column. In this regard, initially, we may apply our GCM type approach to $\boldsymbol{m}_x(H)$ at different horizons.

Remark: $\boldsymbol{m}_x(H)$ captures the total diffusion multiplier effects with respect to $\boldsymbol{x}_t$. It would be an important issue how to decompose the overall diffusion

multipliers into the spatial and dynamic components. The model can be estimated using the QML-type algorithms employed in the spatial literature.

- Add the GCM measures and network/graph approach to the analysis of the dynamic or diffusion multipliers.

5.6.4 STARDL models with observed common factors

We can also add the global factors in a straightforward manner. We now consider the STARDL(p,q) model with the $G\times 1$ vector of observed global factors, $\boldsymbol{g}_t = \left(g_t^1,...,g_t^G\right)'$ (e.g. oil prices, commodity prices or the common currency such as the Euro):[14]

$$y_{it} = \sum_{h=1}^{p}\phi_{ih}y_{i,t-h} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'\boldsymbol{x}_{i,t-h} + \sum_{h=0}^{p}\phi_{ih}^{*}y_{i,t-h}^{*} + \sum_{h=0}^{q}\boldsymbol{\pi}_{ih}'^{*}\boldsymbol{x}_{i,t-h}^{*} + \sum_{h=0}^{q}\boldsymbol{\psi}_{ih}'\boldsymbol{g}_{t-h} + u_{it} \tag{80}$$

Stacking the individual STARDL-F(p,q) regressions, (80), we have:

$$\boldsymbol{y}_t = \sum_{h=1}^{p}\boldsymbol{\Phi}_h\boldsymbol{y}_{t-h} + \sum_{h=0}^{q}\boldsymbol{\Pi}_h\boldsymbol{x}_{t-h} + \sum_{h=0}^{p}\boldsymbol{\Phi}_h^{*}\boldsymbol{W}\boldsymbol{y}_{t-h} + \sum_{h=0}^{q}\boldsymbol{\Pi}_h^{*}\left(\boldsymbol{W}\otimes\boldsymbol{\iota}_K\right)\boldsymbol{x}_{t-h} + \sum_{h=0}^{q}\boldsymbol{\Psi}_h\left(\boldsymbol{i}_N\otimes\boldsymbol{g}_{t-h}\right) + \boldsymbol{u}_t \tag{81}$$

where $\boldsymbol{i}_N$ is an $N\times 1$ vector of unity,

$$\underset{N\times N}{\boldsymbol{\Phi}_h} = \begin{bmatrix} \phi_{1h} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh} \end{bmatrix} \text{ for } h=1,...p, \quad \underset{N\times N}{\boldsymbol{\Phi}_h^{*}} = \begin{bmatrix} \phi_{1h}^{*} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh}^{*} \end{bmatrix} \text{ for } h=0,1,...,p$$

$$\underset{N\times NK}{\boldsymbol{\Pi}_h} = \begin{bmatrix} \boldsymbol{\pi}_{1h}' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\pi}_{Nh}' \end{bmatrix}, \quad \underset{N\times NK}{\boldsymbol{\Pi}_h^{*}} = \begin{bmatrix} \boldsymbol{\pi}_{1h}'^{*} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\pi}_{Nh}'^{*} \end{bmatrix},$$

$$\underset{N\times NG}{\boldsymbol{\Psi}_h} = \begin{bmatrix} \boldsymbol{\psi}_{1h}' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\psi}_{Nh}' \end{bmatrix} \text{ for } h=0,1,...,q.$$

5.6.5 Special Case: STAR models with factors

We consider the STAR model with the heterogeneous parameters given by

$$y_{it} = \phi_i y_{it-1} + \phi_{i0}^{*}y_{it}^{*} + \phi_{i1}^{*}y_{it-1}^{*} + u_{it} \tag{82}$$

where y_{it} is the scalar dependent variable of the ith spatial unit at time t.

[14] For notational simplicity we use the same lag order q for the global factors.

The spatial variable y_{it}^* is defined by

$$y_{it}^* = \sum_{j=1}^{N} w_{ij} y_{jt} = \boldsymbol{w}_i \boldsymbol{y}_t \quad \text{with} \quad \underset{N\times 1}{\boldsymbol{y}_t} = (y_{1t}, ..., y_{Nt})' \tag{83}$$

where $\boldsymbol{w}_i = (w_{i1}, ..., w_{iN})$ denotes a $1 \times N$ vector of spatial weights determined *a priori* with $w_{ii} = 0$. Similarly,

$$y_{i,t-1}^* = \sum_{j=1}^{N} w_{ij} y_{j,t-1} = \boldsymbol{w}_i \boldsymbol{y}_{t-1} \text{ with } \underset{N\times 1}{\boldsymbol{y}_{t-1}} = (y_{1,t-1}, ..., y_{N,t-1})'$$

Then, $\boldsymbol{y}_t^* = (y_{1t}^*, ..., y_{Nt}^*)'$ can be expressed as

$$\underset{N\times 1}{\boldsymbol{y}_t^*} = \begin{bmatrix} \boldsymbol{w}_1 \boldsymbol{y}_t \\ \vdots \\ \boldsymbol{w}_N \boldsymbol{y}_t \end{bmatrix} = \boldsymbol{W} \boldsymbol{y}_t \tag{84}$$

where $\boldsymbol{W}$ is the $N \times N$ matrix of the spatial weights given by

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NN} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix} \text{ with } w_{ii} = 0 \tag{85}$$

Stacking the individual STAR(1) regressions, (82), we have the following generalised spatial representation:

$$\boldsymbol{y}_t = \boldsymbol{\Phi} \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_0^* \boldsymbol{W} \boldsymbol{y}_t + \boldsymbol{\Phi}_1^* \boldsymbol{W} \boldsymbol{y}_{t-1} + \boldsymbol{u}_t \tag{86}$$

where

$$\underset{N\times N}{\boldsymbol{\Phi}} = \begin{bmatrix} \phi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_N \end{bmatrix}, \quad \underset{N\times N}{\boldsymbol{\Phi}_h^*} = \begin{bmatrix} \phi_{1h}^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh}^* \end{bmatrix} \text{ for } h = 0, 1$$

It is straightforward to develop the general STAR(p) model with the heterogeneous parameters as follows:

$$y_{it} = \sum_{h=1}^{p} \phi_{ih} y_{i,t-h} + \sum_{h=0}^{p} \phi_{ih}^* y_{i,t-h}^* + u_{it} \tag{87}$$

Stacking the individual STAR(p) regressions, (87), we have the following generalised spatial representation:

$$\boldsymbol{y}_t = \sum_{h=1}^{p} \boldsymbol{\Phi}_h \boldsymbol{y}_{t-h} + \sum_{h=0}^{p} \boldsymbol{\Phi}_h^* \boldsymbol{W} \boldsymbol{y}_{t-h} + \boldsymbol{u}_t \tag{88}$$

where

$$\underset{N\times N}{\boldsymbol{\Phi}_h} = \begin{bmatrix} \phi_{1h} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh} \end{bmatrix}, h = 1, ...p, \ \underset{N\times N}{\boldsymbol{\Phi}_h^*} = \begin{bmatrix} \phi_{1h}^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{Nh}^* \end{bmatrix}, h = 0, 1, ..., p$$

Remark: **Spatial stability**: The eigenvalues of $\boldsymbol{\Phi}_0^* \boldsymbol{W}$ lie inside the unit circle.

Remark: **Time stability**: We rewrite equation (88) as

$$\boldsymbol{y}_t = \sum_{h=1}^{p} \tilde{\boldsymbol{\Phi}}_h \boldsymbol{y}_{t-h} + \tilde{\boldsymbol{u}}_t, \tag{89}$$

where $\tilde{\boldsymbol{\Phi}}_h = \left(\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W}\right)^{-1} \left(\boldsymbol{\Phi}_h + \boldsymbol{\Phi}_h^* \boldsymbol{W}\right)$, and $\tilde{\boldsymbol{u}}_t = \left(\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W}\right)^{-1} \boldsymbol{u}_t$. The roots of the $N \times N$ matrix polynomial $\tilde{\boldsymbol{\Phi}}(z) = \boldsymbol{I}_N - \sum_{h=1}^{p} \tilde{\boldsymbol{\Phi}}_h z$ lie outside the unit circle.

Suppose that the lag order p is selected sufficiently large in which case u_{it}'s in (87) are free from serial correlations. To deal with the endogeneity of y_{it}^* in (87) we apply the control function approach. Consider the following CF DGP for y_{it}^*:

$$y_{it}^* = \boldsymbol{\varphi}_i' \boldsymbol{z}_{it} + v_{it} \ \text{ with } E\left(\boldsymbol{z}_{it}' v_{it}\right) = \boldsymbol{0} \tag{90}$$

where $\boldsymbol{z}_{it}$ be the $L \times 1$ vector of exogenous variables:

$$\boldsymbol{z}_{it} = \left(\boldsymbol{z}_{it}^{1\prime}, \boldsymbol{z}_{it}^{2\prime}\right)$$

where $\boldsymbol{z}_{it}^1$ be the $L_1 \times 1$ vector of exogenous variables included in (57) and $\boldsymbol{z}_{it}^2$ be the $L_2 \times 1$ vector of exogenous variables excluded.

We assume that the error terms u_{it} and v_{it} have a joint distribution:

$$\begin{pmatrix} u_{it} \\ v_{it} \end{pmatrix} \sim iid\left(0, \Sigma_{uv}\right), \ \ \Sigma_{uv} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \tag{91}$$

Furthermore, $E\left(u_{it}|v_{it}\right) = \rho v_{it}$ and $E\left(u_{it}|v_{it}\right) = \sigma_e^2$. The endogeneity of y_{it}^* comes from the correlation between u_{it} and v_{it}. Using (169), we can construct[15]

$$u_{it} = \rho v_{it} + e_{it} \tag{92}$$

where $\rho = E\left(v_{it} u_{it}\right) / E\left(v_{it}^2\right)$. By construction, $E\left(\boldsymbol{z}_{it}' e_{it}\right) = \boldsymbol{0}$ and $E\left(v_{it} e_{it}\right) = 0$. Replacing u_{it}, we obtain the following transformation of (87):

$$y_{it} = \sum_{h=1}^{p} \phi_{ih} y_{i,t-h} + \sum_{h=0}^{p} \phi_{ih}^* y_{i,t-h}^* + \rho v_{it} + e_{it} \tag{93}$$

[15] In the special case where $(u_{it}, v_{it})'$ has a jointly normal distribution, then

$$u_{it}|v_{it} \sim N\left(\frac{\sigma_{uv}}{\sigma_v^2} v_{it}, \frac{\sigma_{uv}^2}{\sigma_v^2}\right)$$

and e_{it} is independent of v_{it}.

where v_{it} is the control variable, rendering e_{it} uncorrelated with y_{it}^* as well as with v_{it} in (93). In practice, we use the two-step procedure: (i) obtain the reduced form residuals, $\hat{v}_{it} = y_{it}^* - \hat{\boldsymbol{\varphi}}_i' \boldsymbol{z}_{it}$ and (ii) then run the following regression:

$$y_{it} = \sum_{h=1}^{p} \phi_{ih} y_{i,t-h} + \sum_{h=0}^{p} \phi_{ih}^* y_{i,t-h}^* + \rho \hat{v}_{it} + e_{it}^* \tag{94}$$

where $e_{it}^* = e_{it}^* + \rho \left(\hat{\boldsymbol{\varphi}}_i - \boldsymbol{\varphi}_i \right)' \boldsymbol{z}_{it}$ depends on the sampling error in $\hat{\boldsymbol{\varphi}}_i$ unless $\rho = 0$ (exogeneity test). Then, the OLS estimator from (94) will be consistent for all the parameters.

It is straightforward to derive (dynamic) multipliers associated with unit changes in y_t^* on y_t. Rewrite the STAR(p) model, (87) as

$$\phi_i(L) y_{it} = \phi_i^*(L) y_{it}^* + u_{it} \tag{95}$$

where

$$\phi_i(L) = 1 - \sum_{h=1}^{p} \phi_{ih} L^h; \ \phi_i^*(L) = 1 - \sum_{h=0}^{p} \phi_{ih}^* L^h.$$

Premultiplying (95) by the inverse of $\phi_i(L)$, we obtain:

$$y_{it} = \tilde{\phi}_i^*(L) y_{it}^* + \tilde{u}_{it} \tag{96}$$

where $\tilde{\phi}_i^*(L) \left(= \sum_{j=0}^{\infty} \tilde{\phi}_{ij}^* L^j \right) = [\phi_i(L)]^{-1} \phi_i^*(L)$, and $\tilde{u}_{it} = [\phi_i(L)]^{-1} u_{it}$. The dynamic multipliers, $\tilde{\phi}_{ij}^*$, $\tilde{\boldsymbol{\pi}}_{ij}'$ and $\tilde{\boldsymbol{\pi}}_{ij}^{*\prime}$ for $j = 0, 1, ...$, can be evaluated using the following recursive relationships:

$$\tilde{\phi}_{ij}^* = \phi_{i1} \tilde{\phi}_{i,j-1}^* + \phi_{i2} \tilde{\phi}_{i,j-2}^* + \cdots + \phi_{i,j-1} \tilde{\phi}_{i1}^* + \phi_{ij} \tilde{\phi}_{i0}^* + \phi_{ij}^*, \ j = 1, 2, ... \tag{97}$$

where $\phi_{ij} = 0$ for $j < 1$ and $\tilde{\phi}_{i0}^* = \phi_{i0}^*$, $\tilde{\phi}_{ij}^* = 0$ for $j < 0$ by construction.

Define the dynamic multiplier effects as $\frac{\partial y_{i,t+h}}{\partial y_{it}^*}$. Then, the cumulative dynamic multiplier effects of y_{it}^* on $y_{i,t+h}$ for $h = 0, ..., H$, can be evaluated as follows:

$$m_{y_i}(y_i^*, H) = \sum_{h=0}^{H} \tilde{\phi}_{ih}^*, \quad H = 0, 1, ...$$

By construction, as $H \to \infty$,

$$m_{y_i}(y_i^*, H) \to \beta_{yi}^*$$

where β_{yi}^* is the long-run coefficient.

Remark: An important feature of the STAR model is to capture dynamic (spillover) adjustment from initial equilibrium to the new equilibrium following an perturbation with respect to overseas conditions

- An investigation of the key parameters in (57) enables us to categorise the group of countries, say countries that focus on domestic conditions only (e.g. the US), and those that pay attention to both domestic and overseas conditions. The small open economy may be likely to depend more on overseas conditions and *vice versa*. Also differently in the short- and the long-run. Still useful, but not quite sufficiently informative about the spatial and the dynamic multipliers.

We now develop spatial-dynamic multipliers in terms of the spatial system representation (**??**), which we rewrite as

$$\boldsymbol{\Phi}\left(L\right)\boldsymbol{y}_t = \boldsymbol{\Phi}^*\left(L\right)\boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{u}_t \tag{98}$$

where

$$\boldsymbol{\Phi}\left(L\right) = \boldsymbol{I}_N - \sum_{h=1}^{p}\boldsymbol{\Phi}_h L^h; \boldsymbol{\Phi}^*\left(L\right) = \sum_{h=0}^{p}\boldsymbol{\Phi}_h^* L^h$$

Premultiplying (98) by the inverse of $\boldsymbol{\Phi}\left(L\right)$, we obtain:

$$\boldsymbol{y}_t = \tilde{\boldsymbol{\Phi}}^*\left(L\right)\boldsymbol{W}\boldsymbol{y}_t + \tilde{\boldsymbol{u}}_t \tag{99}$$

where $\tilde{\boldsymbol{\Phi}}^*\left(L\right)\left(=\sum_{h=0}^{\infty}\tilde{\boldsymbol{\Phi}}_h^* L^h\right) = \left[\boldsymbol{\Phi}\left(L\right)\right]^{-1}\boldsymbol{\Phi}^*\left(L\right)$, and $\tilde{\boldsymbol{u}}_t = \left[\boldsymbol{\Phi}\left(L\right)\right]^{-1}\boldsymbol{u}_t$. The dynamic multipliers, $\tilde{\boldsymbol{\Phi}}_j^*$ for $j = 0, 1, ...$, can be evaluated using the following recursive relationships:

$$\tilde{\boldsymbol{\Phi}}_j^* = \boldsymbol{\Phi}_1\tilde{\boldsymbol{\Phi}}_{j-1}^* + \boldsymbol{\Phi}_2\tilde{\boldsymbol{\Phi}}_{j-2}^* + \cdots + \boldsymbol{\Phi}_{j-1}\tilde{\boldsymbol{\Phi}}_1^* + \boldsymbol{\Phi}_j\tilde{\boldsymbol{\Phi}}_0^* + \boldsymbol{\Phi}_j^*,\ j = 1, 2, ... \tag{100}$$

where $\boldsymbol{\Phi}_j = 0$ for $j < 1$ and $\tilde{\boldsymbol{\Phi}}_0^* = \boldsymbol{\Phi}_0^*,\ \tilde{\boldsymbol{\Phi}}_j^* = 0$ for $j < 0$ by construction.

Define the dynamic multiplier effects as

$$\frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{y}_t^{*\prime}} = \begin{bmatrix} \frac{\partial y_{1,t+h}}{\partial y_{1t}^*} & \frac{\partial y_{1,t+h}}{\partial y_{2t}^*} & \cdots & \frac{\partial y_{1,t+h}}{\partial y_{Nt}^*} \\ \frac{\partial y_{2,t+h}}{\partial y_{1t}^*} & \frac{\partial y_{2,t+h}}{\partial y_{2t}^*} & \cdots & \frac{\partial y_{2,t+h}}{\partial y_{Nt}^*} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{N,t+h}}{\partial y_{1t}^*} & \frac{\partial y_{N,t+h}}{\partial y_{2t}^*} & \cdots & \frac{\partial y_{N,t+h}}{\partial y_{Nt}^*} \end{bmatrix}_{N\times N}$$

{**Q**} What's $\frac{\partial y_{i,t+h}}{\partial y_{jt}}$? Say,

$$\frac{\partial y_{1,t+h}}{\partial y_{2t}} = \frac{\partial y_{1,t+h}}{\partial y_{1t}^*} \times w_{12} + \frac{\partial y_{1,t+h}}{\partial y_{2t}^*} \times w_{22} + \cdots + \frac{\partial y_{1,t+h}}{\partial y_{Nt}^*} \times w_{n2}$$

Hence,

$$\begin{aligned} \frac{\partial y_{i,t+h}}{\partial y_{jt}} &= \frac{\partial y_{i,t+h}}{\partial y_{1t}^*} \times w_{1j} + \frac{\partial y_{i,t+h}}{\partial y_{2t}^*} \times w_{2j} + \cdots + \frac{\partial y_{i,t+h}}{\partial y_{Nt}^*} \times w_{nj} \\ &= \sum_{k=1}^{N}\frac{\partial y_{i,t+h}}{\partial y_{kt}^*} \times w_{kj} \text{ for } i \neq j \end{aligned}$$

Then, what about $\frac{\partial y_{i,t+h}}{\partial y_{it}}$? Simply set to zero? More??

The matrix of the cumulative dynamic multiplier effects can be evaluated as follows:

$$\boldsymbol{m}_{y^*}(H) = \sum_{h=0}^{H} \frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{y}_t^{*\prime}} = \sum_{h=0}^{H} \tilde{\boldsymbol{\Phi}}_h^*$$

The cumulative dynamic multiplier effects of y_{jt}^* on y_{it} is given by the (i,j)th element of the $N \times N$ matrix $\boldsymbol{m}_{y^*}(H)$.

Remark: $\boldsymbol{m}_{y^*}(H)$ captures the dynamic multiplier effects with respect to $\boldsymbol{y}_t^*$. But, they are block-diagonal by construction:

$$\underset{N\times N}{\boldsymbol{m}_{y^*}(H)} = \begin{bmatrix} m_{y_1}(y_1^*, H) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & m_{y_N}(y_N^*, H) \end{bmatrix}$$

because $\tilde{\boldsymbol{\Phi}}_h^*$ is block-diagonal.

Diffusion IRF and FEVD We rewrite (88) as (see also (89)):

$$\tilde{\boldsymbol{\Phi}}(L)\, \boldsymbol{y}_t = \tilde{\boldsymbol{u}}_t, \tag{101}$$

where $\tilde{\boldsymbol{u}}_t = (\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W})^{-1} \boldsymbol{u}_t$,

$$\tilde{\boldsymbol{\Phi}}(L) = \boldsymbol{I}_N - \sum_{j=1}^{p} \tilde{\boldsymbol{\Phi}}_j L^j \text{ with } \tilde{\boldsymbol{\Phi}}_j = (\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W})^{-1} \left(\boldsymbol{\Phi}_j + \boldsymbol{\Phi}_j^* \boldsymbol{W}\right)$$

Premultiplying (101) by the inverse of $\tilde{\boldsymbol{\Phi}}(L)$, we obtain:

$$\boldsymbol{y}_t = \left[\tilde{\boldsymbol{\Phi}}(L)\right]^{-1} \tilde{\boldsymbol{u}}_t = \left[\tilde{\boldsymbol{\Phi}}(L)\right]^{-1} (\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W})^{-1} \boldsymbol{u}_t \tag{102}$$

from which we can construct (diffusion) IRF and FEVD. $\{\mathbf{Q}\}$ with respect to $\tilde{\boldsymbol{u}}_t$ or $\boldsymbol{u}_t$, probably $\boldsymbol{u}_t$ (can we assume $\boldsymbol{u}_t$ as structural?)

More to follow: **Algebra:**

With Factors In the spatial modelling it is implicitly assumed that u_{it} is iid across spatial units or spatially correlated:

$$u_{it} = \lambda \boldsymbol{W} u_t + \varepsilon_{it}.$$

Now we introduce the common factor structure as in QVAR paper such that[16]

$$u_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{v}_t.$$

[16] Shi and Lee (2017) also postulate that the error term in a SAR panel equation can be decomposed into a common factor component and an idiosyncratic component, where common factors can potentially correlate with included regressors. See also HLP17??

Then, (82) can be extended as

$$y_{it} = \phi_i y_{it-1} + \phi_{i0}^* y_{it}^* + \phi_{i1}^* y_{it-1}^* + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + v_{it} \tag{103}$$

where v_{it} contains the idiosyncratic components which are mutually uncorrelated across (i, j). More generally, we have STAR(p) with (observed) factors:

$$y_{it} = \sum_{h=1}^{p} \phi_{ih} y_{i,t-h} + \sum_{h=0}^{p} \phi_{ih}^* y_{i,t-h}^* + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + v_{it} \tag{104}$$

To deal with the endogeneity of y_{it}^* we apply the control function approach described above. Then, (104) can be expressed as

$$\tilde{\boldsymbol{\Phi}}(L)\, \boldsymbol{y}_t = (\boldsymbol{I}_N - \boldsymbol{\Phi}_0^* \boldsymbol{W})^{-1} (\boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{v}_t) \tag{105}$$

from which we can construct IRF and FEVD.

Remark: This is quite parsimonious specification, implying that we can include the large N spatial units, so another way of circumventing the curse of input dimensionality (e.g. 100 or 1000 asset returns). Further, we may argue that v_{it} would carry certain structural interpretations.

5.7 GVAR-SPVAR Model

Consider a global economy consisting of N economies, indexed by $i = 1, .., N$, and denote the country-specific variables by an $m_i \times 1$ vector $\boldsymbol{y}_{it}$, and the country-specific foreign variables by an $m_i^* \times 1$ vector $\boldsymbol{y}_{it}^* = \sum_{j=1}^{N} w_{ij} \boldsymbol{y}_{jt}$ where $w_{ij} \geq 0$ is the set of granular weights with $\sum_{j=1}^{N} w_{ij} = 1$, and $w_{ii} = 0$ for all i (*e.g.* Pesaran, 2006 and GNS). The country-specific VARX$^*(2, 2)$ model can be written as

$$\boldsymbol{y}_{it} = \boldsymbol{h}_{i0} + \boldsymbol{h}_{i1} t + \boldsymbol{\Phi}_{i1} \boldsymbol{y}_{i,t-1} + \boldsymbol{\Phi}_{i2} \boldsymbol{y}_{i,t-2} + \boldsymbol{\Psi}_{i0} \boldsymbol{y}_{it}^* + \boldsymbol{\Psi}_{i1} \boldsymbol{y}_{i,t-1}^* + \boldsymbol{\Psi}_{i2} \boldsymbol{y}_{i,t-2}^* + \boldsymbol{u}_{it} \tag{106}$$

where the dimension of $\boldsymbol{h}_{ij}$ and $\boldsymbol{\delta}_{ij}$, $j = 0, 1, 2$, is $m_i \times 1$ while the dimensions of $\boldsymbol{\Phi}_{ij}$ and $\boldsymbol{\Psi}_{ij}$, $j = 0, 1, 2$, are $m_i \times m_i$ and $m_i \times m_i^*$. GNS assume that $\boldsymbol{u}_{it} \sim iid(0, \boldsymbol{\Sigma}_{ii})$ where $\boldsymbol{\Sigma}_{ii}$ is an $m_i \times m_i$ positive definite matrix.

5.7.1 Spatial VAR (SPVAR) Models

Beenstock and Felsenstein (2007) propose the panel VAR model with both spatial and time lags, which they refer to as spatial vector autoregressions (SpVAR):

$$Y_{nt} = \mu_n + \theta W Y_{nt} + \sum_{j=1}^{q} \beta_j Y_{n,t-j} + \lambda W Y_{n,t-1} + u_{nt}$$

$$u_{nt} = \rho u_{n,t-1} + \delta W u_{nt} + \gamma W u_{n,t-1} + \varepsilon_{nt} \text{ with } \sigma_{ni} = Cov(\varepsilon_n, \varepsilon_i)$$

The SpVAR resembles the SDPD except that Y_{nt} are allowed to be a $K \times 1$ vector.

The structural multivariate counterpart is (SPSVAR):

$$Y_{knt} = \mu_{kn} + \sum_{i=1}^{K} \left(\alpha_{ki} Y_{int} + \beta_{kj} Y_{in,t-1} + \theta_{ki} Y^*_{int} + \lambda_{ki} Y^*_{in,t-1} \right) + \varepsilon_{knt} \quad (107)$$

where μ's are region-specific effects, α's are within-region contemporaneous causal effects between Y's with $\alpha_{kk} = 0$, θ's are spatial lag coefficients, β's are temporal lag coefficients, and λ's are lagged spatial lag coefficients.

Denote by A, B, Θ and Λ the $K \times K$ coefficient matrices for αs, βs, θs and λs, respectively. We then express (107) as follows:

$$Y_t = \mu + A^* Y_t + B^* Y_{t-1} + \Theta^* Y^*_t + \Lambda^* Y^*_{t-1} + \varepsilon_t \quad (108)$$

where Y is an $NK \times 1$ vector of observations stacked by n, μ is an $NK \times 1$ vector of regional-specific effects,

$$A^* = I_N \otimes A,\ B^* = I_N \otimes B,\ \Theta^* = I_N \otimes \Theta,\ \Lambda^* = I_N \otimes \Lambda,$$

are $NK \times NK$ block diagonal matrices. Y_t and Y^*_t are not independent of ε_t. (108) has a corresponding reduced form:

$$Y_t = \Pi_0 + \Pi_1 Y_{t-1} + \Pi_2 Y^*_t + \Pi_3 Y^*_{t-1} + v_t \quad (109)$$

where $\Pi_0 = (I_{NK} - A^*)^{-1} \mu$, $\Pi_1 = (I_{NK} - A^*)^{-1} B^*$, $\Pi_2 = (I_{NK} - A^*)^{-1} \Theta^*$, $\Pi_3 = (I_{NK} - A^*)^{-1} \Lambda^*$, and $v_t = (I_{NK} - A^*)^{-1} \varepsilon_t$. There are $K(K-1)$ unknown A coefficients, K^2 unknown coefficients for each of B, Θ, and Λ, and there are NK unknown variances for Σ_ε, making a total of $K(4K-1) + K = 4K^2$ unknown structural parameters in (108). In (109) there are $3K^2$ data restrictions from Πs and Σ_v provides $\frac{1}{2}K(K+1)$ further restrictions. Therefore, the SpVAR underidentifies the structural parameters and the identification deficit is $\frac{1}{2}K(K-1)$.

Beenstock and Felsenstein (2007) note that "incidental parameter problem" arises and LSDV estimated B^* has $O(1/T)$ downward bias when T is finite, and further issues arise because q is unknown. When $q = 1$, they propose the bias-corrected estimator following Hsiao (2003). They show that impulse responses of SpVAR help to simulate the spatial-temporal dynamic effects of exogenous shocks.

- Add the review on more recent SPVAR Models: e.g. Mutl (2009) and Spatial Panel Vector Autoregression by Xie (2015).

5.7.2 The spatial representation of the GVAR model

For convenience we assume $m_i = k$ and $m = Nk$. Define the $m \times 1$ vector of the global variables:

$$\underset{m \times 1}{\boldsymbol{y}_t} = (\boldsymbol{y}'_{1t}, ..., \boldsymbol{y}'_{Nt})' \text{ with } \underset{k \times 1}{\boldsymbol{y}_{it}} = (y_{1,it}, ..., y_{k,it})'$$

Define the $N \times N$ weight matrix:

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & & w_{1N} \\ w_{21} & w_{22} & & w_{2N} \\ & & \ddots & \\ w_{N1} & w_{N2} & & w_{NN} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_N \end{bmatrix}$$

Then, we have:

$$\underset{k\times 1}{\boldsymbol{y}_{it}^*} = \left(y_{1,it}^*, ..., y_{k,it}^*\right)' = \left(\mathbf{w}_i \otimes \mathbf{I}_k\right) \boldsymbol{y}_t,\ \underset{m\times 1}{\boldsymbol{y}_t^*} = \left(\boldsymbol{y}_{1t}^{*\prime}, ..., \boldsymbol{y}_{Nt}^{*\prime}\right)' = \left(\mathbf{W} \otimes \mathbf{I}_k\right) \boldsymbol{y}_t$$

Thus, (106) can be written as (drop $\boldsymbol{h}_{i0} + \boldsymbol{h}_{i1}t$ without loss of generality):

$$\boldsymbol{y}_{it} = \boldsymbol{\Phi}_{i1}\boldsymbol{y}_{i,t-1} + \boldsymbol{\Phi}_{i2}\boldsymbol{y}_{i,t-2} + \boldsymbol{\Psi}_{i0}\left(\mathbf{w}_i \otimes \mathbf{I}_{m_i}\right)\boldsymbol{y}_{it} + \boldsymbol{\Psi}_{i1}\left(\mathbf{w}_i \otimes \mathbf{I}_{m_i}\right)\boldsymbol{y}_{it-1} + \boldsymbol{\Psi}_{i2}\left(\mathbf{w}_i \otimes \mathbf{I}_{m_i}\right)\boldsymbol{y}_{it-2} + \boldsymbol{u}_{it} \tag{110}$$

Further,

$$\boldsymbol{y}_{1t} = \boldsymbol{\Phi}_{11}\boldsymbol{y}_{1,t-1} + \boldsymbol{\Phi}_{12}\boldsymbol{y}_{1,t-2} + \boldsymbol{\Psi}_{10}\left(\mathbf{w}_1 \otimes \mathbf{I}_k\right)\boldsymbol{y}_{1t} + \boldsymbol{\Psi}_{11}\left(\mathbf{w}_1 \otimes \mathbf{I}_k\right)\boldsymbol{y}_{1t-1} + \boldsymbol{\Psi}_{12}\left(\mathbf{w}_1 \otimes \mathbf{I}_k\right)\boldsymbol{y}_{1t-2} + \boldsymbol{u}_{1t}$$

$$\vdots$$

$$\boldsymbol{y}_{Nt} = \boldsymbol{\Phi}_{N1}\boldsymbol{y}_{N,t-1} + \boldsymbol{\Phi}_{N2}\boldsymbol{y}_{N,t-2} + \boldsymbol{\Psi}_{N0}\left(\mathbf{w}_N \otimes \mathbf{I}_k\right)\boldsymbol{y}_{Nt} + \boldsymbol{\Psi}_{N1}\left(\mathbf{w}_N \otimes \mathbf{I}_k\right)\boldsymbol{y}_{Nt-1} + \boldsymbol{\Psi}_{N2}\left(\mathbf{w}_N \otimes \mathbf{I}_k\right)\boldsymbol{y}_{Nt-2} + \boldsymbol{u}_{Nt}$$

Stacking these results. we have:

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{y}_{1t} \\ \boldsymbol{y}_{2t} \\ \vdots \\ \boldsymbol{y}_{Nt} \end{bmatrix} = & \begin{bmatrix} \boldsymbol{\Phi}_{11} & 0 & & 0 \\ 0 & \boldsymbol{\Phi}_{21} & & 0 \\ & & \ddots & \\ 0 & 0 & & \boldsymbol{\Phi}_{N1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{1t-1} \\ \boldsymbol{y}_{2t-1} \\ \vdots \\ \boldsymbol{y}_{Nt-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi}_{12} & 0 & & 0 \\ 0 & \boldsymbol{\Phi}_{22} & & 0 \\ & & \ddots & \\ 0 & 0 & & \boldsymbol{\Phi}_{N2} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{1t-2} \\ \boldsymbol{y}_{2t-2} \\ \vdots \\ \boldsymbol{y}_{Nt-2} \end{bmatrix} \\
& + \begin{bmatrix} \boldsymbol{\Psi}_{10} & 0 & & 0 \\ 0 & \boldsymbol{\Psi}_{20} & & 0 \\ & & \ddots & \\ 0 & 0 & & \boldsymbol{\Psi}_{N0} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{1t}^* \\ \boldsymbol{y}_{2t}^* \\ \vdots \\ \boldsymbol{y}_{Nt}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Psi}_{11} & 0 & & 0 \\ 0 & \boldsymbol{\Psi}_{21} & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & & \boldsymbol{\Psi}_{N1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{1t-1}^* \\ \boldsymbol{y}_{2t-1}^* \\ \vdots \\ \boldsymbol{y}_{Nt-1}^* \end{bmatrix} \\
& + \begin{bmatrix} \boldsymbol{\Psi}_{12} & 0 & & 0 \\ 0 & \boldsymbol{\Psi}_{22} & & 0 \\ & & \ddots & \\ 0 & 0 & & \boldsymbol{\Psi}_{N2} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{1t-2}^* \\ \boldsymbol{y}_{2t-2}^* \\ \vdots \\ \boldsymbol{y}_{Nt-2}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_{1t} \\ \boldsymbol{u}_{2t} \\ \vdots \\ \boldsymbol{u}_{Nt} \end{bmatrix}
\end{aligned}$$

Hence, we have:

$$\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2\boldsymbol{y}_{t-2} + \boldsymbol{\Psi}_0\boldsymbol{y}_t^* + \boldsymbol{\Psi}_1\boldsymbol{y}_{t-1}^* + \boldsymbol{\Psi}_2\boldsymbol{y}_{t-2}^* + \boldsymbol{u}_t \qquad (111) \\
&= \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2\boldsymbol{y}_{t-2} + \boldsymbol{\Psi}_0\left(\mathbf{W} \otimes \mathbf{I}_k\right)\boldsymbol{y}_t + \boldsymbol{\Psi}_1\left(\mathbf{W} \otimes \mathbf{I}_k\right)\boldsymbol{y}_{t-1} + \boldsymbol{\Psi}_2\left(\mathbf{W} \otimes \mathbf{I}_k\right)\boldsymbol{y}_{t-2} + \boldsymbol{u}_t
\end{aligned}$$

Alternatively, (111) can be written as

$$\left(\boldsymbol{I}_m - \boldsymbol{\Psi}_0\left(\mathbf{W} \otimes \mathbf{I}_k\right)\right)\boldsymbol{y}_t = \left(\boldsymbol{\Phi}_1 + \boldsymbol{\Psi}_1\left(\mathbf{W} \otimes \mathbf{I}_k\right)\right)\boldsymbol{y}_{t-1} + \left(\boldsymbol{\Phi}_2 + \boldsymbol{\Psi}_2\left(\mathbf{W} \otimes \mathbf{I}_k\right)\right)\boldsymbol{y}_{t-2} + \boldsymbol{u}_t \tag{112}$$

or

$$\left\{\left(\boldsymbol{I}_m - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2\right) - \left(\boldsymbol{\Psi}_0 + \boldsymbol{\Psi}_1 L + \boldsymbol{\Psi}_2 L^2\right)(\mathbf{W} \otimes \mathbf{I}_k)\right\} \boldsymbol{y}_t = \boldsymbol{u}_t \qquad (113)$$

Remark: The above representation clearly shows that SPVAR is the special case of GVAR. Notice in the GVAR modelling that we are interested in IRFs in terms of $\frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{u}_t}$, which are the combination of the spatial and dynamic ones. So it would be an important issue how to decompose the overall IRFs into the spatial and dynamic components.

Remark: We may also be interested in evaluating the dynamic multipliers in terms of $\frac{\partial \boldsymbol{y}_{1,t+h}}{\partial \boldsymbol{y}_{2t}}$ or *vice versa* where $\boldsymbol{y}_t = (\boldsymbol{y}_{1t}' \boldsymbol{y}_{2t}')'$, but this is not quite straightforward (see the STARDL model above). When we add the global factors, denoted $\boldsymbol{f}_t$, such as the oil or commodity prices, it is straightforwrad to derive the dynamic multipliers in terms of $\frac{\partial \boldsymbol{y}_{t+h}}{\partial \boldsymbol{f}_t}$, say.

6 The Joint Modelling of the Spatial Dependence and Unobserved Factors

There has been the massive development of modelling CSD in the double-indexed panel data (e.g. Pesaran, 2006 and Bai, 2009). There have been two main approaches: The spatial effects deal with weak CSD whilst the factor approach accommodates strong CSD. Recently, a few studies attempted to develop a combined approach that can accommodate both weak and strong CSD.

Bailey et al. (2016) develop the multi-step estimation procedure that can distinguish the relationship between spatial units that is purely spatial from that which is due to common factors. Mastromarco et al. (2015) propose the novel technique in modelling technical efficiency of stochastic frontier panels by combining the exogenously driven factor-based approach and an endogenous threshold regime selection advanced by Kapetanios et al. (2014). Gunella et al. (2015) develop the unified framework for modelling multilateral resistance and bilateral heterogeneity simultaneously in panel gravity models. Shi and Lee (2016), Bai and Li (2015) and Kuersteiner and Prucha (2015) have also developed the framework for jointly modelling spatial effects and interactive effects.

YC: update

6.1 The SDPD Models with Interactive Fixed Effects

6.1.1 SDPD Models with Interactive Fixed Effects by Shi and Lee (2017)

Shi and Lee (2017) consider the following SDPD model with large n and T:

$$\boldsymbol{Y}_{nt} = \lambda_0 \boldsymbol{W}_n \boldsymbol{Y}_{nt} + \gamma_0 \boldsymbol{Y}_{n,t-1} + \rho_0 \boldsymbol{W}_n \boldsymbol{Y}_{n,t-1} + \boldsymbol{X}_{nt}\boldsymbol{\beta}_0 + \boldsymbol{\Gamma}_{0n}\boldsymbol{f}_{0t} + \boldsymbol{U}_{nt} \quad (114)$$

$$\boldsymbol{U}_{nt} = \alpha_0 \tilde{\boldsymbol{W}}_n \boldsymbol{U}_{nt} + \boldsymbol{\varepsilon}_{nt}$$

where $\boldsymbol{Y}_{nt}$ is an n-dimensional column vector of dependent variables and $\boldsymbol{X}_{nt}$ is an $n \times (K-2)$ matrix of exogenous regressors, so that the total number of variables is K. The model accommodates two types of cross sectional dependences, namely, local dependence and global (strong) dependence. Individual units are impacted by time varying unknown common factors $\boldsymbol{f}_{0t}$, which capture global (strong) dependence. The effects of the factors can be heterogeneous on the cross section units, Γ_{0n}. For example, in an earnings regression where Y_{nt} is the wage rate, each row of Γ_{0n} may correspond to a vector of an individual's skills and $\boldsymbol{f}_{0t}$ is its time varying premium. The true number of unobserved factors is assumed to be fixed at r_0 that is much smaller than n and T. The matrix of $n \times r_0$ factor loading Γ_{0n} and the $T \times r_0$ factors $\boldsymbol{F}_{0t} = (f_{01}, f_{02}, ..., f_{0t})'$ are treated as fixed effects parameters[17]. The fixed effects approach allows unknown correlation between the time factors and the regressors. The $n \times n$ spatial weights matrices $\boldsymbol{W}_n$ and $\tilde{\boldsymbol{W}}_n$ represent local (spatial) dependence. $\lambda_0 \boldsymbol{W}_n \boldsymbol{Y}_{nt}$ describes the contemporaneous spatial interactions. $\gamma_0 \boldsymbol{Y}_{n,t-1}$ captures the pure dynamic effect. $\rho_0 \boldsymbol{W}_n \boldsymbol{Y}_{n,t-1}$ is a spatial time lag of interactions, which captures diffusion. The idiosyncratic error $\boldsymbol{U}_{nt}$ with elements of ε_{it} being $iid(0, \sigma_0^2)$ also possesses a spatial structure $\tilde{\boldsymbol{W}}_n$.

The model has a rich spatial structure where a spatial weights matrix is specified to measure the relative magnitudes of spatial interactions. They consider a quasi-maximum likelihood estimation and show estimator consistency and characterize its asymptotic distribution. When n and T are comparable, the estimator is $\sqrt{nT}$ consistent. The Monte Carlo experiment shows that the estimator performs well and the proposed bias correction is effective.

QML Estimation The parameters for the model are $\boldsymbol{\theta} = \left(\boldsymbol{\delta}', \lambda, \alpha\right)'$ with $\boldsymbol{\delta} = (\boldsymbol{\gamma}, \rho, \beta)'$, σ^2, $\boldsymbol{\Gamma}_n$ and $\boldsymbol{F}_T$. Collect the exogenous regressors by the $n \times K$ matrix $\boldsymbol{Z}_{nt} = (\boldsymbol{Y}_{n,t-1}, \boldsymbol{W}_n \boldsymbol{Y}_{n,t-1}, \boldsymbol{X}_{nt})$, where $K = k+2$. Denote $\boldsymbol{S}_n(\boldsymbol{\lambda}) = I_n - \blacksquare W_n$ and $\boldsymbol{R}_n(\alpha) = I_n - \alpha \tilde{W}_n$. The sample averaged quasi-log likelihood function is

$$
\begin{aligned}
Q_{nt}(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\Gamma}_n, \boldsymbol{F}_T) \quad = \quad & -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 + \frac{1}{n}\log |\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{R}_n(\alpha)| \\
& -\frac{1}{2\sigma^2 nT} \sum_{t=1}^{T} \left\{ \boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{Y}_{nt} - \boldsymbol{Z}_{nt}\delta - \boldsymbol{\Gamma}_n \boldsymbol{f}_t \right\}' \boldsymbol{R}_n(\alpha)' \boldsymbol{R}_n(\alpha) \left\{ \boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{Y}_{nt} - \boldsymbol{Z}_{nt}\delta - \boldsymbol{\Gamma}_n \boldsymbol{f}_t \right\}
\end{aligned}
$$

[17] Since $\Gamma_n F_t$ is observationally equivalent to $\Gamma_n H H^{-1} f_t$ for any $r \times r$ invertible matrix H, r^2 restrictions are needed to uniquely determine Γ_n and F_t. One set of restrictions has: $\frac{1}{n}\Gamma_n'\Gamma_n = I_r$ and $F_T'F_T$ is diagonal with positive diagonals. The first condition imposes $\frac{1}{2}(r^2 + r)$ restrictions and the second one provides additional $\frac{1}{2}(r^2 - r)$ ones. There is "rotational indeterminacy" in the sense that the order of the factors can be switched. Let K be the matrix of eigenvectors of $F_T'F_T$. Because $F_T'F_T$ is diagonal under restriction, K is a permutation in the columns of I_r. It can be verified that $\tilde{\Gamma}_n$ and $\tilde{F}_t$ satisfy the restrictions above, with $\tilde{\Gamma}_n = \Gamma_n K$ and $\tilde{F}_T = F_T K$. This indeterminacy can be eliminated by requiring the diagonals of $F_T'F_T$ to be in decreasing or increasing order. Here Γ_n and F_t are treated as nuisance parameters and the above restrictions are not imposed.

Concentrating out σ^2 and dropping the overall constant term for simplicity,

$$\begin{aligned} Q_{nt}(\boldsymbol{\theta}, \boldsymbol{\Gamma}_n, \boldsymbol{F}_T) &= \frac{1}{n} \log |\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{R}_n(\alpha)| \\ &\quad -\frac{1}{2} \log \left[\frac{1}{nT} \sum_{t=1}^{T} \{\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{Y}_{nt} - \boldsymbol{Z}_{nt}\delta - \boldsymbol{\Gamma}_n \boldsymbol{f}_t\}' \boldsymbol{R}_n(\alpha)' \boldsymbol{R}_n(\alpha) \{\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{Y}_{nt} - \boldsymbol{Z}_{nt}\delta - \boldsymbol{\Gamma}_n \boldsymbol{f}_t\} \right] \end{aligned}$$

is a concentrated sample averaged log likelihood function of $\boldsymbol{\theta}$, $\boldsymbol{\Gamma}_n$ and $\boldsymbol{F}_T$. Because for our estimation method, no restriction is imposed on $\boldsymbol{\Gamma}_n$ and $R_n(\alpha)$ is assumed invertible for α, optimizing with respect to $\boldsymbol{\Gamma}_n \in R^{n \times r}$ is equivalent to optimizing with respect to the transformed $\tilde{\Gamma}_n$ with $\tilde{\Gamma}_n = R_n(\alpha)\boldsymbol{\Gamma}_n$. The objective function can be equivalently written as

$$\begin{aligned} Q_{nt}(\boldsymbol{\theta}, \tilde{\boldsymbol{\Gamma}}_n, \boldsymbol{F}_T) &= \frac{1}{n} \log |\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{R}_n(\alpha)| \\ &\quad -\frac{1}{2} \log \left[\frac{1}{nT} \sum_{t=1}^{T} \{\boldsymbol{R}_n(\alpha) (\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{Y}_{nt} - \boldsymbol{Z}_{nt}\delta)\}' \{\boldsymbol{R}_n(\alpha) (\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{Y}_{nt} - \boldsymbol{Z}_{nt}\delta)\} \right] \end{aligned}$$

Because the parameter of interest is $\boldsymbol{\theta}$ and it is easier to optimize with respect to a finite dimensional vector, we concentrate out factors and their loadings using the principal component theory: for an $n \times T$ matrix $\boldsymbol{H}_{nT}$,

$$\begin{aligned} &\min_{\boldsymbol{F}_T \in R^{T \times r}, \tilde{\Gamma}_n \in R^{n \times r}} tr \left(\boldsymbol{H}_{nT} - \tilde{\boldsymbol{\Gamma}}_n, \boldsymbol{F}_T' \right) \left(\boldsymbol{H}_{nT} - \tilde{\boldsymbol{\Gamma}}_n, \boldsymbol{F}_T' \right)' \\ = &\min_{\boldsymbol{F}_T \in R^{T \times r}} tr \left(\boldsymbol{H}_{nT} \boldsymbol{M}_{\boldsymbol{F}_T} \boldsymbol{H}_{nT}' \right) = \sum_{i=r+1}^{n} \mu_i \left(\boldsymbol{H}_{nT} \boldsymbol{H}_{nT}' \right) \end{aligned}$$

where the jth column of $\hat{\boldsymbol{F}}_T$ is the eigenvector corresponding to the jth largest eigenvalue of $\boldsymbol{H}_{nT}' \boldsymbol{H}_{nT}$. Because $\tilde{\boldsymbol{\Gamma}}_n \boldsymbol{F}_T'$ cannot be separately identified from $\tilde{\boldsymbol{\Gamma}}_n \boldsymbol{M} \boldsymbol{M}^{-1} \boldsymbol{F}_T'$ for an invertible matrix $\boldsymbol{M}$, the identification conditions that $\boldsymbol{F}_T' \boldsymbol{F}_T = \boldsymbol{I}_r$ and $\tilde{\boldsymbol{\Gamma}}_n' \tilde{\boldsymbol{\Gamma}}_n$ is diagonal have been imposed. The estimated factors and factor loadings will be rotations of their true values. However, the asymptotic distribution of our QML estimator is unaffected by the normalizations because the space spanned by the factors and their loadings do not change. The concentrated log likelihood is

$$\begin{aligned} Q_{nt}(\boldsymbol{\theta}) &= \max_{\boldsymbol{F}_T \in R^{T \times r}, \tilde{\Gamma}_n \in R^{n \times r}} Q_{nt}(\boldsymbol{\theta}, \tilde{\boldsymbol{\Gamma}}_n, \boldsymbol{F}_T) \\ &= \frac{1}{n} \log |\boldsymbol{S}_n(\boldsymbol{\lambda}) \boldsymbol{R}_n(\alpha)| - \frac{1}{2} \log L_{nT}(\boldsymbol{\theta}) \end{aligned}$$

with

$$L_{nT}(\boldsymbol{\theta}) = \mu \left[\frac{1}{nT} \sum_{t=1}^{T} \boldsymbol{R}_n(\alpha) \left(\boldsymbol{S}_n(\boldsymbol{\lambda}) - \sum_{k=1}^{K} \boldsymbol{Z}_k \delta_k \right) \left(\boldsymbol{S}_n(\boldsymbol{\lambda}) - \sum_{k=1}^{K} \boldsymbol{Z}_k \delta_k \right)' \boldsymbol{R}_n(\alpha)' \right]$$

The QML estimator is

$$\hat{\boldsymbol{\theta}}_{nT} = \arg\max_{\blacksquare\in\blacksquare} Q_{nt}(\boldsymbol{\theta}).$$

The estimate for $\tilde{\Gamma}_n$ can be obtained as the eigenvectors associated with the first r largest eigenvalues of $\boldsymbol{R}_n(\alpha)\left(\boldsymbol{S}_n(\boldsymbol{\lambda}) - \sum_{k=1}^K \boldsymbol{Z}_k\delta_k\right)\left(\boldsymbol{S}_n(\boldsymbol{\lambda}) - \sum_{k=1}^K \boldsymbol{Z}_k\delta_k\right)' \boldsymbol{R}_n(\alpha)'$. By switching n and T, the estimate for $\boldsymbol{F}_T$ can be similarly obtained. Note that the estimated $\tilde{\Gamma}_n$ and $\boldsymbol{F}_T$ are not unique, as $\tilde{\boldsymbol{\Gamma}}_n\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{F}_T'$ is observationally equivalent to $\tilde{\boldsymbol{\Gamma}}_n\boldsymbol{F}_T'$ for any invertible $r \times r$ matrix $\boldsymbol{H}$. However, the column spaces of $\tilde{\Gamma}_n$ and $\boldsymbol{F}_T$ are invariant to $\boldsymbol{H}$, hence the projectors $\boldsymbol{M}_{\tilde{\Gamma}_n}$ and $\boldsymbol{M}_{F_T}$ are uniquely determined.

6.1.2 Dynamic spatial panel data models with common shocks by Bai and Li (2015)

Bai and Li (2015) consider jointly modeling spatial interactions, dynamic interactions and common shocks within the following model:

$$y_{it} = \alpha_i + \rho\sum_{j=1}^N w_{ij,N}y_{jt} + \delta y_{it-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + e_{it} \tag{115}$$

where y_{it} is the dependent variable, $\boldsymbol{x}_{it} = (x_{it1}, ..., x_{itk})'$ is a k-dimensional vector of explanatory variables, $\boldsymbol{f}_t$ is an r-dimensional vector of unobservable common shocks; $\boldsymbol{\lambda}_i$ is the heterogenous response to the common shocks, $\boldsymbol{W}_N = (w_{ij,N})_{N\times N}$ is a spatial weights matrix whose diagonal elements $w_{ii,N}$ are 0, and e_{it} are the idiosyncratic errors. $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$ captures the common-effects, $\sum_{j=1}^N w_{ij,N}y_{jt}$ captures the spatial effects, and δy_{it-1} captures the dynamic effects. The joint modeling allows one to test which type of effects is present. We may test $\rho = 0$ while allowing common-shocks effects and dynamic effects, or similarly, we may determine if the number of factors is zero in a model with spatial effects and dynamic effects. The features of model (115) make it flexible enough to cover a wide range of applications.

An additional feature of the model is the allowance of cross sectional heteroskedasticity. If heteroskedasticity exists but homoskedasticity is imposed, MLE can be inconsistent. Under large-N, the consistency analysis for MLE under heteroskedasticity is challenging even for spatial panel models without common shocks, owing to the simultaneous estimation of a large number of variance parameters along with $(\rho, \delta, \boldsymbol{\beta})$. Interestingly, we show that the limiting variance of the MLE is not of a sandwich form if heteroskedasticity is allowed.

The spatial interaction on the dependent variable gives rise to the endogeneity problem while the spatial interaction on the errors does not. As a result, existing estimation methods on the common shocks models such as Pesaran (2006) and Bai (2009) cannot be directly applied to model (115) due to the endogeneity from the spatial interactions.

Real data often have complicated correlation over cross section and time. Modeling, estimating and interpreting the correlations in data are particularly important in economic analysis. This paper integrates several correlation-modeling techniques and propose dynamic spatial panel data models with common shocks to accommodate possibly complicated correlation structure over cross section and time. A large number of incidental parameters exist. The QML is proposed and heteroskedasticity is explicitly estimated. Our analysis indicates that the MLE has a non-negligible bias. We propose a bias correction method. The simulations reveal the excellent finite sample properties of the QMLE after bias correction.

6.1.3 Dynamic Spatial Panel Models: Networks, Common Shocks, and Sequential Exogeneity by Kuersteiner and Prucha (2015)

We consider panel data $\{y_t, x_t, z_t\}_{t=1}^T$, where $y_t = [y_{1t}, ..., y_{nt}]'$, $x_t = [x'_{1t}, ..., x'_{nt}]'$, and $z_t = [z'_{1t}, ..., z'_{nt}]'$ denote the vector of the endogenous variables, and the matrices of k_x weakly exogenous and k_z strictly exogenous variables. The specification allows for temporal dynamics in that x_{it} may include a finite number of time lags of the endogenous variables.

We allow in each period t for the regressors and disturbances to be affected by common shocks. Alternatively we allow for CSD from "spatial lags" in the endogenous variables, the exogenous variables and in the disturbance process. Our specification accommodates higher order spatial lags, as well as time lags in x_{it}. Spatial lags represent weighted cross sectional averages, where the weights will be reflective of some measure of distance between units. The spatial weights will be summarized by $n \times n$ spatial weight matrices denoted as $W_{pt} = (w_{p,ijt})$ and $M_{qt} = (m_{q,ijt})$. $\varepsilon_t = [\varepsilon_{1t}, ..., \varepsilon_{nt}]'$ denotes the vector of regression disturbances, $u_t = [u_{1t}, ..., u_{nt}]'$ denotes the vector of idiosyncratic disturbances, and μ is an $n \times 1$ vector of unobserved factor loadings or individual specific fixed effects, which may be time varying through a common unobserved factor f_t. Let λ and ρ be P, Q dimensional vectors of parameters with typical elements λ_p and ρ_q and define

$$R_t(\lambda) = \sum_{p=1}^{P} \lambda_p W_{pt} \quad \text{for SAR}$$

$$R_t^*(\rho) = I - \sum_{q=1}^{Q} \rho_q M_{qt} \quad \text{for a spatial autoregressive error term}$$

$$R_t^*(\rho) = \left(I + \sum_{q=1}^{Q} \rho_q M_{qt} \right)^{-1} \quad \text{for a spatial moving average error term}$$

Then, the dynamic and cross sectionally dependent panel data model can be written as

$$\begin{aligned} y_t &= R_t(\lambda) y_t + x_t \beta_x + z_t \beta_z + \varepsilon_t = X_t \delta + \varepsilon_t \\ R_t^*(\rho) \varepsilon_t &= \mu f_t + u_t \end{aligned} \tag{116}$$

where $X_t = [M_{1t}y_t, ..., M_{Pt}y_t, x_t, z_t]$ and $\delta = \left[\lambda', \beta'\right]'$ with $\beta = \left(\beta_x', \beta_z'\right)'$. As a normalization we take

$$w_{p,iit} = m_{q,iit} = 0 \text{ and } f_T = 1$$

Note that (116) is a system of n equations describing simultaneous interactions between the individual units. The weighted averages, $\bar{y}_{p,it} = \sum_{j=1}^{n} w_{p,ijt}y_{jt}$ and $\bar{\varepsilon}_{q,it} = \sum_{j=1}^{n} m_{q,ijt}\varepsilon_{jt}$ model contemporaneous direct cross-sectional interactions. We allow the weights to be stochastic and endogenous in that they can depend on $\mu_1, ..., \mu_N$ and u_{it}, and can be correlated with the disturbances ε_t. This extension is important to model sequential group formation as in Carrell et.al. (2013) or endogenous network formation as in Goldsmith-Pinkham and Imbens (2013).

The reduced form is given by

$$y_t = \left(I_n - R_t\left(\lambda\right)\right)^{-1} W_t\delta + \left(I_n - R_t\left(\lambda\right)\right)^{-1} \varepsilon_t \tag{117}$$

$$\varepsilon_t = R_t^*\left(\rho\right)^{-1}\left(\mu f_t + u_t\right)$$

Applying a Cochrane-Orcutt type transformation by premultiplying the first equation in (116) with $R_t^*\left(\rho\right)$ yields

$$R_t^*\left(\rho\right) y_t = R_t^*\left(\rho\right) W_t\delta + \mu f_t + u_t \tag{118}$$

Three examples:

1. The social interactions model by Graham (2008) illustrates the use of both spatial interaction terms and interactive effects in a social interaction model;

2. The analysis of the group level heterogeneity is based on Carrell et. al. (2013), which illustrates the use of higher order, and data-dependent spatial lags to model within-group heterogeneity. By allowing $R_t\left(\lambda\right)$ to depend on predetermined outcomes we can accommodate the fact that group membership is not exogenous;

3. The next example is in the area of health, and considers the spread of an infectious disease.

Kuersteiner and Prucha (2015) consider a class of GMM estimators, allowing for CSD due to spatial lags and due to common shocks. They expand the scope of the existing literature by allowing for endogenous spatial weight matrices, time-varying interactive effects, as well as weakly exogenous covariates. An important area of application is in social interaction and network models where specification can accommodate data dependent network formation. Identification of spatial interaction parameters is achieved through a combination of linear and quadratic moment conditions. They develop an orthogonal forward differencing transformation to aid in the estimation of factor components while maintaining orthogonality of moment conditions. In the social interactions example, orthogonal forward differencing amounts to controlling for unobserved correlated effects by combining multiple outcome measures.

6.2 Other Approaches

Introducing factors into a model is a way to specify possible CSD. Chudik, Pesaran and Tosetti (2011) introduce the concepts of time-specific weak and strong CSD in panel data models. A factor model allows time effects to interact with spatial units with different intensity.

Pesaran and Tosetti (2011) investigate the following model with both common factors and spatial correlation:

$$y_{it} = \alpha_i' \boldsymbol{d}_t + \beta_i' \boldsymbol{x}_{it} + \gamma_i' \boldsymbol{f}_t + e_{it}, \tag{119}$$

where $\boldsymbol{d}_t$ is a $k \times 1$ vector of observed common effects, $\boldsymbol{x}_{it}$ is the $k_x \times 1$ vector of observed individual specific regressors, $\boldsymbol{f}_t$ is an m-dimensional vector of unobservable common factors, and γ_i is the $m \times 1$ vector of factor loadings. The common factors $\boldsymbol{f}_t$ simultaneously affect all cross section units, albeit with different intensity as measured by γ_i'. The e_{it} follows some spatial process, either an SAR or SMA process.

To estimate the parameters of interest (the mean of β_i), we can use the mean group estimator

$$\hat{\beta}_{MG} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_i$$

where $\hat{\beta}_i$ is the OLS estimate of $\boldsymbol{y}_i$ regressed on $\boldsymbol{X}_i$ after the orthogonal projection of data by a $T \times (k + k_x + 1)$ matrix $\boldsymbol{M}_D$ where D consists of observed common factor and cross sectional average of dependent and independent variables, e.g.

$$\hat{\beta}_i = \left(\boldsymbol{X}_i' \boldsymbol{M}_D \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i' \boldsymbol{M}_d \boldsymbol{y}_i$$

where $\boldsymbol{X}_i$ is a $T \times k_x$ vector of regressors for the ith unit, $\boldsymbol{y}_i$ is a $T \times 1$ vector of the dependent variable for the ith unit, and $\boldsymbol{M}_D = \boldsymbol{I}_T - \boldsymbol{D} \left(\boldsymbol{D}'\boldsymbol{D} \right)^{-1} \boldsymbol{D}$. Alternatively, the regression coefficient β can be obtained by a pooled estimate:

$$\hat{\beta}_P = \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{M}_D \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{M}_d \boldsymbol{y}_i$$

Given the spatial structure in e_{it}, Pesaran and Tosetti (2011) derive the asymptotic properties of $\hat{\beta}_{MG}$ and $\hat{\beta}_{MG}$ under difference scenarios such as whether the unobserved common factor is present or not and whether the regression parameters are heterogeneous or not. Pesaran and Tosetti (2011) then consider the Common Correlated Effects (CCE) estimator advanced by Pesaran (2006), which continues to yield estimates of the slope coefficients that are consistent and asymptotically normal. Holly, Pesaran, and Yamagata (2010) use (119) to analyze the changes in real house prices in the US. They have also specified a spatial process in e_{it}, which is shown to be significant after controlling for unobserved common factors.

Chudik, Pesaran, and Tosetti (2011) extend the model by allowing additional weakly dependent common factors. Their model is

$$y_{it} = \alpha_i' d_t + \beta_i' x_{it} + \gamma_i' f_t + \lambda_i' n_t + e_{it}, \tag{120}$$

where n_t, uncorrelated with the regressor x_{it}, is the additional weakly dependent common factors with dimension of the factor loading coefficients λ_i going to infinity. They show that the CCE method still yields consistent estimates of the mean of the slope coefficients and the asymptotic normal theory continues to be applicable.

Holly, Pesaran, and Yamagata (2011) provide a method for the analysis of the spatial and temporal diffusion of shocks in a dynamic system. They generalize VAR panel models with unobserved common factors to incorporate spatial elements in the dynamic coefficient matrix. With coefficients being spatial unit specific, consistent estimation has emphasized on T tending to infinity while the number of spatial units can be moderate or large. They use changes in real house prices within the UK economy, and analyze the effect of shocks using generalized spatiotemporal impulse responses.

The spatial panel data models assume a time invariant spatial weights matrix. When the spatial weights matrix is constructed with economic/socioeconomic distances or demographic characteristics, it can be time varying. For example, Case, Hines, and Rosen (1993) on state spending have weights based on the difference in the percentage of the population that is black. Lee and Yu (2012b) investigate the QML estimation of SDPD models where spatial weights matrices can be time varying. They find that QML estimate is consistent and asymptotically normal. Monte Carlo results show that, when spatial weights matrices are substantially varying over time, a model misspecification of a time invariant spatial weights matrix may cause substantial bias in estimation.

6.2.1 A Nonlinear Panel Data Model of Cross-Sectional Dependence

Kapetanios, Mitchell and Shin (2014) propose a nonlinear panel data model which can endogenously generate both 'weak' and 'strong' CSD. The model's distinguishing characteristic is that a given agent's behaviour is influenced by an aggregation of the views or actions of those around them. The model allows for considerable flexibility in terms of the genesis of this herding or clustering type behaviour. At an econometric level, the model is shown to nest various extant dynamic panel data models. These include panel AR models, spatial models, which accommodate weak dependence only, and panel models where cross-sectional averages or factors exogenously generate strong CSD. An important implication is that the appropriate model for the aggregate series becomes intrinsically nonlinear, due to the clustering behaviour.

We propose dynamic nonlinear panel data models:

$$x_{i,t} = \rho \sum_{j=1}^{N} w_{ij}\left(x_{-i,t-1}, x_{i,t-1}; \gamma\right) x_{j,t-1} + \epsilon_{i,t}, \quad i = 1, ...N, \quad t = 1, ...T, \tag{121}$$

where $x_{-i,t} = (x_{1,t}, x_{2,t}, ..., x_{i-1,t}, x_{i+1,t}, ..., x_{Nt})'$ and $\sum_{j=1}^{N} w_{ij}\left(x_{-i,t-1}, x_{i,t-1}; \gamma\right) = 1$. This form of the model signifies that $x_{i,t}$ depends, possibly in a nonlinear fashion depending on how w_{ij} is parameterised, on weighted averages of past values of $x_t = (x_{1,t}, ..., x_{Nt})'$, where the weights depend on x_{t-1}. One particular motivation is structural and follows from the claim that it mimics structural interactions between economic units. Another, more econometric, justification notes that this model can accommodate generic forms of CSD, including evolving clusters.

The model in (448) encompasses a wide variety of nonlinear specifications. We place particular emphasis on specifications where the weights depend on x_{t-1} only through distances of the form $|x_{j,t-1} - x_{i,t-1}|$. This type of specification is easy to analyse, based on a threshold mechanism, to illustrate the class of models. This model nests a variety of dynamic panel data models, such as panel data AR models and panel models where cross-sectional averages are used to pick up CSD (Pesaran, 2006). Interestingly, it is also closely related to factor or interactive effects models (Bai, 2009).

Our models provide an intuitive means by which many forms of cross-sectional dependence can arise in a large panel dataset comprised of variables of a 'similar' nature that relate to different agents/units. These variables might be the disaggregates underlying often studied macroeconomic or financial aggregates, such as economy-wide inflation or the S&P500 index. In particular, the model allows these different economic units to cluster; and for these clusters (including their number) to evolve over time. Such clustering also has implications when modelling and forecasting the aggregate of these units.

The degree of CSD in our models can vary, from a case where it is similar to standard factor models, for which the largest eigenvalue of the variance covariance matrix of the data tends to infinity at rate N, to the case of very weak or no factor structure where this eigenvalue is bounded as $N \to \infty$. Of course, all intermediate cases can arise as well. Our model constitutes the first attempt to introduce endogenous cross-sectional dependence into a panel modelling framework.

Let $x_{i,t}$ denote the variable of interest, such as the agent's income or the agent's view of the future value of some macroeconomic variable, at time t, for agent i. Then, we specify:

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \epsilon_{i,t}, \quad t = 2, ..., T, \quad i = 1, ..., N, \tag{122}$$

where

$$m_{i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right),$$

$\{\epsilon_{i,t}\}_{t=1}^{T}$ is an error process, $\mathcal{I}(.)$ is the indicator function and $-1 < \rho < 1$. Verbally, the above model states that $x_{i,t}$ is influenced by the cross-sectional average of a selection of $x_{j,t-1}$ and in particular that the relevant $x_{j,t-1}$ are those

that lie closest to $x_{i,t-1}$. The model involves a K nearest neighbor mechanism except that it is in the data generating process and not as a technique to estimate an unknown function. This formalises the intuitive idea that people are affected more by those with whom they share common views or behaviour. The model may be equally viewed as a descriptive model of agents' behaviour, reflecting the fact that 'similar' agents are affected by 'similar' effects, or as a structural model of agents' views whereby agents use the past views of other agents, similar to them to form their own views. The interaction term in (122) may then be thought of capturing the (cross-sectional) local average or common component of their views. This idea of commonality has various clear, motivating and concrete examples in a variety of social science disciplines, such as psychology and politics. In economics and finance, the herding could be rational (imitative herding) or irrational.

A deterministic form of the above model has been analysed previously in the mathematical and system engineering literature. They have analysed a continuous form of the restricted version of (122) given by

$$x_{i,t} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq 1\right) x_{j,t-1}, \quad t = 2, ..., T, \quad i = 1, ..., N, \tag{123}$$

where $m_{i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq 1\right)$.

By setting $r = 0$, we obtain a simple panel autoregressive model:

$$x_{i,t} = \rho x_{i,t-1} + \epsilon_{i,t}. \tag{124}$$

On the other hand, letting $r \to \infty$, we obtain the following model

$$x_{i,t} = \frac{\rho}{N} \sum_{j=1}^{N} x_{j,t-1} + \epsilon_{i,t}, \tag{125}$$

where past cross-sectional averages of opinions inform, in similar fashions, current opinions. Recently, the use of such cross-sectional averages has been advocated by Pesaran (2006) as a means of modelling CSD in the form of unobserved factors. However, in our case, the use of cross-sectional averages is a limiting case of a 'structural' nonlinear model.

Factor models have the property that both the maximum eigenvalue and the row/column sum norm of the covariance matrix of $x_t = (x_{1,t}, ..., x_{N,t})'$ tend to infinity at rate N, as $N \to \infty$. In contrast, for other models such as spatial AR or MA models, these quantities are bounded, implying that they exhibit much lower degrees of CSD than factor models. We show that the column sum norm of the covariance matrix of x_t when x_t follows (122) is $O(N)$. Thus, the model is more similar to factor models than spatial models. Interestingly, there are versions of (122) that resemble spatial models. Another finding is that (125) implies a covariance matrix for x_t with a column sum norm that is $O(1)$. This is surprising, given the similarity that cross-sectional average schemes have with factor models as detailed in Pesaran (2006).

Factor models are intrinsically reduced form; they focus on modelling CSD using an exogenously given number of unobserved factors. Since our model nests (125), it can approximate a factor model when $r \to \infty$. On the other hand, our model has a clear parametric structure, which is a feature shared by dynamic spatial model. But, our models are more general than spatial models, in the sense that the weighting schemes are estimated endogenously, rather than assumed *ex ante*. It is worth noting that the factor model cannot accommodate the weak CSD, in contrast to the extensions of our nonlinear model. The nonlinear model can be seen to lie between the two extremes characterised by weakly cross-sectionally dependent spatial models and strongly cross-sectionally dependent factor models.

Further suggestions on the empirical applications **Inflation expectations**: Here, we have considered the basic model with fixed effects:

$$\pi_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| \leq r\right) \pi_{j,t-1} + \epsilon_{i,t} \tag{126}$$

where π is the one-quarter ahead CPI inflation rate forecast, $\nu_i \sim iid(0, \sigma_\nu^2)$, and obtained the within estimator of ρ along with the consistent estimator of r. We can provide some economic interpretations for $\hat{\rho}$ and $\hat{r}$ in terms of persistence or inertia and the relative distance of similarity. Notable exceptions are to estimate two extreme models, denoted PAR and CSA, respectively:

$$\pi_{i,t} = \nu_i + \rho \pi_{i,t-1} + \epsilon_{i,t} \tag{127}$$

$$\pi_{i,t} = \nu_i + \rho \bar{\pi}_{t-1} + \epsilon_{i,t} \tag{128}$$

It is interesting to see how the estimates of ρ differ for each of three models. Assuming that the overall performance of the model, (126), is superior, we then move to estimate the extensions as discussed in the model of the form:

$$\pi_{i,t} = \nu_i + \rho_1 \tilde{\pi}_{i,t-1} + \rho_2 \tilde{\pi}_{i,t-1}^c + \epsilon_{i,t} \tag{129}$$

where $\tilde{\pi}_{i,t-1}$ and $\tilde{\pi}_{i,t-1}^c$ are the respective cross-section averages related to similar and dissimilar forecasters given by

$$\tilde{\pi}_{i,t-1} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| \leq r\right) \pi_{j,t-1}$$

$$\tilde{\pi}_{i,t-1}^c = \frac{1}{N - m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| > r\right) \pi_{j,t-1}$$

We can also test the hypothesis of $\rho_2 = 0$ (informational contents arising from dissimilar forecasters). If $\rho_2 \neq 0$, what's the prior implication of the sign of ρ_2? In other words, in forming the own forecast, how does each forecaster use any (past) information from others?

Next to find a way of decomposing the (partial) aggregate parameter, ρ in (126) or $\rho_1$in (129) into the own effect and the neighbor effect. One obvious candidate is to consider:

$$\pi_{i,t} = \nu_i + \rho_0 \pi_{i,t-1} + \rho_1 \mathring{\pi}_{i,t-1} + \epsilon_{i,t} \tag{130}$$

$$\pi_{i,t} = \nu_i + \rho_{10} \pi_{i,t-1} + \rho_{11} \mathring{\pi}_{i,t-1} + \rho_2 \tilde{\pi}^c_{i,t-1} + \epsilon_{i,t} \tag{131}$$

where

$$\mathring{\pi}_{i,t-1} = \frac{1}{\mathring{m}_{i,t}} \sum_{j=1, j\neq i}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| \leq r\right) \pi_{j,t-1}$$

Anselin et al. (2008) distinguish spatial dynamic models into four categories based on the following general time-space-dynamic specification:

$$x_{i,t} = \rho_0 x_{i,t-1} + \rho_1 \sum_{j\neq i} w_{ij} x_{j,t-1} + \beta \sum_{j\neq i} w_{ij} x_{j,t} + v_i + \epsilon_{i,t} \tag{132}$$

Here, $\sum_{j\neq i} w_{ij} y_{jt}$ and $\sum_{j\neq i} w_{ij} y_{jt-1}$ are *a first order spatial lag* and its time-lagged value, respectively. The parameter, ρ_0, captures serial dependence of x_{it}, β represents the intensity of a contemporaneous spatial effect and ρ_1 captures *space time autoregressive dependence* (diffusion). Most studies focus on the stable case with $\rho_0 + \rho_1 + \beta < 1$. This specification, (132), includes various special cases:

- if $\rho_0 = \rho_1 = 0$, we obtain a 'pure-space recursive' model in which dependence results from the neighborhood locations in the previous time period;
- if $\beta = 0$, the model is reduced to a 'time space recursive' model in which dependence relates to both the location itself ($x_{i,t-1}$) and its neighbors in the previous time period $\sum_{j\neq i} w_{ij} x_{j,t-1}$;
- if $\rho_1 = 0$, we obtain a 'time space simultaneous' model which includes the time lag ($x_{i,t-1}$) and the spatial lag, $\sum_{j\neq i} w_{ij} x_{j,t}$;
- if $\rho_0 = \rho_1 = 0$, we are dealing with a spatial autoregressive model on panel data, while if $\rho_1 = \beta = 0$ we obtain a 'simple' dynamic model.

According to Anselin (2001) and Abreu et al. (2005), the addition of a spatially lagged dependent variable causes simultaneity and endogeneity problems and thus a candidate consistent estimator should lie between the OLS and within estimates.

Our model, (130), is similar to the time-space recursive model considered in Korniotis (2010), who apply it to investigate the issue of internal versus external habit formation using the annual consumption data for the U.S. states, and find that state consumption growth is not significantly affected by its own (lagged) consumption growth but it is affected by lagged consumption growth of nearby states. Notice that the weight w_{ij} measures the importance of $x_{j,t-1}$ on x_{it}.

The weights are observed quantities, which are known to the econometrician, and therefore exogenous. Because the spatial-time lag, $\sum_{j=1}^{N} w_{ij} x_{j,t-1}$, is a weighted average of past consumption choices of other cross-sectional units, it is the measure of the catching-up habit.

There is a trade-off between our model, (130) and the time space recursive model employed by Korniotis (2010). In (130), the neighbors are selected endogenously but the equal weights are imposed to the selected neighbors. By contrast, in the Korniotis's model, the neighbors are selected exogenously, but the weights are selected in a flexible manner albeit not time-varying. The application of the model, (130) to similar issue of the consumption habit formation will provide an interesting insight.[18]

Unless $\rho_2 = 0$, the model (131) should be more general, and it is interesting to find out the potential application, say GVC, upstream and downstream or asymmetry?

YC: update

It is also interesting to investigate how we relate our approach to the Sias' (2004) approach to an analysis of herding. The idea is to estimate ρ from (126), and find a way to decompose:

$$\rho = \rho_{own} + \rho_{neighbor}$$

following the the Sias' approach. But, the analogy is not quite one-to-one. The potential advantage of this approach is the possible robustness of this measure which can also be used for a finite T, as well.

We generalise (126) and allow different weights to the selected neighbors as follows:

$$x_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) w_{ij} x_{j,t-1} + \epsilon_{i,t} \tag{133}$$

where we may consider the following weights

$$w_{ij} = \frac{d_{ij}^{-2}}{\sum_{j=1}^{N} d_{ij}^{-2}}, \; d_{ij} = |x_{i,t-1} - x_{j,t-1}| \tag{134}$$

(then, how we define w_{ii}, just normalised to 1?). The estimation can be done in two steps: first, the consistent estimate of r is obtained from (126). Then, construct the weights by (134) and the associated cross-section averages, and estimate the model, (133). Or possibly more complicated due to the grid search over r. If successful, then our approach is more general than the spatial models.

Next, we may consider the following extension of (126):

$$x_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(\left|x_{t-1}^{\max} - x_{j,t-1}\right| \leq r\right) x_{j,t-1} + \epsilon_{i,t} \tag{135}$$

[18] We can allow the weights to be inversely proportional to the distance once the threshold parameter is consistently estimated.

where $x_{t-1}^{\max} = \max_j x_{j,t-1}$, such that the distance is measured with respect to the best performer rather than the unit i. The alternative functional type can also be considered.

Remark: From the empirical point of view, the following consideration may be useful: Suppose that the distance between $x_{i,t-1}$ and $x_{j,t-1}$ or more generally between q_{it} and q_{jt}, can be regarded as the sort of similarity measure, and that the parameter may measure the impact of the certain policy. We then estimate the value of ρ's under different values of r, and make a 2-dimensional plot to investigate whether the relationship between ρ and r is monotonic or nonlinear. This approach may be related to the recent GMM analytic approach where the sample moment condition is not equal to zero for over-identified case.

6.2.2 MSS (2015) Approach to Modelling Technical Efficiency in Cross Sectionally Dependent Stochastic Frontier Panels

Mastromarco, Serlenga and Shin (2015) propose a unified framework for accommodating both time- and cross-section dependence in modelling technical efficiency in stochastic frontier models. The proposed approach enables us to deal with both weak and strong forms of CSD by introducing exogenously driven common factors and an endogenous threshold selection mechanism. Using the dataset of 26 OECD countries over the period 1970-2010, we provide the satisfactory estimation results for the production technology parameters and the associated efficiency ranking of individual countries. We find positive spillover effect on efficiency, supporting the hypothesis that knowledge spillover is more likely to be induced by technological proximity. Furthermore, our approach enables us to identify efficiency clubs endogenously.

We begin with the standard Cobb-Douglas production function:

$$y_{it} = \boldsymbol{\beta}' \mathbf{x}_{it} + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{136}$$

where y_{it} is a logarithm of output of country i at time t, $\mathbf{x}_{it}$ a $k \times 1$ vector of (logged) production inputs, $\boldsymbol{\beta}$ a $k \times 1$ vector of structural parameters, and ε_{it} is the composite stochastic errors including the idiosyncratic disturbance (v_{it}) and time varying (logged) technical inefficiency (u_{it}):

$$\varepsilon_{it} = v_{it} - u_{it}. \tag{137}$$

Mastromarco *et al.* (2013) propose the panel stochastic frontier model with unobserved factors for modelling the time-varying technical inefficiency, u_{it}:

$$u_{it} = \alpha_i + \boldsymbol{\lambda}_i' \mathbf{f}_t,\ i = 1, ..., N,\ t = 1, ..., T, \tag{138}$$

where α_i is (unobserved) individual effects, and $\mathbf{f}_t$ is an $r \times 1$ vector of unobserved factors that are expected to provide a proxy for nonlinear and complex trending patterns associated with globalisation and the business-cycle. This factor approach clearly accommodates strong CSD.

Recent literature emphasises that the individual country's total factor productivity (TFP) is likely to be significantly affected by economic performance of

neighboring or frontier countries. To allow for such spatial dependence, Ertur and Koch (2007) develop a growth model in which technological interdependency is specified through spatial externalities. In particular, the productivity shocks in SFA are assumed to be spatially correlated:

$$\boldsymbol{\varepsilon}_t = \rho \mathbf{W} \boldsymbol{\varepsilon}_t + \mathbf{e_t},\ t = 1, ..., T, \tag{139}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$, $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{N}$ is the $N \times N$ spatial weight matrix with diagonal elements equal to zero, ρ is a spatial autoregressive parameter, and $\mathbf{e_t} = (e_{1t}, ..., e_{Nt})'$ the vector of zero-mean idiosyncratic disturbances. The elements in $\mathbf{W}$ are selected exogenously on the basis of geographic or economic proximity measures such as contiguity, physical/economic/climatic distances or dissimilarities.

The spatial models can only control for weak CSD whilst the factor-based models can allow for strong CSD. In this regard the spatial-based approach is likely to produce biased estimates in the presence of strong CSD. While spatial autoregressive models are generally estimated by MLE, Pesaran (2006) and Bai (2009) develop two alternative consistent estimation methodologies in the presence of strong CSD. These studies clearly suggest that factors can play an important role in the cross sectionally correlated panels. Nevertheless, factor-based models impose an assumption that the strong CSD is mainly driven by an exogenously given unobserved factors. Recently, KMS propose an alternative approach that allows the CSD to be determined endogenously.

Suppose that the product of a country i at time t, Y_{it}, is determined by the levels of labor input and private capital, L_{it} and K_{it}. It is also affected by the Hicks-neutral multi-factor productivity TFP:

$$Y_{it} = TFP_{it} F(L_{it}, K_{it}), \tag{140}$$

where TFP_{it} depends on the technological progress. The TFP_{it} component can be decomposed into the level of technology A_{it}, a measurement error w_{it}, and the efficiency measure τ_{it} with $0 < \tau_{it} \leq 1$:

$$TFP_{it} = A_{it} \tau_{it} w_{it}. \tag{141}$$

By writing (140) in log form:

$$y_{it} = \alpha + \beta_1 k_{it} + \beta_2 l_{it} - u_{it} + v_{it}, \tag{142}$$

with the two-way error components structure given by

$$\varepsilon_{it} = v_{it} - u_{it}, \tag{143}$$

where $v_{it} = \ln w_{it}$ and $u_{it} = -\ln(\tau_{it})$ is the term measuring the (time-varying) technical inefficiency.

We propose that innovators consider the behaviour of other agents as:

$$u_{it} = \alpha_i + \rho \tilde{u}_{it}(r) + \boldsymbol{\lambda}_i' \mathbf{f}_t. \tag{144}$$

where

$$\tilde{u}_{it}(r) = \frac{1}{m_{it}} \sum_{j=1}^{N} I\left(\left|u_{t-1}^{*} - u_{jt-1}\right| \leq r\right) u_{jt-1}, \tag{145}$$

and r is the threshold parameter that is determined endogenously and u_{t-1}^{*} is the efficiency of the best performing country and $m_{it} = \sum_{j=1}^{N} I\left(\left|u_{t-1}^{*} - u_{jt-1}\right| \leq r\right)$. The $\tilde{u}_{it}(r)$ is an interaction term that may be thought of capturing the cross sectional local average of the best practices or common technology. The specification in (144) explicitly allows the dynamics of technical inefficiency to be interacted spatially. *A priori*, we expect that such externalities can be captured by a negative ρ. We can also identify the heterogeneous technology clubs that may vary over time and across cross-section units; the frontier cluster formed by technology leading countries and the other group substantially below the frontier. We follow Schmidt and Sickles (1984), Kumbhakar (1990) and attempt to measure individual inefficiency:

$$e_{it} = \max_{i}\left(u_{it}\right) - \left(u_{it}\right) = \max_{i}\left(\alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t\right) - \left(\alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t\right) \tag{146}$$

We discuss how to estimate the proposed model (142-144) by rewriting the models as follows:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{147}$$

$$\varepsilon_{it} = v_{it} - u_{it}, \tag{148}$$

$$u_{it} = \alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t, \tag{149}$$

$$\tilde{u}_{it}(r) = \frac{1}{m_{it}} \sum_{j=1}^{N} I\left(\left|u_{t-1}^{*} - u_{jt-1}\right| \leq r\right) u_{jt-1}, \tag{150}$$

where α_i is (unobserved) individual-specific effect, $\mathbf{f}_t$ is an $r \times 1$ vector of unobserved factors and $\boldsymbol{\lambda}_i$ is an $r \times 1$ vector of the heterogeneous loading, $\tilde{u}_{it}(r)$ represents a cluster effect which is equal to the average efficiency of countries which are close to the frontier where $u_{t-1}^{*} = \min_j\left(u_{jt-1}\right)$, and v_{it} is an idiosyncratic disturbance. The distinguishing feature of our model is the use of unit-specific aggregates, which summaries of past values of efficiency, and connects the units that are close to the technology frontier (the best units).

To obtain consistent estimate of inefficiency in (146), we first estimate $\hat{\boldsymbol{\beta}}$ in (147) by PCCE or IPC, and derive $\hat{\varepsilon}_{it} = y_{it} - \mathbf{x}_{it}\hat{\boldsymbol{\beta}}$ with $\hat{v}_{it} = v_{it} - \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\mathbf{x}_{it} = v_{it} + o_p\left(1\right)$. Then, by normalizing with respect to the maximum, we get a first proxy of inefficiency as $\hat{e}_{it} = \max_i\left(\hat{\varepsilon}_{it}\right) - \left(\hat{\varepsilon}_{it}\right)$. Next, we consider the threshold estimation procedure, where a grid of values for r is constructed. For all values on that grid the model is estimated by least squares to obtain estimates of ρ. Specifically, we estimate $\hat{r}$ and $\hat{\rho}$ jointly by minimising the following criterion

function:

$$\mathbf{V}(r,\rho)=\min_{r,\rho}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{e}_{it}-\rho\frac{1}{m_{it}}\sum_{j=1}^{N}I\left(\left|\hat{e}_{t-1}^{*}-\hat{e}_{jt-1}\right|\leq r\right)\hat{e}_{jt-1}\right)^{2}. \quad (151)$$

The time-varying individual technical inefficiencies can be consistently estimated by

$$\hat{e}_{it}=\max_{i}\left(\hat{u}_{it}\right)-\left(\hat{u}_{it}\right)=\max_{i}\left(\hat{\alpha}_{i}+\hat{\rho}\tilde{u}_{it}(\hat{r})+\hat{\boldsymbol{\lambda}}_{i}^{\prime}\mathbf{f}_{t}\right)-\left(\hat{\alpha}_{i}+\hat{\rho}\tilde{u}_{it}(\hat{r})+\hat{\boldsymbol{\lambda}}_{i}^{\prime}\mathbf{f}_{t}\right) \quad (152)$$

Finally, we will convert $\hat{e}_{it}$ to the time-varying individual technical efficiency by

$$\hat{\tau}_{it}=\exp(-\hat{e}_{it}). \quad (153)$$

For empirical implementations, we follow Bailey *et al.* (2016) who propose a multi-step procedure to deal with both strong and weak forms of CSD as follows:

1. Test for the existence of CSD by applying the Pesaran (2015) CD test;
2. If the null of CSD is rejected, we apply the factor-based model to control for strong CSD.
3. We apply the Pesaran (2015) CD test again to the (de-factored) residuals.
4. If the null of no CSD is rejected, we also apply spatial or network modelling to the residuals (see (149)).

This extended KMS approach enables us to deal with both strong and weak forms of CSD by combining the (exogenously driven) factor-based approach with an endogenous threshold efficiency regime selection mechanism.

7 Nonlinear Regime Switching Models

In practice, we observe many stylised facts about economic time series as:

1. Business cycles are asymmetric in nature, e.g. Burns and Mitchell (1946); namely recessions last longer than expansion.
2. Asset pricing model under noise trading and transaction costs arising from the bid-ask spread: The larger are the pricing errors, the larger is the expected degree of arbitrage and hence the speedier is the price response to disequilibrium and *vice versa*.
3. Asymmetries are intrinsic to microeconomic behavior. For instance, costs of hiring and firing are asymmetric at the firm level.
4. Asymmetries can result from capital constraints on the goods market.

5. Imperfect competition and\or government interventions cause rigidities on credit, goods and labour markets that affect the dynamics of the economy.

However, it is increasingly recognised that the implications of linear modes are problematic in dealing with the above observations reflected in various economics and finance applications. In particular, the followings are questionable:

- Linearity, invariance of dynamic multipliers with respect to the size and the sign of the shock and the history of the system
- Time invariance of the parameters

Consequently, a great deal of interest has recently been made in modelling nonlinearities and asymmetries in economic time series.

Most attention has fallen almost exclusively on regime-switching type models, though there is no established theory suggesting a unique approach for specifying econometric models that embed various types of change in regimes.

- Regime shifts are not considered as singular deterministic events but the unobservable regime is assumed to be governed by an exogenous or predetermined stochastic processes. Thus regime shifts of the past are expected to occur in the future in a similar fashion.
- Regime switching models characterise a nonlinear data generating process as piecewise linear by restricting the process to be linear in each regime.
- The models differ in their assumptions concerning the stochastic process generating the regime; TAR, STAR, MS-AR, etc.

7.1 Structural break models

Suppose that the structural break occurs at time $t = \tau$ and we have

$$y_t = \left\{ \begin{array}{ll} \alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i} + \varepsilon_t & \text{for } t < \tau \\ \alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i} + \varepsilon_t & \text{for } t \geq \tau \end{array} \right\}, \tag{154}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. Then, (154) can be written as

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i} \right) \left(1 - I\left(t; \tau\right)\right) + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i} \right) I\left(t; \tau\right) + \varepsilon_t,$$

where $I\left(t; \tau\right)$ is the indicator function given by

$$I\left(t; \tau\right) = \left\{ \begin{array}{ll} 0 & \text{for } t < \tau \\ 1 & \text{for } t \geq \tau \end{array} \right\}.$$

Two different assumptions have been made:

The break point τ is known So break is deterministic. To estimate (154), split the sample and apply OLS to each regime. Tests of $\beta_{1i} = \beta_{2i}$, $i = 1, ..., p$, will follow the standard χ^2 distribution asymptotically.

See for example Perron (1989) for unit root tests subject to structural breaks.

The break point τ is unknown So break is stochastic and τ needs to be estimated as follow:

$$\begin{aligned} \tau^* &= \arg \min_{\tau \in [0.15T, 0.85T]} RSS(\tau) \\ &= \arg \min_{\tau \in [0.15T, 0.85T]} \left[\tau \hat{\sigma}_1^2(\tau) + (1-\tau) \hat{\sigma}_2^2(\tau) \right], \end{aligned}$$

where RSS stands for residual sum of squares, and the grid search is over $\tau \in [0.15T, 0.85T]$ in practice.

Notice that tests of $\beta_{1i} = \beta_{2i}$, $i = 1, ..., p$, does not follow the standard χ^2 distribution asymptotically, but has a nonstandard asymptotic distribution due to the Davies (1987) problem that nuisance parameter (break point τ here) is not identified under the null. Most solutions to this problem involve integrating out unidentified parameters from the test statistics. This is usually achieved by calculating test statistics over a grid set of possible values of nuisance parameter and then constructing the summary statistics such as sup (maximum) and exponential average, see Andrews and Plobegrer (1994).

7.2 Threshold models

This is a popular class of nonlinear regime-switching models with each regime determined by observed variables.

Threshold Autoregressive (TAR) model Now the regime shifts are triggered by an observable, exogenous transition variable x_t crossing threshold c, and consider the two-regime TAR model:

$$y_t = \left\{ \begin{array}{ll} \alpha_1 + \sum_{i=1}^p \beta_{1i} y_{t-i} + \varepsilon_t & \text{for } x_t \leq c \text{ (regime 1)} \\ \alpha_2 + \sum_{i=1}^p \beta_{2i} y_{t-i} + \varepsilon_t & \text{for } x_t > c \text{ (regime 2)} \end{array} \right\}, \tag{155}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. Alternatively,

$$y_t = \left(\alpha_1 + \sum_{i=1}^p \beta_{1i} y_{t-i} \right) \mathbf{1}\{x_t \leq c\} + \left(\alpha_2 + \sum_{i=1}^p \beta_{2i} y_{t-i} \right) (1 - \mathbf{1}\{x_t \leq c\}) + \varepsilon_t, \tag{156}$$

where $\mathbf{1}\{x_t \leq c\}$ is the indicator function.

Self-Exciting Threshold Autoregressive (SETAR) model If we use as the transition variable a lagged endogenous variable y_{t-d} with delay $d \geq 1$, we

obtain the two-regime SETAR model as follow:

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \mathbf{1}\left\{y_{t-d} \leq c\right\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) \left(1 - \mathbf{1}\left\{y_{t-d} \leq c\right\}\right) + \varepsilon_t, \tag{157}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$.

Notice that (157) can be written alternatively as

$$y_t = \alpha\left(s_t\right) + \sum_{i=1}^{p} \beta_i\left(s_t\right) y_{t-i} + \varepsilon_t, \tag{158}$$

where the probability of the unobservable regime 1 is given by

$$\Pr\left(s_t = 1 | S_{t-1}, Y_{t-1}\right) = \mathbf{1}\left\{y_{t-d} \leq c\right\},$$

where $S_{t-1} = \{s_{t-1}, s_{t-2}, ...\}$ and $Y_{t-1} = \{y_{t-1}, y_{t-2}, ...y_{t-p}\}$. This shows that SETAR and MS-AR models can be observationally equivalent, see Carrasco (1994).

Estimation ML estimation under normality can be carried over the grid search over d and c: select the pair(d, c) that minimises the residual sum of squares:

$$\arg\min_{(d,c)} \sum_{m=1}^{M} T_m \hat{\sigma}_m^2,$$

where T_m and $\hat{\sigma}_m^2$ are the number of observations and the residual variance in regime m. Usually the grid search is restricted such that $\min T_m \geq 0.15T$.

Three-regime SETAR model We now extend to consider the three-regime SETAR model:

$$\begin{aligned} y_t &= \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \mathbf{1}\left\{y_{t-d} \leq c_1\right\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) \mathbf{1}\left\{c_1 < y_{t-d} \leq c_2\right\} \\ &\quad + \left(\alpha_3 + \sum_{i=1}^{p} \beta_{3i} y_{t-i}\right) \mathbf{1}\left\{c_2 < y_{t-d}\right\} + \varepsilon_t, \end{aligned} \tag{159}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$, and c_1 and c_2 are threshold parameters and $c_1 < c_2$.

Example 1 *Trade Cost Model by Sercu, Uppal and van Hulle (1995, Journal of Finance).*

7.3 Smooth Transition Autoregressive (STAR) Models

If our aim is to distinguish between the effects of negative and positive deviations (or large and small) from the equilibrium, then TAR models are appropriate.

Recently, STAR models have attracted more attention in finance. The basic motivation behind it is that prices are expected to adjust more smoothly as is

predicted by TAT models. One explanation is: nonlinear asymmetric behavior of heterogeneous market participants will be smoother at the aggregate level.

Granger and Terasvirta (1993) advance the following STAR model:

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \{1 - F(z_t; \theta, c)\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) F(z_t; \theta, c) + \varepsilon_t, \tag{160}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. The transition function $F(z_t; \theta, c)$ is a continuous function determining the weights of regime and usually bounded between 0 and 1. c and θ are the threshold and smoothness parameters.

The transition variable z_t can be:

- a lagged endogenous variable ($z_t = y_{t-d}$),
- an exogenous variable ($z_t = x_t$),
- or a function such as ($z_t = g(y_{t-d}, x_t)$).
- For $z_t = t$, we obtain a model with smoothly changing parameters, see Lin and Terasvirta (1994).

The STAR model exhibits:

- two regimes associated with the extreme values of the transition function: $F(z_t; \theta, c) = 1$ and $F(z_t; \theta, c) = 0$;
- transition from one regime to the other is gradual, not abrupt as in TAR;
- the regime occurring at time t is observable and determined by $F(z_t; \theta, c)$.

Logistic Smooth Transition Autoregressive (LSTAR) model We consider as the transition function in (160) the logistic CDF:

$$F(z_t; \theta, c) = \frac{1}{1 + \exp\{-\theta(z_t - c)\}}. \tag{161}$$

This model can deal with asymmetric behavior for positive vs negative values of z_t relative to c. We note:

- As $\theta \to \infty$, LSTAR $\to$ TAR, since $F(z_t; \theta, c) = I(z_t > c)$.
- As $\theta \to 0$, LSTAR $\to$ linear AR, since $F(z_t; \theta, c) = 1/2$.

The second order logistic CDF is also considered:

$$F(z_t; \theta, c) = \frac{1}{1 + \exp\{-\theta(z_t - c_1)(z_t - c_2)\}}. \tag{162}$$

We note:

- As $\theta \to \infty$, L2STAR $\to$ 3 regime TAR, since $F(z_t; \theta, c) = 1 - I(c_1 < z_t < c_2)$.
- As $\theta \to 0$, L2STAR $\to$ linear AR, since $F(z_t; \theta, c) = 1/2$.

Exponential Smooth Transition Autoregressive (ESTAR) model We consider as the transition function in (160) the exponential function:

$$F(z_t;\theta,c) = 1 - \exp\left\{-\theta(z_t - c)^2\right\}, \tag{163}$$

where we assume that $\theta \geq 0$ for identification. This model can deal with asymmetric behavior for small vs large deviations of z_t from the threshold c.

The exponential transition function is bounded between zero and 1, *i.e.* $F : \mathbb{R} \to [0,1]$ has the properties:

$$F(0) = 0;\ \lim_{x \to \pm\infty} F(x) = 1,$$

and is symmetrically U-shaped around zero.

As $\theta \to \infty$ and $\theta \to 0$, ESTAR $\to$ linear AR, since $F(z_t;\theta,c) = 1$ and $F(z_t;\theta,c) = 0$, respectively.

Estimation Nonlinear least squares or ML estimation method via numerical optimisation procedure can be applied, but it also involves the grid search over (d,c) as in TAR models.

However, the precise estimation of θ is somewhat difficult in practice.

- For large value of θ, the shape of the logistic function changes only little.
- Accurate estimation of θ requires many observations in the immediate neighborhood of c.
- Insignificance of θ should not be interpreted as evidence against the presence of STAR nonlinearity, see Bates and Watts (1988).

7.4 Markov-Switching Autoregressive (MS-AR) Models

Now the regime s_t is generated by a hidden discrete-state homogeneous and ergodic Markov chain:

$$\Pr(s_t|S_{t-1},Y_{t-1}) = \Pr(s_t|S_{t-1};\rho)$$

defined by the transition probabilities,

$$p_{ij} = \Pr(s_{t+1} = j|s_t = j),$$

where $S_{t-1} = \{s_{t-1}, s_{t-2}, ...\}$, $Y_{t-1} = \{y_{t-1}, y_{t-2}, ...y_{t-p}\}$ and ρ are unknown parameters.

The conditional process is a AR(p) model with

- shift in mean (MSM-AR): once-and-for-all jump in time series:

$$y_t - \mu(s_t) = \sum_{i=1}^{p} \beta_i(s_t)(y_{t-i} - \mu(s_{t-i})) + \varepsilon_t, \tag{164}$$

- shift in intercept (MSI-AR): smooth adjustment of time series:

$$y_t = \alpha(s_t) + \sum_{i=1}^{p} \beta_i(s_t)\, y_{t-i} + \varepsilon_t. \tag{165}$$

Example 2 *MS-AR models of US GNP. Hamilton (1989) consider the 2-regime MS-AR model for the quarterly growth rate of US GNP:*

$$\Delta y_t - \mu(s_t) = \sum_{i=1}^{4} \beta_i(s_t)\left(\Delta y_{t-i} - \mu(s_{t-i})\right) + \varepsilon_t, \tag{166}$$

$$\varepsilon_t | s_t \sim iidN\left(0, \sigma^2\right).$$

Two regimes are defined by

$$\mu(s_t) = \left\{ \begin{array}{ll} \mu_1 > 0 & \text{if } s_t = 1 \text{ (expansion)} \\ \mu_2 < 0 & \text{if } s_t = 2 \text{ (contraction)} \end{array} \right\},$$

which are generated by an ergodic Markov chain

$$p_{12} = \Pr(\text{contraction in } t | \text{expansion in } t-1)$$

$$p_{21} = \Pr(\text{expansion in } t | \text{contraction in } t-1)$$

- The statistical analysis of MS-AR models is based on the state-space form. Then, general concepts such as the likelihood principle and a recursive filtering algorithm can be used.
- In contrast to TAR and STAR models MS-AR models include the possibility that the threshold depends on the last regime, i.e., the threshold staying in regime 2 is different from the threshold for switching from regime 1 to regime 2.

7.5 Linearity Tests for TAR/STAR Specification

Here the null model is that

$$H_0 : y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t, \tag{167}$$

so the model is linear, whilst the alternative models are either

$$H_{1,TAR} : \text{ TAR model given by (156)}, \tag{168}$$

or

$$H_{1,STAR} : \text{ TAR model given by (160)}. \tag{169}$$

More specifically, against the TAR model we have

$$H_0 : \alpha_1 = \alpha_2 \text{ and } \beta_{1i} = \beta_{2i} \text{ for all } i = 1, ..., p, \tag{170}$$

whilst against the STAR model we have

$$H_0 : \theta = 0. \tag{171}$$

However, due to the Davies (1987) problem that nuisance parameters in transition function - namely, threshold parameter c in TAR and smoothness parameter θ and threshold parameter in STAR - are not identified under the null, we could not use the standard asymptotic χ^2 distribution.

1. **Sup test approach for TAR models**: We should obtain a supremum of a number of dependent test statistics over the grid over c: $\sup F$, $\sup Wald$, $\sup LR$ and $\sup LM$ tests with nonstandard limiting distribution. To obtain the p-value, we need to run the bootstrapping simulations. See Hansen (1997,2000).

2. **Taylor approximation approach for STAR models**: Approximate smooth transition function with a first-order expansion around $\theta = 0$. Then, using the derived auxiliary regression, we obtain the LM-type tests with a standard χ^2 limiting distribution. See Luukkonen, Saikkonen and Terasvirta (1988) for LSTAR and Saikkonen and Luukkonen (1988) for ESTAR.

Remark 3 *So far the above tests have been developed under the assumption of stationarity, linear AR model with a unit root* ($\rho = 1$) *being excluded.*

7.6 Nonlinear Unit Root Tests

Balke and Fomby (1997) have popularised a joint analysis of nonstationarity and nonlinearity in the context of threshold cointegration. The threshold cointegrating process is defined as a globally stationary process such that it might follow a unit root in the middle regime, but it is dampened in outer regimes. Importantly, they have shown via Monte Carlo experiments that the power of the DF unit root tests falls dramatically with threshold parameters. See also Pippenger and Goering (1993).

As a response to these problems, there is a growing literature proposing tests for unit roots against threshold autoregressive (TAR) alternatives, *e.g.* Enders and Granger (1998), Caner and Hansen (2001), Kapetanios, Shin and Snell (2003), Bec, Guay and Guerre (2004) and Kapetanios and Shin (2006).

7.6.1 Unit Root Tests in Two-regime TAR Framework

Enders and Granger (1998) have addressed this issue using a two-regime TAR model with implicitly known threshold value,

$$\Delta y_t = \left\{ \begin{array}{ll} \beta_1 y_{t-1} + u_t & if\ y_{t-1} \leq 0 \\ \beta_2 y_{t-1} + u_t & if\ y_{t-1} > 0 \end{array} \right\},\ t = 1, 2, ..., T, \tag{172}$$

and suggested an F-statistic for $\beta_1 = \beta_2 = 0$ in (172).

Despite the main aim to derive a more powerful test, their simulation evidence shows that the proposed F test is less powerful than the DF test that ignores the threshold nature of this two regime alternative. But they also provided simulation results showing that the F-test may have higher power than the DF test against the three regime asymmetric TAR models. See also Berben and van Dijk (1999).

There has also been an alternative line of studies. Caner and Hansen (2001) have considered the following two-regime TAR model:

$$\Delta y_t = \boldsymbol{\theta}_1' \mathbf{x}_{t-1} 1_{\{\Delta y_{t-1} \leq r\}} + \boldsymbol{\theta}_2' \mathbf{x}_{t-1} 1_{\{\Delta y_{t-1} > r\}} + e_t,\ t = 1, 2, ..., T, \tag{173}$$

where $\mathbf{x}_{t-1} = (y_{t-1}, 1, \Delta y_{t-1}, ..., \Delta y_{t-k})'$, r is an unknown threshold parameter, and e_t is an iid error. They have first developed tests for threshold nonlinearity when y_t follows a unit root, and then unit root tests when the threshold nonlinearity is either present or absent. Limitation of this approach is that these tests rely on the stationarity of the transition variable.

7.6.2 Unit Root Tests in Three-regime TAR Framework

Kapetanios and Shin (2006) Suppose that a univariate series y_t follows the three-regime self-exciting threshold autoregressive (SETAR) model:

$$y_t = \left\{ \begin{array}{cc} \phi_1 y_{t-1} + u_t & if\ y_{t-1} \leq r_1 \\ \phi_0 y_{t-1} + u_t & if\ r_1 < y_{t-1} \leq r_2 \\ \phi_2 y_{t-1} + u_t & if\ y_{t-1} > r_2 \end{array} \right\},\ t = 1, 2, ..., T, \tag{174}$$

where u_t is assumed to follow an iid sequence with zero mean, constant variance σ_u^2 and finite $4 + \delta$ moments for some $\delta > 0$, r_1 and r_2 are threshold parameters and $r_1 < r_2$. Here, the lagged dependent variable is used as the transition variable with the delay parameter set to 1 for simplicity. The intuitive appeal of the scheme in (174) is that it allows the speed of adjustment to vary asymmetrically with regimes. Suppose that

$$\phi_0 \geq 1,\ |\phi_1|,\ |\phi_2| < 1. \tag{175}$$

The series are then locally nonstationary, but globally ergodic.

Following the maintained assumption in the literature, we now impose $\phi_0 = 1$ in (174), which implies that y_t follows a random walk in the corridor regime. Then, defining $1_{\{.\}}$ as a binary indicator function, (174) can be compactly written as

$$\Delta y_t = \beta_1 y_{t-1} 1_{\{y_{t-1} \leq r_1\}} + \beta_2 y_{t-1} 1_{\{y_{t-1} > r_2\}} + u_t, \tag{176}$$

where $\beta_1 = \phi_1 - 1$, $\beta_2 = \phi_2 - 1$, and $y_{t-1} 1_{\{y_{t-1} \leq r_1\}}$ and $y_{t-1} 1_{\{y_{t-1} > r_2\}}$ are orthogonal to each other by construction.

We consider the (joint) null hypothesis of unit root as

$$H_0 : \beta_1 = \beta_2 = 0, \tag{177}$$

against the alternative hypothesis of threshold stationarity,

$$H_1 : \beta_1 < 0;\ \beta_2 < 0. \tag{178}$$

Then, the joint null hypothesis of linear unit root against the nonlinear threshold stationarity can be tested using the Wald statistic denoted by $\mathcal{W}_{(r_1,r_2)}$, which has a nonstandard limiting distribution.

In order to deal with the Davies problem that unknown threshold parameters r_1 and r_2 are not defined under the null, we consider the supremum, the average and the exponential average of the Wald statistic defined respectively by

$$\mathcal{W}_{\sup} = \sup_{i\in\Gamma} \mathcal{W}^{(i)}_{(r_1,r_2)},\ \mathcal{W}_{\text{avg}} = \frac{1}{\#\Gamma}\sum_{i=1}^{\#\Gamma} \mathcal{W}^{(i)}_{(r_1,r_2)},\ \mathcal{W}_{\exp} = \frac{1}{\#\Gamma}\sum_{i=1}^{\#\Gamma} exp\left(\frac{\mathcal{W}^{(i)}_{(r_1,r_2)}}{2}\right), \tag{179}$$

where $\mathcal{W}^{(i)}_{(r_1,r_2)}$ is the Wald statistic obtained from the i-th point of the threshold parameters grid set, Γ and $\#\Gamma$ is the number of elements of Γ.

Unlike the stationary TAR models, the selection of the grid of threshold parameters needs more attention. The threshold parameters r_1 and r_2 usually take on the values in the interval

$$(r_1, r_2) \in \Gamma = \{(r_{1,1}, r_{1,2}), ..., (r_{i,1}, r_{i,2}), ..., (r_{\#\Gamma,1}, r_{\#\Gamma,2})\},$$

where $r_{\min} \leq r_{i,1},\ i = 1, ...\#\Gamma$, and $r_{\max} \geq r_{i,2},\ i = 1, ...\#\Gamma$. $r_{\min}$ and $r_{\max}$ are picked so that $\Pr(y_{t-1} < r_{\min}) = \pi_1 > 0$ and $\Pr(y_{t-1} > r_{\max}) = \pi_2 < 1$. The particular choice for π_1 and π_2 is somewhat arbitrary, and in practice must be guided by the consideration that each regime needs to have sufficient observations to identify the underlying regression parameters.

However, since our approach assumes that the coefficient on the lagged dependent variable is set to zero in the corridor regime ($r_1 \leq y_{t-1} < r_2$), we can assign arbitrarily small samples (relative to total sample) to the corridor regime. Notice also that the threshold parameters exist only under the alternative hypothesis in which the process is stationary and therefore bounded in probability. This observation leads us to make an assumption that the grid for unknown threshold parameters should be selected such that the selected corridor regime be of finite width both under the null and under the alternative. Noticing that a random walk process will stay within a corridor regime of finite width for $O_p\left(\sqrt{T}\right)$ periods only, then setting

$$\pi_1 = \bar{\pi} - c/T^{\delta} \text{ and } \pi_2 = \bar{\pi} + c/T^{\delta},$$

where $\bar{\pi}$ is the sample quantile corresponding to zero and $\delta \geq 1/2$, guarantees that the grid set will be of finite width under the null hypothesis. In practice, c can be chosen so as to give a reasonable coverage of each regime in samples of sizes usually encountered. For example, for $T = 100$ and $\delta = 1/2$, c can be set to 3 to give a 60% coverage of the sample for the grid.

The small sample performance of our suggested tests is compared to that of the DF test via Monte Carlo experiments. We find that both average and exponential average tests have reasonably correct size, but the supremum test tends to display significant size distortions in small samples. As expected, both average and exponential average tests eventually dominate the power of the DF test as the threshold band widens.

KS illustrate the usefulness of our proposed tests by examining the stationarity of bilateral real exchange rates for the G7 countries (excluding France). In sum, our proposed (asymmetry) Wald tests reject the null three times out of five cases, while the DF test rejects the null only once.

Bec, Ben Salem and Carrasco (2004) and Bec, Guay and Gurre (2004)
A three-regime SETAR model (174) can be compactly written as

$$\Delta y_t = \beta_1 y_{t-1} 1_{\{y_{t-1} \leq r_1\}} + \beta_0 y_{t-1} 1_{\{r_1 < y_{t-1} < r_2\}} + \beta_2 y_{t-1} 1_{\{y_{t-1} > r_2\}} + u_t, \quad (180)$$

where $1_{\{.\}}$ is a binary indicator function, $\beta_1 = \phi_1 - 1$, $\beta_0 = \phi_0 - 1$, $\beta_2 = \phi_2 - 1$, and $y_{t-1} 1_{\{y_{t-1} \leq r_1\}}$, $y_{t-1} 1_{\{r_1 < y_{t-1} < r_2\}}$, $y_{t-1} 1_{\{y_{t-1} > r_2\}}$ are orthogonal to each other by construction. BBC and BGG have considered the three-regime SETAR model (180), and proposed the supremum-based Wald test procedure for the joint hypothesis of $\beta_1 = \beta_0 = \beta_2 = 0$ in (180).

BBC take the quantile-based approach, assuming that $r = \sqrt{T}\lambda$ where $|r_1| = |r_2| = r$ (symmetric outer regimes).[19] Then they derive the asymptotic distribution of the Wald statistic, denoted by $\mathcal{W}^{BBC}(r)$, for $\beta_1 = \beta_0 = 0$ in (180) after imposing $\beta_1 = \beta_2$, which depends on the nuisance parameter, $\lambda/\hat{\sigma}_{LR}$, where $\hat{\sigma}_{LR}$ is the long-run variance of Δy_t obtained under the null. To avoid the Davies problem, they suggest to use the supremum-based tests defined by

$$\mathcal{W}^{BBC}_{\sup} = \sup_{r \in [r_{\min}, r_{\max}]} \mathcal{W}^{BBC}_i(r). \quad (181)$$

On the other hand, BGG develop an adaptive consistent unit root tests based on the symmetric three regime TAR model (180) with $\beta_1 = \beta_2$ and propose an adaptive choice of the grid set which restricts the grid to remain bounded under the null but to become unbounded under the alternative. They suggest to use the following grid set:

$$r_{\min} = |y|_{(3)} + \frac{\hat{\sigma}_0}{\ell \max(1, t_{ADF})}; \; r_{\max} = |y|_{(3)} + \ell \hat{\sigma}_0 \max(1, t_{ADF}), \quad (182)$$

where $|y|_{(j)}$'s are the ordered variables of $|y_j|$, $j = 1, ..., T-1$, ℓ is a length parameter to be determined empirically, t_{ADF} is the ADF t-statistic, and $\hat{\sigma}_0^2 = \frac{1}{T-1} \sum_{t=1}^{T} \left(y_t - \hat{a} - \hat{\phi} y_{t-1} \right)^2$ with $\hat{a}$ and $\hat{\phi}$ being the OLS estimates. This adaptive choice of the grid set is aimed to boost the power of the tests. First, if the

[19] The assumption that $r = \sqrt{T}\delta$ guarantees that the probability being in the corridor regime is always positive. On the other if r is fixed, this probability becomes zero.

set $\Gamma = [r_{\min}, r_{\max}]$ is small under the null, then the associated critical values of the $\mathcal{W}_{\sup}^{BBC}$ test statistic will be small too. Second, it will make a larger class of alternatives including the linear stationary model. Using the simulation evidence BGG argue that the former provides the most important contribution of the improved power performance of the $\mathcal{W}_{\sup}^{BBC}$ test when the grid set is selected by (182).

Example 4 *See Dutta and Leon (2002) for three regime model in exchange rate dynamics.*

7.6.3 Unit Root Tests in ESTAR Framework (Kapetanios, Shin and Snell, 2003)

Consider a univariate smooth transition autoregressive of order 1, ESTAR(1) model,

$$y_t = \beta y_{t-1} + \gamma y_{t-1}\left[1 - \exp\left(-\theta y_{t-d}^2\right)\right] + \varepsilon_t, \tag{183}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$, β and γ are unknown parameters, and we assume that $\theta \geq 0$, and $d \geq 1$ is the delay parameter. (183) can be conveniently reparameterised as:

$$\Delta y_t = \phi y_{t-1} + \gamma y_{t-1}\left[1 - \exp\left(-\theta y_{t-d}^2\right)\right] + \varepsilon_t, \tag{184}$$

where $\phi = \beta - 1$. If θ is positive, then it effectively determines the speed of mean reversion. The representation (184) makes economic sense in that many economic models predict that the underlying system tends to display a dampened behavior towards an attractor when it is (sufficiently far) away from it, but that it shows some instability within the locality of that attractor.

We prove under $\theta > 0$ that the condition we need for geometric ergodicity of the model (183) or (184) is in fact $|\beta + \gamma| < 1$ or $|\phi + \gamma| < 0$.

Remark 5 *The application that motivates our model is that of Sercu et al. (1995) and of Michael et al. (1997). These authors analyse nonlinearities in the PPP relationship. They adopt a null of a unit root for real exchange rates and have an alternative hypothesis of stationarity i.e. the long run PPP. Their theory suggests that the larger the deviation from PPP, the stronger the tendency to move back to equilibrium. In the context of our model, this would imply that while $\phi \geq 0$ is possible, we must have $\gamma < 0$ and $\phi + \gamma < 0$ for the process to be globally stationary. Under these conditions, the process might display unit root or explosive behaviour in the middle regime for small y_{t-d}^2, but for large y_{t-d}^2, it has stable dynamics and as a result is geometrically ergodic. They claim that the ADF test may lack power against such stationary alternatives and one of the contributions of this paper is to provide an alternative test designed to have a power against such an ESTAR processes.*

Imposing $\phi = 0$ and $d = 1$ gives our specific ESTAR model (184) as

$$\Delta y_t = \gamma y_{t-1}\left\{1 - \exp\left(-\theta y_{t-1}^2\right)\right\} + \varepsilon_t. \tag{185}$$

Our test directly focuses on a specific parameter, θ, which is zero under the null and positive under the alternative. Hence we test

$$H_0 : \theta = 0, \tag{186}$$

against the alternative

$$H_1 : \theta > 0. \tag{187}$$

Obviously, testing the null hypothesis (198) directly is not feasible, since γ is not identified under the null.

To overcome this problem we follow Luukkonen *et al.* (1988), and derive a t-type test statistic. If we compute a first-order Talyor series approximation to the $ESTAR$ model under the null we get the auxiliary regression

$$\Delta y_t = \delta y_{t-1}^3 + error. \tag{188}$$

This suggests that we could obtain the t-statistic for $\delta = 0$ against $\delta < 0$ as

$$t_{NL} = \widehat{\delta} / s.e. \left(\widehat{\delta} \right), \tag{189}$$

where $\widehat{\delta}$ is the OLS estimate of δ and $s.e. \left(\widehat{\delta} \right)$ is the standard error of $\widehat{\delta}$. Our test is motivated by the fact that the auxiliary regression is testing the significance of the score vector from the quasi-likelihood function of the $ESTAR$ model, evaluated at $\theta = 0$.

Unlike the case of testing linearity against nonlinearity for the stationary process, the t_{NL} test does not have an asymptotic standard normal distribution.

KSS find *inter alia* that under the alternative of a globally stationary ESTAR process, our test has better power in cases where the nonlinear adjustment is relatively important.

KSS also provide an application to *ex post* real interest rates and bilateral real exchange rates from eleven major OECD countries, and in particular find that our proposed test is able to reject a unit root in some cases where the linear ADF tests fails to do so, providing a limited evidence of of nonlinear mean-reversion in both real interest and exchange rates.

7.7 Nonlinear Error Correction Models

Clearly, many stylised facts can be evoked to account for the asymmetric adjusting behavior. For example, in financial markets prices are constrained to persistent short-run disequilibria due to information barriers, transaction costs, noise trading, market segmentation, etc.

- A first strand of the literature is based on a generalisation of the usual concept of cointegration. Notions such as 'attractors', 'transients', 'Lyapunov stability', 'equilibration' have been introduced in an attempt to capture richer dynamics than is allowed by linear cointegration models.

- Another approach aims to clarify the concept of cointegration. If the two processes have the same order of integration, they may be cointegrated if their combination (either linear or nonlinear) is mixing. But, one must impose the bound conditions on the nonlinear functions.

- A third part of the literature is centred on nonlinear co-trending. Nonlinear trends are modelled as general polynomial functions that allows multiple representations of nonlinear trends. Co-trending means that the combination of nonlinear trends provides linear trends.

We assume that the attractor is linear but that the adjustment towards the long-run equilibrium is nonlinear. The NEC model is written as

$$\Delta y_t = \sum_{i=1}^{p} \delta_i' \Delta y_{t-i} + \sum_{i=1}^{q} \gamma_i' \Delta x_{t-i} + \lambda z_{t-1} + f\left(z_{t-1}, \theta\right) + u_t, \tag{190}$$

$$\Delta y_t = v_t,$$
$$z_t = y_t - \beta' x_t.$$

Assume that (i) u_t and v_t are mixing processes with finite second-order moments and cross moments; (ii) f is a nonlinear function that is continuously differentiable and satisfies some regularity condition:

$$-1 < \frac{\partial f\left(z_{t-1}, \theta\right)}{\partial z_{t-1}} < 1;$$

(iii) the roots of $\left|1 - \sum_{i=1}^{p} \delta_i L^i\right| = 0$ all lie outside the unit circle; and (iv) u_t is a martingale difference sequence with zero mean and constant variance.

Under this assumption Escribano and Mira (2002) prove that z_t is NED and y_t and x_t are cointegrated. The cointegration hypothesis is tested as follows:

$$H_0 : f\left(z_{t-1}, \theta\right) = 0 \text{ against } H_1 : f\left(z_{t-1}, \theta\right) \neq 0.$$

H_0 means that the adjustment mechanism is linear. Under H_1: it is not sufficient that $f\left(z_{t-1}, \theta\right) \neq 0$, but this function must characterize an EC mechanism (hence the importance of the stability condition of f).

Estimation of (190) can be done in 4-steps:

1. Obtain the OLS estimate of $\boldsymbol{\beta}$ from the regression of y_t on $\mathbf{x}_t$. Construct the estimate of error correction term by $\hat{z}_t = y_t - \hat{\boldsymbol{\beta}}' \mathbf{x}_t$.

2. Substitute $\hat{z}_{t-1}$ for z_{t-1} in $f\left(z_{t-1}, \theta\right)$.

3. Use the NLS method to find an estimate of $\boldsymbol{\theta}$.

4. Estimate other coefficients of the model (190) by OLS.

In practice the great difficulty lies in finding an appropriate function that satisfies the stability condition defined in Assumption (ii). The following functional forms are employed in Dufrénot and Mignon (2002):

- Logistic Smooth Transition Regression:

$$LF\left(z_{t-1}\right)=\left[1+\exp\left(-\theta\left(z_{t-1}-c\right)\right)\right]$$

- Cubic Polynomial:

$$LF\left(z_{t-1}\right)=\delta_1 z_{t-1}+\delta_2 z_{t-1}^2+\delta_3^3 z_{t-1}$$

- Rational Polynomial:

$$LF\left(z_{t-1}\right)=\frac{\left(z_{t-1}+\gamma_1\right)^3+\gamma_2}{\left(z_{t-1}+\gamma_3\right)^3+\gamma_4}.$$

Example 6 *Rational or irrational bubbles? Many empirical studies find that there is no cointegration between stock prices and dividends. This may imply that the fluctuations in asset prices are too large to reflect changes occurring in the fundamentals (here dividends). This excess volatility can be regarded as a consequence of the presence of a (possibly) nonstationary bubble. This paves the way for a nonlinear dynamic analysis as to how to add to the usual arbitrage equation a nonlinear component reflecting the complexity of the short term dynamics between both variables.*

7.7.1 Asymmetric TAR NEC Models

See Balke and Fomby (1997).

7.7.2 Asymmetric STR NEC Models

We begin with the following general nonlinear vector error correction model for the $n\times 1$ vector of I(1) stochastic processes, $\mathbf{z}_t$:

$$\Delta\mathbf{z}_t=\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{z}_{t-1}+g\left(\boldsymbol{\beta}'\mathbf{z}_{t-1}\right)+\sum_{i=1}^{p}\boldsymbol{\Gamma}_i\Delta\mathbf{z}_{t-i}+\boldsymbol{\varepsilon}_t,\ t=1,2,...,T, \tag{191}$$

where $\boldsymbol{\alpha}$ $(n\times r)$, $\boldsymbol{\beta}$ $(n\times r)$ and $\boldsymbol{\Gamma}_i$ $(n\times n)$ are parameter matrices with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of full column rank and $g:\mathbb{R}^r\rightarrow\mathbb{R}^n$ is a nonlinear function. See Saikkonen (2004).

We aim to analyse at most one conditional long-run cointegrating relationship between y_t and $\mathbf{x}_t$, and focus on the conditional modelling of the scalar variable y_t given the k-vector $\mathbf{x}_t$ $(k=n-1)$ and the past values of $\mathbf{z}_t$ and $\mathbf{Z}_0$, where we decompose $\mathbf{z}_t=\left(y_t,\mathbf{x}_t'\right)'$. For this we rewrite (191) as

$$\Delta\mathbf{z}_t=\boldsymbol{\alpha}u_{t-1}+g\left(u_{t-1}\right)+\sum_{i=1}^{p}\boldsymbol{\Gamma}_i\Delta\mathbf{z}_{t-i}+\boldsymbol{\varepsilon}_t,\ t=1,2,...,T, \tag{192}$$

where $\boldsymbol{\alpha}$ is an $n\times 1$ vector of adjustment parameters, and

$$u_t=y_t-\boldsymbol{\beta}_x'\mathbf{x}_t, \tag{193}$$

with $\boldsymbol{\beta}_x$ being a $k \times 1$ vector of cointegrating parameters.

We now make the following assumption:

- 2**(i)** Partition $\boldsymbol{\alpha} = (\phi, \boldsymbol{\alpha}_x')'$ and $\boldsymbol{\varphi} = (\gamma, \boldsymbol{\varphi}_x')'$ conformably with $\mathbf{z}_t = (y_t, \mathbf{x}_t')'$. Then, $\boldsymbol{\alpha}_x = \boldsymbol{\varphi}_x = \mathbf{0}$.
- 2**(ii)** There is no cointegration among the k-vector of I(1) variables, $\mathbf{x}_t$.
- 2**(iii)** $g(\bullet)$ follows the exponential smooth transition regressive (ESTR) functional form,[1]

$$g(u_{t-1}) = \boldsymbol{\varphi} u_{t-1} \left(1 - e^{-\theta(u_{t-1}-c)^2}\right), \tag{194}$$

  where we assume $\theta \geq 0$ for identification purpose and c is a transition parameter.

Assumption 2(i) and (ii) imply that the process $\mathbf{x}_t$ are weakly exogenous and therefore the parameters of interest in (196) are variation-free from the parameters in (197), see Pesaran *et al.* (2001).

Next, partitioning $\boldsymbol{\varepsilon}_t$ conformably with $\mathbf{z}_t$ as $\boldsymbol{\varepsilon}_t = (\varepsilon_{yt}, \boldsymbol{\varepsilon}_{xt}')'$ and its variance matrix as $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{yy} & \boldsymbol{\sigma}_{yx} \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}$, we may express ε_{yt} conditionally in terms of $\boldsymbol{\varepsilon}_{xt}$ as

$$\varepsilon_{yt} = \boldsymbol{\sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\varepsilon}_{xt} + e_t, \tag{195}$$

where $e_t \sim iid(0, \sigma_e^2)$, $\sigma_e^2 \equiv \sigma_{yy} - \boldsymbol{\sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$ and e_t is uncorrelated with $\boldsymbol{\varepsilon}_{xt}$ by construction. Substituting (195) and (194) into (192), partitioning $\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{yi}', \boldsymbol{\Gamma}_{xi}')'$, $i = 1, ..., p$, and under Assumption 2, we obtain the following conditional nonlinear error correction model for Δy_t and the marginal VAR model for $\boldsymbol{\Delta x_t}$:

$$\Delta y_t = \phi u_{t-1} + \gamma u_{t-1} \left(1 - e^{-\theta(u_{t-1}-c)^2}\right) + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i' \Delta \mathbf{z}_{t-i} + e_t, \tag{196}$$

$$\Delta \mathbf{x}_t = \sum_{i=1}^{p} \boldsymbol{\Gamma}_{xi} \Delta \mathbf{z}_{t-i} + \boldsymbol{\varepsilon}_{xt}, \tag{197}$$

where $\boldsymbol{\omega} \equiv \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$ and $\boldsymbol{\psi}_i' \equiv \boldsymbol{\gamma}_{yi} - \boldsymbol{\omega}' \boldsymbol{\Gamma}_{xi}$, $i = 1, ..., p$.

We call (196) the (conditional) nonlinear STR error correction model. The representation (196) makes economic sense in that many economic models predict that the underlying system tends to display a dampened behavior towards an attractor when it is (sufficiently far) away from it, but shows some instability within the locality of that attractor.

7.7.3 Testing for Cointegration under STR ECM

To fix ideas for the motivation of the tests, we follow Kapetanios, Shin and Snell (2003, hereafter KSS) and impose $\phi = 0$ in (196), implying that u_t follows a unit root process in the middle regime, see also Balke and Fomby (1997) in the context of threshold error correction models. Note that for the operational versions of the tests we suggest below we consider both the case $\phi = 0$ and $\phi \neq 0$. It is then straightforward to show that the test of the null of no cointegration against the alternative of globally stationary cointegration can be based on the null hypothesis of no cointegration as

$$H_0 : \theta = 0, \tag{198}$$

against the alternative of nonlinear ESTR cointegration of $H_1 : \theta > 0$, where the positive value of θ determines the stationarity properties of u_t.

We propose a number of operational versions of the cointegration test under the nonlinear STR-ECM framework given by (196). To this end we follow Engle and Granger (1987) and take a pragmatic residual-based two step approach. In the first stage, we obtain the residuals, $\hat{u}_t = y_t - \hat{\boldsymbol{\beta}}_x' \mathbf{x}_t$ with $\hat{\boldsymbol{\beta}}_x$ being the OLS estimate of $\boldsymbol{\beta}_x$. In the second stage and in order to overcome the Davies problem that γ in (196) is not identified under the null, we follow Luukkonen, Saikkonen and Teräsvirta (1988) and KSS and approximate (196) by a first-order Taylor series approximation to $\left(1 - e^{-\theta(u_{t-1}-c)^2}\right)$, while allowing $\phi \neq 0$ under the alternative hypothesis, to get

$$\Delta y_t = \delta_1 u_{t-1} + \delta_2 u_{t-1}^2 + \delta_3 u_{t-1}^3 + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i' \Delta \mathbf{z}_{t-i} + e_t. \tag{199}$$

For this model, we consider an F-type test for $\delta_1 = \delta_2 = \delta_3 = 0$ given by

$$F_{NEC} = \frac{(SSR_0 - SSR_1)/3}{SSR_0/(T - 4 - p)}, \tag{200}$$

where SSR_0 and SSR_1 are the sum of squared residuals obtained from the specification with and without imposing the restrictions $\delta_1 = \delta_2 = \delta_3 = 0$ in (199), respectively.

There are prior theoretical justifications for restricting the switch point, c to be zero in many economic and financial applications in the ESTR function (194), in which case we obtain the following restricted auxiliary testing regression:

$$\Delta y_t = \delta_1 \hat{u}_{t-1} + \delta_2 \hat{u}_{t-1}^3 + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i' \Delta \mathbf{z}_{t-i} + e_t, \tag{201}$$

and obtain the following F-type statistic:

$$F_{NEC}^{*} = \frac{(SSR_0 - SSR_1)/2}{SSR_0/(T - 3 - p)}, \tag{202}$$

where SSR_0 and SSR_1 are the sum of squared residuals obtained from the specification with and without imposing $\delta_1 = \delta_2 = 0$ in (201), respectively.

Finally, under the further assumption that $\phi = 0$ (which is the maintained assumption made in KSS), (201) is simplified to

$$\Delta y_t = \delta u_{t-1}^3 + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i' \Delta \mathbf{z}_{t-i} + e_t. \tag{203}$$

For this model, we propose a t-type statistic for $\delta = 0$ (no cointegration) against $\delta < 0$ (ESTR cointegration), denoted by t_{NEC}.

The asymptotic distributions of all these tests are nonstandard, and the associated critical values have been tabulated via stochastic simulations.

The small sample performance of the suggested tests is compared to that of the linear EG and Johansen (1995) tests via Monte Carlo experiments. We find that our proposed nonlinear tests have good size and superior power properties compared to the linear tests. In particular, both F_{NEC} and t_{NEC} tests are superior to both linear or nonlinear EG tests when the regressors are weakly exogenous in a cointegrating regression. This supports similar findings made in linear models that the EG test loses power relative to ECM-based cointegration tests because of the loss of potentially valuable information from the correlation between the regressors and the underlying disturbances.

KSS provide an application to investigating the presence of cointegration of asset prices and dividends for eleven stock portfolios allowing for nonlinear STR adjustment to equilibrium. Interestingly, our new tests are able to reject the null of no cointegration in majority cases whereas the linear EG test rejects only twice. We also estimate adjustment parameters under the alternative, and find that these estimates are well defined in all cases. We further evaluate the impulse response functions of the error correction term with respect to initial impulses of 1-4 standard deviation shocks. The striking finding is that the time taken to recover one half of a one standard deviation shock varies between five and twenty years, whereas the time taken to recover one half of larger shocks varies between just 4 to 18 months. This implies that data periods dominated by extreme volatility may display substantial reversion of prices towards their NPV relationship, while in "calmer" times where the error in the NPV relationship takes on smaller values, the process driving it may well look like a unit root.

7.7.4 MS NEC Models

Psaradakis, Sola and Spagnolo (2004) consider the following single equation-based MS NEC model:

$$y_t + \alpha x_t = z_t,\ z_t = \phi_{s_t} z_{t-1} + \varepsilon_{1t}, \tag{204}$$

$$y_t + \beta x_t = u_t,\ u_t = u_{t-1} + \varepsilon_{2t}, \tag{205}$$

where $\alpha \neq 0$, $\beta \in R$, $\phi_{s_t} \in (-1, 1]$ and s_t are the latent random variables on $\{0, 1\}$. Suppose that

$$\phi_{s_t} = \phi_0 + (\phi_1 - \phi_0)\, s_t,\ |\phi_0| < 1,\ \phi_1 = 1, \tag{206}$$

where $\{s_t\}$ is a homogeneous irreducible and aperiodic Markov chain of order 1with state-space, $S = \{0, 1\}$ and transition probabilities

$$p_{ij} = \Pr\{s_t = j | s_{t-1} = i\},\ i, j \in S. \tag{207}$$

Deviations from equilibrium tend to decay to the mean level of 0 as long as $s_t = 0$; otherwise z_t behaves like a nonstationary process. Despite the occasional nonstaitoanry behavior of $\{z_t\}$ when $s_t = 1$, the eq error can be globally stationary, provided that p_{00}, p_{11}, ϕ_0 and ϕ_1 satisfy appropriate restrictions. A necessary and sufficient condition is given by (Franq and Zakoian, 2001)

$$p_{00}\phi_0^2 + p_{11}\phi_1^2 + (1 - p_{00} - p_{11})\,\phi_0^2\phi_1^2 < 1 \text{ and } p_{00}\phi_0^2 + p_{11}\phi_1^2 < 2. \tag{208}$$

For an irreducible and aperiodic Markov chain, these conditions are easily satisfied when $|\phi_0| < 1$ and $\phi_1 = 1$. See an application to the relationship between stock prices and dividends in Psaradakis, Sola and Spagnolo (2004).

We could also allow for $\{z_t\}$ to evolve according to the Markov switching ARMA model,

$$z_t = c_{s_t} + \sum_{i=1}^{m} \phi_{s_t}^{(i)} z_{t-i} + \sigma_{s_t}\xi_t + \sum_{j=1}^{q} \psi_{s_t}^{(i)} \sigma_{s_{t-j}}\xi_{t-j}, \tag{209}$$

where ξ_t is a white noise with $E\xi_t = 0$ and $E\xi_t^2 = 0$. A sufficient condition for the 2nd order stationarity is that all the eigenvalues of the $2m^2 \times 2m^2$ matrix,

$$\Lambda = \begin{bmatrix} p_{00}\,(\mathbf{\Phi}_0 \otimes \mathbf{\Phi}_0) & p_{10}\,(\mathbf{\Phi}_0 \otimes \mathbf{\Phi}_0) \\ p_{01}\,(\mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1) & p_{11}\,(\mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1) \end{bmatrix}$$

lie on the open disk where

$$\mathbf{\Phi}_h = \begin{bmatrix} \phi_h^{(1)} & \phi_h^{(2)} & \cdots & \phi_h^{(m-1)} & \phi_h^{(m)} \\ 1 & 0 & & 0 & 0 \\ 1 & 0 & & 0 & 0 \\ & & & & \\ 0 & 0 & & 1 & 0 \end{bmatrix},\ h \in S.$$

Another useful extension is that it is reasonable to expect that the further away from the equilibrium of the system is the higher the probability of switching from an unstable noncorrecting regime to a stable error correcting one. This allows the transition probabilities of the hidden Markov chain to depend on the extent to which the system is out of long-run equilibrium. Therefore,

$$\Pr\{s_t = i | s_{t-1} = i, z_{t-1}\} = \frac{\exp(a_i + b_i z_{t-1})}{1 + \exp(a_i + b_i z_{t-1})},\ i \in S, \tag{210}$$

$$\Pr\{s_t = j | s_{t-1} = i, z_{t-1}\} = 1 - \Pr\{s_t = i | s_{t-1} = i, z_{t-1}\},\ i \in S,\ i \neq j \tag{211}$$

It is natural to consider testing the null of single-regime/no-coinetgration against the alternative of cointegration with MEC adjustment. The testing problem is nonstandard due to the presence of unit roots and the unidentifiability of the transition probabilities under the null.

8 Regime Switching Panel Data Models

8.1 Panel Threshold Regression Models

This is a summary of the paper by B. Hansen (1999, JOE, 93: 345-368). The structural equation is

$$y_{it} = \mu_i + \beta_1' x_{it} I\left(q_{it} \leq \gamma\right) + \beta_2' x_{it} I\left(q_{it} > \gamma\right) + e_{it}, \tag{212}$$

which can be written as

$$y_{it} = \mu_i + \beta' x_{it}\left(\gamma\right) + e_{it}, \tag{213}$$

where $x_{it}\left(\gamma\right) = \left(\begin{array}{c} x_{it} I\left(q_{it} \leq \gamma\right) \\ x_{it} I\left(q_{it} > \gamma\right) \end{array} \right)$ and $\beta = \left(\beta_1', \beta_2'\right)'$. For identification it is required that x_{it} are not time invariant. e_{it} is assumed to be iid, which excludes lagged dependent variables from x_{it}. The analysis is asymptotic with fixed T as $n \to \infty$.

8.1.1 Estimation

Taking averages of (227),

$$\bar{y}_i = \mu_i + \beta' \bar{x}_i\left(\gamma\right) + \bar{e}_i, \tag{214}$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{i=1}^{T} y_{it};\ \bar{x}_i\left(\gamma\right) = \frac{1}{T}\sum_{i=1}^{T} x_{it}\left(\gamma\right) = \frac{1}{T}\sum_{i=1}^{T} \left(\begin{array}{c} x_{it} I\left(q_{it} \leq \gamma\right) \\ x_{it} I\left(q_{it} > \gamma\right) \end{array} \right);$$

and taking the difference between (228) and (229),

$$y_{it}^* = \beta' x_{it}^*\left(\gamma\right) + e_{it}^*, \tag{215}$$

where

$$y_{it}^* = y_{it} - \bar{y}_i;\ x_{it}^*\left(\gamma\right) = x_{it}\left(\gamma\right) - \bar{x}_i\left(\gamma\right).$$

Let

$$y_i^* = \left[\begin{array}{c} y_{i2}^* \\ \\ y_{iT}^* \end{array} \right];\ x_i^*\left(\gamma\right) = \left[\begin{array}{c} x_{i2}^{*\prime}\left(\gamma\right) \\ \\ x_{iT}^{*\prime}\left(\gamma\right) \end{array} \right]$$

denote the stacked data with one time period deleted. Then let Y^*

$$Y^* = \left[\begin{array}{c} y_1^* \\ \\ y_n^* \end{array} \right];\ X^*\left(\gamma\right) = \left[\begin{array}{c} x_1^*\left(\gamma\right) \\ \\ x_n^*\left(\gamma\right) \end{array} \right].$$

denote the data stacked over all individuals. Then

$$Y^* = X^*\left(\gamma\right)\beta + e^*. \tag{216}$$

For given γ, β can be estimated by OLS;

$$\hat{\beta}(\gamma) = \left(X^*(\gamma)' X^*(\gamma)\right)^{-1} X^*(\gamma)' Y^*. \tag{217}$$

Chan (1993) and Hansen (1999) recommend estimation of γ by LS. The LSE of γ is

$$\hat{\gamma} = \arg\min_{\gamma} S_1(\gamma), \tag{218}$$

where

$$S_1(\gamma) = \hat{e}^*(\gamma)' \hat{e}^*(\gamma) = Y^{*'} \left(I - X^*(\gamma)\left(X^*(\gamma)' X^*(\gamma)\right)^{-1} X^*(\gamma)'\right) Y^* \tag{219}$$

$$\hat{e}^*(\gamma) = Y^* - X^*(\gamma)\hat{\beta}(\gamma).$$

Once $\hat{\gamma}$ is obtained,

$$\hat{\beta} = \hat{\beta}(\hat{\gamma});\ \hat{e}^* = \hat{e}^*(\hat{\gamma});$$

$$\hat{\sigma}^2 = \frac{1}{n(T-1)} \hat{e}^{*\prime} \hat{e}^* = \frac{1}{n(T-1)} S_1(\hat{\gamma}). \tag{220}$$

Since $S_1(\gamma)$ depends only on γ through the indicator function, the sum of SSE is a step function with most nT steps with the steps occurring at distinct values of the observed threshold variable q_{it}. Thus the minimisation problem can be reduced to searching over the values of γ equalling the (at most nT) distinct values of q_{it} in the sample.

Sort the distinct values of the observations on q_{it}. Eliminate the smallest and largest η%. The remaining N values constitute the values of γ which can be searched for $\hat{\gamma}$. Foe each of these N values regression are estimated yielding the SSE. The smallest value yields the estimate $\hat{\gamma}$.

A simplifying shortcut is to restrict search to a smallest set of values of γ. The search may be limited to specific quintiles, perhaps integer valued. This reduces the number of regressions performed in the search. The estimation from such an approximation are likely to be sufficiently precise. For the empirical work we used the grid {1%, 1.25%, 1.5%, 1.75%, 2%,...,99% } which contains 393 quantiles.

8.1.2 Inference

The hypothesis of no threshold is:

$$H_0 : \beta_1 = \beta_2.$$

The FE (230) fall in the class of models considered by Hansen (1996) who suggested a bootstrap to simulate the asymptotic distribution of the LR test. Under the null of no threshold, the model is

$$y_{it} = \mu_i + \beta_1' x_{it} + e_{it}, \tag{221}$$

after the FE transformation, we have

$$y_{it}^{*} = \beta_1' x_{it}^{*} + e_{it}^{*}, \tag{222}$$

from which we obtain: $\tilde{\beta}_1$, $\tilde{e}_{it}^{*}$ and $S_0 = \tilde{e}^{*\prime}\tilde{e}^{*}$. The LR test is based on

$$F_1 = \frac{S_0 - S_1(\hat{\gamma})}{\hat{\sigma}^2}. \tag{223}$$

Hansen (1996) shows that a bootstrap procedure attains the first-order asymptotic distribution, so p-values are asymptotically valid.

Treat x_{it} and q_{it} as given. Take the residuals, $\hat{e}_{it}^{*}$ and group them by individual: $\hat{e}_i^{*} = (\hat{e}_{i1}^{*}, ..., \hat{e}_{iT}^{*})$. Treat $(\hat{e}_i^{*}, ..., \hat{e}_n^{*})$ as the empirical distribution to be used for bootstrapping. Draw (with replacement) a sample of size n from the empirical distribution and use these errors to create a bootstrap sample under H_0. Using the bootstrap sample estimate the model under the null and the alternative and calculate the bootstrap value of the LR test F_1. Repeat this procedure a large number of times and calculate the percentage of draws for which the simulated statistic exceeds the actual. This is the bootstrap estimate of the asymptotic p-value for F_1 under H_0.

Hansen (1999) argues that the best way to form CI for γ is to form the no-rejection region using the LR test. Toe test $H_0 : \gamma = \gamma_0$, we have

$$LR_1(\gamma) = \frac{S_1(\gamma) - S_1(\hat{\gamma})}{\hat{\sigma}^2}. \tag{224}$$

Note that the statistic (224) is testing a different hypothesis from (223).

Theorem 1. Under Assumptions 1-8 and $H_0 : \gamma = \gamma_0$

$$LR_1(\gamma) \to_d \xi,$$

as $n \to \infty$, where ξ is a random variable with distribution function,

$$P(\xi \leq x) = \left\{1 - \exp\left(-\frac{x}{2}\right)\right\}^2. \tag{225}$$

Since the asymptotic distribution in Theorem 1 is pivotal, it may be used to form valid asymptotic CIs. The distribution function (225) has the inverse:

$$c(\alpha) = -2\log\left(1 - \sqrt{1-\alpha}\right) \tag{226}$$

from which it is easy to calculate critical values.

To form an asymptotic CI for γ the non-rejection region of CI level $1-\alpha$ is the set of values of γ such that $LR_1(\gamma) \leq c(\alpha)$. This is a natural by-product of model estimation. To find LSE of γ the sequence of $S_1(\gamma)$ were calculated. $LR_1(\gamma)$ is a simple renormalization of these numbers and require no further computation.

Chan and Hansen show that the dependence on the threshold estimate is not of first-order asymptotic importance, so inference on β can proceed as if $\hat{\gamma}$ were true value. Hence,

$$\hat{\beta} \overset{a}{\sim} N(\beta, V),$$

where

$$\hat{V} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*} \left(\hat{\gamma} \right) x_{it}^{*} \left(\hat{\gamma} \right)' \right)^{-1} \hat{\sigma}^{2}.$$

If the errors are allowed to be conditional heteroskedastic, then

$$\begin{aligned} \hat{V} &= \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*} \left(\hat{\gamma} \right) x_{it}^{*} \left(\hat{\gamma} \right)' \right)^{-1} \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*} \left(\hat{\gamma} \right) x_{it}^{*} \left(\hat{\gamma} \right)' \hat{e}_{it}^{*\,2} \right) \\ &\quad \times \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*} \left(\hat{\gamma} \right) x_{it}^{*} \left(\hat{\gamma} \right)' \right)^{-1}. \end{aligned}$$

8.1.3 Investment and financing constraints

We use the multiple threshold regression model:

$$\begin{aligned} I_{it} &= \mu_i + \theta_1 Q_{it-1} + \theta_2 Q_{it-1}^2 + \theta_3 Q_{it-1}^3 + \theta_4 D_{it-1} + \theta_5 Q_{it-1} D_{it-1} \\ &\quad + \beta_1 CF_{it-1} I\left(D_{it-1} \leq \gamma_1 \right) + \beta_2 CF_{it-1} I\left(\gamma_1 < D_{it-1} \leq \gamma_2 \right) \\ &\quad + \beta_3 CF_{it-1} I\left(D_{it-1} > \gamma_2 \right) + e_{it} \end{aligned}$$

where I_{it} is the ratio of investment to capital, Q_{it} is the ratio of total market value to assets, CF_{it} is the ratio of cash flow to assets and D_{it} is the ratio of long term debt to assets, where the stock variables are defined at the end of year. See Table 5 (What is unexpected is that the firm with the highest debt levels have the smallest coefficient. Also in all three cases the coefficients on cash flows are positive.)

8.2 Panel Smooth Transition Regression Models

See Gonzalez, Terasvirta and van Dijk (2005).

8.3 Threshold Regression in Dynamic Panels

See Dang, Kim and Shin (2012, JEF) and Seo and Shin (2016, JoE).

8.4 PMG Estimation of Threshold Dynamic (Heterogeneous) Panels

See Shin (prep...).

8.5 Panel Threshold Regression Models in the Presence of CSD (Still very preliminary)

Another important issue is how to model the spatial dependence, the spatial heterogeneity and the spatial nonliearity, simultaneously. Here we provide a review of the limited studies so far...

8.5.1 Hacoglu Hoke and Kapetanios (2017)

Consider the panel data model where y_{it} is generated by the following data generating process (DGP),

$$y_{it} = \beta_{i1} x_{it} + \beta_{i2} x_{it} g\left(q_{it} : \gamma, c\right) + \phi_{y_{i1}} f_{1t} + \phi_{y_{i2}} f_{2t} + e_{it}$$

$$x_{it} = \phi_{i1} f_{1t} + \phi_{i3} f_{3t} + v_{it}$$

$$g\left(q_{it} : \gamma, c\right) = \frac{1}{1 + \exp\left(-\gamma\left(q_{it} - c\right)\right)}$$

where x_{it} is the observable regressors on the ith cross-sectional dimension at time t for $i = 1, ..., N, t = 1, ..., T$; f_{jt} is the unobserved factors, $e_{it} \sim IIDN(0, 1)$ and $q_{it} = x_{it}$ for $i = 1, ..., N$. We address the non-linear term $x_{it} g\left(q_{it} : \gamma, c\right)$ as w_{it}.

The function $g()$ is a logistic function used in Gonzalez, Terasvirta, and van Dijk (2005) and previously by Terasvirta (1994). The slope parameter γ and the location parameter c are estimated endogenously. Function $g()$ is bounded between 0 and 1. As $\gamma \to \infty$, $g()$ becomes an indicator function so the smooth transition model reduces to a threshold model with two regimes as in Hansen (1999), i.e. higher slope parameter leads to faster transition. When $\gamma \to 0$, the model reduces to a linear panel regression with fixed e ffcts. We take the parameters of $g()$ constant throughout the simulations, and over time and cross sections, i.e. $\gamma = 1$ and $c = 0$. We discuss the implications of heterogeneous parameters in Appendix B.

We explore two options for the correction of cross-sectional dependence:

Option 1: correction with $\bar{y}$, $\bar{x}$, and $\bar{w}$,

Option 2: correction with $\bar{y}$, $\bar{x}$.

We repeat the simulations for each pair of $(N, T) = 20, 50, 100, 200, 400$ with 2000 replications. Note that the sample we use for options 1 and 2 is the same within a replication for each pair of N and T.

8.5.2 Chudik, Mohaddes, Pesaran and Raiss (2017)

To explore the importance of heterogeneities, simultaneous determination of debt and growth, and dynamics, we begin with the following baseline autoregressive distributed lag (ARDL) specification:

$$\Delta y_{it} = c_i + \phi' g\left(d_{it}, \tau\right) + \sum_{\ell=1}^{p} \lambda_{i\ell} \Delta y_{i,t-\ell} + \sum_{\ell=0}^{p} \beta_{i\ell} \Delta d_{it-\ell} + v_{it},$$

and, following Chudik et al. (2016), we also consider the alternative approach of estimating the long-run effects using the distributed lag (DL) counterpart, given by

$$\Delta y_{it} = c_i + \theta' g\left(d_{it}, \tau\right) + \phi_i \Delta d_{i,t} + \sum_{\ell=0}^{p} \alpha_{i\ell} \Delta^2 d_{it-\ell} + v_{it},$$

where $g(d_{it},\tau)$ consists of up to two threshold variables: $g_1(d_{it},\tau) = I[d_{it} > \ln(\tau)]$ and/or $g_2(d_{it},\tau) = I[d_{it} > \ln(\tau)] \times \max(0, \Delta d_{it})$. The threshold variable $g_1(d_{it},\tau)$ takes the value of 1 if debt-to-GDP ratio is above the given threshold value of τ and 0 otherwise. The interactive threshold term, $g_2(d_{it},\tau)$, is nonzero only if $\Delta d_{it} > 0$, and $d_{it} > ln(\tau)$. As before, y_{it} is the log of real GDP and d_{it} is the log of debt-to-GDP. In addition to assuming a common threshold, τ, specifications also assume that the coefficients of the threshold variables, ϕ and θ, are the same across all countries whose debt-to-GDP ratio is above the common threshold τ. We test for the threshold effects not only in the full sample of forty countries but also for the two subsamples of advanced and developing countries, assuming homogeneous thresholds within each group, but allowing for the threshold parameters to vary across the country groupings.

Given the strong evidence of error cross-sectional dependence and as shown in section III, the panel threshold tests based on ARDL and DL regressions that do not allow for error cross-sectional dependence can yield incorrect inference regarding the presence of threshold effects. To address this problem, we employ the CS-ARDL and CSDL approaches, based on Chudik and Pesaran (2015a) and Chudik et al. (2016), which augment the ARDL and DL regressions with cross-sectional averages of the regressors, the dependent variable, and their lags. Specifically, the cross sectionally augmented ARDL (CS-ARDL) specification is given by

$$\Delta y_{it} = c_i + \phi' g(d_{it},\tau) + \sum_{\ell=1}^{p} \lambda_{i\ell} \Delta y_{i,t-\ell} + \sum_{\ell=0}^{p} \beta_i \Delta d_{i,t-\ell} + \sum_{\ell=0}^{p} \omega_{i\ell,h} \bar{h}_{t-\ell} + \omega'_{i,g} \bar{g}_t(\tau) + u_{it},$$

where $\bar{h}_t = \left(\Delta \bar{y}_t, \Delta \bar{d}_t\right)$, $\Delta \bar{y}_t$ and $\Delta \bar{d}_t$ are defined as averages of output growth and debt-to-GDP growth across countries and other variables are defined as before. The cross-sectionally augmented DL (CS-DL) specification is defined by

$$\begin{aligned} \Delta y_{it} &= c_i + \theta' g\left(d_{it},\tau\right) + \phi_i \Delta d_{i,t} + \sum_{\ell=0}^{p} \alpha_{i\ell} \Delta^2 d_{it-\ell} \\ &\quad + \omega_{i,y} \Delta \bar{y}_t + \sum_{\ell=0}^{p} \omega_{i\ell,d} \Delta \bar{d}_{t-\ell} + \omega'_{i,g} \bar{g}_t(\tau) + v_{it}, \end{aligned}$$

Compared to the CS-ARDL approach, the CS-DL method has better small sample performance for moderate values of T, which is often the case in applied work (see Chudik et al., 2016). Furthermore, it is robust to a number of departures from the baseline specification, such as residual serial correlation, and possible breaks in the error processes.

Omay and Kan (2010) Consider the following non-linear model with a single factor:

$$y_{it} = \mu_i + \beta_i' x_{it} + F\left(s_{it},\gamma,c\right) \tilde{\beta}_i' x_{it} + u_{it}$$

where

$$F\left(s_{it},\gamma,c\right) = \frac{1}{1 + \exp\left(-\gamma\left(s_{it} - c\right)\right)}$$

$$u_{it} = \varphi_i f_t + \varepsilon_{it}$$
$$x_{it} = \delta_i \tilde{f}_t + v_{it}$$

Notice here that f_t and $\tilde{f}_t$ are different factor variables that affect dependent, independent and state dependent variables, respectively. Now suppose that u_{it} and x_{it} specifications are plugged into original Eq. (15). Thus, we have:

$$y_{it} = \mu_i + \beta_i' \delta_i \tilde{f}_t + \beta_i' v_{it} + F(s_{it}, \gamma, c)\, \tilde{\beta}_i' \delta_i \tilde{f}_t + F(s_{it}, \gamma, c)\, \tilde{\beta}_i' v_{it} + \varphi_i f_t + \varepsilon_{it}$$

$\tilde{f}_t$ can be removed by using proxy variable, $\bar{x}_t$ where $\bar{x}_t = N^{-1} \sum_{i=1}^{N} x_{it}$ which can be obtained through taking the average of x_{it}:

$$\bar{x}_t = \bar{\delta} \tilde{f}_t + \bar{v}_t$$

which implies that

$$\tilde{f}_t = \frac{\bar{x}_t - \bar{v}_t}{\bar{\delta}}$$

Substituting this into Eq. (16) we obtain:

$$y_{it} = \mu_i + \beta_i' \delta_i \frac{\bar{x}_t - \bar{v}_t}{\bar{\delta}} + \beta_i' v_{it} + F(s_{it}, \gamma, c)\, \tilde{\beta}_i' \delta_i \frac{\bar{x}_t - \bar{v}_t}{\bar{\delta}} + F(s_{it}, \gamma, c)\, \tilde{\beta}_i' v_{it} + \varphi_i f_t + \varepsilon_{it}$$

In order to remove the factor ft from Eq. (17), we first take the averages of the above equation and obtain f_t appropriately:

$$\bar{y}_t = \bar{\mu} + \overline{\beta' \delta} \frac{\bar{x}_t - \bar{v}_t}{\bar{\delta}} + \overline{\beta' v_t} + \bar{F}(s_{it}, \gamma, c)\, \overline{\tilde{\beta}_i' \delta_i} \frac{\bar{x}_t - \bar{v}_t}{\bar{\delta}} + \overline{F(s_{it}, \gamma, c)\, \tilde{\beta}_i' v_{it}} + \bar{\varphi} f_t + \bar{\varepsilon}_t$$

then, with some algebra, $\bar{y}_t$ can be written as:

$$\bar{y}_t = \bar{\mu} + \overline{\beta} \bar{x}_t + \overline{\beta' v_t} + \bar{F}(s_{it}, \gamma, c)\, \overline{\tilde{\beta}_i'} \bar{x}_t + \bar{\varphi} f_t + \bar{\varepsilon}_t$$

Hence f_t is:

$$f_t = \frac{\bar{y}_t - \bar{\mu} - \overline{\beta} \bar{x}_t - \overline{\beta' v_t} - \bar{F}(s_{it}, \gamma, c)\, \overline{\tilde{\beta}_i'} \bar{x}_t - \bar{\varepsilon}_t}{\bar{\varphi}}$$

we obtain f_t from Eq. (20) and substitute it in Eq. (15):

$$y_{it} = \mu_i + \beta_i' x_{it} + F(s_{it}, \gamma, c)\, \tilde{\beta}_i' x_{it} + \frac{\varphi_i}{\varphi} \left(\frac{\bar{y}_t - \bar{\mu} - \overline{\beta} \bar{x}_t - \overline{\beta' v_t} - \bar{F}(s_{it}, \gamma, c)\, \overline{\tilde{\beta}_i'} \bar{x}_t - \bar{\varepsilon}_t}{\bar{\varphi}} \right) + \eta_{it}$$

again with relevant algebra, we obtain the auxiliary regression:

$$y_{it} = \tilde{\mu}_i + \beta_i' x_{it} + F(s_{it}, \gamma, c)\, \tilde{\beta}_i' x_{it} + a_i \bar{y}_t + b_i \bar{x}_t + F(.)\, c_i \bar{x}_t + \eta_{it}$$

The nonlinear model with a single factor is specified as follows:

$$y_{it} = \mu_i + \beta_0' x_{it} + \beta_1' x_{it} F(s_{1,it}, \gamma_1, c_1) + \beta_2' x_{it} G(s_{2,it}, \gamma_2, c_2) + \beta_3' x_{it} F(s_{1,it}, \gamma_1, c_1)\, G(s_{2,it}, \gamma_2, c_2) + u_{it}$$

where

$$F\left(s_{it},\gamma,c\right)=\frac{1}{1+\exp\left(-\gamma\left(s_{it}-c\right)\right)}$$
$$u_{it}=\varphi_i f_t+\varepsilon_{it}$$
$$x_{it}=\delta_i\tilde{f}_t+v_{it}$$

Notice here that f_t and $\tilde{f}_t$ are different factor variables that affect dependent, independent, and state-dependent variables, respectively. By applying the relevant algebra, we obtain the auxiliary regression in line with Omay and Kan (2010) and Omay (2014):

$$\begin{aligned} y_{it} &= \tilde{\mu}_i+\beta_0'x_{it}+\beta_1'x_{it}F\left(s_{1,it},\gamma_1,c_1\right)+\beta_2'x_{it}G\left(s_{2,it},\gamma_2,c_2\right)+\beta_3'x_{it}F\left(s_{1,it},\gamma_1,c_1\right)G\left(s_{2,it},\gamma_2,c_2\right) \\ &\quad +a\bar{y}_1+b\bar{x}_t+c_1\bar{F}\left(.\right)\bar{x}_t+c_2\bar{G}\left(.\right)\bar{x}_t+c_3\bar{F}\left(.\right)\bar{G}\left(.\right)\bar{x}_t+\eta_{it} \end{aligned}$$

Now we can estimate the models by this transformation in order to eliminate the cross-section dependence.

8.5.3 Eberhardt and Presbitero (2015)

The basic equation of the debt–growth nexus is a log-linearised Cobb–Douglas production function augmented with a debt stock:

$$y_{it}=\beta_i^K capi_t+\beta_i^D debt_{it}+u_{it};\ u_{it}=\alpha_i+\lambda_i'f_t+\varepsilon_{it}$$

where y is aggregate GDP, cap is capital stock and debt is the total debt stock — all variables are in logarithms of per capita terms. The parameter β_i^j (for $j=K,D$) heterogeneity is a central feature of our empirical setup as motivated above. Eq. (1) also includes country-specific intercepts (α_i) and a set of unobserved common factors f_t with country-specific 'factor loadings' λ_i to account for evolution of unobserved Total Factor Productivity (TFP), respectively. **Allowing the common factors to be nonstationary** has important implications, since all observable and unobservable processes are now integrated and standard inference is invalid (Kao, 1999). These common factors not only drive output, but also the capital and debt stocks, in line with the standard assumption of endogenous inputs to production.

We employ an error correction model (ECM) representation. This offers three advantages over static models and restricted dynamic specifications: (i) we can readily distinguish short-run from long-run behaviour; (ii) we can investigate the error correction term and deduce the speed of adjustment for the economy to the long-run equilibrium; and (iii) we can test for cointegration in the ECM by closer investigation of the statistical significance of the error correction term. The ECM representation is as follows:

$$\Delta y_{it}=\alpha_i+\rho_i\left(y_{i,t-1}-\beta_i^K cap_{i,t-1}-\beta_i^D debt_{i,t-1}-\lambda_i'f_{t-1}\right)+\gamma_i^K\Delta cap_{it}+\gamma_i^D\Delta debt_{it}+\gamma_i^{F\prime}\Delta f_t+\varepsilon_{it}$$

$$\Leftrightarrow$$

$$\Delta y_{it}=\pi_{0i}+\pi_i^{EC}y_{i,t-1}+\pi_i^K cap_{i,t-1}+\pi_i^D debt_{i,t-1}+\pi_i^{F\prime}f_{t-1}+\gamma_i^K\Delta cap_{it}+\gamma_i^D\Delta debt_{it}+\gamma_i^{F\prime}\Delta f_t+\varepsilon_{it}$$

where the β_i^j represent the long-run equilibrium relationship between GDP (y) and the measures for capital and debt in our model, while the γ_i^j represent the short-run relations. The ρ_i indicate the speed of convergence of the economy to its long-run equilibrium. **We included the common factors f in our long-run equation, which implies thatwe seek to investigate an equilibrium relationship between output, capital, debt and TFP**. In Eq. (4) we have relaxed the restrictions between the parameters ρ_i and β_i implicit in Eq. (3) and reparameterized the model.

Following Pesaran (2006) and Banerjee and Carrion-i-Silvestre (2011) we employ cross-section averages of all variables in the model to capture unobservables and omitted elements of the cointegration relationship. Recent work by Chudik and Pesaran (in press) has highlighted that in a dynamic panel this approach is subject to small sample bias. **They suggest to include further lags of the cross-section averages in addition to the cross-section averages of all model variables.** Our estimation equation is thus

$$\begin{aligned}\Delta y_{it} &= \pi_{0i} + \pi_i^{EC} y_{i,t-1} + \pi_i^K cap_{i,t-1} + \pi_i^D debt_{i,t-1} + \gamma_i^K \Delta cap_{it} + \gamma_i^D \Delta debt_{it} \\ &+\pi_{1i}^{CA}\overline{\Delta y_t} + \pi_{2i}^{CA}\overline{y_{t-1}} + \pi_{3i}^{CA}\overline{cap}_{t-1} + \pi_{4i}^{CA}\overline{debt}_{t-1} + \pi_{5i}^{CA}\overline{\Delta cap}_t + \pi_{6i}^{CA}\overline{\Delta debt}_t \\ &+ \sum_{l=2}^{p} \pi_{7il}^{CA}\overline{\Delta y_{t-l}} + \sum_{l=1}^{p} \pi_{8il}^{CA}\overline{\Delta cap}_{t-l} + \sum_{l=1}^{p} \pi_{9il}^{CA}\overline{\Delta debt}_{t-l} + \varepsilon_{it}\end{aligned}$$

Asymmetric dynamic model We follow Shin et al. (2013) and define the asymmetric long-run regression model:

$$y_{it} = \beta_i^K capi_t + \beta_i^{D+} debt_{it}^+ + \beta_i^{D-} debt_{it}^- + \lambda_i' f_t + \varepsilon_{it}$$

where debt stock has been decomposed into

$$debt_{it} = debt_{i0} + debt_{it}^+ + debt_{it}^-$$

The latter two terms are partial sums of values above and below a specific threshold, debt_{i0} has been subsumed into the constant term. The ECM version of our asymmetric dynamic model is then

$$\begin{aligned}\Delta y_{it} &= \pi_{0i} + \pi_i^{EC} y_{i,t-1} + \pi_i^K cap_{i,t-1} + \pi_i^{D+} debt_{i,t-1}^+ + \pi_i^{D-} debt_{i,t-1}^- + \pi_i^{F\prime} f_{t-1} \\ &+\pi_i^k \Delta cap_{it} + \pi_i^{d+} \Delta debt_{it}^+ + \pi_i^{d-} \Delta debt_{it}^- + \gamma_i^{F\prime} \Delta f_t + \varepsilon_{it}\end{aligned}$$

Finally, we obatin the CS augmented model by

$$\begin{aligned}\Delta y_{it} &= \pi_{0i} + \pi_i^{EC} y_{i,t-1} + \pi_i^K cap_{i,t-1} + \pi_i^{D+} debt_{i,t-1}^+ + \pi_i^{D-} debt_{i,t-1}^- \\ &+\pi_i^k \Delta cap_{it} + \pi_i^{d+} \Delta debt_{it}^+ + \pi_i^{d-} \Delta debt_{it}^- \\ &+\pi_{1i}^{CA}\overline{\Delta y_t} + \pi_{2i}^{CA}\overline{y_{t-1}} + \pi_{3i}^{CA}\overline{cap}_{t-1} + \pi_{4i}^{CA}\overline{debt^+}_{t-1} + \pi_{5i}^{CA}\overline{debt^-}_{t-1} \\ &+\pi_{6i}^{CA}\overline{\Delta cap}_t + \pi_{7i}^{CA}\overline{\Delta debt^+}_t + \pi_{8i}^{CA}\overline{\Delta debt^-}_t \\ &+ \sum_{l=2}^{p} \pi_{9il}^{CA}\overline{\Delta y_{t-l}} + \sum_{l=1}^{p} \pi_{10il}^{CA}\overline{\Delta cap}_{t-l} + \sum_{l=1}^{p} \pi_{11il}^{CA}\overline{\Delta debt^+}_{t-l} + \sum_{l=1}^{p} \pi_{12il}^{CA}\overline{\Delta debt^-}_{t-l} + \varepsilon_{it}\end{aligned}$$

8.5.4 PTAR extensions

Panel Threshold Regression Models This is a summary of the paper by B. Hansen (1999, JOE, 93: 345-368). The structural equation is

$$y_{it} = \mu_i + \beta_1' x_{it} I(q_{it} \leq \gamma) + \beta_2' x_{it} I(q_{it} > \gamma) + e_{it}, \tag{227}$$

which can be written as

$$y_{it} = \mu_i + \beta' x_{it}(\gamma) + e_{it}, \tag{228}$$

where $x_{it}(\gamma) = \begin{pmatrix} x_{it} I(q_{it} \leq \gamma) \\ x_{it} I(q_{it} > \gamma) \end{pmatrix}$ and $\beta = (\beta_1', \beta_2')'$. e_{it} is assumed to be iid, which excludes lagged dependent variables from x_{it}. The analysis is asymptotic with fixed T as $n \to \infty$.

Estimation: Taking averages of (227),

$$\bar{y}_i = \mu_i + \beta' \bar{x}_i(\gamma) + \bar{e}_i, \tag{229}$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{i=1}^{T} y_{it};\ \bar{x}_i(\gamma) = \frac{1}{T}\sum_{i=1}^{T} x_{it}(\gamma) = \frac{1}{T}\sum_{i=1}^{T} \begin{pmatrix} x_{it} I(q_{it} \leq \gamma) \\ x_{it} I(q_{it} > \gamma) \end{pmatrix};$$

and taking the difference between (228) and (229),

$$y_{it}^* = \beta' x_{it}^*(\gamma) + e_{it}^*, \tag{230}$$

where

$$y_{it}^* = y_{it} - \bar{y}_i;\ x_{it}^*(\gamma) = x_{it}(\gamma) - \bar{x}_i(\gamma).$$

Let

$$y_i^* = \begin{bmatrix} y_{i2}^* \\ \vdots \\ y_{iT}^* \end{bmatrix};\ x_i^*(\gamma) = \begin{bmatrix} x_{i2}^{*\prime}(\gamma) \\ \vdots \\ x_{iT}^{*\prime}(\gamma) \end{bmatrix}$$

denote the stacked data with one time period deleted. Then let Y^*

$$Y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix};\ X^*(\gamma) = \begin{bmatrix} x_1^*(\gamma) \\ \vdots \\ x_n^*(\gamma) \end{bmatrix}.$$

denote the data stacked over all individuals. Then

$$Y^* = X^*(\gamma)\beta + e^*. \tag{231}$$

For given γ, β can be estimated by OLS;

$$\hat{\beta}(\gamma) = \left(X^*(\gamma)' X^*(\gamma)\right)^{-1} X^*(\gamma)' Y^*. \tag{232}$$

Chan (1993) and Hansen (1999) recommend estimation of γ by LS. The LSE of γ is

$$\hat{\gamma} = \arg\min_{\gamma} S_1(\gamma), \tag{233}$$

where

$$S_1(\gamma) = \hat{e}^*(\gamma)' \hat{e}^*(\gamma) = Y^{*'} \left(I - X^*(\gamma) \left(X^*(\gamma)' X^*(\gamma) \right)^{-1} X^*(\gamma)' \right) Y^* \tag{234}$$

$$\hat{e}^*(\gamma) = Y^* - X^*(\gamma) \hat{\beta}(\gamma).$$

Once $\hat{\gamma}$ is obtained,

$$\hat{\beta} = \hat{\beta}(\hat{\gamma}); \ \hat{e}^* = \hat{e}^*(\hat{\gamma});$$

$$\hat{\sigma}^2 = \frac{1}{n(T-1)} \hat{e}^{*\prime} \hat{e}^* = \frac{1}{n(T-1)} S_1(\hat{\gamma}). \tag{235}$$

Since $S_1(\gamma)$ depends only on γ through the indicator function, the sum of SSE is a step function with most nT steps with the steps occurring at distinct values of the observed threshold variable q_{it}. Thus the minimisation problem can be reduced to searching over the values of γ equalling the (at most nT) distinct values of q_{it}. A simplifying shortcut is to restrict search to a smallest set of values of γ. The search may be limited to specific quintiles, perhaps integer valued. This reduces the number of regressions performed in the search.

Inference to follow:

We now extend the model and allow for the unobserved factors such that

$$y_{it} = \beta_1' x_{it} I(q_{it} \leq \gamma) + \beta_2' x_{it} I(q_{it} > \gamma) + u_{it}, \tag{236}$$

$$u_{it} = \mu_i + \lambda_i' f_t + e_{it}$$

where e_{it} is assumed to be iid. Consider the following (linear) data generating process for x_{it}:

$$x_{it} = \mu_{xi} + \lambda_{xi}' f_t + v_{it}, \tag{237}$$

where v_{it} is the generic stationary process, assumed to be uncorrelated with u_{it}. Define

$$x_{1,it}(\gamma) = x_{it} I(q_{it} \leq \gamma) \text{ and } x_{2,it}(\gamma) = x_{it} I(q_{it} > \gamma)$$

such that

$$x_{it} = x_{1,it}(\gamma) + x_{2,it}(\gamma)$$

$$x_{1,it} = \left(\mu_{xi} + \lambda_{xi}' f_t + v_{it} \right) I(q_{it} \leq \gamma) = \mu_{x1i}(\gamma) + \lambda_{x1i}'(\gamma) f_t + v_{i1t}(\gamma)?? \tag{238}$$

$$x_{2,it} = \left(\mu_{xi} + \lambda_{xi}' f_t + v_{it} \right) I(q_{it} > \gamma) = \mu_{x2i}(\gamma) + \lambda_{x2i}'(\gamma) f_t + v_{i2t}(\gamma)?? \tag{239}$$

Combining (286), (238) and (239), we have:

$$\boldsymbol{z}_{it} = \begin{pmatrix} y_{it} \\ x_{1,it} \\ x_{2,it} \end{pmatrix} = \boldsymbol{\mu}_i + \boldsymbol{\Phi}_i \boldsymbol{f}_t + \boldsymbol{\varepsilon}_{it} \tag{240}$$

where

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i + \beta_1' \mu_{x1i}(\gamma) + \beta_2' \mu_{x2i}(\gamma) \\ \mu_{x1i}(\gamma) \\ \mu_{x2i}(\gamma) \end{bmatrix}, \ \boldsymbol{\Phi}_i = \begin{bmatrix} \lambda_i' + \beta_1' \lambda_{x1i}(\gamma) + \beta_2' \lambda_{x2i}(\gamma) \\ \lambda_{x1i}'(\gamma) \\ \lambda_{x2i}'(\gamma) \end{bmatrix}$$

$$\boldsymbol{\varepsilon}_{it} = \begin{bmatrix} e_{it} + \beta_1' v_{i1t}(\gamma) + \beta_2' v_{i2t}(\gamma) \\ v_{i1t}(\gamma) \\ v_{i2t}(\gamma) \end{bmatrix}$$

Taking CS average of (340), we have:

$$\bar{\boldsymbol{z}}_t = \begin{pmatrix} \bar{y}_t \\ \bar{x}_{1,t} \\ \bar{x}_{2,t} \end{pmatrix} = \bar{\boldsymbol{\mu}} + \bar{\boldsymbol{\Phi}} \boldsymbol{f}_t + \bar{\boldsymbol{\varepsilon}}_t \tag{241}$$

Suppose that the rank condition holds. Then, from (241), we have:

$$\boldsymbol{f}_t = \left(\bar{\boldsymbol{\Phi}}' \bar{\boldsymbol{\Phi}} \right)^{-1} \bar{\boldsymbol{\Phi}} \left(\bar{\boldsymbol{z}}_t - \bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\varepsilon}}_t \right)$$

As $N \to \infty$,

$$\bar{\boldsymbol{\varepsilon}}_t = O_p\left(\frac{1}{N}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right),$$

hence

$$\boldsymbol{f}_t - \left(\bar{\boldsymbol{\Phi}}' \bar{\boldsymbol{\Phi}} \right)^{-1} \bar{\boldsymbol{\Phi}} \left(\bar{\boldsymbol{z}}_t - \bar{\boldsymbol{\mu}} \right) = O_p\left(\frac{1}{N}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right).$$

This suggests that we can use $\bar{\boldsymbol{z}}_t$ as observable proxies for $\boldsymbol{f}_t$. Then, we can consistently estimate the individual slope coefficients, β_1 and β_2 (or β_{1i} and β_{2i} as well as their means) by augmenting the regression, (286) with the cross-section averages $\bar{\boldsymbol{z}}_t$.

Alternatively, we apply the within transformation to (286) first and obtain:

$$y_{it}^* = \beta_1' x_{1,it}^*(\gamma) + \beta_2' x_{2,it}^*(\gamma) + \lambda_i' f_t^* + e_{it}^*, \tag{242}$$

where

$$y_{it}^* = y_{it} - \bar{y}_i, \ x_{1,it}^*(\gamma) = x_{1,it}(\gamma) - \bar{x}_{1i}(\gamma), \ x_{2,it}^*(\gamma) = x_{2,it}(\gamma) - \bar{x}_{2i}(\gamma),$$

$$f_t^* = f_t - \bar{f}, \ e_{it}^* = e_{it} - \bar{e}_i.$$

Applying the within transformation to (339), we have:

$$x_{it}^* = \lambda_{xi}' f_t^* + v_{it}^*, \tag{243}$$

where $v_{it}^* = v_{it} - \bar{v}_i$. Then,

$$x_{1,it}^* = \left(\lambda_{xi}' f_t^* + v_{it}^* \right) I(q_{it} \leq \gamma) = \lambda_{x1i}'(\gamma) f_t^* + v_{i1t}^*(\gamma)?? \tag{244}$$

$$x_{2,it}^* = \left(\lambda_{xi}' f_t^* + v_{it}^* \right) I(q_{it} > \gamma) = \lambda_{x2i}'(\gamma) f_t^* + v_{i2t}^*(\gamma)?? \tag{245}$$

Combining (242), (244) and (245), we have:

$$\boldsymbol{z}_{it}^{*}=\left(\begin{array}{c} y_{it}^{*} \\ x_{1,it}^{*} \\ x_{2,it}^{*} \end{array}\right)=\boldsymbol{\Phi}_{i}\boldsymbol{f}_{t}^{*}+\boldsymbol{\varepsilon}_{it}^{*} \tag{246}$$

where

$$\boldsymbol{\Phi}_{i}=\left[\begin{array}{c} \lambda_{i}^{\prime}+\beta_{1}^{\prime}\lambda_{x1i}\left(\gamma\right)+\beta_{2}^{\prime}\lambda_{x2i}\left(\gamma\right) \\ \lambda_{x1i}^{\prime}\left(\gamma\right) \\ \lambda_{x2i}^{\prime}\left(\gamma\right) \end{array}\right],\ \boldsymbol{\varepsilon}_{it}^{*}=\left[\begin{array}{c} e_{it}+\beta_{1}^{\prime}v_{i1t}^{*}\left(\gamma\right)+\beta_{2}^{\prime}v_{i2t}^{*}\left(\gamma\right) \\ v_{i1t}^{*}\left(\gamma\right) \\ v_{i2t}^{*}\left(\gamma\right) \end{array}\right]$$

Taking CS average of (246), we have:

$$\bar{\boldsymbol{z}}_{t}^{*}=\left(\begin{array}{c} \bar{y}_{t}^{*} \\ \bar{x}_{1,t}^{*} \\ \bar{x}_{2,t}^{*} \end{array}\right)=\bar{\boldsymbol{\Phi}}\boldsymbol{f}_{t}^{*}+\bar{\boldsymbol{\varepsilon}}_{t}^{*} \tag{247}$$

Suppose that the rank condition holds. Then, from (247), we have:

$$\boldsymbol{f}_{t}^{*}=\left(\bar{\boldsymbol{\Phi}}^{\prime}\bar{\boldsymbol{\Phi}}\right)^{-1}\bar{\boldsymbol{\Phi}}\left(\bar{\boldsymbol{z}}_{t}^{*}-\bar{\boldsymbol{\varepsilon}}_{t}^{*}\right)$$

As $N\rightarrow\infty$,

$$\bar{\boldsymbol{\varepsilon}}_{t}^{*}=O_{p}\left(\frac{1}{N}\right)+O_{p}\left(\frac{1}{\sqrt{NT}}\right),$$

hence

$$\boldsymbol{f}_{t}^{*}-\left(\bar{\boldsymbol{\Phi}}^{\prime}\bar{\boldsymbol{\Phi}}\right)^{-1}\bar{\boldsymbol{\Phi}}\bar{\boldsymbol{z}}_{t}^{*}=O_{p}\left(\frac{1}{N}\right)+O_{p}\left(\frac{1}{\sqrt{NT}}\right).$$

This suggests that we can use $\bar{\boldsymbol{z}}_{t}^{*}$ as observable proxies for $\boldsymbol{f}_{t}^{*}$. Then, we can consistently estimate the individual slope coefficients, β_{1} and β_{2} (or β_{1i} and β_{2i} as well as their means) by augmenting the regression, (242) with the cross-section averages $\bar{\boldsymbol{z}}_{t}^{*}$.

Dynamic Panel Threshold Regression Models to follow:

8.5.5 Spatial TAR or STAR models

Surprisingly limited number of studies, e.g. Pede, Florax and Lambert (2014)... The spatial STAR model with a spatially lagged exogenous variable in the transition function, combined with a spatially lagged dependent variable and spatially autoregressive errors is:

$$y=\rho Wy+X\alpha+X\delta\circ G\left(Wx,\gamma,c\right)+\varepsilon$$

$$\varepsilon=\lambda W\varepsilon+\mu$$

where ρ and λ are spatial autoregressive parameters pertaining to the lag and/or the error, μ are independent and identically distributed disturbances, and the

other symbols are as defined before. Three different spatial STAR models are nested in the general ARAR STAR model depending on the value of the spatial autoregressive parameters. The restriction $\lambda = 0$ leads to the spatial lag STAR model, $\rho = 0$ to the spatial error STAR model, and $\rho = \lambda = 0$ to the basic spatial STAR model. Each model can be estimated with maximum likelihood (ML).

9 Multi-dimensional Panel Data Modelling in the Presence of Cross Sectional Dependence

Given the growing availability of the dataset which contain information on multiple dimensions, the recent literature has focused more on extending the two-way fixed effects models to the multidimensional setting. The 'triple-way specification' has been popularised by Mátyás (1997), where time, source (exporter) and destination (importer) fixed effects are specified respectively as unobservable. Baltagi et al. (2003) propose an extended specification with fixed exporter-time, importer-time, and country-pair effects. The triple-index models can be applied to the number of bilateral flows between countries or regions such as trade, FDI, capital or migration flows (e.g. Feenstra, 2005; Bertoli and Fernandez-Huertas Moraga, 2013; Gunnella et al., 2015), but also to a variety of matched dataset which may link the employer-the employee and pupils-teachers (e.g. Abowd et al., 1999; Kramarz et al., 2008).

Balazsi, Mathyas and Wansbeek (2015, BMW) introduce the 3D within estimators and analyse the behavior of these estimators in the cases of no self-flow data, unbalanced data, and dynamic autoregressive models.[20] Consider the 3D country-time fixed effects panel data model:

$$y_{ijt} = \boldsymbol{\beta}'\mathbf{x}_{ijt} + \boldsymbol{\gamma}'\mathbf{s}_{it} + \boldsymbol{\delta}'\mathbf{d}_{jt} + \boldsymbol{\kappa}'\mathbf{q}_t + \boldsymbol{\varphi}'\mathbf{z}_{ij} + u_{ijt}, \tag{248}$$

for $i = 1, ..., N_1, j = 1, ..., N_2, t = 1, ..., T$, with errors:

$$u_{ijt} = \mu_{ij} + v_{it} + \zeta_{jt} + \varepsilon_{ijt} \tag{249}$$

where y_{ijt} is the dependent variable across 3 indices (e.g. the import of country j from country i at period t); $\mathbf{x}_{ijt}$, $\mathbf{s}_{it}$, $\mathbf{d}_{jt}$, $\mathbf{q}_t$, $\mathbf{z}_{ij}$ are the $k_x \times 1$, $k_s \times 1$, $k_d \times 1$, $k_q \times 1$, $k_z \times 1$ vectors of covariates covering all measurements across 3indices; The multi-error components contain pair-fixed effects $\left(\mu_{ij}\right)$ as well as origin and destination CTFEs, v_{it} and ζ_{jt}.

Unobserved fixed effects, μ_{ij}, v_{it} and ζ_{jt}, can be modelled by adding the following $N^2T \times (N^2 + 2NT)$ matrix of the dummies:

$$D = \left(\left(I_N \otimes I_N \otimes l_T\right), \left(I_N \otimes l_N \otimes I_T\right), \left(l_N \otimes I_N \otimes I_T\right)\right) \ \left(N^2T \times \left(N^2 + 2NT\right)\right)$$

[20] See also Balazsi Mathyas and Pus (2015) for the random effect approach and Baltagi et al. (2015) for the host of issues related to the panel data gravity models of trade.

where I_N is the $N \times N$ identity matrix and l_N is the $N \times 1$ vector of ones (similarly for I_T and l_T). To remove all unobserved FEs, BMW derive the 3D within transformation:

$$\tilde{y}_{ijt} = y_{ijt} - \bar{y}_{ij\cdot} - \bar{y}_{\cdot jt} - \bar{y}_{i\cdot t} + \bar{y}_{\cdot\cdot t} + \bar{y}_{\cdot j\cdot} + \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots} \tag{250}$$

where $\bar{y}_{ij\cdot} = T^{-1}\sum_{t=1}^{T} y_{ijt}$, $\bar{y}_{\cdot jt} = N^{-1}\sum_{i=1}^{N} y_{ijt}$, $\bar{y}_{i\cdot t} = N^{-1}\sum_{j=1}^{N} y_{ijt}$, $\bar{y}_{\bullet\bullet t} = N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ijt}$, $\bar{y}_{\bullet j\bullet} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{ijt}$, $\bar{y}_{i\bullet\bullet} = (NT)^{-1}\sum_{j=1}^{N}\sum_{t=1}^{T} y_{ijt}$ and $\bar{y}_{\cdots} = \left(N^2T\right)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T} y_{ijt}$.

We can estimate consistently $\boldsymbol{\beta}$ from the transformed regression:

$$\tilde{y}_{ijt} = \boldsymbol{\beta}'\tilde{\mathbf{x}}_{ijt} + \tilde{\varepsilon}_{ijt}, \tag{251}$$

where $\tilde{\mathbf{x}}_{ijt} = \mathbf{x}_{ijt} - \bar{\mathbf{x}}_{ij\cdot} - \bar{\mathbf{x}}_{\cdot jt} - \bar{\mathbf{x}}_{i\cdot t} + \bar{\mathbf{x}}_{\cdot\cdot t} + \bar{\mathbf{x}}_{\cdot j\cdot} + \bar{\mathbf{x}}_{i\cdot\cdot} - \bar{\mathbf{x}}_{\cdots}$. We write (251) compactly as

$$\tilde{\mathbf{Y}}_{ij} = \tilde{\mathbf{X}}_{ij}\boldsymbol{\beta} + \tilde{\mathbf{E}}_{ij} \tag{252}$$

$$\underset{T\times 1}{\tilde{\mathbf{Y}}_{ij}} = \begin{bmatrix} \tilde{y}_{ij1} \\ \vdots \\ \tilde{y}_{ijT} \end{bmatrix}, \underset{T\times k_x}{\tilde{\mathbf{X}}_{ij}} = \begin{bmatrix} \tilde{\mathbf{x}}'_{ij1} \\ \vdots \\ \tilde{\mathbf{x}}'_{ijT} \end{bmatrix}, \underset{T\times 1}{\tilde{\mathbf{E}}_{ij}} = \begin{bmatrix} \tilde{\varepsilon}_{ij1} \\ \vdots \\ \tilde{\varepsilon}_{ijT} \end{bmatrix}.$$

The 3D-within estimator of $\boldsymbol{\beta}$ is obtained by

$$\hat{\boldsymbol{\beta}}_W = \left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\tilde{\mathbf{X}}'_{ij}\tilde{\mathbf{X}}_{ij}\right)^{-1}\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\tilde{\mathbf{X}}'_{ij}\tilde{\mathbf{Y}}_{ij}\right). \tag{253}$$

As $(N_1, N_2, T) \to \infty$,

$$\begin{aligned} &\sqrt{N_1N_2T}\left(\hat{\boldsymbol{\beta}}_W - \boldsymbol{\beta}\right) \\ &\overset{a}{\sim} N\left(\mathbf{0}, \sigma^2_{\varepsilon}\lim_{(N_1,N_2,T)\to\infty}\left(\frac{1}{N_1N_2T}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\tilde{\mathbf{X}}'_{ij}\tilde{\mathbf{X}}_{ij}\right)^{-1}\right). \end{aligned} \tag{254}$$

The 3D within transformation wipes out all other covariates, $\mathbf{x}_{it}$, $\mathbf{x}_{jt}$, $\mathbf{x}_t$, and $\mathbf{x}_{ij}$. It would be worthwhile to develop an extension of the Hausman-Taylor (1981) estimation, popular in the 2D panels in the presence of CSD (e.g. Serlenga and Shin, 2007). Balazsi, Bun, Chan and Harris (2017) develop an extended HT estimator for multi-dimensional panels.

9.1 Research Extensions

We now propose a number of important research extensions.

- First, the 3D within transformation, (250) wipes out the regressors x_{it}, x_{jt}, x_t and x_{ij}. But, we are also interested in uncovering the effects of those covariates, e.g. the impacts of measured trade costs in the structural gravity model. In order to recover these coefficients, we wish to develop an extension of the Hausman-Taylor type estimation or the Mundlak transformations.

- Second and more importantly, we aim to introduce the cross-section dependence (CSD) explicitly within the three-way error component specifications. This makes the timely contribution to the literature, given the massive development of modelling CSD in the 2D panels, e.g. Pesaran (2006) and Bai (2009, 2014).
  - What's the nature of CSD, strong, semi-strong or weak? As α_{it} and α_{jt}^{*} in the CTFE model (**??**) are related to the local time factors, their CSD may be semi-strong.
  - To introduce the strong CSD, we may add the common unobserved global factor, λ_t:

$$u_{ijt} = \mu_{ij} + v_{it} + \zeta_{jt} + \lambda_t + \varepsilon_{ijt} \tag{255}$$

    However, the 3D-within transformation also removes λ_t, because λ_t is proportional to $\sum_{i=1}^{N_1} v_{it}$ or $\sum_{j=1}^{N_2} \zeta_{jt}$.
  - Consider the following error components:

$$u_{ijt} = \mu_{ij} + \pi_{ij}\lambda_t + \varepsilon_{ijt} \tag{256}$$

    which is the two-way version of the factor model considered by Serlenga and Shin (2007). More generally, we will examine:

$$u_{ijt} = \mu_{ij} + v_{it} + \zeta_{jt} + \pi_{ij}\lambda_t + \varepsilon_{ijt} \tag{257}$$

  - Kapetanios, Serlenga and Shin (2017) model the error components, u_{ijt} to follow the hierarchical multi-factor structure:

$$u_{ijt} = \boldsymbol{\gamma}_{ij}'\boldsymbol{f}_t + \boldsymbol{\gamma}_{\circ j}'\boldsymbol{f}_{i\circ t} + \boldsymbol{\gamma}_{i\circ}'\boldsymbol{f}_{\circ jt} + \varepsilon_{ijt}, \tag{258}$$

    where $\boldsymbol{f}_t$, $\boldsymbol{f}_{\circ jt}$ and $\boldsymbol{f}_{i\circ t}$ are respectively $m_f \times 1$, $m_{\circ\bullet} \times 1$ and $m_{\bullet\circ} \times 1$ vectors of unobserved common effects, and ε_{ijt} are idiosyncratic errors. Exporter i reacts heterogeneously to the common import market condition $\boldsymbol{f}_{\circ jt}$ and importer j reacts heterogeneously to the common export market condition $\boldsymbol{f}_{i\circ t}$. Both exporter i and importer j reacts heterogeneously to the common global market condition $\boldsymbol{f}_t$. Within this model, we can distinguish between three types of CSD: (i) the strong global factor, $\boldsymbol{f}_t$ which influences the (ij) pairwise interactions (of N^2 dimension); (ii) the semi-strong local factors, $\boldsymbol{f}_{i\circ t}$ and $\boldsymbol{f}_{\circ jt}$, which influence origin or destination separately (each of N dimension); and (iii) the weak CSD idiosyncratic errors, ε_{ijt}.

- We expect that this kind of generalisation would be most natural within the 3D panel data models.

- In such a case, the 3D within estimator is unable to estimate β's consistently due to the nonzero correlation between omitted interactive effects $(\pi_{ij}\lambda_t)$ and the regressors.

- The full estimation of (**??**) with (256), (257) or (338) can be feasible by combining the Pesaran or the Bai type estimation procedures with the Hausman-Taylor extension.
- Finally, there are a few studies which attempt to develop a general approach that can accommodate both weak and strong CSD in modelling cross-sectionally correlated panels.
  - Bailey et al. (2015) develop the multi-step estimation procedure that can distinguish the relationship between spatial units that is purely spatial from that which is due to the effect of common factors.
  - Kapetanios et al. (2014) advance a flexible nonlinear panel data model, which can generate strong and/or weak CSD endogenously. Furthermore, Mastromarco et al. (2015) propose the novel technique for allowing weak and strong CSD in modelling technical efficiency of stochastic frontier panels by combining the exogenously driven factor-based approach by SS and an endogenous threshold regime selection by KMS. See also Gunella et al. (2015).
  - Bai and Li (2015) and Shi and Lee (2017) develop the framework for jointly modelling spatial effects and interactive effects. See also Kuersteiner and Prucha (2015).
  - Hence, the extension of such joint modelling to the multidimensional data would be challenging but shed further lights on the understanding the complex structure of CSD.

9.2 The Econometrics of Multi-dimensional Panel Data Modelling in the Presence of Cross-sectional Error Dependences by Kapetanios, Mastromarco, Serlenga and Shin (2017, KMSS)

There has been no study to address an issue of controlling cross-section (error) dependence (CSD) in 3D or higher-dimensional data, despite strong CSD evidence in 2D panels (Pesaran, 2015). Two main approaches in modelling CSD in 2D panels are: (i) the factor-based approach (Pesaran, 2006; Bai, 2009) and (ii) the spatial econometrics techniques (Baltagi, 2005; Behrens et al., 2012).

The factor-based models exhibit strong CSD while the spatial models weak CSD only (Chudik et al., 2011). See also Bailey et al. (2016), Le Gallo and Pirotte (2017), and Baltagi, Egger and Erhardt (2017).

We develop the 3D models with strong CSD. We generalise the multi-dimensional error components by incorporating unobserved heterogeneous global factors. The country-time fixed (CTFE) and random effects (CTRE) estimators fail to remove heterogenous global factors; inconsistent in the presence of nonzero correlation between the regressors and unobserved global factors.

We develop the 2-step estimation procedure. Following Pesaran (2006), we augment the 3D model with cross-section averages of dependent variable and regressors, as proxies for unobserved global factors. We then apply the 3D-within

transformation to the augmented specification and obtain consistent estimators (the 3D-PCCE estimator). Our approach is the first attempt to accommodate strong CSD in multi-dimensional panels.

We discuss the extent of CSD under 3 different error components with CTFE, with 2-way heterogeneous factor, and with both. We develop a diagnostic test for the null of (pairwise) residual cross-section independence or weak dependence, which is a modified CD test in 2D panels proposed by Pesaran (2015). We also provide extensions into unbalanced panels and 4D models.

Monte Carlo studies confirm that the 3D-PCCE estimators perform well when the 3D panels are subject to heterogeneous global factors. On the contrary CTFE displays severe biases and size distortions. We apply the 3D PCCE estimation, together with the 2-way FE and CTFE estimators, to the dataset over 1960-2008 for 91 country-pairs amongst 14 EU countries. Based on the CD test results, and the predicted signs and statistical significance of the coefficients, we find that the 3D PCCE estimation results are most reliable and satisfactory. The trade effect of currency union is rather modest. This suggests that the trade increase within the Euro area may reflect a continuation of a long-run historical trend linked to the broader set of EU's economic integration policies.

Throughout we adopt the following standard notations. $\mathbf{I}_N$ is an $N \times N$ identity matrix, $\mathbf{J}_N$ the $N \times N$ identity matrix of ones, and $\boldsymbol{\iota}_N$ the $N \times 1$ vector of ones, respectively. $\mathbf{M}_A$ projects the $N \times N$ matrix $\mathbf{A}$ into its null-space, i.e., $\mathbf{M}_A = \mathbf{I}_N - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. Finally, $y_{.jt} = N_1^{-1}\sum_{i=1}^{N_1} y_{ijt}$, $y_{i.t} = N_2^{-1}\sum_{j=1}^{N_2} y_{ijt}$ and $y_{ij.} = T^{-1}\sum_{t=1}^{T} y_{ijt}$ denote the average of y over the index i, j and t, respectively, with the definition extending to other quantities such as $y_{..t}$, $y_{.j.}$, $y_{i..}$ and $y_{...}$.

9.2.1 The 3D Models with CSD

We first consider the following error components specification:

$$u_{ijt} = \mu_{ij} + \pi_{ij}\lambda_t + \varepsilon_{ijt}. \tag{259}$$

Similar to the 2-way heterogeneous factor model by Serlenga and Shin (2007). We apply the cross-section averages of (248) over i and j:

$$\bar{y}_{..t} = \boldsymbol{\beta}'\bar{\mathbf{x}}_{..t} + \boldsymbol{\gamma}'\bar{\mathbf{s}}_{.t} + \boldsymbol{\delta}'\bar{\mathbf{d}}_{.t} + \boldsymbol{\kappa}'\mathbf{q}_t + \boldsymbol{\varphi}'\bar{\mathbf{z}}_{..} + \bar{\mu}_{..} + \bar{\pi}_{..}\lambda_t + \bar{\varepsilon}_{..t} \tag{260}$$

Hence, we have:

$$\lambda_t = \frac{1}{\bar{\pi}_{..}}\left\{\bar{y}_{..t} - \left(\boldsymbol{\beta}'\bar{\mathbf{x}}_{..t} + \boldsymbol{\gamma}'\bar{\mathbf{s}}_{.t} + \boldsymbol{\delta}'\bar{\mathbf{d}}_{.t} + \boldsymbol{\kappa}'\mathbf{q}_t + \boldsymbol{\varphi}'\bar{\mathbf{z}}_{..} + \bar{\mu}_{..} + \bar{\varepsilon}_{..t}\right)\right\}$$

We augment the model (248) with the cross-section averages:

$$y_{ijt} = \boldsymbol{\beta}'\mathbf{x}_{ijt} + \boldsymbol{\gamma}'\mathbf{s}_{it} + \boldsymbol{\delta}'\mathbf{d}_{jt} + \boldsymbol{\psi}_{ij}'\mathbf{f}_t + \tau_{ij} + \mu_{ij}^* + \varepsilon_{ijt}^*, \tag{261}$$

where

$$\boldsymbol{\psi}_{ij}' = \left(\frac{\pi_{ij}}{\bar{\pi}_{..}}, \frac{-\pi_{ij}\boldsymbol{\beta}'}{\bar{\pi}_{..}}, \frac{-\pi_{ij}\boldsymbol{\gamma}'}{\bar{\pi}_{..}}, \frac{-\pi_{ij}\boldsymbol{\delta}'}{\bar{\pi}_{..}}, \left(1 - \frac{\pi_{ij}}{\bar{\pi}_{..}}\right)\boldsymbol{\kappa}'\right)$$

$$\mathbf{f}_t = \left(\bar{y}_{..t}, \bar{\mathbf{x}}'_{..t}, \bar{\mathbf{s}}'_{.t}, \bar{\mathbf{d}}'_{.t}, \mathbf{q}'_t\right)' \tag{262}$$

$$\tau_{ij} = \boldsymbol{\varphi}'\mathbf{z}_{ij} - \frac{-\pi_{ij}}{\bar{\pi}_{..}}\boldsymbol{\varphi}'\mathbf{z}_{..},\ \mu^*_{ij} = \mu_{ij} - \frac{\pi_{ij}\mu_{..}}{\bar{\pi}_{..}},\ \varepsilon^*_{ijt} = \varepsilon_{ijt} - \frac{\pi_{ij}}{\bar{\pi}_{..}}\bar{\varepsilon}_{..t}.$$

We write (261) compactly as

$$\mathbf{Y}_{ij} = \mathbf{W}_{ij}\boldsymbol{\theta} + \mathbf{H}\boldsymbol{\psi}^*_{ij} + \mathbf{E}^*_{ij},\ i = 1, ..., N_1, j = 1, ..., N_2 \tag{263}$$

$$\underset{T\times 1}{\mathbf{Y}_{ij}} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijT} \end{bmatrix},\ \underset{T\times k_x}{\mathbf{X}_{ij}} = \begin{bmatrix} \mathbf{x}'_{ij1} \\ \vdots \\ \mathbf{x}'_{ijT} \end{bmatrix},\ \underset{T\times k_s}{\mathbf{S}_i} = \begin{bmatrix} \mathbf{s}'_{i1} \\ \vdots \\ \mathbf{s}'_{iT} \end{bmatrix},$$

$$\underset{T\times k_d}{\mathbf{D}_j} = \begin{bmatrix} \mathbf{d}'_{j1} \\ \vdots \\ \mathbf{d}'_{jT} \end{bmatrix},\ \underset{T\times k_f}{\mathbf{F}} = \begin{bmatrix} \mathbf{f}'_1 \\ \vdots \\ \mathbf{f}'_T \end{bmatrix},\ \underset{T\times 1}{\mathbf{E}^*_{ij}} = \begin{bmatrix} \varepsilon^*_{ij1} \\ \vdots \\ \varepsilon^*_{ijT} \end{bmatrix},$$

$\mathbf{W}_{ij} = \left(\mathbf{X}_{ij}, \mathbf{S}_i, \mathbf{D}_j\right)$, $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta}' & \boldsymbol{\gamma}' & \boldsymbol{\delta}' \end{pmatrix}'$, $\boldsymbol{\psi}^*_{ij} = \left(\boldsymbol{\psi}'_{ij}, \left(\tau_{ij} + \mu^*_{ij}\right)\right)'$ and $\mathbf{H} = [\mathbf{F}, \boldsymbol{\iota}_T]$.

We derive the 3D-PCCE estimator of $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}}_{PCCE} = \left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\mathbf{W}'_{ij}\mathbf{M}_H\mathbf{W}_{ij}\right)^{-1}\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\mathbf{W}'_{ij}\mathbf{M}_H\mathbf{Y}_{ij}\right) \tag{264}$$

where $\mathbf{M}_H = \mathbf{I}_T - \mathbf{H}\left(\mathbf{H}'\mathbf{H}\right)^{-1}\mathbf{H}'$.Following Pesaran (2006), it is straightforward to show that as $(N_1, N_2, T) \to \infty$,

$$\sqrt{N_1N_2T}\left(\hat{\boldsymbol{\theta}}_{PCCE} - \boldsymbol{\theta}\right) \overset{a}{\sim} N\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right), \tag{265}$$

where the (robust) consistent estimator of $\boldsymbol{\Sigma}_\theta$ is given by

$$\hat{\boldsymbol{\Sigma}}_\theta = \frac{1}{N_1N_2}\mathbf{S}_\theta^{-1}\mathbf{R}_\theta\mathbf{S}_\theta^{-1}, \tag{266}$$

$$\mathbf{R}_\theta = \frac{1}{N_1N_2 - 1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\frac{\mathbf{W}'_{ij}\mathbf{M}_H\mathbf{W}_{ij}}{T}\right)\left(\hat{\boldsymbol{\theta}}_{ij} - \hat{\boldsymbol{\theta}}_{MG}\right)\left(\hat{\boldsymbol{\theta}}_{ij} - \hat{\boldsymbol{\theta}}_{MG}\right)'\left(\frac{\mathbf{W}'_{ij}\mathbf{M}_H\mathbf{W}_{ij}}{T}\right),$$

$$\mathbf{S}_\theta = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\frac{\mathbf{W}'_{ij}\mathbf{M}_H\mathbf{W}_{ij}}{T}\right),\ \hat{\boldsymbol{\theta}}_{MG} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\hat{\boldsymbol{\theta}}_{ij},$$

where $\hat{\boldsymbol{\theta}}_{ij}$ is the (ij) pairwise OLS estimator obtained from the individual regression of $\mathbf{Y}_{ij}$ on $(\mathbf{W}_{ij}, \mathbf{H})$ in (263).

Next, we consider the 3D model (248) with the general errors:

$$u_{ijt} = \mu_{ij} + v_{it} + \zeta_{jt} + \pi_{ij}\lambda_t + \varepsilon_{ijt}. \tag{267}$$

The 3D-within transformation fails to remove $\pi_{ij}\lambda_t$, because

$$\tilde{u}_{ijt} = \tilde{\pi}_{ij}\tilde{\lambda}_t + \tilde{\varepsilon}_{ijt}$$

where $\tilde{\lambda}_t = \lambda_t - \bar{\lambda}$ with $\bar{\lambda} = T^{-1}\sum_{t=1}^{T}\lambda_t$ and $\tilde{\pi}_{ij} = \pi_{ij} - \pi_{.j} - \pi_{i.} + \pi_{..}$ with $\pi_{.j} = N_1^{-1}\sum_{i=1}^{N_1}\pi_{ij}$ and $\pi_{i.} = N_2^{-1}\sum_{j=1}^{N_2}\pi_{ij}$.[21] In the presence of the nonzero correlation between $\mathbf{x}_{ijt}$ and λ_t, the 3D-within estimator of $\boldsymbol{\beta}$ is biased.

We develop the two-step consistent estimation procedure. First, taking the cross-section averages of (**??**) over i and j,

$$\bar{y}_{..t} = \boldsymbol{\beta}'\bar{\mathbf{x}}_{..t} + \boldsymbol{\gamma}'\bar{\mathbf{s}}_{.t} + \boldsymbol{\delta}'\bar{\mathbf{d}}_{.t} + \boldsymbol{\kappa}'\mathbf{q}_t + \boldsymbol{\varphi}'\mathbf{z}_{..} + \mu_{..} + \bar{v}_{.t} + \bar{\zeta}_{.t} + \bar{\pi}_{..}\lambda_t + \bar{\varepsilon}_{..t} \quad (268)$$

where $\bar{v}_{.t} = N_1^{-1}\sum_{i=1}^{N_1}v_{it}$, $\bar{\zeta}_{.t} = N_2^{-1}\sum_{j=1}^{N_2}\zeta_{jt}$. We augment the model (248) with the cross-section averages:

$$y_{ijt} = \boldsymbol{\beta}'\mathbf{x}_{ijt} + \boldsymbol{\gamma}'\mathbf{s}_{it} + \boldsymbol{\delta}'\mathbf{d}_{jt} + \boldsymbol{\psi}_{ij}'\mathbf{f}_t + \tau_{ij} + \mu_{ij}^{*} + v_{ijt}^{*} + \zeta_{ijt}^{*} + \varepsilon_{ijt}^{*}, \quad (269)$$

where $v_{ijt}^{*} = v_{it} - \frac{\pi_{ij}\bar{v}_{.t}}{\bar{\pi}_{..}}$, $\zeta_{ijt}^{*} = \zeta_{jt} - \frac{\pi_{ij}\bar{\zeta}_{.t}}{\bar{\pi}_{..}}$.

We rewrite (269) as

$$y_{ijt} = \boldsymbol{\beta}'\mathbf{x}_{ijt} + \boldsymbol{\gamma}'\mathbf{s}_{it} + \boldsymbol{\delta}'\mathbf{d}_{jt} + \boldsymbol{\psi}_{ij}'\mathbf{f}_t + \tau_{ij} + \mu_{ij}^{*} + v_{it} + \zeta_{jt} + \varepsilon_{ijt}^{**}, \quad (270)$$

where $\varepsilon_{ijt}^{**} = \varepsilon_{ijt} - \frac{\pi_{ij}}{\bar{\pi}_{..}}\bar{\varepsilon}_{..t} - \frac{\pi_{ij}\bar{v}_{.t}}{\bar{\pi}_{..}} - \frac{\pi_{ij}\bar{\zeta}_{.t}}{\bar{\pi}_{..}}$. As $N_1, N_2 \to \infty$, $\varepsilon_{ijt}^{**} \to_p \varepsilon_{ijt}$.

We apply the 3D-within transformation (250) to (270):

$$\tilde{y}_{ijt} = \boldsymbol{\beta}'\tilde{\mathbf{x}}_{ijt} + \tilde{\boldsymbol{\psi}}_{ij}'\tilde{\mathbf{f}}_t + \tilde{\varepsilon}_{ijt}^{**}, \quad (271)$$

where $\tilde{\boldsymbol{\psi}}_{ij} = \boldsymbol{\psi}_{ij} - \boldsymbol{\psi}_{.j} - \boldsymbol{\psi}_{j.} + \boldsymbol{\psi}_{..}$, $\tilde{\mathbf{f}}_t = \mathbf{f}_t - \bar{\mathbf{f}}$ with $\bar{\mathbf{f}} = T^{-1}\sum_{t=1}^{T}\mathbf{f}_t$.Rewriting (271) compactly as

$$\tilde{\mathbf{Y}}_{ij} = \tilde{\mathbf{X}}_{ij}\boldsymbol{\beta} + \tilde{\mathbf{F}}\tilde{\boldsymbol{\psi}}_{ij} + \tilde{\mathbf{E}}_{ij}^{**},\ i = 1, ..., N_1, j = 1, ..., N_2 \quad (272)$$

$$\underset{T\times 1}{\tilde{\mathbf{Y}}_{ij}} = \begin{bmatrix} \tilde{y}_{ij1} \\ \vdots \\ \tilde{y}_{ijT} \end{bmatrix}, \underset{T\times k_x}{\tilde{\mathbf{X}}_{ij}} = \begin{bmatrix} \tilde{\mathbf{x}}_{ij1}' \\ \vdots \\ \tilde{\mathbf{x}}_{ijT}' \end{bmatrix}, \underset{T\times k_f}{\tilde{\mathbf{F}}} = \begin{bmatrix} \tilde{\mathbf{f}}_1' \\ \vdots \\ \tilde{\mathbf{f}}_T' \end{bmatrix}, \tilde{\mathbf{E}}_{ij}^{**} = \begin{bmatrix} \tilde{\varepsilon}_{ij1}^{**} \\ \vdots \\ \tilde{\varepsilon}_{ijT}^{**} \end{bmatrix}.$$

The 3D-PCCE estimator of $\boldsymbol{\beta}$ is obtained by

$$\hat{\boldsymbol{\beta}}_{PCCE} = \left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{X}}_{ij}\right)^{-1}\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{Y}}_{ij}\right) \quad (273)$$

where $\mathbf{M}_{\tilde{F}} = \mathbf{I}_T - \tilde{\mathbf{F}}\left(\tilde{\mathbf{F}}'\tilde{\mathbf{F}}\right)^{-1}\tilde{\mathbf{F}}'$ is the $T\times T$ idempotent matrix.As $(N_1, N_2, T) \to \infty$,

$$\sqrt{N_1N_2T}\left(\hat{\boldsymbol{\beta}}_{PCCE} - \boldsymbol{\beta}\right) \overset{a}{\sim} N\left(\mathbf{0}, \boldsymbol{\Sigma}_{\beta}\right), \quad (274)$$

[21] Unless $\tilde{\pi}_{ij} = 0$,$\tilde{u}_{ijt} \neq \tilde{\varepsilon}_{ijt}$. This holds only if factor loadings, π_{ij} are homogeneous.

where the (robust) consistent estimator of $\mathbf{\Sigma}_\beta$ is given by

$$\hat{\mathbf{\Sigma}}_\beta = \frac{1}{N^2}\mathbf{S}_\beta^{-1}\mathbf{R}_\beta\mathbf{S}_\beta^{-1}, \tag{275}$$

$$\mathbf{R}_\beta = \frac{1}{N_1N_2-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\frac{\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{X}}_{ij}}{T}\right)\left(\hat{\boldsymbol{\beta}}_{ij}-\hat{\boldsymbol{\beta}}_{MG}\right)\left(\hat{\boldsymbol{\beta}}_{ij}-\hat{\boldsymbol{\beta}}_{MG}\right)'\left(\frac{\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{X}}_{ij}}{T}\right),$$

$$\mathbf{S}_\beta = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\frac{\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{X}}_{ij}}{T}\right),\ \hat{\boldsymbol{\beta}}_{MG} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\hat{\boldsymbol{\beta}}_{ij},$$

where $\hat{\boldsymbol{\beta}}_{ij}$ is the (ij) pairwise OLS estimator from the individual regression of $\tilde{\mathbf{Y}}_{ij}$ on $\left(\tilde{\mathbf{X}}_{ij},\tilde{\mathbf{F}}\right)$ in (272).

We extend to the 3D panels with heterogeneous slope parameters:

$$y_{ijt} = \boldsymbol{\beta}_{ij}'\mathbf{x}_{ijt} + \boldsymbol{\gamma}_j'\mathbf{s}_{it} + \boldsymbol{\delta}_i'\mathbf{d}_{jt} + \boldsymbol{\kappa}_{ij}'\mathbf{q}_t + \boldsymbol{\varphi}'\mathbf{z}_{ij} + u_{ijt} \tag{276}$$

We develop the mean group estimators:

$$\hat{\boldsymbol{\beta}}_{W,MG} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\tilde{\mathbf{X}}_{ij}'\tilde{\mathbf{X}}_{ij}\right)^{-1}\left(\tilde{\mathbf{X}}_{ij}'\mathbf{Y}_{ij}\right) \tag{277}$$

$$\hat{\boldsymbol{\theta}}_{MGCCE} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\mathbf{W}_{ij}'\mathbf{M}_H\mathbf{W}_{ij}\right)^{-1}\left(\mathbf{W}_{ij}'\mathbf{M}_H\mathbf{Y}_{ij}\right) \tag{278}$$

$$\hat{\boldsymbol{\beta}}_{MGCCE} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left(\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{X}}_{ij}\right)^{-1}\left(\tilde{\mathbf{X}}_{ij}'\mathbf{M}_{\tilde{F}}\tilde{\mathbf{Y}}_{ij}\right) \tag{279}$$

9.2.2 Cross-section Dependence (CD) Test

The extent of CSD is captured by non-zero covariance between u_{ijt} and $u_{i'j't}$, which relates to rate at which $\frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sigma_{ijt,u}$ declines with N_1N_2. CTFE accommodates non-zero covariance locally, but imposes the same covariance for all $i=1,...,N_1$ and $j=1,...,N_2$. Such restrictions are too strong. Our proposed error components (267) accommodates non-zero covariances both locally and globally.

The diagnostic test for the null hypothesis of residual cross-section independence in the 3D panels using the residuals, $\mathbf{e}_{ij}=(e_{ij1},...,e_{ijT})'$. We have $\mathbf{e}_{ij}=\tilde{\mathbf{Y}}_{ij}-\tilde{\mathbf{X}}_{ij}\hat{\boldsymbol{\beta}}_W$ for the model (252), $\mathbf{e}_{ij}=\mathbf{M}_H\mathbf{Y}_{ij}-\mathbf{M}_H\mathbf{W}_{ij}\hat{\boldsymbol{\theta}}_{PCCE}$ for (263), and $\mathbf{e}_{ij}=\mathbf{M}_{\tilde{F}}\tilde{\mathbf{Y}}_{ij}-\mathbf{M}_{\tilde{F}}\tilde{\mathbf{X}}_{ij}\hat{\boldsymbol{\beta}}_{PCCE}$ for (272). The cross-section dependence (CD) test is a modified counterpart of an existing CD test by Pesaran (2015).

We compute the pair-wise residual correlations between n and n' cross-section units by

$$\hat{\rho}_{nn'} = \frac{\mathbf{e}_n'\mathbf{e}_{n'}}{\sqrt{\left(\mathbf{e}_n'\mathbf{e}_n\right)\left(\mathbf{e}_{n'}'\mathbf{e}_{n'}\right)}},\ n,n'=1,...,N_1N_2 \text{ and } n\neq n',$$

where we represent $\mathbf{e}_{ij}$ as the (ij) pair using the single index $n = 1, ..., N_1N_2$. We construct the CD statistic by

$$CD = \sqrt{\frac{2}{N_1N_2\left(N_1N_2 - 1\right)}} \sum_{n=1}^{N_1N_2-1} \sum_{n'=n+1}^{N_1N_2} \sqrt{T}\hat{\rho}_{nn'} \tag{280}$$

CD test has the limiting $N(0,1)$ distribution under the null $H_0 : \hat{\rho}_{nn'} = 0$ for all $n, n' = 1, ..., N_1N_2$ and $n \neq n'$ (Pesaran, 2015).

9.2.3 Monte Carlo Study

We construct DGP1 by

$$y_{ijt} = \beta' x_{ijt} + \mu_{ij} + \pi_{ij}\lambda_t + \varepsilon_{ijt}, \tag{281}$$

$$x_{ijt} = \mu^x_{ij} + \mu_{ij} + \pi^x_{ij}\lambda_t + v_{ijt}, \tag{282}$$

for $i = 1, ..., N_1$, $j = 1, ..., N_2$, and $t = 1, ..., T$. The global factor, λ_t and idiosyncratic errors, ε_{ijt} and v_{ijt} are generated independently as *iid* processes

$$\lambda_t \sim iidN\left(0,1\right),\ \varepsilon_{ijt} \sim iidN\left(0,1\right),\ v_{ijt} \sim iidN\left(0,1\right).$$

We generate pairwise individual effects independently as

$$\mu_{ij} \sim iidN\left(0,1\right),\ \mu^x_{ij} \sim iidN\left(0,1\right).$$

Factor loadings, π_{ij} and π^x_{ij}, are independently generated from $U[1,2]$.

Next, we construct DGP2 by

$$y_{ijt} = \beta' x_{ijt} + \mu_{ij} + v_{it} + \zeta_{jt} + \pi_{ij}\lambda_t + \varepsilon_{ijt}. \tag{283}$$

$$x_{ijt} = \mu^x_{ij} + \mu_{ij} + \pi^x_{ij}\lambda_t + v_{ijt},, \tag{284}$$

for $i = 1, ..., N_1$, $j = 1, ..., N_2$, and $t = 1, ..., T$. In addition, we generate v_{it} and ζ_{jt} independently as:

$$v_{it} \sim U\left(-1,1\right) \text{ and } \zeta_{jt} \sim U\left(-1,1\right)$$

In both DGP1 and DGP2 we set $\beta = 1$.

In Table 1 biases of the 2D PCCE and 3D PCCE estimators of β are mostly negligible even for $(N_1, N_2, T) = (25, 25, 50)$. The CTFE estimator displays substantial biases. RMSE results are qualitatively similar to the bias pattern. CTFE over-rejects the null in all cases and tends to 1 even as $N_1\,(N_2)$ or T rises. The size of the 2D PCCE is close to the nominal 5% while 3D PCCE slightly over-rejects when N_1 or N_2 is small. Overall performance of the 2D PCCE estimator is the best under DGP1.

Simulation results in Table 2 are qualitatively similar to those in Table 1. Biases of PCCE are almost negligible and their RMSEs decrease rapidly with $N_1\,(N_2)$ or T. Empirical sizes are still close to the nominal 5% level. CTFE suffers from substantial biases and size distortions, and its performance does not improve in large samples. Good performance of the 2D PCCE is rather surprising as the 3D PCCE estimator is expected to dominate. Overall simulation results support the simulation findings reported under the 2D panels by Kapetanios and Pesaran (2005) and Pesaran (2006).

9.2.4 The Gravity Model of the Intra-EU Trade

Anderson and van Wincoop (2003): "The gravity equation tells us that bilateral trade, after controlling for size, depends on the bilateral trade barriers but relative to the product of their Multilateral Resistance Indices (MTR)." Omitting MTR induces severe bias (e.g. Baldwin and Taglioni, 2006). Subsequent research focused on estimating the model with directional country-specific fixed effects to control for unobservable MTRs (e.g. Feenstra, 2004).

A large number of studies established an importance of taking into account multilateral resistance and bilateral heterogeneity in the 2D panels. Serlenga and Shin (2007) is the first to develop the panel gravity model by incorporating observed and unobserved factors. Behrens et al. (2012) develop the spatial econometric specification, to control for multilateral cross-sectional correlations across trade flows. Mastromarco et al. (2015) compare the factor- and the spatial-based gravity models to investigate the Euro impact on intra-EU trade flows over 1960-2008 for 190 country-pairs of 14 EU and 6 non-EU OECD countries. The CD test confirms that the factor-based model is more appropriate for controlling for CSD.

For the 3D models, we should control for source of biases presented by unobserved time-varying MTRs. Baltagi et al. (2003) propose the 3D model (251) with CTFE specification (**??**). This approach popular in measuring the impacts of MTRs of the exporters and the importers in the structural gravity studies (e.g. Baltagi et al., 2015). CTFE or CTRE estimators fail to accommodate (strong and heterogeneous) CSD. The presence of CSD across (ij) pairs suggests that the appropriate econometric techniques be required.

We apply our approach to the dataset covering the period 1960-2008 (49 years) for 182 country-pairs amongst 14 EU member countries (Austria, Belgium-Luxemburg, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, Sweden, United Kingdom).

Consider the generalised panel gravity specification:

$$\begin{aligned}\ln EXP_{ijt} = {} & \beta_0 + \beta_1 CEE_{ijt} + \beta_2 EMU_{ijt} + \beta_3 SIM_{ijt} + \beta_4 RLF_{ijt} \\ & + \beta_5 \ln GDP_{it} + \beta_6 \ln GDP_{jt} + \beta_7 RER_t \\ & + \gamma_1 DIS_{ij} + \gamma_2 BOR_{ij} + \gamma_3 LAN_{ij} + u_{ijt}\end{aligned} \tag{285}$$

The dependent variable EXP_{ijt} is the export flow from country i to country j at time t; CEE and EMU are dummies for European Community membership and European Monetary Union; SIM and RLF measure similarity in size and difference in relative factor endowments; RER represents the logarithm of common real exchange rates; GDP_{it} and GDP_{jt} are logged GDPs of exporter and importer; The logarithm of geographical distance (DIS) and the dummies for common language (LAN) and for common border (BOR) represent time-invariant bilateral barriers.

We report the CD test results for the residuals and the estimates of the CSD exponent (α). Our focus is on the impacts of t_{ij} that contain both barriers and incentives to trade. We focus on the two dummy variables; CEE (one when both

countries belong to the European Community); EMU (one when both adopt the same currency). Both are expected to exert a positive impact on bilateral export flows.

The empirical evidence is mixed. Rose (2001), Frankel and Rose (2002), Glick and Rose (2002) and Frankel (2008), document a huge positive effect; A number of studies report negative or insignificant effects (Persson, 2001, Pakko and Wall, 2002, De Nardis and Vicarelli, 2003). Recent studies by Serlenga and Shin (2007), Mastromarco et al. (2015) and Gunnella et al. (2015) finding a small but significant effect (7 to 10%) of the euro on intra-EU trade, after controlling for strong CSD.

Table 3 reports the estimation results. The two-way FE estimation results are statistically significant except RER. The impacts of home and foreign GDPs on exports are positive, but surprisingly, the former is twice larger than the latter. The impact of SIM is negative and significant, inconsistent with *a priori* expectations. CEE and EMU significantly boost exports, but their magnitudes seem to be too high. The CD test rejects the null of no or weak CSD convincingly. $\hat{\alpha}$ is 0.99 with CI containing unity; the residuals strongly correlated and the FE results biased and unreliable. This supports our main concern that upward trends in omitted trade determinants may cause them to be upward-biased.

We turn to the CTFE estimation results. CD test results indicate that the CTFE residuals do not suffer from any strong CSD. This rather surprising result is not supported by $\hat{\alpha} = 0.91$ (pretty close to 1). All the coefficients become insignificant except for CEE. The CEE is still substantial (0.29) while the EMU turns negligible (-0.011). Overall CTFE results are unreliable.

The 2D PCCE results are significant with the expected signs except for EMU. The impact of foreign GDP on exports is substantially larger than home GDP. The RER is positive, confirming that a depreciation of the home currency increases exports. The CEE is smaller (0.186), but EMU is insignificant and negligible (0.017). The 2D PCCE suffers from strong CSD residuals with $\hat{\alpha} = 0.87$.

Finally, the 3D PCCE results show that all the coefficients are significant with the expected signs. The CD test fails to strongly reject the null, supported by the smaller estimate of $\hat{\alpha} = 0.77$, close to a moderate range of weak CSD.[22] CEE still substantial (0.335) while the EMU modest at 0.081, close to the consensus reported in the 2D panel studies (e.g. Baldwin, 2006, Gunnella et al., 2015). The 3D PCCE results are mostly reliable, suggesting that the trade-boosting effect of the Euro should be viewed in the long-run historical and multilateral perspectives rather than simply focusing on the formation of a monetary union as an isolated event.

The CTFE estimator is proposed to capture (unobserved) multilateral resistance terms and trade costs, but it fails to accommodate strong CSD among MTRs, clearly present in our sample of the EU countries (confirmed by CD

[22] BHP show that the values of $\alpha \in [1/2, 3/4)$ represent a moderate degree of CSD.

Table 1: 3D panel gravity model estimation results for bilateral export flows

	FE			CTFE		
	Coeff	se	t-ratio	Coeff	se	t-ratio
gdph	2.185	0.041	52.97			
gdpf	1.196	0.041	28.98			
sim	-0.263	0.052	-5.069	-0.055	0.074	-0.754
rlf	0.031	0.006	5.011	0.006	0.005	1.294
rer	0.005	0.007	0.791	0.031	0.072	0.436
cee	0.302	0.014	22.05	0.290	0.017	16.99
emu	0.204	0.019	10.71	-0.011	0.036	-0.315
CD stat	620.1			-2.676		
	$\alpha_{0.05}$	α	$\alpha_{0.95}$	$\alpha_{0.05}$	α	$\alpha_{0.95}$
CSD exponent	0.925	0.992	1.059	0.865	0.914	0.963
	2D PCCE			3D PCCE		
	Coeff	se	t-ratio	Coeff	se	t-ratio
gdph	0.289	0.095	3.033			
gdpf	1.491	0.095	15.69			
sim	0.042	0.105	0.401	1.032	0.111	9.290
rlf	0.007	0.005	1.420	-0.004	0.005	-0.748
rer	0.144	0.019	7.427	0.168	0.114	1.471
cee	0.187	0.014	13.20	0.335	0.022	15.10
emu	0.018	0.015	1.160	0.081	0.045	1.793
CD stat	76.11			-4.19		
	$\alpha_{0.05}$	α	$\alpha_{0.95}$	$\alpha_{0.05}$	α	$\alpha_{0.95}$
CSD exponent	0.837	0.867	0.897	0.724	0.775	0.826

Notes: Using the annual dataset over 1960-2008 for 182 country-pairs amongst 14 EU member countries, we estimate the generalised panel gravity specification, (335). FE stands for the standard two-way fixed effects estimator with country-pair and time fixed effects. CTFE refers to the 3D within estimator given by (252). 2D PCCE is the PCCE estimator given by (263) with factors $\mathbf{f}_t = \left\{\overline{gdp}_{.t}, \overline{sim}_{..t}, \overline{rlf}_{..t}, \overline{cee}_{..t}, \overline{rer}_t, t\right\}$. 3D PCCE is the PCCE estimator given by (272) with factors $\mathbf{f}_t = \left\{\overline{sim}_{..t}, \overline{rlf}_{..t}, \overline{rer}_t\right\}$. CD test refers to testing the null hypothesis of residual cross-sectional error independence or weak dependence and is defined in (280). CSD exponent denotes the point estimate of the exponents of CSD α and the 90% level confidence bands.

tests and CSD exponent estimates). We should model the time-varying interdependency of bilateral export flows in a flexible manner than simply introducing deterministic country-time specific dummies. MTRs arise from the bilateral country-pair specific reactions to global shocks or the local spillover effects or both.

9.2.5 Conclusion

We propose novel estimation techniques to accommodate CSD within the 3D panel data models. Our framework is a generalisation of the multidimensional country-time fixed and random effects estimators. Our approach is the first attempt to introduce strong CSD into the multi-dimensional error components. We develop the two-step estimation procedure, the 3D-PCCE estimator. The empirical usefulness of the 3D-PCCE estimator is demonstrated via the Monte Carlo studies and the empirical application to the gravity model of the intra-EU trade.

Extensions and generalisations. First, we develop the multi-dimensional heterogenous panel data models with hierarchical multi-factor error structure (KSS). Next, we aim to develop the challenging models by combining the spatial- and the factor-based techniques. Bailey et al. (2016) develop the multi-step estimation procedure that can distinguish the relationship between spatial units from that which is due to the effect of common factors. Mastromarco et al. (2015) propose the technique for allowing weak and strong CSD in stochastic frontier panels by combining the exogenously driven factor-based approach and an endogenous threshold regime selection by Kapetanios et al. (2014, KMS). Bai and Li (2015) and Shi and Lee (2014,5) developed the framework for jointly modelling spatial effects and interactive effects. See also Gunnella et al. (2015) and Kuersteiner and Prucha (2015).

9.3 The Multi-dimensional Heterogeneous Panel Data with the Hierarchical Multi-factor Error Structure by Kapetanios, Serlenga and Shin (2017, KMS)

9.3.1 The Model

Consider the triple-index heterogeneous panel data model:

$$y_{ijt} = \boldsymbol{\beta}_{ij}' \boldsymbol{x}_{ijt} + \boldsymbol{\delta}_{ij}' \boldsymbol{d}_t + u_{ijt},\ i,j = 1,...,N,\ t = 1,...,T, \tag{286}$$

where y_{ijt} is the dependent variable observed across 3 indices, i the origin, j the destination at period t (say, the export from country i to j at t); $\boldsymbol{x}_{ijt}$ is the $m_x \times 1$ vector of covariates; $\boldsymbol{d}_t$ is the $m_d \times 1$ vector of observed common effects such as constants and trends. $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\delta}_{ij}$ are the $m_x \times 1$ and $m_d \times 1$ vectors of parameters.

We allow u_{ijt} to follow the hierarchical multi-factor structure:

$$u_{ijt} = \boldsymbol{\gamma}_{ij}' \boldsymbol{f}_t + \boldsymbol{\gamma}_{\circ j}' \boldsymbol{f}_{i\circ t} + \boldsymbol{\gamma}_{i\circ}' \boldsymbol{f}_{\circ jt} + \varepsilon_{ijt} \tag{287}$$

where $\boldsymbol{f}_t$, $\boldsymbol{f}_{\circ jt}$ and $\boldsymbol{f}_{i\circ t}$ are $m_f \times 1$, $m_{\circ\bullet} \times 1$ and $m_{\bullet\circ} \times 1$ vectors of unobserved common effects; ε_{ijt} are idiosyncratic errors distributed independently of $(\boldsymbol{x}_{ijt}, \boldsymbol{d}_t)$.

$\boldsymbol{f}_t$ are the global factors affecting all of the bilateral pairs; $\boldsymbol{f}_{i\circ t}$ and $\boldsymbol{f}_{\circ jt}$ are local origin i and destination j factors. They are designed to account for commonality in y_{ijt}; CSD between a given flow and a flow from the exporting region's commonality to the importing region (exporting-based dependence) and another flow from the exporting region to the importing region's commonality (importing-based dependence). This can provide a natural alternative to the existing literature. Business cycles can be decomposed into world, region and country-specific factors (Kose et al. 2003). See also Choi et al. (2016).

To deal with the general case where $\boldsymbol{f}_t$, $\boldsymbol{f}_{\circ jt}$ and $\boldsymbol{f}_{i\circ t}$, are correlated with $(\boldsymbol{x}_{ijt}, \boldsymbol{d}_t)$, we consider the following DGP for $\boldsymbol{x}_{ijt}$:

$$\boldsymbol{x}_{ijt} = \mathcal{D}_{ij}\boldsymbol{d}_t + \boldsymbol{\Gamma}_{ij}\boldsymbol{f}_t + \boldsymbol{\Gamma}_{\circ j}\boldsymbol{f}_{i\circ t} + \boldsymbol{\Gamma}_{i\circ}\boldsymbol{f}_{\circ jt} + \boldsymbol{v}_{ijt}, \tag{288}$$

where $\mathcal{D}_{ij}$ is the $(m_x \times m_d)$ parameter matrix, $\boldsymbol{\Gamma}_{ij}$, $\boldsymbol{\Gamma}_{\circ j}$ and $\boldsymbol{\Gamma}_{i\circ}$ are $(m_x \times m_f)$, $(m_x \times m_{\bullet\circ})$, $(m_x \times m_{\circ\bullet})$ factor loading matrices, and $\boldsymbol{v}_{ijt}$ are the idiosyncratic errors.

Combining (286)-(339), we have:

$$\boldsymbol{z}_{ijt} = \begin{pmatrix} y_{ijt} \\ \boldsymbol{x}_{ijt} \end{pmatrix} = \boldsymbol{\Xi}_{ij}\boldsymbol{d}_t + \boldsymbol{\Phi}_{ij}\boldsymbol{f}_t + \boldsymbol{\Phi}_{\circ j}\boldsymbol{f}_{i\circ t} + \boldsymbol{\Phi}_{i\circ}\boldsymbol{f}_{\circ jt} + \boldsymbol{u}_{ijt} \tag{289}$$

$$\boldsymbol{\Xi}_{ij} = \begin{pmatrix} \boldsymbol{\delta}'_{ij} + \boldsymbol{\beta}'_{ij}\mathcal{D}_{ij} \\ \mathcal{D}_{ij} \end{pmatrix}, \boldsymbol{\Phi}_{ij} = \begin{pmatrix} \boldsymbol{\gamma}'_{ij} + \boldsymbol{\beta}'_{ij}\boldsymbol{\Gamma}_{ij} \\ \boldsymbol{\Gamma}_{ij} \end{pmatrix}, \boldsymbol{\Phi}_{i\circ} = \begin{pmatrix} \boldsymbol{\gamma}'_{i\circ} + \boldsymbol{\beta}'_{ij}\boldsymbol{\Gamma}_{i\circ} \\ \boldsymbol{\Gamma}_{i\circ} \end{pmatrix}, \tag{290}$$

$$\boldsymbol{\Phi}_{\circ j} = \begin{pmatrix} \boldsymbol{\gamma}'_{\circ j} + \boldsymbol{\beta}'_{ij}\boldsymbol{\Gamma}_{\circ j} \\ \boldsymbol{\Gamma}_{\circ j} \end{pmatrix}, \boldsymbol{u}_{ijt} = \begin{pmatrix} \varepsilon_{ijt} + \boldsymbol{\beta}'_{ij}\boldsymbol{v}_{ijt} \\ \boldsymbol{v}_{ijt} \end{pmatrix}.$$

The ranks of $\boldsymbol{\Phi}_{ij}$, $\boldsymbol{\Phi}_{i\circ}$ and $\boldsymbol{\Phi}_{\circ j}$ determined by the ranks of

$$\underset{(m_x+1)\times m_f}{\tilde{\boldsymbol{\Gamma}}_{ij}} = \begin{pmatrix} \boldsymbol{\gamma}'_{ij} \\ \boldsymbol{\Gamma}_{ij} \end{pmatrix}, \underset{(m_x+1)\times m_{\circ\bullet}}{\tilde{\boldsymbol{\Gamma}}_{i\circ}} = \begin{pmatrix} \boldsymbol{\gamma}'_{i\circ} \\ \boldsymbol{\Gamma}_{i\circ} \end{pmatrix}, \underset{(m_x+1)\times m_{\bullet\circ}}{\tilde{\boldsymbol{\Gamma}}_{\circ j}} = \begin{pmatrix} \boldsymbol{\gamma}'_{\circ j} \\ \boldsymbol{\Gamma}_{\circ j} \end{pmatrix}.$$

Rewrite (286) and (340) in the matrix notation:

$$\boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{ij} + \boldsymbol{D}\boldsymbol{\delta}_{ij} + \boldsymbol{F}\boldsymbol{\gamma}_{ij} + \boldsymbol{F}_{i\circ}\boldsymbol{\gamma}_{\circ j} + \boldsymbol{F}_{\circ j}\boldsymbol{\gamma}_{i\circ} + \boldsymbol{\varepsilon}_{ij}, \tag{291}$$

$$\boldsymbol{z}_{ij} = \boldsymbol{D}\boldsymbol{\Xi}_{ij} + \boldsymbol{F}\boldsymbol{\Phi}_{ij} + \boldsymbol{F}_{i\circ}\boldsymbol{\Phi}_{\circ j} + \boldsymbol{F}_{\circ j}\boldsymbol{\Phi}_{i\circ} + \boldsymbol{u}_{ij}, \tag{292}$$

where

$$\underset{T\times 1}{\boldsymbol{y}_{ij}} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijT} \end{bmatrix}, \underset{T\times m_x}{\boldsymbol{X}_{ij}} = \begin{bmatrix} \boldsymbol{x}'_{ij1} \\ \vdots \\ \boldsymbol{x}'_{ijT} \end{bmatrix}, \underset{T\times m_d}{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{d}'_1 \\ \vdots \\ \boldsymbol{d}'_T \end{bmatrix}, \underset{T\times(m_x+1)}{\boldsymbol{z}_{ij}} = \begin{bmatrix} \boldsymbol{z}'_{ij1} \\ \vdots \\ \boldsymbol{z}'_{ijT} \end{bmatrix}, \tag{293}$$

$$\underset{T\times m_f}{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{f}'_1 \\ \vdots \\ \boldsymbol{f}'_T \end{bmatrix}, \underset{T\times m_{\bullet\circ}}{\boldsymbol{F}_{i\circ}} = \begin{bmatrix} \boldsymbol{f}'_{i\circ 1} \\ \vdots \\ \boldsymbol{f}'_{i\circ T} \end{bmatrix}, \underset{T\times m_{\circ\bullet}}{\boldsymbol{F}_{\circ j}} = \begin{bmatrix} \boldsymbol{f}'_{\circ j1} \\ \vdots \\ \boldsymbol{f}'_{\circ jT} \end{bmatrix}, \underset{T\times 1}{\boldsymbol{\varepsilon}_{ij}} = \begin{bmatrix} \varepsilon_{ij1} \\ \vdots \\ \varepsilon_{ijT} \end{bmatrix}, \underset{T\times(m_x+1)}{\boldsymbol{u}_{ij}} = \begin{bmatrix} \boldsymbol{u}'_{ij1} \\ \vdots \\ \boldsymbol{u}'_{ijT} \end{bmatrix}$$

Assumption 1. *Common Effects*: The $(m_d + m_f + m_{\bullet\circ} + m_{\circ\bullet}) \times 1$ vector of common factors $\boldsymbol{g}_t = \left(\boldsymbol{d}_t', \boldsymbol{f}_t', \boldsymbol{f}_{i\circ t}', \boldsymbol{f}_{\circ jt}'\right)'$, is covariance stationary with absolute summable autocovariances, distributed independently of $\varepsilon_{ijt'}$ and $\boldsymbol{v}_{ijt'}$ for all i, j, t and t'.

Assumption 2. *Individual-specific Errors*: ε_{ijt} and $\boldsymbol{v}_{ijt'}$ are distributed independently for all i, j, t and t', and they are distributed independently of $\boldsymbol{x}_{ijt}$ and $\boldsymbol{d}_t$.

Assumption 3. *Factor Loadings*: Unobserved factor loadings are independently and identically distributed across (i, j), and of ε_{ijt}, $\boldsymbol{v}_{ijt}$, $\boldsymbol{g}_t$ for all i, j, t, with finite means and variances. In particular,

$$\boldsymbol{\gamma}_{ij} = \boldsymbol{\gamma}_{\circ\circ} + \boldsymbol{\eta}_{ij}, \quad \boldsymbol{\gamma}_{i\circ} = \boldsymbol{\gamma}_{\bullet\circ} + \boldsymbol{\eta}_{i\circ}, \quad \boldsymbol{\gamma}_{\circ j} = \boldsymbol{\gamma}_{\circ\bullet} + \boldsymbol{\eta}_{\circ j}, \tag{294}$$

$$\boldsymbol{\Gamma}_{ij} = \boldsymbol{\Gamma}_{\circ\circ} + \boldsymbol{\xi}_{ij}, \quad \boldsymbol{\Gamma}_{i\circ} = \boldsymbol{\Gamma}_{\bullet\circ} + \boldsymbol{\xi}_{i\circ}, \quad \boldsymbol{\Gamma}_{\circ j} = \boldsymbol{\Gamma}_{\circ\bullet} + \boldsymbol{\xi}_{\circ j}, \tag{295}$$

where $\boldsymbol{\eta}_{ij} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\eta}_{\circ\circ}}\right)$, $\boldsymbol{\xi}_{ij} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\xi}_{\circ\circ}}\right)$, $\boldsymbol{\eta}_{i\circ} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\eta}_{\bullet\circ}}\right)$, $\boldsymbol{\xi}_{i\circ} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\xi}_{\bullet\circ}}\right)$, $\boldsymbol{\eta}_{\circ j} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\eta}_{\circ\bullet}}\right)$ and $\boldsymbol{\xi}_{\circ j} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\xi}_{\circ\bullet}}\right)$. Further, $\|\boldsymbol{\gamma}_{\circ\circ}\| < K$, $\|\boldsymbol{\gamma}_{\bullet\circ}\| < K$, $\|\boldsymbol{\gamma}_{\circ\bullet}\| < K$, $\|\boldsymbol{\Gamma}_{\circ\circ}\| < K$, $\|\boldsymbol{\Gamma}_{\bullet\circ}\| < K$, and $\|\boldsymbol{\Gamma}_{\circ\bullet}\| < K$ for positive constant $K < \infty$.

Assumption 4. *Random Slope Coefficients*:

$$\boldsymbol{\beta}_{ij} = \boldsymbol{\beta} + \boldsymbol{\nu}_{i\circ} + \boldsymbol{\nu}_{\circ j} + \boldsymbol{\nu}_{ij}, \boldsymbol{\nu}_{i\circ} \sim iid\left(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\bullet\circ}}\right), \boldsymbol{\nu}_{\circ j} \sim iid\left(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ\bullet}}\right), \boldsymbol{\nu}_{ij} \sim iid\left(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ\circ}}\right) \tag{296}$$

where $\|\boldsymbol{\beta}\| < K$ and $\boldsymbol{\nu}_{ij}$, $\boldsymbol{\nu}_{i\circ}$, $\boldsymbol{\nu}_{\circ j}$ are distributed independently of one another, and of $\boldsymbol{\gamma}_{ij}$, $\boldsymbol{\Gamma}_{ij}$, ε_{ijt}, $\boldsymbol{v}_{ijt}$ and $\boldsymbol{g}_t$ for all i, j and t.

Assumption 5. *Identification of $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\beta}$*: Construct the cross-section averages of $\boldsymbol{z}_{ijt}$ by

$$\bar{\boldsymbol{z}}_t = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{z}_{ijt}, \ \bar{\boldsymbol{z}}_{i\circ t} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{z}_{ijt} \text{ and } \bar{\boldsymbol{z}}_{\circ jt} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_{ijt} \tag{297}$$

Let $\bar{\boldsymbol{Z}}_{ij} = \left(\bar{\boldsymbol{Z}}, \bar{\boldsymbol{Z}}_{i\circ}, \bar{\boldsymbol{Z}}_{\circ j}\right)$ and $\bar{\boldsymbol{H}}_{ij} = \left(\boldsymbol{D}, \bar{\boldsymbol{Z}}_{ij}\right)$, where

$$\underset{T\times(m_x+1)}{\bar{\boldsymbol{Z}}} = \begin{bmatrix} \bar{\boldsymbol{z}}_1' \\ \vdots \\ \bar{\boldsymbol{z}}_T' \end{bmatrix}, \ \underset{T\times(m_x+1)}{\bar{\boldsymbol{Z}}_{i\circ}} = \begin{bmatrix} \bar{\boldsymbol{z}}_{i\circ 1}' \\ \vdots \\ \bar{\boldsymbol{z}}_{i\circ T}' \end{bmatrix}, \ \underset{T\times(m_x+1)}{\bar{\boldsymbol{Z}}_{i\circ}} = \begin{bmatrix} \bar{\boldsymbol{z}}_{\circ j1}' \\ \vdots \\ \bar{\boldsymbol{z}}_{\circ j1}' \end{bmatrix},$$

and construct the idempotent matrix:

$$\bar{\boldsymbol{M}}_{ij} = \boldsymbol{I}_T - \bar{\boldsymbol{H}}_{ij}\left(\bar{\boldsymbol{H}}_{ij}'\bar{\boldsymbol{H}}_{ij}\right)^{-1}\bar{\boldsymbol{H}}_{ij}' \tag{298}$$

(i) Identification of $\boldsymbol{\beta}_{ij}$: The $m_x \times m_x$ matrices, $\bar{\boldsymbol{\Psi}}_{ij,T} = T^{-1}\left(\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}\right)$ are nonsingular, and $\bar{\boldsymbol{\Psi}}_{ij,T}^{-1}$ have finite second-order moments.
(ii) Identification of $\boldsymbol{\beta}$: The $m_x \times m_x$ matrix, $\bar{\boldsymbol{\Psi}} = N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{\boldsymbol{\Psi}}_{ij,T}$ is nonsingular.

Remark 1: It is challenging to develop an appropriate model for accommodating CSD within the multilevel dataset. LeSage and Llano (2015) propose a spatial econometric methodology that introduces spatially-structured origin and destination effects, in such a way that regions treated as origins (destinations) exhibit similar effects to neighbors of the origins (destinations). Choi et al. (2016) develop a multilevel factor model with global and country factors, and propose a sequential principal component estimation procedure. KMSS address an important issue of controlling CSD in 3D panels by adding unobserved heterogeneous global factors to the CTFE specification, and propose the 3D PCCE estimator. The hierarchical multi-factor error model is more parsimonious and structural.

Remark 2: The weights are not necessarily unique. One could use the equal weight, $1/N$ for reasonably large N. Alternatively, the economic distance-based or time-varying measures could be considered.

Remark 3: The number of observed factors and the number of individual-specific regressors are fixed and known. The number of unobserved factors, $m = m_f + m_{\bullet\circ} + m_{\circ\bullet}$, is assumed fixed, but needs not to be known.

Represent hierarchical cross-section averages as follows:

$$\bar{\boldsymbol{z}}_t = \bar{\boldsymbol{\Xi}}_{\circ\circ}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}_{\circ\circ}\boldsymbol{f}_t + \bar{\boldsymbol{\Phi}}_{\circ\bullet}\boldsymbol{f}_{\bullet\circ t} + \bar{\boldsymbol{\Phi}}_{\bullet\circ}\boldsymbol{f}_{\circ\bullet t} + \bar{\boldsymbol{u}}_t, \tag{299}$$

$$\bar{\boldsymbol{z}}_{i\circ t} = \bar{\boldsymbol{\Xi}}_{i\circ}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}_{i\circ}\boldsymbol{f}_t + \bar{\boldsymbol{\Phi}}_{\circ\bullet}\boldsymbol{f}_{i\circ t} + \boldsymbol{\Phi}_{i\circ}\boldsymbol{f}_{\circ\bullet t} + \bar{\boldsymbol{u}}_{i\circ t}, \tag{300}$$

$$\bar{\boldsymbol{z}}_{\circ jt} = \bar{\boldsymbol{\Xi}}_{\circ j}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}_{\circ j}\boldsymbol{f}_t + \boldsymbol{\Phi}_{\circ j}\boldsymbol{f}_{\bullet\circ t} + \bar{\boldsymbol{\Phi}}_{\bullet\circ}\boldsymbol{f}_{\circ jt} + \bar{\boldsymbol{u}}_{\circ jt}, \tag{301}$$

$$\bar{\boldsymbol{\Xi}}_{\circ\circ} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Xi}_{ij}, \quad \bar{\boldsymbol{\Xi}}_{i\circ} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\Xi}_{ij}, \quad \bar{\boldsymbol{\Xi}}_{\circ j} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Xi}_{ij},$$

$$\bar{\boldsymbol{\Phi}}_{\circ\circ} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_{ij}, \quad \bar{\boldsymbol{\Phi}}_{i\circ} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_{ij}, \quad \bar{\boldsymbol{\Phi}}_{\circ j} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}_{ij}, \tag{302}$$

$$\bar{\boldsymbol{\Phi}}_{\circ\bullet} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_{\circ j} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_{\circ j}, \quad \bar{\boldsymbol{\Phi}}_{\bullet\circ} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_{i\circ} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}_{i\circ}, \tag{303}$$

$$\bar{\boldsymbol{u}}_t = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{u}_{ijt}, \quad \bar{\boldsymbol{u}}_{i\circ t} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{u}_{ijt}, \quad \bar{\boldsymbol{u}}_{\circ jt} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}_{ijt}$$

$$\boldsymbol{f}_{\circ\bullet t} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{f}_{\circ jt}, \quad \boldsymbol{f}_{\bullet\circ t} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{f}_{i\circ t}$$

Combining (299)-(301), we have:

$$\bar{\boldsymbol{z}}_{ijt} = \bar{\boldsymbol{\Xi}}_{ij}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}_{ij}\boldsymbol{f}_{ijt} + \bar{\boldsymbol{u}}_{ijt}, \tag{304}$$

where

$$\underset{3(m_x+1)\times 1}{\bar{\boldsymbol{z}}_{ijt}}=\begin{bmatrix}\bar{\boldsymbol{z}}_t\\ \bar{\boldsymbol{z}}_{i\circ t}\\ \bar{\boldsymbol{z}}_{\circ jt}\end{bmatrix},\ \underset{m\times 1}{\boldsymbol{f}_{ijt}}=\begin{bmatrix}\boldsymbol{f}_t\\ \boldsymbol{f}_{i\circ t}\\ \boldsymbol{f}_{\circ jt}\end{bmatrix},\bar{\boldsymbol{u}}_{ijt}=\begin{bmatrix}\bar{\boldsymbol{u}}_t+\bar{\boldsymbol{\Phi}}_{\circ\bullet}\boldsymbol{f}_{\bullet\circ t}+\bar{\boldsymbol{\Phi}}_{\bullet\circ}\boldsymbol{f}_{\circ\bullet t}\\ \bar{\boldsymbol{u}}_{i\circ t}+\boldsymbol{\Phi}_{i\circ}\boldsymbol{f}_{\circ\bullet t}\\ \bar{\boldsymbol{u}}_{\circ jt}+\boldsymbol{\Phi}_{\circ j}\boldsymbol{f}_{\bullet\circ t}\end{bmatrix}$$

$$\underset{3(m_x+1)\times m_d}{\bar{\boldsymbol{\Xi}}_{ij}}=\begin{bmatrix}\bar{\boldsymbol{\Xi}}_{\circ\circ}\\ \bar{\boldsymbol{\Xi}}_{i\circ}\\ \bar{\boldsymbol{\Xi}}_{\circ j}\end{bmatrix},\quad \underset{3(m_x+1)\times m}{\bar{\boldsymbol{\Phi}}_{ij}}=\begin{bmatrix}\bar{\boldsymbol{\Phi}}_{\circ\circ} & \mathbf{0} & \mathbf{0}\\ \bar{\boldsymbol{\Phi}}_{i\circ} & \bar{\boldsymbol{\Phi}}_{\circ\bullet} & \mathbf{0}\\ \bar{\boldsymbol{\Phi}}_{\circ j} & \mathbf{0} & \bar{\boldsymbol{\Phi}}_{\bullet\circ}\end{bmatrix},\tag{305}$$

with $m=m_f+m_{\circ\bullet}+m_{\bullet\circ}$.

Using (290) and (346), $\bar{\boldsymbol{\Phi}}_{ij}$ can be represented as follows:

$$\underset{(k+1)\times m_f}{\bar{\boldsymbol{\Phi}}_{\circ\circ}}=\tilde{\boldsymbol{B}}\tilde{\boldsymbol{\Gamma}}_{\circ\circ}+\begin{pmatrix}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\boldsymbol{\nu}_{i\circ}+\boldsymbol{\nu}_{\circ j}+\boldsymbol{\nu}_{ij}\right)'\boldsymbol{\Gamma}_{ij}\\ \mathbf{0}\end{pmatrix}\tag{306}$$

$$\underset{(k+1)\times m_{\circ\bullet}}{\bar{\boldsymbol{\Phi}}_{\circ\bullet}}=\tilde{\boldsymbol{B}}\tilde{\boldsymbol{\Gamma}}_{\circ\bullet}+\begin{pmatrix}\frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{\nu}_{i\circ}+\boldsymbol{\nu}_{\circ j}+\boldsymbol{\nu}_{ij}\right)'\boldsymbol{\Gamma}_{\circ j}\\ \mathbf{0}\end{pmatrix}\tag{307}$$

$$\underset{(k+1)\times m_{\bullet\circ}}{\bar{\boldsymbol{\Phi}}_{\bullet\circ}}=\tilde{\boldsymbol{B}}\tilde{\boldsymbol{\Gamma}}_{\bullet\circ}+\begin{pmatrix}\frac{1}{N}\sum_{=1}^{N}\left(\boldsymbol{\nu}_{i\circ}+\boldsymbol{\nu}_{\circ j}+\boldsymbol{\nu}_{ij}\right)'\boldsymbol{\Gamma}_{i\circ}\\ \mathbf{0}\end{pmatrix}\tag{308}$$

$$\underset{(k+1)\times m_f}{\bar{\boldsymbol{\Phi}}_{i\circ}}=\tilde{\boldsymbol{B}}\tilde{\boldsymbol{\Gamma}}_{i\circ}+\begin{pmatrix}\frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{\nu}_{i\circ}+\boldsymbol{\nu}_{\circ j}+\boldsymbol{\nu}_{ij}\right)'\boldsymbol{\Gamma}_{ij}\\ \mathbf{0}\end{pmatrix}\tag{309}$$

$$\underset{(k+1)\times m_f}{\bar{\boldsymbol{\Phi}}_{\circ j}}=\tilde{\boldsymbol{B}}\tilde{\boldsymbol{\Gamma}}_{\circ j}+\begin{pmatrix}\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\nu}_{i\circ}+\boldsymbol{\nu}_{\circ j}+\boldsymbol{\nu}_{ij}\right)'\boldsymbol{\Gamma}_{ij}\\ \mathbf{0}\end{pmatrix}\tag{310}$$

$$\tilde{B}=\begin{pmatrix}1 & \boldsymbol{\beta}'\\ 0 & \boldsymbol{I}_k\end{pmatrix},\tilde{\Gamma}_{\circ\circ}=\begin{pmatrix}\bar{\boldsymbol{\gamma}}'_{\circ\circ}\\ \bar{\boldsymbol{\Gamma}}_{\circ\circ}\end{pmatrix},\tilde{\Gamma}_{\circ\bullet}=\begin{pmatrix}\bar{\boldsymbol{\gamma}}'_{\circ\bullet}\\ \bar{\boldsymbol{\Gamma}}_{\circ\bullet}\end{pmatrix},\tilde{\Gamma}_{\bullet\circ}=\begin{pmatrix}\bar{\boldsymbol{\gamma}}'_{\bullet\circ}\\ \bar{\boldsymbol{\Gamma}}_{\bullet\circ}\end{pmatrix},\tilde{\Gamma}_{i\circ}=\begin{pmatrix}\bar{\boldsymbol{\gamma}}'_{i\circ}\\ \bar{\boldsymbol{\Gamma}}_{i\circ}\end{pmatrix},\tilde{\Gamma}_{\circ j}=\begin{pmatrix}\bar{\boldsymbol{\gamma}}'_{\circ j}\\ \bar{\boldsymbol{\Gamma}}_{\circ j}\end{pmatrix}$$

and $\bar{\boldsymbol{\Gamma}}_{\circ\circ}$, $\bar{\boldsymbol{\gamma}}_{\circ\circ}$, $\bar{\boldsymbol{\Gamma}}_{\circ\bullet}$, $\bar{\boldsymbol{\gamma}}_{\circ\bullet}$, $\bar{\boldsymbol{\Gamma}}_{\bullet\circ}$, $\bar{\boldsymbol{\gamma}}_{\bullet\circ}$, $\bar{\boldsymbol{\Gamma}}_{i\circ}$, $\bar{\boldsymbol{\gamma}}_{i\circ}$, $\bar{\boldsymbol{\Gamma}}_{\circ j}$ and $\bar{\boldsymbol{\gamma}}_{\circ j}$ are defined similarly to $\bar{\boldsymbol{\Phi}}_{\circ\circ}$, $\bar{\boldsymbol{\Phi}}_{i\circ}$, $\bar{\boldsymbol{\Phi}}_{\circ j}$, $\bar{\boldsymbol{\Phi}}_{\circ\bullet}$ and $\bar{\boldsymbol{\Phi}}_{\bullet\circ}$.

Suppose that the rank condition holds:

$$Rank\left(\bar{\boldsymbol{\Phi}}_{ij}\right)=m \text{ for all } (ij).\tag{311}$$

Then, we obtain from (304):

$$\boldsymbol{f}_{ijt}=\left(\bar{\boldsymbol{\Phi}}'_{ij}\bar{\boldsymbol{\Phi}}_{ij}\right)^{-1}\bar{\boldsymbol{\Phi}}'_{ij}\left(\bar{\boldsymbol{z}}_{ijt}-\bar{\boldsymbol{\Xi}}_{ij}\boldsymbol{d}_t-\bar{\boldsymbol{u}}_{ijt}\right)\tag{312}$$

It is easily seen that

$$\bar{\boldsymbol{u}}_{ijt}=O_p\left(\frac{1}{N}\right)+O_p\left(\frac{1}{\sqrt{NT}}\right) \text{ for each } t, \text{ as } N\to\infty.$$

Therefore,

$$\boldsymbol{f}_{ijt}-\left(\bar{\boldsymbol{\Phi}}'_{ij}\bar{\boldsymbol{\Phi}}_{ij}\right)^{-1}\bar{\boldsymbol{\Phi}}'_{ij}\left(\bar{\boldsymbol{z}}_{ijt}-\bar{\boldsymbol{\Xi}}_{ij}\boldsymbol{d}_t\right)=O_p\left(\frac{1}{N}\right)+O_p\left(\frac{1}{\sqrt{NT}}\right).$$

We can use $\bar{\boldsymbol{h}}_{ijt}=\left(\boldsymbol{d}'_t,\bar{\boldsymbol{z}}'_{ijt}\right)'$ as observable proxies for $\boldsymbol{f}_{ijt}$, and consistently estimate $\boldsymbol{\beta}_{ij}$ and their mean $\boldsymbol{\beta}$ by augmenting the regression, (286) with $\boldsymbol{d}_t$ and $\bar{\boldsymbol{z}}_{ijt}$. These are referred to as the 3DCCE estimators.

9.3.2 3D Common Correlated Effects Estimator

Individual Specific Coefficients The 3DCCE estimator of $\boldsymbol{\beta}_{ij}$ is given by

$$\hat{\boldsymbol{b}}_{ij} = \left(\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}\right)^{-1}\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{y}_{ij} \tag{313}$$

We show the dependence of $\hat{\boldsymbol{b}}_{ij}$ on the unobserved factors as:

$$\hat{\boldsymbol{b}}_{ij} - \boldsymbol{\beta}_{ij} = \left(\frac{\boldsymbol{X}_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)^{-1}\frac{\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{F}_{ij}}{T}\boldsymbol{\gamma}_{ij}^{*} + \left(\frac{\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)^{-1}\frac{\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{\varepsilon}_{ij}}{T} \tag{314}$$

$$= \left(\frac{\boldsymbol{X}_{ij}\boldsymbol{M}_{Q,ij}\boldsymbol{X}_{ij}}{T}\right)^{-1}\frac{\boldsymbol{X}_{ij}'\boldsymbol{M}_{Q,ij}\boldsymbol{F}_{ij}}{T}\boldsymbol{\gamma}_{ij}^{*} + \left(\frac{\boldsymbol{X}_{ij}'\boldsymbol{M}_{Q,ij}\boldsymbol{X}_{ij}}{T}\right)^{-1}\frac{\boldsymbol{X}_{ij}'\boldsymbol{M}_{Q,ij}\boldsymbol{\varepsilon}_{ij}}{T}$$

$$+ O_p\left(\frac{1}{N}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right)$$

where $\boldsymbol{F}_{ij} = (\boldsymbol{F}, \boldsymbol{F}_{i\circ}, \boldsymbol{F}_{\circ j})$, $\boldsymbol{\gamma}_{ij}^{*} = \left(\boldsymbol{\gamma}_{ij}', \boldsymbol{\gamma}_{i\circ}', \boldsymbol{\gamma}_{\circ j}'\right)'$and $\boldsymbol{M}_{Q,ij} = \boldsymbol{I}_T - \boldsymbol{Q}_{ij}\left(\boldsymbol{Q}_{ij}'\boldsymbol{Q}_{ij}\right)^{-1}\boldsymbol{Q}_{ij}'$ with $\boldsymbol{Q}_{ij} = \left(\boldsymbol{F}\bar{\boldsymbol{\Phi}}_{\circ\circ}', \boldsymbol{F}_{i\circ}\bar{\boldsymbol{\Phi}}_{\circ\bullet}', \boldsymbol{F}_{\circ j}\bar{\boldsymbol{\Phi}}_{\bullet\circ}'\right)$.

Suppose that the rank condition (311), is satisfied. Then,

Theorem 7 *Consider the triple-index heterogeneous panel data model, (286)-(339). Suppose that Assumptions 1-4 and 5(a) hold. Then, the 3DCCE estimator of the individual slope coefficients given by (313) is consistent. Further, as $N, T \to \infty$ and $T/N \to K < \infty$,*

$$\sqrt{T}\left(\hat{\boldsymbol{b}}_{ij} - \boldsymbol{\beta}_{ij}\right) \to^{d} N(0, \boldsymbol{V}_{ij}), \tag{315}$$

where $\boldsymbol{V}_{ij} = \boldsymbol{\Sigma}_{v,ij}^{-1}\boldsymbol{\Sigma}_{ij\varepsilon}\boldsymbol{\Sigma}_{v,ij}^{-1}$, $\boldsymbol{\Sigma}_{v,ij} = Var(\boldsymbol{v}_{ijt})$, $\boldsymbol{\Sigma}_{ij\varepsilon} = p\lim_{T\to\infty}\left[\frac{\boldsymbol{X}_{ij}'\boldsymbol{M}_{F,ij}\boldsymbol{\Omega}_{ij\varepsilon}\boldsymbol{M}_{F,ij}\boldsymbol{X}_{ij}}{T}\right]$, and $\boldsymbol{\Omega}_{ij\varepsilon} = E\left(\boldsymbol{\varepsilon}_{ij}'\boldsymbol{\varepsilon}_{ij}\right)$.

Remark: If the rank condition (311) does not hold, we need to show that $\frac{1}{T}\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\left(\boldsymbol{F}\boldsymbol{\gamma}_{ij} + \boldsymbol{F}_{i\circ}\boldsymbol{\gamma}_{\circ j} + \boldsymbol{F}_{\circ j}\boldsymbol{\gamma}_{i\circ}\right)$ converges to zero. We can establish that

$$\hat{\boldsymbol{b}}_{ij} - \boldsymbol{\beta}_{ij} = \left(\frac{\boldsymbol{X}_{ij}'\boldsymbol{M}_{Q,ij}\boldsymbol{X}_{ij}}{T}\right)^{-1}\frac{\boldsymbol{X}_{ij}'\boldsymbol{M}_{Q,ij}\boldsymbol{\varepsilon}_{ij}}{T} + o_p(1).$$

$\sqrt{T}\left(\hat{\boldsymbol{b}}_{ij} - \boldsymbol{\beta}_{ij}\right)$ will be asymptotically normal if $\sqrt{T}/N \to 0$ as $N, T \to \infty$.

3D Common Correlated Effects Mean Group Estimator The 3DCCEMG estimator is an average of the individual $\hat{\boldsymbol{b}}_{ij}$:

$$\hat{\boldsymbol{b}}_{MG} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\boldsymbol{b}}_{ij}, \tag{316}$$

Under Assumption 4 and using (314), we decompose $\sqrt{N}\left(\hat{\boldsymbol{b}}_{MG}-\boldsymbol{\beta}\right)$ and analyse each terms to obtain the following Theorem.

Theorem 8 *Consider the 3D model, (286)-(339). Suppose that Assumptions 1-4 and 5(a) hold. Then, the 3D CCEMG, $\hat{\boldsymbol{b}}_{MG}$ is consistent. As $N,T\to\infty$,*

$$\sqrt{N}\left(\bar{\boldsymbol{b}}_{MG}-\boldsymbol{\beta}\right)\to^{d}N(0,\boldsymbol{V}_{MG}),\boldsymbol{V}_{MG}=\boldsymbol{\Omega}_{\boldsymbol{\nu}\bullet\circ}+\boldsymbol{\Omega}_{\boldsymbol{\eta}\bullet\circ}+\boldsymbol{\Omega}_{\boldsymbol{\nu}\circ\bullet}+\boldsymbol{\Omega}_{\boldsymbol{\eta}\circ\bullet} \quad (317)$$

$$\Omega_{\boldsymbol{\nu}\bullet\circ}=\lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}E\left(\boldsymbol{A}_{1,i,NT}\boldsymbol{\Omega}_{\boldsymbol{\nu}_{i\circ}}\boldsymbol{A}_{1,i,NT}'\right),\Omega_{\boldsymbol{\eta}\bullet\circ}=\lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}E\left(\boldsymbol{A}_{2,i,NT}\boldsymbol{\Omega}_{\boldsymbol{\eta}_{i\circ}}\boldsymbol{A}_{2,i,NT}'\right)$$

$$\Omega_{\boldsymbol{\nu}\circ\bullet}=\lim_{N,T\to\infty}\frac{1}{N}\sum_{j=1}^{N}E\left(\boldsymbol{A}_{1,j,NT}\boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ j}}\boldsymbol{A}_{1,j,NT}'\right),\Omega_{\boldsymbol{\eta}\circ\bullet}=\lim_{N,T\to\infty}\frac{1}{N}\sum_{j=1}^{N}E\left(\boldsymbol{A}_{2,j,NT}\boldsymbol{\Omega}_{\boldsymbol{\eta}_{\circ j}}\boldsymbol{A}_{2,j,NT}'\right)$$

$\boldsymbol{V}_{MG}$ can be consistently estimated by

$$\hat{V}_{MG}=\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{b}}_{i}-\hat{\boldsymbol{b}}_{MG}\right)\left(\hat{\boldsymbol{b}}_{i}-\hat{\boldsymbol{b}}_{MG}\right)'+\frac{1}{N-1}\sum_{j=1}^{N}\left(\hat{\boldsymbol{b}}_{j}-\hat{\boldsymbol{b}}_{MG}\right)\left(\hat{\boldsymbol{b}}_{j}-\hat{\boldsymbol{b}}_{MG}\right)', \quad (318)$$

where $\hat{\boldsymbol{b}}_{i}=\frac{1}{N}\sum_{j=1}^{N}\hat{\boldsymbol{b}}_{ij}$ and $\hat{\boldsymbol{b}}_{j}=\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{b}}_{ij}$.

An important finding is that the dominant terms of $\sqrt{N}\left(\bar{\boldsymbol{b}}_{MG}-\boldsymbol{\beta}\right)$ are those that involve $\boldsymbol{\nu}_{i\circ}$, $\boldsymbol{\nu}_{i\circ}$, $\boldsymbol{\eta}_{i\circ}$ and $\boldsymbol{\eta}_{\circ j}$ only, because the terms associated with $\boldsymbol{\nu}_{ij}$ and $\boldsymbol{\eta}_{ij}$ are asymptotically negligible. This explains the $N^{1/2}$ rate of convergence. The nonparametric variance estimator $\frac{1}{N^{2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)'$ used by Pesaran (2006), is not consistent since it gives equal weights to the terms containing $\boldsymbol{\nu}_{i\circ}$, $\boldsymbol{\nu}_{i\circ}$, $\boldsymbol{\eta}_{i\circ}$ and $\boldsymbol{\eta}_{\circ j}$, and those containing $\boldsymbol{\nu}_{ij}$ and $\boldsymbol{\eta}_{ij}$. The consistent nonparametric estimator, $\hat{\boldsymbol{V}}_{MG}$ in (318) ensures that $\boldsymbol{\nu}_{ij}$ and $\boldsymbol{\eta}_{ij}$ are averaged out by the use of $\hat{\boldsymbol{b}}_{i}$ and $\hat{\boldsymbol{b}}_{j}$.

Remark Theorem 2 does not require the rank condition to hold as long as the number of unobserved factors m is fixed. We do not require any restriction on the relative rate of N and T.

3D Common Correlated Effects Pooled Estimator Consider the special case where $\boldsymbol{\beta}_{ij}$ are homogeneous, where efficiency gains from pooling can be achieved. We still allow the coefficients on observed and unobserved common effects to differ across (ij). We derive the pooled estimator of $\boldsymbol{\beta}$, referred to as the 3D CCEP estimator by

$$\hat{\boldsymbol{b}}_{P}=\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}\right)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{y}_{ij}, \quad (319)$$

Theorem 9 *Consider the 3D model, (286)-(339). Suppose that Assumptions 1-4 and 5(b) hold. Then,*

$$\sqrt{N}\left(\hat{\boldsymbol{b}}_P-\boldsymbol{\beta}\right)\rightarrow^d N(0,\boldsymbol{\Psi}^{-1}\boldsymbol{R}\boldsymbol{\Psi}^{-1}) \tag{320}$$

$$\boldsymbol{\Psi}=\lim_{N\rightarrow\infty}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Psi}_{ij}\ \textit{with}\ \boldsymbol{\Psi}_{ij}=E\left[\left(\frac{\boldsymbol{X}_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)^{-1}\right] \tag{321}$$

$$\boldsymbol{R}=\tilde{\boldsymbol{\Omega}}_{\boldsymbol{\nu}\bullet\circ}+\tilde{\boldsymbol{\Omega}}_{\boldsymbol{\eta}\bullet\circ}+\tilde{\boldsymbol{\Omega}}_{\boldsymbol{\nu}\circ\bullet}+\tilde{\boldsymbol{\Omega}}_{\boldsymbol{\eta}\circ\bullet}. \tag{322}$$

$$\tilde{\Omega}_{\boldsymbol{\nu}\bullet\circ}=\lim_{N,T\rightarrow\infty}\frac{1}{N}\sum_{i=1}^{N}E\left(\tilde{\boldsymbol{A}}_{1,i,NT}\boldsymbol{\Omega}'_{\boldsymbol{\nu}_{i\circ}}\tilde{\boldsymbol{A}}'_{1,i,NT}\right),\tilde{\Omega}_{\boldsymbol{\eta}\bullet\circ}=\lim_{N,T\rightarrow\infty}\frac{1}{N}\sum_{i=1}^{N}E\left(\tilde{\boldsymbol{A}}_{2,i,NT}\boldsymbol{\Omega}'_{\boldsymbol{\eta}_{i\circ}}\tilde{\boldsymbol{A}}_{2,i,NT}\right)$$

$$\tilde{\Omega}_{\boldsymbol{\nu}\circ\bullet}=\lim_{N,T\rightarrow\infty}\frac{1}{N}\sum_{j=1}^{N}E\left(\tilde{\boldsymbol{A}}_{1,j,NT}\boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ j}}\boldsymbol{A}'_{1,j,NT}\right),\tilde{\Omega}_{\boldsymbol{\eta}\circ\bullet}=\lim_{N,T\rightarrow\infty}\frac{1}{N}\sum_{j=1}^{N}E\left(\tilde{\boldsymbol{A}}_{2,j,NT}\boldsymbol{\Omega}_{\boldsymbol{\eta}_{\circ j}}\tilde{\boldsymbol{A}}'_{2,j,NT}\right)$$

where $\tilde{\boldsymbol{A}}_{1,i,NT}$, $\tilde{\boldsymbol{A}}_{2,i,NT}$, $\tilde{\boldsymbol{A}}_{1,j,NT}$ *and* $\tilde{\boldsymbol{A}}_{2,j,NT}$ *are defined in (??)-(??). The variance* $\boldsymbol{\Psi}^{-1}\boldsymbol{R}\boldsymbol{\Psi}^{-1}$ *can be consistently estimated by* $\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{R}}\hat{\boldsymbol{\Psi}}^{-1}$ *where*

$$\hat{\boldsymbol{\Psi}}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\boldsymbol{X}'_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}, \tag{323}$$

and

$$\hat{R}=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\boldsymbol{X}'_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)\right]\left[\frac{1}{N}\sum_{j=1}^{N}\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)'\left(\frac{\boldsymbol{X}'_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)\right]$$
$$+\frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\boldsymbol{X}'_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)\right]\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)'\left(\frac{\boldsymbol{X}'_{ij}\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)\right]$$

Remark The asymptotic variance matrix of $\hat{\boldsymbol{b}}_P$ depends on unobserved factors and loadings, but it is possible to estimate it consistently along lines similar to 3DCCEMG.

The Special Cases Better convergence rates can be achieved if the hierarchical structure is simplified. We focus on two special cases:

$$\text{Condition S1}:\boldsymbol{\eta}_{i\circ}=\boldsymbol{\eta}_{\circ j}=\boldsymbol{\nu}_{i\circ}=\boldsymbol{\nu}_{\circ j}=\boldsymbol{0}$$

$$\text{Condition S2}:\boldsymbol{F}_{i\circ}=\boldsymbol{F}_{\circ j}=0.$$

S2 is more restrictive and considered by KMSS.

Under Condition S2, the setup is similar to that of Pesaran (2006) because there is no hierarchical factor structure. We can treat the dataset as a $T \times N^2$ panel by amalgamating the two cross-section dimensions into one and applying the 2D CCE estimation procedure. The $\sqrt{N}$ rate will be replaced by N, and all the results of Pesaran (2006) and others analysing CCE estimator hold.

Next, consider the case where S1 holds but not S2. Then,

$$\sqrt{N}\left(\hat{\boldsymbol{b}}_{MG}-\boldsymbol{\beta}\right)=\frac{1}{N^{3/2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\nu}_{ij}+\frac{1}{N^{2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Psi}_{ijT}^{-1}\left(\frac{\sqrt{N}\boldsymbol{X}_{ij}'\boldsymbol{M}_{Q,ij}\boldsymbol{\varepsilon}_{ij}}{T}\right) \tag{324}$$

$$+\frac{1}{N^{3/2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\chi}_{ij,\circ\bullet}+O_p\left(\frac{1}{N}\right)+O_p\left(\frac{1}{\sqrt{NT}}\right)$$

From the proof of Theorem 2, the magnitude of all terms on the RHS of (324) is still N as long as $N/T \to 0$, since $\frac{1}{\sqrt{NT}}=o\left(\frac{1}{N}\right)$. Normality does not follow since the $O_p\left(\frac{1}{N}\right)$ term in RHS of (324) is not negligible. In this case, the asymptotic variance estimators in Pesaran (2006) become relevant only if normality holds. In particular,

$$\hat{\boldsymbol{V}}_{MG}=\frac{1}{N^{2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)\left(\hat{\boldsymbol{b}}_{ij}-\hat{\boldsymbol{b}}_{MG}\right)', \tag{325}$$

for the mean group estimator, and

$$\hat{\boldsymbol{R}}=\frac{1}{N^{2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right)\left(\bar{\boldsymbol{b}}_{ij}-\bar{\boldsymbol{b}}_{MG}\right)\left(\bar{\boldsymbol{b}}_{ij}-\bar{\boldsymbol{b}}_{MG}\right)'\left(\frac{\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}}{T}\right), \tag{326}$$

for the pooled estimator.

If Condition S1 is considered too restrictive - as it implies homogeneity for the coefficients of the factors in e_{ijt}, we may entertain the more general setup:

$$\boldsymbol{\gamma}_{i\circ}=\boldsymbol{\gamma}_{ij\circ}=\boldsymbol{\gamma}_{\bullet\circ}+\boldsymbol{\eta}_{ij\circ},\quad \boldsymbol{\gamma}_{\circ j}=\boldsymbol{\gamma}_{\circ ij}=\boldsymbol{\gamma}_{\circ\bullet}+\boldsymbol{\eta}_{\circ ij}.$$

Because of the double cross-sectional averaging, $\boldsymbol{\eta}_{ij\circ}$ and $\boldsymbol{\eta}_{\circ ij}$ are negligible since terms associated with $\boldsymbol{\chi}_{ij,\bullet\circ}$ and $\boldsymbol{\chi}_{ij,\circ\bullet}$ decay sufficiently fast to give the same fast convergence rate as under S1.

9.3.3 Monte Carlo Study

We generate y_{ijt} and x_{ijt} as follows:

$$y_{ijt}=\beta_{ij}x_{ijt}+\gamma_{1,ij}f_{1,t}+\gamma_{2,ij}f_{2,t}+\gamma_{1,\circ j}f_{1,i\circ t}+\gamma_{2,\circ j}f_{2,i\circ t}+\gamma_{1,i\circ}f_{1,\circ jt}+\gamma_{2,i\circ}f_{2,\circ jt}+\varepsilon_{ijt}, \tag{327}$$

$$x_{ijt} = \Gamma_{1,ij} f_{1,t} + \Gamma_{2,ij} f_{2,t} + \Gamma_{1,\circ j} f_{1,i\circ t} + \Gamma_{2,\circ j} f_{2,i\circ t} + \Gamma_{1,i\circ} f_{1,\circ jt} + \Gamma_{2,i\circ} f_{2,\circ jt} + v_{ijt}, \tag{328}$$

We set $m_d = 0$, $m_x = 1$ and $m_f = m_{\circ\bullet} = m_{\bullet\circ} = 2$.

$\boldsymbol{f}_t$, $\boldsymbol{f}_{\circ jt}$, $\boldsymbol{f}_{i\circ t}$ are generated independently as stationary AR processes with zero mean and unit variance:

$$f_{h,t} = \rho_{f_h} f_{h,t-1} + v_{f_h t} \text{ with } v_{f_h t} \sim iidN\left(0, 1-\rho_{f_h}^2\right) \text{ for } h = 1, 2$$

$$f_{h,i\circ t} = \rho_{f_{h,i\circ}} f_{h,i\circ,t-1} + v_{f_{h,i\circ} t} \text{ with } v_{f_{h,i\circ} t} \sim iidN\left(0, 1-\rho_{f_{h,i\circ}}\right) \text{ for } h = 1, 2$$

$$f_{h,\circ jt} = \rho_{f_{h,\circ j}} f_{h,\circ j,t-1} + v_{f_{h,\circ j} t} \text{ with } v_{f_{h,\circ j} t} \sim iidN\left(0, 1-\rho_{f_{h,\circ j}}^2\right) \text{ for } h = 1, 2$$

ε_{ijt} and v_{ijt}, are generated independently as

$$\varepsilon_{ijt} = \rho_\varepsilon \varepsilon_{ij,t-1} + e_{\varepsilon,ijt} \text{ with } e_{\varepsilon,ijt} \sim iidN\left(0, 1-\rho_\varepsilon^2\right)$$

$$v_{ijt} = \rho_v v_{ij,t-1} + e_{v,ijt} \text{ with } e_{v,ijt} \sim iidN\left(0, 1-\rho_v^2\right)$$

We set $\rho_{f_h} = \rho_{f_{h,i\circ}} = \rho_{f_{h,\circ j}} = \rho_\varepsilon = \rho_v = \{0, 0.5\}$.

2 experiments: Experiment A with the full rank and Experiment B with the rank condition (311) violated. For x_{ijt} in (328), we draw the factor loadings independently by

$$\Gamma_{1,ij} \sim iidN\left(0.5, 0.5\right) \text{ and } \Gamma_{2,ij} \sim iidN\left(0, 0.5\right) \text{ for } i, j = 1, ..., N$$

$$\Gamma_{1,\circ j} \sim iidN\left(0.5, 0.5\right) \text{ and } \Gamma_{2,\circ j} \sim iidN\left(0, 0.5\right) \text{ for } j = 1, ..., N$$

$$\Gamma_{1,i\circ} \sim iidN\left(0.5, 0.5\right) \text{ and } \Gamma_{2,i\circ} \sim iidN\left(0, 0.5\right) \text{ for } i = 1, ..., N$$

For y_{ijt} in (327), we consider two experiments. For experiment A,

$$\gamma_{1,ij} \sim iidN\left(1, 0.2\right) \text{ and } \gamma_{2,ij} \sim iidN\left(1, 0.2\right) \text{ for } i, j = 1, ..., N$$

$$\gamma_{1,\circ j} \sim iidN\left(1, 0.2\right) \text{ and } \gamma_{2,\circ j} \sim iidN\left(1, 0.2\right) \text{ for } j = 1, ..., N$$

$$\gamma_{1,i\circ} \sim iidN\left(1, 0.2\right) \text{ and } \gamma_{2,i\circ} \sim iidN\left(1, 0.2\right) \text{ for } i = 1, ..., N.$$

For experiment B, we generate:

$$\gamma_{1,ij} \sim iidN\left(1, 0.2\right) \text{ and } \gamma_{2,ij} \sim iidN\left(0, 1\right) \text{ for } i, j = 1, ..., N$$

$$\gamma_{1,\circ j} \sim iidN\left(1, 0.2\right) \text{ and } \gamma_{2,\circ j} \sim iidN\left(0, 1\right) \text{ for } j = 1, ..., N$$

$$\gamma_{1,i\circ} \sim iidN\left(1, 0.2\right) \text{ and } \gamma_{2,i\circ} \sim iidN\left(0, 1\right) \text{ for } i = 1, ..., N$$

Consider Case 1 with the heterogeneous slopes:

$$\beta_{ij} = \beta + \nu_{i\circ} + \nu_{\circ j} + \nu_{ij}, \nu_{i\circ} \sim iidN\left(0, 1\right), \nu_{\circ j} \sim iidN\left(0, 1\right), \nu_{ij} \sim iidN\left(0, 1\right)$$

and Case 2 with the homogeneous slopes $\beta_{ij} = \beta = 1$.

We consider the two-way within estimator with $u_{ijt} = \alpha_{ij} + \theta_t + \varepsilon_{ijt}$, and the three versions of 3D estimators: the $3DCCE_G$ with $u_{ijt} = \alpha_{ij} + \boldsymbol{\gamma}_{ij}' \boldsymbol{f}_t + \varepsilon_{ijt}$

where we approximate the heterogenous global factors only by $\bar{\boldsymbol{z}}_t = (\bar{y}_t, \bar{x}_t)'$; the $3DCCE_L$ estimator with $u_{ijt} = \alpha_{ij} + \boldsymbol{\gamma}'_{\circ j}\boldsymbol{f}_{i\circ t} + \boldsymbol{\gamma}'_{i\circ}\boldsymbol{f}_{\circ jt} + \varepsilon_{ijt}$ where we approximate the heterogenous local factors only by $\bar{\boldsymbol{z}}_{i\circ t}$ and $\bar{\boldsymbol{z}}_{\circ jt}$ and the $3DCCE_{GL}$ estimator with $u_{ijt} = \alpha_{ij} + \boldsymbol{\gamma}'_{\circ j}\boldsymbol{f}_{i\circ t} + \boldsymbol{\gamma}'_{i\circ}\boldsymbol{f}_{\circ jt} + \boldsymbol{\gamma}'_{ij}\boldsymbol{f}_t + \varepsilon_{ijt}$ where we approximate the heterogenous global and local factors by $\bar{\boldsymbol{z}}_t$, $\bar{\boldsymbol{z}}_{i\circ t}$ and $\bar{\boldsymbol{z}}_{\circ jt}$. We consider both mean group and pooled estimators. We report the bias, the root mean squared error and coverage rates at the 95% confidence with 1,000 replications for (N, T) pairs with $N = \{10, 25, 100\}$ and $T = \{50, 100\}$.

Table 1: simulation results for Experiment A (the full rank) with heterogeneous coefficients (Case 1). The biases of $3DCCE_{GL}$ are mostly negligible even for the relatively small samples. The performance of both pooled and mean group estimators is almost identical. Both FE and $3DCCE_G$ estimators suffer from severe biases. The biases of the $3DCCE_L$ are much smaller than those of the $3DCCE_G$, showing that the local factor approximations seem to be more effective than the global counterpart, though they are still non-negligible even for large N and T. The CCE estimator advanced by Pesaran (2006) in 2D panels fail to remove correlations between local factors and regressors. These provide strong support for our theoretical predictions that the joint approximations of the heterogenous global and local factors can only provide consistent estimation of $E(\beta)$ in the presence of the hierarchical multifactors. We find the similar patterns of RMSE. The RMSEs of $3DCCE_L$ and $3DCCE_{GL}$ estimators are significantly lower than those of FE and $3DCCE_G$ estimators. The difference between $3DCCE_L$ and $3DCCE_{GL}$ is mostly negligible, but the RMSEs of $3DCCE_{GL}$ tends to decline slightly faster with sample sizes. Turning to the coverage rates, $3DCCE_L$ and the $3DCCE_{GL}$ estimators perform better than FE and $3DCCE_G$ estimators. Coverage rates of $3DCCE_{GL}$ estimator only tend to the nominal 95% as N or T rises.

Table 2 presents simulation results for Experiment A (the full rank) with homogeneous coefficients (Case 2). We find qualitatively similar results for the biases to Table 1, confirming that the $3DCCE_{GL}$ estimator is most reliable. RMSEs of FE and $3DCCE_G$ estimators are significantly higher than those of $3DCCE_L$ and $3DCCE_{GL}$ estimators. RMSEs of $3DCCE_{GL}$ is significantly lower than those of $3DCCE_L$, but they also fall sharply with sample sizes. The relative performance of both pooled and mean group estimators is qualitatively similar. Surprisingly, all the estimators produce unsatisfactory coverage rates. FE, $3DCCE_G$ and $3DCCE_L$ estimators tend to under-estimate coverage substantially even as N rises whilst $3DCCE_{GL}$ over-estimate it.

Table 3 presents simulation results for Experiment B (the rank deficiency) with heterogeneous coefficients (Case 1). We find qualitatively similar results to Table 1; the performance of the estimators are not affected significantly by the rank deficiency; confirming that the $3DCCE_{GL}$ estimator is most reliable. Table 4 presents simulation results for Experiment B (the rank deficiency) with homogeneous coefficients (Case 2). We find qualitatively similar results to Table 2, and conclude that the $3DCCE_{GL}$ estimator is still most reliable, though it tends to over-estimate coverage rates.

We conduct the additional simulations under Conditions S1 and S2 described

in Section 3.4 We find the results confirming that the faster convergence rates are achieved in both cases (available on online supplement).

Table 2: Simulation results for Case 1 - Full Rank (Experiment A)

		FE		$3DCCE_G$		$3DCCE_L$		$3DCCE_{LG}$	
		Panel A: Bias							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.203	0.216	0.232	0.251	0.064	0.079	-0.006	0.008
0	25	0.250	0.217	0.263	0.272	0.062	0.07	-0.004	0.004
	50	0.227	0.220	0.285	0.279	0.077	0.072	0.008	0.002
	100	0.222	0.224	0.282	0.285	0.074	0.076	0.002	0.003
	10	0.233	0.210	0.302	0.253	0.115	0.067	0.04	-0.008
0.5	25	0.220	0.236	0.265	0.289	0.049	0.074	-0.022	0.004
	50	0.251	0.251	0.301	0.29	0.079	0.068	0.005	-0.006
	100	0.244	0.245	0.300	0.300	0.077	0.078	-0.002	-0.001
		Mean Group Estimator							
	10	0.201	0.212	0.219	0.239	0.046	0.063	0.001	0.017
0	25	0.242	0.211	0.249	0.258	0.053	0.061	0.003	0.012
	50	0.219	0.213	0.268	0.264	0.067	0.062	0.011	0.007
	100	0.214	0.216	0.265	0.268	0.062	0.065	0.002	0.005
	10	0.227	0.205	0.292	0.247	0.103	0.056	0.053	0.006
0.5	25	0.214	0.230	0.254	0.276	0.041	0.063	-0.012	0.010
	50	0.241	0.241	0.287	0.276	0.068	0.056	0.009	-0.003
	100	0.236	0.237	0.286	0.287	0.064	0.065	-0.001	0.000
		Panel B: RMSE							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.522	0.524	0.544	0.53	0.488	0.474	0.483	0.468
0	25	0.377	0.361	0.389	0.396	0.293	0.296	0.286	0.287
	50	0.300	0.292	0.348	0.345	0.213	0.216	0.198	0.203
	100	0.263	0.262	0.316	0.316	0.160	0.156	0.141	0.137
	10	0.518	0.511	0.557	0.536	0.479	0.475	0.466	0.469
0.5	25	0.366	0.373	0.390	0.412	0.290	0.300	0.288	0.291
	50	0.325	0.322	0.369	0.353	0.225	0.211	0.211	0.200
	100	0.283	0.284	0.332	0.332	0.162	0.163	0.142	0.143
		Mean Group Estimator							
	10	0.513	0.520	0.528	0.523	0.480	0.469	0.478	0.464
0	25	0.370	0.356	0.379	0.385	0.289	0.292	0.284	0.286
	50	0.293	0.288	0.333	0.332	0.208	0.212	0.197	0.203
	100	0.256	0.255	0.301	0.301	0.154	0.151	0.141	0.137
	10	0.507	0.503	0.544	0.525	0.466	0.463	0.458	0.460
0.5	25	0.360	0.369	0.382	0.400	0.287	0.297	0.285	0.290
	50	0.317	0.314	0.356	0.340	0.220	0.207	0.209	0.199
	100	0.275	0.277	0.319	0.320	0.155	0.157	0.141	0.143
		Panel C: Coverage rate at 95 confidence level							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.973	0.969	0.965	0.953	0.903	0.903	0.895	0.886
0	25	0.960	0.965	0.951	0.947	0.924	0.919	0.924	0.929
	50	0.941	0.948	0.911	0.898	0.922	0.919	0.94	0.932
	100	0.881	0.876	0.792	0.775	0.917	0.924	0.947	0.959
	10	0.977	0.968	0.965	0.973	0.898	0.912	0.892	0.890
0.5	25	0.963	0.957	0.96	0.952	0.925	0.931	0.935	0.927
	50	0.932	0.928	0.899	0.907	0.914	0.927	0.935	0.945
	100	0.855	0.836	0.790	0.745	0.915	0.912	0.952	0.947
		Mean Group Estimator							
	10	0.943	0.934	0.951	0.939	0.891	0.900	0.888	0.891
0	25	0.914	0.910	0.931	0.928	0.923	0.914	0.925	0.931
	50	0.818	0.810	0.876	0.863	0.933	0.925	0.946	0.931
	100	0.842	0.830	0.724	0.713	0.929	0.929	0.947	0.961
	10	0.942	0.945	0.954	0.954	0.904	0.908	0.903	0.901
0.5	25	0.912	0.904	0.936	0.926	0.932	0.934	0.935	0.936
	50	0.799	0.798	0.849	0.850	0.917	0.937	0.933	0.949
	100	0.805	0.779	0.691	0.677	0.926	0.920	0.949	0.946

Notes:FE is the the two-way within estimator, $3DCCE_G$ is the 3D CCE estimator with the global factors approximation only, $3DCCE_L$ is the 3D CCE estimator with the local factors approximation only, and $3DCCE_{GL}$ is the 3D CCE estimator with both global and local factors approximation. $3DCCE$ estimators are defined in (316) and (319). We consider both mean group and pooled estimators. The variance of $3DCCE_G$ is estimated by (325) for the mean group and (326) for the pooled estimator. The variances of $3DCCE_L$ and $3DCCE_{GL}$ are given by (318) for the mean group and (323)-(**??**) for the pooled estimator.

9.3.4 Empirical Application

Anderson and van Wincoop (2003) show that bilateral trade depends on the bilateral trade barriers but relative to the product of their Multilateral Resistance Indices: bilateral barrier relative to average trade barriers that both regions face with all their trading partners. They derive the following system of the structural gravity equations:

$$X_{ij} = \frac{Y_i Y_j}{Y} \left(\frac{t_{ij}}{\Pi_i P_j} \right)^{1-\sigma} \tag{329}$$

$$\Pi_i^{1-\sigma} = \sum_j \left(\frac{t_{ij}}{P_j} \right)^{1-\sigma} \frac{Y_j}{Y} \text{ and } P_j^{1-\sigma} = \sum_i \left(\frac{t_{ij}}{\Pi_i} \right)^{1-\sigma} \frac{Y_i}{Y} \tag{330}$$

where X_{ij} are exports from i to j, Y_i, Y_j and Y are GPD for i (exporter), j (importer) and the world, $t_{ij}\,(>1)$ is one plus the tariff equivalent of trade costs of imports of j from i, $\sigma\,(>1)$ is the elasticity of substitution with CES preference, Π_i is ease of access of exporter i, and P_j is the ease of access of importer j. P_j and Π_i are called inward and outward multilateral resistance. Omitting MTR induces potentially severe bias.

Consider the log-linearised specification of (329):

$$\ln X_{ij} = \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_j + \beta_3 \ln t_{ij} + \beta_4 \ln P_i + \beta_5 \ln P_j + \varepsilon_{ij} \tag{331}$$

where P_i and P_j are unobservable MTRs, and t_{ij} contain both barriers and incentives to trade between i and j. Subsequent research has focused on estimating (331) with replacing unobservable MTRs by N country-specific dummies, μ_i and μ_j. We extend (331) into 3D panels:

$$\ln X_{ijt} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 \ln Y_{jt} + \beta_3 \ln t_{ijt} + \beta_4 \ln P_{it} + \beta_5 \ln P_{jt} + \varepsilon_{ijt}, \tag{332}$$

where we should allow MTRs to vary over time. Baltagi et al. (2003) propose:

$$u_{ijt} = \alpha_{ij} + \theta_{it} + \theta^*_{jt} + \varepsilon_{ijt}, \tag{333}$$

which contains bilateral pair-fixed effects α_{ij} as well as origin (exporter) and destination (importer) country-time fixed effects (CTFE) θ_{it} and θ^*_{jt}. This approach popular in measuring the impacts of MTRs of exporters and importers in the structural gravity studies.

The main drawback of the CTFE approach lies in the assumption that bilateral trade flows are independent of what happens to the rest of the trading world. Recently, KMSS extend the 3D panel data model (286) with the more general error components:

$$u_{ijt} = \alpha_{ij} + \theta_{it} + \theta^*_{jt} + \pi_{ij}\theta_t + \varepsilon_{ijt}, \tag{334}$$

that attempts to model residual CSD via unobserved heterogeneous global factor θ_t in addition to CTFEs. The CTFE estimator is biased because it fails

Table 3: Simulation Results for Experiment A (Full Rank) with Homogeneous Coefficients

		FE		$3DCCE_G$		$3DCCE_L$		$3DCCE_{LG}$	
		Panel A: Bias							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.204	0.199	0.240	0.246	0.067	0.073	-0.001	0.006
0	25	0.214	0.215	0.269	0.270	0.068	0.068	0.001	0.002
	50	0.218	0.219	0.278	0.278	0.071	0.070	0.001	0.001
	100	0.220	0.220	0.281	0.282	0.074	0.074	0.000	0.000
	10	0.228	0.225	0.267	0.261	0.076	0.072	0.002	-0.003
0.5	25	0.240	0.245	0.292	0.292	0.075	0.075	0.003	0.004
	50	0.245	0.245	0.300	0.299	0.077	0.076	0.002	0.002
	100	0.248	0.248	0.304	0.303	0.081	0.080	0.001	0.000
		Mean Group Estimator							
	10	0.202	0.198	0.230	0.235	0.057	0.062	0.005	0.011
0	25	0.207	0.208	0.255	0.256	0.059	0.060	0.006	0.006
	50	0.211	0.212	0.262	0.262	0.060	0.060	0.002	0.003
	100	0.212	0.212	0.265	0.265	0.061	0.062	0.000	0.001
	10	0.224	0.223	0.256	0.252	0.067	0.064	0.010	0.006
0.5	25	0.233	0.236	0.279	0.278	0.066	0.065	0.008	0.008
	50	0.236	0.237	0.286	0.286	0.066	0.066	0.004	0.004
	100	0.239	0.239	0.289	0.289	0.068	0.067	0.002	0.001
		Panel B: RMSE							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.236	0.230	0.268	0.273	0.129	0.136	0.110	0.112
0	25	0.220	0.220	0.274	0.274	0.083	0.082	0.045	0.044
	50	0.219	0.220	0.279	0.279	0.074	0.074	0.022	0.021
	100	0.221	0.220	0.282	0.282	0.075	0.075	0.010	0.010
	10	0.257	0.256	0.293	0.289	0.136	0.137	0.113	0.114
0.5	25	0.245	0.250	0.296	0.296	0.090	0.088	0.046	0.045
	50	0.247	0.247	0.301	0.300	0.081	0.080	0.022	0.022
	100	0.249	0.249	0.304	0.303	0.082	0.081	0.011	0.010
		Mean Group Estimator							
	10	0.230	0.225	0.255	0.259	0.118	0.121	0.103	0.104
0	25	0.212	0.212	0.259	0.260	0.073	0.073	0.042	0.042
	50	0.212	0.213	0.263	0.263	0.063	0.064	0.020	0.021
	100	0.213	0.212	0.265	0.265	0.062	0.063	0.010	0.010
	10	0.249	0.249	0.279	0.276	0.123	0.122	0.103	0.103
0.5	25	0.238	0.240	0.283	0.282	0.078	0.077	0.042	0.041
	50	0.237	0.238	0.286	0.287	0.070	0.069	0.021	0.021
	100	0.239	0.239	0.289	0.289	0.069	0.068	0.011	0.010
		Panel C: Coverage rate at 95 confidence level							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.997	1.000	1.000	1.000	0.947	0.940	0.964	0.971
0	25	0.994	0.997	1.000	1.000	0.866	0.864	0.985	0.985
	50	0.686	0.661	0.988	0.987	0.376	0.363	0.992	0.993
	100	0.812	0.760	0.001	0.000	0.000	0.000	0.999	0.995
	10	0.998	1.000	1.000	1.000	0.944	0.938	0.976	0.973
0.5	25	0.993	0.992	1.000	1.000	0.838	0.839	0.986	0.990
	50	0.666	0.584	0.970	0.955	0.314	0.284	0.990	0.990
	100	0.326	0.162	0.000	0.000	0.000	0.000	0.989	0.995
		Mean Group Estimator							
	10	0.999	0.998	1.000	1.000	0.958	0.949	0.963	0.970
0	25	0.992	0.994	1.000	1.000	0.886	0.870	0.987	0.987
	50	0.438	0.427	0.973	0.973	0.426	0.416	0.989	0.995
	100	0.488	0.460	0.000	0.000	0.001	0.000	0.999	0.991
	10	0.996	0.997	1.000	1.000	0.949	0.944	0.976	0.971
0.5	25	0.988	0.984	1.000	1.000	0.853	0.853	0.989	0.987
	50	0.344	0.264	0.892	0.861	0.331	0.315	0.985	0.991
	100	0.049	0.028	0.000	0.000	0.000	0.000	0.992	0.991

Notes: See notes to Table 1.

Table 4: Simulation results for Case 1 - Rank Deficiency- (Experiment A)

		FE		$3DCCE_G$		$3DCCE_L$		$3DCCE_{LG}$	
		Panel A: Bias							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.227	0.213	0.220	0.242	0.077	0.099	-0.005	0.018
0	25	0.214	0.222	0.223	0.239	0.050	0.066	-0.017	-0.001
	50	0.217	0.208	0.248	0.246	0.061	0.061	0.002	0.001
	100	0.232	0.222	0.260	0.249	0.067	0.057	0.012	0.001
	10	0.227	0.208	0.242	0.249	0.076	0.084	-0.004	0.003
0.5	25	0.246	0.227	0.273	0.274	0.076	0.079	0.009	0.012
	50	0.247	0.241	0.258	0.266	0.049	0.058	-0.012	-0.003
	100	0.246	0.245	0.272	0.271	0.056	0.055	0.000	-0.002
		Mean Group Estimator							
	10	0.223	0.210	0.211	0.234	0.040	0.062	0.008	0.030
0	25	0.204	0.217	0.214	0.228	0.028	0.043	-0.006	0.009
	50	0.209	0.202	0.237	0.235	0.047	0.046	0.009	0.008
	100	0.224	0.214	0.248	0.238	0.057	0.046	0.010	0.005
	10	0.227	0.207	0.230	0.241	0.041	0.054	0.006	0.019
0.5	25	0.241	0.220	0.263	0.263	0.059	0.060	0.023	0.024
	50	0.239	0.234	0.247	0.254	0.038	0.045	-0.003	0.005
	100	0.237	0.236	0.260	0.259	0.047	0.046	0.003	0.002
		Panel B: RMSE							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.497	0.498	0.509	0.520	0.467	0.469	0.461	0.460
0	25	0.357	0.359	0.364	0.362	0.294	0.279	0.290	0.271
	50	0.292	0.290	0.313	0.322	0.202	0.216	0.192	0.208
	100	0.273	0.263	0.297	0.286	0.158	0.152	0.144	0.141
	10	0.520	0.502	0.517	0.520	0.461	0.463	0.454	0.457
0.5	25	0.388	0.368	0.397	0.396	0.298	0.297	0.289	0.285
	50	0.317	0.312	0.326	0.335	0.205	0.212	0.199	0.205
	100	0.283	0.284	0.306	0.306	0.151	0.152	0.140	0.143
		Mean Group Estimator							
	10	0.490	0.488	0.503	0.510	0.459	0.457	0.457	0.454
0	25	0.352	0.356	0.358	0.354	0.288	0.274	0.286	0.270
	50	0.286	0.285	0.304	0.313	0.197	0.211	0.191	0.207
	100	0.266	0.256	0.287	0.276	0.154	0.148	0.144	0.141
	10	0.513	0.493	0.501	0.509	0.447	0.452	0.444	0.450
0.5	25	0.383	0.361	0.388	0.388	0.291	0.291	0.286	0.285
	50	0.310	0.304	0.317	0.326	0.202	0.208	0.198	0.203
	100	0.275	0.275	0.295	0.295	0.146	0.149	0.139	0.142
		Panel C: Coverage rate at 95 confidence level							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.984	0.973	0.967	0.957	0.896	0.893	0.891	0.894
0	25	0.972	0.970	0.967	0.971	0.927	0.944	0.925	0.951
	50	0.953	0.944	0.936	0.909	0.938	0.920	0.949	0.927
	100	0.844	0.870	0.801	0.833	0.922	0.930	0.941	0.952
	10	0.969	0.978	0.967	0.971	0.901	0.912	0.897	0.902
0.5	25	0.965	0.967	0.954	0.955	0.912	0.931	0.926	0.938
	50	0.943	0.929	0.943	0.912	0.937	0.926	0.942	0.934
	100	0.847	0.851	0.815	0.795	0.932	0.935	0.958	0.942
		Mean Group Estimator							
	10	0.939	0.927	0.956	0.949	0.905	0.907	0.898	0.909
0	25	0.903	0.905	0.948	0.949	0.934	0.951	0.935	0.953
	50	0.824	0.801	0.904	0.879	0.946	0.928	0.953	0.932
	100	0.788	0.831	0.742	0.768	0.932	0.933	0.949	0.952
	10	0.951	0.948	0.962	0.965	0.922	0.923	0.919	0.922
0.5	25	0.886	0.895	0.926	0.928	0.931	0.941	0.931	0.944
	50	0.818	0.792	0.898	0.869	0.940	0.937	0.940	0.935
	100	0.787	0.781	0.727	0.724	0.937	0.936	0.957	0.943

Notes: See notes to Table 1.

Table 5: Simulation results for Case 2 - Rank Deficiency - (Experiment B)

		FE		$3DCCE_G$		$3DCCE_L$		$3DCCE_{LG}$	
		Panel A: Bias							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.209	0.205	0.231	0.223	0.076	0.072	0.003	-0.001
0	25	0.215	0.214	0.243	0.243	0.063	0.063	0.002	0.002
	50	0.219	0.219	0.247	0.248	0.056	0.058	0.000	0.001
	100	0.220	0.220	0.250	0.251	0.054	0.054	0.000	0.001
	10	0.227	0.230	0.250	0.250	0.078	0.078	0.004	0.004
0.5	25	0.241	0.242	0.270	0.266	0.067	0.064	0.004	0.001
	50	0.246	0.247	0.272	0.271	0.059	0.059	0.001	0.001
	100	0.249	0.248	0.276	0.275	0.056	0.056	0.001	0.000
		Mean Group Estimator							
	10	0.204	0.201	0.222	0.215	0.053	0.047	0.012	0.006
0	25	0.209	0.208	0.232	0.234	0.048	0.049	0.008	0.009
	50	0.211	0.212	0.236	0.237	0.046	0.047	0.005	0.005
	100	0.212	0.213	0.238	0.239	0.045	0.046	0.002	0.003
	10	0.224	0.227	0.242	0.246	0.057	0.060	0.013	0.016
0.5	25	0.233	0.234	0.258	0.256	0.054	0.052	0.011	0.010
	50	0.236	0.238	0.261	0.260	0.051	0.050	0.007	0.006
	100	0.239	0.239	0.264	0.264	0.049	0.049	0.004	0.004
		Panel B: RMSE							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.240	0.235	0.258	0.251	0.136	0.132	0.111	0.110
0	25	0.220	0.220	0.248	0.247	0.078	0.077	0.045	0.044
	50	0.220	0.220	0.249	0.249	0.061	0.062	0.023	0.023
	100	0.220	0.221	0.250	0.251	0.055	0.056	0.012	0.011
	10	0.258	0.258	0.278	0.276	0.138	0.137	0.115	0.112
0.5	25	0.246	0.246	0.275	0.271	0.082	0.079	0.047	0.046
	50	0.247	0.248	0.273	0.273	0.064	0.064	0.024	0.024
	100	0.249	0.248	0.276	0.276	0.058	0.057	0.012	0.012
		Mean Group Estimator							
	10	0.230	0.228	0.245	0.239	0.116	0.112	0.103	0.101
0	25	0.213	0.213	0.236	0.238	0.063	0.064	0.042	0.042
	50	0.213	0.213	0.237	0.238	0.051	0.051	0.021	0.021
	100	0.212	0.213	0.238	0.239	0.047	0.047	0.011	0.010
	10	0.248	0.249	0.264	0.266	0.118	0.118	0.105	0.103
0.5	25	0.237	0.238	0.261	0.259	0.068	0.066	0.042	0.042
	50	0.238	0.239	0.262	0.261	0.055	0.054	0.022	0.021
	100	0.239	0.239	0.264	0.264	0.050	0.050	0.011	0.011
		Panel C: Coverage rate at 95 confidence level							
		Pooled Estimator							
ρ	N/T	50	100	50	100	50	100	50	100
	10	0.998	1.000	1.000	1.000	0.945	0.953	0.978	0.965
0	25	0.997	0.996	1.000	1.000	0.893	0.876	0.989	0.983
	50	0.613	0.511	0.999	0.996	0.605	0.555	0.991	0.987
	100	0.701	0.612	0.075	0.035	0.041	0.017	0.993	0.992
	10	0.999	0.999	1.000	1.000	0.945	0.948	0.966	0.976
0.5	25	0.998	0.994	1.000	1.000	0.871	0.887	0.991	0.987
	50	0.532	0.431	0.988	0.984	0.557	0.557	0.991	0.990
	100	0.170	0.079	0.004	0.000	0.034	0.024	0.993	0.992
		Mean Group Estimator							
	10	0.999	1.000	1.000	1.000	0.962	0.969	0.978	0.973
0	25	0.989	0.991	1.000	0.999	0.938	0.938	0.984	0.983
	50	0.301	0.209	0.992	0.992	0.691	0.661	0.988	0.982
	100	0.342	0.246	0.005	0.003	0.063	0.040	0.987	0.992
	10	0.996	1.000	1.000	1.000	0.953	0.960	0.969	0.980
0.5	25	0.983	0.979	1.000	1.000	0.913	0.912	0.989	0.983
	50	0.161	0.124	0.938	0.949	0.586	0.612	0.985	0.991
	100	0.007	0.007	0.000	0.000	0.031	0.029	0.989	0.988

Notes: See notes to Table 1.

to remove heterogenous global factors that would be correlated with covariates. KMSS develop the two-step consistent 3D-PCCE estimation procedure by approximating global factors with double cross-section averages of dependent variable and regressors and applying the 3D-within transformation. Following this research trend, in this paper, we develop the hierarchical multi-factor error components specification, (338), which is more structural and parsimonious than (334).

The data We collect the dataset over the period 1970-2013 (44 years), and consider two control groups: the 210 country-pairs of the EU15 member countries with 11 Euro countries (Austria, Belgium-Luxemburg, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain) and 4 control countries (Denmark, Norway, Sweden, the UK); the 320 country-pairs among 19 countries with the EU15 countries and 4 non-EU OECD countries (Australia, Canada, Japan and the US).

We collect the bilateral export flow from IMF. The Data starts from 1970 as the German data are unavailable in the 60s. There are no missing data so we consider the balanced panel. Our sample period consists of several important economic integrations, such as the European Monetary System in 1979 and the Single Market in 1993, all of which can be regarded as promoting intra-EU trades.

Empirical specification: We consider the 3D panel gravity specification:

$$\begin{aligned}\ln EXP_{ijt} = \beta_0 + \beta_1 CEE_{ijt} + \beta_2 EMU_{ijt} + \beta_3 SIM_{ijt} + \beta_4 RLF_{ijt} + \beta_5 \ln GDP_{it} \\ + \beta_6 \ln GDP_{jt} + \beta_7 RER_t + \gamma_1 DIS_{ij} + \gamma_2 BOR_{ij} + \gamma_3 LAN_{ij} + u_{ijt}\end{aligned} \tag{335}$$

the dependent variable, EXP_{ijt} is the export flow from country i to country j at time t; CEE and EMU are dummies for European Community membership and for European Monetary Union; SIM is the logarithm of an index that captures the relative size of two countries and bounded between zero (absolute divergence) and 0.5 (equal size); RLF is the logarithm of the absolute value of the difference between per capita GDPs of trading countries; RER represents the logarithm of common real exchange rates; GDP_{it} and GDP_{jt} are logged GDPs of exporter and importer; the logarithm of geographical distance (DIS) and the dummies for common language (LAN) and for common border (BOR) represent time-invariant bilateral barriers.

We apply the four estimators considered in the MC simulations, namely the two-way within estimator and the three versions of 3D CCEP estimators. We also report the CD results applied to the residuals and the estimates of the CSD exponent (α). We focus on investigating the impacts of t_{ij} that contain both barriers and incentives to trade; the two dummy variables CEE (equal to one when both countries belong to the European Community) and EMU (equal to one when both trading partners adopt the same currency). Both are expected to exert a positive impact on export flows. The empirical evidence is mixed,

though recent studies by Mastromarco et al. (2015), and Gunnella et al. (2015) that control for strong CSD in 2D panels, find modest but significant effects (7 to 10%) of the euro on intra-EU trade flows. KMSS (2017) apply a 3D PCCE estimator, finding that the EMU impact on exports is about 8%.

The Estimation Results for the EU15 Countries Table 5 reports the panel gravity estimation results for the 210 country-pairs among the EU15 member countries over the period 1970-2013 (44 years). The FE estimator suffers from strong CSD while the $3DCCEP$ estimators display lower degree CSD. CD diagnostic test by Pesaran (2015) fails to reject the null of weak CSD for both $3DCCEP_L$ and $3DCCEP_{GL}$. This is also supported by the smaller estimates of α for $3DCCEP_L$ (0.624) and $3DCCEP_{GL}$ (0.609), close to a moderate range of weak CSD.

Factor approximations

Theoretically, we should employ the entire set of cross-section averages to approximate heterogeneous global and local factors. In practice, this may raise an issue of multicollinearity. Further, to avoid the curse of dimensionality, we search for an optimal subset of cross sectional averages. In the aftermath of the global financial crisis, export flows display a negative average growth as shown below:

Export Growth	70/80	80/90	90/00	00/10	10/13
EU15 + 4 OECD	7.06	6.25	4.35	2.16	-0.34
EU15	8.86	7.37	3.92	2.82	-2.05

Hence, we also add t^2 as an observed factor, which helps to capture the confounding effect of the crisis.

We focus on the $3DCCEP_{GL}$ estimation results with the lowest degree of CSD. All the coefficients are significant and their signs are consistent with our *a priori* expectations. The effect of the foreign GDP is substantially higher than the home GDP. The effects of SIM and RER are positive while a depreciation of the home currency leads to a significant increase in exports. SIM boosts real export flows, which suggests that the intra-industry trade is the main part of the trade in the EU. Importantly, the impacts of EMU and CEE are significant, but substantially smaller than the potentially biased FE estimates. Both Euro and CEE impacts drop sharply from 0.099 and 0.074 to 0.03 and 0.05. Other estimators provide rather unreliable results.[23]

The 3DCCEP estimator wipes out the time invariant regressors. Following the 2-step approach as in Serlenga and Shin (2007), we can estimate $\boldsymbol{\gamma}$ by the between estimator:

$$d_{ijt} = \alpha_{ij} + \gamma_1 DIS_{ij} + \gamma_2 BOR_{ij} + \gamma_3 LAN_{ij} + u_{ijt} \tag{336}$$

[23] For example, the impacts of home GDP on exports is surprisingly larger than the foreign impact while both Euro and CEE impacts seem to be rather high for the FE. The RER coefficient is significantly negative for the CCEP with the global approximation only whereas the CEE impact is insignificant and the Euro impact is almost negligible for the CCEP with the local approximation only.

Table 6: Table 5: Estimation Results for 15 EU Countries

	FE	$3DCCEP_G$	$3DCCEP_L$	$3DCCEP_{GL}$
gdph	1.517	0.230	0.023	0.342
	(0.044)	(0.036)	(0.037)	(0.124)
gdpf	0.953	1.478	0.779	1.498
	(0.044)	(0.037)	(0.057)	(0.031)
sim	-0.045	0.639	-0.012	0.197
	(0.060)	(0.069)	(0.056)	(0.075)
rlf	0.030	-0.002	0.002	0.006
	(0.006)	(0.005)	(0.002)	(0.004)
rer	0.012	-0.046	0.016	0.103
	(0.008)	(0.007)	(0.004)	(0.010)
euro	0.099	0.030	0.012	0.030
	(0.016)	(0.003)	(0.003)	(0.003)
cee	0.074	0.066	0.007	0.050
	(0.014)	(0.007)	(0.007)	(0.013)
CD stat	206.6	4.67	2.33	2.72
α	0.91 (0.90-0.93)	0.78 (0.72-0.84)	0.62 (0.59-0.66)	0.61 (0.57-0.65)

Notes: FE is the two-way fixed effect estimator. $3DCCEP_G$ is the CCEP estimator with only the global factors approximated by $\mathbf{f}_t = \left\{\overline{export}_{..t}, \overline{gdp}_{..t}, \overline{sim}_{..t}, \overline{rlf}_{..t}, \overline{cee}_{..t}, t, t^2\right\}$ $3DCCEP_L$ is the CCEP estimator with only the local factors approximated by $\mathbf{f}_t = \mathbf{f}_{iot} = \left\{\overline{y}_{i.t}, \overline{gdp}_{i.t}\right\}$ and $\mathbf{f}_{ojt} = \left\{\overline{sim}_{.jt}, \overline{rlf}_{.jt}\right\}$. $3DCCEP_{GL}$ is the CCEP estimator with both global and local factors approximated by $\mathbf{f}_t = \left\{\overline{export}_{..t}, \overline{gdp}_{..t}, \overline{sim}_{..t}, \overline{rlf}_{..t}, \overline{cee}_{..t}, t, t^2\right\}$ and $\mathbf{f}_{iot} = \left\{\overline{sim}_{i.t}, \overline{rlf}_{i.t}, \overline{rer}_{i.t}\right\}$. * and ** stand for significance at 5% and 10% level. CD test refers to testing the null hypothesis of residual cross-section independence or weak dependence (Pesaran, 2015). α is the estimate of CSD exponent with 90% confidence bands inside parenthesis.

where $d_{ijt} = y_{ijt} - \widehat{\boldsymbol{\beta}}' \mathbf{x}_{ijt}$ with $\widehat{\boldsymbol{\beta}}$ being the 3D CCEP estimator. We test the validity of the hypothesis: if the Euro had a positive effect on the EU trade by reducing bilateral barriers and eliminating exchange-related uncertainties and transaction costs, this caused a decrease in trade impacts of bilateral barriers (e.g. Cafiso, 2010). A declining trend after 1999 will support the hypothesis that the Euro helps to promote more EU integration. To this end we estimate (336) by the cross-section regression at each period, and produce time-varying coefficients of $\boldsymbol{\gamma}$.

Figure 1 shows the time varying estimates of $\boldsymbol{\gamma}$ in (336), using the CCEP estimator with both local and global factors approximation. The border effect has been declining until the mid 1980's, and quite stable except the slight dip during the global financial crisis albeit statistically insignificant. The language effect steadily decreasing until the end of 1980's, reflecting the progressive lessening of restrictions on labor mobility within the EU, that encouraged migration and reduced the relative importance of trade costs and cultural difference. Since the introduction of the Euro in 1999, both language and border effects became flat, suggesting that the EU integration may reach near-completion stage. This is consistent with the currency union formation hypothesis by Frankel (2005) that countries, which decide to join a currency union, are self-selected on the basis of distinctive features shared by EU members. The effect of distance has been on a declining trend from the mid 80's, but started to rise slightly after 1999.

The Estimation Results for the EU15 plus 4 OECD Countries Table 3 reports the estimation results for an enlarged sample of the 342 country-pairs among the EU15 member countries plus four more countries (Australia, Canada, Japan and the US). Again we focus on the estimation results for the $3DCCEP_{GL}$, which shows the lowest degree of CSD. The results are qualitatively similar to those reported in Table 2. All the coefficients are significant with expected signs. The effect of the foreign GDP is substantially higher than the home GDP effect. The effects of SIM is slightly higher (from 0.2 to 0.22) but RLF becomes negligibly negative. The impact of RER is stronger (from 0.10 to 0.18), implying a stronger terms of trade effect.

The impacts of EMU and CEE are significant, though their magnitudes become smaller than those with the EU15 countries, namely from 3% to 1.5% and from 5% to 3%. The smaller effects for the enlarged sample might reflect the trade diversion between the Euro and non-Euro area. The effects of the EMU on trade will differ with respect to the selected control group and depend on the composition of treatment and control groups (e.g. Baier and Bergstrand, 2009). Other estimators provide rather misleading results. In particular, the FE estimation provides an opposite result that both Euro and CEE impacts increase substantially, from 0.099 and 0.074 to 0.258 and 0.161.

Figure 2 displays time varying estimates of $\boldsymbol{\gamma}$, using the CCEP estimator with both local and global factors approximation. Both border and language effect show similar pattern to the case with the EU15 countries. Again, we

Table 7: Table 6: Estimation Results for 15 EU plus 4 OECD countries

	FE	$3DCCEP_G$	$3DCCEP_L$	$3DCCEP_{GL}$
gdph	1.066	0.531	0.069	0.169
	(0.019)	(0.016)	(0.010)	(0.055)
gdpf	0.904	1.419	1.262	1.417
	(0.020)	(0.020)	(0.015)	(0.017)
sim	0.332	0.109	0.100	0.220
	(0.029)	(0.021)	(0.013)	(0.023)
rlf	0.027	-0.008	0.010	-0.004
	(0.004)	(0.002)	(0.001)	(0.002)
rer	0.058	0.086	0.074	0.179
	(0.008)	(0.004)	(0.002)	(0.005)
euro	0.258	0.012	0.012	0.014
	(0.009)	(0.003)	(0.001)	(0.002)
cee	0.161	0.021	0.007	0.030
	(0.028)	(0.017)	(0.001)	(0.012)
CD stat	243.33	3.272	2.331	3.201
α	0.90 (0.88-0.920)	0.74 (0.69-0.76)	0.65 (0.61-0.69)	0.62 (0.57-0.66)

Notes: FE is the two-way fixed effect estimator. $3DCCEP_G$ is the CCEP estimator with only the global factors approximated by $\mathbf{f}_t = \left\{\overline{export}_{..t}, \overline{gdp}_{..t}, \overline{sim}_{..t}, \overline{rlf}_{..t}, \overline{cee}_{..t}, t\right\}$. $3DCCEP_L$ is the CCEP estimator with only the local factors approximated by $\mathbf{f}_t = \mathbf{f}_{iot} = \left\{\overline{y}_{i.t}, \overline{gdp}_{i.t}\right\}$ and $\mathbf{f}_{ojt} = \left\{\overline{sim}_{.jt}, \overline{rlf}_{.jt}\right\}$. $3DCCEP_{GL}$ is the CCEP estimator with both global and local factors approximated by $\mathbf{f}_t = \left\{\overline{export}_{..t}, \overline{gdp}_{..t}, \overline{sim}_{..t}, \overline{rlf}_{..t}, \overline{cee}_{..t}, t\right\}$ and $\mathbf{f}_{iot} = \left\{\overline{sim}_{i.t}, \overline{rlf}_{i.t}, \overline{rer}_{i.t}\right\}$. * and ** stand for significance at 5% and 10% level. CD test refers to testing the null hypothesis of residual cross-section independence or weak dependence (Pesaran, 2015). α is the estimate of CSD exponent with 90% confidence bands inside parenthesis.

do not observe any evidence in favour of the Euro effect on trade integration, consistent with the currency union formation hypothesis by Frankel (2005). The effect of distance has been slightly increasing over the whole period. This is consistent with the meta-study by Disdier and Head (2008), who document that the trade elasticity with respect to distance has not declined, but rather increased recently.

10 The 3D Panels with Regional/Global factor specification

We consider the 3D heterogeneous panel data model given by

$$y_{ijt} = \boldsymbol{\beta}_{ij}' \boldsymbol{x}_{ijt} + \boldsymbol{\delta}_{ij}' \boldsymbol{d}_t + e_{ijt},\ i = 1, ..., R,\ j = 1, ..., N_i,\ t = 1, ..., T, \quad (337)$$

where y_{ijt} is the dependent variable observed across three indices, $i = 1, ..., R$ denotes the ith region, $j = 1, ..., n_i$ denotes the j'th individual variable of region i (say, the GDP growth of country j in the region i), $\boldsymbol{x}_{ijt}$ is the $m_x \times 1$ vector of covariates and $\boldsymbol{d}_t$ is the $m_d \times 1$ vector of observed common effects including deterministic components such as constants and trends. $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\delta}_{ij}$ are $m_x \times 1$ and $m_d \times 1$ vectors of parameters.

We allow e_{ijt} to follow the multi-level factor structure:

$$e_{ijt} = \boldsymbol{\gamma}_{ij}^{g\prime} \boldsymbol{g}_t + \boldsymbol{\gamma}_{ij}^{i\prime} \boldsymbol{f}_t^{\,i} + \varepsilon_{ijt}, \quad (338)$$

where $\boldsymbol{g}_t = \left(g_{1t}, ..., g_{m_g t}\right)'$ is a $m_g \times 1$ vector of global factors, $\boldsymbol{f}_t^{\,i} = \left(f_{1t}^i, ..., f_{m_i t}^i\right)'$ collects the $m_i \times 1$ vector of local factors in region $i = 1, ..., R$, $\boldsymbol{\gamma}_{ij}^g$ and $\boldsymbol{\gamma}_{ij}^i$ are $m_g \times 1$ and $m_i \times 1$ vectors of heterogenous loadings, and ε_{ijt} are idiosyncratic errors. Define the total observation by $N = \sum_{i=1}^R N_i$.

Extensions here...

10.0.5 The plans

Here we may develop and make the multiple projects and papers.

- Consider the simpler model given by (337) and (338). Here we may develop the 3D CCE and/or the 3D PCA extensions (see below).
- An extension to the general case

  $$y_{ijt} = \boldsymbol{\beta}_{ij}' \boldsymbol{x}_{ijt} + \boldsymbol{\beta}_j' \boldsymbol{x}_{it} + \boldsymbol{\beta}_i' \boldsymbol{x}_{jt} + \boldsymbol{\delta}_{ij}' \boldsymbol{d}_t + u_{ijt}$$

  would be straightforward, though we stick to the simpler setting for tractability.

- Consider the different hierarchical factor structure, (358) (say, source-destination countries) considered by KMS (2018) and Lu and Su (2018) and/or the three level or overlapping factor structure, (381) analysed by Breitung and Eickmeier (2016). Here we may also develop the 3D CCE and/or the 3D PCA extensions by developing different modelling techniques.

- Eventually, we wish to develop alternative approaches by combining the (local) spatial effects and global (or multi-level) factors. Indeed, this would make my ultimate goal. I consider extending the QML-EM algorithms proposed by Bai and Li in a sequence of papers (all published in top journals).

- There will be abundant empirical topics in Economics, Finance, Social Science and so on. One of my ph.d students, Rui has been working on this application in extending the 2D CAPM model into 3D CAPM or APT models.

10.1 The CCE extensions

bla bla

10.2 The KMS CCE extension

Unobserved factors, $\boldsymbol{f}_t^g$ and $\boldsymbol{f}_{it}^i$, are likely to be correlated with $\boldsymbol{x}_{ijt}$. Thus, we consider the data generating process for $\boldsymbol{x}_{ijt}$ as follows:

$$\boldsymbol{x}_{ijt} = \mathcal{D}_{ij}\boldsymbol{d}_t + \boldsymbol{\Gamma}_{ij}^g \boldsymbol{g}_t + \boldsymbol{\Gamma}_{ij}^i \boldsymbol{f}_t^i + \boldsymbol{v}_{ijt}, \tag{339}$$

where $\mathcal{D}_{ij}$ is the $(m_x \times m_d)$ parameter matrix on observed common effects, $\boldsymbol{\Gamma}_{ij}^g$ and $\boldsymbol{\Gamma}_{ij}^i$ are $(m_x \times m_g)$ and $(m_x \times m_i)$ factor loading matrices, and $\boldsymbol{v}_{ijt}$ are the idiosyncratic errors.

Combining (337)-(339), we have:

$$\boldsymbol{z}_{ijt} = \begin{pmatrix} y_{ijt} \\ \boldsymbol{x}_{ijt} \end{pmatrix} = \boldsymbol{\Xi}_{ij}\boldsymbol{d}_t + \boldsymbol{\Phi}_{ij}^g \boldsymbol{g}_t + \boldsymbol{\Phi}_{ij}^i \boldsymbol{f}_t^i + \boldsymbol{u}_{ijt} \tag{340}$$

where

$$\boldsymbol{\Xi}_{ij} = \begin{pmatrix} \boldsymbol{\delta}_{ij}' + \boldsymbol{\beta}_{ij}'\mathcal{D}_{ij} \\ \mathcal{D}_{ij} \end{pmatrix}, \boldsymbol{\Phi}_{ij}^g = \begin{pmatrix} \boldsymbol{\gamma}_{ij}^{g\prime} + \boldsymbol{\beta}_{ij}'\boldsymbol{\Gamma}_{ij}^g \\ \boldsymbol{\Gamma}_{ij}^g \end{pmatrix}, \boldsymbol{\Phi}_{ij}^r = \begin{pmatrix} \boldsymbol{\gamma}_{ij}^{i\prime} + \boldsymbol{\beta}_{ij}'\boldsymbol{\Gamma}_{ij}^i \\ \boldsymbol{\Gamma}_{ij}^i \end{pmatrix}, \boldsymbol{u}_{ijt} = \begin{pmatrix} \varepsilon_{ijt} + \boldsymbol{\beta}_{ij}'\boldsymbol{v}_{ijt} \\ \boldsymbol{v}_{ijt} \end{pmatrix}$$

The ranks of $\boldsymbol{\Phi}_{ij}^g$ and $\boldsymbol{\Phi}_{ij}^i$ are determined by the ranks of the following matrices:

$$\underset{(m_x+1)\times m_g}{\tilde{\boldsymbol{\Gamma}}_{ij}^g} = \begin{pmatrix} \boldsymbol{\gamma}_{ij}^{g\prime} \\ \boldsymbol{\Gamma}_{ij}^g \end{pmatrix}, \quad \underset{(m_x+1)\times m_i}{\tilde{\boldsymbol{\Gamma}}_{ij}^r} = \begin{pmatrix} \boldsymbol{\gamma}_{ij}^{i\prime} \\ \boldsymbol{\Gamma}_{ij}^i \end{pmatrix}.$$

For each (i,j), we rewrite (337) and (340) in matrix notation:

$$\boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{ij} + \boldsymbol{D}\boldsymbol{\delta}_{ij} + \boldsymbol{G}\boldsymbol{\gamma}^{g}_{ij} + \boldsymbol{F}^{i}\boldsymbol{\gamma}^{i}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{341}$$

$$\boldsymbol{z}_{ij} = \boldsymbol{D}\boldsymbol{\Xi}_{ij} + \boldsymbol{G}\boldsymbol{\Phi}^{g}_{ij} + \boldsymbol{F}^{i}\boldsymbol{\Phi}^{i}_{ij} + \boldsymbol{u}_{ij}, \tag{342}$$

where

$$\underset{T\times 1}{\boldsymbol{y}_{ij}} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijT} \end{bmatrix}, \underset{T\times m_x}{\boldsymbol{X}_{ij}} = \begin{bmatrix} \boldsymbol{x}'_{ij1} \\ \vdots \\ \boldsymbol{x}'_{ijT} \end{bmatrix}, \underset{T\times m_d}{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{d}'_1 \\ \vdots \\ \boldsymbol{d}'_T \end{bmatrix}, \underset{T\times(m_x+1)}{\boldsymbol{z}_{ij}} = \begin{bmatrix} \boldsymbol{z}'_{ij1} \\ \vdots \\ \boldsymbol{z}'_{ijT} \end{bmatrix}, \tag{343}$$

$$\underset{T\times m_g}{\boldsymbol{G}} = \begin{bmatrix} \boldsymbol{g}'_1 \\ \vdots \\ \boldsymbol{g}'_T \end{bmatrix}, \underset{T\times m_i}{\boldsymbol{F}^{i}} = \begin{bmatrix} \boldsymbol{f}^{i\prime}_{1} \\ \vdots \\ \boldsymbol{f}^{i\prime}_{T} \end{bmatrix}, \underset{T\times 1}{\boldsymbol{\varepsilon}_{ij}} = \begin{bmatrix} \varepsilon_{ij1} \\ \vdots \\ \varepsilon_{ijT} \end{bmatrix}, \underset{T\times(m_x+1)}{\boldsymbol{u}_{ij}} = \begin{bmatrix} \boldsymbol{u}'_{ij1} \\ \vdots \\ \boldsymbol{u}'_{ijT} \end{bmatrix}$$

We develop the estimation and inference theory for $E\left(\boldsymbol{\beta}_{ij}\right) = \boldsymbol{\beta}$ as well as individual coefficients, $\boldsymbol{\beta}_{ij}$. We make the following assumptions:

- **Check the assumptions??**

Assumption 1. *Common Effects*: The $\left(m_d + m_g + \sum_{i=1}^{R} m_i\right) \times 1$ vector of common factors $\left(\boldsymbol{d}'_t, \boldsymbol{g}'_t, \boldsymbol{f}^{1\prime}_{t}, ..., \boldsymbol{f}^{R\prime}_{t}\right)'$, is covariance stationary with absolute summable autocovariances, distributed independently of $\varepsilon_{ijt'}$ and $\boldsymbol{v}_{ijt'}$ for all N, i, j, t and t'. $\left(\boldsymbol{g}'_t, \boldsymbol{f}^{1\prime}_{t}, ..., \boldsymbol{f}^{R\prime}_{t}\right)' \left(\boldsymbol{f}^{g\prime}_{t}, \boldsymbol{f}^{r\prime}_{1t}, ..., \boldsymbol{f}^{r\prime}_{Rt}\right)'$ are zero mean process and are mutually uncorrelated.

Assumption 2. *Individual-specific Errors*: ε_{ijt} and $\boldsymbol{v}_{ijt'}$ are distributed independently for all i, j, t and t', and they are distributed independently of $\boldsymbol{x}_{ijt}$ and $\boldsymbol{d}_t$.

Assumption 3. *Factor Loadings*: The factor loadings are independently and identically distributed across (i,j), and of the individual- specific errors ε_{ijt} and $\boldsymbol{v}_{ijt}$, the common factors for all i, j and t, with finite means and finite variances. In particular, we have:

$$\boldsymbol{\gamma}^{g}_{ij} = \boldsymbol{\gamma}^{g} + \boldsymbol{\eta}^{g}_{ij},\ \boldsymbol{\gamma}^{i}_{ij} = \boldsymbol{\gamma}^{i}_{i} + \boldsymbol{\eta}^{i}_{i} \text{ or } \boldsymbol{\gamma}^{i}_{ij} = \boldsymbol{\gamma}^{i}_{i} + \boldsymbol{\eta}^{i}_{ij}, \tag{344}$$

$$\boldsymbol{\Gamma}^{g}_{ij} = \boldsymbol{\Gamma}^{g} + \boldsymbol{\xi}^{g}_{ij},\ \boldsymbol{\Gamma}^{r}_{ij} = \boldsymbol{\Gamma}^{r}_{i} + \boldsymbol{\xi}^{r}_{i} \text{ or } \boldsymbol{\Gamma}^{r}_{ij} = \boldsymbol{\Gamma}^{r}_{i} + \boldsymbol{\xi}^{r}_{ij} \tag{345}$$

where $\boldsymbol{\eta}^{g}_{ij} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\eta}^{g}}\right)$, $\boldsymbol{\xi}^{g}_{ij} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\xi}^{g}}\right)$, $\boldsymbol{\eta}^{r}_{i} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\eta}_{i}}\right)$, ?? $\boldsymbol{\xi}_{i\circ} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\xi}_{\bullet\circ}}\right)$, $\boldsymbol{\eta}_{\circ j} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\eta}_{\circ\bullet}}\right)$ and $\boldsymbol{\xi}_{\circ j} \sim iid\left(0, \boldsymbol{\Omega}_{\boldsymbol{\xi}_{\circ\bullet}}\right)$. Further, $\|\boldsymbol{\gamma}_{\circ\circ}\| < K$, $\|\boldsymbol{\gamma}_{\bullet\circ}\| < K$, $\|\boldsymbol{\gamma}_{\circ\bullet}\| < K$, $\|\boldsymbol{\Gamma}_{\circ\circ}\| < K$, $\|\boldsymbol{\Gamma}_{\bullet\circ}\| < K$, and $\|\boldsymbol{\Gamma}_{\circ\bullet}\| < K$ for some positive constant $K < \infty$.

Assumption 4. *Random Slope Coefficients*: $\boldsymbol{\beta}_{ij}$ follow the random coefficient specification: {??}

$$\boldsymbol{\beta}_{ij} = \boldsymbol{\beta} + \boldsymbol{\nu}_{i\circ} + \boldsymbol{\nu}_{\circ j} + \boldsymbol{\nu}_{ij} \text{ with } \boldsymbol{\nu}_{i\circ} \sim iid\left(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\bullet\circ}}\right),\ \boldsymbol{\nu}_{\circ j} \sim iid\left(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ\bullet}}\right),\ \boldsymbol{\nu}_{ij} \sim iid\left(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ\circ}}\right) \tag{346}$$

or

$$\boldsymbol{\beta}_{ij} = \boldsymbol{\beta} + \boldsymbol{\nu}_{\circ j} + \boldsymbol{\nu}_{ij} \text{ with } \boldsymbol{\nu}_{\circ j} \sim iid\left(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ\bullet}}\right),\ \boldsymbol{\nu}_{ij} \sim iid\left(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\nu}_{\circ\circ}}\right)$$

where $\|\boldsymbol{\beta}\| < K$ and $\boldsymbol{\nu}_{ij}$, $\boldsymbol{\nu}_{i\circ}$, $\boldsymbol{\nu}_{\circ j}$ are distributed independently of one another, and of $\boldsymbol{\gamma}_{ij}$, $\boldsymbol{\Gamma}_{ij}$, ε_{ijt}, $\boldsymbol{v}_{ijt}$ and $\boldsymbol{g}_t$ for all i, j and t.

Assumption 5. *Identification of $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\beta}$*: Let

$$\bar{\boldsymbol{z}}_t = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{z}_{ijt},\ \bar{\boldsymbol{z}}_{i\cdot t} = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{z}_{ijt},\ i = 1, ..., R \tag{347}$$

and $\bar{\boldsymbol{Z}}_{ij} = \left(\bar{\boldsymbol{Z}}, \bar{\boldsymbol{Z}}_{i\cdot}\right)$ and $\bar{\boldsymbol{H}}_{ij} = \left(\boldsymbol{D}, \bar{\boldsymbol{Z}}_{ij}\right)$, where

$$\underset{T\times(1+m_x)}{\bar{\boldsymbol{Z}}} = \begin{bmatrix} \bar{\boldsymbol{z}}_1' \\ \vdots \\ \bar{\boldsymbol{z}}_T' \end{bmatrix},\ \underset{T\times(1+m_x)}{\bar{\boldsymbol{Z}}_{i\cdot}} = \begin{bmatrix} \bar{\boldsymbol{z}}_{i\cdot 1}' \\ \vdots \\ \bar{\boldsymbol{z}}_{i\cdot T}' \end{bmatrix},$$

Further,

$$\bar{\boldsymbol{M}}_{ij} = \boldsymbol{I}_T - \bar{\boldsymbol{H}}_{ij}\left(\bar{\boldsymbol{H}}_{ij}'\bar{\boldsymbol{H}}_{ij}\right)^{-1}\bar{\boldsymbol{H}}_{ij}'. \tag{348}$$

(i) Identification of $\boldsymbol{\beta}_{ij}$: The $m_x \times m_x$ matrices, $\bar{\boldsymbol{\Psi}}_{ij,T} = T^{-1}\left(\boldsymbol{X}_{ij}'\bar{\boldsymbol{M}}_{ij}\boldsymbol{X}_{ij}\right)$ are nonsingular, and $\bar{\boldsymbol{\Psi}}_{ij,T}^{-1}$ have finite second-order moments for all (i, j).
(ii) Identification of $\boldsymbol{\beta}$: The $m_x \times m_x$ matrix, $N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{\boldsymbol{\Psi}}_{ij,T}$ is nonsingular.

Remark 10 *The factors are assumed to have zero mean for simplicity. Any means can be subsumed in $\boldsymbol{\delta}_{ij}$. Further, they are assumed mutually uncorrelated to ensure that cross-sectional averages of local factors converge to zero. This is another simplifying assumption, since some weak cross-sectional dependence across local factors could be allowed, in exact analogy to weak cross sectional dependence across idiosyncratic shocks.*

Remark 11 *The weights are not necessarily unique, but they do not affect the asymptotic results advanced in this paper (see also (?)). We focus, for simplicity, on equal weights, $1/N$. Alternatively, economic distance-based or time-varying measures such as trade weights or input-output shares could be considered (e.g. (?); (?)). The number of observed factors, m_d and the number of individual-specific regressors, m_x are assumed fixed. The number of unobserved factors, $m = m_g + m_i$, is assumed fixed, but need not to be known.*

Using (340), we represent the hierarchical cross-section averages in (347) as follows:

$$\begin{aligned} \bar{\boldsymbol{z}}_{i\cdot t} &= \frac{1}{N_i}\sum_{j=1}^{N_i}\left(\boldsymbol{\Xi}_{ij}\boldsymbol{d}_t + \boldsymbol{\Phi}_{ij}^{g}\boldsymbol{g}_t + \boldsymbol{\Phi}_{ij}^{i}\boldsymbol{f}_t^{i} + \boldsymbol{u}_{ijt}\right) \qquad (349) \\ &= \bar{\boldsymbol{\Xi}}_{i\cdot}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}_{i\cdot}^{g}\boldsymbol{g}_t + \bar{\boldsymbol{\Phi}}_{i\cdot}^{i}\boldsymbol{f}_t^{i} + \bar{\boldsymbol{u}}_{i\cdot t} = \bar{\boldsymbol{\Xi}}_{i\cdot}\boldsymbol{d}_t + \left(\bar{\boldsymbol{\Phi}}_{i\cdot}^{g}, \bar{\boldsymbol{\Phi}}_{i\cdot}^{i}\right)\begin{pmatrix} \boldsymbol{g}_t \\ \boldsymbol{f}_t^{i} \end{pmatrix} + \bar{\boldsymbol{u}}_{i\cdot t} \end{aligned}$$

where

$$\bar{\boldsymbol{\Xi}}_{i\cdot} = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{\Xi}_{ij}, \bar{\boldsymbol{\Phi}}^{g}_{i\cdot} = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{\Phi}^{g}_{ij}, \bar{\boldsymbol{\Phi}}^{i}_{i\cdot} = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{\Phi}^{i}_{ij}, \bar{\boldsymbol{u}}_{i\cdot t} = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{u}_{ijt}$$

and

$$\begin{aligned}\bar{\boldsymbol{z}}_t &= \frac{1}{R}\sum_{i=1}^{R}\frac{1}{N_i}\sum_{j=1}^{N_i}\left(\boldsymbol{\Xi}_{ij}\boldsymbol{d}_t + \boldsymbol{\Phi}^{g}_{ij}\boldsymbol{g}_t + \boldsymbol{\Phi}^{i}_{ij}\boldsymbol{f}^{i}_{t} + \boldsymbol{u}_{ijt}\right) \qquad (350)\\ &= \bar{\boldsymbol{\Xi}}_{\cdot\cdot}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}^{g}_{\cdot\cdot}\boldsymbol{g}_t + \left(\frac{1}{R}\sum_{i=1}^{R}\bar{\boldsymbol{\Phi}}^{i}_{i\cdot}\boldsymbol{f}^{i}_{t}\right) + \bar{\boldsymbol{u}}_{\cdot\cdot t}\end{aligned}$$

where

$$\bar{\boldsymbol{\Xi}}_{\cdot\cdot} = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{\Xi}_{ij}, \bar{\boldsymbol{\Phi}}^{g}_{\cdot\cdot} = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{\Phi}^{g}_{ij}, \bar{\boldsymbol{u}}_{\cdot\cdot t} = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{u}_{ijt}$$

Combining (349) and (350), we have:

$$\bar{\boldsymbol{z}}_{it} = \bar{\boldsymbol{\Xi}}_{ij}\boldsymbol{d}_t + \bar{\boldsymbol{\Phi}}_{ij}\boldsymbol{h}_t + \bar{\boldsymbol{u}}_{ijt}, \qquad (351)$$

where

$$\underset{2(m_x+1)\times 1}{\bar{\boldsymbol{z}}_{it}} = \begin{bmatrix}\bar{\boldsymbol{z}}_t\\ \bar{\boldsymbol{z}}_{i\cdot t}\end{bmatrix}, \underset{(m_g+m_i)\times 1}{\boldsymbol{h}_t} = \begin{bmatrix}\boldsymbol{g}_t\\ \boldsymbol{f}^{i}_{t}\end{bmatrix}, \underset{2(m_x+1)\times 1}{\bar{\boldsymbol{u}}_{ijt}} = \begin{bmatrix}\bar{\boldsymbol{u}}_{\cdot\cdot t} + \frac{1}{R}\sum_{i=1}^{R}\bar{\boldsymbol{\Phi}}^{r}_{i\cdot}\boldsymbol{f}^{r}_{it}\\ \bar{\boldsymbol{u}}_{i\cdot t}\end{bmatrix}$$

$$\underset{3(m_x+1)\times m_d}{\bar{\boldsymbol{\Xi}}_{ij}} = \begin{bmatrix}\bar{\boldsymbol{\Xi}}_{\cdot\cdot}\\ \bar{\boldsymbol{\Xi}}_{i\cdot}\end{bmatrix}, \underset{2(m_x+1)\times(m_g+m_i)}{\bar{\boldsymbol{\Phi}}_{ij}} = \begin{bmatrix}\bar{\boldsymbol{\Phi}}^{g}_{\cdot\cdot} & \boldsymbol{0}\\ \bar{\boldsymbol{\Phi}}^{g}_{i\cdot} & \bar{\boldsymbol{\Phi}}^{i}_{i\cdot}\end{bmatrix}.$$

Then, we obtain from (351):

$$\boldsymbol{h}_t = \left(\bar{\boldsymbol{\Phi}}'_{ij}\bar{\boldsymbol{\Phi}}_{ij}\right)^{-1}\bar{\boldsymbol{\Phi}}'_{ij}\left(\bar{\boldsymbol{z}}_{it} - \bar{\boldsymbol{\Xi}}_{ij}\boldsymbol{d}_t - \bar{\boldsymbol{u}}_{ijt}\right) \qquad (352)$$

It is easily seen, by applying Lemma 1 of (**?**), to $\bar{\boldsymbol{u}}_{\cdot\cdot t}$, $\bar{\boldsymbol{u}}_{i\cdot t}$ and $\frac{1}{R}\sum_{i=1}^{R}\bar{\boldsymbol{\Phi}}^{i}_{i\cdot}\boldsymbol{f}^{i}_{t}$,[24]

[24] Alternatively, do we assume $\frac{1}{R}\sum_{i=1}^{R}\bar{\boldsymbol{\Phi}}^{r}_{i\cdot}\boldsymbol{f}^{r}_{it} = 0$ for identification?? How to make it sure under the zero mean assumption for $\boldsymbol{f}^{i}_{t}$? If $\frac{1}{R}\sum_{i=1}^{R}\bar{\boldsymbol{\Phi}}^{r}_{i\cdot}\boldsymbol{f}^{r}_{it} \neq 0$, then

$$\frac{1}{R}\sum_{i=1}^{R}\bar{\boldsymbol{\Phi}}^{i}_{i\cdot}\boldsymbol{f}^{i}_{t} = \frac{1}{R}\left(\bar{\boldsymbol{\Phi}}^{i}_{1\cdot}, ..., \bar{\boldsymbol{\Phi}}^{i}_{R\cdot}\right)\begin{pmatrix}\boldsymbol{f}^{1}_{t}\\ \vdots\\ \boldsymbol{f}^{R}_{t}\end{pmatrix}. \qquad (353)$$

In this case, instead of (351), we have

$$\bar{\boldsymbol{z}}_{it} = \begin{pmatrix}\bar{\boldsymbol{\Phi}}^{g}_{\cdot\cdot} & \tilde{\boldsymbol{\Phi}}^{i}_{1\cdot} & \cdots & \tilde{\boldsymbol{\Phi}}^{i}_{i\cdot} & \cdots & \tilde{\boldsymbol{\Phi}}^{i}_{R\cdot}\\ \bar{\boldsymbol{\Phi}}^{g}_{i\cdot} & 0 & \cdots & \bar{\boldsymbol{\Phi}}^{r}_{i\cdot} & \cdots & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{f}^{g}_{t}\\ \boldsymbol{f}^{1}_{t}\\ \vdots\\ \boldsymbol{f}^{R}_{t}\end{pmatrix} + \bar{\boldsymbol{u}}_{ijt}, \qquad (354)$$

where $\tilde{\boldsymbol{\Phi}}^{i}_{i\cdot} = \frac{1}{R}\bar{\boldsymbol{\Phi}}^{i}_{i\cdot}$. And the analysis will be more complicated.

that for each t, as $N \to \infty$,

$$\bar{\boldsymbol{u}}_{ijt} = O_p\left(\frac{1}{\sqrt{N}}\right)??$$

Therefore, we establish that

$$\boldsymbol{h}_t - \left(\bar{\boldsymbol{\Phi}}_{ij}'\bar{\boldsymbol{\Phi}}_{ij}\right)^{-1}\bar{\boldsymbol{\Phi}}_{ij}'\left(\bar{\boldsymbol{z}}_{it} - \bar{\boldsymbol{\Xi}}_{ij}\,\boldsymbol{d}_t\right) = O_p\left(\frac{1}{\sqrt{N}}\right)??$$

This suggests that we can use $\bar{\boldsymbol{h}}_{i\circ t} = \left(\boldsymbol{d}_t', \bar{\boldsymbol{z}}_{it}'\right)'$ as observable proxies for $\boldsymbol{h}_t$. Then, we can consistently estimate the individual slope coefficients, $\boldsymbol{\beta}_{ij}$ and their means $\boldsymbol{\beta}$ by augmenting the regression, (337) with $\boldsymbol{d}_t$ and the cross-section averages $\bar{\boldsymbol{z}}_{it}$. This estimator is referred to as the 3D CCE estimator??

- **Remark**: The multi-level factor structure in (338) is different from KMS, and it is more suitable for an analysis of the nested multi-level panel data (here 2-level).
- ??Further, the local factors are allowed to have the loadings heterogeneous across (i, j). Consider the special case with the within-region homogeneity:

$$e_{ijt} = \boldsymbol{\gamma}_{ij}^{g\prime}\boldsymbol{g}_t + \boldsymbol{\gamma}_i^{i\prime}\boldsymbol{f}_t^{\,i} + \varepsilon_{ijt} \tag{355}$$

  In this case we have (351) with

$$\bar{\boldsymbol{u}}_{ijt} = \begin{bmatrix} \bar{\boldsymbol{u}}_{\cdot\cdot t} + \frac{1}{R}\sum_{i=1}^{R}\boldsymbol{\Phi}_{i\cdot}^{r}\boldsymbol{f}_t^{\,i} \\ \bar{\boldsymbol{u}}_{i\cdot t} \end{bmatrix}, \bar{\boldsymbol{\Phi}}_{ij} = \begin{bmatrix} \bar{\boldsymbol{\Phi}}_{\cdot\cdot}^{g} & \mathbf{0} \\ \bar{\boldsymbol{\Phi}}_{i\cdot}^{g} & \boldsymbol{\Phi}_{i\cdot}^{i} \end{bmatrix}$$

  or see the footnote.

10.3 Panel Data with Cross-Sectional Dependence Characterized by a Multi-Level Factor Structure by Rodríguez-Caballero (2016)

This is also the CCE extension.

Consider the linear heterogeneous panel data model for $i = 1, ..., N$, $t = 1, ..., T$, $r = 1, ..., R$: (**NB: abuse of notations**)

$$y_{r,it} = \alpha_{r,i}d_{r,t} + \beta_{r,i}'x_{r,it} + e_{r,it}$$

$$e_{r,it} = \mu_{r,i}'G_t + \lambda_{r,i}'F_{r,t} + \epsilon_{r,it}$$

where $d_{r,t}$ is a $RN \times 1$ vector of observed common effects in the block r, $x_{r,it}$ is a $k \times 1$ vector of observed individual-specific regressors, G_t is the $r_G \times 1$ vector of unobserved global factors, $F_{r,t}$ is the $r_F \times 1$ vector of unobserved regional factors, and $\epsilon_{r,it}$ are the individual-specific idiosyncratic errors. **Assume the number of blocks, R, to be fixed**.

I adopt the following specification for the individual specific regressors:

$$x_{r,it} = A'_{r,i}d_{r,t} + M'_{r,i}G_t + \Lambda'_{r,i}F_{r,t} + v_{r,it}$$

where $A_{r,i}$, $M_{r,i}$ and $\Lambda_{r,i}$ are $N \times k$, $r_G \times k$, and $r_f \times k$ matrices, $v_{r,it}$ are errors of $x_{r,it}$ distributed independently of the factor structure and across i and r.

The model can be re-written for the specific block r as

$$\underset{(k+1)\times 1}{\mathbf{z}_{r,it}} = \underset{(k+1)\times N}{\mathbf{B}'_{r,i}}\ \underset{N\times 1}{\mathbf{d}_{r,t}} + \underset{(k+1)\times r_G}{\mathbf{C}'_{r,i}}\ \underset{r_G\times 1}{\mathbf{G}_t} + \underset{(k+1)\times r_F}{\mathbf{D}'_{r,i}}\ \underset{r_F\times 1}{\mathbf{F}_{r,t}} + \mathbf{u}_{r,it}$$

where

$$\mathbf{z}_{r,it} = \begin{pmatrix} y_{r,it} \\ \mathbf{x}_{r,it} \end{pmatrix},\ \mathbf{B}_{r,i} = (\alpha_{r,i}, A_{r,i}) \begin{pmatrix} 1 & 0 \\ \beta_{r,i} & I_k \end{pmatrix}$$

$$\mathbf{C}_{r,i} = \left(\mu_{r,i}, M_{r,i}\right) \begin{pmatrix} 1 & 0 \\ \beta_{r,i} & I_k \end{pmatrix}, \mathbf{D}_{r,i} = (\lambda_{r,i}, \Lambda_{r,i}) \begin{pmatrix} 1 & 0 \\ \beta_{r,i} & I_k \end{pmatrix}$$

$$\mathbf{u}_{r,it} = \begin{bmatrix} \epsilon_{r,it} + \beta'_{r,i}\mathbf{v}_{r,it} \\ \mathbf{v}_{r,it} \end{bmatrix}$$

The rank of matrices $\mathbf{C}_{r,i}$ and $\mathbf{D}_{r,i}$ are determined by the rank of the $r_G \times (k+1)$ and $r_F \times (k+1)$ matrices of the unobserved top-level and block-specific factor loadings, respectively.

The model simplifies to that proposed by Pesaran (2006) and Bai (2009): i) When $R = 1$. ii) When there are no block-specific factors, i.e. $\sum_{r=1}^{R} r_{F_r} = 0$.

Assumption A. **Observed Common Effects:**

The $RN \times 1$ vector of observed common effects $d_{r,t}$ is covariance stationary with absolute summable autocovariances, distributed independently of the individual specific errors $\epsilon_{r,it}$ and $\mathbf{v}_{r,it}$. Each d_r is orthogonal to G.

Assumption B. **Unobserved Common Factors**:

B1 Block-specific factors $F_{r,t}$ are covariance stationary such that

$$E\left\|F_{r,t}\right\|^4 \leq M < \infty \ with\ T^{-1}\sum_{t=1}^{T} F_{r,t}F'_{r,t} \to_p \Sigma_{F_r}$$

for some $r_F \times r_F$ positive definite matrix Σ_{F_r}, $r = 1, ..., R$.

B2 The pervasive top-level factor G_t is covariance stationary such that

$$E\left\|G_t\right\|^4 \leq M < \infty \ with\ T^{-1}\sum_{t=1}^{T} G_tG'_t \to_p \Sigma_G$$

for some $r_G \times r_G$ positive definite matrix Σ_G.

B3 Define $H_t = \left[G'_t, F'_{r,t}\right]'$. For a fixed r, assume that

$$T^{-1}\sum_{t=1}^{T} H_tH'_t \to_p \Sigma_H$$

for some positive-definite matrix Σ_H with rank $r_G + r_1 + \cdots + r_R$.

B4 Factors have zero mean, and

$$\sum_{t=1}^{T} G_t F'_{r,t} = 0 \text{ for } r = 1, ..., R$$

Assumption C. Individual-specific Errors:

C1 $\epsilon_{r,it}$, $r = 1, ..., R, i = 1, ..., N, t = 1, ..., T$, are independently across r, i, and t with zero mean and variance σ_i^2, and have a finite fourth-moment.

C2 $\mathbf{v}_{r,it}$, $r = 1, ..., R, i = 1, ..., N, t = 1, ..., T$, are independently across r, i, and t with zero mean and variance $\Sigma_i > 0$, and $\sup_{r;it} E \left\|\mathbf{v}_{r,it}\right\|^4 < \infty$.

C3 $\epsilon_{r,it}$ and $\mathbf{v}_{r,jt'}$ are distributed independently for all r, i, j, t and t'. For each r and i, $\epsilon_{r,it}$ and $\mathbf{v}_{r,it}$ follow linear stationary processes with absolute summable autocovariances.

Assumption D. Factor loadings:

The unobserved factor loadings $\mu_{r,i}, \lambda_{r,i}$ $M_{r,i}$ and $\Lambda_{r,i}$ are independently and identically distributed across r, i, and of the individual specific errors $\epsilon_{r,it}$ and $\mathbf{v}_{r,jt}$, the common observable factors $d_{r,t}$ and the unobserved common factors $(G_t, F_{r,t})$ for all r, i, j, and t with fixed means μ, λ,M and Λ, and finite variances. In particular,

D1 $\lambda_{r,i}$ is either deterministic such that $\left\|\lambda_{r,i}\right\| \leq M < \infty$, or it is stochastic such that $E \left\|\lambda_{r,i}\right\|^4 \leq M < \infty$. In the latter case,

$$N_r^{-1} \Lambda'_r \Lambda_r \to_p \Sigma_{\Lambda_r} > 0$$

for an $r_F \times r_F$ non-random matrix Σ_{Λ_r} for all $r = 1, ..., R$ with a positive constant M.

D2 $\mu_{r,i}$ is either deterministic such that $\left\|\mu_{r,i}\right\| \leq M < \infty$, or it is stochastic such that $E \left\|\mu_{r,i}\right\|^4 \leq M < \infty$ with

$$N_r^{-1} \mu'_r \mu_r \to_p \Sigma_{\mu_r} > 0$$

for an $r_G \times r_G$ non-random matrix Σ_{μ_r} for all $r = 1, ..., R$.

Assumption E. Random Slope Coefficients:

The slope coefficients $\beta_{r,i}$ follow the random coefficient model

$$\beta_{r,i} = \beta + v_{r,i}, \ \mathbf{v}_{r,i} \sim iid\,(0, \Omega_v)$$

for $i = 1, ..., N$ and $r = 1, ..., R$ where $\left\|\beta\right\| < K$, $\left\|\Omega_v\right\| < K$, Ω_v is $k \times k$ symmetric nonnegative definite matrix, and $\mathbf{v}_{r,i}$ are distributed independently of $\mu_{r,j}, \lambda_{r,j}$ $M_{r,j}$ and $\Lambda_{r,j}$, $\epsilon_{r,jt}$, $\mathbf{v}_{r,jt}$, $d_{r,t}$, $F_{r,t}$, and G_t for all r, i, j and t.

Assumption F. Identification of $\beta_{r,i}$:

Identification of the slope coefficients are given by a two-step procedure of cross-sectional averages. In the first step, consider the cross-sectional averages of the individual-specific variables $\mathbf{z}_{r,it}$ in each one of R blocks separately. Define by

$$\overline{\mathbf{z}}_{r,it} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{z}_{r,jt}$$

and let

$$\overline{W_r} = I_T - \overline{H_r}\left(\overline{H_r}'\overline{H_r}\right)^{-}\overline{H_r}'$$

$$W_{\mathbb{F}_r} = I_T - \mathbb{F}_r\left(\mathbb{F}_r'\mathbb{F}_r\right)^{-}\mathbb{F}_r'$$

where

$$\overline{H_r} = \left(D_r, \overline{Z_r}\right) \text{ with } \underset{T\times N}{D_r} = (d_{r,1}, ..., d_{r,T})'$$

and $Z_r = (\bar{\mathbf{z}}_{r,1}, ..., \bar{\mathbf{z}}_{r,T})'$ is the $T \times (k+1)$ matrix of time observations on the cross-sectional averages for each region r and

$$\mathbb{F}_r = (D_r, F_r)$$

where $F_r = (F_{r,1}, ..., F_{r,T})'$ is $T \times r_{F_r}$ data matrices on unobserved block specific factors.

For the second step,

$$\underset{(k+1)RN\times T}{Z^*} = \left[\left(\underset{(k+1)N\times T}{Z_1}\ \underset{T\times T}{\overline{W_1}}\right)', \left(Z_2\overline{W_2}\right)', ..., \left(Z_R\overline{W_R}\right)'\right]'$$

and consider the cross-sectional average of the complete panel and let

$$\overline{W}^* = I_T - \overline{H}^*\left(\overline{H}^{*\prime}\overline{H}^*\right)^{-}\overline{H}^{*\prime}$$

$$\bar{W}_G = I_T - G\left(G'G\right)^{-1}G'$$

where $\overline{H}^* = \overline{Z^*}$, with $\overline{Z^*} = (\bar{\mathbf{z}}_1^*, ..., \bar{\mathbf{z}}_T^*)'$ is the $T \times (k+1)$ matrix of observations on the cross-sectional averages. denote $\bar{W}_{r,t}G_t'$ as G_t^*, then $G^* = (G_1^*, ..., G_T^*)'$ is $T \times r_G$ data matrix on unobserved top-level factors.

F1 Identification of $\beta_{r,i}$:

The $k \times k$ matrix $\hat{\Psi}_{r,iT} = \left(\frac{\mathbf{X}_{r,i}^{*\prime}\overline{W}^*\mathbf{X}_{r,i}^*}{T}\right)$ and $\hat{\Psi}_{r,iG} = \left(\frac{\mathbf{X}_{r,i}^{*\prime}W_G\mathbf{X}_{r,i}^*}{T}\right)$ are nonsingular, and $\hat{\Psi}_{r,iT}, \hat{\Psi}_{r,iG}$ have finite second-order moments for all r and i.

Assumption E can be relaxed allowing for $\beta_{r,i} = \beta_r + v_{r,i}$, $v_{r,i} \sim IID(0, \Omega_v)$ implying that the slope can vary among regions but being the same within the specific block r. Assumption F details the identification strategy of the slope coefficients.

I propose an extended CCE procedure to filter out the full factor space involved in a model whose cross-sectional dependence is driven by a multi-level factor structure. The **Multi-Level Common Correlated Effect Estimators (MLCCE)** consists of two steps.

In the first step, consider each of the $\mathbf{z}_{r,it}$ vectors for each block $r = 1, ..., R$. Note that the cross-sectional dependence in block r is only driven by the block-specific observable common factors, $d_{r,t}$ as well as the unobservable common factors $F_{r,t}$. **The pervasive common factors G_t do not play a role hitherto since none of the remaining blocks are considered** {**why??**}. In

this sense, $\mu_{r,i} = 0$ and $M_{r,i} = 0$ lead to $C_{r,i} = 0$ for all $i = 1, ..., N$. Then, we have

$$\mathbf{z}_{r,it} = \mathbf{B}'_{r,i}\mathbf{d}_{r,t} + \mathbf{D}'_{r,i}\mathbf{F}_{r,t} + \mathbf{u}_{r,it}$$

The cross-sectional averages of the observables of $\bar{\mathbf{z}}_{r,it}$ can work as a proxy for these block-specific factors:

$$\bar{\mathbf{z}}_{r,t} = \bar{\mathbf{B}}'_r\mathbf{d}_{r,t} + \bar{\mathbf{D}}'_r\mathbf{F}_{r,t} + \bar{\mathbf{u}}_{r,t}$$

where

$$\bar{\mathbf{z}}_{r,t} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_{r,it}; \bar{\mathbf{B}}_r = \frac{1}{N}\sum_{i=1}^{N}\mathbf{B}_{r,i}; \bar{\mathbf{D}}_r = \frac{1}{N}\sum_{i=1}^{N}\mathbf{D}_{r,i}; \bar{\mathbf{u}}_{r,t} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_{r,it}$$

Assuming

$$rk\left(\bar{\mathbf{D}}\right) = r_{F_r} < k + 1$$

then

$$F_{r,t} = \left(\bar{\mathbf{D}}_r\bar{\mathbf{D}}'_r\right)^{-1}\bar{\mathbf{D}}_r\left(\bar{\mathbf{z}}_{r,t} - \bar{\mathbf{B}}'_r\mathbf{d}_{r,t} - \bar{\mathbf{u}}_{r,t}\right)$$

Lemma 1 in Pesaran (2006) shows that $\bar{\mathbf{u}}_{r,t} \to_{qm} 0$ as $N \to \infty$, for each t, which implies

$$F_{r,t} = \left(\mathbf{D}\mathbf{D}'\right)^{-1}\mathbf{D}\left(\bar{\mathbf{z}}_{r,t} - \bar{\mathbf{B}}'_r\mathbf{d}_{r,t}\right) \to_{qm} 0$$

where

$$\mathbf{D} = \lim_{N\to\infty}\bar{\mathbf{D}}_r = \tilde{\Lambda}\begin{pmatrix} 1 & 0 \\ \beta_{r,i} & I_k \end{pmatrix}$$

$$\tilde{\Lambda} = E\left(\lambda_{r,i}, \Lambda_{r,i}\right) = \left(\lambda_{r,i}, \Lambda_{r,i}\right) \text{ and } \beta_r = E\left(\beta_{r,i}\right).$$

Therefore, the block-specific factors, $F_{r,t}$, can be approximated by a linear combination of $\mathbf{d}_r$, the cross-sectional averages of the dependent variable, $\bar{y}_{r,t} = \frac{1}{N}\sum_{i=1}^{N} y_{r,it}$, and of the individual-specific regressors, $\bar{\mathbf{x}}_{r,t} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{r,it}$.

Let $\mathbf{Z}^*$ as

$$\underset{(k+1)RN\times T}{\mathbf{Z}^*} = \begin{bmatrix} \underset{(k+1)N\times T}{Z_1} & \underset{T\times T}{\overline{W_1}} \\ Z_2\overline{W_2} \\ \vdots \\ Z_R\overline{W_R} \end{bmatrix}$$

where

$$\underset{T\times T}{\bar{W}_r}\underset{T\times(k+1)N}{Z'_r} = \left(I_T - \overline{H_r}\left(\overline{H_r}'\overline{H_r}\right)^{-}\overline{H_r}'\right)Z'_r$$

which is the residual from the regression of Z'_r on $\overline{H_r} = \left(D_r, \overline{Z_r}\right)$. And similarly,

$$G^* = \left(G\bar{W}_r\right)' = \left\{I_T - \overline{H_r}\left(\overline{H_r}'\overline{H_r}\right)^{-}\overline{H_r}'\right\}G';\ U^* = \left[\left(U_1\bar{W}_1\right)', ..., \left(U_R\bar{W}_R\right)'\right]'$$

where $\overline{W_r} = I_T - \overline{H_r}\left(\overline{H_r}'\overline{H_r}\right)^{-}\overline{H_r}'$. Then, we have:

$$\mathbf{z}_{it}^{*} = \mathbf{C}_i'\mathbf{G}_t^{*} + \mathbf{u}_{it}^{*} \tag{356}$$

after partialing out the effects of the block-specific factors from all the blocks $r = 1, ..., R$ by using the orthogonal projection matrix $\overline{W_r}$. The cross-section averages will result in

$$\bar{\mathbf{z}}_t^{*} = \bar{\mathbf{C}}'\mathbf{G}_t^{*} + \bar{\mathbf{u}}_t^{*}$$

where

$$\bar{\mathbf{z}}_t^{*} = \frac{1}{RN}\sum_{i=1}^{RN}\mathbf{z}_{it}^{*}$$

In the second step, $\bar{H}^{*} = \bar{\mathbf{Z}}^{*}$ becomes an observable proxy for unobserved global factor G_t.

The Common Correlated Effects Mean Group (CCEMG) is an average of the individual MLCCE estimators, $\hat{\beta}_{r,i}$,

$$\hat{\beta}_{CCEMG} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{r,i}$$

where

$$\begin{aligned}
\hat{\beta}_{r,i} &= \left(\mathbf{X}_{r,i}^{*\prime}\bar{\mathbf{W}}^{*}\mathbf{X}_{r,i}^{*}\right)^{-1}\mathbf{X}_{r,i}^{*\prime}\bar{\mathbf{W}}^{*}y_{r,i}^{*} \qquad (357)\\
\mathbf{X}_{r,i}^{*} &= \left(\mathbf{x}_{r,i1}\bar{W}_{r,1}, \mathbf{x}_{r,i2}\bar{W}_{r,2}, ..., \mathbf{x}_{r,iT}\bar{W}_{r,T}\right)\\
y_{r,i}^{*} &= \left(y_{r,i1}\bar{W}_{r,1}, y_{r,i2}\bar{W}_{r,2}, ..., y_{r,iT}\bar{W}_{r,T}\right)
\end{aligned}$$

Theorem 4.1. Under Assumptions A-C, and F1, as $(N,T)_j \to \infty$ and the rank conditions, then $\hat{\beta}_{r,i}$ is a consistent estimator of $\beta_{r,i}$. Furthermore, assuming $\frac{\sqrt{T}}{N} \to 0$ as $(N,T)_j \to \infty$ in the block r, then

$$\sqrt{T}\left(\hat{\beta}_{r,i} - \beta_{r,i}\right) \to_d N\left(0, \Sigma_{\beta_{r,i}}\right)$$

where

$$\Sigma_{\beta_{r,i}} = \sigma_{r,i}^{2}\Sigma_{r,iv}^{-1}(0)\,\Sigma_{r,iv}^{-1}(0)$$

It is possible to consider a CCEMG estimator either for the specific block r or for the full panel. I focus on the more general setting.

Theorem 4.2. Under Assumptions A-E, and F1, then $\hat{\beta}_{CCEMG}$ is asymptotically unbiased for fixed R, and T and as $N \to \infty$. Furthermore, as$(N,T)_j \to \infty$ with R fixed,

$$\sqrt{RN}\left(\hat{\beta}_{CCEMG} - \beta\right) \to_d N\left(0, \Sigma_{CCEMG}\right)$$

where Σ_{CCEMG} can be consistently estimated non-parametrically by

$$\Sigma_{CCEMG} = \frac{1}{R-1}\sum_{r=1}^{R}\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\beta}_{r,i} - \hat{\beta}_{CCEMG}\right)\left(\hat{\beta}_{r,i} - \hat{\beta}_{CCEMG}\right)'$$

Note that the block-specific and global rank conditions are necessary in the Theorem 4.1 but not in 4.2 (**how??**).

When $\beta_{r,i} = \beta$ for all r and i, pooled estimators would gain efficiency. Similar asymptotic analysis can be done. I avoid these details to focus only on the heterogeneity assumption which seems to be more appropriate in macroeconomics and financial applications.

- **Remark**: The MLCCE first augment the regional block regression with $\overline{H_r} = \left(D_r, \overline{Z_r}\right)$, where $\overline{Z_r}$ are designed to approximate the local factors, see the regression (356). Then, augment the global regression (356) with $\bar{\mathbf{z}}_t^*$, where $\bar{\mathbf{z}}_t^*$ are designed to approximate the global factors. He showed via MC that the performance of the MLCCE is satisfactory... See MC section and we may consider similar designs?

- First, I wonder whether **this 2-step approach in (357) is equivalent to our algorithm developed in the previous section ?? Does this cover the model in (351) or in (??)? which is more efficient?**

10.4 The Multi-level PCA approach

- **What about applying the PCA following the international BC literature, say Breitung and Eickmeier (2016) and Choi et al. (2016)? This is quite feasible and worthwhile along with developing the information criteria for determining the number of both global and local factors, though I prefer the approach not affected by the consistent estimation of the number of true factors...**

Here we aim to adopt and extend the approaches advanced by Breitung and Eickmeier (2016) and Choi et al. (2017). We first review a couple of existing approaches (incomplete or not rigorous yet).

10.5 Three-Dimensional Panel Data Models with Factor Structures by Lu and Su (2018)

We consider general three-dimensional panel data models with factor structures, which include both global factors and local factors. This type of model can be applied to many fields in economics, such as international trade, macroeconomics and finance, among others. We will propose a method to determine the number of factors and provide estimators for the factors and factor loadings. We will study the asymptotic properties of our methods, show the consistency and derive the asymptotic distribution of our estimators. Monte Carlo simulations will be used to demonstrate the finite sample performance of our methods. Empirical applications to international trade will also be provided.

Latent factor models have received considerable attention in econometrics. Most factor models have been applied to two-dimensional (2D) panel data. However, more three-dimensional (3D) panel datasets have become available (e.g.,

Mátyás (2017) for a recent review). In this project, we consider estimation and inference for factor models with 3D panel data.

We first consider a 3D pure factor model without any exogenous regressors:

$$y_{ijt} = \lambda_{ij}^{(0)\prime} f_t^{(0)} + \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + \lambda_{ij}^{(2)\prime} f_{jt}^{(2)} + u_{ijt}; i = 1, ..., N; j = 1, ..., M_i, t = 1, ..., T; \tag{358}$$

where y_{ijt} is the observable data, $f_t^{(0)}$ is the global factor, $f_{it}^{(1)}$ and $f_{jt}^{(2)}$ are local factors that depend on i and j, respectively, $\left(\lambda_{ij}^{(0)\prime}, \lambda_{ij}^{(1)\prime}, \lambda_{ij}^{(2)\prime}\right)'$ are their corresponding factor loadings and u_{ijt} is the idiosyncratic error. We assume that the dimensions of $f_t^{(0)}$, $f_{it}^{(1)}$ and $f_{jt}^{(2)}$ are $r^0 \times 1$, $r_i^1 \times 1$ and $r_j^2 \times 1$, respectively and $\lambda_{ij}^{(0)}$, $\lambda_{ij}^{(1)}$, and $\lambda_{ij}^{(2)}$ have the same corresponding dimensions. That is, there are r^0, r_i^1 and r_j^2 global factors, i-specific local factors and j-specific local factors, respectively.

We can also extend to allow the exogenous regressors

$$y_{ijt} = \beta' x_{ijt} + \lambda_{ij}^{(0)\prime} f_t^{(0)} + \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + \lambda_{ij}^{(2)\prime} f_{jt}^{(2)} + u_{ijt}, \tag{359}$$

where x_{ijt} is a $k \times 1$ vector of observable regressors and β is the corresponding slope coefficients. For example, in the international trade application above, x_{ijt} could be the trading costs from country i to country j at year t: Another example of x_{ijt} is the lagged dependent variable, i.e., $x_{ijt} = y_{ij,t-1}$ in the dynamic model. This model is often referred to as panel data models with interactive fixed effects, as we allow the regressor x_{ijt} to be correlated with the unobservable factors.

The model can be used in various ways. The most general 3D linear fixed effect model considered in the literature is

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij} + \alpha_{it} + \alpha_{jt}^* + u_{ijt},$$

where $\gamma_{ij}, \alpha_{it}, \alpha_{jt}^*$ are fixed effects (e.g., Lu, Miao and Su (2017)). This model can be thought of as a special case of our Model (359). We consider large panels where the three dimensions (N, M, T) go to infinity jointly. The goal of this project is to determine $\left(r^0, r_i^1, r_j^2\right)$ and estimate $\left(f_t^{(0)}, f_{it}^{(1)}, f_{jt}^{(2)}\right)$, $\left(\lambda_{ij}^{(0)\prime}, \lambda_{ij}^{(1)\prime}, \lambda_{ij}^{(2)\prime}\right)'$ up to a rotation matrix.

Literature Review There is a large literature on the factor models for 2D panel data. Omitting the j index, models (358) and (359) reduce to

$$y_{it} = \lambda_i' f_t + u_{it}$$

$$y_{it} = \beta' x_{it} + \lambda_i' f_t + u_{it}$$

where f_t and λ_i are $r \times 1$ vector of factor and factor loadings. The literature on factor models for 2D panel data has been developed rapidly. For a detailed review, see Bai and Wang (2016). For the pure factor models, Bai (2003) considers estimation based on principal components analysis (PCA) and develops the

inference theory assuming the number of factor r is known. For the model with exogenous regressors, Bai (2009) and Moon and Weidner (2015,2017) provide estimators based on Gaussian quasi-maximum likelihood estimation (QMLE) and PCA. Pesaran (2006) proposes common correlated effects (CCE) estimators for estimating β.

There are limited number of papers on 3D factor models. See Breitung and Eickmeier (2015), Wang (2016), and Choi et al. (2017). All these papers consider a special case of our pure factor model where there is only one local component, i.e.,

$$y_{ijt} = \lambda_{ij}^{(0)\prime} f_t^{(0)} + \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + u_{ijt}; i = 1, ..., N; j = 1, ..., M_i; t = 1, ..., T;$$

This model is often referred to as **a multi-level factor model**. There are certain limitations of the existing methods. First, most papers assume that numbers of global factors $r^{(0)}$ and local factors $r_i^{(1)}$ are known, e.g., Breitung and Eickmeier (2015) and Wang (2016). Choi et al. (2017) provide information criteria for selecting the number of local factors $r_i^{(1)}$ but assume the number of global factors $r^{(0)}$ is known. Second, there is no inference theory for the estimated factor and factor loading. These papers only establish the consistency of their estimators. Third, usually strong assumptions are imposed on local factors, $f_{it}^{(1)}$. For example, Choi et al. (2017) assume that the local factors are uncorrelated, that is,

$$cov\left(f_{i_1 t}^{(1)}, f_{i_2 t}^{(1)}\right) = 0 \ for \ i_1 \neq i_2;$$

which may not be satisfied in practice. Fourth, these papers only consider pure factor models that do not allow exogenous regressors, x_{ijt}.

Estimation of Pure Factor Models We impose the key assumptions.

Assumption A.1. $f_t^{(0)}, f_{it}^{(1)}$, and $f_{jt}^{(2)}$ are random variables such that (i) $E\left(f_t^{(0)}\right) = 0$, $E\left(f_{it}^{(1)}\right) = 0$ and $E\left(f_{jt}^{(2)}\right) = 0$ for all i, j and t; and (ii) $E\left(f_t^{(0)} f_{it}^{(1)\prime}\right) = 0$; $E\left(f_t^{(0)} f_{jt}^{(2)\prime}\right) = 0$ and $E\left(f_{it}^{(1)} f_{it}^{(1)\prime}\right) = 0$ for all i, j and t:

The uncorrelated assumption allows us to separate the three factors. Define $M = \max\{M_i, i = 1, ..., N\}$: By symmetry, we assume that for each j, the i index runs from $1, ..., N_j$. We can rewrite Model (358) as

$$y_{ijt} = \lambda_{ij}^{(0)\prime} f_t^{(0)} + u_{ijt}^{(0)}, \ u_{ijt}^{(0)} = \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + \lambda_{ij}^{(2)\prime} f_{jt}^{(2)} + u_{ijt} \tag{360}$$

$$y_{ijt}^{(1)} = \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + u_{ijt}^{(1)}, \ y_{ijt}^{(1)} = y_{ijt} - \lambda_{ij}^{(0)\prime} f_t^{(0)}, \ u_{ijt}^{(1)} = \lambda_{ij}^{(2)\prime} f_{jt}^{(2)} + u_{ijt} \tag{361}$$

$$y_{ijt}^{(2)} = \lambda_{ij}^{(2)\prime} f_{jt}^{(2)} + u_{ijt}, \ y_{ijt}^{(2)} = y_{ijt} - \lambda_{ij}^{(0)\prime} f_t^{(0)} - \lambda_{ij}^{(1)\prime} f_{it}^{(1)} \tag{362}$$

In matrix form, we can write

$$Y^{(0)} = F^{(0)} \Lambda^{(0)\prime} + U^{(0)}; \tag{363}$$

$$Y_i^{(1)} = F_i^{(1)}\Lambda_i^{(1)\prime} + U_i^{(1)};\ i = 1, ..., N \tag{364}$$

$$Y_j^{(2)} = F_j^{(2)}\Lambda_j^{(2)\prime} + U_j^{(2)};\ j = 1, ..., M \tag{365}$$

where $y_{ij} = (y_{ij1}, ..., y_{ijT})'$, $y_{ij}^{(1)} = \left(y_{ij1}^{(1)}, ..., y_{ijT}^{(1)}\right)'$, $y_{ij}^{(2)} = \left(y_{ij1}^{(2)}, ..., y_{ijT}^{(2)}\right)'$, $Y_i^{(1)} = \left(y_{i1}^{(1)}, ..., y_{iM_i}^{(1)}\right)$, $Y_j^{(2)} = \left(y_{1j}^{(2)}, ..., y_{N_j j}^{(2)}\right)$, $Y^{(0)} = (y_{11}, y_{12}, ..., y_{1M_1}, ..., y_{N1}, ..., y_{NM_N})$, $F^{(0)} = (f_1, f_2, ..., f_T)'$, $F_i^{(1)} = \left(f_{i1}^{(1)}, f_{i2}^{(1)}, ..., f_{iT}^{(1)}\right)'$, $F_j^{(2)} = \left(f_{j1}^{(2)}, f_{j2}^{(2)}, ..., f_{jT}^{(2)}\right)'$, $\Lambda_i^{(1)} = \left(\lambda_{i1}^{(1)}, \lambda_{i2}^{(1)}, ..., \lambda_{iM_i}^{(1)}\right)'$, $\Lambda_j^{(2)} = \left(\lambda_{1j}^{(2)}, \lambda_{2j}^{(2)}, ..., \lambda_{N_j j}^{(2)}\right)'$, and $\Lambda^{(0)} = \left(\lambda_{11}^{(0)}, \lambda_{12}^{(0)}, ..., \lambda_{1M_1}^{(0)}, ..., \lambda_{N1}^{(0)}, ..., \lambda_{NM_N}^{(0)}\right)$. Definitions of $U^{(0)}$, $U_i^{(1)}$ and $U_j^{(2)}$ are similar to $Y^{(0)}$, $Y_i^{(1)}$ and $Y_j^{(2)}$.

We identify the factors and factor loading sequentially based on the three equations above. We first identify the global factor and global factor loadings based on Model (360). Note that we can treat $U^{(0)}$ or $u_{ijt}^{(0)}$ as a new error term, as $\left(\lambda_{ij}^{(1)\prime} f_{it}^{(1)} + \lambda_{ij}^{(2)\prime} f_{jt}^{(2)}\right)$ is uncorrelated with $f_t^{(0)}$. If we stack the (i, j) indices to one index as in (363), then we can view Model (360) as a standard 2D factor model and apply the PCA method as in Bai (2003). We can determine the number of factors $r^{(0)}$ by maximizing the ratio of two adjacent eigenvalues, as suggested by Ahn and Horenstein (2013). To identify $F^{(0)}$ and $\Lambda^{(0)}$ we need to impose certain normalization condition:

$$\frac{1}{T} F^{(0)\prime} F^{(0)} = I_{r^{(0)}}$$

(an $r^{(0)} \times r^{(0)}$ identity matrix) and $\Lambda^{(0)\prime}\Lambda^{(0)}$ is an $r^{(0)} \times r^{(0)}$ diagonal matrix. Then, we can identify $r^{(0)}$, $\lambda_{ij}^{(0)\prime} F_t^{(0)}$ and a rotated version of $\lambda_{ij}^{(0)}$ and $F_t^{(0)}$.

After identifying $\lambda_{ij}^{(0)\prime} F_t^{(0)}$, we can identify $y_{ijt}^{(1)} = y_{ijt} - \lambda_{ij}^{(0)\prime} F_t^{(0)}$ in Model (361). For each i, we impose the normalization condition that

$$\frac{1}{T} F_i^{(1)\prime} F_i^{(1)} = I_{r^{(1)}}$$

and $\Lambda_i^{(1)\prime}\Lambda_i^{(1)}$ is an $r_i^{(1)} \times r_i^{(1)}$ diagonal matrix. Then for each i, we can identify $r_i^{(1)}$, $\lambda_{ij}^{(1)\prime} F_{it}^{(1)}$ and a rotated version of $\lambda_{ij}^{(1)}$ and $F_{it}^{(1)}$.

After achieving the identification of $\lambda_{ij}^{(0)\prime} F_t^{(0)}$ and $\lambda_{ij}^{(1)\prime} F_{it}^{(1)}$, $y_{ijt}^{(2)}$ in Model (362) is identified. With the normalization condition that $\frac{1}{T} F_j^{(2)\prime} F_j^{(2)} = I_{r^{(2)}}$ and $\Lambda_j^{(2\prime}\Lambda_j^{(2)}$ is an $r_j^{(2)} \times r_j^{(2)}$ diagonal matrix for each j, $r_j^{(2)}$, $\lambda_{ij}^{(2)\prime} F_{jt}^{(2)}$ and a rotated version of $\lambda_{ij}^{(2)}$ and $F_{jt}^{(2)}$ are identified.

We can obtain the initial consistent estimators based on the discussion above. Then we can achieve more efficient estimators through iterations. The detailed estimation algorithm is described in the appendix.

Remark 2.1.1 To ensure A.1(i), we need to demean the data first. Assuming that the data are weakly stationary along the time dimension, we can apply our estimation method to the demeaned data: $y_{ijt} - T^{-1}\sum_{t=1}^{T} y_{ijt}$:

Remark 2.1.2 A.1(ii) is crucial for the identification and often assumed in multi-level factor models, see Assumption 1(ii) in Choi et al. (2017). It can be thought of as a normalization, as it can be satisfied by linear projections and redefining factors and factor loadings.

Estimation of Models with Exogenous Regressors We propose the following iteration estimation method for β: Define the factor components as

$$C_{ijt} = \lambda_{ij}^{(0)\prime} f_t^{(0)} + \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + \lambda_{ij}^{(2)\prime} f_{jt}^{(2)}$$

Estimation for β: Start with an initial estimator of β, iterate the following two steps until certain convergence criterion is satisfied.

Step 1: For a given estimator of $\beta, \hat{\beta}$, apply Algorithm 1 in the appendix to $y_{ijt} - \hat{\beta}' x_{ijt}$ to obtain estimator of C_{ijt}:

$$\hat{C}_{ijt} = \hat{\lambda}_{ij}^{(0)\prime} \hat{f}_t^{(0)} + \hat{\lambda}_{ij}^{(1)\prime} \hat{f}_{it}^{(1)} + \hat{\lambda}_{ij}^{(2)\prime} \hat{f}_{jt}^{(2)}$$

Step 2: Given the estimator of $\hat{C}_{ijt}$, run a regression of $y_{ijt} - \hat{C}_{ijt}$ on x_{ijt} to obtain the OLS estimator $\hat{\beta}$:

The convergence criterion could be such that

$$\frac{\left\|\hat{\beta}_\ell - \hat{\beta}_{\ell-1}\right\|}{\left\|\hat{\beta}_{\ell-1}\right\|} < \epsilon_0 \text{ and } \frac{\sum_{i=1}^N \sum_{j=1}^{M_i} \left\|\hat{\mathbb{C}}_{ij,\ell} - \hat{\mathbb{C}}_{ij,\ell-1}\right\|}{\sum_{i=1}^N \sum_{j=1}^{M_i} \left\|\hat{\mathbb{C}}_{ij,\ell-1}\right\|} < \epsilon_0$$

where ϵ_0 is a small number, $\hat{\mathbb{C}}_{ij,\ell} = \left(\hat{C}_{ij1}, ..., \hat{C}_{ijT}\right)'$ and the subscript ℓ represent the ℓth iteration.

Algorithm Pick up a reasonably large r_{max}; which is the largest number of factors we allow.

1. Initial estimation: Step 01: apply PCA to y_{ijt} based on (2.4).

Let $\tilde{\mu}_k^{(0)}$ be the kth largest eigenvalue of $Y^{(0)}Y^{(0)\prime}/(NMT)$ (a $T \times T$ matrix). The estimated r^0 is

$$r^{(0)} = \arg\max_k \left\{ \frac{\tilde{\mu}_k^{(0)}}{\tilde{\mu}_{k+1}^{(0)}}; k = 1, ..., r_{max} \right\}.$$

The estimated factor $F^{(0)}$ and factor loading $\Lambda^{(0)}$ are respectively. $\tilde{F}^{(0)} = \sqrt{T}$ times eigenvectors corresponding to the $\tilde{r}^{(0)}$ largest eigenvalues of $Y^{(0)}Y^{(0)\prime}$, and

$$\tilde{\Lambda}^{(0)} = \frac{\tilde{F}^{(0)\prime} Y^{(0)}}{T}$$

Let $\tilde{\lambda}_{ij}^{(0)}$ and $\tilde{f}_t^{(0)}$ be the elements of $\tilde{\Lambda}^{(0)}$ and $\tilde{F}^{(0)}$, respectively.

Step 02: For each i apply PCA to $\tilde{y}_{ijt}^{(1)} = y_{ijt} - \tilde{\lambda}_{ij}^{(0)\prime} \tilde{f}_t^{(0)}$ based on (2.5).

For a given i, define the data matrix $\tilde{Y}_i^{(1)} = \left(\tilde{y}_{i1}^{(1)}, ..., \tilde{y}_{iM_i}^{(1)}\right)$ where $\tilde{y}_{ij}^{(1)} = \left(\tilde{y}_{ij1}^{(1)}, ..., \tilde{y}_{ijT}^{(1)}\right)'$. Let $\tilde{\mu}_k^{(1)}$ be the kth largest eigenvalue of $Y_i^{(1)} Y_i^{(1)\prime}/(M_i T)$ (a $T \times T$ matrix). The estimated $r_i^{(1)}$ is

$$\tilde{r}_i^{(1)} = \arg\max_k \left\{ \frac{\tilde{\mu}_k^{(1)}}{\tilde{\mu}_{k+1}^{(1)}}; k = 1, ..., r_{max} \right\}.$$

The estimated factor $F_i^{(1)}$ and factor loading $\Lambda_i^{(1)}$ are respectively: $\tilde{F}_i^{(1)} = \sqrt{T}$ times eigenvectors corresponding to the $\tilde{r}_i^{(1)}$ largest eigenvalues of $Y_i^{(1)} Y_i^{(1)\prime}$, and

$$\tilde{\Lambda}_i^{(1)} = \frac{\tilde{F}_i^{(1)\prime} Y_i^{(1)}}{T}$$

Let $\tilde{\lambda}_{ij}^{(1)}$ and $\tilde{f}_{it}^{(1)}$ be the elements of $\tilde{\Lambda}_i^{(1)}$ and $\tilde{F}_i^{(1)}$, respectively.

Step 03: For each j, apply PCA to $\tilde{y}_{ijt}^{(2)} = y_{ijt} - \tilde{\lambda}_{ij}^{(0)\prime} \tilde{f}_t^{(0)} - \tilde{\lambda}_{ij}^{(1)\prime} \tilde{f}_{it}^{(1)}$ it based on (2.6).

For a given j, define the data matrix $\tilde{Y}_j^{(2)} = \left(\tilde{y}_{ij}^{(2)}, ..., \tilde{y}_{N_j j}^{(2)}\right)$ where $\tilde{y}_{ij}^{(2)} = \left(\tilde{y}_{ij1}^{(2)}, ..., \tilde{y}_{ijT}^{(2)}\right)'$. Let $\tilde{\mu}_k^{(2)}$ be the kth largest eigenvalue of $Y_j^{(2)} Y_j^{(2)\prime}/(N_j T)$ (a $T \times T$ matrix). The estimated $r_j^{(2)}$ is

$$\tilde{r}_j^{(2)} = \arg\max_k \left\{ \frac{\tilde{\mu}_k^{(2)}}{\tilde{\mu}_{k+1}^{(2)}}; k = 1, ..., r_{max} \right\}.$$

The estimated factor $F_j^{(2)}$ and factor loading $\Lambda_j^{(2)}$ are respectively: $\tilde{F}_j^{(2)} = \sqrt{T}$ times eigenvectors corresponding to the $\tilde{r}_j^{(2)}$ largest eigenvalues of $Y_j^{(2)} Y_j^{(2)\prime}$, and

$$\tilde{\Lambda}_j^{(2)} = \frac{\tilde{F}_j^{(2)\prime} Y_j^{(2)}}{T}$$

Let $\tilde{\lambda}_{ij}^{(2)}$ and $\tilde{f}_{jt}^{(2)}$ be the elements of $\tilde{\Lambda}_j^{(2)}$ and $\tilde{F}_j^{(2)}$, respectively.

2. Iteration: We use the subscript ℓ denote the ℓth iteration, $\ell = 1, ...$

Step ℓ1: apply PCA to data

$$\tilde{y}_{ijt,\ell}^{(0)} = y_{ijt} - \tilde{\lambda}_{ij,\ell-1}^{(1)\prime} \tilde{f}_{it,\ell-1}^{(1)} - \tilde{\lambda}_{ij,\ell-1}^{(2)\prime} \tilde{f}_{jt,\ell-1}^{(2)}$$

as in Step 01.

Step ℓ2: apply PCA to

$$\tilde{y}_{ijt,\ell}^{(1)} = y_{ijt} - \tilde{\lambda}_{ij,\ell-1}^{(0)\prime} \tilde{f}_{t,\ell}^{(0)} - \tilde{\lambda}_{ij,\ell-1}^{(2)\prime} \tilde{f}_{jt,\ell-1}^{(2)}$$

as in Step 02.

Step 3: apply PCA to data

$$\tilde{y}_{ijt,\ell}^{(2)} = y_{ijt} - \tilde{\lambda}_{ij,\ell-1}^{(0)\prime} \tilde{f}_{t,\ell}^{(0)} - \tilde{\lambda}_{ij,\ell-1}^{(1)\prime} \tilde{f}_{it,\ell-1}^{(1)}$$

as in Step 03.

The algorithm stops if certain convergence criterion is satisfied, e.g.

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left\|\tilde{\mathbb{C}}_{ij,\ell} - \tilde{\mathbb{C}}_{ij,\ell-1}\right\|}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left\|\tilde{\mathbb{C}}_{ij,\ell-1}\right\|} < \epsilon_0$$

where ϵ_0 is a small number,

$$\tilde{C}_{ijt} = \tilde{\lambda}_{ij}^{(0)\prime} \tilde{f}_t^{(0)} + \tilde{\lambda}_{ij}^{(1)\prime} \tilde{f}_{it}^{(1)} + \tilde{\lambda}_{ij}^{(2)\prime} \tilde{f}_{jt}^{(2)}$$

and $\tilde{\mathbb{C}}_{ij} = \left(\tilde{C}_{ij1}, ..., \tilde{C}_{ijT}\right)'$.

10.6 A panel data model with interactive effects characterized by multilevel non-parallel factors by Li and Yang (2017)

Consider a two-level factor model:

$$\varepsilon_{it}^{s} = \gamma_i' g_t + \lambda_i^{s\prime} f_t^s + u_{it}^s,\ s = 1, ..., S$$

where the second-level factors, f_t^s $(r_s \times 1)$ are parallel to each other and nested under the global factor g_t $(r \times 1)$. If regional shocks are also considered, the factor model is extended to

$$\varepsilon_{it}^{sd} = \gamma_i' g_t + \lambda_i^{s\prime} f_t^s + \delta_i^{d\prime} h_t^d + u_{it}^{sd},\ s = 1, ..., S, d = 1, ..., D \tag{366}$$

where h_t^d $(r_d \times 1)$ are regional factors with loading parameters δ_i^d. h_t^d are neither parallel to nor nested under the industrial factors f_t^s. **The estimators of Wang (2008), Moench, Ng, and Potter (2013) and Bai and Wang (2015) cannot be used.**

The vector form of the multilevel non-parallel factor model is given by

$$\begin{bmatrix} \varepsilon_{it}^{11} & \varepsilon_{it}^{12} & \cdots & \varepsilon_{it}^{1D} \\ \varepsilon_{it}^{21} & \varepsilon_{it}^{22} & \cdots & \varepsilon_{it}^{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{it}^{S1} & \varepsilon_{it}^{S2} & \cdots & \varepsilon_{it}^{SD} \end{bmatrix}_{S\times D} = \begin{bmatrix} \gamma_i' & \lambda_i^{1\prime} & 0 & \cdots & 0 \\ \gamma_i' & 0 & \lambda_i^{2\prime} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_i' & 0 & 0 & \cdots & \lambda_i^{S\prime} \end{bmatrix}_{S\times r} \left(e_D' \otimes \begin{bmatrix} g_t \\ f_t^1 \\ \vdots \\ f_t^S \end{bmatrix} \right)_{r\times D} \tag{367}$$

$$+ \left(e_S \otimes \begin{bmatrix} \delta_i^{1\prime} & \delta_i^{2\prime} & \cdots & \delta_i^{D\prime} \end{bmatrix}\right)_{S\times r^D} \begin{bmatrix} h_t^1 & 0 & \cdots & 0 \\ 0 & h_t^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_t^D \end{bmatrix}_{r^D\times D}$$

$$+ \begin{bmatrix} u_{it}^{11} & u_{it}^{12} & \cdots & u_{it}^{1D} \\ u_{it}^{21} & u_{it}^{22} & \cdots & u_{it}^{2D} \\ \vdots & \vdots & \ddots & \vdots \\ u_{it}^{S1} & u_{it}^{S2} & \cdots & u_{it}^{SD} \end{bmatrix}_{S\times D}$$

where

$$\underset{S\times 1}{e_S} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \ \underset{D\times 1}{e_D} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Remark: Li and Yang (2017) propose the iterative PC algorithm following Bai (2009) with simulation evidence. But, I believe that the model (366) is basically the same as (381). Esp. when expressing (367) with more careful notations, it would be equivalent to (383). Still, it is worthwhile to consider their panel estimator and extend it more rigorously. See also BE2016...

10.7 The 3D Panel Data Models with the Multi-level Factor Structure

- **Here we develop the 3D PCA following Breitung and Eickmeier (2016) and Choi et al. (2017)...**
- **This is quite feasible and worthwhile along with developing the information criteria for determining the number of both global and local factors.**
- **One of my ph.d students, Rui has been working on this extension, mostly focussing on the simulation performance yet.**
- **I will provide more details soon.**

3D Panel Data Models with the Multi-level Factor Structure: An iterative approach Consider the following multi-dimensional factor model:

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij}' G_t + \lambda_{ij}' F_{it}^{(1)} + u_{ijt}, \ i = 1, ..., R, \ j = 1, ..., M_i, \ t = 1, ..., T \tag{368}$$

where $i = 1, ..., R$ indicates the country, $j = 1, ..., M_i$ denotes the sector and $t = 1, ..., T$ is the time period. x_{ijt} is a $k \times 1$ vector of regressors and β is a conformably defined vector containing homogeneous coefficients. $G_t = (G_{1t}, ..., G_{r_{0t}})'$ comprises the $r_0 \times 1$ global factors, and F_{it} collects the $r_i \times 1$ vector of sectoral factors in country i. Stacking (368) for $j = 1, ..., M_i$, we obtain:

$$y_{i.t} = x_{i.t}\beta + (\Gamma_i, \Lambda_i) \begin{pmatrix} G_t \\ F_{it} \end{pmatrix} + u_{i.t} \tag{369}$$

where

$$\underset{M_i \times 1}{y_{i.t}} = \begin{bmatrix} y_{i1t} \\ y_{i2t} \\ \vdots \\ y_{iM_it} \end{bmatrix}, \ \underset{M_i \times 1}{u_{i.t}} = \begin{bmatrix} u_{i1t} \\ u_{i2t} \\ \vdots \\ u_{iM_it} \end{bmatrix}, \ \underset{M_i \times r_0}{\Gamma_i} = \begin{bmatrix} \gamma_{i1}' \\ \gamma_{i2}' \\ \vdots \\ \gamma_{iM_i}' \end{bmatrix}, \ \underset{M_i \times r_i}{\Lambda_i} = \begin{bmatrix} \lambda_{i1}' \\ \lambda_{i2}' \\ \vdots \\ \lambda_{iM_i}' \end{bmatrix}, \ \underset{M_i \times k}{x_{i.t}} = \begin{bmatrix} x_{i1t}' \\ x_{i2t}' \\ \vdots \\ x_{iM_it}' \end{bmatrix}$$

The entire system becomes:

$$y_t = x_t\beta + \Lambda^* F_t^* + u_t \tag{370}$$

where

$$\underset{N \times 1}{y_t} = \begin{bmatrix} y_{1.t} \\ y_{2.t} \\ \vdots \\ y_{R.t} \end{bmatrix}, \ \underset{N \times 1}{u_t} = \begin{bmatrix} u_{1.t} \\ u_{2.t} \\ \vdots \\ u_{R.t} \end{bmatrix}, \ \underset{r^* \times 1}{F_t^*} = \begin{pmatrix} G_t \\ F_{1t} \\ F_{2t} \\ \vdots \\ F_{Rt} \end{pmatrix}$$

$$\underset{N \times k}{x_t} = \begin{bmatrix} x_{1.t} \\ x_{2.t} \\ \vdots \\ x_{R.t} \end{bmatrix}, \underset{N \times r^*}{\Lambda^*} = \begin{pmatrix} \Gamma_1 & \Lambda_1 & 0 & \cdots & 0 \\ \Gamma_2 & 0 & \Lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Gamma_R & 0 & 0 & \cdots & \Lambda_R \end{pmatrix},$$

and

$$N = \sum_{i=1}^{R} M_i, \ r^* = r_0 + \sum_{i=1}^{R} r_i$$

Then, we write the full system as

$$Y = \left(I_T \otimes \beta'\right) X + F^* \Lambda^{*\prime} + U \tag{371}$$

where

$$\underset{T \times N}{Y} = \begin{bmatrix} y_1' \\ \vdots \\ y_T' \end{bmatrix}, \ \underset{T \times r^*}{F^*} = \begin{bmatrix} F_1^{*\prime} \\ \vdots \\ F_T^{*\prime} \end{bmatrix}, \ \underset{T \times N}{U} = \begin{bmatrix} u_1' \\ \vdots \\ u_T' \end{bmatrix}, \underset{Tk \times N}{X} = \begin{bmatrix} x_1' \\ \vdots \\ x_T' \end{bmatrix}.$$

Alternatively, we transpose equation (369) and stack over time as

$$y_i = \left(I_T \otimes \beta'\right) x_i + G\Gamma_i' + F_i\Lambda_i' + u_i$$

where

$$\underset{T\times M_i}{y_i} = \begin{bmatrix} y'_{i.1} \\ y'_{i.2} \\ \vdots \\ y'_{i.T} \end{bmatrix}, \underset{Tk\times M_i}{x_i} = \begin{bmatrix} x'_{i.1} \\ x'_{i.2} \\ \vdots \\ x'_{i.T} \end{bmatrix}, u_i = \begin{bmatrix} u'_{i.1} \\ u'_{i.2} \\ \vdots \\ u'_{i.T} \end{bmatrix}$$

Estimation: The estimation involves two types of iterations, outer iteration and inner iteration. Outer iteration runs between beta and factors while inner iteration runs between country factors and sectoral factors. We use the subscript ℓ denote the ℓth outer iteration, $\ell = 1, ...$

Outer iteration:

Step 1: Run the OLS regression by ignoring factors using the data

$$y_{ijt} = \beta' x_{ijt} + \varepsilon_{ijt}$$

where $\varepsilon_{ijt} = \gamma'_{ij} G_t + \lambda'_{ij} F_{it} + u_{ijt}$, and obtain the initial estimator of β denoted as $\tilde{\beta}^{(0)}$.

Step 2: Apply one of the multi-level factor estimation techniques to

$$\hat{y}^{(0)}_{ijt} = \gamma'_{ij} G_t + \lambda'_{ij} F_{it} + u_{ijt}$$

where $\hat{y}^{(0)}_{ijt} = y_{ijt} - \tilde{\beta}^{(0)\prime} x_{ijt}$. Denote the initial estimators of γ_{ij}, G_t, λ_{ij} and F_{it} as $\tilde{\gamma}^{(0)}_{ij}$, $\tilde{G}^{(0)}_t$, $\tilde{\lambda}^{(0)}_{ij}$ and $\tilde{F}^{(0)}_{it}$.

Step 3: Construct

$$\tilde{y}^{(0)}_{ijt} = y_{ijt} - \tilde{\gamma}^{(0)\prime}_{ij} G^{(0)}_t - \tilde{\lambda}^{(0)}_{ij} \tilde{F}^{(0)}_{it}$$

by subtracting the (estimated) factors from the data. Then, run the following OLS regression:

$$\tilde{y}^{(0)}_{ijt} = \beta' x_{ijt} + \varepsilon_{ijt}$$

and obtain the updated estimator of β denoted as $\tilde{\beta}^{(1)}$.

Step 4: Next, construct the updated residuals by

$$\hat{y}^{(1)}_{ijt} = y_{ijt} - \tilde{\beta}^{(1)\prime} x_{ijt}$$

Then, apply the multi-level factor estimation to

$$\hat{y}^{(1)}_{ijt} = \gamma'_{ij} G_t + \lambda'_{ij} F_{it} + u_{ijt}$$

and the updated estimates of factors and factor loadings, denoted $\tilde{\gamma}^{(1)}_{ij}$, $\tilde{G}^{(1)}_t$, $\tilde{\lambda}^{(1)}_{ij}$ and $\tilde{F}^{(1)}_{it}$.

Repeat Steps 3 and 4 until convegrence. After the ℓth iteration, we have the estimators $\tilde{\beta}^{(\ell)}$, $\tilde{\gamma}_{ij}^{(\ell)}$, $\tilde{G}_t^{(\ell)}$, $\tilde{\lambda}_{ij}^{(\ell)}$ and $\tilde{F}_{it}^{(\ell)}$. The algorithm stops if certain convergence criterion is satisfied, e.g.

$$\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left\|\tilde{\mathbb{C}}_{ij,\ell}-\tilde{\mathbb{C}}_{ij,\ell-1}\right\|<\epsilon_0$$

where ϵ_0 is a small number,

$$\tilde{C}_{ijt}=\tilde{\lambda}_{ij(\ell)}^{(0)\prime}\tilde{f}_{t(\ell)}^{(0)}+\tilde{\lambda}_{ij(\ell)}^{(1)\prime}\tilde{f}_{it(\ell)}^{(1)}$$

and $\tilde{\mathbb{C}}_{ij}=\left(\tilde{C}_{ij1},...,\tilde{C}_{ijT}\right)'$. We follow two different approaches BE16 and Choi et al. (2017) to estimate multi-level factors and describe their algorithm respectively.

The multi-level factor estimation by Choi et al. (2017) Suppose that r_0 and all r_i are given. As in Step 2 and 4, we need to estimate the factors and factor loadings from

$$\begin{aligned}\hat{y}_{ijt}^{(\ell)}&=\gamma_{ij}'G_t+\lambda_{ij}'F_{it}+u_{ijt}\\&=\left[\gamma_{ij}',\lambda_{ij}'\right]\begin{bmatrix}G_t\\F_{it}\end{bmatrix}+u_{ijt}\end{aligned}$$

where $\hat{y}_{ijt}^{(\ell)}=y_{ijt}-\tilde{\beta}^{(\ell)\prime}x_{ijt}$.

Stacking the above equation for $j=1,...,M_i$, we obtain:

$$\begin{aligned}\hat{y}_{i.t}^{(\ell)}&=[\Gamma_i,\Lambda_i]\begin{bmatrix}G_t\\F_{it}\end{bmatrix}+u_{i.t}\\&=\Theta_iK_{it}+u_{i.t}\end{aligned}\tag{372}$$

where $\Theta_i=[\Gamma_i,\Lambda_i]$ and $K_{it}=[G_t,F_{it}]'$. Transpose (372) and further stack over time we obtain

$$\underset{T\times M_i}{\hat{y}_i^{(\ell)}}=\underset{T\times r_0}{G}\,\Gamma_i'+\underset{T\times r_i}{F_i}\,\Lambda_i'+\underset{T\times M_i}{u_i}\tag{373}$$

where $\hat{y}_i^{(\ell)}=[\hat{y}_{i.1}^{(\ell)},...,\hat{y}_{i.T}^{(\ell)}]$, $G=[G_1,...,G_T]'$, $F_i=[F_{i1},...,F_{iT}]'$ and $u_i=[u_{i.1},...,u_{i.T}]'$.

Step $\ell,0$: Choose two sectors, say 1 and 2, apply the PCE to (372) and obtain the estimators of $\underset{T\times r_1^*}{K_1}$ and K_2, denoted by $\hat{K}_1$ and $\hat{K}_2$. Calculate the sample covariance matrices between $\hat{K}_1$ and $\hat{K}_2$ as Σ_{ab} $(a,b=1,2)$. Denote the $r_1^*\times 1$ eigenvector corresponding to an mth largest eigenvalue, μ_m of the canonical correlation matrix, $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$ as p_m. Collect all r_0 eigenvectors as $\underset{r_1^*\times r_0}{p}=[p_1,\ldots,p_{r^{(0)}}]$. Then, we obtain the initial estimate of G by

$$\underset{T\times r_0}{\hat{G}^{(\ell,0)}}=\hat{K}_1p,$$

where $(\ell, 0)$ means the initial estimator in ℓth step of the outer iteration. This is the canonical correlation analysis (CCA). CCA searches for a linear combination of two sets of variables which yields the mth maximum squared correlation equal to the mth largest eigenvalue of $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$. The corresponding eigenvector is the coefficients of such combination. Therefore, if K_1 and K_2 contain the r_0 same global factors, then there must be r_0 eigenvalues which are equal to 1. **BE16 suggests choosing the sector pairs who have the largest canonical correlation. Choi et al. suggests choosing the pairs whose the sample mean of canonical correlations is largest (but in their code they choose the pairs with largest canonical correlation).**

Step ℓ1: After obtaining the initial country factors $\hat{G}^{(\ell,0)}$, we can estimate the sectoral factors by projecting out $\hat{G}^{(\ell,0)}$. Specifically, the initial estimator of ith sectoral factor, $\hat{F}_i^{(\ell,0)}$ is given by $\sqrt{T}$ times the eigenvectors matrix corresponding to the r_i largest eigenvalues of the matrix, $\hat{y}_{Hi}^{(\ell)}\hat{y}_{Hi}^{(\ell)\prime}$, where $\hat{y}_{Hi}^{(\ell)} = H\hat{y}_i^{(\ell)}$ and $H = I - \hat{G}^{(\ell,0)}(\hat{G}^{(\ell,0)}\hat{G}^{(\ell,0)})^{-1}\hat{G}^{(\ell,0)}$. The initial factor loading matrix can be estimated by $\hat{\Lambda}_i^{(\ell,0)} = (1/T)\hat{y}_{Hi}^{(\ell,0)\prime}\hat{F}_i^{(\ell,0)}$.

Step ℓ2: Now, we can update the country factors and factor loadings using the results of the previous step. Rewrite (372) as

$$\begin{aligned}\hat{y}_{i.t}^{(\ell)} - \hat{\Lambda}_i^{(\ell,0)}\hat{F}_{i.t}^{(\ell,0)} &= \Gamma_i G_t + u_{i.t} - (\hat{\Lambda}_i^{(\ell,0)}\hat{F}_{i.t}^{(\ell,0)} - \Lambda_i F_{i.t}) \\ &= \Gamma_i G_t + \dot{u}_{i.t}\end{aligned}$$

Stacking them across sectors we obtain:

$$\underset{(N\times 1)}{\begin{bmatrix}\hat{y}_{1.t}^{(\ell)} - \hat{\Lambda}_1^{(\ell,0)}\hat{F}_1^{(\ell,0)} \\ \vdots \\ \hat{y}_{N.t}^{(\ell)} - \hat{\Lambda}_N^{(\ell,0)}\hat{F}_N^{(\ell,0)}\end{bmatrix}} = \underset{(N\times r_0)}{\begin{bmatrix}\Gamma_1 \\ \vdots \\ \Gamma_N\end{bmatrix}} G_t + \underset{(N\times 1)}{\begin{bmatrix}\dot{u}_{1.t} \\ \vdots \\ \dot{u}_{N.t}\end{bmatrix}}$$

Then we apply PCE to above equation and therefore update $\hat{G}^{(\ell,1)}$ and obtain its factor loading matrix $\hat{\Gamma}_i^{(\ell,1)}$. They will be the final estimate of country factor and loading for ℓth outer iteration.

Step ℓ3: Finally we use the updated country factors and factor loadings to update the sectoral factors and corresponding factor loadings using PCE for each sector using

$$\hat{y}_{i.t}^{(\ell)} - \hat{\Gamma}_i^{(\ell,1)}\hat{G}_t^{(\ell,1)} = \Lambda_i F_{i.t} + \ddot{u}_{i.t}.$$

The estimators $\hat{F}^{(\ell,1)}$ and $\hat{\Lambda}_i^{(\ell,1)}$ will be the final estimator for sectoral factors and factor loadings for ℓth outer iteration. Rewrite each element of $\hat{\Gamma}_i^{(\ell,1)}$, $\hat{G}^{(\ell,1)}$, $\hat{\Lambda}_i^{(\ell,1)}$ and $\hat{F}_i^{(\ell,1)}$ as $\tilde{\Gamma}_i^{(\ell)}$, $\tilde{G}^{(\ell)}$, $\tilde{\Lambda}_i^{(\ell)}$ and $\tilde{F}_i^{(\ell)}$ and they will be inputs in the $\ell+1$th iteration.

The multi-level factor estimation by BE16 In step 4 of the ℓth iteration, given the estimator of $\tilde{\beta}^{(\ell)}$ we estimate the following model:

$$\hat{y}_{ijt}^{(\ell)} = \gamma_{ij}' G_t + \lambda_{ij}' F_{it} + u_{ijt}. \tag{374}$$

Assume that the idiosyncratic components are identically and independent normally distributed (i.i.d.) across i, j and t with $E(u_{ijt}^2) = \sigma^2$. The estimator remains consistent if the errors are heteroskedastic and autocorrelated, cf. Wang (2010). Using equation (371), treating the factors and factor loadings as unknown parameters yields the log-likelihood function

$$\mathcal{L} = const - \frac{N^*T}{2} log(\sigma^2) - \frac{1}{2\sigma^2} tr[(\tilde{Y}^{(\ell)} - F^*\Lambda^{*\prime})(\tilde{Y}^{(\ell)} - F^*\Lambda^{*\prime})']$$

where $\tilde{Y}^{(\ell)} = Y - \left(I_T \otimes \beta^{(\ell)\prime}\right) X$. The maximization of the likelihood function is equivalent to minimizing the sum of squared residuals (RSS)

$$\begin{aligned} S(F, \Lambda) &= \sum_{t=1}^{T} (\tilde{y}_t^{(\ell)} - \Lambda^* F_t^*)' (\tilde{y}_t^{(\ell)} - \Lambda^* F_t^*) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{t=1}^{T} (\hat{y}_{ijt}^{(\ell)} - \gamma_{ij}' G_t - \lambda_{ij}' F_{it})^2 \end{aligned} \tag{375}$$

We use sequential sub-iteration to minimize the objective function.

Inner iteration: Step $\ell 0$: Employ the same procedure as Step $\ell 0$ and Step $\ell 1$ described above so that we obtain the initial estimators $\hat{G}^{(\ell,0)}$ and $\hat{F}_i^{(\ell,0)}$, the same as in Choi et al.'s approach

Step $\ell 1$: Run N time series regressions of the form

$$\hat{y}_{ijt}^{(\ell)} = \gamma_{ij}' \hat{G}_t^{(\ell,0)} + \lambda_{ij}' \hat{F}_{it}^{(\ell,0)} + u_{ijt}.$$

Collecting all the estimated factor loadings $\gamma_{ij}^{(\ell,0)}$ and $\lambda_{ij}^{(\ell,0)}$ we can construct the loading matrix for the full system as

$$\underset{N \times r^*}{\hat{\Lambda}^{*(\ell,0)}} = \begin{pmatrix} \hat{\Gamma}_1^{(\ell,0)} & \hat{\Lambda}_1^{(\ell,0)} & 0 & \cdots & 0 \\ \hat{\Gamma}_2^{(\ell,0)} & 0 & \hat{\Lambda}_2^{(\ell,0)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Gamma}_R^{(\ell,0)} & 0 & 0 & \cdots & \hat{\Lambda}_R^{(\ell,0)} \end{pmatrix}$$

Step $\ell 2$: For each t, run the following cross section regression similar to equation (370) treating $\hat{\Lambda}^{*(\ell,0)}$ as the data matrix

$$\hat{y}_t^{(\ell)} = \hat{\Lambda}^{*(\ell,0)} F_t^* + u_t.$$

Then we can obtain the OLS estimator of F_t^* as

$$\hat{F}_t^{(\ell,1)} = \begin{pmatrix} \hat{G}_t^{(\ell,1)} \\ \hat{F}_{1t}^{(\ell,1)} \\ \hat{F}_{2t}^{(\ell,1)} \\ \vdots \\ \hat{F}_{Rt}^{(\ell,1)} \end{pmatrix} = (\hat{\Lambda}^{*(\ell,0)\prime} \hat{\Lambda}^{*(\ell,0)\prime})^{-1} \hat{\Lambda}^{*(\ell,0)\prime} \hat{y}_t^{(\ell)}.$$

Step $\ell 3$: After obtaining updated factors, we can use them to update the factor loadings by running N time series regressions as in Step $\ell 1$. Denote the updated factor loadings as $\hat{\gamma}_{ij}^{(\ell,1)}$, $\hat{\lambda}_{ij}^{(\ell,1)}$ and the full loading matrix as $\hat{\Lambda}^{*(\ell,1)}$.

Repeat Step $\ell 1$ to $\ell 3$ until the RSS no longer decreases. It is easy to see that

$$S(\hat{F}^{*(\ell,0)}, \hat{\Lambda}^{*(\ell,0)}) \geq S(\hat{F}^{*(\ell,1)}, \hat{\Lambda}^{*(\ell,1)}) \geq S(\hat{F}^{*(\ell,2)}, \hat{\Lambda}^{*(\ell,2)}) \geq \dots$$

since in each step the previous estimators are contained in the parameter space of the subsequent least-squares estimators. Hence the next estimation step cannot yield a larger RSS. Any fixed point is characterized by the condition

$$\hat{\Lambda}^{*\prime}\hat{Y}^{(\ell)\prime}\hat{Y}^{(\ell)}(I - \hat{\Lambda}^{*}(\hat{\Lambda}^{*\prime}\hat{\Lambda}^{*})^{-1}\hat{\Lambda}^{*\prime}) = 0$$

which results from the fact that the sum of squared residuals does no longer decrease whenever the estimated factors and factor loadings are orthogonal to the residuals of the previous step. In order to ensure the uniqueness of the solution we need to adopt the same restrictions as in PC. We need to make country factors and sectoral factors orthogonal and all factors should have unit variance. Suppose that the sub-iteration stops at κth iteration, we first regress each sectoral factors on country factors as

$$\hat{F}_{izt}^{(\ell,\kappa)} = b'\hat{G}_{t}^{(\ell,\kappa)} + v_{izt}, \quad i = 1, \dots, R, z = 1, \dots, r_i.$$

We use residuals of these regressions as updated sectoral factors denoted as $\bar{F}_{it}^{(\ell,\kappa)}$. Now country factors are orthogonal to all sectoral factors. Given $\hat{G}_{t}^{(\ell,\kappa)}$ and $\bar{F}_{it}^{(\ell,\kappa)}$, the corresponding factor loadings $\bar{\Gamma}_{i}^{(\ell,\kappa)}$ and $\bar{\Lambda}_{i}^{(\ell,\kappa)}$ can be updated using OLS regression as in step $\ell 1$. The normalized country factors can be obtained as the r_0 PCs of the estimated common components resulting from the nonzero eigenvalues and the associated eigenvectors of the matrix

$$\bar{\Gamma}_{i}^{(\ell,\kappa)} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{G}_{t}^{(\ell,\kappa)} \hat{G}_{t}^{(\ell,\kappa)'} \right) \bar{\Gamma}_{i}^{(\ell,\kappa)'}.$$

We denote the final estimator of the country factors and factor loadings as $\tilde{G}_{t}^{(\ell)}$. Similarly for each sector the normalized sectoral factors can be obtained as the r_i PCs of the estimated common components resulting from the nonzero eigenvalues and the associated eigenvectors of the matrix

$$\bar{\Lambda}_{i}^{(\ell,\kappa)} \left(\frac{1}{T} \sum_{t=1}^{T} \bar{F}_{t}^{(\ell,\kappa)} \bar{F}_{t}^{(\ell,\kappa)'} \right) \bar{\Lambda}_{i}^{(\ell,\kappa)'}.$$

The final estimators of the sectoral factors are denoted by $\tilde{F}_{it}^{(\ell)}$. Finally run OLS regressions as in Step $\ell 1$ we obtain the final estimators of factor loadings, denoted by $\tilde{\gamma}_{ij}^{(\ell)}$ and $\tilde{\lambda}_{ij}^{(\ell)}$.

10.8 Monte Carlo Simulation A

We consider three different estimations, OLS top-down approach and Choi et al's approach. We drop BE's approach since the algorithm is extremely time consuming. In our MC simulation, we set $M_1 = M2 = ...M_N = M$ so that sectors have the same number of individuals. We consider the combination of $M = 10, 50, 100$, $R = 2, 10$ and $T = 50, 100$. The number country and sectoral factors r_0 and r_i are all set to be 1. The data are generated by

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij} G_t + \lambda_{ij} F_{it} + u_{ijt}$$

where $\beta_1 = 2$ and $\beta_2 = 1$ and the second regressor $x_{ij,2t}$ is exogenous via $x_{ij,2t} \sim U(0, 10)$, while the first regressor $x_{ij,1t}$ is generated by

$$x_{ij,1t} = \gamma_{ij} G_t + \lambda_{ij} F_{it} + \gamma_{ij} \phi_{1t} + \lambda_{ij} \phi_{2t} + \phi_{3,ij} G_t + \phi_{4,ij} F_{it} + w_{ijt}$$

ϕ_{1t}, ϕ_{2t}, $\phi_{3,ij}$ and $\phi_{4,ij} \sim U(0.5, 1.5)$, $w_{ijt} \sim N(0, 1)$. The other parameters except factors are set as γ_{ij}, $\lambda_{ij} \sim U(0.2, 0.5)$ and $u_{ijt} \sim N(0, 1)$. Both country and sectoral factors are generated by independent AR(1) processes as

$$G_t = \alpha^G G_{t-1}^{(0)} + v_t^G$$
$$F_{it} = \alpha_i^F F_{i,t-1}^{(1)} + v_{it}^F$$

where α^G, $\alpha_i^F \sim U(0.2, 0.5)$ and v_t^G, $v_{it}^F \sim N(0, 1)$.

(lr)5-6

ts top-down PC approach and Choi represents Choi et al. (2017) approach described above. The simulation is repeated 1000 times.

10.9 Monte Carlo Simulation B

We consider three different estimations, OLS top-down approach and Choi et al's approach. We drop BE's approach since the algorithm is extremely time consuming. IN our MC simulation, we set $M_1 = M2 = ...M_N = M$ so that sectors have the same number of individuals. We consider the combination of $M = 10, 50, 100$, $R = 2, 20$ and $T = 50, 100$. The number country and sectoral factors r_0 and r_i are all set to be 1. The data are generated by

$$y_{ijt} = \beta' x_{ijt} + \gamma_{ij} G_t + \lambda_{ij} F_{it} + u_{ijt}$$

where $\beta_1 = 2$ and $\beta_2 = 1$ and the second regressor $x_{ij,2t}$ is exogenous via $x_{ij,2t} \sim N(1,1)$, while the first regressor $x_{ij,1t}$ is generated as

$$x_{ij,1t} = \zeta_{ij} G_t + \eta_{ij} F_{it} + w_{ijt}$$

We allow ζ_{ij} and γ_{ij} are correlated and so are η_{ij} and λ_{ij} via

$$\begin{bmatrix} \gamma_{ij} \\ \zeta_{ij} \end{bmatrix} \sim MN\left(\iota, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}\right), \begin{bmatrix} \lambda_{ij} \\ \eta_{ij} \end{bmatrix} \sim MN\left(\iota, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}\right)$$

where $\iota = [1, 1]'$. The error terms $u_{ijt}, w_{ijt} \sim N(0, 1)$. Both country and sectoral factors are generated by stationary AR(1) processes as

$$G_t = \alpha^G G_{t-1} + v_t^G$$
$$F_{it} = \alpha_i^F F_{i,t-1} + v_{it}^F$$

where α^G, $\alpha_i^F = 0.5$. We draw the error terms from a multivariate normal distribution as $[v_t^G, v_{1t}^F, \ldots, v_{Rt}^F] \sim MN(0, \Sigma_v)$, where

$$\Sigma_v = \begin{bmatrix} 1-(\alpha^G)^2 & 0 & 0 & \ldots & 0 \\ 0 & 1-(\alpha_1^F)^2 & 0.2 & \ldots & 0.2 \\ 0 & 0.2 & 1-(\alpha_2^F)^2 & \ldots & 0.2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0.2 & 0.2 & \ldots & 1-(\alpha_R^F)^2 \end{bmatrix}$$

so that each factor has unit variance while country factor is orthogonal to sectoral factors and sectoral factors are mutually correlated. In addition to β, we also pay attention to the estimated factor $\tilde{G}$ and $\tilde{F}_i$. The precision of the estimation for factors is measured by trace ratio. The trace ratio can be expressed as

$$tr(\tilde{G}) = \frac{tr(G'\tilde{G}(\tilde{G}'\tilde{G})^{-1}\tilde{G}'G)}{tr(G'G)}$$

and

$$tr(\tilde{F}_i) = \frac{tr(F_i'\tilde{F}_i(\tilde{F}_i'\tilde{F}_i)^{-1}\tilde{F}_i'F_i)}{tr(F_i'F_i)}$$

for country and sectoral factors respectively, where G and F_i are the true factors, $\tilde{G}$ and $\tilde{F}_i$ are the estimators. A value of the trace ratio closer to one implies better performance of the factor estimates. We also consider the size and power of the t test for β. The standard error for the residual is estimated from the residuals of the pooled regression after correcting for unobserved factors as

$$y_{ijt} - \tilde{\gamma}_{ij}^{(\ell)}\tilde{G}_t^{(\ell)} - \tilde{\lambda}_{ij}^{(\ell)}\tilde{F}_{it}^{(\ell)} = \beta' x_{ijt} + u_{ijt} \tag{376}$$

where $\tilde{\gamma}_{ij}^{(\ell)}$, $\tilde{G}_t^{(\ell)}$, $\tilde{\lambda}_{ij}^{(\ell)}$ and $\tilde{F}_{it}^{(\ell)}$ are the final estimates of the factors and factor loadings from the outer iteration. The estimated standard error for the residual is therefore

$$\tilde{\sigma} = \sqrt{\frac{\sum_{i=1}^{R}\sum_{j=1}^{M_i}\sum_{t=1}^{T}\tilde{u}_{ijt}^2}{NT-k}}$$

where $\tilde{u}_{ijt}$ is the residual from regression (376). Then the standard error can be estimated from

$$\tilde{\Sigma}_\beta = \left[(X'^{-1}\tilde{\sigma}^2\right]^{-1/2}$$

sents Choi et al. (2017) approach described above. The simulation is repeated 1000 times. Time indicates the average computational time fo

al. (2017) approach described above. The simulation is repeated 1000 times. The trace ratio for local factors are the trace ratios averaged a

					TD		Choi	
(lr)5-6(lr)7-8 M	R	T	null	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
10	2	50	1.8	1	1	1	1
			1.85	1	1	1	1
			1.9	0.984	0.955	0.983	0.952
			1.95	0.769	0.605	0.792	0.599
			2	0.493	0.228	0.491	0.229
			2.05	0.761	0.629	0.735	0.625
			2.1	0.974	0.961	0.966	0.961
			2.15	0.998	1	0.999	0.999
			2.2	1	1	1	1
10	2	100	1.8	1	1	1	1
			1.85	1	1	1	1
			1.9	0.998	0.998	0.997	0.998
			1.95	0.901	0.824	0.911	0.817
			2	0.494	0.223	0.486	0.229
			2.05	0.905	0.824	0.891	0.823
			2.1	1	0.997	0.999	0.998
			2.15	1	1	1	1
			2.2	1	1	1	1
10	10	50	1.8	1	1	1	1
			1.85	1	1	1	1
			1.9	1	1	1	1
			1.95	1	1	1	1
			2	0.419	0.194	0.422	0.192
			2.05	1	1	1	1
			2.1	1	1	1	1
			2.15	1	1	1	1
			2.2	1	1	1	1
10	10	100	1.8	1	1	1	1
			1.85	1	1	1	1
			1.9	1	1	1	1
			1.95	1	1	1	1
			2	0.46	0.242	0.481	0.247
			2.05	1	1	1	1
			2.1	1	1	1	1
			2.15	1	1	1	1
			2.2	1	1	1	1

Size and power

					TD		Choi	
(lr)5-6(lr)7-8	M	R	T	null	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
	50	2	50	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	0.993	0.983	0.993	0.983
				2	0.442	0.179	0.455	0.177
				2.05	0.994	0.989	0.994	0.989
				2.1	1	1	1	1
				2.15	1	1	1	1
				2.2	1	1	1	1
	50	2	100	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	1	1	1
				2	0.436	0.211	0.43	0.207
				2.05	1	0.998	1	0.998
				2.1	1	1	1	1
				2.15	1	1	1	1
Size and power				2.2	1	1	1	1
	50	10	50	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	1	1	1
				2	0.425	0.176	0.432	0.178
				2.05	1	1	1	1
				2.1	1	1	1	1
				2.15	1	1	1	1
				2.2	1	1	1	1
	50	10	100	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	1	1	1
				2	0.419	0.194	0.422	0.192
				2.05	1	1	1	1
				2.1	1	1	1	1
				2.15	1	1	1	1
				2.2	1	1	1	1

					TD		Choi	
(lr)5-6(lr)7-8	M	R	T	null	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
	100	2	50	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	0.999	1	0.998
				2	0.405	0.21	0.41	0.211
				2.05	1	1	1	1
				2.1	1	1	1	1
				2.15	1	1	1	1
				2.2	1	1	1	1
	100	2	100	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	1	1	1
				2	0.413	0.174	0.408	0.172
				2.05	1	1	1	1
				2.1	1	1	1	1
				2.15	1	1	1	1
Size and power				2.2	1	1	1	1
	100	10	50	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	1	1	1
				2	0.414	0.172	0.424	0.174
				2.05	1	1	1	1
				2.1	1	1	1	1
				2.15	1	1	1	1
				2.2	1	1	1	1
	100	10	100	1.8	1	1	1	1
				1.85	1	1	1	1
				1.9	1	1	1	1
				1.95	1	1	1	1
				2	0.403	0.184	0.409	0.186
				2.05	1	1	1	1
				2.1	1	1	1	1
				2.15	1	1	1	1
				2.2	1	1	1	1

10.10 Potential Dataset

The monthly housing price data of 70 cities from Nation Bureau of Statistics of China. For Y, the dependent variable, we have a panel of $T = 91$ with different N for different grouping schemes. For X, the regressors, we have candidates such as the average disposable income, the quantity of loan etc. X is yet to be col-

Different Grouping Scheme	8*Housing Price for 70 Cities	First Tier City Second Tier City Third Tier City
		Newly Built Second hand
		Large ($Size > 144m^2$) Medium ($90m^2 < Size \leq 144m^2$) Small ($Size \leq 90m^2$)

10.11 Digressions: Factor Representations

The following factor representations will be useful for enhancing our understanding.

The benchmark case of Lu and Xu Consider the following generic multi-dimensional factor model:

$$y_{ijt} = \lambda_{ij}^{(0)\prime} f_t^{(0)} + \lambda_{ij}^{(1)\prime} f_{it}^{(1)} + \lambda_{ij}^{(2)\prime} f_{jt}^{(2)} + u_{ijt},\ i = 1, ..., N,\ j = 1, ..., M,\ t = 1, ..., T \tag{377}$$

where $i = 1, ..., N$ indicates the source country, the index $j = 1, ..., M$ denotes the destination country, and $t = 1, ..., T$ stands for the time period. (For simplicity we fix the number of i and j indices at N and M, which can be generalised in the different applications.) $f_t^{(0)} = \left(f_{1t}^{(0)}, ..., f_{r^0 t}^{(0)}\right)'$ comprises the $r^{(0)} \times 1$ global factors, $f_{it}^{(1)}$ collects the $r_i^{(1)} \times 1$ vector of source-factors in country i, and $f_{jt}^{(2)}$ collects the $r_j^{(2)} \times 1$ vector of destination factors in country j. The idiosyncratic component, u_{ijt} is assumed to satisfy an approximate factor model. Stacking (377) for $j = 1, ..., M$, we obtain:

$$\begin{aligned}
\begin{bmatrix} y_{i1t} \\ y_{i2t} \\ \vdots \\ y_{iMt} \end{bmatrix} &= \begin{bmatrix} \lambda_{i1}^{(0)\prime} \\ \lambda_{i2}^{(0)\prime} \\ \vdots \\ \lambda_{iM}^{(0)\prime} \end{bmatrix} f_t^{(0)} + \begin{bmatrix} \lambda_{i1}^{(1)\prime} \\ \lambda_{i2}^{(1)\prime} \\ \vdots \\ \lambda_{iM}^{(1)\prime} \end{bmatrix} f_{it}^{(1)} + \begin{bmatrix} \lambda_{i1}^{(2)\prime} f_{1t}^{(2)} \\ \lambda_{i2}^{(2)\prime} f_{2t}^{(2)} \\ \vdots \\ \lambda_{iM}^{(2)\prime} f_{Mt}^{(2)} \end{bmatrix} + \begin{bmatrix} u_{i1t} \\ u_{i2t} \\ \vdots \\ u_{iMt} \end{bmatrix} \\
&= \begin{bmatrix} \lambda_{i1}^{(0)\prime} \\ \lambda_{i2}^{(0)\prime} \\ \vdots \\ \lambda_{iM}^{(0)\prime} \end{bmatrix} f_t^{(0)} + \begin{bmatrix} \lambda_{i1}^{(1)\prime} \\ \lambda_{i2}^{(1)\prime} \\ \vdots \\ \lambda_{iM}^{(1)\prime} \end{bmatrix} f_{it}^{(1)} + \begin{bmatrix} \lambda_{i1}^{(2)\prime} & 0 & \cdots & 0 \\ 0 & \lambda_{i2}^{(2)\prime} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{iM}^{(2)\prime} \end{bmatrix} \begin{bmatrix} f_{1t}^{(2)} \\ f_{2t}^{(2)} \\ \vdots \\ f_{Mt}^{(2)} \end{bmatrix} + \begin{bmatrix} u_{i1t} \\ u_{i2t} \\ \vdots \\ u_{iMt} \end{bmatrix}
\end{aligned}$$

which can be compactly written as

$$y_{i.t} = \Lambda_i^{(0)} f_t^{(0)} + \Lambda_i^{(1)} f_{it}^{(1)} + \hat{\Lambda}_i^{(2)} F_t^{(2)} + u_{i.t} = \left(\Lambda_i^{(0)}, \Lambda_i^{(1)}, \hat{\Lambda}_i^{(2)}\right) \begin{pmatrix} f_t^{(0)} \\ f_{it}^{(1)} \\ F_t^{(2)} \end{pmatrix} + u_{i.t} \tag{378}$$

where $r^{(2)} = \sum_{j=1}^{M} r_j^{(2)}$,

$$\underset{M\times 1}{y_{i.t}} = \begin{bmatrix} y_{i1t} \\ y_{i2t} \\ \vdots \\ y_{iMt} \end{bmatrix}, \quad \underset{M\times 1}{u_{i.t}} = \begin{bmatrix} u_{i1t} \\ u_{i2t} \\ \vdots \\ u_{iMt} \end{bmatrix}, \quad \underset{r^{(2)}\times 1}{F_t^{(2)}} = \begin{bmatrix} f_{1t}^{(2)} \\ f_{2t}^{(2)} \\ \vdots \\ f_{Mt}^{(2)} \end{bmatrix}$$

$$\underset{M\times r^{(0)}}{\Lambda_i^{(0)}} = \begin{bmatrix} \lambda_{i1}^{(0)\prime} \\ \lambda_{i2}^{(0)\prime} \\ \vdots \\ \lambda_{iM}^{(0)\prime} \end{bmatrix}, \quad \underset{M\times r_i^{(1)}}{\Lambda_i^{(1)}} = \begin{bmatrix} \lambda_{i1}^{(1)\prime} \\ \lambda_{i2}^{(1)\prime} \\ \vdots \\ \lambda_{iM}^{(1)\prime} \end{bmatrix}, \quad \underset{M\times r^{(2)}}{\mathring{\Lambda}_i^{(2)}} = \begin{bmatrix} \lambda_{i1}^{(2)\prime} & 0 & \cdots & 0 \\ 0 & \lambda_{i2}^{(2)\prime} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{iM}^{(2)\prime} \end{bmatrix}$$

The entire system representing all $i = 1, ..., N$ source countries becomes:

$$\begin{pmatrix} y_{1.t} \\ y_{2.t} \\ \vdots \\ y_{N.t} \end{pmatrix} = \begin{pmatrix} \Lambda_1^{(0)} & \Lambda_1^{(1)} & 0 & \cdots & 0 & \mathring{\Lambda}_1^{(2)} \\ \Lambda_2^{(0)} & 0 & \Lambda_2^{(1)} & \cdots & 0 & \mathring{\Lambda}_2^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \Lambda_N^{(0)} & 0 & 0 & \cdots & \Lambda_N^{(1)} & \mathring{\Lambda}_N^{(2)} \end{pmatrix} \begin{pmatrix} f_t^{(0)} \\ f_{1t}^{(1)} \\ f_{2t}^{(1)} \\ \vdots \\ f_{Nt}^{(1)} \\ F_t^{(2)} \end{pmatrix} + \begin{pmatrix} u_{1.t} \\ u_{2.t} \\ \vdots \\ u_{N.t} \end{pmatrix}$$

which can be compactly written as

$$y_t = \Lambda F_t + u_t$$

where $r = r^{(0)} + r^{(1)} + r^{(2)}$ with $r^{(1)} = \sum_{i=1}^{N} r_i^{(1)}$,

$$\underset{NM\times 1}{y_t} = \begin{bmatrix} y_{1.t} \\ y_{2.t} \\ \vdots \\ y_{M.t} \end{bmatrix}, \quad \underset{NM\times 1}{u_t} = \begin{bmatrix} u_{1.t} \\ u_{2.t} \\ \vdots \\ u_{M.t} \end{bmatrix}, \quad \underset{r\times 1}{F_t} = \begin{pmatrix} f_t^{(0)} \\ f_{1t}^{(1)} \\ f_{2t}^{(1)} \\ \vdots \\ f_{Nt}^{(1)} \\ F_t^{(2)} \end{pmatrix}$$

$$\underset{NM\times r}{\Lambda} = \begin{pmatrix} \Lambda_1^{(0)} & \Lambda_1^{(1)} & 0 & \cdots & 0 & \mathring{\Lambda}_1^{(2)} \\ \Lambda_2^{(0)} & 0 & \Lambda_2^{(1)} & \cdots & 0 & \mathring{\Lambda}_2^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \Lambda_N^{(0)} & 0 & 0 & \cdots & \Lambda_N^{(1)} & \mathring{\Lambda}_N^{(2)} \end{pmatrix}$$

Then, we write the full system as

$$Y = F\Lambda' + U$$

where

$$\underset{T\times NM}{Y}=\begin{bmatrix} y_1' \\ \vdots \\ y_T' \end{bmatrix},\ \underset{T\times r}{F}=\begin{bmatrix} F_1' \\ \vdots \\ F_T' \end{bmatrix},\ \underset{T\times NM}{U}=\begin{bmatrix} u_1' \\ \vdots \\ u_T' \end{bmatrix}$$

Special case: the country-industry hierarchical model of Choi et al. (2017) Consider the following multi-dimensional factor model:

$$y_{ijt}=\lambda_{ij}^{(0)\prime}f_t^{(0)}+\lambda_{ij}^{(1)\prime}f_{it}^{(1)}+u_{ijt},\ i=1,...,N,\ j=1,...,M_i,\ t=1,...,T \quad (379)$$

where $i=1,...,N$ indicates the country, $j=1,...,M$ denotes the sector and $t=1,...,T$ is the time period. $f_t^{(0)}=\left(f_{1t}^{(0)},...,f_{r^0t}^{(0)}\right)'$ comprises the $r^{(0)}\times 1$ global factors, and $f_{it}^{(1)}$ collects the $r_i^{(1)}\times 1$ vector of sectoral factors in country i. Stacking (379) for $j=1,...,M_i$, we obtain:

$$y_{i.t}=\Lambda_i^{(0)}f_t^{(0)}+\Lambda_i^{(1)}f_{it}^{(1)}+u_{i.t}=\left(\Lambda_i^{(0)},\Lambda_i^{(1)}\right)\begin{pmatrix} f_t^{(0)} \\ f_{it}^{(1)} \end{pmatrix}+u_{i.t} \quad (380)$$

where

$$\underset{M_i\times 1}{y_{i.t}}=\begin{bmatrix} y_{i1t} \\ y_{i2t} \\ \vdots \\ y_{iMt} \end{bmatrix},\ \underset{M_i\times 1}{u_{i.t}}=\begin{bmatrix} u_{i1t} \\ u_{i2t} \\ \vdots \\ u_{iMt} \end{bmatrix},\ \underset{M_i\times r^{(0)}}{\Lambda_i^{(0)}}=\begin{bmatrix} \lambda_{i1}^{(0)\prime} \\ \lambda_{i2}^{(0)\prime} \\ \vdots \\ \lambda_{iM}^{(0)\prime} \end{bmatrix},\ \underset{M_i\times r_i^{(1)}}{\Lambda_i^{(1)}}=\begin{bmatrix} \lambda_{i1}^{(1)\prime} \\ \lambda_{i2}^{(1)\prime} \\ \vdots \\ \lambda_{iM}^{(1)\prime} \end{bmatrix}$$

The entire system becomes:

$$\begin{pmatrix} y_{1.t} \\ y_{2.t} \\ \vdots \\ y_{N.t} \end{pmatrix}=\begin{pmatrix} \Lambda_1^{(0)} & \Lambda_1^{(1)} & 0 & \cdots & 0 \\ \Lambda_2^{(0)} & 0 & \Lambda_2^{(1)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_N^{(0)} & 0 & 0 & \cdots & \Lambda_N^{(1)} \end{pmatrix}\begin{pmatrix} f_t^{(0)} \\ f_{1t}^{(1)} \\ f_{2t}^{(1)} \\ \vdots \\ f_{Nt}^{(1)} \end{pmatrix}+\begin{pmatrix} u_{1.t} \\ u_{2.t} \\ \vdots \\ u_{N.t} \end{pmatrix}$$

which can be compactly written as

$$y_t=\Lambda F_t+u_t$$

where

$$\underset{N^*\times 1}{y_t}=\begin{bmatrix} y_{1.t} \\ y_{2.t} \\ \vdots \\ y_{N.t} \end{bmatrix},\ \underset{N^*\times 1}{u_t}=\begin{bmatrix} u_{1.t} \\ u_{2.t} \\ \vdots \\ u_{N.t} \end{bmatrix},\ \underset{r^*\times 1}{F_t}=\begin{pmatrix} f_t^{(0)} \\ f_{1t}^{(1)} \\ f_{2t}^{(1)} \\ \vdots \\ f_{Nt}^{(1)} \end{pmatrix}$$

$$\underset{N^* \times r^*}{\Lambda} = \begin{pmatrix} \Lambda_1^{(0)} & \Lambda_1^{(1)} & 0 & \cdots & 0 \\ \Lambda_2^{(0)} & 0 & \Lambda_2^{(1)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_N^{(0)} & 0 & 0 & \cdots & \Lambda_N^{(1)} \end{pmatrix}$$

and

$$N^* = \sum_{i=1}^{N} M_i, \; r^* = r^{(0)} + \sum_{i=1}^{N} r_i^{(1)}$$

Then, we write the full system as

$$Y = F\Lambda' + U$$

where

$$\underset{T \times N^*}{Y} = \begin{bmatrix} y_1' \\ \vdots \\ y_T' \end{bmatrix}, \; \underset{T \times r^*}{F} = \begin{bmatrix} F_1' \\ \vdots \\ F_T' \end{bmatrix}, \; \underset{T \times N^*}{U} = \begin{bmatrix} u_1' \\ \vdots \\ u_T' \end{bmatrix}$$

The extension to the three level or overlapping factor model (BE16)
Consider the following three-level factor model:

$$y_{ik,jt} = \lambda_{ik,j}^{(0)\prime} f_t^{(0)} + \lambda_{ik,j}^{(1)\prime} f_{it}^{(1)} + \lambda_{ik,j}^{(2)\prime} f_{kt}^{(2)} + u_{ik,jt}, \; i = 1, ..., N, k = 1, ..., K, \; j = 1, ..., M, \; t = 1, ..., T \tag{381}$$

where $i = 1, ..., N$ indicates the country, $j = 1, ..., M$ denotes the sector, $k = 1, ..., K$ is another classification (could be overlapping, e.g. style, value-growth or the market, nasdaq, dowjones). $f_t^{(0)} = \left(f_{1t}^{(0)}, ..., f_{r^0 t}^{(0)}\right)'$ comprises the $r^{(0)} \times 1$ global factors, $f_{it}^{(1)}$ collects the $r_i^{(1)} \times 1$ vector of industry factors, and $f_{kt}^{(2)}$ collects the $r_k^{(2)} \times 1$ vector of another classification factors. Stacking (377) for $j = 1, ..., M$, we obtain:

$$\begin{bmatrix} y_{ik,1t} \\ \vdots \\ y_{iK,Mt} \end{bmatrix} = \begin{bmatrix} \lambda_{ik,1}^{(0)\prime} \\ \vdots \\ \lambda_{iM}^{(0)\prime} \end{bmatrix} f_t^{(0)} + \begin{bmatrix} \lambda_{ik,1}^{(1)\prime} \\ \vdots \\ \lambda_{ik,M}^{(1)\prime} \end{bmatrix} f_{it}^{(1)} + \begin{bmatrix} \lambda_{ik,1}^{(2)\prime} \\ \vdots \\ \lambda_{ik,M}^{(2)\prime} \end{bmatrix} f_{kt}^{(2)} + \begin{bmatrix} u_{ik,t} \\ \vdots \\ u_{iMt} \end{bmatrix}$$

which can be compactly written as

$$y_{ik,.t} = \Lambda_{ik}^{(0)} f_t^{(0)} + \Lambda_{ik}^{(1)} f_{it}^{(1)} + \Lambda_{ik}^{(2)} f_{kt}^{(2)} + u_{ik,.t} = \left(\Lambda_{ik}^{(0)}, \Lambda_{ik}^{(1)}, \Lambda_{ik}^{(2)}\right) \begin{pmatrix} f_t^{(0)} \\ f_{it}^{(1)} \\ f_{kt}^{(2)} \end{pmatrix} + u_{ik,.t} \tag{382}$$

where

$$\underset{M \times 1}{y_{ik,.t}} = \begin{bmatrix} y_{i1t} \\ \vdots \\ y_{iMt} \end{bmatrix}, \; \underset{M \times 1}{u_{ik,.t}} = \begin{bmatrix} u_{i1t} \\ \vdots \\ u_{iMt} \end{bmatrix},$$

$$\underset{M\times r^{(0)}}{\Lambda_{ik}^{(0)}} = \begin{bmatrix} \lambda_{ik,1}^{(0)\prime} \\ \vdots \\ \lambda_{iM}^{(0)\prime} \end{bmatrix}, \quad \underset{M\times r_i^{(1)}}{\Lambda_{ik}^{(1)}} = \begin{bmatrix} \lambda_{ik,1}^{(1)\prime} \\ \vdots \\ \lambda_{ik,M}^{(1)\prime} \end{bmatrix}, \quad \underset{M\times r_k^{(2)}}{\Lambda_{ik}^{(2)}} = \begin{bmatrix} \lambda_{ik,1}^{(2)\prime} \\ \vdots \\ \lambda_{ik,M}^{(2)\prime} \end{bmatrix}$$

Next, the entire system for $i = 1, ..., N$ and $k = 1, ..., K$, becomes:

$$\begin{pmatrix} y_{11,.t} \\ \vdots \\ y_{N1,.t} \\ y_{12,.t} \\ \vdots \\ y_{N2,.t} \\ \vdots \\ y_{1K,.t} \\ \vdots \\ y_{NK,.t} \end{pmatrix} = \begin{pmatrix} \Lambda_{11}^{(0)} & \Lambda_{11}^{(1)} & 0 & \cdots & 0 & \Lambda_{11}^{(2)} & 0 & & 0 \\ \Lambda_{21}^{(0)} & 0 & \Lambda_{21}^{(1)} & \cdots & 0 & \Lambda_{21}^{(2)} & 0 & & 0 \\ \vdots & & & \ddots & & \vdots & \vdots & \ddots & \vdots \\ \Lambda_{N1}^{(0)} & 0 & 0 & \cdots & \Lambda_{N1}^{(1)} & \Lambda_{N1}^{(2)} & 0 & & 0 \\ \Lambda_{12}^{(0)} & \Lambda_{12}^{(1)} & 0 & \cdots & 0 & 0 & \Lambda_{12}^{(2)} & & 0 \\ \Lambda_{22}^{(0)} & 0 & \Lambda_{22}^{(1)} & \cdots & 0 & 0 & \Lambda_{22}^{(2)} & & 0 \\ \vdots & & & \ddots & & & \vdots & \ddots & \vdots \\ \Lambda_{N2}^{(0)} & 0 & 0 & \cdots & \Lambda_{N2}^{(1)} & 0 & \Lambda_{N2}^{(2)} & & 0 \\ \vdots & & & & & & \vdots & & \vdots \\ \Lambda_{1K}^{(0)} & \Lambda_{1K}^{(1)} & & \cdots & & 0 & 0 & & \Lambda_{1K}^{(2)} \\ \Lambda_{2K}^{(0)} & & \Lambda_{2K}^{(1)} & \cdots & & 0 & 0 & & \Lambda_{2K}^{(2)} \\ \vdots & & & \ddots & & \vdots & \vdots & \ddots & \vdots \\ \Lambda_{NK}^{(0)} & & & \cdots & \Lambda_{2K}^{(1)} & 0 & 0 & & \Lambda_{NK}^{(2)} \end{pmatrix} \begin{pmatrix} f_t^{(0)} \\ f_{1t}^{(1)} \\ \vdots \\ f_{Nt}^{(1)} \\ f_{1t}^{(2)} \\ \vdots \\ f_{Kt}^{(2)} \end{pmatrix} + \begin{pmatrix} u_{11,.t} \\ \vdots \\ u_{N1,.t} \\ \vdots \\ u_{1K,.t} \\ \vdots \\ u_{NK,.t} \end{pmatrix} \tag{383}$$

which can be compactly written as

$$y_t = \Lambda F_t + u_t$$

where

$$\underset{NKM\times 1}{y_t} = \begin{pmatrix} y_{11,.t} \\ \vdots \\ y_{N1,.t} \\ \vdots \\ y_{1K,.t} \\ \vdots \\ y_{NK,.t} \end{pmatrix}, \quad \underset{NKM\times 1}{u_t} = \begin{pmatrix} u_{11,.t} \\ \vdots \\ u_{N1,.t} \\ \vdots \\ u_{1K,.t} \\ \vdots \\ u_{NK,.t} \end{pmatrix}, \quad \underset{r\times 1}{F_t} = \begin{pmatrix} f_t^{(0)} \\ f_{1t}^{(1)} \\ \vdots \\ f_{Nt}^{(1)} \\ f_{1t}^{(2)} \\ \vdots \\ f_{Kt}^{(2)} \end{pmatrix}$$

$$\underset{NKM\times r}{\Lambda} = \begin{pmatrix}
\Lambda_{11}^{(0)} & \Lambda_{11}^{(1)} & 0 & \cdots & 0 & \Lambda_{11}^{(2)} & 0 & & 0 \\
\Lambda_{21}^{(0)} & 0 & \Lambda_{21}^{(1)} & \cdots & 0 & \Lambda_{21}^{(2)} & 0 & & 0 \\
\vdots & & & \ddots & & \vdots & \vdots & \ddots & \vdots \\
\Lambda_{N1}^{(0)} & 0 & 0 & \cdots & \Lambda_{N1}^{(1)} & \Lambda_{N1}^{(2)} & 0 & & 0 \\
\Lambda_{12}^{(0)} & \Lambda_{12}^{(1)} & 0 & \cdots & 0 & 0 & \Lambda_{12}^{(2)} & & 0 \\
\Lambda_{22}^{(0)} & 0 & \Lambda_{22}^{(1)} & \cdots & 0 & 0 & \Lambda_{22}^{(2)} & & 0 \\
\vdots & & & \ddots & & & \vdots & \ddots & \vdots \\
\Lambda_{N2}^{(0)} & 0 & 0 & \cdots & \Lambda_{N2}^{(1)} & 0 & \Lambda_{N2}^{(2)} & & 0 \\
\vdots & & & & & & \vdots & & \vdots \\
\Lambda_{1K}^{(0)} & \Lambda_{1K}^{(1)} & & \cdots & & 0 & 0 & & \Lambda_{1K}^{(2)} \\
\Lambda_{2K}^{(0)} & & \Lambda_{2K}^{(1)} & \cdots & & 0 & 0 & & \Lambda_{2K}^{(2)} \\
\vdots & & & \ddots & & \vdots & \vdots & \ddots & \vdots \\
\Lambda_{NK}^{(0)} & & & \cdots & \Lambda_{2K}^{(1)} & 0 & 0 & & \Lambda_{NK}^{(2)}
\end{pmatrix}$$

Then, we write the full system as

$$Y = F\Lambda' + U$$

where

$$\underset{T\times NKM}{Y} = \begin{bmatrix} y_1' \\ \vdots \\ y_T' \end{bmatrix}, \ \underset{T\times r}{F} = \begin{bmatrix} F_1' \\ \vdots \\ F_T' \end{bmatrix}, \ \underset{T\times NKM}{U} = \begin{bmatrix} u_1' \\ \vdots \\ u_T' \end{bmatrix}$$

Remark: See the estimation algorithm in BE16... But, this is not quite three-level model.We assume that $E\left(f_{kt}^{(2)} f_{kt}^{(2)\prime}\right) = I_{r_k^{(2)}}$ as well as $E\left(f_{kt}^{(2)} f_t^{(0)\prime}\right) = 0$ and $E\left(f_{kt}^{(2)} f_{it}^{(1)\prime}\right) = 0$. The least-squares can be applied to estimate the factors and factor loadings, where the iteration adopts a sequential estimation of the factors $f_t^{(0)}, f_{1t}^{(1)}, ..., f_{Nt}^{(1)}$ and $f_{1t}^{(2)}, ..., f_{Kt}^{(2)}$. In what follows we focus on the sequential LS procedure. Consistent starting values can be obtained from a CCA of the relevant subfactors (see below). Let $\hat{f}_t^{0(0)}, \hat{f}_{1t}^{0(1)}, ..., \hat{f}_{Nt}^{0(1)}$ and $\hat{f}_{1t}^{0(2)}, ..., \hat{f}_{Kt}^{0(2)}$ denote the initial estimators. The loading matrices can be estimated by running regressions of $y_{ik,jt}$ on the initial factor estimates $\hat{f}_t^{0(0)}, \hat{f}_{1t}^{0(1)}, ..., \hat{f}_{Nt}^{0(1)}$ and $\hat{f}_{1t}^{0(2)}, ..., \hat{f}_{Kt}^{0(2)}$. The resulting LS estimators of the loading coefficients are organised as in the matrix Λ, yielding the estimator $\hat{\Lambda}$. An update of the factor estimates is obtained by running a regression of y_t on $\hat{\Lambda}$ yielding the updated vector of factors, $\hat{f}_t^{1(0)}, \hat{f}_{1t}^{1(1)}, ..., \hat{f}_{Nt}^{1(1)}$ and $\hat{f}_{1t}^{1(2)}, ..., \hat{f}_{Kt}^{1(2)}$. With updated estimates of factors we obtain improved estimates of the loading coefficients by running again regressions of $y_{ik,jt}$ on the estimated factors. This sequential LS estimation procedure continues until convergence. The last step involves orthogonalising the local factors, $\left(\hat{f}_{1t}^{0(1)}, ..., \hat{f}_{Nt}^{0(1)}\right)$ and $\left(\hat{f}_{1t}^{0(2)}, ..., \hat{f}_{Kt}^{0(2)}\right)$. Although this orthogonalisation step is not necessary for identification of the factors, it

enables us to perform a variance decomposition of individual variables with respect to the factors. Orthogonalising the factors can be achieved by regressing $\left(\hat{f}_{1t}^{0(1)}, ..., \hat{f}_{Nt}^{0(1)}\right)$ on $\left(\hat{f}_{1t}^{0(2)}, ..., \hat{f}_{Kt}^{0(2)}\right)$ (or *vice versa*) and taking the residuals as new estimates of $\left(f_{1t}^{(1)}, ..., f_{Nt}^{(1)}\right)$ or $f_{1t}^{(2)}, ..., f_{Kt}^{(2)}$.

Remark: The initialization for the three-level factor model works as follows. We first estimate the global factor as the first $r^{(0)}$ PCs and the global factors are eliminated from the variables by running least-squares regressions of the variables on the estimated global factors.[25] In the next step the CCA is employed to extract the common component among the $r_i^{(1)} + r_k^{(2)}$ estimated factors from region i, group k and the estimated vectors from the same region i but different group k'. This common component is the estimated regional factor. Similarly, the estimated factor $f_{kt}^{(2)}$ obtained from a CCA of the factor of region i, group k and a different region i' but the same group. These initial estimates are used to start the sequential LS procedure. The overall estimation procedure outlined for the three-level factor model with an overlapping factor structure can be generalised to allow for further levels of factors (provided that the number of units in each group is sufficiently large). Furthermore, the levels may be specified as a hierarchical structure (e.g. Moench et al. (2013)), that is, the second level of factors (e.g. regions) is divided into a third level of factors (e.g. countries) such that each third level group is uniquely assigned to one second level group. For such hierarchical structures the CCA can be adapted to yield a consistent initial estimator for a sequential estimation procedure that switches between estimating the factors and (restricted) loadings.

More to come...

10.12 Alternative approaches based on the (local) spatial effects and global factors

To be completed:

- Indeed, this would make my ultimate goal. I have some preliminary notes and consider extending the QML-EM algorithms poposed by Bai and Li in a seuqnce of papers (all publised in top journals)**.**
- **I will provide more details soon.**

[25] Alternatively, a CCA between (i) the variables in region i and group k and (ii) the variables in group i' and k' with $i \neq i'$ and $k \neq k'$ may be employed to extract the common factors. In our experience the two-step top-down estimator used in our simulation performs similarly and has the advantage that the starting values are invariant with respect to a reorganization of the levels (that is interchanging regions and groups).

11 Spatial Weights Matrix and KMS[26]

The choice of appropriate spatial weights is a central component of spatial models as it assumes *a priori* a structure of spatial dependence, which may or may not correspond closely to reality. The spatial weights are interpreted as functions of relevant measures of geographic or economic distance (Anselin, 1988, 2002). The choice of weights is frequently arbitrary, there is substantial uncertainty regarding the choice, and empirical results vary considerably.

The spatial panel data models assume a time invariant spatial weights matrix. When the spatial weights matrix is constructed with economic/socioeconomic distances or demographic characteristics, it can be time varying (e.g. Case, Hines, and Rosen, 1993). Lee and Yu (2012b) investigate the QML estimation of SDPD models with time varying spatial weights matrices.Monte Carlo results show that, when spatial weights matrices are substantially varying over time, a model misspecification of a time invariant spatial weights matrix may cause substantial bias in estimation.

Define the $N \times N$ matrix of the spatial weights:[27]

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NN} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix} \text{ with } w_{ii} = 0 \tag{384}$$

The jth element of $\boldsymbol{w}_i$, w_{ij}, represents the link (or distance) between the neighbor j and the spatial unit i. It is a common practice to have $\boldsymbol{W}$ having a zero diagonal and being row-normalized such that the sum of elements in each row of $\boldsymbol{W}$ is unity. The ith row $\boldsymbol{w}_i$ may be constructed as $\boldsymbol{w}_i = (d_{i1}, d_{i2}, ..., d_{in}) / \sum_{j=1}^{n} d_{ij}$, where $d_{ij} \geq 0$ represents a function of the spatial distance between ith and jth units. The weighting operation may be interpreted as an average of neighboring values.

Shi (2016) defines the endogenous and time-varying spatial weights matrix satisfies the following assumption (see below for details):

Assumption 6. (1) The spatial weights satisfy $w_{ij,t} \geq 0$, $w_{ij,t} = 0$, and $w_{ij,t} = 0$ if $\rho_{ij,t} > \rho_c$, i.e., there exists a threshold $\rho_c > 1$ such that the weight is zero if the geographic distance exceeds ρ_c. For $i \neq j$,

$$w_{ij,t} = h_{ij}\left(z_{it}, z_{jt}\right) I\left(\rho_{ij,t} < \rho_c\right) \tag{385}$$

or the row normalized version that

$$w_{ij,t} = \frac{h_{ij}\left(z_{it}, z_{jt}\right) I\left(\rho_{ij,t} < \rho_c\right)}{\sum_{k=1}^{n} h_{ik}\left(z_{it}, z_{kt}\right) I\left(\rho_{ik,t} < \rho_c\right)}$$

where $h_{ij}()$'s are nonnegative, uniformly bounded functions. (2) The function $h_{ij}(.)$ satisfies the Lipschitz condition,

$$\left|h_{ij}\left(a_1, b_1\right) - h_{ij}\left(a_2, b_2\right)\right| \leq c_0\left(\left|a_1 - a_2\right| + \left|b_1 - b_2\right|\right)$$

[26] I will make the use of more clear and consitent notations after your feedbacks.

[27] We can easily construct the time-varying matrix, W_t.

for some finite constant c_0.

For convenience consider the simple SAR model:[28]

$$y_{it} = \lambda \sum_{j=1}^{n} w_{ij,t} y_{jt} + v_{it}. \tag{386}$$

We consider the different specifications for the spatial weights matrix as follows:

- Fixed weights based on the physical or economic distance. Assuming the threshold is known or trying a number of different thresholds. This exogenous assumption may hold when spatial weights are constructed using predetermined geographic distances.
- If "economic distance" such as the relative GDP or trade volume is used to construct the weight matrix, then it is likely that these elements are correlated with the final outcome.
- BHP: using the spatial correlation-based adjacency or spatial weights matrix, subject to sparsity issues, rendering the estimation less reliable., e.g. in the UK house price data by Beulah's thesis, the correlation-based weights matrix contains only 0.3% nonzero weights, rendering too many zero weights such that effective sample in each region is mostly 1 or 2.
- Under (385), Qu and Lee (2014) and Shi (2016) consider (see below for details):[29]

  $$h_{ij}\left(z_{it}, z_{jt}\right) = \frac{1}{|z_{it} - z_{jt}|} \tag{387}$$

  where $|z_{it} - z_{jt}|$ measures the economic distance, and $I\left(\rho_{ij,t} < \rho_c\right)$ is predetermined based on geographic distance, ρ_c such that $I\left(\rho_{ij,t} < \rho_c\right) = 1$ if the two locations are neighbors and otherwise 0.
- Using (385) and (387), the model (386) can be written as

  $$y_{it} = \lambda \sum_{j=1}^{n} \frac{1}{|z_{it} - z_{jt}|} I\left(\rho_{ij,t} < \rho_c\right) y_{jt} + v_{it}. \tag{388}$$

- Assuming that ρ_c is known and z_{it} is correlated with v_{it}, Shi extends Qu and Lee (2014) and develops the control function approach in panels:[30]

  $$y_{it} = \lambda \sum_{j=1}^{n} w_{ij,t} y_{jt} + \left(Z_{nt} - X_{2nt}\Gamma\right)_{it}' \delta + \xi_{it}. \tag{389}$$

[28] It is straightforward to develop the extended model with covariates and error components with unobserved fixed effects and factors or interactive effects.

[29] Qu and Lee (2014) also suggest the higher and nonlienar orders, $h_{ij}\left(z_{it}, z_{jt}\right) = \sum_{d=1}^{D} \frac{1}{|z_{it} - z_{jt}|^d}$

[30] It is unclear how to write the z specification for each spatial unit. See (406) and (407) below. The basic idea seems to get some information about the single covariate z_{it}, z_{jt} and their difference $z_{it} - z_{jt}$ from the large system equations.

where

$$w_{ij,t} = \frac{1}{|z_{it} - z_{jt}|} I\left(\rho_{ij,t} < \rho_c\right)$$

- Consider (406) below, and assume the single covariate in z:

$$z_{it} = x_{it}'\beta_z + \gamma_{iz}' f_{zt} + \epsilon_{it} \tag{390}$$

where x_{it} are $k_z \times 1$ regressors with coefficient vector β_z, and unobservables have two components, f_{zt} consisting of $R_z \times 1$ time factors with loading γ_{izl}' and ϵ_{itl} is idiosyncratic error. Then, (389) can be written as

$$y_{it} = \lambda \sum_{j=1}^{n} w_{ij,t} y_{jt} + \left(z_{it} - x_{it}'\beta_z - \gamma_{iz}' f_{zt}\right)' \delta + \xi_{it}. \tag{391}$$

- The structure of the model (388) is similar to KMS, which is given by

$$y_{it} = \lambda \frac{1}{m_{it}} \sum_{j=1}^{n} I\left(\rho_{ij,t} < \rho_c\right) y_{jt} + v_{it}. \tag{392}$$

where $m_{it} = \sum_{j=1}^{n} I\left(\rho_{ij,t} < \rho_c\right)$. (388) assumes the known threshold, ρ_c whilst (392) estimates λ and ρ_c jointly, then imposing the equal weight once y_{jt} is selected.

- More generally, we consider:

$$w_{ij,t} = \frac{1}{\rho_{ij,t}} I\left(\rho_{ij,t} < \rho_c\right) \ with\ \rho_{ij,t} = |z_{it} - z_{jt}| \tag{393}$$

Alternatively, we may consider the row-normalisation version as

$$w_{ij,t} = \frac{\frac{1}{\rho_{ij,t}} I\left(\rho_{ij,t} < \rho_c\right)}{\sum_{j=1}^{n} \frac{1}{\rho_{ij,t}} I\left(\rho_{ij,t} < \rho_c\right)} \ with\ \rho_{ij,t} = |z_{it} - z_{jt}| \tag{394}$$

Notice that Qu and Lee conjecture that **another issue that needs future research is to consider an endogenous spatial weight matrix purely constructed with economic distances**. This could be a technical challenging issue as the near-epoch assumption may not be met. {**WHY??**} Alternative large sample theorems may need to be developed.

- We can consider (i) the exogenous case, $E\left(z_{it}' v_{it}\right) = 0$ and (ii) the endogenous case, $E\left(z_{it}' v_{it}\right) \neq 0$. We may also consider the time-invariant case using

$$w_{ij} = \frac{1}{\rho_{ij}} I\left(\rho_{ij} < \rho_c\right) \ with\ \rho_{ij} = |z_i - z_j| \ \text{ or } \rho_{ij} = |\bar{z}_i - \bar{z}_j| \tag{395}$$

where we consider the time-invariant covariate, z_i or the time-average, $\bar{z}_i = T^{-1} \sum z_{it}$.

- I conjecture that the KMS algorithm to construct the endogenous spatial weights or selection matrix would be useful, makeing contribution to the literature. In particular, we consider VAR as the DGP for the $N \times 1$ vector, $\mathbf{z}_t = (z_{1t}, ..., z_{Nt})'$, say

$$\mathbf{z}_t = \sum_{j=1}^{p} \Phi_j \mathbf{z}_{t-j} + \boldsymbol{\epsilon}_t$$

and derive the CF:

$$\mathbf{v}_t = \left(\mathbf{z}_t - \sum_{j=1}^{p} \Phi_j \mathbf{z}_{t-j} \right)' \boldsymbol{\delta} + \boldsymbol{\xi}_t.$$

Then, the final model will become:

$$y_{it} = \lambda \sum_{j=1}^{n} w_{ij,t} y_{jt} + \boldsymbol{\epsilon}_{it}' \delta + \xi_{it}, \tag{396}$$

where $w_{ij,t}$ is defined in (393) or (394).

- Horrace et al. (2015) consider a firm with n workers. When the manager allocates workers to projects (peer groups) in each time period, $t = 1, ..., T$, she specifies an $n \times n$ adjacency matrix which determines the interrelatedness of the workers' productivity. Let the adjacency matrix be denoted by $A_t^o = [a_{ij,t}^o]$, where $a_{ij,t}^o = 1$ if workers i and j are assigned to the same project and $a_{ij,t}^o = 0$ otherwise. We set $a_{ii,t}^o = 0$. Let the row-normalized A_t^o be $A_t = [a_{ij,t}]$, where $a_{ij,t} = a_{ij,t}^o / \sum_{k=1}^{n} a_{ik,t}^o$. Then productivity of the worker i in period t is given by

$$y_{it} = \rho \sum_{j=1}^{n} a_{ij,t} y_{jt} + x_{it}\beta + u_{it}. \tag{397}$$

The dependent variable y_{it} is the productivity of worker i in period t. $\sum_{j=1}^{n} a_{ij,t} y_{jt}$ is the average productivity of worker i's co-workers assigned to the same project with ρ capturing the peer effect. x_{it} is a $1 \times k_x$ vector of exogenous variables. u_{it} is the disturbance. In this setting, the marginal product across workers in period t is $\rho a_{ij,t}$ when the workers are on the same project and 0 otherwise.
If A_t is exogenous so that $E(U_t | A_t, X_t) = 0$, then model (397) can be estimated using spatial panel data methods. However, it is reasonable to believe that the manager's **choices of how to allocate workers to projects may be correlated with** U_t. Then $E(U_t | A_t, X_t) \neq 0$ and A_t is endogenous.
Let d_{it} be an indicator variable such that $d_{it} = 1$ if worker is assigned to the project and $d_{it} = 0$ otherwise. Suppose m_t workers are allocated to

the project. Then, for worker i assigned to the project (i.e. $d_{it} = 1$), (397) can be written as

$$y_{it} = \rho \frac{1}{m_t} \sum_{j=1}^{n} d_{jt} y_{jt} + x_{it}\beta + E\left(u_{it}|D_t\right) + u_{it}^*. \tag{398}$$

where $D_t = (d_{1t}, ..., d_{nt})'$ and $u_{it}^* = u_{it} - E(u_{it}|D_t)$. By construction, $E\left(u_{it}^*|D_t\right) = 0$ and the weights d_{jt} in the peer effect regressor can be considered exogenous. We refer to $E\left(u_{it}|D_t\right)$ as the **selectivity bias**. As m_t is often predetermined (e.g., in sports games, the number of active players m_t is fixed), d_{it} is not independent across i. Hence, instead of modeling the probability that a certain worker is assigned to a project (i.e. $\Pr(d_{it} = 1)$), we consider the probability of a set of workers is assigned to a project.

- Notice that the structure of the model (398) is also similar to KMS, but it imposes that d_{jt}'s are known a priori. So this model is more restrictive than (388).
- **MORE discussions...**
- CCE approximation possible for spatial and factors?
- Heterogeneous extension of KMS?
- How to estimate the weights and thresholds together in KMS with and without endogeneity?

11.1 Qu and Lee (2014)

Consider the output equation of a cross-sectional SAR model:

$$Y_n = \lambda W_n Y_n + X_{1n}\beta + V_n, \tag{399}$$

where $Y_n = (y_{1,n}, ..., y_{n,n})'$ is an $n \times 1$ vector, X_{1n} is an $n \times k_1$ matrix with its elements $\{x_{1,in}; l(i) \in D_n, n \in N\}$ being bounded for all i and n, $V_n = (v_{1,n}, ..., v_{n,n})'$, λ is a scalar, and $\beta = \left(\beta_1, ..., \beta_{k_1}\right)'$ is a $k_1 \times 1$ vector of coefficients. $W_n = (w_{ij,n})$ is an $n \times n$ nonnegative weights matrix with zero diagonals and its elements constructed by Z_n:

$$w_{ij,n} = h_{ij}(Z_n, \rho_{ij}) \ for \ i, j = 1, ..., n; \ i \neq j, \tag{400}$$

where $h(\cdot)$ is a bounded function. Finally, we consider the DGP for Z_n:

$$\underset{n \times p_2}{Z_n} = \underset{n \times k_2}{X_{2n}} \underset{k_2 \times p_2}{\Gamma} + \varepsilon_n, \tag{401}$$

where X_{2n} is an $n \times k_2$ matrix with $\{x_{2,in}; l(i) \in D_n, n \in N\}$ bounded for all i and n, Γ is a $k_2 \times p_2$ matrix of coefficients, $\varepsilon_n = (\varepsilon_{1,n}, ..., \varepsilon_{n,n})'$ is an $n \times p_2$ matrix

of disturbances with $\varepsilon_{i,n} = (\varepsilon_{1,in}, ..., \varepsilon_{p_2,in})'$ being p_2 dimensional column vectors, and $Z_n = (z_{1,n}, ..., z_{n,n})'$ is an $n \times p_2$ matrix with $z_{i,n} = (z_{1,in}, ..., z_{p_2,in})'$.

$$\underset{p_2\times 1}{z_{i,n}} = \begin{bmatrix} z_{1,in} \\ \vdots \\ z_{p_2,in} \end{bmatrix}, \ \underset{n\times p_2}{Z_n} = \begin{bmatrix} z'_{1,n} \\ \vdots \\ z'_{n,n} \end{bmatrix} = \begin{bmatrix} z_{1,1n} & & z_{p_2,1n} \\ & \ddots & \\ z_{1,nn} & & z_{p_2,nn} \end{bmatrix},$$

$$\underset{p_2\times 1}{z_{i}} = \begin{bmatrix} z_{1i} \\ \vdots \\ z_{p_2 i} \end{bmatrix}, \ \underset{n\times p_2}{Z_n} = \begin{bmatrix} z'_{1} \\ \vdots \\ z'_{n} \end{bmatrix} = \begin{bmatrix} z_{11} & & z_{p_2 1} \\ & \ddots & \\ z_{1n} & & z_{p_2 n} \end{bmatrix},$$

$$\underset{k_2\times 1}{x_{2i}} = \begin{bmatrix} x_{2,1i} \\ \vdots \\ x_{2,k_2 i} \end{bmatrix}, \ \underset{n\times k_2}{X_{2n}} = \begin{bmatrix} x'_{21} \\ \vdots \\ x'_{2n} \end{bmatrix} = \begin{bmatrix} x_{2,11} & & x_{2,k_2 1} \\ & \ddots & \\ x_{2,1n} & & x_{2,k_2 n} \end{bmatrix},$$

$$\underset{k_2\times p_2}{\Gamma} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{k_2} \end{bmatrix} = \begin{bmatrix} \gamma_{11} & & \gamma_{1p_2} \\ & \ddots & \\ \gamma_{k_2 1} & & \gamma_{k_2 p_2} \end{bmatrix}$$

$$Z_n = X_{2n}\Gamma + \varepsilon_n$$

$$\begin{bmatrix} z_{1,1n} & & z_{p_2,1n} \\ & \ddots & \\ z_{1,nn} & & z_{p_2,nn} \end{bmatrix} = \begin{bmatrix} x_{2,11} & & x_{2,k_2 1} \\ & \ddots & \\ x_{2,1n} & & x_{2,k_2 n} \end{bmatrix} \begin{bmatrix} \gamma_{11} & & \gamma_{1p_2} \\ & \ddots & \\ \gamma_{k_2 1} & & \gamma_{k_2 p_2} \end{bmatrix}$$

$$= \begin{bmatrix} x_{2,11}\gamma_{11} + \cdots + x_{2,k_2 1}\gamma_{k_2 1} & & x_{2,11}\gamma_{1p_2} + \cdots + x_{2,k_2 1}\gamma_{k_2 p_2} \\ & \ddots & \\ x_{2,1n}\gamma_{11} + \cdots + x_{2,k_2 n}\gamma_{k_2 1} & & x_{2,1n}\gamma_{1p_2} + \cdots + x\gamma_{k_2 p_2} \end{bmatrix}$$

Let $\left\{ \left(\varepsilon_{l(i),n}, v_{l(i),n} \right); l(i) \in D_n, n \in N \right\}$ be a triangular double array of real random variables defined on a probability space $(\Omega; F; P)$, where the index set $D_n \subset D$ is a finite set.

Remark: We consider n agents in an area where each agent i is endowed with a predetermined location $l(i)$. Any two agents are separated away by a distance of at least 1. Due to some competition or spillover effects, each agent i has an outcome $y_{i,n}$ directly affected by its neighbors' outcomes $y_{j,n}$'s. The spatial weight $w_{ij,n}$ is a measure of the relative strength of linkage between agents i and j, and the spatial coefficient λ provides a multiplier for the spillover effects. However, **the spatial weight $w_{ij,n}$ is not predetermined but depends on some observable random variable Z_n.** We can think of $z_{i,n}$ as some economic variables at location $l(i)$ such as GDP, consumption and economic growth rate which influence strength of links across units. This specification has been used in the literature, and it may introduce endogeneity into the spatial weight

matrix. **Case et al. (1993)** consider the weights of the form (before row normalization):

$$w_{ij,n} = \frac{1}{|z_{i,n} - z_{j,n}|}$$

where $z_{i,n}$ and $z_{j,n}$ are observations on "meaningful" socioeconomic characteristics.

We have the following moment assumption.

Assumption 2. The error terms $v_{i,n}$ and $\varepsilon_{i,n}$, have a joint distribution:

$$\left(v_{i,n}, \varepsilon'_{i,n}\right)' \sim iid\left(0, \Sigma_{v\varepsilon}\right), \quad \Sigma_{v\varepsilon} = \begin{bmatrix} \sigma_v^2 & \sigma'_{v\varepsilon} \\ \sigma_{v\varepsilon} & \Sigma_\varepsilon \end{bmatrix}$$

where $\Sigma_{v\varepsilon}$ is positive definite,σ_v^2 is a scalar variance, covariance $\sigma_{v\varepsilon} = (\sigma_{v\varepsilon 1}, ..., \sigma_{v\varepsilon p_2})'$ is a p_2 dimensional vector, and Σ_ε is a $p_2 \times\ p_2$ matrix. The $\sup_{i,n} E|v_{i,n}|^{4+\delta_\varepsilon}$ and $\sup_{i,n} E\left\|\varepsilon_{i,n}\right\|^{4+\delta_\varepsilon}$ exist for some $\delta_\varepsilon > 0$. Furthermore, $E(v_{i,n}|\varepsilon_{i,n}) = \varepsilon'_{i,n}\delta$ and $Var(v_{i,n}|\varepsilon_{i,n}) = \sigma_\xi^2$.[31]

The endogeneity of W_n comes from the correlation between $v_{i,n}$ and $\varepsilon_{i,n}$. Using Assumption 2, we can construct[32]

$$\xi_n = V_n - \varepsilon_n \delta,$$

where

$$\delta = \Sigma_\varepsilon^{-1}\sigma_{v\varepsilon} \text{ and } \xi_n \sim \left(0, \sigma_\xi^2 I_n\right) \text{ with } \sigma_\xi^2 = \sigma_v^2 - \sigma'_{v\varepsilon}\Sigma_\varepsilon^{-1}\sigma_{v\varepsilon}.$$

In particular, ξ_n are uncorrelated with ε_n. The outcome equation (399) becomes:

$$Y_n = \lambda W_n Y_n + X_{1n}\beta + \left(Z_n - X_{2n}\Gamma\right)\delta + \xi_n, \tag{402}$$

with $E\left(\xi_{i,n}|\varepsilon_{i,n}\right) = 0$ and $E\left(\xi_{i,n}^2|\varepsilon_{i,n}\right) = \sigma_\xi^2$ and $\xi_{i,n}$'s are iid across i. Our subsequent asymptotic analysis will rely on (402), where $(Z_n - X_{2n}\Gamma)$ are control variables to control the endogeneity of W_n.

Remark: Qu and Lee (2014) propose 3 estimators. If error terms are jointly normally distributed, QMLE becomes MLE and achieves the asymptotic efficiency. The 2SIV estimation is based on linear moments only and therefore asymptotically less efficient. For the GMM estimator based on some proper linear and quadratic moment conditions, it can be asymptotically as efficient as the ML estimator under normality. In the absence of normality, QMLE is no longer asymptotic efficient. The optimum GMM estimator based on linear and quadratic moment conditions might be asymptotically more efficient than QMLE, because it adopts the best weighting matrix for those moment conditions, while QMLE

[31] This conditional homoskedasticity condition is required for the QMLE theory. For the IV or GMM estimations, we can relax this assumption.

[32] In the special case that $(v_{i,n}, \varepsilon'_{i,n})'$ has a jointly normal distribution, then

$$v_{i,n}|\varepsilon_{i,n} \sim N(\sigma'_{v\varepsilon}\Sigma_\varepsilon^{-1}\varepsilon_{i,n}, \sigma_v^2 - \sigma'_{v\varepsilon}\Sigma_\varepsilon^{-1}\sigma_{v\varepsilon})$$

and ξ_n is independent of ε_n.

gives each moment an equal weight. However, the asymptotic efficiency of the 2SIV and QMLE cannot be directly compared. **Therefore, the 2SIV and GMM methods have the merit of computational simplicity and robustness.**

Assumption 4. We consider two cases of W_n:

(4.1) Case 1: The spatial weight $w_{ij,n} = h_{ij}(z_{i,n}, z_{j,n}, \rho_{ij})$ for $i \neq j$, where $h_{ij}(\cdot)$'s are non-negative, uniformly bounded functions of some observable variable Z_n. $0 \leq w_{ij,n} \leq c_1 \rho_{ij}^{-c_3 d_0}$ for some $0 \leq c_1$ and $c_3 > 3$.[33] Furthermore, there exist at most K ($K \geq 1$) columns of W_n that the column sum exceeds c_w, where K is a fixed number that does not depend on n.[34]

(4.2) Case 2: The spatial weight $w_{ij,n} = 0$ if $\rho_{ij} > \rho_c$, i.e., there exists a threshold $\rho_c > 1$ and if the geographic distance exceeds ρ_c , then the weight is zero. For $i \neq j$,

$$w_{ij,n} = h_{ij}\left(z_{i,n}, z_{j,n}\right) I\left(\rho_{ij} \leq \rho_c\right) \text{ or } w_{ij,n} = \frac{h_{ij}(z_{i,n}, z_{j,n}) I(\rho_{ij} \leq \rho_c)}{\sum_{\rho_{ij} \leq \rho_c} h_{ik}(z_{i,n}, z_{k,n})},$$

where $h_{ij}(.)$'s are non-negative, uniformly bounded functions.

Remark: Assumption 4 provides the essential features of the weights matrix. The geographic distance plays an important role in constraining magnitudes of spatial weights. The spatial weight of two locations would be larger if they were closer to each other **or when their economic indices were more similar**, but their weights would become smaller when two units are further apart. Assumption (4.1) allows the situation that all agents are spatially correlated but the spatial weight decreases sufficiently fast at a certain rate as physical distances increase. Symmetry is not imposed on the spatial weight matrix. If W_n is symmetric, the second part on the column sum norm condition in (4.1) will not be needed. For an asymmetric W_n, the second part of (4.1) limits the number of columns which have large magnitudes relative to the row sum norm. For example, big countries may have great impact on small countries, but those small countries may have little or zero influence on big countries. **In this example, we have some "stars" whose row sums are bounded by c_w, while their column sums can be much larger.** Assumption (4.1) assumes that the number of such stars can only be finite and bounded. Assumption (4.2) allows for a row-normalized spatial weight matrix. In this case, $w_{ij,n}$ might have agents linked in an area, but once the geographic distance between two agents exceeds a threshold, the two units are not spatially interacted.

In the MC they construct the endogenous, row-normalized $W_n = (w_{ij,n})$ as follows:

1. Generate bivariate normal random variables $(v_{i,n}, \varepsilon_{i,n})$ from $iidN\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ as disturbances in the outcome equation and the spatial weights equation.

[33] For example, $w_{ij,n} = \min\left(\frac{1}{\|z_{i,n} - z_{j,n}\|_p}, c_1 \rho_{ij}^{-c_3 d_0}\right)$.

[34] As $c_0^{-\rho_{ij}}$ decreases faster than $\rho_{ij}^{-c_3 d_0}$, all the results hold for the case of $0 \leq w_{ij,n}^d \leq c_1 c_0^{-\rho_{ij}}$ with some $c_1 \geq 0$ and $c_0 > 1$.

2. Construct the spatial weight matrix as the Hadamard product $W_n = W_n^d \circ W_n^e$, i.e., $w_{ij,n} = w_{ij,n}^d w_{ij,n}^e$, where W_n^d is a predetermined matrix based on geographic distance: $w_{ij,n}^d = 1$ if the two locations are neighbors and otherwise 0; W_n^e is a matrix based on economic similarity: $w_{ij,n}^e = 1/|z_{i,n} - z_{j,n}|$ if $i \neq j$ and $w_{ii,n}^e = 0$, where elements of Z_n are generated by $z_{i,n} = 1 + 0.8x_{i2,n} + \varepsilon_{i,n}$.

3. Row-normalize W_n.

Remark: See also Assumption 6 in Shi (2016) that considers only Assumption (4.2). Notice that $h_{ij}(z_{i,n}, z_{j,n})$ measures economic similarity while $I(\rho_{ij} \leq \rho_c)$ depends on the physical distance**. In our case we consider** $h_{ij}(z_{i,n}, z_{j,n})$ measures economic similarity and $I(\rho_{ij} \leq \rho_c)$ also depends on the economic distance.

11.2 Shi (2106)

A data set contains n individuals indexed by i for T time periods indexed by t. Individuals are located on a unevenly spaced lattice $D \subset R^{d_0}$ with $d_0 \geq 1$. The location $l : \{1, ..., n\} \to D_n \subset D$ is a mapping of individual i to its location $l(i) \in D \subset R^{d_0}$. Let $D_T \subset Z$ denote the set of time indexes. This is an adaptation of the topological structure of spatial processes in Jenish and Prucha (2009, 2012) and Qu and Lee (2015) to the panel data setting. Define the metric

$$\rho_{it,js} = \rho(it, js) = \max\left\{\max_{1\leq k\leq d_0}\left(|l(i)_k - l(j)_k|\right), |t-s|\right\}$$

where $l(i)_k$ is the kth component of the $d_0 \times 1$ vector $l(i)$.

Assumption 1. The lattice $D \subset R^{d_0}$ with $d_0 \geq 1$, is infinitely countable. All elements in D are located at distances of at least $\rho_0 > 0$ from each other. We assume that $\rho_0 = 1$.

The equation of interest is (**bad notations)**

$$y_{nt} = \lambda W_{nt} y_{nt} + X_{nty}\beta_y + \Gamma_{ny} f_{yt} + v_{nt}, \tag{403}$$

where y_{nt} is $n \times 1$, X_{nty} is $n \times k_y$ matrix of exogenous regressors. W_{nt} is an $n \times n$ spatial weights matrix with elements, $w_{ij,t}$.

$$\underset{n\times 1}{y_{nt}} = \begin{bmatrix} y_{1t} \\ \vdots \\ y_{nt} \end{bmatrix}, \underset{n\times k_y}{X_{nty}} = \begin{bmatrix} X_{1t,1} & & X_{1t,k_y} \\ \vdots & \ddots & \vdots \\ X_{nt,1} & & X_{nt,k_y} \end{bmatrix}, \underset{k_y\times 1}{\beta_y} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{k_y} \end{bmatrix}$$

$$\underset{n\times R_y}{\Gamma_{ny}} = \begin{bmatrix} \Gamma_{11} & & \Gamma_{1R_y} \\ \vdots & \ddots & \vdots \\ \Gamma_{n1} & & \Gamma_{nR_y} \end{bmatrix} = \begin{bmatrix} \Gamma_1 \\ \vdots \\ \Gamma_n \end{bmatrix}, \underset{R_y\times 1}{f_{yt}} = \begin{bmatrix} f_{1t} \\ \vdots \\ f_{R_y t} \end{bmatrix}, \underset{n\times 1}{v_{nt}} = \begin{bmatrix} v_{1t} \\ \vdots \\ v_{nt} \end{bmatrix},$$

$$\underset{n\times n}{W_{nt}} = \begin{bmatrix} w_{11,t} & \cdots & w_{1n,t} \\ \vdots & \ddots & \vdots \\ w_{n1,t} & \cdots & w_{nn,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{1,t} \\ \vdots \\ \boldsymbol{w}_{n,t} \end{bmatrix} \text{ with } w_{ii,t} = 0 \tag{404}$$

The **spatial weights matrix measures the degree of connections between spatial units** with zero diagonals. Let y_{it} denote the i-th element of y_{nt}. $\lambda w_{ij,t}$ measures the impact of y_{jt} on y_{it} and λ is the spatial interactions coefficient. Economic theory may suggest how the spatial weights matrix is constructed.

Example: Kelejian and Piras (2014): Weights are based on relative prices of cigarettes between two neighboring states. We assume the following model:

$$\ln C_{it} = \beta_1 \ln C_{i,t-1} + \beta_2 \ln p_{it} + \beta_3 \ln I_{it} + \lambda \left[\sum_{j=1}^{46} \frac{p_{jt}}{p_{it}} d_{ijt} \ln C_{jt} \right] + \mu_i + \delta_t + u_{it} \tag{405}$$

where $i = 1, \ldots, 46$ denotes states, $t = 1, \ldots, 29$ denotes time periods, and the disturbance term u_{it} has the non-parametric specification.[35] C_{it} is cigarette sales to persons of smoking age in packs per capita in state i at time t. p_{it} is the average retail price per pack of cigarettes. I_{it} is per capita disposable income. All values are measured in real terms. The spatial lag accounts for cross border cigarette shopping, or bootlegging. d_{ijt} is a dummy variable which indicates the desirability of cross border shopping. In particular, $d_{ijt} = 1$ if i and j are border states and $p_{jt} < p_{it}$; $d_{ijt} = 0$ if i and j are not border states, or if $p_{jt} > p_{it}$. The multiplication of d_{ijt} by the price ratio indicates that the price ratio p_{jt}/p_{it} will only be considered by cigarette consumers in state i if $p_{jt} < p_{it}$. Since the price of cigarettes is endogenous in a demand model for cigarettes, our weighting matrix is endogenous. We expect λ to be positive because the higher is the price ratio, the less attractive is cross state shopping. See also an example (subsection 3.2.3 in Shi) about the average effect of a treatment on the treated (ATT).

Sources of endogeneity The spatial weights matrix may be constructed from covariate that may correlate with v_{nt} together with common factors. Suppose that there are p variables, $z_{it1}, ..., z_{itp}$ that are used to construct W_{nt}. Consider the regression equation for z_{itl}:

$$z_{itl} = x'_{itzl}\beta_{zl} + \gamma'_{izl} f_{zt} + \epsilon_{itl},\ l = 1, ..., p \tag{406}$$

where x_{itzl} are $k_{zl} \times 1$ regressors with coefficient vector β_{zl}, and unobservables have two components, f_{zt} consisting of $R_z \times 1$ time factors with loading γ'_{izl} and ϵ_{itl} is idiosyncratic error. Stacking it across i and then over l for time period

[35] We assume the following non-parametric specification:

$$u_N = R_N \varepsilon_N$$

where R_N is an unknown $NT \times NT$ non-stochastic matrix, and ε_N is an $NT \times 1$ random vector whose mean is zero and VC is I_{NT} with $E(u_N) = 0$ and $(u_N u'_N) = R_N R'_N$.

t, we have $z_{nt} = (z_{1t1}, ..., z_{nt1}, z_{1t2}, ..., z_{ntp})'$ an $np \times 1$ vector which has the following structure:

$$z_{nt} = X_{ntz}\beta_z + \Gamma_{nz} f_{zt} + \epsilon_{nt} \tag{407}$$

where X_{ntz} is $np \times k_z$ with $k_z = k_{z1} + ... + k_{zp}$, $\Gamma_{nz} = (\gamma_{1z1}, ..., \gamma_{nz1}, \gamma_{1z2}, ..., \gamma_{nzp})'$ is $np \times R$ and ϵ_{nt} is defined similarly. Note that X_{nty} and X_{ntz} may have some common regressors.

Remark: Notations in Shi are generally unclear and need to be improved. In constructing z_{nt}, he allows the dimension of X regressors is different for $l = 1, ..., p$. Hence, the definition of the $np \times k_z$ matrix, X_{ntz} is quite unclear. In MC and empirical applications he used the univariate construction of z_i variable. Suppose that we use the difference between univariate variable, say $\frac{1}{|z_{it} - z_{jt}|}$ to construct the spatial weights. In such case the use of large dimensional specification for z_{nt} in (407) seems to be redundant (also computationally infeasible). If so, I conjecture that the use of **VAR or SPVAR seems to be more sensible, say**

$$z_{nt} = \sum_{j=1}^{p} \Phi_j z_{n,t-j} + \Gamma_{nz} f_{zt} + \epsilon_{nt} \tag{408}$$

Still an important issue of how to construct the spatial weights from (408), which I will discuss more later. Remind that the VAR cannot be employed in the cross-section studies by Qu and Lee (2015).

The following conditional moment assumption specifies that the disturbance terms in z_{nt} may correlate with v_{it} in the y_{nt} equation.

Assumption 2 (Qu and Lee (2015)). The error terms v_{it} and ϵ_{it} are independently distributed over i and t, and have a joint distribution

$$(v_{it}, \epsilon_{it}')' \sim (0, \Sigma_{v\varepsilon}), \quad \Sigma_{v\varepsilon} = \begin{bmatrix} \sigma_v^2 & \sigma_{v\varepsilon}' \\ \sigma_{v\varepsilon} & \Sigma_\varepsilon \end{bmatrix}$$

where $\Sigma_{v\varepsilon}$is positive definite, σ_v^2 is a scalar variance, the covariance $\sigma_{v\varepsilon} = (\sigma_{v\varepsilon 1}, ..., \sigma_{v\varepsilon p})$ is a $p \times 1$ vector, and Σ_ε is a $p \times p$ matrix. Furthermore, ϵ_{it} is generated as $\epsilon_{it} = \Sigma_\epsilon^{1/2} e_{it}$ where e_{itp} is independently distributed across i, t and p with $E(e_{itp}) = 0$, $E(e_{itp}^2) = 1$. Furthermore, $\sup_{n,T} \sup_{i,t} E|v_{it}|^{4+\delta_\epsilon}$ and $\sup_{n,T} \sup_{i,t} E\|v_{it}\|^{4+\delta_\epsilon}$ exist for some $\delta_\epsilon > 0$. Denote

$$E(v_{it}|\epsilon_{it}) = \epsilon_{it}'\delta$$

and define

$$\xi_{it} = v_{it} - \epsilon_{it}'\delta.$$

Assuming that $E(\xi_{it}^2 |\epsilon_{it}|) = E(\xi_{it}^2) = \sigma_\xi^2$, $E(\xi_{it}^3 |\epsilon_{it}|) = E(\xi_{it}^3)$ and $E(\xi_{it}^4 |\epsilon_{it}|) = E(\xi_{it}^4)$.

If $\sigma_{v\epsilon} \neq 0$, the spatial weights matrix correlates with disturbances in the outcome equation of the same period. Hence, using the assumption 2, the outcome equation (403) becomes:

$$y_{nt} = \lambda W_{nt} y_{nt} + X_{nty}\beta_y + \Gamma_{ny} f_{yt} + (z_{nt} - X_{ntz}\beta_z - \Gamma_{nz} f_{zt})'\delta + \xi_{nt}, \tag{409}$$

Shi develops the QML estimator of (409), though (409) is not clearly defined.

Remark: The use of control function approach is found to be robust to nonlinear model, e.g. Wooldridge or Blundell?? Suppose in the linear model that $W_{nt}y_{nt}$ and/or X_{nty} are endogenous, then we use as CF

$$\left(W_{nt}y_{nt} - z_{nt}\gamma_y\right)'\delta_y \text{ or } \left(X_{nty} - z_{nt}\gamma_x\right)'\delta_x$$

Similarly, when treating $w_{ij,t}$ as the variable, then we may consider the following CF:

$$\left(W_{nt} - z_{nt}\gamma_W\right)'\delta_W$$

Here, Qu and Lee and Shi assume that spatial weights follow the smooth function of z_{nt}, and employ the CF in terms of $\left(z_{nt} - X_{ntz}\beta_z - \Gamma_{nz}f_{zt}\right)'\delta$ in (409). So we should think about this issue more carefully.

MC Consider the main outcome equation:

$$y_{it} = \lambda\sum_{j=1}^{n} w_{ij,t}y_{jt} + x_{it}\beta_y + \gamma_{yi}'f_{yt} + v_{it}.$$

Two unobserved factors (f_{yt}) and factor loadings (γ_{yi}') are generated independently from $U[-2,2]$. The observed regressor (x_{it}) is a scalar and correlates with factors through

$$x_{it} = \frac{1}{3}\left\{\gamma_{yi}'f_{yt} + \gamma_{yi}'\ell_2 + f_{yt}'\ell_2\right\} + \eta_{it},$$

where $\eta_{it} \sim U[-2,2]$ and ℓ_2 is a 2×1 vector of 1's.

The spatial weights are correlated with v_{it} according to the following process.

1. Individuals indexed by 1 to n are located successively on a chessboard of dimension $\sqrt{n}\times\sqrt{n}$. The neighborhood structure follows the rook pattern. Individuals in the interior of the chessboard have 4 neighbors, and those on the border and the corner have 3 and 2 neighbors, respectively. Let $w_{ij}^d = 1$ if i and j are neighbors and $w_{ij}^d = 0$ otherwise, and denote the $n\times n$ matrix $W_d = \left[w_{ij}^d\right]$.

2. Let
$$z_{it} = x_{it}\beta_z + \gamma_{zi}f_{zt} + \epsilon_{it}.$$
with one unobserved factor f_{zt} being the first element of f_{yt}. The scalar γ_{zi} is generated from $U[-2,2]$, and . Let
$$w_{ij,t}^e = w_{ij}^d \times \min\left(\frac{1}{|z_{it} - z_{jt}|}, 2\right)$$
The spatial dependence is stronger for neighbors with similar z's. $w_{it,jt}^e$ is capped at 2 such that individuals whose z's are very similar ($|z_{it} - z_{jt}| <$

0.5) have the same degree of spatial effect. The spatial weights in the outcome equation are row-normalized,

$$w_{it,jt} = \frac{w^e_{it,jt}}{\sum_{j=1}^n w^e_{it,jt}}.$$

3. The idiosyncratic errors v_{it} and ϵ_{it} are generated from i.i.d. bivariate normal random variables,

$$N\left(0, \frac{4}{3}\vartheta \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

The inverse of ϑ is a measure of the signal to noise ratio.

Application Let y_{it} denote the HECM origination rate, defined as the number of newly originated HECM loans in state i and quarter t as a percentage of the senior population (age 65 plus) from the 2010 census. There are two $n \times n$ spatial weights matrices, $W_{1,n} = (w_{1,ij})$ and $W_{2,nt} = (w_{2,it,jt})$. $w_{1,ij} = 1$ if states i and j share the same border and $w_{1,ij} = 0$ otherwise. $W_{2,nt}$ captures the different spillover effect from large lenders. $w_{2,it,jt} = w_{1,ij} z_{it} z_{jt}$, where z_{it} is the share of the HECM loans originated by the large lenders (defined as being one of the top 10 largest lenders. A larger weight is given to a state with more dominant large lenders. House price dynamic variables are constructed using the Federal Housing Finance Agency's quarterly all-transactions house price indexes (HPI) deflated by the CPI, and include deviations from the previous 9 year averages (hpi_dev), standard deviations of house price changes in the previous 9 years (hpi_v) and the interaction between the two.

Our empirical application considers multiple spatial weights matrices such that

$$z_{it} = hpi\_dev_{it}\beta_{z1} + hpi\_v_{it}\beta_{z2} + (hpi\_dev_{it} \times hpi\_v_{it})\beta_{z3} + \gamma'_{zi} f_{zt} + \epsilon_{it}; \quad (410)$$

$$y_{it} = \lambda_1 \sum_{j=1}^n w_{1,ij} y_{jt} + \lambda_2 \sum_{j=1}^n w_{it,jt} y_{jt} + hpi\_dev_{it}\beta_{y1} + hpi\_v_{it}\beta_{y2} + (hpi\_dev_{it} \times hpi\_v_{it})\beta_{y3} + \gamma'_{yi} f_{yt} + v_{it}; \quad (411)$$

where ϵ_{it} and v_{it} have variances σ^2_ϵ and σ^2_v with correlation ρ. According to the eigenvalue ratio criterion, z_{it} has one unobserved factor and so is y_{it}, and the growth ratio criterion gives the same result. Table 3.7 reports the estimation results.

Remark: As discussed above, Shi used the relatively simple specification. There is only single z variable, and x regressors are common in (410) and (411). There is no details of whether there are common factors between f_{zt} and f_{yt}. Also unclear of whether one unobserved factor estimated for z_{it} and y_{it} is the same or different. In general, unsure about the computational details... Further, an intuition about $w_{2,it,jt} = w_{1,ij} z_{it} z_{jt}$ is unclear too, though it is designed to capture spillovers due to large lenders. Overall, this line of research will be more important and they provide the first research step heading into such directions but still there are many ambiguous issues in the current study.

11.3 Horrace et al. (2016)

Suppose there are q_t possible lineups denoted by L_s for $s = 1, ..., q_t$. The manager allocates lineup L_s to the project if and only if

$$d^*_{st} > \max_{r \neq s} d^*_{rt},$$

where

$$d^*_{st} = \pi_{st} + \xi_{st}, \quad s = 1, ..., q_t$$

where π_{st} is the deterministic component of d^*_{st} and ξ_{st} is a random innovation with zero mean and unit variance. Let d_{st} be a dummy such that $d_{st} = 1$ if the lineup L_s is chosen in period t and $d_{st} = 0$ otherwise. Then, $d_{st} = 1$ if and only if

$$\epsilon_{st} < 0 \text{ with } \epsilon_{st} = \max_{r \neq s} d^*_{rt} - d^*_{st}.$$

The productivity of L_s is given by the following model:

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta + U_{st}, \ s = 1, ..., q_t, t = 1, ..., T. \tag{412}$$

$Y_{st} = [y_{it}]_{i \in L_s}$ is an $m_t \times 1$ vector of the dependent variable of the workers in L_s. W_t is a constant weighting matrix given by (complete network)

$$\begin{aligned} W_t &= \frac{1}{m_t - 1}\left(1_{m_t}1'_{m_t} - I_{m_t}\right) \\ &= \frac{1}{m_t - 1}\left(\begin{bmatrix} 1 & & 1 \\ & \ddots & \\ 1 & & 1 \end{bmatrix} - \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}\right) = \frac{1}{m_t - 1}\begin{bmatrix} 0 & & 1 \\ & \ddots & \\ 1 & & 0 \end{bmatrix}. \end{aligned}$$

$$W_t Y_{st} = \frac{1}{m_t - 1}\begin{bmatrix} 0 & & 1 \\ & \ddots & \\ 1 & & 0 \end{bmatrix}\begin{bmatrix} y_{(1)t} \\ \vdots \\ y_{(m_t)t} \end{bmatrix} = \begin{bmatrix} \frac{1}{m_t - 1}\sum_{j=2}^{m_t} y_{(j)t} \\ \vdots \\ \frac{1}{m_t - 1}\sum_{j=1}^{m_t - 1} y_{(j)t} \end{bmatrix}$$

$W_t Y_{st}$ measures the average productivity of a worker's co-workers in lineup L_s, with its coefficient ρ capturing the peer effect. $X_{st} = [x_{it}]_{i \in L_s}$ is an $m_t \times k_x$ matrix of k_x exogenous variables of the workers in L_s. U_{st} is an $m_t \times 1$ vector of disturbances such that $U_{st} \sim iid(0, \Sigma)$. We allow for possible correlation between U_{st} and $\xi_t = \left(\xi_{1t}, ..., \xi_{q_t t}\right)$ such that

$$E(U_{st}|d_{st} = 1, \pi_t) = \lambda_s(\pi_t)\, 1_{m_t} \text{ with } \pi_t = (\pi_{1t}, ..., \pi_{q_t t}) \tag{413}$$

Then, the model (412) can be written as[36]

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta + \lambda_s(\pi_t)1_{m_t} + U^*_{st}, \tag{414}$$

where $U^*_{st} = U_{st} - \lambda_s(\pi_t)1_{m_t}$. We consider three different approaches for estimation of (414).

[36] The selectivity bias $\lambda_s(\pi_t)$ introduces a group correlated effect (Manski, 1993). Semiparametric estimation of ρ and β along with the unknown $\lambda_s(.)$ would face the "the curse of dimensionality."

The parametric selection correction approach Let $F_{st}(.|\pi_t)$ denote the conditional distribution function of $\epsilon_{st} = \max_{r \neq s} d^*_{rt} - d^*_{st}$. Let $\Phi(.)$ and $\phi(.)$ denote the standard normal distribution and density. Lee (1983) suggests using the transformation:

$$J_{st}(.) \equiv \Phi^{-1}(F_{st}(.|\pi_t))$$

to reduce the dimensionality of the selectivity bias. In this case the selectivity bias is given by

$$E(U_{st}|d_{st} = 1, \pi_t) = E[U_{st}|J_{st}(\epsilon_{st}) < J_{st}(0), \pi_t] = E[U_{st}|J_{st}(\epsilon_{st}) < J_{st}(0)],$$

where Lee (1983) implicitly assumes that the joint distribution of U_{st} and $J_{st}(\epsilon_{st})$ does not depend on π_t. Further, we make the following assumption that U_{st} and $J_{st}(\epsilon_{st})$ are i.i.d. with a joint normal distribution given by

$$\begin{bmatrix} U_{st} \\ J_{st}(\epsilon_{st}) \end{bmatrix} \sim N\left[0, \begin{pmatrix} \Sigma & \sigma_{12}1_{m_t} \\ \sigma_{12}1'_{m_t} & 1 \end{pmatrix}\right]$$

The selectivity bias is then given by

$$E(U_{st}|d_{st} = 1, \pi_t) = -\sigma_{12}\frac{\phi(J_{st}(0))}{F_{st}(0|\pi_t)}1_{m_t}. \tag{415}$$

Thus, from (413) and (415), we have:

$$\lambda_s(\pi_t) = -\sigma_{12}\frac{\phi\left(\Phi^{-1}(P_{st})\right)}{P_{st}}, \tag{416}$$

where we use

$$J_{st}(0) = \Phi^{-1}(F_{st}(0|\pi_t)) \text{ and } P_{st} = F_{st}(0|\pi_t).$$

Substitution of (416) into (414) gives

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta - \sigma_{12}\frac{\phi\left(\Phi^{-1}(P_{st})\right)}{P_{st}}1_{m_t} + U^*_{st}. \tag{417}$$

For the network model, Lee's approach can be implemented as follows.

- **Step 1**: Let $\pi_{st} = z_{st}\gamma$, where z_{st} is a $1 \times k_z$ vector of exogenous variables. Then, γ can be estimated by maximizing the likelihood function:

  $$\ln L = \sum_{t=1}^{T}\sum_{s=1}^{q_t} d_{st} \ln P_{st}.$$

  It proves convenient to assume that ξ_{st} is independently and identically Gumbel distributed so that

  $$P_{st} = \exp(z_{st}\gamma)/\sum_{r=1}^{q_t}\exp(z_{rt}\gamma)$$

  Then, γ can be estimated by a conditional logit estimator $\hat{\gamma}$ (McFadden, 1974).

- **Step 2**: With the predicted probabilities

$$\hat{P}_{st} = \exp\left(z_{st}\hat{\gamma}\right) / \sum_{r=1}^{q_t} \exp\left(z_{st}\hat{\gamma}\right),$$

we consider the feasible counterpart of (14)

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta - \sigma_{12}\frac{\phi\left(\Phi^{-1}(\hat{P}_{st})\right)}{\hat{P}_{st}} 1_{m_t} + U_{st}^{**}$$

and estimate $(\rho, \beta', \sigma_{12})'$ by 2SLS estimator with linearly independent columns in $W_t X_{st}$ as instruments for $W_t Y_{st}$.

The semi-parametric selection correction approach Following Dahl (2002), we impose the following assumption to reduce the dimensionality of the selectivity bias:

$$\lambda_s\left(\pi_t\right) = \mu(P_{st}).$$

Thus, (414) becomes:

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta + \mu(P_{st}) 1_{m_t} + U_{st}^{*}.$$

The semi-parametric selection correction approach can be implemented in a similar two-step procedure.

- **Step 1**: We obtain the predicted probabilities $\hat{P}_{st}$ from, say, a conditional logit regression.
- **Step 2**: We replace $\mu(P_{st})$ by its (feasible) series approximation

$$\sum_{k=1}^{K} \kappa_k b_k\left(\hat{P}_{st}\right),$$

where $b_k(.)$ are the basis functions, and estimate $(\rho, \beta')'$ together with κ_k by the 2SLS estimator with linearly independent columns in $W_t X_{st}$ as instruments for $W_t Y_{st}$.

The fixed-effect approach The selectivity bias $\lambda_s\left(\pi_t\right)$ can be considered as a time-varying lineup-specific fixed effect. To avoid estimating the unknown function $\lambda_s\left(.\right)$, we can apply a within transformation to eliminate this term. Suppose $X_{st} = [X_{1,st}, 1_{m_t} x_{2,st}]$, where $X_{1,st}$ is an $m_t \times k_1$ matrix of k_1 individual-varying exogenous variables and $x_{2,st}$ is a $1 \times k_2$ vector of individual-invariant exogenous variables ($k_1 + k_2 = k_x$). Then, (414) can be written as

$$Y_{st} = \rho W_t Y_{st} + X_{1,st}\beta_1 + 1_{m_t} x_{2,st}\beta_2 + \lambda_s\left(\pi_t\right) 1_{m_t} + U_{st}^{*}. \tag{418}$$

Let $Q_t = I_{m_t} - \frac{1}{m_t} 1_{m_t} 1'_{m_t}$ denote the within-transformation projector. As $Q_t 1_{m_t} = 0$ and $Q_t U^*_{st} = Q_t U_{st}$, premultiplication of (414) by Q_t gives:

$$Q_t Y_{st} = \rho Q_t W_t Y_{st} + Q_t X_{1,st} \beta_1 + Q_t U_{st}. \tag{419}$$

Then, ρ and β_1 can be estimated from the within model (419) by the conditional maximum likelihood (CML) in Lee (2007).

The fixed-effect approach does not impose any restrictions on $\lambda_s(\pi_t)$. However, the within transformation may cause an identification problem esp. if $m_t = m$ for all t, similar to the one studied in Lee (2007). The fixed-effect approach can be implemented by the following steps.

- **Step 1**: We estimate the within Eq. (419) by the CML estimator in Lee (2007).
- **Step 2**: We obtain the predicted probabilities $\hat{P}_{st}$ from, say, a conditional logit regression.
- **Step 3**: Let

$$\hat{r}_{st} = \frac{1}{m_t} 1'_{m_t} \left(Y_{st} - \hat{\rho} W_t Y_{st} - X_{1,st} \hat{\beta}_1 \right),$$

  where $\hat{\rho}$ and $\hat{\beta}_1$ are the first-step estimates. We consider the regression:

$$\hat{r}_{st} = x_{2,st} \beta_2 + \mu\left(\hat{P}_{st}\right) + \zeta_{st},$$

  where the selectivity bias $\mu\left(\hat{P}_{st}\right)$) is either given by $-\sigma_{12}\phi\left(\Phi - 1(\hat{P}_{st})\right)/\hat{P}_{st}$ in the parametric approach or approximated by $\sum_{k=1}^{K} \kappa_k b_k(\hat{P}_{st})$ in the semi-parametric approach. We estimate β_2 together with the unknown parameters in $\mu\left(\hat{P}_{st}\right)$ by the OLS estimator.
- **Remark**: The parametric and semi-parametric approaches have the advantage of computational simplicity. However, both approaches impose strong restrictions on the selectivity bias $\lambda_s(\pi_t)$ to reduce its dimensionality. Because of the endogeneity of the peer effect regressor, the model needs to be estimated by the 2SLS estimator that relies on the existence of valid instruments. This may be quite challenging in empirical applications. On the other hand, the fixed effect approach does not impose any restrictions on $\lambda_s(\pi_t)$. We can use the CML or GMM estimator, which exploit both linear and quadratic moment conditions, and may outperform the 2SLS estimator that only uses linear moment conditions. However, the within transformation makes the identification of the peer effect more challenging. In particular, the within equation is not identified if m_t does not vary over time. In this case identification can be achieved by imposing exclusion restrictions through heterogeneous peer effects.

- **Comments**:
  - No CSD...
  - "The selectivity bias $\lambda_s(\pi_t)$ can be considered as a time-varying lineup-specific fixed effect. To avoid estimating the unknown function $\lambda_s(.)$, we can apply a within transformation to eliminate this term." Is this sufficiently general? In other words, how is the endogeneity correction in (413) is general? Suppose that $\lambda_s(\pi_t)$ is time-varying, still the within transformation will remove $\lambda_s(\pi_t) 1_{m_t}$ for all t?
  - In the threshold model, Kourtellos et al. (2015) have addressed an issue of endogenous threshold variable in the single equation context. "Our estimation of the threshold parameter is based on a two-stage concentrated least squares method that involves an inverse Mills ratio bias correction term in each regime." So can we adopt the approach by Horrace et al. (2015) by extending the threshold model to the panel context. This would be an alternative approach to Seo and Shin (2015).
  - Notice that the model (398) is also similar to KMS, but it imposes that d_{jt}'s are known a priori. So we rewrite (398) using the KMS threshold selection mechanism as

$$y_{it} = \rho \frac{1}{m_{it}} \sum_{j=1}^{n} I(|z_{it} - z_{jt}| < r) y_{jt} + x_{it}\beta + E(u_{it}|\mathbf{z}_t) + u_{it}^*, \quad (420)$$

    where $\mathbf{z}_t = (z_{1t}, ..., z_{nt})'$. As $m_{it} = \sum_{j=1}^{n} I(|z_{it} - z_{jt}| < r)$, the equal weights are row-normalised by construction. In this regard, (420) is a natural generalisation of (398), where all networks are generally unobserved a priori or where the network members are potentially randomly chosen individuals. "Bandiera et al. (2009) investigate how social connections between workers and managers affect the productivities of fruit pickers in the UK. Their measure of social connectedness is based on similarities of worker/manager characteristics, and there are multiple managers whose worker assignments change daily." **We may consider and follow this line of the research trend.**
  - Notice that Horrace et al. (2015) argue that m_t is often predetermined (e.g., in sports games, the number of active players m_t is fixed). Next, workers work on the project to produce output for a given time period. For the population of n workers, the $n \times n$ adjacency matrix across all projects is potentially endogenous. By focusing on a single project of interest, we have an $m \times m$ submatrix of the adjacency matrix which is exogenous conditional on selection into the specific project. Thus, the network endogeneity is reduced to a selectivity bias, which can be corrected using a fixed effect estimator or a polychotomous Heckman-type bias correction procedure due to Lee (1983) and Dahl (2002).

- Our setup allows us to give a structural interpretation to the selectivity bias correction term that links the model to the productive efficiency literature in that the bias can viewed as "managerial (in)competence" or (in)efficiency, depending on the sign of the estimate. **So in this regard, we may consider the stochastic frontier type application or extension?**
- KMS extension seems to be complicated. Here we have N individuals, and select the m_t members out of N-idnviduals at time $\dot{t}$. HLP assume that the selction is pre-determined and then control for selctivity bias. Now, we wish to endogenise the selection process using the threshold mechanism advanced by KMS. Then, how? Next, we control endogeneity of selection by using the generic control fucntion approach. Here the main aim is to control for endogeneity, not interested in measuring the selection bias parametrically or semiparametrically.

- Consider the following KMS type model:

$$y_{it} = \rho \frac{1}{m_{it}} \sum_{j=1}^{n} I\left(|q_{it} - q_{jt}| < r\right) y_{jt} + \boldsymbol{\beta}' \mathbf{x}_{it} + u_{it}, \tag{421}$$

where $I(A)$ is an indicator function, taking unity if the event A is true and 0 otherwise, and $m_{it} = \sum_{j=1}^{n} I\left(|q_{it} - q_{jt}| < r\right)$ with r being a threshold parameter to be estimated. So the individual productivity is spatially affected by peers or members. Now, q_{it} is likely to be correlated. To control for such endogeneity, we consider the following control function:

$$q_{it} = \boldsymbol{\lambda}' \mathbf{q}_t^{(i)} + v_{it}, \tag{422}$$

where

$$\underset{(N-1)\times 1}{\mathbf{q}_t^{(i)}} = \left[\begin{array}{ccc} q_{1t} & \cdots & q_{Nt} \end{array} \right], \ \underset{(N-1)\times 1}{\boldsymbol{\lambda}} = \left[\begin{array}{c} \lambda_1 \\ \vdots \\ \lambda_N \end{array} \right]$$

Assume now:

$$\left(\begin{array}{c} u_{it} \\ v_{it} \end{array} \right) \sim \left(0, \left(\begin{array}{cc} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{array} \right) \right) \tag{423}$$

then we construct

$$u_{it} = \delta v_{it} + e_{it}$$

where

$$\delta = \frac{\sigma_{uv}}{\sigma_v^2}, \ e_{it} \sim \left(0, \sigma_e^2 = \sigma_u^2 - \frac{\sigma_{uv}^2}{\sigma_v^2} \right)$$

Hence, we rewrite (421) as

$$y_{it} = \rho \frac{1}{m_{it}} \sum_{j=1}^{n} I\left(|q_{it} - q_{jt}| < r\right) y_{jt} + \boldsymbol{\beta}' \mathbf{x}_{it} + \delta \left(q_{it} - \boldsymbol{\lambda}' \mathbf{q}_t^{(i)} \right) + e_{it}, \tag{424}$$

Now, the selection is exogneous with respect to e_{it}, so that we follow the KMS algorithm.

- My main query is how to extend the above logic from the individal level to the line-up level analysed as in your paper?

- A few more questions:

  - I am not still 100% convinced about your estimation details.
  - You now estimate eq. (22) in empirical applications. How do you construct, Y_{st}? Is it an $m \times 1$ vector? Then what is the data demension? How do you write the model in matrix form in terms of $Y_t = \left(Y'_{1t}, ...Y'_{q_t t}\right)'$? Also, the data seems to be unbalanced as the specific line-up (a total of 79 different line-ups) does not play at all time period. Then, not sure how to handle such unblanced data?
  - You propose the within estimator with $Q_t = I_{m_t} - \frac{1}{m_t} 1_{m_t} 1'_{m_t}$ denoting the within-transformation projector, since the selectivity bias $\lambda_s(\pi_t)$ can be considered as a time-varying lineup-specific fixed effect. But this is a bit confusing, as it looks like the time effect, and Q_t constructs the deviation from the cross-section or group average rather than the indvidual mean over time.
  - In Section 5.3.1, "we can use player dummies to control for unobserved player-specific characteristics." I am a bit unclear about how to combine them... Is it feasible to introduce the line-up specific individual effects from eq. (7)? More generally, why not controlling for interactive effects by

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta + \varepsilon_{st}, \ \ s = 1, ..., q_t, t = 1, ..., T.$$

$$\varepsilon_{st} = \boldsymbol{\lambda}'_s \mathbf{f}_t + U_{st}$$

    so that the model can also allow for strong form of cross-section dependence?

- Initially, I thought that the KMS extension seems to be straighforward. However, here we have N individuals, and select the m_t members out of N-individuals at every time $\dot{t}$. You assume that the selction is pre-determined and then control for selctivity bias. Now, we wish to endogenise the selection process using the threshold mechanism advanced by KMS. Then, how we move from an individual selection to the group selection? If feasible, my idea is to control endogeneity of selection by using the generic control fucntion approach. Here remind that the main aim is to control for endogeneity, not interested in measuring the selection bias parametrically or semi-parametrically.

11.4 Threshold model extensions

11.4.1 The single-equation threshold model considered by Kourtellos et al. (2015)

Let $\{y_i, z_i, x_i, q_i\}_{i=1}^n$ be an *i.i.d* or a weakly dependent observed sample, where y_i is real valued, z_i is an $l \times 1$ vector, x_i is a $p \times 1$ vector such that $l \geq p$, and q_i is a scalar. Consider the following structural threshold regression model:

$$y_i = \beta'_{x1} x_i + u_i, \ q_i \leq \gamma, \tag{425}$$

$$y_i = \beta'_{x2} x_i + u_i, \ q_i > \gamma, \tag{426}$$

where q_i is the threshold variable that splits the sample into two regimes. In each of the two linear models, y_i is a dependent variable, x_i is a vector of slope variables (regressors) including an intercept, and u_i is the equation error with $E(u_i|\mathcal{F}_{i-1}) = 0$, where $\mathcal{F}_{i-1} = \{z_{i-j}, x_{i-1-j}, q_{i-1-j}, u_{i-1-j} : j \geq 0\}$.

Consider the case where x_i is a vector of strictly exogenous regressors and a strict subset of z_i.[37] Then endogeneity bias arises when u_i is correlated with q_i. Consider the reduced form model for the threshold variable q_i given by

$$q_i = \pi'_q z_i + v_{qi} \tag{427}$$

where $E(v_{qi}|\mathcal{F}_{i-1}) = 0$. Then, endogeneity amounts to $E(u_i|\mathcal{F}_{i-1}, v_{qi}) \neq 0$. (427) is analogous to a selection equation in the literature on limited dependent variable models (Heckman, 1979). The main difference is that while limited dependent variable models treat q_i as latent and the sample split as observed, here we treat the sample split value as an unknown parameter.

Assumption 1.
1.1 $E(u_i|\mathcal{F}_{i-1}) = 0$
1.2 $E(v_{qi}|\mathcal{F}_{i-1}) = 0$
1.3 $E(u_i|\mathcal{F}_{i-1}, v_{qi}) = E(u_i|v_{qi})$
1.4 $E(u_i|v_{qi}) = \kappa v_{qi}$
1.5 $v_{qi} \sim N(0, 1)$

Assumption 1.4 assumes a linear conditional expectation between the errors of the structural and the reduced form equations.[38] Assumptions 1.4 and 1.5 can be relaxed and **the bias correction terms can be estimated by semi-parametric methods such as a series approximation**; see Kourtellos et al. (2015).

Using Assumption 1 we get:

$$\begin{aligned} E\left(u_i|\mathcal{F}_{i-1}, v_{qi} \leq \gamma - \pi'_q z_i\right) &= \kappa E\left(v_{qi}|v_{qi} \leq \gamma - \pi'_q z_i\right) \\ &= \kappa \int_{-\infty}^{\gamma - \pi'_q z_i} v_q f\left(v_q|v_q \leq \gamma - \pi'_q z_i\right) dv_q = \kappa \lambda_1\left(\gamma - \pi'_q z_i\right) \end{aligned}$$

[37] It is straightforward to allow endogenous regressors, x_i.

[38] Not sure sufficiently general? For example, when they allow for regime-specific inverse Mills ratios, why not allowing for regime-specific conditional means here too?

$$\begin{aligned} E\left(u_i|\mathcal{F}_{i-1}, v_{qi} > \gamma - \pi_q' z_i\right) &= \kappa E\left(v_{qi}|v_{qi} > \gamma - \pi_q' z_i\right) \\ &= \kappa \int_{\gamma-\pi_q' z_i}^{\infty} v_q f\left(v_q|v_q > \gamma - \pi_q' z_i\right) dv_q = \kappa\lambda_2\left(\gamma - \pi_q' z_i\right) \end{aligned}$$

where

$$\lambda_1\left(\gamma - \pi_q' z_i\right) = \frac{-\phi\left(\gamma - \pi_q' z_i\right)}{\Phi\left(\gamma - \pi_q' z_i\right)} \; and \; \lambda_2\left(\gamma - \pi_q' z_i\right) = \frac{\phi\left(\gamma - \pi_q' z_i\right)}{1 - \Phi\left(\gamma - \pi_q' z_i\right)}$$

are the inverse Mills ratio terms. $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal pdf and cdf. Note that the normality of v_{qi} is key for the derivation of the inverse Mills ratio terms.

Denote the inverse Mills ratio terms at the true value π_{q0} as

$$\lambda_{1i}(\gamma) = \lambda_1\left(\gamma - \pi_{q0}' z_i\right), \quad \lambda_{2i}(\gamma) = \lambda_2\left(\gamma - \pi_{q0}' z_i\right) \tag{428}$$

Then taking conditional expectations yields

$$E\left(y_i|\mathcal{F}_{i-1}, v_{qi} \leq \gamma - \pi_{q0}' z_i\right) = \beta_{x1}' x_i + E\left(u_i|\mathcal{F}_{i-1}, v_{qi} \leq \gamma - \pi_{q0}' z_i\right) = \beta_{x1}' x_i + \kappa\lambda_{1i}(\gamma)$$

$$E\left(y_i|\mathcal{F}_{i-1}, v_{qi} > \gamma - \pi_{q0}' z_i\right) = \beta_{x2}' x_i + E\left(u_i|\mathcal{F}_{i-1}, v_{qi} > \gamma - \pi_{q0}' z_i\right) = \beta_{x2}' x_i + \kappa\lambda_{2i}(\gamma)$$

The STR model is then defined by

$$y_i = \beta_{x1}' x_i + \kappa\lambda_{1i}(\gamma) + \varepsilon_{1i}, \; q_i \leq \gamma,$$

$$y_i = \beta_{x2}' x_i + \kappa\lambda_{2i}(\gamma) + \varepsilon_{2i}, \; q_i > \gamma,$$

where

$$\varepsilon_{1i} = -\kappa\lambda_{1i}(\gamma) + u_i \; and \; \varepsilon_{2i} = -\kappa\lambda_{2i}(\gamma) + u_i.$$

It is useful to write the model in a single equation by making the following definitions

$$I(\cdot) = 1 \; iff \; q_i \leq \gamma, \; 0 \;\; iff \; q_i > \gamma$$

$$\Lambda_i(\gamma) = \lambda_{1i}(\gamma) I(q_i \leq \gamma) + \lambda_{2i}(\gamma) I(q_i > \gamma)$$

$$\varepsilon_i = \varepsilon_{1i} I(q_i \leq \gamma) + \varepsilon_{2i} I(q_i > \gamma)$$

We can then express equations as

$$y_i = \beta_{x1}' x_i I(q_i \leq \gamma) + \beta_{x2}' x_i I(q_i > \gamma) + \kappa\Lambda_i(\gamma) + \varepsilon_i, \tag{429}$$

where $E\left(\varepsilon_i|\mathcal{F}_{i-1}\right) = 0$. The STR model, (429) nests the threshold regression model of Hansen (2000) with $\kappa = 0$. Another difference is that the presence of different inverse Mills ratios in each of the regimes in STR necessarily implies the presence of regime-specific heteroskedasticity.

Endogeneity in Both the Threshold and Slope Variables When x_i's are also endogenous and not a subset of z_i, the reduced form model for x_i is:

$$x_i = \Pi_x' z_i + v_{xi}, \tag{430}$$

where $E(v_{xi}|\mathcal{F}_{i-1}) = 0$ and Π_x is a $l \times p$ matrix of unknown parameters. Denote the conditional expectation at the true value Π_{x0} as

$$g_{xi} = E(x_i|\mathcal{F}_{i-1}) = \Pi_{x0}' z_i$$

Assumptions 1.1–1.5 augmented with
1.6 $E(v_{xi}|\mathcal{F}_{i-1}) = 0$.
1.7 $v_{xi} \perp I\left(v_{qi} \leq \gamma - \pi_q' z_i\right)|\mathcal{F}_{i-1}$
Assumptions 1.6 and 1.7 allow us to write[39]

$$E\left(x_i|\mathcal{F}_{i-1}, v_{qi} \leq \gamma - \pi_q' z_i\right) = E(x_i|\mathcal{F}_{i-1}) = \Pi_{x0}' z_i$$

$$E\left(x_i|\mathcal{F}_{i-1}, v_{qi} > \gamma - \pi_q' z_i\right) = E(x_i|\mathcal{F}_{i-1}) = \Pi_{x0}' z_i$$

Then, we have:

$$E\left(y_i|\mathcal{F}_{i-1}, v_{qi} \leq \gamma - \pi_{q0}' z_i\right) = \beta_{x1}' g_{xi} + \kappa\lambda_{1i}(\gamma)$$

$$E\left(y_i|\mathcal{F}_{i-1}, v_{qi} > \gamma - \pi_{q0}' z_i\right) = \beta_{x2}' g_{xi} + \kappa\lambda_{2i}(\gamma)$$

STR model that allows for endogeneity in both threshold and slope variables can be written as:

$$y_i = \beta_{x1}' g_{x_i} I\left(q_i \leq \gamma\right) + \beta_{x2}' g_{x_i} I\left(q_i > \gamma\right) + \kappa\Lambda_i(\gamma) + e_i^*,$$

where

$$e_i^* = \beta_{x1}' v_{x_i} I\left(q_i \leq \gamma\right) + \beta_{x2}' v_{x_i} I\left(q_i > \gamma\right) + \varepsilon_i$$

with $E(e_i^*|\mathcal{F}_{i-1}) = 0$.[40]

We proceed in three steps to estimate the model: a two-step concentrated LS method to estimate the threshold parameter and an additional step to produce estimates of the slope coefficients.

YC Remark: read the paper for detailed estimation procedure, but I find their proposed estimation technique seems rather complicated and cumbersome.

[39] One could allow dependence between v_{xi} and v_{qi} by assuming that $E(v_{xi}|F_{i-1}, v_{qi})$ is a linear function of v_{qi}, which implies the need for an additional inverse Mills ratio term in each regime.

[40] One possible concern is the assumption of linearity in the reduced form of x_i. This assumption can be relaxed to allow for nonlinearities such as a threshold regression in the first stage. However, **this extension is not trivial**.

11.4.2 Semiparametric Threshold Regression based on nonparametric CF by Kourtellos et al. (2016) Incomplete

We propose a semiparametric approach to deal with the endogeneity of threshold variable and regressors that relaxes the parametric assumptions of Kourtellos, Stengos, and Tan (2015). Specifically, we propose to estimate the threshold parameter using a concentrated least squares (CLS) which includes a regime specific control function estimated by series estimation method based on polynomial and splines.

Consider the basic parametric structural threshold regression (or STR) model:

$$y_t = x_t'\beta_1 + \sigma_1 u_t, q_t \leq \gamma_0$$

$$y_t = x_t'\beta_2 + \sigma_2 u_t, q_t > \gamma_0$$

for $t = 1, 2, ..., n$, where y_t is the log income per capita in country t, q_t is an endogenous threshold variable (such as the quality of institutions) with γ_0 being the sample split value, x_t is a $d_x \times 1$ vector of growth determinants, β_1 and β_2 are regime-specific slope coefficients, and u_t is an error with zero mean and unit variance.

A reduced form equation for q_t is given by

$$q_t = z_t'\pi_q + v_{q,t}, t = 1, 2, ..., n,$$

where $E(v_{q,t}, z_t) = 0$ for all t. Assuming

$$E(u_t|x_t, z_t, v_{q,t}) = E(u_t|v_{q,t}) = g(v_{q,t})$$

almost surely, where $g(\cdot)$ is a smooth unknown function to be estimated. Letting F_v be the cdf of $v_{q,t}$, we obtain

$$E(u_t|v_{q,t} \leq \gamma_0 - z_t'\pi_q) = \frac{E[g(v_{q,t})I(v_{q,t} \leq \gamma_0 - z_t'\pi_q)]}{F_v(\gamma_0 - z_t'\pi_q)} \equiv h_1(\gamma_0 - z_t'\pi_q),$$

$$E(u_t|v_{q,t} > \gamma_0 - z_t'\pi_q) = \frac{E[g(v_{q,t})I(v_{q,t} > \gamma_0 - z_t'\pi_q)]}{1 - F_v(\gamma_0 - z_t'\pi_q)} \equiv h_2(\gamma_0 - z_t'\pi_q),$$

Therefore, we can rewrite the model as

$$y_t = x_t'\beta_1 + \sigma_1 h_1(\gamma_0 - z_t'\pi_q) + \varepsilon_{1t}, q_t \leq \gamma_0$$

$$y_t = x_t'\beta_2 + \sigma_2 h_2(\gamma_0 - z_t'\pi_q) + \varepsilon_{2t}, q_t > \gamma_0$$

where

$$\varepsilon_{jt} = \sigma_j[u_t - h_j(\gamma_0 - z_t'\pi_q)] \; for \; j = 1, 2.$$

And, combining together gives a single equation:

$$y_t = x_t'\beta_2 + x_t'\delta I(q_t \leq \gamma_0) + h(\gamma_0 - z_t'\pi_q) + \varepsilon_t, \tag{431}$$

where $\delta = \beta_1 - \beta_2$, the regression error,

$$\varepsilon_t = \varepsilon_{1t} I(q_t \leq \gamma_0) + \varepsilon_{2t} I(q_t > \gamma_0),$$

$$h\left(\gamma_0 - z_t'\pi_q\right) = \sigma_1 h_1\left(\gamma_0 - z_t'\pi_q\right) + \sigma_2 h_1\left(\gamma_0 - z_t'\pi_q\right),$$

We need to impose an identification condition as we cannot identify $(\gamma, \sigma_1, \sigma_2)$ from the unknown functions $h(\cdot)$, $h1(\cdot)$, and $h2(\cdot)$. Therefore, we propose the following model:

$$y_t = x_t'\beta_2 + x_t'\delta I\left(q_t \leq \gamma_0\right) + h\left(\gamma_0 - z_t'\pi_q\right) + \varepsilon_t,$$

$$\begin{aligned} h\left(z_t'\pi_q\right) &= h_1\left(z_t'\pi_q\right) I\left(q_t \leq \gamma_0\right) + h_2\left(z_t'\pi_q\right) I\left(q_t > \gamma_0\right) \\ &= h_2\left(z_t'\pi_q\right) + \eta\left(z_t'\pi_q\right) I\left(q_t \leq \gamma_0\right) \end{aligned}$$

where

$$\eta\left(z_t'\pi_q\right) = h_1\left(z_t'\pi_q\right) - h_2\left(z_t'\pi_q\right)$$

captures the endogenous threshold effect. We set

$$h(0) = h_1(0) = h_2(0) = 0$$

for identification purpose when x_t contains a constant term one.

Let $\{\phi_1(w), \phi_2(w), ...\}$ be a sequence of orthonormal basis functions in $L_2(-\infty, \infty)$ space if q_t takes value from the real line or $L_2[0,1]$ space if q_t has a finite support. We approximate $h\left(z_t'\pi_q\right)$ by

$$h^*\left(z_t'\pi_q\right) = h_1^*\left(z_t'\pi_q\right) I(q_t \leq \gamma_0) + h_2^*\left(z_t'\pi_q\right) I(q_t > \gamma_0) = h_2^*\left(z_t'\pi_q\right) + \eta^*\left(z_t'\pi_q\right) I(q_t \leq \gamma_0),$$

where we denote an $L_n \times 1$ vector, $\Phi_{L_n}(w) = \{\phi_1(w), ..., \phi_{L_n}(w)\}'$, and $h_j^*(w) = \alpha_{L_n,j}'\Phi_{L_n}(w)$ for $j = 1, 2$, and $\eta^*(w) = (\alpha_{L_n,1} - \alpha_{L_n,2})'\Phi_{L_n}(w)$.

Our four-step estimation procedure is given as follows.

Step 1. For a given $\gamma \in [\underline{\gamma}, \bar{\gamma}]$, we estimate $\theta = \left(\beta_1', \alpha_{L_n,1}', \beta_2', \alpha_{L_n,2}'\right)'$ from the objective function

$$\hat{\theta} = a\arg\min_{\theta} \sum_{t=1}^{n} \left[y_t - x_{-,t}'\beta_1 - \alpha_{L_n,1}'\Phi_{L_n,\gamma}^{-}\left(\hat{q}_t\right) - x_{+,t}'\beta_2 - \alpha_{L_n,2}'\Phi_{L_n,\gamma}^{+}\left(\hat{q}_t\right)\right]^2,$$

where we denote $x_{-,t} = x_t I(q_t \leq \gamma)$, $x_{+,t} = x_t I(q_t > \gamma)$, $\Phi_{L_n,\gamma}^{-}\left(\hat{q}_t\right) = \Phi_{L_n,\gamma}\left(\hat{q}_t\right) I(q_t \leq \gamma)$ and $\Phi_{L_n,\gamma}^{+}\left(\hat{q}_t\right) = \Phi_{L_n,\gamma}\left(\hat{q}_t\right) I(q_t > \gamma)$. Denoting $\mathcal{X}_{\blacksquare} = [\mathcal{X}_{-,\gamma}, \mathcal{X}_{+,\gamma}]$, where $\mathcal{X}_{-,\gamma}$ stacks up $\left[x_{-,t}', \Phi_{L_n,\gamma}^{-\prime}\left(\hat{q}_t\right)\right]$ and $\mathcal{X}_{+,\gamma}$ stacks up $\left[x_{+,t}', \Phi_{L_n,\gamma}^{+\prime}\left(\hat{q}_t\right)\right]$, and solving (2.14) give

$$\hat{\theta}(\gamma) = \left(\mathcal{X}_\gamma'\mathcal{X}_\gamma\right)^{-1}\mathcal{X}_\gamma' y.$$

Step 2. We estimate the threshold parameter γ by minimizing the concentrated least squares criterion:

$$\hat{\gamma} = \arg \min_{\gamma \in [\underline{\gamma}, \bar{\gamma}]} \sum_{t=1}^{n} \left[y_t - \mathcal{X}'_{t,\gamma} \hat{\theta}(\gamma) \right]^2$$

Step 3. Calculate

$$\hat{y}_t = y_t - \hat{\alpha}'_{L_n,1}(\hat{\gamma}) \Phi^{-}_{L_n,\gamma}(\hat{q}_t) - \hat{\alpha}'_{L_n,2}(\hat{\gamma}) \Phi^{+}_{L_n,\gamma}(\hat{q}_t),$$

We then run a linear regression model,

$$\hat{y}_t = x'_t \beta_2 + \delta'_n x_t I(q_t \leq \hat{\gamma}) + error_t$$

and obtain the OLS estimator $\tilde{\beta}_2$ and $\tilde{\delta}_n$ for β_2 and δ_n respectively.

Step 4. Calculate

$$\check{y}_t = y_t - x'_{-,t} \hat{\beta}_1 - x'_{+,t} \hat{\beta}_2.$$

We then re-estimate $h_2(w)$ and $\eta_n(w)$ by the local linear regression approach from

$$\check{y}_t = h_2(\hat{q}_t) + \eta_n(\hat{q}_t) I(q_t \leq \hat{\gamma}) + error_t, \ t = 1, 2, ..., n.$$

We denote the estimator for $\psi(w) = [h_2(w), \eta_n(w)]'$ by $\tilde{\psi}(w) = [\tilde{h}_2(w), \tilde{\eta}_n(w)]'$.

11.4.3 YC's hunch

Rewrite (425) and (426) as

$$\begin{aligned} y_i &= \beta'_{x1} x_i (1 - I(q_i > \gamma)) + \beta'_{x2} x_i I(q_i > \gamma) + u_i \\ &= \beta'_{x1} x_i + \left(\beta'_{x2} - \beta'_{x1}\right) x_i I(q_i > \gamma) + u_i \end{aligned} \tag{432}$$

We follow Qu and Lee (2015) and make the following assumptions:

$$\begin{pmatrix} u_i \\ v_{qi} \end{pmatrix} \sim \left(0, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}\right) \tag{433}$$

then we construct

$$u_i = \delta v_{qi} + e_i$$

where

$$\delta = \frac{\sigma_{uv}}{\sigma_v^2}, \ e_i \sim \left(0, \sigma_e^2 = \sigma_u^2 - \frac{\sigma_{uv}^2}{\sigma_v^2}\right)$$

Hence, we rewrite (432) simply as

$$y_i = \beta'_{x1} x_i + \left(\beta'_{x2} - \beta'_{x1}\right) x_i I(q_i > \gamma) + \delta \left(q_i - \pi'_q z_i\right) + e_i. \tag{434}$$

I trust that there is no longer endogeneity of q_i in (434) as it is uncorrelated with e_i by construction. There is no need to allow for regime-dependent inverse Mills ratios for the paramertic model or regime-dependent series approximations for the semiparamertic model. As

briefly discussed, the use of CF seems to be robust to any nonlinearity associated with the model. (Pls check!) Further, the assumption in (433) does not require that $\begin{pmatrix} u_i \\ v_{qi} \end{pmatrix}$ **are normally distributed.**

- **So the bottom line is that if the CF transformation in (434) renders** q_i **exogenous, then we may proceed to develop the associated estimation and inference from (434) in a rather straightforward manner. If the main aim is to make the transition variable** q_i **exogenous, then I trust that the use of (434) is sufficient for consistent estimation of regime-dependent slope parameters,** β_1 **and** β_2**. Notice that the approximation by the inverse Mills-ratio used in (429) is valid only if the joint normality condition is satisfied. So the comparison of the estimator obtained from (429) and (434) is similar to the RE and FE estimators. If the joint normality is valid, the estimator by (429) is consistent and more efficient than the estimator by (434). If not, only the estimator by (434) is consistent. Also, here, we are not interested in finding out the exact magnitude of the (specific) selection bias terms. So our motivation to provide much simpler but robust approach. So could you please check the validity of this conjecture?**

Along similar logics, we can also control for endogenous regressors. We now consider the case with endogenous regressors as in (430). We combine (427) and (430) and obtain:

$$\begin{pmatrix} q_i \\ x_i \end{pmatrix} = \begin{pmatrix} \pi_q' \\ \Pi_x' \end{pmatrix} z_i + \begin{pmatrix} v_{qi} \\ v_{xi} \end{pmatrix} \tag{435}$$

which can be written as[41]

$$\underset{(p+1)\times 1}{X_i} = \underset{(p+1)\times l_l \times 1}{\Pi} z_i + \mathbf{v}_i \tag{436}$$

We thus modify the assumption, (433) as follows:

$$\begin{pmatrix} u_i \\ \mathbf{v}_i \end{pmatrix} \sim \left(0, \begin{pmatrix} \sigma_u^2 & \Sigma_{uv} \\ \Sigma_{uv}' & \Sigma_{vv} \end{pmatrix}\right) \tag{437}$$

then we construct

$$u_i = \boldsymbol{\delta}' \mathbf{v}_i + e_i$$

where

$$\boldsymbol{\delta} = \Sigma_{vv}^{-1}\Sigma_{uv},\ e_i \sim \left(0, \sigma_e^2 = \sigma_u^2 - \Sigma_{uv}'\Sigma_{vv}^{-1}\Sigma_{uv}\right)$$

Hence, we rewrite (432) as

$$y_i = \beta_{x1}' x_i + \left(\beta_{x2}' - \beta_{x1}'\right) x_i I(q_i > \gamma) + \boldsymbol{\delta}' \left(X_i - \Pi z_i\right) + e_i. \tag{438}$$

[41] In this way we can allow for correlation between v_{qi} and v_{xi}.

Again I trust that there is no longer endogeneity of q_i and x_i in (438) as they are uncorrelated with e_i by regression construction. So if (438) works, it does not impose any (potentially strong) assumptions as in say, Kourtellos et al. (2015).

11.4.4 The panel data extension

Next, we develop the panel data extension along with dynamic modelling, also paying more attention to the FE approaches by Horrace et al. (2015).

Panel Threshold Regression Models with Endogenous Threshold Variables by Wang and Lin (2010) This is a panel extension of Kourtellos et al. (2015). Consider the panel threshold model:

$$y_{it} = x_{it} I\left(q_{it} \leq \theta\right) \beta_1 + x_{it} I\left(q_{it} > \theta\right) \beta_2 + e_{it} \tag{439}$$

$$q_{it} = z_{it}\pi + u_{it}; \tag{440}$$

where q_{it} is an observed threshold variables, θ is an unknown threshold parameter, and z_{it} is a vector of instruments.

Assumption 2.1. $\{y_{it}, x_{it}, q_{it}, e_{it}\}$ is strictly stationary, ergodic.

Assumption 2.2. $E\left|x_{it}\right|^4 < \infty$ and $E\left|e_{it}\right|^4 < \infty$.

Assumption 2.3. $n \to \infty$ and T is xed.

Assumption 2.4. For some fixed number $G < \infty$ and $0 < \alpha < 1/2$, $\delta = \beta_2 - \beta_1 = n^{-\alpha} G$.

Assumption 2.5. Let $f_t\left(\theta\right)$ denote the density function of q_{it}. Define

$$D\left(\theta\right) = \sum_{t=1}^{T} E(Gx_{it}|q_{it} = \theta) f_t\left(\theta\right)$$

$D\left(\theta\right)$ is continuous at $\theta = \theta_0$.

Assumption 2.6. $D\left(\theta\right) = D, 0 < D < 1$.

Assumption 2.7. $u_{it}|z_{it} \sim N(0,1)$.

Assumption 2.8. The joint distribution between e_{it} and u_{it} is defined as:

$$\left(\begin{array}{c} e_{it} \\ u_{it} \end{array}\right) |x_{it}, z_{it} = N\left(0, \left(\begin{array}{cc} \sigma_e^2 & \gamma_j \\ \gamma_j & 1 \end{array}\right)\right)$$

where γ_j is the covariance between e_{it} and u_{it}; $\gamma_j = \gamma_1$ when $q_{it} \leq \theta$ and $\gamma_j = \gamma_2$ when $q_{it} > \theta$.

- **Remark: why is covariance regime-dependent?**

Assumptions (2.7) and (2.8) impose the correlation relationship between threshold variable and panel threshold errors. We are able to find the accurate functional form of the **generated regressor** for a two-stage bias correction

estimator with Assumption (2.7). Assumption (2.8) describes the endogeneity structure of the panel threshold model.

From Assumption (2.8) and KST (2007), it is useful to decompose e_{it} into two parts:

$$\begin{pmatrix} \varepsilon_{it} \\ u_{it} \end{pmatrix} = \begin{pmatrix} 1 & -\gamma_j \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix}$$

meaning

$$e_{it} = \gamma_1 u_{it} I\left(q_{it} \leq \theta\right) + \gamma_2 u_{it} I\left(q_{it} > \theta\right) + \varepsilon_{it} \tag{441}$$

- **Remark: This is more general than Assumption 1.4 in Kourtellos et al. (2015). Still not quite intuitive why we consider the same threhsold model for the regression of e_{it} (main equation) on u_{it} (transition variable regression)? Any meaningful economic example?**

We then get the joint distribution of ε_{it} and u_{it}:

$$\begin{pmatrix} \varepsilon_{it} \\ u_{it} \end{pmatrix} | x_{it}, z_{it} = N\left(0, \begin{pmatrix} \sigma_e^2 - \gamma_j^2 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Under Assumptions (2.1)-(2.8), the effect introduced by endogenous threshold variables, $\gamma_j u_{it}$ enters Equation (1) linearly.[42] When $q_{it} \leq \theta$, the conditional expectation of panel threshold model is:

$$E\left[y_{it} | x_{it}, z_{it}, q_{it} \leq \theta\right] = E\left[y_{it} | x_{it}, z_{it}, u_{it} \leq \theta - z_{it}\pi\right] = x_{it}\beta_1 + \lambda_1\left(\theta - z_{it}\pi\right)$$

When $q_{it} > \theta$, we have that:

$$E\left[y_{it} | x_{it}, z_{it}, q_{it} > \theta\right] = E\left[y_{it} | x_{it}, z_{it}, u_{it} > \theta - z_{it}\pi\right] = x_{it}\beta_2 + \lambda_2\left(\theta - z_{it}\right)$$

where

$$\lambda_1\left(\theta - z_{it}\pi\right) = -\frac{\phi\left(\theta - z_{it}\pi\right)}{\Phi\left(\theta - z_{it}\pi\right)},\ \lambda_2\left(\theta - z_{it}\pi\right) = -\frac{\phi\left(\theta - z_{it}\pi\right)}{1 - \Phi\left(\theta - z_{it}\pi\right)}$$

$\phi\left(\right)$ and $\Phi\left(\right)$ denote the density and cumulated density function of a standard normal distribution. We can rewrite panel threshold model with endogenous threshold variables:

$$y_{it} = x_{it} I\left(q_{it} \leq \theta\right)\beta_1 + x_{it} I\left(q_{it} > \theta\right)\beta_2 + \psi\left(q_{it}, z_{it}, \theta, \pi\right) + \varepsilon_{it} \tag{442}$$

where

$$\psi\left(q_{it}, z_{it}, \theta, \pi\right) = \gamma_1\lambda_1\left(\theta - z_{it}\pi\right) I\left(q_{it} \leq \theta\right) + \gamma_2\lambda_2\left(\theta - z_{it}\pi\right) I\left(q_{it} > \theta\right)$$

or

$$\psi\left(q_{it}, z_{it}, \theta, \pi\right) = \gamma_1\lambda_1\left(\theta - z_{it}\pi\right) + \gamma_2\lambda_2\left(\theta - z_{it}\pi\right)$$

[42] To simplify our analysis, we consider that the correlation between e_{it} and u_{it} is fixed across i and t; i.e., γ_j in this paper. This setting can be relaxed in the future study.

Our estimation procedure proceeds in three steps: First, we estimate the parameter in (440) by OLS. Second, we estimate the threshold parameter by minimizing a concentrated least square criterion using ^from first stage.

$$S^{CLS}\left(\beta_i\left(\theta\right),\gamma_i\left(\theta\right),\theta\right)=\arg\min\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}I\left(q_{it}\leq\theta\right)\beta_1-x_{it}I\left(q_{it}>\theta\right)\beta_2-\psi\left(q_{it},z_{it},\theta,\hat{\pi}\right)\right)^2$$

Third, we estimate the parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ by LS based on the split samples.

- **Remark: still weak theory and weak MC results with no application, that's why still unpublished?**

YC's proposed CF-based modelling Rewrite the panel threshold model,

$$y_{it}=x_{it}\beta_1+x_{it}I\left(q_{it}>\theta\right)\left(\beta_2-\beta_1\right)+u_{it} \tag{443}$$

and we allow both q_{it} and x_{it} to be endogenous such that

$$\left(\begin{array}{c}q_{it}\\x_{it}\end{array}\right)=\left(\begin{array}{c}\pi_q'\\\Pi_x'\end{array}\right)z_{it}+\left(\begin{array}{c}v_{qi,t}\\v_{xi,t}\end{array}\right) \tag{444}$$

which can be written as

$$\underset{(p+1)\times1}{X_{it}}=\underset{(p+1)\times ll\times1}{\Pi}z_{it}+\mathbf{v}_{it} \tag{445}$$

Assume:

$$\left(\begin{array}{c}u_{it}\\\mathbf{v}_{it}\end{array}\right)\sim\left(0,\left(\begin{array}{cc}\sigma_u^2&\Sigma_{uv}\\\Sigma_{uv}'&\Sigma_{vv}\end{array}\right)\right) \tag{446}$$

then we construct

$$u_{it}=\boldsymbol{\delta}'\mathbf{v}_{it}+e_{it}$$

where

$$\boldsymbol{\delta}=\Sigma_{vv}^{-1}\Sigma_{uv},\ e_i\sim\left(0,\sigma_e^2=\sigma_u^2-\Sigma_{uv}'\Sigma_{vv}^{-1}\Sigma_{uv}\right)$$

Hence, we rewrite (443) as

$$y_{it}=x_{it}\beta_1+x_{it}I\left(q_{it}>\theta\right)\left(\beta_2-\beta_1\right)+\boldsymbol{\delta}'\left(X_{it}-\Pi z_{it}\right)+e_{it}. \tag{447}$$

Again I trust that there is no longer endogeneity of q_{it} and x_{it} in (438) as they are uncorrelated with e_i by regression construction. So if (447) works, it does not impose any (potentially strong) assumptions as above.

- **I will work on more general case with fixed effects and cross-section dependence, also paying attention to the FE approaches by Horrace et al. (2015).**

11.4.5 Digressions to FE models

We consider the error components-based panel:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it},\ i = 1, 2, \ldots, N; t = 1, 2, \ldots, T, \tag{448}$$

$$\varepsilon_{it} = \alpha_i + u_{it}. \tag{449}$$

Here we assume:

- u_{it}'s are $N(0, \sigma_u^2)$.
- α_i's and u_{it}' are correlated with $\mathbf{x}_{it}$.

Further,

- α_i's and u_{it}' are not correlated with $\mathbf{x}_{jt}$ for $i \neq j$. (sensible??) maybe in the case where α_i uncorrelated with α_j and u_{it} uncorrelated with u_{jt}.
- We may decompose the $Nk \times 1$ vector, $\mathbf{x}_t = (\mathbf{x}'_{1t}, ..., \mathbf{x}'_{Nt})'$ into the two groups, say $\mathbf{x}_t = \left(\mathbf{x}_t^{(1)\prime}, \mathbf{x}_t^{(2)\prime}\right)$ with $\mathbf{x}_t^{(1)} = \left(\mathbf{x}'_{1t}, ..., \mathbf{x}'_{N_1 t}\right)'$ and $\mathbf{x}_t^{(2)} = \left(\mathbf{x}'_{N_1+1,t}, ..., \mathbf{x}'_{Nt}\right)'$. Then, assume that $\mathbf{x}_t^{(2)}$ are uncorrelated with α_i and u_{it} if $\mathbf{x}_{it}$ does not belong to $\mathbf{x}_t^{(2)}$ or $\mathbf{x}_{-i,t}^{(2)}$ are uncorrelated with α_i and u_{it} if $\mathbf{x}_{it}$ does not belong to $\mathbf{x}_t^{(2)}$ and $\mathbf{x}_{-i,t}^{(2)}$ indicates the vector excluding $\mathbf{x}_{it}$ from $\mathbf{x}_t^{(2)}$. e.g. $\mathbf{x}_t^{(1)}$ influential countries and $\mathbf{x}_t^{(2)}$ small countries. or $\mathbf{x}_t^{(1)}$ neighbor (similar) countries and $\mathbf{x}_t^{(2)}$ remote countries. not sure yet...

Here we consider the few candidates for the control function for $\mathbf{x}_{it}$: First, we consider the following time series regression for the $k \times 1$ vector $\mathbf{x}_{it}$:[43]

$$\mathbf{x}_{it} = \mathbf{x}_{-i,t}\boldsymbol{\Phi}_i + \mathbf{v}_{it},\quad t = 1, ..., T \tag{450}$$

where

$$\underset{k\times(N-1)}{\mathbf{x}_{-i,t}} = \left[\begin{array}{ccc} \mathbf{x}'_{1t} & \cdots & \mathbf{x}'_{Nt} \end{array}\right],\ \underset{k(N-1)\times 1}{\boldsymbol{\Phi}_i} = \left[\begin{array}{c} \boldsymbol{\Phi}_{i1} \\ \vdots \\ \boldsymbol{\Phi}_{iN} \end{array}\right]$$

Assume now:

$$\left(\begin{array}{c} u_{it} \\ \mathbf{v}_{it} \end{array}\right) \sim \left(0, \left(\begin{array}{cc} \sigma_u^2 & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}'_{uv} & \boldsymbol{\Sigma}_{vv} \end{array}\right)\right),\quad t = 1, ..., T \tag{451}$$

then we construct the following orthogonal projection:

$$u_{it} = \boldsymbol{\delta}'\mathbf{v}_{it} + e_{it},\quad t = 1, ..., T \tag{452}$$

where

$$\boldsymbol{\delta} = \boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{uv},\ e_{it} \sim \left(0, \sigma_e^2 = \sigma_u^2 - \Sigma'_{uv}\Sigma_{vv}^{-1}\Sigma_{uv}\right)$$

[43] We can easily allow unobserved individual heterogeneities in (450).

Hence, using (452), we rewrite (448) as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \boldsymbol{\delta}'\mathbf{v}_{it} + \alpha_i + e_{it} \tag{453}$$

where $\mathbf{x}_{it}$ is uncorrelated with e_{it}. Applying the within transformation to (453), we obtain:

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}\boldsymbol{\beta} + \boldsymbol{\delta}'\tilde{\mathbf{v}}_{it} + \tilde{e}_{it} \tag{454}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$ with $\bar{y}_i = T^{-1}\sum y_{it}$ and similarly for $\tilde{\mathbf{x}}_{it}$, $\tilde{\mathbf{v}}_{it}$ and $\tilde{e}_{it}$.

Construction of $\mathbf{v}_{it}$ in full: write (450) in the matrix notation:

$$\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\Phi}_i + \mathbf{V}_i \tag{455}$$

where

$$\underset{T\times k}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{iT} \end{bmatrix}, \quad \underset{T\times(N-1)k}{\mathbf{X}_{-i}} = \begin{bmatrix} \mathbf{x}'_{-i,1} \\ \vdots \\ \mathbf{x}'_{-i,T} \end{bmatrix}, \quad \underset{T\times k}{\mathbf{V}_i} = \begin{bmatrix} \mathbf{v}'_{i1} \\ \vdots \\ \mathbf{v}'_{iT} \end{bmatrix}$$

Next, we need to construct $\mathbf{v}_{it}$, and stacking (455) for $i = 1, ..., N$ by

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{X}_{-1}\boldsymbol{\Phi}_1 + \mathbf{V}_1 \\ &\vdots \\ \mathbf{X}_N &= \mathbf{X}_{-N}\boldsymbol{\Phi}_N + \mathbf{V}_N \end{aligned}$$

Thus, we obtain the final sepcification:

$$\mathbf{X} = \mathbf{X}_-\boldsymbol{\Phi} + \mathbf{V} \tag{456}$$

where

$$\underset{NT\times k}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}, \quad \underset{NT\times N(N-1)k}{\mathbf{X}_-} = \begin{bmatrix} \mathbf{X}_{-1} & & \\ & \ddots & \\ & & \mathbf{X}_{-N} \end{bmatrix}, \quad \underset{N(N-1)k\times 1}{\boldsymbol{\Phi}} = \begin{bmatrix} \boldsymbol{\Phi}_1 \\ \vdots \\ \boldsymbol{\Phi}_N \end{bmatrix}, \quad \underset{NT\times k}{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_N \end{bmatrix}$$

In the case where $N > T$, the estimatin of (450) may be subject to the incidental parameters problem. Hence, to reduce the dimensionality of $\boldsymbol{\Phi}$, we may consider the following simpler and practical approaches: first consider

$$\mathbf{x}_{it} = \bar{\mathbf{x}}_{-i,t}\bar{\boldsymbol{\Phi}}_i + \mathbf{v}_{it}, \tag{457}$$

where

$$\underset{k\times 1}{\bar{\mathbf{x}}_{-i,t}} = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N}\mathbf{x}_{jt}, \quad \underset{k\times 1}{\bar{\boldsymbol{\Phi}}_i} = \begin{bmatrix} \bar{\Phi}_{i1} \\ \vdots \\ \bar{\Phi}_{ik} \end{bmatrix}$$

In this case we have

$$\mathbf{X}_i = \bar{\mathbf{X}}_{-i}\bar{\boldsymbol{\Phi}}_i + \mathbf{V}_i \tag{458}$$

where

$$\underset{T\times k}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{x}_{i1}' \\ \vdots \\ \mathbf{x}_{iT}' \end{bmatrix}, \underset{T\times k}{\bar{\mathbf{X}}_{-i}} = \begin{bmatrix} \bar{\mathbf{x}}_{-i,1}' \\ \vdots \\ \bar{\mathbf{x}}_{-i,T}' \end{bmatrix}, \underset{T\times k}{\mathbf{V}_i} = \begin{bmatrix} \mathbf{v}_{i1}' \\ \vdots \\ \mathbf{v}_{iT}' \end{bmatrix}$$

and

$$\mathbf{X} = \bar{\mathbf{X}}_{-}\bar{\mathbf{\Phi}} + \mathbf{V} \tag{459}$$

where

$$\underset{NT\times k}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}, \underset{NT\times Nk}{\bar{\mathbf{X}}_{-}} = \begin{bmatrix} \bar{\mathbf{X}}_{-1} & & \\ & \ddots & \\ & & \bar{\mathbf{X}}_{-N} \end{bmatrix}, \underset{Nk\times 1}{\bar{\mathbf{\Phi}}} = \begin{bmatrix} \bar{\mathbf{\Phi}}_1 \\ \vdots \\ \bar{\mathbf{\Phi}}_N \end{bmatrix}, \underset{NT\times k}{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_N \end{bmatrix}$$

Alternatively, we follows KMS and consider:

$$\mathbf{x}_{it} = \breve{\mathbf{x}}_t^{(i)}\breve{\mathbf{\Phi}}^{(i)} + \mathbf{v}_{it}, \tag{460}$$

where

$$\underset{km\times 1}{\breve{\mathbf{x}}_t^{(i)}} = \frac{1}{m}\sum_{j=1,j\neq i}^{N} \mathbf{x}_{jt} 1\left\{|q_{it} - q_{jt}| \leq r\right\}, \underset{km\times 1}{\breve{\mathbf{\Phi}}^{(i)}} = \begin{bmatrix} \breve{\mathbf{\Phi}}_1^{(i)} \\ \vdots \\ \breve{\mathbf{\Phi}}_m^{(i)} \end{bmatrix}$$

Or maybe LASSO or SAR... Then, we will modify the final regression, (453), accordingly.

Next, more generally, $\mathbf{x}_{it}$ is assumed to be the smooth function of the $kN\times 1$ vector, $\mathbf{x}_t = (\mathbf{x}_{1t}, ..., \mathbf{x}_{Nt})'$. In this case the control function for $\mathbf{x}_{it}$ follows the VAR(1):

$$\mathbf{x}_t = \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{v}_t.$$

Assume now:

$$\begin{pmatrix} u_{it} \\ \mathbf{v}_{it} \end{pmatrix} \sim \left(0, \begin{pmatrix} \sigma_u^2 & \mathbf{\Sigma}_{uv} \\ \mathbf{\Sigma}_{uv}' & \mathbf{\Sigma}_{vv} \end{pmatrix}\right) \tag{461}$$

then we construct

$$u_{it} = \boldsymbol{\delta}'\mathbf{v}_{it} + e_{it}$$

where

$$\boldsymbol{\delta} = \mathbf{\Sigma}_{vv}^{-1}\mathbf{\Sigma}_{uv},\ e_{it} \sim \left(0, \sigma_e^2 = \sigma_u^2 - \Sigma_{uv}'\Sigma_{vv}^{-1}\Sigma_{uv}\right)$$

Hence, we rewrite (448) as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \boldsymbol{\delta}'\left(\mathbf{x}_t - \mathbf{\Phi}\mathbf{x}_{t-1}\right) + \alpha_i + e_{it} \tag{462}$$

or

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}\boldsymbol{\beta} + \boldsymbol{\delta}'\left(\mathbf{x}_t - \mathbf{\Phi}\mathbf{x}_{t-1}\right) + e_{it} \tag{463}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$. To reduce the dimensionality of $\mathbf{\Phi}^{(i)}$??

Mundlak (1978) argued that the dichotomy between fixed effects and random effects models disappears if we make the assumption that α_i depend on the mean values of $\mathbf{x}_i$, an assumption he regards as reasonable in many problems. As before, consider the error components model,

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\alpha}_i + \mathbf{u}_i, \tag{464}$$

but now assume

$$\alpha_i = \bar{\mathbf{x}}_i\boldsymbol{\pi} + w_i,$$

where w_i has the same properties that α_i was assumed to have; that is,

1. w_i's are $iidN(0, \sigma_w^2)$.
2. w_i's are uncorrelated with u_{jt} for all i, j, t, that is, $E[w_i u_{jt}] = 0$ for all i, j, t.
3. w_i's are uncorrelated with x_{jt} for all i, j, t, that is, $E[w_i x_{jt}] = 0$ for all i, j, t.

12 Concluding Remarks

Eventually, this project aims to

- develop the general econometrics specifications that can accommodate the spatial and factor dependence, the spatial heterogeneity, the endogenous spatial weights matrix as well as the spatial nonlinearity in dynamic heterogenous panels in a rather unified framework by combining all the recent advances made in the related literature.
- extend all the advances to the multi-dimensional dataset, separately and jointly. As the dimension grows, it would be more complicated and challenging to develop the hierarchical and strucutal structue of the spatial effects and factors, jointly.
- These works will be of great applicability to a variety of the big dataset, not only the health economics data...

13 References

Abowd, J.M., F. Kramarz and D.N. Margolis (1999): "High Wage Workers and High Wage Firms", *Econometrica* 67: 251-333.

Ahn S. and A. Horenstein (2013): Eigenvalue ratio test for the number of factors. *Econometrica* 81: 1203-1227

Ahn, S.C., Y.H. Lee, and P.Schmidt (2013): Panel data models with multiple time-varying individual effects. *Journal of Econometrics* 174: 1-14.

Anderson, T.W. and C. Hsiao (1981): Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76: 598-606.

Anderson, J. and E. van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review* 93: 170-92.

Andrews, D.W.K. (2005): Cross-section regression with common shocks. *Econometrica* 73: 1551-1585.

Anselin, L. (1988): *Spatial Econometrics: Methods and Models*. Kluwer Academic, Boston, MA.

Arellano, M. and S. Bond (1991): Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277-297.

Bai, J. (2003): Inferential theory for factor models of large dimensions. *Econometrica* 71: 135-171.

Bai, J. (2009): Panel data models with interactive fixed effects. *Econometrica* 77: 1229-1279.

Bai, J. (2013): "Likelihood Approach to Dynamic Panel Models with Interactive Effects," *mimeo.*, Columbia University.

Bai, J. and K. Li (2014): "Spatial Panel Data Models with Common Shocks," *mimeo.*, Columbia University.

Bai, J. and K. Li (2015): "Dynamic Spatial Panel Data Models with Common Shocks," *mimeo.*, Columbia University.

Baier, S.L. and J.H. Bergstrand (2007): "Do Free Trade Agreements Actually Increase Members International Trade?" *Journal of International Economics* 71:72-95.

Bailey, N., S. Holly and M.H. Pesaran (2016): "A Two-Stage Approach to Spatio-Temporal Analysis with Strong and Weak Cross-sectional Dependence," forthcoming in *Journal of Applied Econometrics*.

Bailey, N., G. Kapetanios and M.H. Pesaran (2016): "Exponent of Cross-sectional Dependence: Estimation and Inference," forthcoming in *Journal of Applied Econometrics*.

Balazsi, L., M. Bun, F. Chan and M.N. Harris (2016): "Models with Endogenous Regressors," Chapter 3 in *The Econometrics of Multi-dimensional Panels* ed. by L. Mathyas.

Balazsi, L., B.H. Baltagi, L. Mathyas and D. Pus (2016): "Modelling Multidimensional Panel Data: A Random Effects Approach", *mimeo.*, Central European University.

Balazsi, L., L. Mathyas and T. Wansbeek (2015): "The Estimation of Multidimensional Fixed Effects Panel Data Models", forthcoming in *Econometric Reviews*.

Baldwin, R.E. (2006): *In or Out: Does it Matter? An Evidence-Based Analysis of the Euro's Trade Effects*. Centre for Economic Policy Research.

Baldwin, R.E. and D. Taglioni (2006): "Gravity for Dummies and Dummies for Gravity Equations," NBER Working Paper 12516.

Baltagi B.H. (2005): *Econometric Analysis of Panel Data*, 3rd edition. Wiley: Chichester.

Baltagi B.H and G. Bresson (2016): "Modelling Housing Using Multi-dimensional Panel Data," Chapter 12 in *The Econometrics of Multi-dimensional Panels* ed. by L. Mathyas.

Baltagi, B.H., P. Egger, P. and M. Pfaffermayr (2003): "A Generalized Design for Bilateral Trade Flow Models," *Economics Letters* 80: 391-397.

Baltagi, B., P. Egger and M. Pfaffermayr (2008): Estimating regional trade agreement effects on FDI in an interdependent world. *Journal of Econometrics* 145: 194-208.

Baltagi, B.H., P. Egger, P. and M. Pfaffermayr (2015): "Panel Data Gravity Models of International Trade," in *The Oxford Handbook of Panel Data* ed. by B.H. Baltagi.

Behrens, K., C. Ertur andW. Kock (2012): "Dual Gravity: Using Spatial Econometrics to Control For Multilateral Resistance," *Journal of Applied Econometrics* 27: 773-794.

Bertoli, S. and J. Fernandez-Huertas Moraga (2013): "Multilateral Resistance to Migration," *Journal of Development Economics* 102: 79-100.

Breitung, Jorg and Sandra Eickmeier (2016): Analyzing international business and financial cycles using multi-level factor models: A comparison of alternative approaches, in *Dynamic Factor Models*, Chap. 5, pp. 177–214. Emerald Group Publishing Limited.

Chamberlain, G. (1984): Panel Data, in *Handbook of Econometrics*, Volume 2, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, 1247-1318.

Choi I. (2012): Efficient estimation of factor models. *Econometric Theory* 28: 274-308.

Choi, In, Dukpa Kim, Yun Jung Kim, and Noh-Sun Kwark (2017) A Multilevel Factor Model: Identification, Asymptotic Theory and Applications. *Journal of Applied Econometrics*.

Chudik, A., M.H. Pesaran and E. Tosetti (2011): Weak and strong cross section dependence and estimation of large panels. *The Econometrics Journal* 14: C45-C90.

Chudik, A., K. Mohaddes, M.H. Pesaran and M. Raissi (2017): "Is there a debt-thrshold effect on output growth?" *Review of Economics and Statistics* 99: 135-150.

Cliff, A.D. and J.K. Ord (1973): *Spatial Autocorrelation*. Pion, London.

Conley, T.G. and B. Dupor (2003): A spatial analysis of sectoral complementarity. *Journal of Political Economy* 111: 311-352.

Dang, V.A., M. Kim and Y. Shin (2012): Asymmetric Capital Structure Adjustments: New Evidence from Dynamic Panel Threshold Models. *Journal of Empirical Finance* 19: 465-482.

Defever, F., B. Heid and M. Larch (2015): Spatial exporters. *Journal of International Economics* 95: 145-C156.

De Nardis, S. and C. Vicarelli (2003): "Currency Unions and Trade: The Special Case of EMU," *World Review of Economics* 139: 625-49.

Eberhardt, M and A.F. Presbitero (2015): "Public Debt and Growth: Heterogeneity and Non-linearity," *Journal of International Economics* 97: 45-58.

Eberhardt, M., C. Helmers and H. Strauss (2013): Do spillovers matter when estimating private returns to R&D?. *Review of Economics and Statistics* 95: 436-448.

Egger, P. and M. Pfaffermayr (2003): "The Proper Econometric Specification of the Gravity Equation: 3-Way Model with Bilateral Interaction Effects," *Empirical Economics* 28: 571-580.

Elhorst, J.P. (2005): Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis* 37: 85-106.

Elhorst, J.P. (2010): Dynamic panels with endogenous interaction effects when T is small. *Regional Science and Urban Economics* 40: 272-282.

Elliott, M., B. Golub and M.O. Jackson (2014): Financial Networks and Contagion. *American Economic Review* 104: 3115-53.

Fan, J., Y. Liao, Y and M. Mincheva (2011): High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* 39: 3320.

Feenstra, R.C. (2004): *Advanced International Trade*. Princeton University Press: Princeton, NJ.

Forni, M., M. Hallin, M. Lippi and L. Reichlin (2004): The generalized dynamic factor model consistency and rates. *Journal of Econometrics* 119: 231-255.

Frankel, J.A. and A.K. Rose (2002): "An Estimate of the Effect of Common Currencies on Trade and Income," *Quarterly Journal of Economics* 117: 437-466.

Glick R. and A.K. Rose (2002): "Does a Currency Union Affect Trade? The Time Series Evidence," *European Economic Review* 46: 1125-1151.

Gunnella V., C. Mastromarco, L. Serlenga and Y. Shin (2015): "The Euro Effects on Intra-EU Trade Flows and Balances: Evidence from the Cross Sectionally Correlated Panel Gravity Models," *mimeo.*, University of York.

Hausman, J.A. and W.E. Taylor (1981): "Panel Data and Unobservable Individual Effect," *Econometrica* 49: 1377-1398.

Hahn, J. and G. Kuersteiner (2002): Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica* 70: 1639-1657.

Helpman, E. (1987): "Imperfect competition and international trade: evidence from fourteen industrialized countries, " *Journal of the Japanese and International Economies*, 1: 62?81.

Herwartz, H. and H. Weber (2010): "The Euro's Trade Effect under Cross-sectional Heterogeneity and Stochastic Resistance", Kiel Working Paper No. 1631, Kiel Institute for the World Economy, Germany.

Kapetanios, G. and M.H. Pesaran (2005): "Alternative Approaches to Estimation and Inference in Large Multifactor Panels: Small Sample Results with an Application to Modelling of Asset Returns,", CESifo Working Paper Series 1416, CESifo Group Munich.

George Kapetanios, Laura Serlenga, and Yongcheol Shin (2018): Estimation and Inference for Multi-dimensional Heterogeneous Panel Datasets with Hierarchical Multi-factor Error Structure, *mimeo.*, University of York.

Kato, T. (1995): *Perturbation Theory for Linear Operators*. Springer, Berlin.

Kelejian, H.H. and I.R. Prucha (2010): Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances.

Journal of Econometrics 157: 53-67.

Kose M., C. Otrok and C. Whiteman (2003): International business cycles: World, region, and country-specific factors. *American Economic Review* 93: 1216-1239

Kramarz, F., S.J. Machin and A. Ouazad (2008): "What Makes a Test Score? The Respective Contributions of Pupils, Schools, and Peers in Achievement in English Primary Education," INSEAD Working Paper.

Kuersteiner, G.M. and I.R. Prucha (2015): "Dynamic Spatial Panel Models: Networks, Common Shocks, and Sequential Exogeneity," *mimeo.*, University of Maryland.

Krugman, PR. (1997): "Increasing Returns, Monopolistic Competition and International Trade," *Journal of International Economics*, 9: 469-479.

Lee, L.F. (2004): Asymptotic distributions of quasi-maximum likelihood estimator for spatial autoregressive models. *Econometrica* 72: 1899-1925.

Lee, L.f. and J. Yu (2010a): Some recent developments in spatial panel data models. *Regional Science and Urban Economics* 40: 255-271.

Lee, L.f. and J. Yu (2010b): A spatial dynamic panel data model with both time and individual fixed effects. *Econometric Theory* 26: 564-597.

Lee, L.f. and J. Yu (2013): Identification of spatial durbin panel models. Working paper; forthcoming in *Journal of Applied Econometrics*.

Lee, L.f. and J. Yu (2015): Spatial panel data models, chapter 12, in: Baltagi, B.H. (Ed.), *The Oxford Handbook of Panel Data*. Oxford University Press, New York, NY.

Li, J. and Z. Yang (2017): A panel data model with interactive effects characterized by multilevel non-parallel factors. *Applied Economics Letters*.

Lin, X. and L.F. Lee (2010): GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics* 157: 34-52.

Liu, X. and L.F. Lee (2010): GMM estimation of social interaction models with centrality. *Journal of Econometrics* 159: 99-115.

Lu, X. and L. Su (2015): Shrinkage estimation of dynamic panel data models with interactive fixed effects. Working paper. Singapore Management School

Lu, X. and Su, S. (2018): Three-Dimensional Panel Data Models with Factor Structures, *mimeo.*, Hong Kong University of Science and Technology.

Manski, C.F. (1993): Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60: 531-542.

Mastromarco, C., L. Serlenga and Y. Shin (2016): "Modelling Technical Inefficiency in Cross Sectionally Dependent Stochastic Frontier Panels," forthcoming in *Journal of Applied Econometrics* 31: 281-297.

Mastromarco, C., L. Serlenga and Y. Shin, (2015): "Multilateral Resistance and Euro Effects on Trade Flows," forthcoming in *Spatial Econometric Interaction Modelling* eds. by G. Arbia and R. Patuelli. Springer: Berlin.

Matyas, L. (1997): "Proper Econometric Specification of the Gravity Model," *The World Economy* 20: 363-369.

Moon, H. and M. Weidner (2014): Dynamic linear panel regression models with interactive fixed effects. Working paper

Moon, H. and M. Weidner (2015): Linear regression for panel with unknown number of factors as interactive fixed effects. Forthcoming in *Econometrica*.

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

Nauges, C. and A. Thomas (2003). Consistent estimation of dynamic panel data models with time-varying individual effects. *Annales d'Economie et de Statistique* 70: 53-74.

Omay T. and E. Kan (2010): "Re-examining the threshold effects in the inflation–growth nexuswith cross-sectionally dependent non-linear panel: Evidence from six industrialized economies," *Economic Modelling* 27: 996–1005.

Onatski A. (2009): A formal statistical test for the number of factors in the approximate factor models. *Econometrica* 77: 1447-1479

Onatski A. (2010): Determining the number of factors from the empirical distribution of eigenvalues. *Review of Economics and Statistics* 92: 1004-1016

Ord, K. (1975): Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70: 120-297.

Pakko, M.R. and H.J. Wall (2002): "Reconsidering the Trade-creating Effects of a Currency Union," *Federal Reserve Bank of St Louis Review* 83: 37-46.

Pesaran, M.H. (2006): Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74: 967-1012.

Pesaran, M.H. (2015): "Testing Weak Cross-sectional Dependence in Large Panels," *Econometric Reviews* 34: 1089-1117.

Pesaran, M.H. and E. Tosetti (2011): Large panels with common factors and spatial correlation. *Journal of Econometrics* 161: 182-202.

Persson T. (2001): "Currency Union and Trade: How Large is the Treatment Effect?" *Economy Policy* 33: 435-448.

Rodríguez-Caballero, C.V. (2016): Panel Data with Cross-Sectional Dependence Characterized by a Multi-Level Factor Structure, *mimeo.*, CREATES, Aarhus University.

Rose, A.K. (2001): "Currency Unions and Trade: The Effect is Large," *Economic Policy* 33: 449-61.

Serlenga, L. and Y. Shin (2007): "Gravity Models of Intra-EU Trade: Application of the CCEP-HT Estimation in Heterogeneous Panels with Unobserved Common Time-specific Factors," *Journal of Applied Econometrics* 22: 361-381.

Seo, M and Y. Shin (2016): Inference for Dynamic Panels with Threshold Effect and Endogeneity. *Journal of Econometrics* 195: 169-186.

Shi, W. and L.F. Lee (2017): "Spatial Dynamic Panel Data Models with Interactive Fixed Effects," *Journal of Econometrics* 197: 323-347.

Stock J. and M.W. Watson (2010): Dynamic factor models. *Oxford Handbook of Economic Forecasting*. Michael P. Clements and David F. Hendry (eds), Oxford University Press. DOI: 10.1093/oxfordhb/9780195398649.013.0003

Su L., S. Jin S. and Y. Zhang (2015): Specification test for panel data models with interactive fixed effects. *Journal of Econometrics* 186: 222-244.

Su, L. and Z. Yang (2015): QML estimation of dynamic panel data models with spatial errors. *Journal of Econometrics* 185: 230-258.