Estimating Spatial Dynamic Panel Data Models with Unobserved Common Factors in Stata

Sebastian Kripfganz
University of Exeter Business School

Vasilis Sarafidis
Brunel University London
and
BI Norwegian Business School

Abstract

This article introduces the **spxtivdfreg** package in Stata, which implements a general instrumental variables (IV) approach for estimating dynamic spatial panel data models with unobserved common factors or interactive effects, when the number of both cross-sectional and time series observations is large. The estimator has been developed in a recent paper by Cui, Sarafidis, and Yamagata (2023). The underlying idea is to project out the common factors from exogenous covariates using principal components analysis, and to run IV regression in both of two stages, using defactored covariates (and their spatial counterparts) as instruments. The resulting two-stage IV estimator is valid for models with homogeneous slope coefficients, and has several advantages relative to existing popular approaches. In addition, the **spxtivdfreg** package allows estimation of short-run and long-run direct and indirect effects, as well as total effects, accounting for the cumulative effects over time and across space. Standard errors for such effects are computed using the Delta method. Last, the **spxtivdfreg** package allows for heterogeneous slope coefficients, as in Chen, Cui, Sarafidis, and Yamagata (2023). In particular, we construct a "mean group" IV estimator, which involves averaging first-step IV estimates of individual-specific slope coefficients.

Keywords: panel data, longitudinal models, time lags, spatial lags, unobserved common factors, cross-sectional dependence, instrumental variables, heterogeneous coefficients, Stata.

1. Introduction

Panel data (also known as longitudinal data) relate to both space and time, and arise by following the same (N) subjects over multiple (T) time periods. The analysis of such data is important for modeling human behavior across many fields of research. In particular, the temporal dimension allows one to identify how current behavior is influenced by past own behavior. In economics, such dependence over time is due to habit formation, costs of adjustment, and uncertainty, among other factors (Bun and Sarafidis 2015). On the other hand, the spatial dimension allows one to identify the extent to which an economic agent's own behavior is also influenced by the behavior of other agents, typically their peers. Such phenomenon is due to social interactions, network linkages and spillover effects; e.g., see the pioneering work of Case (1991) and Manski (1993). Finally, the combination of space and time allows one to control for richer sources of unobserved heterogeneity compared to cross-sectional or time series data alone. For instance, since agents inhabit a common economic environment, their behavior is often influenced by common economy-wide "shocks" that hit all individual entities, albeit with different intensities. Examples of such common shocks include technological disruptions, natural disasters, financial crises, geopolitical conflicts, global pandemics, and so on. Even if these attributes can be proxied by certain observed variables, their influence on each individual is typically unknown and idiosyncratic, leading to unobserved nonlinear heterogeneity. In econometrics, a prominent methodology for dealing with such heterogeneity structure is the so-called *common-factor* approach (Sarafidis and Wansbeek 2021). This assumes the presence of a linear combination of a finite number of latent (time-specific) variables, called factors, interacted with individual-specific variables that are also unobserved, known as factor loadings.

Most "large-T" panel data methods available for estimating models with the afore-described features are based on quasi-maximum likelihood estimation (QMLE); e.g., see Yu, de Jong, and Lee (2008), Shi and Lee (2017) and Bai and Li (2021).[1] Recently, Cui *et al.* (2023) developed an instrumental-variables (IV) approach, which is appealing both from a computational as well as from a theoretical point of view. To begin with, their approach is linear in the parameters of interest and therefore computationally inexpensive. In addition, unlike QMLE methodologies, their IV estimator can deal with endogenous covariates, where endogeneity arises due to, say, reverse causality or measurement error.

Finally, the IV estimator of Cui *et al.* (2023) is asymptotically unbiased. Therefore, their method does not require any bias correction for asymptotically valid inferences (see Cui, Norkute, Sarafidis, and Yamagata (2022) for more details). In contrast, existing QMLE methodologies are subject to the so-called "incidental-parameters problem". This arises because an increasing number of nuisance parameters (the factors and factor loadings) needs to be estimated as either N or T increases. Unfortunately, approximate bias correction procedures, aiming to re-center the asymptotic distribution of the estimator, may fail to remove all bias terms, particularly those of higher order. Moreover, in practice the number of factors is typically unknown. When the number of factors is overestimated, bias correction can perform poorly. Both of these issues can result in significant size distortions.

The present paper introduces the **spxtivdfreg** package in Stata, which implements the IV approach of Cui *et al.* (2023), and extends it in two major ways. Firstly, the algorithm allows estimation of direct and indirect effects, as well as total effects. Direct effects are those attributed solely to changes in one's own behavior. Indirect effects are those attributed to changes in the behavior of one's peers. Total effects are the sum of the two; e.g., see LeSage and Pace (2009) and Elhorst (2014b). Standard errors for all effects are computed in the **spxtivdfreg** package using the Delta method.[2]

Secondly, the algorithm allows for heterogeneous slope coefficients, as in a recent working paper by Chen *et al.* (2023). In particular, we construct a "mean group" IV estimator, which involves averaging first-step IV estimates of individual-specific slope coefficients.

The class of estimators implemented by the **spxtivdfreg** package is valid under a "large N, large T" framework. In practice, this implies that N and T need to be of comparable magnitude; neither dimension can be considered negligible in comparison to the other. There are several potential applications where the "large N, large T" framework is relevant. Exam-

[1] A notable exception is Chen, Shin, and Zheng (2022), who put forward an IV estimator for static panels (i.e., without any dynamics), based on the "common correlated effects" approach of Pesaran (2006).

[2] Although standard errors for direct/indirect effects can be computed using the method of bootstrap, this is computationally highly intensive, especially for large datasets, making it less practical in the present situation.

ples include (i) the analysis of macroeconomic variables (e.g., gross domestic product (GDP), inflation, unemployment) across a large number of countries over an extended time period, (ii) longitudinal health studies investigating health outcomes for individuals examined over a prolonged timeframe, (iii) the analysis of financial markets, studying the behavior of financial instruments across a large number of firms observed over a long time horizon, and (iv) studies in political science, where political behavior, voter preferences, or policy outcomes are evaluated across a large number of jurisdictions over many election cycles, to mention a few.

Since version 15, Stata is shipped with several packages for the estimation of spatial autoregressive models: **spregress** for generalized spatial two-stage least squares and maximum likelihood estimation, **spivregress** for IV estimation with endogenous regressors, and **spxtregress** for fixed-effects and random-effects panel data estimation. They are accompanied by tools for spatial data preparation, choropleth maps graphing, and spatial-weights matrix manipulation. These official packages heavily build on earlier community contributions – most notably the **spreg**, **spivreg**, and **spmat** packages (Drukker, Peng, Prucha, and Raciborski 2013a; Drukker, Prucha, and Raciborski 2013b,c), **spmap** (Pisati 2007), as well as a collection of further tools for spatial data analysis discussed by Pisati (2001). When it comes to estimating dynamic spatial panel models, only the community-contributed package **xsmle**, developed and described in Belotti, Hughes, and Mortari (2017), is available in Stata, but without capabilities for instrumental variables.

Spatial econometric models can also be estimated with other statistical software. In MATLAB, the econometrics toolbox, documented in LeSage and Pace (2009), provides an extensive set of functions for the estimation of conventional spatial models, with a particular focus on Bayesian estimation techniques. Elhorst (2014a) provides functions for spatial panel models. The **spreg** package within the Python library **PySal** (Rey and Anselin 2010; Rey, Anselin, Amaral, Arribas-Bel, Cortes, Gaboardi, Kang, Knaap, Li, Lumnitz, Oshan, Shao, and Wolf 2022) has similar capabilities as the official Stata commands. The library contains further packages that can assist with the analysis and visualization of spatial data, and the implementation of generalized regression techniques. **spreg** evolved from the self-contained GeoDa software (Anselin, Li, and Koschinsky 2006). The latter's recent implementation as a library, **libgeoda**, enables integration into other software environments (Anselin, Li, and Koschinsky 2022).

In R, the **splm** package by Millo and Piras (2012) implements maximum likelihood and generalized method of moments estimators for static panel models, together with related test statistics. The packages **spdep** (Bivand 2023) and **spatialreg** (Bivand and Piras 2023) provide a complementary collection of functions for modeling and analyzing spatial dependence in a cross-sectional context. A large number of further R packages provides additional specialized functionality for spatial data analysis. For an overview of software for spatial econometrics and statistics, see Pebesma, Bivand, and Ribeiro (2015), Bivand and Piras (2015), and Bivand (2022).

Crucially, none of the existing packages is able to combine a spatial model with the common-factor approach. Spatial models and common-factor models capture different aspects of dependence across individuals. Overlooking either aspect may result in misinterpretations of the underlying dependence structure. As an example, to the extent that spatial interactions are driven by a few dominant units in the population, the common-factor approach may be more appropriate than a spatial model.

Moreover, all of the existing packages impose that the slope coefficients (including the spatial parameter) are homogeneous across all individuals. In practice, however, the strength of spatial interactions can vary across individuals, depending on unobserved individual-specific attributes. Therefore, the restriction of slope parameter homogeneity can pose a major limitation. In comparison, the Stata package **spxtivdfreg** allows not only for common factors in the residuals, but also for potential heterogeneity in the slope coefficients. Thus, it permits a richer formulation and estimation of spatial panel models.

A further advantage of **spxtivdfreg** is that it is computationally easy to implement compared to existing methods. In particular, both **spxtregress** and **xsmle** require prior knowledge of Stata's programming tools for the creation and management of spatial weights matrices, such as **spmatrix** or **spmat** (Drukker *et al.* 2013a). While the aforementioned spatial tools can be used with **spxtivdfreg** as well, our command also allows generating the spatial weights matrix in an Excel file or in a delimited text file. Thus, no priori knowledge of Stata's spatial tools is required, which implies zero "sunk cost" for empirical practitioners. The computational simplicity of the **spxtivdfreg** package is enhanced by the fact that the IV estimator of Cui *et al.* (2023) requires no bias correction to deal with incidental parameters.

The **spxtivdfreg** package is a generalization of the existing community-contributed command **xtivdfreg**, developed by Kripfganz and Sarafidis (2021). The latter does not allow for spatial variables in the model, nor for the computation of long-run marginal effects. The extension to spatial models is far from trivial. For instance, as a standard feature, Stata estimation commands enable predictions of the dependent variable. However, the naive linear prediction ignores the dynamic nature of the spatial linkages. Instead, predictions of the reduced-form mean are required. After solving for the endogenous spatial spillovers, this conditional expectation is nonlinear in the parameters. Similarly, complications also arise in the computation of (short- and long-run) direct, indirect, and total effects, and especially their standard errors. Given the novelty of this IV approach, it's important to highlight that it has yet to be implemented in any other statistical software.

The remainder of the paper is organized as follows. Section 2 outlines the methods developed by Cui *et al.* (2023) and Chen *et al.* (2023). Section 3 introduces the **spxtivdfreg** in Stata. An application illustrating the use of the package is provided in Section 4. Section 5 concludes.

2. Spatial models

2.1. Homogeneous spatial model

We consider the following spatial panel data model with N cross-sectional units and T time periods:

$$y_{it} = \psi \sum_{j=1}^{N} w_{ij} y_{jt} + \rho y_{it-1} + \mathbf{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{\gamma}_{y,i}^\top \mathbf{f}_{y,t} + \varepsilon_{it}, \tag{1}$$

$i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, where y_{it} denotes the observation on the dependent variable for individual unit i at time period t, and $\mathbf{x}_{it}$ is a $k \times 1$ vector of covariates with slope coefficients $\boldsymbol{\beta}$. The lagged dependent variable y_{it-1} captures dynamic or temporal effects due to state dependence. The error term of the model is composite: $\mathbf{f}_{y,t}$ and $\boldsymbol{\gamma}_{y,i}$ denote $r_y \times 1$ vectors of latent factors and factor loadings respectively, and ε_{it} is an idiosyncratic error. We note that the lagged dependent variable is endogenous w.r.t. the common-factor component, unless one is willing to impose highly restrictive assumptions on the time series properties of $\mathbf{f}_{y,t}$.

The "spatial-lag" variable $\sum_{j=1}^{N} w_{ij} y_{jt}$ is a weighted average of the outcome variable in the neighboring locations for individual i. w_{ij} denotes the (i,j)th element of the $N \times N$ spatial weights matrix $\mathbf{W}_N$. w_{ij} is inversely related to the distance between units i and j. $\mathbf{W}_N$ is assumed to be known. The spatial-lag variable is endogenous by construction, as it essentially represents the formal specification for the equilibrium outcome of a spatial interaction process, in which the value of y_{it} for one agent is simultaneously determined with that of neighboring agents (Elhorst 2014b). Such spatial interlinkages among agents may be attributed to peer effects, spillovers or contagion, as well as strategic interactions, to mention a few.

It is common practice to normalize the spatial weights matrix prior to estimation such that its largest eigenvalue equals 1. This implies that the spatial-lag coefficient ψ is bounded by 1 as well and its magnitude can be interpreted similarly to the time-lag coefficient ρ (LeSage and Pace 2009; Elhorst 2014b). An innocuous way of achieving this is the spectral standardization, which simply re-scales all elements of the spatial weights matrix by the same proportional factor. Another common practice in empirical studies is row standardization, wherein each element of the spatial weights matrix is divided by the sum of its respective row. The strength of a network link is then measured relative to the sum of all links a unit has with other units. It must be noted that a row standardization generally alters the implied network structure. For example, a spatial weights matrix that was symmetric before row standardization may no longer be so afterwards. As emphasized by Kelejian and Prucha (2010), this can lead to model misspecification if row standardization is not justified on economic grounds.[3]

As an example of the model in Equation 1, in Section 4 we analyse panel data on a sample of US banking institutions, each one observed over 35 consecutive quarters. The dependent variable captures a measure of credit risk and the vector $\mathbf{x}_{it}$ contains a number of risk-taking determinants, such as bank size, profitability, asset quality and liquidity. In this case, the lagged dependent variable, y_{it-1}, may absorb idiosyncratic risk vulnerabilities that build up over time, whereas the spatial variable $\sum_{j=1}^{N} w_{ij} y_{jt}$ captures endogenous spillover effects that may arise due to multiple balance sheet interdependencies among financial institutions. The common factors can absorb (among other things) changes in the regulatory framework within the banking industry during the sample period, as well as market risks and economic conditions, such as interest rate volatility, equity and currency risks, business cycle fluctuations, to mention a few. These factors may hit the population of all banks simultaneously, albeit with different intensities, depending on individual bank characteristics, such as the quality of corporate governance and other sources of latent risk exposure.

To ensure that the covariates are correlated with the factor component (a third source of endogeneity in the model, in addition to endogeneity in y_{it-1} and $w_{ij} y_{jt}$), we impose the following reduced-form data-generating process for $\mathbf{x}_{it}$:

$$\mathbf{x}_{it} = \mathbf{\Gamma}_{x,i}^{\top} \mathbf{f}_{x,t} + \mathbf{v}_{it}, \tag{2}$$

where $\mathbf{f}_{x,t}$ denotes an $r_x \times 1$ vector of latent factors, $\mathbf{\Gamma}_{x,i}$ denotes an $r_x \times k$ factor loading matrix, while $\mathbf{v}_{it}$ is an idiosyncratic disturbance of dimension $k \times 1$.[4] Note that $\mathbf{f}_{y,t}$ and $\mathbf{f}_{x,t}$

[3] Because the standardization is a user's modeling choice that is not necessary for estimation, our `spxtivdfreg` `Stata` command neither offers any options nor carries out any checks in this regard. It only checks whether the dimensions of $\mathbf{W}_N$ are $N \times N$ and the main-diagonal elements are all zero. `Stata`'s `spmatrix` command can be used to normalize the spatial weights matrix before feeding it into `spxtivdfreg`.

[4] The above linear factor structure can be viewed as restrictive at first glance. However, as argued by

can be identical, share some common factors, or they can be completely different but mutually correlated. Similarly, $\boldsymbol{\gamma}_{y,i}$ and $\boldsymbol{\Gamma}_{x,i}$ can be mutually correlated. This way, $\mathbf{x}_{it}$ is endogenous w.r.t. the common-factor component of the error. On the other hand, to simplify exposition, we shall assume that $\mathbf{x}_{it}$ is strictly exogenous w.r.t. the purely idiosyncratic error, ε_{it}. This allows the formulation of internal instruments based on the existing model covariates. The case where strict exogeneity w.r.t. ε_{it} is violated, is discussed in Section 2.4.

Stacking the T observations for each i yields

$$\begin{aligned} \mathbf{y}_i &= \psi \mathbf{Y}\mathbf{w}_i + \rho \mathbf{y}_{i,-1} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{F}\boldsymbol{\gamma}_{y,i} + \boldsymbol{\varepsilon}_i; \\ \mathbf{X}_i &= \mathbf{F}_x \boldsymbol{\Gamma}_{x,i} + \mathbf{V}_i, \end{aligned} \tag{3}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})^\top$, $\mathbf{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})^\top$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iT})^\top$ denote $T \times 1$ vectors, $\mathbf{X}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT})^\top$ and $\mathbf{V}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT})^\top$ are matrices of order $T \times k$, while $\mathbf{F}_y = (\mathbf{f}_{y,1}, \cdots, \mathbf{f}_{y,T})^\top$ and $\mathbf{F}_x = (\mathbf{f}_{x,1}, \cdots, \mathbf{f}_{x,T})^\top$ are of dimensions $T \times r_y$ and $T \times r_x$, respectively. Finally, $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)^\top$ denotes a $T \times N$, matrix and the $N \times 1$ vector $\mathbf{w}_i$ represents the ith row of $\mathbf{W}_N$. More succinctly, the model in Equation 3 can be written as

$$\mathbf{y}_i = \mathbf{C}_i \boldsymbol{\theta} + \mathbf{u}_i, \tag{4}$$

where $\mathbf{C}_i = (\mathbf{Y}\mathbf{w}_i, \mathbf{y}_{i,-1}, \mathbf{X}_i)$, $\boldsymbol{\theta} = \left(\psi, \rho, \boldsymbol{\beta}^\top\right)^\top$ and $\mathbf{u}_i = \mathbf{F}_y \boldsymbol{\gamma}_{y,i} + \boldsymbol{\varepsilon}_i$.

The IV approach of Cui *et al.* (2023) involves two stages. In the first stage, the common factors in $\mathbf{X}_i$ are asymptotically projected out using principal components analysis, as in Bai (2003). Subsequently, the resulting, "defactored" covariates are used as instruments to obtain consistent estimates of the model parameters, $\boldsymbol{\theta}$. In the second stage, the factors entering $\mathbf{u}_i$ are projected out from the model, based on the first-stage IV residuals; next, a second IV regression is implemented using the same instruments as in stage one.

In particular, define $\mathbf{X}_{i,-\tau} \equiv L^\tau \mathbf{X}_i$, where L^τ denotes the time series lag operator of order τ. We shall make use of the convention $\mathbf{X}_{i,-0} = \mathbf{X}_i$. Moreover, let $\widehat{\mathbf{F}}_{x,-\tau}$ be defined as $\sqrt{T}$ times the eigenvectors corresponding to the r_x largest eigenvalues of the $T \times T$ matrices $(NT)^{-1} \sum_{i=1}^N \mathbf{X}_{i,-\tau} \mathbf{X}_{i,-\tau}^\top$, for $\tau = 0, 1$. The matrices that project out $\widehat{\mathbf{F}}_x$ and $\widehat{\mathbf{F}}_{x,-1}$ from $\mathbf{X}_i$ and $\mathbf{X}_{i,-1}$ respectively, are given by

$$\mathbf{M}_{\widehat{\mathbf{F}}_x} = \mathbf{I}_T - \widehat{\mathbf{F}}_x \left(\widehat{\mathbf{F}}_x^\top \widehat{\mathbf{F}}_x\right)^{-1} \widehat{\mathbf{F}}_x^\top; \quad \text{and} \quad \mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}} = \mathbf{I}_T - \widehat{\mathbf{F}}_{x,-1} \left(\widehat{\mathbf{F}}_{x,-1}^\top \widehat{\mathbf{F}}_{x,-1}\right)^{-1} \widehat{\mathbf{F}}_{x,-1}^\top. \tag{5}$$

The matrix of instruments is formulated as follows:

$$\widehat{\mathbf{Z}}_i = \left(\sum_{j=1}^N w_{ij} \mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_j, \quad \mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}} \mathbf{X}_{i,-1}, \quad \mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_i \right), \tag{6}$$

which is of dimension $T \times 3k$. Loosely speaking, the term $\sum_{j=1}^N w_{ij} \mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_j$ instruments $\mathbf{Y}\mathbf{w}_i$, the term $\mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}} \mathbf{X}_{i,-1}$ instruments $\mathbf{y}_{i,-1}$, and $\mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_i$ instruments $\mathbf{X}_i$. All these three terms are asymptotically uncorrelated with the common-factor component and therefore, under strict exogeneity w.r.t. $\boldsymbol{\varepsilon}_i$, they constitute valid instruments.

Freeman and Weidner (2023), such structure can also approximate nonlinear functions, by letting the number of estimated factors increase. For example a quadratic form in $\mathbf{f}_{x,t}$ can be dealt with using r_x additional factors, $\mathbf{f}_{x,t}^+$, where in the true DGP $\mathbf{f}_{x,t}^+ = \mathbf{f}_{x,t}^2$.

The first-stage IV estimator of $\boldsymbol{\theta}$ is defined as:

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\mathbf{A}}^\top \widehat{\mathbf{B}}^{-1} \widehat{\mathbf{A}}\right)^{-1} \widehat{\mathbf{A}}^\top \widehat{\mathbf{B}}^{-1} \widehat{\mathbf{c}}_y, \tag{7}$$

where

$$\widehat{\mathbf{A}} = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \mathbf{C}_i; \quad \widehat{\mathbf{B}} = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \widehat{\mathbf{Z}}_i; \quad \widehat{\mathbf{c}}_y = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \mathbf{y}_i\,.$$

Under certain regularity conditions, although $\widehat{\boldsymbol{\theta}}$ is $\sqrt{NT}$ consistent, the underlying limiting distribution of the estimator will not be centered around the true value; that is, asymptotic biases exist (see Cui *et al.* (2023) for more details). Instead of attempting to correct the asymptotic bias of this estimator, the authors put forward a second-stage estimator, which is free from asymptotic bias and is potentially more efficient.[5]

To implement the second stage, the space spanned by $\mathbf{F}_y$ is estimated from the first-stage IV residuals; i.e., $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{C}_i\hat{\boldsymbol{\theta}}$. Subsequently, $\mathbf{F}_y$ is asymptotically eliminated by pre-multiplying the data by the following projection matrix:

$$\mathbf{M}_{\widehat{\mathbf{F}}_y} = \mathbf{I}_T - \widehat{\mathbf{F}}_y \left(\widehat{\mathbf{F}}_y^\top \widehat{\mathbf{F}}_y\right)^{-1} \widehat{\mathbf{F}}_y^\top, \tag{8}$$

where $\widehat{\mathbf{F}}_y$ is defined as $\sqrt{T}$ times the eigenvectors corresponding to the r_y largest eigenvalues of the $T \times T$ matrices $(NT)^{-1}\sum_{i=1}^{N} \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i^\top$.

The second-stage IV (2SIV) estimator for $\boldsymbol{\theta}$ is defined as follows:

$$\widetilde{\boldsymbol{\theta}} = (\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{B}}^{-1} \widetilde{\mathbf{A}})^{-1} \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{B}}^{-1} \widetilde{\mathbf{c}}_y, \tag{9}$$

where

$$\widetilde{\mathbf{A}} = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \mathbf{M}_{\widehat{\mathbf{F}}_y} \mathbf{C}_i; \quad \widetilde{\mathbf{B}} = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \mathbf{M}_{\widehat{\mathbf{F}}_y} \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i^\top \mathbf{M}_{\widehat{\mathbf{F}}_y} \widehat{\mathbf{Z}}_i; \quad \widetilde{\mathbf{c}}_y = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \mathbf{M}_{\widehat{\mathbf{F}}_y} \mathbf{y}_i\,.$$

Theorem 3.2 in Cui *et al.* (2023) shows that, as $N, T \to \infty$ such that $N/T \to c$ where $0 < c < \infty$, not only the 2SIV estimator is consistent and asymptotically normally distributed with variance-covariance matrix $\boldsymbol{\Psi}$, it is also correctly centered around the true value of the parameter vector $\boldsymbol{\theta}$.[6]

In order to allow for heteroskedasticity, the following variance estimator is recommended:

$$\widetilde{\boldsymbol{\Psi}} = \left(\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{B}}^{-1} \widetilde{\mathbf{A}}\right)^{-1} \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{B}}^{-1} \widehat{\boldsymbol{\Omega}} \widetilde{\mathbf{B}}^{-1} \widetilde{\mathbf{A}} \left(\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{B}}^{-1} \widetilde{\mathbf{A}}\right)^{-1}, \tag{10}$$

where

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{NT}\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i^\top \mathbf{M}_{\widehat{\mathbf{F}}_y} \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i^\top \mathbf{M}_{\widehat{\mathbf{F}}_y} \widehat{\mathbf{Z}}_i. \tag{11}$$

Standard errors for $\widetilde{\boldsymbol{\theta}}$ are directly available, based on the square root of the diagonal entries of the above variance estimator.

[5]The only other estimator available in the literature that is free from asymptotic bias due to incidental parameters is the GMM estimator of Juodis and Sarafidis (2022a). However, this estimator does not allow for spatial lags.

[6]An intuition of this result is provided in Cui *et al.* (2022).

A particularly useful diagnostic in this framework is the so-called overidentifying restrictions (J) test statistic.[7] In the present context, the J statistic is given by

$$J = \frac{1}{NT}\left(\sum_{i=1}^{N}\widetilde{\mathbf{u}}_i^{\top}\mathbf{M}_{\widehat{\mathbf{F}}_y}\widehat{\mathbf{Z}}_i\right)\widehat{\boldsymbol{\Omega}}^{-1}\left(\sum_{i=1}^{N}\widehat{\mathbf{Z}}_i^{\top}\mathbf{M}_{\widehat{\mathbf{F}}_y}\widetilde{\mathbf{u}}_i\right), \tag{12}$$

where $\widetilde{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{C}_i\tilde{\boldsymbol{\theta}}$. The null hypothesis postulates that the moment conditions are valid, i.e., $E\left(\sum_{i=1}^{N}\mathbf{Z}_i^{\top}\mathbf{M}_{\mathbf{F}_y}\mathbf{u}_i\right) = \mathbf{0}$, where

$$\mathbf{Z}_i = \left(\sum_{j=1}^{N} w_{ij}\mathbf{M}_{\mathbf{F}_x}\mathbf{X}_j, \quad \mathbf{M}_{\mathbf{F}_{x,-1}}\mathbf{X}_{i,-1}, \quad \mathbf{M}_{\mathbf{F}_x}\mathbf{X}_i\right), \tag{13}$$

and $\mathbf{M}_{\mathbf{F}_y} = \mathbf{I}_T - \mathbf{F}_y\left(\mathbf{F}_y^{\top}\mathbf{F}_y\right)^{-1}\mathbf{F}_y^{\top}$. Under this null hypothesis, the J statistic is asymptotically χ^2 distributed with $m-g$ degrees of freedom, where m denotes the number of instruments used (i.e., the number of columns in $\widehat{\mathbf{Z}}_i$), and g equals the total number of slope parameters estimated in the model. For the standard specification analyzed above, $m = 3k$ and $g = k+2$. The J statistic can be used to test whether e.g., the covariates are strictly exogenous w.r.t. ε_{it}, or whether the model parameters are indeed homogeneous across i. Violation of either of these assumptions invalidates the moment conditions of the model.

2.2. Decomposition of direct and indirect effects

When an explanatory variable in a particular cross-sectional unit changes, not only will the dependent variable of that unit itself change, but the dependent variables of its neighbouring units may also change, depending on the extent of their interaction. The first feature is known in the spatial literature as a "direct effect" and the second one as an "indirect effect".[8]

To facilitate exposition, we shall reconsider the original model in Equation 1 and stack the N observations for each t as follows:

$$\begin{aligned}\mathbf{y}_{(t)} &= \rho\mathbf{y}_{(t-1)} + \psi\mathbf{W}_N\mathbf{y}_{(t)} + \sum_{\ell=1}^{k}\beta_\ell\mathbf{x}_{\ell(t)} + \mathbf{u}_{(t)};\\ \mathbf{u}_{(t)} &= \boldsymbol{\Gamma}_y\mathbf{f}_{y,t} + \boldsymbol{\varepsilon}_{(t)},\end{aligned} \tag{14}$$

where $\mathbf{y}_{(t)}$ is of dimension $N\times 1$, and similarly for the other variables. $\boldsymbol{\Gamma}_y = \left(\boldsymbol{\gamma}_{y,1}, \ldots, \boldsymbol{\gamma}_{y,N}\right)^{\top}$, denotes an $N \times r_y$ matrix of factor loadings.

Solving the model above yields

$$\mathbf{y}_{(t)} = \left[(1-\rho L)\mathbf{I}_N - \psi\mathbf{W}_N\right]^{-1}\left(\sum_{\ell=1}^{k}\beta_\ell\mathbf{x}_{\ell(t)}\right) + \left[(1-\rho L)\mathbf{I}_N - \psi\mathbf{W}_N\right]^{-1}\mathbf{u}_{(t)}. \tag{15}$$

The matrix of partial derivatives of the expected value of $\mathbf{y}$ with respect to the ℓth covariate in the *long-run* is given by:

$$\left[\frac{\partial E(\mathbf{y})}{\partial x_{\ell 1}} \cdots \frac{\partial E(\mathbf{y})}{\partial x_{\ell N}}\right] = \left[(1-\rho)\mathbf{I}_N - \psi\mathbf{W}_N\right]^{-1}\beta_\ell. \tag{16}$$

[7]See Juodis and Sarafidis (2022b) for details regarding the usefulness of the J test statistic in model specification involving latent common factors.

[8]See Elhorst (2012) for further details.

The equivalent expression for the *short-run* is obtained by setting $\rho = 0$ above. Individual-specific direct effects of x_ℓ are given by the main-diagonal elements of that matrix. The off-diagonal elements provide directional pair-specific indirect effects. Following LeSage and Pace (2009) and Debarsy, Ertur, and LeSage (2012), it is conventional to report an average of both direct and indirect effects. The average total effect is then the sum of the average direct and indirect effects. Standard errors for all of these effects are obtained with the Delta method from the variance estimator in Equation 10.

For the respective effects to be meaningful, the model parameters need to obey a restriction for dynamic stability. Let ω be the largest eigenvalue of $\mathbf{W}_N$. For short-run stability, the spatial lag coefficient needs to obey $\psi < 1/\omega$. For stability in the long-run, $\rho/(1-\psi\omega) < 1$ is required. When the spatial weights matrix is suitably normalized such that the largest eigenvalue equals one, this conveniently restricts $\psi < 1$ in the short-run – similar to the usual requirement on ρ for a stable time series autoregression – and $\rho + \psi < 1$ in the long-run. Our Stata package **spxtivdfreg** still estimates the regression coefficients when those conditions are violated, but it no longer reports direct, indirect, and total effects.

2.3. Heterogeneous spatial model

We now extend the original model in Equation 4, allowing for heterogeneous slope coefficients. In particular, the model in compact form is now written as

$$\mathbf{y}_i = \mathbf{C}_i \boldsymbol{\theta}_i + \mathbf{u}_i, \tag{17}$$

where $\boldsymbol{\theta}_i = \left(\psi_i, \rho_i, \boldsymbol{\beta}_i^\top\right)^\top$. We shall employ a random-coefficients assumption, such that $\boldsymbol{\theta}_i = \boldsymbol{\theta} + \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ denotes a random error that is independently and identically distributed across i with mean zero and variance $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, such that $E(\boldsymbol{\theta}_i) = \boldsymbol{\theta}$.

As before, we use the IV method in order to estimate $\boldsymbol{\theta}_i$. Specifically, the matrix of instruments is formulated as follows:

$$\widehat{\mathbf{Z}}_i = \left(\sum_{j=1}^{N} w_{ij} \mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_j, \quad \mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}} \mathbf{X}_{i,-1}, \quad \mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_i \right), \tag{18}$$

which remains of dimension $T \times 3k$. The resulting IV estimator for $\boldsymbol{\theta}_i$ is given by

$$\widehat{\boldsymbol{\theta}}_i = \left(\widehat{\mathbf{A}}_i^\top \widehat{\mathbf{B}}_i^{-1} \widehat{\mathbf{A}}_i \right)^{-1} \widehat{\mathbf{A}}_i^\top \widehat{\mathbf{B}}_i^{-1} \widehat{\mathbf{c}}_{y,i}, \tag{19}$$

where

$$\widehat{\mathbf{A}}_i = T^{-1} \widehat{\mathbf{Z}}_i^\top \mathbf{C}_i, \quad \widehat{\mathbf{B}}_i = T^{-1} \widehat{\mathbf{Z}}_i^\top \widehat{\mathbf{Z}}_i, \quad \widehat{\mathbf{c}}_{y,i} = T^{-1} \widehat{\mathbf{Z}}_i^\top \mathbf{y}_i. \tag{20}$$

Under certain regularity conditions, Theorem 3.1 in Chen *et al.* (2023) shows that as $N, T \to \infty$ with $T/N^2 \to 0$, then $\widehat{\boldsymbol{\theta}}_i$ is consistent and asymptotically normally distributed. Once the individual-specific estimates of $\boldsymbol{\theta}_i$ are obtained, the mean-group estimator of the average of $\boldsymbol{\theta}_i$ is constructed as follows:

$$\widehat{\boldsymbol{\theta}}_{MG} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\boldsymbol{\theta}}_i. \tag{21}$$

In Theorem 3.2 of Chen *et al.* (2023), it is demonstrated that $\widehat{\boldsymbol{\theta}}_{MG}$ is consistent for the population mean $\boldsymbol{\theta}$, for $N, T \to \infty$ with $T/N^2 \to 0$. Furthermore, if $N/T^{6/5} \to 0$, then the mean-group estimator has the following asymptotic distribution:

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_{MG} - \boldsymbol{\theta}) \xrightarrow{d} N\left(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right), \quad \text{as } N, T \to \infty.$$

Standard errors for $\widehat{\boldsymbol{\theta}}_{MG}$ are obtained as the square root of the diagonal elements of the variance estimator

$$\widehat{\Sigma}_{\boldsymbol{\eta}} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{MG}\right) \left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{MG}\right)^{\top}. \tag{22}$$

2.4. Further generalizations

Under the same theoretical setting and derivations in Cui *et al.* (2023), the **spxtivdfreg** package in Stata can also allow for spatial-time lags, spatial lags of the covariates, or further autoregressive terms. That is, the original model in Equation 1 can be extended as follows:

$$y_{it} = \sum_{\tau=0}^{S} \psi_\tau \sum_{j=1}^{N} w_{ij} y_{jt-\tau} + \sum_{\tau=1}^{S} \rho_\tau y_{it-\tau} + \mathbf{x}_{it}^{\top} \boldsymbol{\beta} + \sum_{j=1}^{N} w_{ij} \mathbf{x}_{jt}^{\top} \boldsymbol{\delta} + \boldsymbol{\gamma}_{y,i}^{\top} \mathbf{f}_{y,t} + \varepsilon_{it}, \tag{23}$$

$i = 1, 2, \ldots, N$, $t = 2 - S, \ldots, T$.[9]

The model above can be viewed as a dynamic spatial Durbin model with a generalized autoregressive structure. The special case where $\rho_\tau = 0$ for $\tau > 1$ is discussed in detail by Elhorst (2012). The term $\sum_{j=1}^{N} w_{ij} \mathbf{x}_{jt}$ implies the presence of "exogenous network effects" or "contextual effects", as discussed in Manski (1993).

In vector form, the model above can be written as

$$\mathbf{y}_i = \sum_{\tau=0}^{S} \psi_\tau \mathbf{Y}_{-\tau} \mathbf{w}_i + \sum_{\tau=1}^{S} \rho_\tau \mathbf{y}_{i,-\tau} + \mathbf{X}_i \boldsymbol{\beta} + \sum_{\ell=1}^{k} \mathbf{X}_{(\ell)} \mathbf{w}_i \delta_{(\ell)} + \mathbf{F} \boldsymbol{\gamma}_{y,i} + \boldsymbol{\varepsilon}_i, \tag{24}$$

where $\mathbf{Y}_{-\tau} = (\mathbf{y}_{1-\tau}, \cdots, \mathbf{y}_{T-\tau})^{\top}$ denotes a $T \times N$ matrix, and $\mathbf{X}_{(\ell)}$ collects the observations for the ℓth covariate in a $T \times N$ matrix. Equation 16 for the long-run impacts becomes

$$\left[\frac{\partial E\left(\mathbf{y}\right)}{\partial x_{\ell 1}} \cdots \frac{\partial E\left(\mathbf{y}\right)}{\partial x_{\ell N}}\right] = \left[\left(1 - \sum_{\tau=1}^{S} \rho_\tau\right) \mathbf{I}_N - \sum_{\tau=0}^{S} \psi_\tau \mathbf{W}_N\right]^{-1} \beta_\ell. \tag{25}$$

The short-run impacts are computed as before by setting $\rho_\tau = \psi_\tau = 0$ for all $\tau > 0$.

Additional instruments for (i) $\mathbf{Y}_{-1}\mathbf{w}_i$; (ii) $\mathbf{y}_{i,-\tau}$; and (iii) $\mathbf{X}_{(\ell)}\mathbf{w}_i$ can be employed in a straightforward manner. For instance, $\mathbf{Y}_{-1}\mathbf{w}_i$ can be instrumented by $\sum_{j=1}^{N} w_{ij} \mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}} \mathbf{X}_{j-1}$, whereas $\mathbf{y}_{i,-\tau}$ can be instrumented by $\mathbf{M}_{\widehat{\mathbf{F}}_{x,-\tau}} \mathbf{X}_{i,-\tau}$. Lastly, $\mathbf{X}_{(\ell)}\mathbf{w}_i$ can be instrumented either by $\mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_{(\ell)} \widetilde{\mathbf{w}}_i$ or by further lags of $\mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_{(\ell)} \mathbf{w}_i$, where $\widetilde{\mathbf{w}}_i$ denotes the ith row of $\mathbf{W}_N \mathbf{W}_N$. When $\mathbf{W}_N$ is sparse, the choice of $\mathbf{M}_{\widehat{\mathbf{F}}_x} \mathbf{X}_{(\ell)} \widetilde{\mathbf{w}}_i$ amounts to using as instruments those (defactored) covariates corresponding to the "2nd-order" neighbours of individual i; i.e., the neighbours of the neighbours of unit i.

Finally, so far we have assumed that $\mathbf{X}_i$ is exogenous w.r.t. the purely idiosyncratic error, $\boldsymbol{\varepsilon}_i$. In practice however, this assumption can often be violated, due to (say) reverse causality

[9] The total number of spatial-time lags and autoregressive lags can be different. Here we set both equal to S only for ease of exposition.

or measurement error. Unlike QMLE procedures, the present method can accommodate such sources of endogeneity, so long as external instruments are available. To see this, let $\mathbf{X}_i = (\mathbf{X}_i^{(exog)}, \mathbf{X}_i^{(endog)})$, where $\mathbf{X}_i^{(exog)}$ and $\mathbf{X}_i^{(endog)}$ refer to those sets of regressors that are strictly exogenous and endogenous, respectively, w.r.t. $\boldsymbol{\varepsilon}_i$. Note that $\mathbf{X}_i^{(exog)}$ and $\mathbf{X}_i^{(endog)}$ are of dimension $T \times k^{(exog)}$ and $T \times k^{(endog)}$, respectively. Furthermore, let $\mathbf{X}_i^{+} = (\mathbf{X}_i^{(exog)}, \mathbf{X}_i^{(ext)})$, a $T \times k^{+}$ matrix with $k^{+} = k^{(exog)} + k^{(ext)}$, where $\mathbf{X}_i^{(ext)}$ denotes the matrix of external exogenous covariates. We note that $\mathbf{X}_i^{(ext)}$ can still be correlated with the factor component, i.e., it may be subject to a similar data generating process as in Equation 2, so long as it remains strictly exogenous w.r.t. $\boldsymbol{\varepsilon}_i$. Define $\widehat{\mathbf{F}}_x^{+}$ as $\sqrt{T}$ times the eigenvectors corresponding to the m_x^{+} largest eigenvalues of the $T \times T$ matrix $\sum_{i=1}^{N} \mathbf{X}_i^{+} \left(\mathbf{X}_i^{+}\right)^{\top} / NT$. The corresponding projection matrices are defined in the same way as in Equation 8 mutatis mutandis. In this case, the matrix of instruments becomes

$$\widehat{\mathbf{Z}}_i = \left(\sum_{j=1}^{N} w_{ij} \mathbf{M}_{\widehat{\mathbf{F}}_x^{+}} \mathbf{X}_j^{+}, \quad \mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}^{+}} \mathbf{X}_{i,-1}^{+}, \quad \mathbf{M}_{\widehat{\mathbf{F}}_x^{+}} \mathbf{X}_i^{+} \right), \tag{26}$$

In conclusion, the **spxtivdfreg** package can deal with all four types of endogeneity discussed above in reference to the model in Equation 1, namely (i) endogeneity due to the presence of a spatial lag ($\psi \neq 0$); (ii) endogeneity due to the presence of a lagged dependent variable ($\rho \neq 0$); (iii) endogeneity due to potential correlations between the covariates and the common-factor component; and (iv) endogeneity due to potential correlations between the covariates and the idiosyncratic error, as in the case of reverse causality.

3. The spxtivdfreg package

3.1. Syntax

The general syntax of **spxtivdfreg** is similar to standard estimation commands in Stata:

```
spxtivdfreg depvar indepvars, options
```

where `depvar` and `indepvars` are to be replaced by the respective names of the dependent and independent variables, and `options` can be selected from the list below. All variable lists allow Stata's factor variable notation for indicator variables and interaction terms, provided that the package **ftools** (Correia 2016) is installed, as well as time series operators. To restrict the estimation sample, Stata's standard `if` or `in` qualifiers can be used following the list of independent variables. We note that **spxtivdfreg** requires that the panel dataset is balanced. Otherwise, the connectedness structure may change over time.

3.2. Options

A key option is `spmatrix()`, which locates the spatial weights matrix. This option is compulsory. It can be specified in one of following ways:

- `spmatrix(filename, import)` imports a spatial weights matrix from an Excel file or a delimited text file by specifying the respective filename;

- `spmatrix(name, spmatrix)` or `spmatrix(name)` declare the spatial weights matrix to be an SP matrix with the specified name. This must have already been created and stored in memory with Stata's official **spmatrix** command (introduced in Stata 15);

- `spmatrix(name, stata)` or `spmatrix(name, mata)` declare the spatial weights matrix to be a conventional Stata or Mata matrix, respectively.

The command carries out some basic checks on the supplied matrix. The matrix must be square, have no missing values, and all diagonal entries must equal zero. Importantly, the command does not check whether the elements w_{ij} in the matrix are in the correct order corresponding to the spatial units in the data set; this remains a responsibility of the users.

The dynamic model components can be specified with the following options:

- `splag` requests to include a spatial lag of the dependent variable as an additional regressor; i.e., $\sum_{j=1}^{N} w_{ij} y_{jt}$;

- `tlags(#)` requests to include # time lags of the dependent variable as additional regressors; i.e., $y_{it-1}, \ldots, y_{it-\#}$;

- `sptlags(#)` requests to include # spatial time lags of the dependent variable as additional regressors; i.e., $\sum_{j=1}^{N} w_{ij} y_{jt-1}, \ldots, \sum_{j=1}^{N} w_{ij} y_{jt-\#}$;

- `spindepvars(varlist)` requests to include spatial lags of the specified variable list as additional regressors; i.e., $\sum_{j=1}^{N} w_{ij} \mathbf{x}_{jt}^{\top}$.

The estimation relies on sufficiently many instrumental variables. These must be specified with the `iv()` option, which has several suboptions for adding spatial components and controlling the defactorization process:

- `iv(varlist, suboptions)` declares a variable list to be used as instrumental variables;

  - Suboption `splags` requests to include spatial lags of the specified variables as additional instruments;
  - Suboption `spiv(varlist)` adds spatial lags of further variables not yet included in the list of instruments. This allows the use of instruments in spatially lagged form only;
  - Suboption `fvar(varlist)` specifies a list of variables from which to extract the factors. By default, factors are extracted from all instrumental variables;
  - Suboption `lags(#)` requests to add # time lags of the instrumental variables. For each lag order, from 0 to #, factors are computed separately. The default is `lags(0)`;
  - Suboption `factmax(#)` declares the maximum number of factors. The default is `factmax(4)`;
  - Suboption `noeigratrio` requests using a fixed number of factors according to suboption `factmax(#)`, while suboption `eigratio` requests to use the Ahn-Horenstein eigenvalue ratio test to compute the number of factors. The latter is the default;

- Suboptions `std` and `nostd` requests extracting factors from standardized or non-standardized variables, respectively. The latter is the default. Standardization can sometimes help to stabilize the estimation since principal component analysis can be sensitive to the scale of the data, as it is well-known;
- Suboptions `doubledefact` and `nodoubledefact` request to either include or not include the variables in `fvar(varlist)` in a second defactorization stage. This is asymptotically redundant for the model with homogeneous slopes but can improve efficiency in the model with heterogeneous slopes. The default is `nodoubledefact` for the former and `doubledefact` for the latter. See Kripfganz and Sarafidis (2021) for more details regarding this option.

The `iv()` option can be specified multiple times if it is desired to extract factors separately from different sets of instruments. The options `factmax(#)`, `noeigratio`, `std`, and `doubledefact/nodoubledefact` also exist as standalone options, which can be used to alter the default for the respective `iv()` suboptions. Further options are:

- `absorb(varlist)` specifies categorical variables that identify the fixed effects to be absorbed. This option requires the packages **reghdfe** (Correia 2014) and **ftools** (Correia 2016) to be installed. Typical use is `absorb(panelvar)` or `absorb(panelvar timevar)` for one-way or two-way fixed effects, respectively, where `panelvar` and `timevar` are to be replaced by the respective names of the panel and time identifier variables. These are typically the variables used to declare the data to be panel data with the command **xtset**;
- `fstage` requests the first-stage IV estimator to be computed instead of the second-stage estimator. For the model with homogeneous slopes, the first-stage estimator is asymptotically biased and inefficient. For the model with heterogeneous slopes, this option is implied because no efficient second-stage estimator exists;
- `mg` requests to compute the mean-group estimator for the model with heterogeneous slopes. Group-specific coefficients and standard errors are stored in matrices `e(b_mg)` and `e(se_mg)`, respectively;
- `mg(#)` requests to display the group-specific estimates for group # instead of the mean-group estimates. This must be a factor level of the panel identifier variable;
- `noconstant` suppresses the regression intercept;
- `level(#)` sets the confidence level for the confidence interval in the coefficient output. The default is `level(95)`;
- `noheader` and `notable` suppress display of the coefficient table header and the coefficient table itself, respectively.

Further standard options are allowed to alter the appearance of the coefficient table; see the help file for details. To estimate a model without spatial components, the command **xtivdfreg** (Kripfganz and Sarafidis 2021) can be used. Aside from the options for the spatial model components and the time lags, the syntax and options are similar to those described above.

3.3. Stored results

As is standard for estimation commands in Stata, the stored results can subsequently be recovered with the `e()` function:

- The following scalars are stored:
  `e(N)`: number of observations;
  `e(df_m)`: model degrees of freedom;
  `e(N_g)`: number of groups;
  `e(g_min)`: smallest group size;
  `e(g_avg)`: average group size;
  `e(g_max)`: largest group size;
  `e(sigma2u)`: variance of error term $u_{it} = f_{it} + e_{it}$;
  `e(sigma2f)`: variance of factor error component f_{it};
  `e(rho)`: fraction of variance due to factor component;
  `e(chi2_J)`: Hansen's J-statistic;
  `e(df_J)`: degrees of freedom of Hansen's J-test;
  `e(p_J)`: p-value of Hansen's J-test;
  `e(rank)`: rank of e(V);
  `e(zrank)`: number of instruments;
  `e(fact1)`: number of factors in the first stage;
  `e(fact2)`: number of factors in the second stage;
  `e(mg_id)`: group ID for displayed group-specific estimates; only saved with option `mg(#)`;
  `e(splag)`: spatial lag of the dependent variable;
  `e(tlags)`: time lags of the dependent variable;
  `e(sptlags)`: spatial time lags of the dependent variable;
  `e(maxeig)`: maximum eigenvalue of spatial weights matrix.

- The following local macro variables are stored:
  `e(cmd)`: the command name, **spxtivdfreg**;
  `e(cmdline)`: command as typed;
  `e(ivar)`: variable denoting groups;
  `e(tvar)`: variable denoting time;
  `e(estat_cmd)`: the program name for postestimation statistics, **spxtivdfreg_estat**;
  `e(predict)`: the program name for postestimation predictions, **spxtivdfreg_p**;
  `e(marginsok)`: predictions allowed by **margins**;
  `e(vcetype)`: title used to label Std. Err.;
  `e(estimator)`: `fstage`, `sstage`, or `mg`;
  `e(properties)`: `b V`;
  `e(depvar)`: name of dependent variable.

- The following Stata matrices are stored:
  `e(b)`: coefficient vector;
  `e(V)`: variance-covariance matrix of the estimators;
  `e(factnum)`: variable-specific number of factors in the first stage;
  `e(b_mg)`: matrix of group-specific coefficiens; only saved with option `mg`;
  `e(se_mg)`: matrix of group-specific standard errors; only saved with option `mg`.

- Finally, an indicator for the observations used in the current estimation sample is stored in the following function:
`e(sample)`: marks estimation sample.

3.4. Postestimation tools

Below the coefficient table, **spxtivdfreg** displays the Hansen test of the overidentifying restrictions. This test can be re-displayed later by using the **estat overid** command, which does not require further arguments or options.

For spatial panel data models, direct, indirect, and total impacts based on Equation 16 or Equation 25 are often the key quantities of interest. These can be computed with the postestimation command **estat impact** once the model was estimated with **spxtivdfreg**. If the model contains time lags or spatial time lags of the dependent variable, we can further distinguish between short-run and long-run impacts. To compute short-run (`sr`) and long-run (`lr`) impacts for a selected list of independent variables, respectively type

```
estat impact varlist, sr
estat impact varlist, lr
```

Specifying a variable list is optional. By default, impacts are computed for all independent variables, excluding the regression intercept. To include the intercept, option `constant` needs to be added. If the estimated coefficients for the spatial lag, time lags, and spatial time lags violate the model's dynamic stability condition, **estat impact** stops with an error message. Although not recommended, the option `force` can be used in such a situation to ignore the violation of the stability condition. Importantly, for the correct computation of the long-run effects, time lags and spatial time lags of the dependent variable must be specified with the **spxtivdfreg** options `tlags()` and `sptlags()`, and not directly in the list of independent variables.

If further analysis shall be performed on the computed impacts, the option `post` can be used with **estat impact** to replace the coefficient vector `e(b)` in the stored estimation results by the vector of direct, indirect, and total impacts, and accordingly for the variance-covariance matrix in `e(V)`. This allows to subsequently compute linear combinations of the impacts with Stata's **lincom** command, or to test linear hypotheses with the `test` command. To recover the original estimation results, the model either needs to be refit with **spxtivdfreg** or the estimation results should be stored with **estimates store** before invoking **estat impact**.

The **predict** command can be used for predictions after **spxtivdfreg**:

```
predict newvar, statistic
```

where `newvar` is the name for a new variable to be created. `if` or `in` qualifiers can be used in the standard way to select the relevant observations. `statistic` can be one of the following:

- `rform` computes the reduced-form prediction, which is the predicted mean of the dependent variable conditional on the independent variables and any spatial lags of the independent variables. This is the default;

- `direct` computes the prediction of the direct mean, which is a unit's predicted contribution to its reduced-form mean;

- `indirect` computes the prediction of the indirect mean, which is the predicted contribution of all other units to the reduced-form mean; i.e., the predicted contribution to the direct mean subtracted from the reduced-form prediction;
- `naive` computes the naive-form prediction, which is the linear prediction from the fitted model;
- `xb` computes the linear prediction from the fitted model, ignoring the spatial lag of the dependent variable;
- `residuals` calculates the residuals; i.e., the naive-form prediction substracted from the dependent variable.

4. Illustration

As an illustration of our **spxtivdfreg** package, we utilize a subset of the data used by Cui *et al.* (2023). It comprises of 350 U.S. banking institutions, each one observed over the period 2006:Q1 to 2014:Q4. The sampling period is quite rich as it overlaps with the GFC (global financial crisis) during 2007-2008, but it also spans the period of increased capital requirements, which were introduced worldwide in the early 2010s, following the collapse of major banks. In the U.S. the resulting regulatory framework is known as the “Dodd-Frank Wall Street Reform and Consumer Act”.

4.1. Model specification

We shall estimate the same model as in Equation 1, where $y_{it} \equiv \text{NPL}_{it}$ denotes the ratio of non-performing loans to total loans for bank i at time period t. This is a popular measure of credit risk. Higher values of the NPL ratio indicate that banks ex ante took higher lending risk and therefore they have accumulated ex post more bad loans (see Ding and Sickles 2019); $x_{1it} \equiv \text{INEFF}_{it}$ denotes the time-varying operational inefficiency of bank i at period t, which is constructed using a cost frontier model with a translog functional form; $x_{2it} \equiv \text{CAR}_{it}$ stands for “capital adequacy ratio”, proxied by the ratio of core capital over risk-weighted assets; $x_{3it} \equiv \text{SIZE}_{it}$ is proxied by the natural logarithm of banks’ total assets; $x_{4it} \equiv \text{BUFFER}_{it}$ denotes the amount of capital buffer, and it is computed by subtracting from the core capital (leverage) ratio the value of the minimum regulatory capital ratio (8%); profitability, $x_{5it} \equiv \text{PROFIT}_{it}$, is proxied by the return on equity (ROE), defined as annualized net income expressed as a percentage of average total equity on a consolidated basis; $x_{6it} \equiv \text{QUALITY}_{it}$ is computed as the total amount of loan loss provisions (LLP) expressed as a percentage of assets; and $x_{7it} \equiv \text{LIQUIDITY}_{it}$ is proxied by the loan-to-deposit (LTD) ratio. When this ratio is too high, banks may not have enough liquidity to meet unforeseen funding requirements.

The spatial weights matrix has been constructed in three steps. First, we computed Spearman’s correlation coefficient corresponding to a bank’s debt ratio. Second, in order to focus only on the strongest linkages, we set the spatial weights equal to 1 if the correlation exceeds the 95th percentile within a given row, and 0 otherwise. By convention, the diagonal elements of $\mathbf{W}_N$ are set equal to zero, in order to ensure that no individual is treated as its own neighbor. Finally, each of the rows of $\mathbf{W}_N$ has been divided by the sum of its corresponding elements so that $\sum_j w_{ij} = 1$ for all i.

The model is estimated using the 2SIV estimator in Equation 9, combined with the robust variance estimator in Equation 10. A priori, we remove the firm-specific averages over time, $\overline{\mathbf{x}}_{\ell i}$, in order to allow explicitly for individual-specific effects; this is done with the command's `absorb(id)` option, where `id` is the name of the firm identifier variable. INEFF is treated as endogenous with respect to ε_{it} due to reverse causality, and thereby it is instrumented by the ratio of interest expenses paid on deposits over the value of total deposits, $\widetilde{x}_{1it} \equiv \text{INTEREST}_{it}$. Reverse causality arises because higher levels of risk imply additional costs and managerial efforts incurred by banks in order to improve existing loan underwriting and monitoring procedures. The remaining covariates are treated as exogenous with respect to ε_{it}. However, these covariates can be endogenous w.r.t. the common-factor component, $\boldsymbol{\gamma}_{y,i}^{\top}\mathbf{f}_{y,t}$. The matrix of instruments is of the same form as in Equation 18 with $\mathbf{X}_i$ replaced by $\underline{\widetilde{\mathbf{X}}}_i \equiv (\underline{\widetilde{\mathbf{x}}}_{1i}, \underline{\mathbf{x}}_{2i}, \ldots, \underline{\mathbf{x}}_{7i})$, a matrix of order $T \times 7$, where $\underline{\mathbf{x}}_{\ell i} = \mathbf{x}_{\ell i} - \overline{\mathbf{x}}_{\ell i}$ is a $T \times 1$ vector that denotes the ℓth de-meaned covariate corresponding to β_ℓ, for $\ell = 2, \ldots, k$, whereas $\underline{\widetilde{\mathbf{x}}}_{1i} = \widetilde{\mathbf{x}}_{1i} - \overline{\widetilde{\mathbf{x}}}_{1i}$ refers to the external instrument. Thus, we make use of 28 moment conditions in total. With 9 estimated parameters the model is overidentified with 19 degrees of freedom. Additionally, because principal components analysis can be sensitive to the scale of the data, the variables are standardized for the computation of matrix $\widehat{\mathbf{F}}_x$ using option `std`.

4.2. Results

The Stata command line for the estimation is

```
spxtivdfreg NPL INEFF CAR SIZE BUFFER PROFIT QUALITY LIQUIDITY,
  absorb(ID) splag tlags(1) spmatrix("W.csv", import)
  iv(INTEREST CAR SIZE BUFFER PROFIT QUALITY LIQUIDITY,
  splags lag(1)) std
```

where options `splag` and `tlags(1)` specify that the model should include a spatial-lag variable and a time lag of the dependent variable; the first two right-hand side components in Equation 1. The spatial weights matrix is imported from a comma-separated text file. This command line delivers output similar to standard Stata estimation commands; see Table 1.

Besides information about the sample size, the output header indicates the number of instruments and – importantly – the estimated number of factors $\widehat{r}_x = 2$ and $\widehat{r}_y = 1$. A standard table of coefficient estimates and standard errors follows, together with the z-statistic and p-value for the null hypothesis of regressor irrelevance, and associated confidence interval. The first coefficient, `L1.NPL`, refers to the lagged dependent variable, where `L1` is Stata's usual indicator for the first-order time lag. The section titled `W` lists coefficients of spatially lagged variables. In the present illustration, only the dependent variable has been spatially lagged. The final three parameter estimates are self-explanatory from the description in the output. Here, we see that the factors explain a relevant portion – about one third – of the residual variance. Finally, the J-test statistic and its corresponding p-value are reported.

In order to illustrate the practical importance of allowing for unobserved factors in the model, we re-estimate an alternative specification imposing no factors. This is easily done by adding the option `factmax(0)` to the above command line. That is,

```
spxtivdfreg NPL INEFF CAR SIZE BUFFER PROFIT QUALITY LIQUIDITY,
  absorb(ID) splag tlags(1) spmatrix("W.xlsx", import)
```

```
Defactored instrumental variables estimation

Group variable: ID                          Number of obs        =     12250
Time variable: TIME                         Number of groups     =       350

Number of instruments  =     28             Obs per group     min =        35
Number of factors in X =      2                               avg =        35
Number of factors in u =      1                               max =        35

Second-stage estimator (model with homogeneous slope coefficients)
------------------------------------------------------------------------------
             |               Robust
         NPL | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         NPL |
         L1. |   .2898521   .0543794     5.33   0.000     .1832704    .3964339
             |
       INEFF |   .4473777   .1045636     4.28   0.000     .2424368    .6523186
         CAR |   .0305078   .0057852     5.27   0.000      .019169    .0418465
        SIZE |   .2225966   .0941614     2.36   0.018     .0380436    .4071496
      BUFFER |  -.0545049   .0118678    -4.59   0.000    -.0777653   -.0312445
      PROFIT |  -.0053351   .0018411    -2.90   0.004    -.0089437   -.0017266
     QUALITY |   .1830412   .0307657     5.95   0.000     .1227415    .2433408
   LIQUIDITY |   2.452391   .2696471     9.09   0.000     1.923892    2.980889
       _cons |  -4.510715   1.311453    -3.44   0.001    -7.081115   -1.940315
-------------+----------------------------------------------------------------
W            |
         NPL |   .3943206   .0848856     4.65   0.000     .2279479    .5606932
-------------+----------------------------------------------------------------
     sigma_f |  .64162366   (std. dev. of factor error component)
     sigma_e |  .90381799   (std. dev. of idiosyncratic error component)
         rho |  .33509009   (fraction of variance due to factors)
------------------------------------------------------------------------------
Hansen test of the overidentifying restrictions       chi2(19)    =   18.8250
H0: overidentifying restrictions are valid            Prob > chi2 =    0.4681
```

Table 1: Stata regression output for the full model with homogeneous slopes.

```
iv(INTEREST CAR SIZE BUFFER PROFIT QUALITY LIQUIDITY,
splags lag(1)) std factmax(0)
```

Essentially, in this case estimation is implemented in a single stage, as in Equation 7, except the instruments are based on untransformed covariates rather than on "defactored" ones.

That is, the matrix of instruments is given by

$$\widehat{\mathbf{Z}}_i = \left(\sum_{j=1}^{N} w_{ij}\mathbf{X}_j, \quad \mathbf{X}_{i,-1}, \quad \mathbf{X}_i \right). \tag{27}$$

For comparison purposes, we also estimate the original model without a spatially lagged dependent variable; that is, in Equation 1 we impose $\psi = 0$, and the matrix of instruments is formulated as follows:

$$\widehat{\mathbf{Z}}_i = \left(\mathbf{M}_{\widehat{\mathbf{F}}_{x,-1}}\mathbf{X}_{i,-1}, \quad \mathbf{M}_{\widehat{\mathbf{F}}_x}\mathbf{X}_i \right). \tag{28}$$

To estimate this specification, we can remove the option `splag` from the command line and the suboption `splags` from the `iv()` option.[10] Alternatively, we can directly use the **xtivdfreg** command, on which **spxtivdfreg** is based. Note that **xtivdfreg** does not have a `tlags()` option, so that time-lagged regressors need to be specified explicitly with the lag-operator notation:

```
xtivdfreg NPL L.NPL INEFF CAR SIZE BUFFER PROFIT QUALITY
  LIQUIDITY, absorb(ID) iv(INTEREST CAR SIZE BUFFER
  PROFIT QUALITY LIQUIDITY, lag(1)) std
```

The advantage of using **spxtivdfreg** over **xtivdfreg** even in the case without spatial lags is that it enables the computation of long-run impacts with the postestimation command `estat impact`, which is demonstrated further below for the case with spatially lagged dependent variable. Due to the absence of indirect impacts in the non-spatial case, the direct and total short-run impacts simply equal the regression coefficients. Instead of showing similar Stata output, we compare the coefficient estimates and standard errors in Table 2.

To begin with, consider the spatial autoregressive specifications with and without a factor structure in the first two columns. Both $\widehat{\psi}$ and $\widehat{\rho}$ are statistically significant, providing evidence for endogenous spatial interactions and state dependence. However, there are significant differences in the magnitude of the two estimated coefficients, as well as in terms of the standard errors of these estimates. The model without factors underestimates the magnitude of the spatial interactions but grossly overestimates the degree of state dependence. This affects substantially the estimates of direct/indirect effects, to be reported below. At the same time, the model without factors incorrectly suggests far more precision in the estimates of these parameters. Regarding the model without spatial lags, we notice that some coefficient estimates remain fairly similar. However, changes in others are non-negligible. Most notably, we observe a larger estimate of the inefficiency coefficient when spatial terms are ignored. The number of estimated factors remains the same as in the initial model.

The positive coefficient of operational inefficiency provides support for the so-called "bad management hypothesis" (e.g., see Fiordelisi, Marques-Ibanez, and Molyneux 2011), which postulates that managers' failure to control costs efficiently, can result in poor monitoring of loans and thereby higher default rates. When factors are accounted for, the estimated effect is stronger than without factors. Other notable differences between the two specifications include: (i) while all other coefficients are statistically significant at least at the 5% level when factors are accounted for, the estimated effect of size becomes negligible when factors are ignored; (ii) the estimated coefficient for liquidity is almost three times as large in the

[10]Similarly, to estimate a model without a time lag, the option `tlags(1)` can be removed.

	Full model	Without factors	Without spatial lag
$\widehat{\psi}$ ($\mathbf{W}_N$ NPL_t)	0.394***	0.288***	
	(0.085)	(0.038)	
$\widehat{\rho}$ (NPL_{t-1})	0.290***	0.594***	0.323***
	(0.054)	(0.034)	(0.055)
$\widehat{\beta}_1$ (INEFF_t)	0.447***	0.366***	0.638***
	(0.105)	(0.107)	(0.116)
$\widehat{\beta}_2$ (CAR_t)	0.031***	0.017***	0.030***
	(0.006)	(0.004)	(0.006)
$\widehat{\beta}_3$ (SIZE_t)	0.223**	0.089	0.346***
	(0.094)	(0.061)	(0.096)
$\widehat{\beta}_4$ (BUFFER_t)	-0.055***	-0.025**	-0.045***
	(0.012)	(0.010)	(0.016)
$\widehat{\beta}_5$ (PROFIT_t)	-0.005***	-0.006***	-0.004**
	(0.002)	(0.002)	(0.002)
$\widehat{\beta}_6$ (QUALITY_t)	0.183***	0.283***	0.183***
	(0.031)	(0.029)	(0.036)
$\widehat{\beta}_7$ (LIQUIDITY_t)	2.452***	0.843***	2.534***
	(0.270)	(0.180)	(0.311)
$\widehat{r}_x$	2	0	2
$\widehat{r}_y$	1	0	1
J-test	18.825	48.151	8.174
	[0.468]	[0.000]	[0.226]

Table 2: Coefficient estimates; see Section 4.1 for details on the model specification.

former specification compared to no factors. In the latter specification, this risk-enhancing effect of a high LTD ratio is thus substantially underestimated.

The J-test indicates that the null hypothesis of valid overidentifying restrictions cannot be rejected at conventional significance levels in the model with factors. In contrast, when we do not allow for latent factors, the p-value of the J-test statistic decreases drastically, and the null hypothesis of correct model specification is rejected. This demonstrates the importance of allowing for common factors in the residuals in this model. Note also that failure to reject the null hypothesis in the first specification provides support to the homogeneous model. This is because pooling the model across i when the parameters are individual-specific, renders the instruments used invalid.[11]

The coefficient estimates discussed above can be regarded as immediate effects before any spillovers or adjustments to deviations from a long-run equilibrium are taken into account. Long-run direct, indirect, and total effects are obtained by simply typing

```
estat impact, lr
```

which yields the output in Table 3 for the model with factors.[12]

[11]See Juodis and Sarafidis (2022a) for a recent example.

[12]The short-run results are qualitatively similar, and so we do not provide them here to save space. They

Long-run impacts

	Impact	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
direct						
INEFF	.6470588	.1593924	4.06	0.000	.3346554	.9594623
CAR	.0441245	.0092325	4.78	0.000	.0260292	.0622198
SIZE	.3219497	.1416728	2.27	0.023	.044276	.5996233
BUFFER	-.0788324	.0183176	-4.30	0.000	-.1147342	-.0429306
PROFIT	-.0077164	.0023773	-3.25	0.001	-.0123757	-.003057
QUALITY	.2647392	.0466629	5.67	0.000	.1732816	.3561968
LIQUIDITY	3.546983	.4454284	7.96	0.000	2.673959	4.420007
indirect						
INEFF	.7694677	.3352809	2.29	0.022	.1123291	1.426606
CAR	.0524719	.0237326	2.21	0.027	.0059569	.0989868
SIZE	.3828552	.1975749	1.94	0.053	-.0043845	.770095
BUFFER	-.0937457	.0428643	-2.19	0.029	-.1777581	-.0097333
PROFIT	-.0091761	.0046348	-1.98	0.048	-.0182603	-.000092
QUALITY	.3148218	.1408165	2.24	0.025	.0388266	.590817
LIQUIDITY	4.217992	1.742264	2.42	0.015	.8032163	7.632767
total						
INEFF	1.416526	.4274849	3.31	0.001	.5786715	2.254382
CAR	.0965964	.0291942	3.31	0.001	.0393768	.1538159
SIZE	.7048049	.3099048	2.27	0.023	.0974027	1.312207
BUFFER	-.1725781	.0541498	-3.19	0.001	-.2787098	-.0664465
PROFIT	-.0168925	.0063692	-2.65	0.008	-.0293759	-.0044091
QUALITY	.579561	.1670612	3.47	0.001	.2521271	.9069949
LIQUIDITY	7.764974	1.90367	4.08	0.000	4.033851	11.4961

Table 3: Stata output for the long-run impacts.

The output is separated into three sections for the direct, indirect, and total impacts, respectively. Again, to save space, we do not provide output in the same format for the restricted model without factor structure; however, we compare the estimates of the two models in Table 4. Total effects are simply the sum of direct and indirect effects. These results provide a more complete picture of the cumulative effects over time and across banks. Due to the stronger persistence estimate, $\widehat{\rho} = 0.594$ compared to $\widehat{\rho} = 0.290$, those cumulative effects now tend to be higher for most variables in the model without factors. While the immediate impact of inefficiency is estimated to be smaller when factors are left out, the total long-run impact is

can be easily recovered by replacing the `lr` option with the `sr` option in the `estat impact` postestimation command line. The respective output table has the same structure as the one for the long-run impacts.

actually more than double the impact corresponding to the model with factors. Similarly, the differences in the liquidity impacts are much less pronounced over the long-run than indicated by the regression coefficients above.

	With factors			Without factors		
	Direct	Indirect	Total	Direct	Indirect	Total
INEFF	0.647***	0.769**	1.417***	0.959***	2.157**	3.117**
CAR	0.044***	0.052**	0.097***	0.045***	0.100**	0.145**
SIZE	0.322**	0.383*	0.705**	-0.233	0.523	0.756
BUFFER	-0.079***	-0.094**	-0.173***	-0.065**	-0.147	-0.212*
PROFIT	-0.008***	-0.009**	-0.017***	-0.016***	-0.036**	-0.053***
QUALITY	0.265***	0.315**	0.580***	0.741***	1.666***	2.407***
LIQUIDITY	3.547***	4.218**	7.765***	2.209***	4.967**	7.176**

Table 4: Decomposition of long-run effects.

Still, considerable differences persist between the two specifications. In the light of the earlier J-test result, the impact estimates from the model with factors remain more reliable. Across variables, the long-run indirect impacts appear stronger than the direct impacts, highlighting the importance of accounting for spatial spillover effects.

Even though the J-test does not reject the model with homogeneous slopes, it can still be instructive to run the estimation with heterogeneous slopes. This is easily achieved by adding the `mg` option to the earlier command line. The `Stata` output is shown in Table 5. We notice – somewhat unexpectedly – that some of the coefficient estimates differ substantially from the earlier ones, even at an order of magnitude for the coefficients of size and capital buffer, which is at odds with the earlier J-test result. In addition, the spatial spillover effects disappear under heterogeneous slopes. Consequently, computing direct, indirect, and total spillovers does not provide any additional insights here. A possible explanation for these results could be that the MG estimator is susceptible to outliers. While there is insufficient parameter heterogeneity to trigger a rejection of the J-test in finite samples, a few outliers can result in seemingly contradicting results from the MG estimator. The larger standard errors of the MG estimates compared to the estimation with homogeneous coefficients might also be indicative of outliers. However, an increase in standard errors is expected because the MG estimator is only $\sqrt{N}$-consistent and thereby less efficient than 2SIV.[13] After estimation, group-specific coefficients and standard errors can be extracted from the matrices `e(b_mg)` and `e(se_mg))`, respectively, which could then be used for further analysis. For a specific group – say, the bank with ID number 101 – the estimates can alternatively be displayed as standard estimation output with option `mg(101)`. However, while it would be worth investigating the potential influence of outliers, a further exploration is beyond the scope of this paper.

5. Summary and discussion

[13]MG estimation only proceeds in one stage because the second stage would require factors to be estimated from the residuals of group-specific estimations when coefficients are heterogeneous; see footnote 9 in Kripfganz and Sarafidis (2021). This would likely be even more inefficient.

```
Defactored instrumental variables estimation

Group variable: ID                          Number of obs        =     12250
Time variable: TIME                         Number of groups     =       350

Number of instruments  =     28             Obs per group     min =        35
Number of factors in X =      2                               avg =        35
                                                              max =        35

Mean-group estimator (model with heterogeneous slope coefficients)
------------------------------------------------------------------------------
             |               Robust
         NPL | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         NPL |
         L1. |   .3005247   .0148501    20.24   0.000      .271419    .3296303
             |
       INEFF |   .7587664   .1583511     4.79   0.000     .4484039    1.069129
         CAR |    .218054   .0262755     8.30   0.000      .166555    .2695531
        SIZE |   2.004026   .3385335     5.92   0.000     1.340513     2.66754
      BUFFER |  -.3763774   .0420252    -8.96   0.000    -.4587453   -.2940095
      PROFIT |  -.0179663    .005944    -3.02   0.003    -.0296164   -.0063161
     QUALITY |   .2872525   .1386973     2.07   0.038     .0154107    .5590942
   LIQUIDITY |   6.330179   .5059499    12.51   0.000     5.338536    7.321823
       _cons |  -29.01259   4.166689    -6.96   0.000    -37.17915   -20.84603
-------------+----------------------------------------------------------------
W            |
         NPL |    .031593   .0511028     0.62   0.536    -.0685667    .1317528
------------------------------------------------------------------------------
```

Table 5: Stata regression output for the model with heterogeneous slopes.

The package **spxtivdfreg** introduces two IV estimators for estimating large-N spatial, dynamic panel data models with unobserved common factors. The slope coefficients can be either homogeneous or heterogeneous. The command accommodates a flexible specification of instruments. The spatial weights matrix can also be generated in an Excel file or a delimited text file and subsequently imported directly into Stata using **spxtivdfreg**.

Version requirements

The **spxtivdfreg** command is part of the **xtivdfreg** package (Kripfganz and Sarafidis 2021, version 1.4.2 or newer), which can be installed from the Statistical Software Components (SSC) archive using the following command line in Stata:

```
ssc install xtivdfreg
```

The package requires at least Stata version 13.0. Some features require newer versions, such as importing a spatial weights matrix from a file (Stata 14.0), using a spatial weights matrix created with **spmatrix** (Stata 15.0), and standardizing variables for the factor extraction with option `std` (Stata 16.1). The use of option `absorb()` requires the packages **reghdfe** (Correia 2014, version 6.12.3) and **ftools** (Correia 2016, version 2.49.1) to be installed, which are available from SSC as well.

Acknowledgments

We would like to thank seminar participants at the University of Bologna, the BI Norwegian Business School, and the 2023 SNDE Symposium. Vasilis Sarafidis gratefully acknowledges financial support from the Australian Research Council under research grant number DP-170103135.

References

Anselin L, Li X, Koschinsky J (2006). "GeoDa: An Introduction to Spatial Data Analysis." *Geographical Analysis*, **38**(1), 5–22. doi:https://doi.org/10.1111/j.0016-7363.2005.00671.x.

Anselin L, Li X, Koschinsky J (2022). "GeoDa, From the Desktop to an Ecosystem for Exploring Spatial Data." *Geographical Analysis*, **54**(3), 439–466. doi:https://doi.org/10.1111/gean.12311.

Bai J, Li K (2021). "Dynamic Spatial Panel Data Models with Common Shocks." *Journal of Econometrics*, **224**(1), 134–160. doi:10.1016/j.jeconom.2020.12.002.

Belotti F, Hughes G, Mortari AP (2017). "Spatial Panel-Data Models Using Stata." *The Stata Journal*, **17**(1), 139–180. doi:10.1177/1536867X1701700109.

Bivand R (2022). "R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data." *Geographical Analysis*, **54**(3), 488–518. doi:10.18637/jss.v063.i18.

Bivand R (2023). ***spdep**: Spatial Dependence: Weighting Schemes, Statistics.* R package version 1.2-8, URL https://cran.r-project.org/package=spdep.

Bivand R, Piras G (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software*, **63**(18). doi:10.18637/jss.v063.i18.

Bivand R, Piras G (2023). ***spatialreg**: Spatial Regression Analysis.* R package version 1.2-9, URL https://cran.r-project.org/package=spatialreg.

Bun MJG, Sarafidis V (2015). "Dynamic Panel Data Models." In BH Baltagi (ed.), *The Oxford Handbook of Panel Data*, chapter 3. Oxford University Press, Oxford. doi:10.1093/oxfordhb/9780199940042.013.0003.

Case A (1991). "Spatial Patterns in Household Demand." *Econometrica*, **59**(4), 953–965. doi:10.2307/2938168.

Chen J, Cui G, Sarafidis V, Yamagata T (2023). "IV Estimation of Heterogeneous Spatial Dynamic Panel Models with Interactive Effects." Mimeo.

Chen J, Shin Y, Zheng C (2022). "Estimation and Inference in Heterogeneous Spatial Panels with a Multifactor Error Structure." *Journal of Econometrics*, **229**(1), 55–79. doi:10.1016/j.jeconom.2021.05.003.

Correia S (2014). "**REGHDFE**: Stata Module to Perform Linear or Instrumental-Variable Regression Absorbing Any Number of High-Dimensional Fixed Effects." *Statistical Software Components S457874*, Boston College Department of Economics. Version 5.7.3, Revised 18 Nov 2019, URL https://ideas.repec.org/c/boc/bocode/s457874.html.

Correia S (2016). "**FTOOLS**: Stata Module to Provide Alternatives to Common Stata Commands Optimized for Large Datsets." *Statistical Software Components S458213*, Boston College Department of Economics. Version 2.37.0, Revised 18 Nov 2019, URL https://ideas.repec.org/c/boc/bocode/s458213.html.

Cui G, Norkute M, Sarafidis V, Yamagata T (2022). "Two-Stage Instrumental Variable Estimation of Linear Panel Data Models with Interactive Effects." *The Econometrics Journal*, **25**(2), 340–361. doi:10.1093/ectj/utab029.

Cui G, Sarafidis V, Yamagata T (2023). "IV Estimation of Spatial Dynamic Panels with Interactive Effects: Large Sample Theory and an Application on Bank Attitude Towards Risk." *The Econometrics Journal*, **26**(2), 124–146. doi:10.1093/ectj/utac026.

Debarsy N, Ertur C, LeSage JP (2012). "Interpreting Dynamic Space-Time Panel Data Models." *Statistical Methodology*, **9**(1–2), 158–171. doi:10.1016/j.stamet.2011.02.002.

Ding D, Sickles R (2019). "Capital Regulation, Efficiency, and Risk Taking: A Spatial Panel Analysis of US Banks." In M Tsionas (ed.), *Panel Data Econometrics*, chapter 13, pp. 405–466. Elsevier, London. doi:10.1016/B978-0-12-815859-3.00013-5.

Drukker D, Peng H, Prucha IR, Raciborski R (2013a). "Creating and Managing Spatial-Weighting Matrices with the **spmat** Command." *The Stata Journal*, **13**(2), 242–286. doi:10.1177/1536867X1301300202.

Drukker D, Prucha IR, Raciborski R (2013b). "A Command for Estimating Spatial-Autoregressive Models with Spatial-Autoregressive Disturbances and Additional Endogenous Variables." *The Stata Journal*, **13**(2), 287–301. doi:10.1177/1536867X1301300203.

Drukker D, Prucha IR, Raciborski R (2013c). "Maximum Likelihood and Generalized Spatial Two-Stage Least-Squares Estimators for a Spatial-Autoregressive Model with Spatial-Autoregressive Disturbances." *The Stata Journal*, **13**(2), 221–241. doi:10.1177/1536867X1301300201.

Elhorst JP (2012). "Dynamic Spatial Panels: Models, Methods, and Inferences." *Journal of Geographical Systems*, **14**, 5–28. doi:10.1007/s10109-011-0158-4.

Elhorst JP (2014a). "MATLAB Software for Spatial Panels." *International Regional Science Review*, **37**(3), 247–405. doi:https://doi.org/10.1177/0160017612452429.

Elhorst JP (2014b). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer-Verlag, London. doi:10.1007/978-3-642-40340-8.

Fiordelisi F, Marques-Ibanez D, Molyneux P (2011). "Efficiency and Risk in European Banking." *Journal of Banking & Finance*, **35**(5), 1315–1326. doi:10.1016/j.jbankfin.2010.10.005.

Freeman H, Weidner M (2023). "Linear Panel Regressions with Two-Way Unobserved Heterogeneity." *Journal of Econometrics*, **237**(1), 105498. doi:10.1016/j.jeconom.2023.105498.

Juodis A, Sarafidis V (2022a). "An Incidental Parameters Free Inference Approach for Panels with Common Shocks." *Journal of Econometrics*, **229**(1), 19–54. doi:10.1016/j.jeconom.2021.03.011.

Juodis A, Sarafidis V (2022b). "A Linear Estimator for Factor-Augmented Fixed-T Panels With Endogenous Regressors." *Journal of Business & Economic Statistics*, **40**(1), 1–15. doi:10.1080/07350015.2020.1766469.

Kelejian HH, Prucha IR (2010). "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances." *Journal of Econometrics*, **157**(1), 53–67. doi:10.1016/j.jeconom.2009.10.025.

Kripfganz S, Sarafidis V (2021). "Instrumental-Variable Estimation of Large-T Panel-Data Models with Common Factors." *The Stata Journal*, **21**(3), 659–686. doi:10.1177/1536867X211045558.

LeSage JP, Pace RK (2009). *Introduction to Spatial Econometrics*. CRC Press, Boca Raton.

Manski CF (1993). "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies*, **60**(3), 531–542. doi:10.2307/2298123.

Millo G, Piras G (2012). "**splm**: Spatial Panel Data Models in R." *Journal of Statistical Software*, **47**(1). doi:10.18637/jss.v047.i01.

Pebesma E, Bivand R, Ribeiro PJ (2015). "Software for Spatial Statistics." *Journal of Statistical Software*, **63**(11). doi:10.18637/jss.v063.i01.

Pesaran MH (2006). "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica*, **74**(4), 967–1012. doi:10.1111/j.1468-0262.2006.00692.x.

Pisati M (2001). "**sg162**: Tools for Spatial Data Analysis." *Stata Technical Bulletin*, **60**, 21–37.

Pisati M (2007). "**SPMAP**: Stata Module to Visualize Spatial Data." *Statistical Software Components S456812*, Boston College Department of Economics. Version 1.3.2, Revised 18 Jan 2018, URL https://ideas.repec.org/c/boc/bocode/s456812.html.

Rey SJ, Anselin L (2010). "**PySal**: A Python Library of Spatial Analytical Methods." In MM Fischer, A Getis (eds.), *Handbook of Applied Spatial Analysis*. Springer-Verlag, Berlin. doi:https://doi.org/10.1007/978-3-642-03647-7_11.

Rey SJ, Anselin L, Amaral P, Arribas-Bel D, Cortes RX, Gaboardi JD, Kang W, Knaap E, Li Z, Lumnitz S, Oshan TM, Shao H, Wolf LJ (2022). "The **PySal** Ecosystem: Philosophy and Implementation." *Geographical Analysis*, **54**(3), 467–487.

Sarafidis V, Wansbeek TJ (2021). "Celebrating 40 Years of Panel Data Analysis: Past, Present and Future." *Journal of Econometrics*, **220**(2), 215–226. doi:10.1016/j.jeconom.2020.06.001.

Shi W, Lee LF (2017). "Spatial Dynamic Panel Data Models with Interactive Fixed Effects." *Journal of Econometrics*, **197**(2), 323–347. doi:10.1016/j.jeconom.2016.12.001.

Yu JR, de Jong R, Lee LF (2008). "Quasi-Maximum Likelihood Estimators for Spatial Dynamic Panel Data with Fixed Effects When Both n and T are Large." *Journal of Econometrics*, **146**(1), 118–134. doi:10.1016/j.jeconom.2008.08.002.

Affiliation:

Sebastian Kripfganz
University of Exeter Business School
Exeter, EX4 4PU, United Kingdom
E-mail: S.Kripfganz@exeter.ac.uk
URL: https://www.kripfganz.de/

Vasilis Sarafidis
Brunel University London
Uxbridge, UB8 3PH, United Kingdom
E-mail: vasilis.sarafidis@brunel.ac.uk
URL: https://sites.google.com/view/vsarafidis