Joint Modelling and Estimation of Global and Local Cross-Sectional Dependence in Large Panels[1]

by Siem Jan Koopman[a,b,c], Julia Schaumburg[a,b] and Quint Wiersma[a,b]

[a] *Vrije Universiteit Amsterdam, the Netherlands*
[b] *Tinbergen Institute, Amsterdam, the Netherlands*
[c] *CREATES, Aarhus University, Denmark*

November, 2021

We propose a new unified approach to identifying and estimating spatio-temporal dependence structures in large panels. The model accommodates global cross-sectional dependence due to global dynamic factors as well as local cross-sectional dependence, which may arise from local network structures. Model selection, filtering of the dynamic factors, and estimation are carried out iteratively using a new algorithm that combines the Expectation-Maximization algorithm with coordinate descent and gradient descent, allowing us to efficiently maximize an ℓ_1- and ℓ_2-penalized state space likelihood function. A Monte Carlo simulation study illustrates the good performance of the algorithm in terms of determining the presence and magnitude of global and/or local cross-sectional dependence. In an empirical application, we investigate monthly US interest rate data on 15 maturities over almost 40 years. We find that besides a changing number of global dynamic factors, there is heterogeneous local dependence among neighboring maturities. Taking this heterogeneity into account substantially improves out-of-sample forecasting performance.

Keywords: high-dimensional factor model, Lasso, spatial error model, yield curve

JEL Classification: C32, C33, C38

1 ‖ Introduction

We propose a new unified approach to identifying and estimating spatio-temporal dependence structures in large panel data sets. Conditionally on a set of unobserved factors that impact many but not necessarily all cross-sectional units, outcome observations may depend on the outcomes of neighboring units. Furthermore, our framework accommodates weakly exogenous covariates, as well as heteroskedasticity. Model selection and

[1]Email addresses: s.j.koopman@vu.nl (Siem Jan Koopman), j.schaumburg@vu.nl (Julia Schaumburg), q.wiersma@vu.nl (Quint Wiersma)

estimation are conducted using a new algorithm that allows us to efficiently maximize a penalized state space likelihood function with many unknown coefficients.

Dissecting several sources of cross-sectional dependence and heterogeneity is particularly useful when modelling and forecasting the term structure of interest rates. Apart from well-known factors such as level, slope, and curvature, the yield curve has been shown to react to macroeconomic shocks. Furthermore, the preferred habitat theory of the yield curve suggests the presence of a more local network structure between neighboring maturities. Our framework is flexible enough to incorporate all these potential features in an empirical model, while providing a means for automatic data-driven deselection of variables. We investigate how the best model for a cross-section of yields changes over time in an empirical study of high-dimensional US interest rate data.

Models for large panel data sets with cross-sectional dependence have been and are being addressed in a growing number of studies. Among the most widely used tools to introduce cross-sectional dependence to panel data models are spatial models. Spatial models traditionally rely on an observed, exogenous weights matrix that defines the neighborhood structure among units, and that enters the model with a scalar unknown intensity parameter, see Anselin (1988), LeSage and Pace (2008) and Elhorst (2014) for textbook treatments. Recently, this rigid structure has been relaxed in several ways: Aquaro et al. (2020) allow for a vector of spatial intensity parameters, Lam and Souza (2019) provide a framework in which the spatial weights matrix can be estimated, and Kuersteiner and Prucha (2020) incorporate the possibility of endogenous formation of the weights matrix. On the other hand, Bai and Li (2015) discusses quasi-maximum likelihood estimation of spatial lag models with observed regressors and common factors.

Our modeling approach combines the frameworks of Aquaro et al. (2020) and Bai and Li (2015) and extends to high dimensions: While allowing for the possibility of heterogeneous network intensity parameters and regression coefficients as well as heteroscedasticity, a Lasso type penalty term ensures tractability and serves as a built-in model selection device for the individual static coefficients. An additional group Lasso penalty term can be used for choosing the number of factors by (de-)selecting entire columns of the loading matrix. While Hirose and Konishi (2012) and Lu and Su (2016) use group Lasso for selecting the number of factors in a static factor model, to the best of our knowledge, this procedure has not been used in the context of the dynamic factor model. In the absence of regressors and spatial dependence, our model framework also nests a high-dimensional factor model in state space form. Bayesian sparse factor models have been analyzed in Frühwirth-Schnatter and Lopes (2018), Kaufmann and Schumacher (2017), Kaufmann and Schumacher (2019). They require computationally intensive MCMC algorithms for estimation. In contrast, we combine filtering and esti-

mation in an efficient iterative procedure that combines the Expectation-Maximization algorithm of Dempster et al. (1977) with coordinate and gradient descent.

The coordinate descent proximal Expectation-Conditional Maximization (CDPECM) algorithm developed to estimate the dynamic factor model with spatial errors and exogeneous regressors is build upon two algorithms. First, blockwise coordinate descent update steps are derived for a multiple response generalized elastic net group Lasso regression, based on the work of Simon et al. (2013). Second, we use a proximal gradient descent type algorithm for ℓ_1 regularized objective functions (see e.g. Parikh and Boyd, 2014).

As pointed out in Chudik et al. (2011), it is important to disentangle different sources of cross-sectional dependence in panel data, in particular global dependence introduced by factors, and local or weak dependence that may be due to local network structures. In a Monte Carlo study, we investigate the ability of our method to distinguish between the two types of dependence in a variety of settings. We also focus on model selection ability, as well as estimation precision for the nonzero parameters in the model, including factor loadings, slope coefficients, spatial intensity parameters, and unit-specific variances. Overall, we can find that the method performs well for settings that are comparable to sample sizes of real data sets.

Finally, we contribute to the empirical literature on modeling the yield curve of interest rates across time and different monetary policy regimes, see, for instance, Diebold and Li (2006), Härdle and Majer (2016), Eo and Kang (2020), and many others. The conditional mean specification of our empirical model features the well-known factor structure including level, slope and curvature as well as a set of macroeconomic variables, as it has been shown in several studies that the term structure of interest rates reacts to macroeconomic shocks, see, for instance, Ludvigson and Ng (2009), Coroneo et al. (2016) and Bianchi et al. (2020). Furthermore, we allow for shock spillovers among neighboring maturities, to capture the possibility of segmented investors that target specific maturities as suggested in Vayanos and Vila (2009). Using a rolling window analysis, we find substantial variation in the number of factors and the magnitude of factor loadings. Beyond the common factors, we find evidence for local dependence in the error terms. In terms of out-of-sample performance our method performs very well. For the majority of maturities, it significantly outperforms the widely used dynamic Nelson-Siegel model of Diebold and Li (2006).

The remainder of the paper is organized as follows. Section 2 introduces the dynamic factor model with covariates and spatial errors as well as some extensions. In Section 3, we present the details of our new algorithm, allowing us to simultaneously conduct estimation, filtering, and model selection in a high-dimensional setting. An extensive

Monte Carlo study can be found in Section 4. We present the empirical results in Section 5. Section 6 concludes.

2 ‖ Dynamic Factor Model with Spatial Errors

2.1 ‖ Model

The model combines the spatial error model with a dynamic factor model and exogenous regressors. It is given by

$$\begin{aligned} y_t &= X_t\beta + \Lambda f_t + \xi_t\,, \\ f_{t+1} &= \phi f_t + \eta_t\,, \qquad & \eta_t &\sim \mathcal{N}(0,\Sigma_\eta)\,, \\ \xi_t &= \rho W\xi_t + \varepsilon_t\,, \qquad & \varepsilon_t &\sim \mathcal{N}(0,\Sigma_\varepsilon)\,, \end{aligned} \tag{1}$$

for $t = 1,\dots,T$ where T is the length of the time series, $y_t = (y_{1t},\dots,y_{Nt})'$ is a N-dimensional time series, X_t is a $N \times K$ matrix of exogenous regressors, β is a K-dimensional vector of unknown coefficients, f_t is a r-dimensional vector of factors, Λ is an unknown $N \times r$ loading matrix with factor loadings Λ_{ij}, ϕ is a $r \times r$ autoregressive coefficient matrix, and Σ_η is a $r \times r$ diagonal covariance matrix. Furthermore, the scalar coefficient ρ captures the spatial dependence, W is a $N \times N$ exogenous matrix of spatial weights, and Σ_ε is a diagonal $N \times N$ covariance matrix.

We assume that the r factors are independent, hence, ϕ is a diagonal matrix with autoregressive coefficients ϕ_i, $i = 1,...,r$. Moreover, we restrict Σ_η to be an $r \times r$ identity matrix. These two restrictions combined are sufficient to overcome the rotational identification issues of the loading matrix Λ (see Appendix A).

Following Bailey et al. (2016) and Aquaro et al. (2019) we also allow for heterogeneous spatial dependence. The spatial error equation of the model outlined in (1) is then modified to

$$\xi_t = PW\xi_t + \varepsilon_t\,,$$

where $P = \mathrm{diag}(\rho) = \mathrm{diag}(\rho_1,\dots,\rho_N)$.

We can write the model more compactly by combining the first and last lines of (1) to obtain a more familiar representation for the observation equation of a state space model

$$y_t = X_t\beta + \Lambda f_t + G\varepsilon_t\,,$$

with $G = (I - \rho W)^{-1}$. Using this formulation, signal extraction of the factors f_t can be based on the Kalman filter and smoother routines. The Kalman filter and smoother

equations and their derivations are given by, among others, Harvey (1989) and Durbin and Koopman (2012).

The log-likelihood function can readily be calculated using the Kalman filter techniques. For a given parameters vector θ, which collects the unknown parameters β, Λ, ϕ, ρ, and Σ_ε, the log-likelihood is given by

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{t=1}^{T} \mathcal{L}_t(\theta) \\ &= -\frac{NT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\log|F_t| + v_t' F_t^{-1} v_t\right) ,\end{aligned}$$

where v_t is the forecast error and F_t the forecast error variance. Both objects are provided by the Kalman filter.

2.2 ‖ Cross-Sectional Dependence in Panels

Cross-sectional dependence may be due to common effects stemming from local and/or common factors. The common factors are pervasive in nature and reflect aggregate shocks. In our framework these common factors are modelled by the dynamic factors f_t. The local factors reflect spatial interactions, that generate spill-over effects, between units which are 'close' together. In this case 'close' can refer to actual geographical distance, economic relationships, or any measure, for that matter, of closeness in terms of interactions. These local factors are captured by the spatial errors in (1). As pointed out in Chudik et al. (2011), it is important to disentangle different sources of cross-sectional dependence in panel data.

We will refer to the two types of cross-sectional dependence as global (pervasive in nature) and local (not pervasive in nature). These types of cross-sectional dependence are closely related to the notions of strong and weak cross-sectional dependence as defined by Chudik et al. (2011).

The very general framework of Chudik and Pesaran (2013), which includes static factors and a weak cross-sectional dependent error process, can allow for the same dependency structures as our framework. However, Chudik and Pesaran (2013) are mainly interested in the slope coefficients of the exogenous regressors. In this paper we are directly interested in the unobserved dynamic factors, their loadings, and the spatial structure in the errors.

2.2.1 ‖ *Identifying Types of Cross-Sectional Dependence*

As outlined before the dynamic factor model with spatial errors can accommodate both global and local cross-sectional dependence. Bailey et al. (2016) propose to identify the

presence of strong and weak cross-sectional dependence using a two-step procedure. This two-step procedure tests, in the first step, whether weak-cross sectional dependence is present and, if not, proceeds by estimating a static factor model. In the second step the "de-factored" residuals are tested for weak cross-sectional dependence and, if present, a spatial autoregressive model is fitted.

In contrast, we aim at identifying whether the data exhibits local cross-sectional dependence, global cross-sectional dependence, or both, jointly with estimation of the parameters. Regularization techniques enable us to do so. As the notions of strong and weak cross-sectional dependence rely on the cross-sectional dimension tending to infinity we use the terms global and local dependence, in order to apply our regularization approach in settings with finite N.

By estimating the model using standard regularization techniques, such as Lasso and group Lasso, the approach is able to differentiate between the two forms of cross-sectional dependence in a data-driven way. We allow for a sparse, or even zero, loading matrix and a potentially zero coefficient for the spatial dependence parameter. The algorithm will be discussed in more detail in the next section.

3 ‖ Estimation

In order to estimate a dynamic factor model with spatial errors, where the types of cross-sectional dependence are identified during estimation, we optimize the regularized log-likelihood via an algorithm that is inspired by the EM algorithm of Dempster et al. (1977). One of the first state space applications of the EM algorithm is by Shumway and Stoffer (1982). The idea of the EM algorithm is as follows. If we knew the latent/missing data, standard estimation techniques can be deployed to estimate the model parameters. Once we update the model parameters we can make a much better estimate/guess of the latent variables. The EM algorithm iterates between these two steps to obtain the maximum likelihood estimates.

Our EM algorithm is not tailored towards optimizing a standard log-likelihood, the focus is on a penalized likelihood, where the penalization has the form of an ℓ_1- and an additional ℓ_2-norm penalty on the parameters. Due to the non-smooth nature of the objective function, arising from the ℓ_1- and ℓ_2-norms, score-driven (numerical) optimization techniques/approaches to optimize the penalized likelihood directly are cumbersome, difficult, and very sensitive. The reason why the EM algorithm is better suited for optimizing a penalized log-likelihood of state space models in general is due to the fact that after taking the conditional expectation (E-step) of the complete data likelihood we can exploit the linear nature of the model very efficiently. In terms of dealing

with the non-smooth part of the objective function, a fast and reliable algorithm is devised on the basis of coordinate descent steps (for a standard linear regression framework suggested and introduced by van der Kooij (2007), Friedman et al. (2007), and Friedman et al. (2010)) and a proximal gradient algorithm for non-smooth constrained optimization (see Parikh and Boyd, 2014).

Meng and Rubin (1993) show that the EM algorithm can be sped up significantly by their expectation conditional maximization (ECM) approach. The ECM algorithm differs from the standard EM algorithm in the way the maximization step is conducted. In the EM algorithm the maximization step is conducted by optimizing over the entire set of parameters. The conditional maximization step of the ECM algorithm consists of solving several, potentially simpler, conditional maximization problems. These conditional maximization problems consist of maximizing subsets of the parameters conditional on the other parameters.

The algorithm is outlined below as a one-step procedure. However, in the Monte Carlo simulation study and empirical application we use an adaptive version of the algorithm, because of the improved asymptotic properties of adaptive procedures, see Zou (2006). The adaptive version of the algorithm uses as first-step estimates, to determine the adaptive weights, the estimates from a first pass of the algorithm.

3.1 ‖ Coordinate Descent Proximal ECM Algorithm

As discussed previously, we will focus on the optimization of the penalized log-likelihood. Hence, the objective function of interest is, considering heterogeneous spatial dependence, as follows,

$$\begin{aligned}\mathcal{L}(\theta) = &-\frac{NT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\log|F_t| + v_t' F_t^{-1} v_t\right) \\ &- \sum_{i=1}^{N}\gamma_{\rho,i}|\rho_i| - \sum_{k=1}^{N}\gamma_{\beta,k}|\beta_k| - \sum_{i=1}^{N}\sum_{j=1}^{r}\gamma^{l}_{\Lambda,ij}|\Lambda_{ij}| - \sum_{j=1}^{r}\sqrt{N}\gamma^{gl}_{\Lambda,j}\|\Lambda_{\bullet j}\|_2\,,\end{aligned} \tag{2}$$

where $\Lambda_{\bullet j}$ denotes the j-th column of Λ. The amount of penalization, and hence, sparsity in the solution is controlled by $\gamma_{\rho,i}$, $\gamma_{\beta,k}$, $\gamma^{l}_{\Lambda,ij}$, and $\gamma^{gl}_{\Lambda,j}$. Optimizing this non-smooth objective function with a numerical score algorithm is very cumbersome, hence, we opt for the ECM algorithm. In order to introduce our ECM algorithm we have to

introduce the penalized complete data log-likelihood, which is given by

$$\begin{aligned}\tilde{\mathcal{L}}(\theta) = &- \frac{r(T-1)}{2}\log(2\pi) - \frac{T-1}{2}\log|\Sigma_\eta| - \frac{1}{2}\sum_{t=2}^{T}(f_t - \phi f_{t-1})'\,\Sigma_\eta^{-1}\,(f_t - \phi f_{t-1}) \\ &- \frac{NT}{2}\log(2\pi) + T\log|I - PW| \\ &- \frac{T}{2}\log|\Sigma_\varepsilon| - \frac{1}{2}\sum_{t=1}^{T}(y_t - X_t\beta - \Lambda f_t)'\,\Sigma_\xi^{-1}\,(y_t - X_t\beta - \Lambda f_t) \\ &- \sum_{i=1}^{N}\gamma_{\rho,i}|\rho_i| - \sum_{k=1}^{N}\gamma_{\beta,k}|\beta_k| - \sum_{i=1}^{N}\sum_{j=1}^{r}\gamma^{l}_{\Lambda,ij}|\Lambda_{ij}| - \sum_{j=1}^{r}\sqrt{N}\gamma^{gl}_{\Lambda,j}\|\Lambda_{\bullet j}\|_2\,,\end{aligned} \tag{3}$$

with $\Sigma_\xi^{-1} = (I - PW)'\Sigma_\varepsilon^{-1}(I - PW)$.

3.1.1 ‖ *Expectation Step*

The conditional expectation of this complete data likelihood w.r.t. the observed series $y_1, \ldots, y_T$ and current parameter estimates is denoted by $Q(\theta^{(k+1)}|\theta^{(k)})$. Before we are able to write down the conditional expectation of the likelihood we introduce some notation. Let the conditional mean be denoted by

$$\hat{f}_t = \mathbb{E}[f_t \,|\, y_1, \ldots, y_T]\,,$$

similarly we define the covariance functions as

$$V_t = \mathrm{Var}(f_t \,|\, y_1, \ldots, y_T)$$

and

$$V_{t,t+h} = \mathrm{Cov}(f_t, f_{t+h} \,|\, y_1, \ldots, y_T)\,,$$

where we dropped the conditioning on the current parameter values $\theta^{(k)}$. Hence, $\hat{f}_t$ is the smoothed estimate and V_t the corresponding covariance matrix, which are provided by the Kalman filter and smoother recursions. $V_{t,t+h}$ is the smoothed autocovariance matrix at lag h, which is provided by additional smoother recursions, see Appendix B.

The expectation step of the ECM algorithm consists of taking conditional expecta-

tions in (3), yielding

$$
\begin{aligned}
Q\left(\theta^{(k+1)} \,|\, \theta^{(k)}\right) = & -\frac{r(T-1)}{2}\log(2\pi) - \frac{NT}{2}\log(2\pi) - \frac{T-1}{2}\log|\Sigma_\eta| - \frac{T}{2}\log|\Sigma_\varepsilon| \\
& + T\log|I - PW| - \frac{1}{2}\operatorname{tr}\left(\Sigma_\eta^{-1}\left[V_0^* - V_{0,-1}^*\phi' - \phi(V_{0,-1}^*)' + \phi V_{-1}^*\phi'\right]\right) \\
& - \frac{1}{2}\operatorname{tr}\left(\Sigma_\xi^{-1}\sum_{t=1}^{T}\left[(y_t - X_t\beta - \Lambda\hat{f}_t)(y_t - X_t\beta - \Lambda\hat{f}_t)' + \Lambda V_t\Lambda'\right]\right) \\
& - \sum_{i=1}^{N}\gamma_{\rho,i}|\rho_i| - \sum_{k=1}^{N}\gamma_{\beta,k}|\beta_k| - \sum_{i=1}^{N}\sum_{j=1}^{r}\gamma_{\Lambda,ij}^{l}|\Lambda_{ij}| \\
& - \sum_{j=1}^{r}\sqrt{N}\gamma_{\Lambda,j}^{gl}\|\Lambda_{\bullet j}\|_2\,,
\end{aligned}
\tag{4}
$$

where tr denotes the trace,

$$V_{-1}^* = \sum_{t=2}^{T} V_{t-1} + \hat{f}_{t-1}\hat{f}_{t-1}'\,,$$

$$V_{0,-1}^* = \sum_{t=2}^{T} V_{t,t-1} + \hat{f}_t\hat{f}_{t-1}'\,,$$

$$V_0^* = \sum_{t=2}^{T} V_t + \hat{f}_t\hat{f}_t'\,,$$

and the superscript $(k+1)$ of the parameters is dropped for brevity. The Kalman smoother estimates are calculated conditional on the current parameter values $\theta^{(k)}$.

3.1.2 ‖ *Conditional Maximization Step*

The maximization step of the EM algorithm consists now of maximizing $Q\left(\theta^{(k+1)} \,|\, \theta^{(k)}\right)$ w.r.t. to the parameter vector $\theta^{(k+1)}$. In our ECM algorithm we solve several simpler conditional optimization problems, more specifically we optimize $Q\left(\theta_s^{(k+1)} \,|\, \theta^{(k)}, \theta_{-s}^{(k+1)}\right)$ w.r.t. to the subset parameter vector $\theta_s^{(k+1)}$. Here $\theta_s^{(k+1)}$ corresponds to each of the distinct model components captured in the parameter vector θ and $\theta_{-s}^{(k+1)}$ captures all parameters expect those in subset s.

The benefits of using the ECM algorithm become clear at this stage. Due to the ECM algorithm we are able to exploit the linear nature of the model efficiently in this conditional maximization step. First of all, because we observe that the two parts of the

likelihood (one corresponding to the state equation and the other to the measurement equation) are separated we can simply optimize them independently. So, we get

$$\phi^{(k+1)} = V^*_{0,-1}(V^*_{-1})^{-1} \quad \text{and} \quad \Sigma^{(k+1)}_{\eta} = \frac{1}{T-1}(V^*_0 - V^*_{0,-1}(V^*_{-1})^{-1}(V^*_{0,-1})')\,. \tag{5}$$

As we assumed that ϕ and Σ_η are diagonal matrices we have that optimizing the likelihood w.r.t. ϕ and Σ_η boils down to maximizing a quadratic scalar function. Hence, (5) simplifies to

$$\phi^{(k+1)}_{ii} = V^*_{0,-1,ii}(V^*_{-1,ii})^{-1} \quad \text{and} \quad \sigma^{(k+1)}_{\eta,ii} = \frac{1}{T-1}(V^*_{0,ii} - (V^*_{0,-1,ii})^2/V^*_{-1,ii})\,. \tag{6}$$

Another convenient aspect of the ECM algorithm is that we can now directly recognize a linear regression structure in the part of the likelihood originating from the measurement equation, more specifically a generalized ridge regression. This is convenient as the penalty term only relates to the parameters of the measurement equation and efficient algorithms exist for solving a ℓ_1-norm combined with an ℓ_2-norm penalized linear regression framework, i.e. sparse group Lasso, such as the blockwise proximal gradient descent algorithm of Simon et al. (2013).

Similar in spirit to the estimation of the penalized loading matrix, we can exploit a penalized linear regression structure for the slope coefficients β. The coordinate descent update formula for the slope coefficients is a multiple response equivalent of the coordinate descent steps outlined in Friedman et al. (2007) and Friedman et al. (2010) and is derived, for heterogeneous slopes, in Lee and Liu (2012) and Schnücker (2017). The update steps for the homogeneous case have the following form

$$\tilde{\beta}_k \leftarrow S(\bar{\beta}_k, \tilde{\gamma}_{\beta,k})\,,$$

with

$$\tilde{\gamma}_{\beta,k} = \frac{\gamma_{\beta,k}}{\sum_{i=1}^N \omega_{ii} X'_{ik} X_{ik}}$$

and

$$\bar{\beta}_k = \frac{\sum_{i=1}^N \sum_{j=i}^N \omega_{ij} X'_{ik} e_j}{\sum_{i=1}^N \omega_{ii} X'_{ik} X_{ik}} + \beta_k\,,$$

where $e_j = (e_{j1}, \ldots, e_{jT})'$ with $e_{jt} = y^{**}_{jt} - X_{jt}\beta_j$ and $y^{**}_{jt} = y_{jt} - \sum_{q=1}^r \Lambda_{jq} f_{qt}$.

The difficult part of the optimization procedure is maximizing the objective function $Q(\cdot|\cdot)$ w.r.t. the spatial dependence parameters ρ. The maximizer is given implicitly by

$$\begin{aligned} \rho^{(k+1)} &= \arg\min_{\rho} Q_\rho(\rho) \\ &= \arg\min_{\rho} - T\log|I - PW| + \frac{1}{2}\operatorname{tr}\left((I-PW)'\Sigma_\varepsilon^{-1}(I-PW)D\right) \\ &\qquad + \sum_{i=1}^N \gamma_{\rho,i}|\rho_i| \end{aligned} \tag{7}$$

with $P = \text{diag}(\rho)$ and

$$D = \sum_{t=1}^{T} \left[(y_t - X_t\beta - \Lambda \hat{f}_t)(y_t - X_t\beta - \Lambda \hat{f}_t)' + \Lambda V_t \Lambda' \right] .$$

Deriving closed form solutions for the maximizer is not possible and, hence, we have to rely on numerical optimization methods. An extra difficulty arises as the spatial dependence parameter is penalized. Therefore, we rely on a proximal gradient descent algorithm (see e.g. Parikh and Boyd, 2014) to solve (7).

Finally, maximizing the objective function $Q(\cdot|\cdot)$ w.r.t. the covariance matrix, Σ_ε, of the measurement equation yields

$$\Sigma_\varepsilon^{(k+1)} = \frac{1}{T}(I - PW)D(I - PW)', \tag{8}$$

or when Σ_ε is assumed diagonal

$$\sigma_{\varepsilon,ii}^{(k+1)} = \frac{1}{T}\left[(I - PW)D(I - PW)'\right]_{ii} . \tag{9}$$

3.1.3 ‖ *Algorithm*

Combining all parts, as discussed above, the coordinate descent proximal ECM (CD-PECM) algorithm becomes as reported in Algorithm 1.

4 ‖ Simulation Study

4.1 ‖ Simulation design

In this section, we investigate the performance of the Coordinate Descent Proximal ECM Algorithm in terms of estimation and model selection accuracy, in a variety of settings. In particular, we simulate from a dynamic factor model with spatially dependent errors[2], which is given by

$$y_t = \Lambda f_t + \xi_t , \quad f_{t+1} = \phi f_t + \eta_t , \quad \xi_t = PW\xi_t + \varepsilon_t. \tag{10}$$

The autoregressive parameter matrix ϕ is set equal to $0.9 \cdot \text{I}_r$, where r denotes the number of factors. The error terms η_t and ϵ_t are normally distributed with zero mean. The errors of the dynamic factors are set to have a covariance matrix equal to the identity matrix in order to solve rotational identification. For the idiosyncratic errors we use $\Sigma_\varepsilon = \text{diag}(\sigma_{\varepsilon,11}, \ldots, \sigma_{\varepsilon,NN})$ with $\sigma_{\varepsilon,ii} = \mathcal{U}(1/2, 1)$. The matrix of spatial dependence parameters is either given by $P = \rho \cdot \text{I}_N$ with $\rho = 0.4$ (low spillover intensity) or $\rho = 0.9$ (high spillover intensity) or it exhibits a two-group structure. In the latter

[2]For simplicity, we abstract from including exogenous regressors in the simulation study.

Algorithm 1: Coordinate Descent Proximal ECM Algorithm for Sparse Estimation

Input: Tuning parameters $\gamma^{l}_{\Lambda,ij} = \gamma^{l,*}_{\Lambda,ij}$, $\gamma^{gl}_{\Lambda,ij} = \gamma^{gl,*}_{\Lambda,ij}$, $\gamma_{\beta,k} = \gamma^{*}_{\beta,k}$, and $\gamma_{\rho,i} = \gamma^{*}_{\rho,i}$ and initial parameter vector $\theta^{(0)}$.

Iterate until convergence: For iteration l.

The E-step

- Run the Kalman filter and smoother recursions to obtain $\hat{f}_t^{(l)}$, $V_t^{(l)}$, and $V_{t,t-1}^{(l)}$ (based on the estimates of iteration $l-1$).
- Calculate $V_{-1}^{*,(l)}$, $V_{0,-1}^{*,(l)}$, and $V_0^{*,(l)}$.

The CM-step

- Estimate $\phi^{(l)}$ and $\Sigma_\eta^{(l)}$ according to either (5) or (6).
- Estimate $\Lambda^{(l)}$, $\beta^{(l)}$, $\Sigma_\varepsilon^{(l)}$, and $\rho^{(l)}$ as follows:
  - Given the current estimates $\beta^{(l-1)}$, $\Sigma_\varepsilon^{(l-1)}$ and $\rho^{(l-1)}$, estimate $\Lambda^{(l)}$ as follows:

    Iterate until convergence: For iteration n.
    - Cycle through columns $j \in \{1, \ldots, r\}$ and update $\Lambda_{\bullet j}^{(l,n)}$ using the block-wise proximal gradient descent algorithm.
  - Similarly, given the current estimates $\Lambda^{(l)}$, $\Sigma_\varepsilon^{(l-1)}$ and $\rho^{(l-1)}$, estimate $\beta^{(l)}$ as follows:

    Iterate until convergence: For iteration n.
    - Cycle through $k \in \{1, \ldots, K\}$ and update $\beta_k^{(l,n)}$ according to (**??**).
  - Next, given the current estimates $\Lambda^{(l)}$, $\beta^{(l)}$ and $\rho^{(l-1)}$ update $\Sigma_\varepsilon^{(l)}$ according to either (8) or (9).
  - Finally, given the current estimates $\Lambda^{(l)}$, $\beta^{(l)}$ and $\Sigma_\varepsilon^{(l)}$ estimate $\rho^{(l)}$ according to (7) using the proximal gradient descent algorithm.

case, we consider two options as well: $\rho_1 = 0.4$ for $i = 1, ..., N/2$ and $\rho_2 = 0.9$ for $i = N/2 + 1, ..., N$, or $\rho_1 = 0.4$ for half the units and $\rho_2 = 0$ for the other half. In all cases, the spatial weight matrix corresponds to the case of two-way spatial effects for direct neighbors in space, assuming a spherical world. For the matrix of factor loadings, we use

$$\Lambda_{ij} = \begin{cases} 0 & \text{if} \begin{cases} i < (\#_{nz} - \#_o) \cdot (j-1) + 1 \\ \text{or} \quad (i > (\#_{nz} - \#_o) \cdot j + \#_0 \text{ and } j < r)\,, \end{cases} \\ \mathcal{U}(1/5, 1) & \text{otherwise}\,, \end{cases}$$

where $\#_{nz}$ is the number of non-zero loadings and $\#_o$ the number of overlap in the loadings between neighboring factors.

The considered sample sizes are $N = 20$ or $N = 50$ in the cross-section and $T \in \{200, 500, 1000\}$. Furthermore, the true number of factors r is either set to 2 or to 4, while the maximum number of included factors is always 5. Together with the four options for the structure of P in (10), this results in 48 different simulation settings.

Throughout the simulation study, we use 1000 replications. The optimal tuning parameters are identified using a 7-dimensional grid search, γ_ρ, γ_ρ^{adp}, γ_Λ^l, $\gamma_\Lambda^{l,\text{adp}}$, γ_Λ^{gl}, $\gamma_\Lambda^{gl,\text{adp}}$, and τ (adaptive Lasso weight), combined with the generalized information criteria (GIC) of Fan and Tang (2013). Due to the high computational cost, this 7-dimensional grid search is employed for the first 50 simulations, and for the remaining simulations, the average value of the previously found tuning parameters is used. The loading matrix is initialized using sparse PCA (Zou et al., 2006), the autoregressive parameter is initialized using the factors obtained from a sparse PCA analysis, the spatial dependence parameters is initialized as the OLS estimate based on the errors of the sparse PCA analysis, Σ_ε is initialized with the variance of the errors of the spatial dependence parameter regression.

4.2 ‖ Simulation results

We use the Coordinate Descent Proximal ECM Algorithm to select the sparsity pattern and to estimate the nonzero static parameters of the model. In particular, the set of penalized coefficients includes the entries of the loading matrix and the spatial dependence parameter(s). The autoregressive coefficients in the factor transition equation and the diagonal elements of the covariance matrix of the idiosyncratic errors are estimated within the CM-step of the algorithm without penalization.

Tables I and II present the simulation outcomes for the 48 settings we consider. We distinguish between homogeneous (Table I) and grouped (Table II) spatial intensity parameters. Estimation precision is measured using the average Frobenius norm of

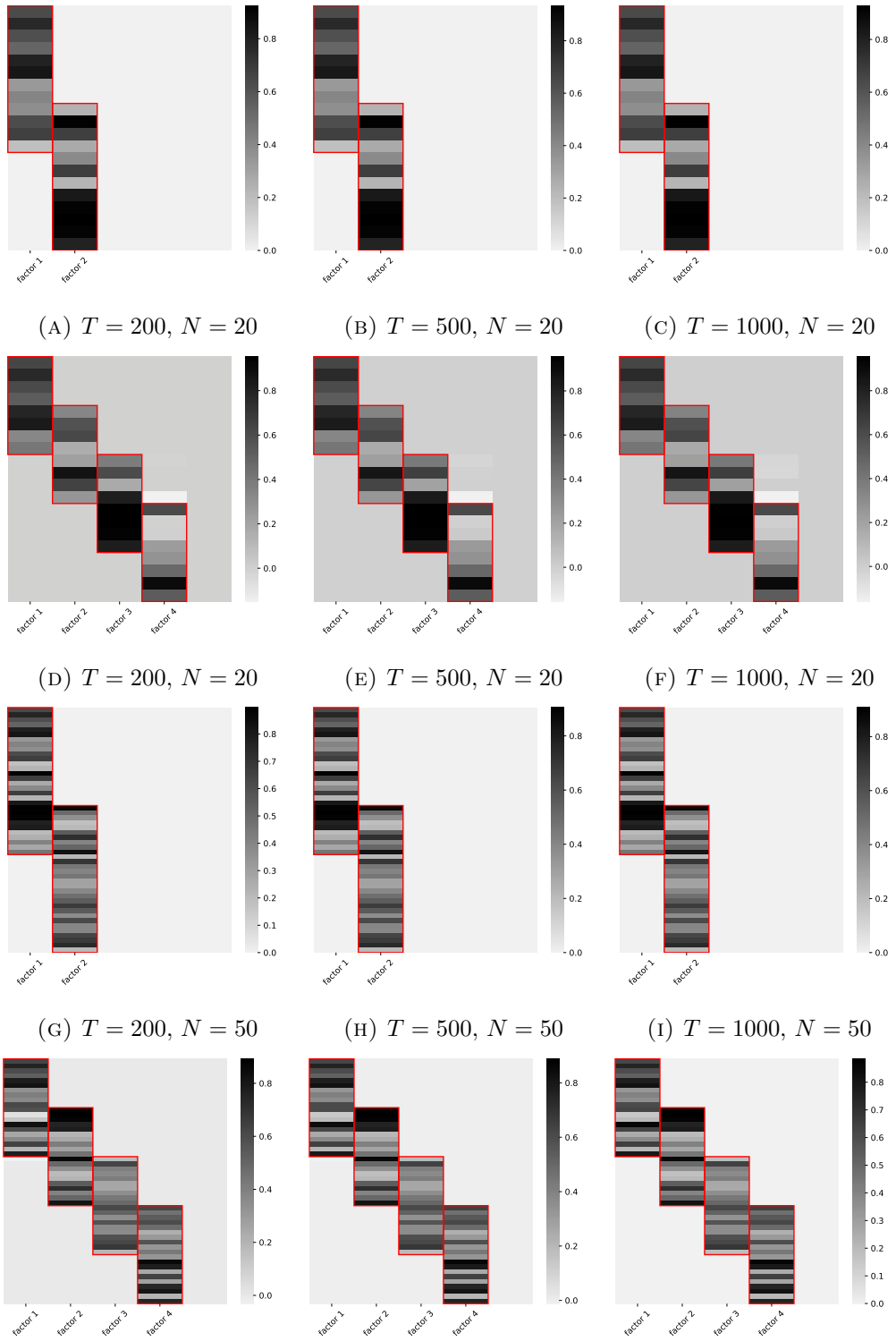
the difference between estimated and true loading matrix. For the spatial dependence parameter, we use the root mean squared error (RMSE). To illustrate the model selection performance of our algorithm, we also report the fraction of correctly identified zeros in the loading matrix, averaged across simulation runs, as well as the fraction of cases in which the number of factors is chosen correctly.

We find that our approach performs well. As expected, estimation becomes more precise and model selection improves as N and T increase, which is apparent from the decreases in Frobenius norms for Λ and the decreasing RMSEs for ρ, as well as the increases in the fraction of correctly identified zero coefficients. We also observe that estimation becomes more challenging the more factors are present, and the higher the spatial dependence parameters. This is to be expected, as for a given sample size, more factors imply a higher number of nonzero loading coefficients, which leads to more estimation uncertainty. Furthermore, if spatial dependence is high, it is harder to disentangle global and local cross-sectional dependence. However, even in the most challenging settings with a large number of factors ($r = 4$) and high spillover intensity ($\rho = 0.9$), the fraction of correctly identified zero coefficients equals 75% for the smallest sample size ($N = 20$ and $T = 200$), and it goes up to 91% for $N = 50$ and $T = 1000$. In contrast, when the spillover intensity is low, the method's model selection ability is almost perfect even in small sample sizes and when $r = 4$.

The overall patterns are similar for the case of grouped spatial dependence parameters (Table II). As before, including some units exhibiting high spillover intensity leads to higher estimation uncertainty and lower fractions of correct zeros. However, the performance improves quickly as sample sizes increase. Furthermore, we find that the presence of different values of ρ does not impair model selection or estimation results. In particular, the method has no problem identifying when there is no weak dependence for some units, and consequently setting ρ_2 to zero.

To provide more insight into the different simulation settings and the corresponding performance in terms of model selection, we also show some heatmaps of the estimated loading matrices in Figure I.[3] Each subfigure refers to the element-wise median outcome across simulation runs in one of our settings. Bars correspond to entries in the loading matrix – the darker the shading, the larger the coefficient, implying that light areas match the parts of the loading matrices with coefficients set to zero. The red boxes mark the areas which are truly nonzero. We observe that for all the sample sizes, the method correctly identifies the number of factors by setting the irrelevant parts of the loadings to zero. Furthermore, on average, the correct sparsity patterns within columns are found in almost all cases.

[3]Here, we show the plots for homogeneous spatial intensity with $\rho = 0.4$. The plots for the other settings look very similar and are available upon request.



(A) $T = 200$, $N = 20$ (B) $T = 500$, $N = 20$ (C) $T = 1000$, $N = 20$

(D) $T = 200$, $N = 20$ (E) $T = 500$, $N = 20$ (F) $T = 1000$, $N = 20$

(G) $T = 200$, $N = 50$ (H) $T = 500$, $N = 50$ (I) $T = 1000$, $N = 50$

(J) $T = 200$, $N = 50$ (K) $T = 500$, $N = 50$ (L) $T = 1000$, $N = 50$

FIGURE I

HEATMAP OF MEDIAN MONTE CARLO SIMULATION RESULTS OF Λ FOR $T = 200, 500, 1000$ AND $N = 20, 50$, FOR THE CASE OF LOW SPATIAL SPILLOVER INTENSITY ($\rho = 0.4$).

TABLE I

SIMULATION OUTCOMES: HOMOGENEOUS SPATIAL INTENSITY

		low spillover intensity ($\rho = 0.4$)				high spillover intensity ($\rho = 0.9$)			
		Frob. norm	% correct zeros	% correct factors	RMSE	Frob. norm	% correct zeros	% correct factors	RMSE
					$r = 2$				
	$T = 200$	0.0155	0.9849	0.9850	0.0155	0.1185	0.8379	0.6070	0.0061
$N = 20$	$T = 500$	0.0108	0.9986	0.9990	0.0102	0.0900	0.8920	0.7790	0.0047
	$T = 1000$	0.0091	1.0000	1.0000	0.0073	0.0706	0.9111	0.8440	0.0040
	$T = 200$	0.0103	0.9862	0.9890	0.0092	0.0609	0.9456	0.9600	0.0063
$N = 50$	$T = 500$	0.0080	0.9991	0.9990	0.0057	0.0575	0.9657	0.9590	0.0055
	$T = 1000$	0.0072	1.0000	1.0000	0.0041	0.0565	0.9725	0.9770	0.0054
					$r = 4$				
	$T = 200$	0.0362	0.9127	0.9930	0.0212	0.1234	0.7535	0.8470	0.0070
$N = 20$	$T = 500$	0.0307	0.9553	0.9960	0.0149	0.0939	0.8304	0.8480	0.0055
	$T = 1000$	0.0297	0.9597	0.9860	0.0119	0.0834	0.8632	0.8270	0.0049
	$T = 200$	0.0242	0.9246	0.9800	0.0111	0.0834	0.8209	0.7040	0.0069
$N = 50$	$T = 500$	0.0145	0.9823	0.9950	0.0069	0.0771	0.8869	0.8330	0.0064
	$T = 1000$	0.0127	0.9920	1.0000	0.0055	0.0733	0.9109	0.9080	0.0064

Table II

Simulation outcomes: Grouped spatial intensities

		grouped spillover intensity ($\rho_1 = 0.4$, $\rho_2 = 0.9$)					grouped spillover intensity ($\rho_1 = 0.4$, $\rho_2 = 0$)				
		Frob. norm	% correct zeros	% correct factors	RMSE (ρ_1)	RMSE (ρ_2)	Frob. norm	% correct zeros	% correct factors	RMSE (ρ_1)	RMSE (ρ_2)
						$r = 2$					
	$T = 200$	0.0631	0.9255	0.7480	0.0221	0.0099	0.0162	0.9858	0.9930	0.0263	0.0247
$N = 20$	$T = 500$	0.0637	0.9441	0.6990	0.0145	0.0071	0.0116	0.9991	1.0000	0.0147	0.0000
	$T = 1000$	0.0684	0.9472	0.6670	0.0105	0.0064	0.0103	1.0000	1.0000	0.0105	0.0000
	$T = 200$	0.0428	0.9636	0.9810	0.0127	0.0075	0.0117	0.9919	0.9990	0.0131	0.0030
$N = 50$	$T = 500$	0.0424	0.9845	0.9940	0.0080	0.0071	0.0097	0.9997	1.0000	0.0082	0.0007
	$T = 1000$	0.0426	0.9894	1.0000	0.0058	0.0072	0.0091	1.0000	1.0000	0.0060	0.0000
						$r = 4$					
	$T = 200$	0.0521	0.8483	0.9080	0.0256	0.0086	0.0350	0.9243	0.9970	0.0276	0.0024
$N = 20$	$T = 500$	0.0449	0.9275	0.9440	0.0170	0.0065	0.0302	0.9612	0.9920	0.0181	0.0000
	$T = 1000$	0.0434	0.9529	0.9590	0.0140	0.0056	0.0288	0.9645	0.9930	0.0141	0.0000
	$T = 200$	0.0541	0.8610	0.8010	0.0247	0.0084	0.0232	0.9398	0.9950	0.0150	0.0080
$N = 50$	$T = 500$	0.0488	0.9374	0.8840	0.0190	0.0076	0.0172	0.9884	0.9930	0.0092	0.0024
	$T = 1000$	0.0461	0.9588	0.9310	0.0154	0.0073	0.0157	0.9970	1.0000	0.0069	0.0009

5 ‖ Empirics

5.1 ‖ Data

We use monthly constant maturity zero-coupon Treasury yield data from the data set of Liu and Wu (2020). Using a nonparametric kernel smoothing method introduced in Linton et al. (2001), they extract yield data for maturities between 1 and up to 360 months, with starting dates for the low maturity time series going back to the 1960s. Thus, accurate data on both the short and the (very) long end of the yield curve are available, which is in contrast to other widely used yield curve data sets, such as Fama and Bliss (1987) and Gürkaynak et al. (2007).

To strike a balance between data availability in the time and cross-sectional dimensions, we choose July 1981 as starting month. From this time onward, yield data with maturities from 1 to 240 months are available. In particular, we use a subset of 15 time series in our empirical analysis: Maturities of 1, 3, 6, and 9 months capture the short end of the yield curve; maturities of 12, 24, 36, 48, and 60 months form the medium range, and maturities of 84, 120, 150, 180, 210, and 240 months are our group of long-term interest rates. In total our sample contains 462 monthly observations, covering July 1981 until December 2019.

In our empirical analysis below, we allow for heterogeneous spatial dependence parameters across the three groups of maturities. The weight matrix has a simple form in which entries are equal to one for direct neighbors, and zero otherwise, which implies a two-sided spatial AR(1) process if the parameter ρ is non-zero. In this framework, shocks to a particular maturity or maturity group can spill over to other maturities and, eventually throughout the entire yield curve. This possibility may be seen as an empirical approximation of the arguments put forward by Vayanos and Vila (2009) in their theoretical "preferred habitat model". Local spillovers between neighboring maturities may account for investors who target certain groups of maturities, such as pension funds, as well as arbitrageurs, who see to the shock transmission. In a recent paper, Crump and Gospodinov (2019) also propose a spatial AR(1) process for excess bond returns of neighboring maturities.

Macroeconomic variables have been found to add useful information to yield curve prediction models, see, for instance Ludvigson and Ng (2009), Coroneo et al. (2016), and Bianchi et al. (2020). These findings are in contrast to the "spanning hypothesis" stating that all information on the future yield curve are included in current yield data. To account for this possibility, we include production growth and the federal funds rate as regressors. These are the two macroeconomic variables that were found to be most informative in the study of Coroneo et al. (2016). The data are extracted from the

FRED-MD data base and transformed in the way suggested by McCracken and Ng (2016).

5.2 ‖ Global and local cross-sectional dependence in the term structure of interest rates

5.2.1 ‖ *Full-sample estimation*

We estimate the high-dimensional dynamic factor model with spatial errors for the 15 time series of monthly yields for the full sample spanning 38.5 years. We allow for heterogeneous spatial dependence in the errors for three groups of maturities (short, medium, and long), and we include the two macroeconomic variables IP growth and federal funds rate.

Using our estimation and model selection procedure[4], we find that five factors are needed to capture the comovements of yields for the full sample. Figure II shows plots of the five smoothed factors, as well as heatmaps of the estimated factor loadings. We observe that the loadings on the first factor are very similar across maturities, clearly suggesting an interpretation as level of the yield curve. For the second factor, loadings are particularly high at the short end of the yield curve and declining towards the medium term, which is consistent with the interpretation of the second factor as slope. However, the loadings increase again as maturities become longer. The third factor appears to capture comovements of the very short and medium maturities, while the fourth factor loads on the medium-term yields and the long end. Finally, the presence of the fifth factor seems necessary to describe the dynamics of the very long maturity yields.

Besides the rich factor structure, we find strong evidence for heterogeneous local dependence in the error terms. The group of short-term yields shows particularly high spillover dependence. On the other hand, the two macroeconomic factors do not appear to have an impact in the full sample. Their coefficients are set to zero by the Lasso. Table III lists the estimated coefficients.

Table III

Full sample coefficient estimates

β_{IP}	β_{fed}	ρ_{short}	ρ_{medium}	ρ_{long}	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
0.0	0.0	0.9817	0.7427	0.6662	0.9870	0.9771	0.8815	0.8772	0.9350

[4]Initialization and model selection are identical to the ones used for the Monte Carlo simulation study as described in Section 4.

5.2.2 ‖ *Rolling window analysis*

We also conduct a dynamic investigation of the estimated model coefficients, the number of factors, and the strength of local spillovers. We use rolling windows of 15 years (180 monthly observations), and then re-estimate the model once a year, resulting in 24 subsamples. It is well-established in the literature that the yield curve is driven by at least three factors, see Diebold and Li (2006), Härdle and Majer (2016), and many others. However, the appropriate number of factors at all time points has been a point of debate, see, for instance, Crump and Gospodinov (2019). Indeed, looking at the heatmaps in Figure III, we observe time variation in the structure of the loading matrices, and, in particular, also in the number of included factors, i.e. nonzero columns.

As in the full-sample case, we also find evidence for heterogeneous local spillover intensities, which are measured by the rolling window estimates of the group-specific spatial parameters ρ shown in panel (a) of Figure IV. More specifically, the plot reveals that throughout the sample, the coefficient is estimated to be at the boundary of 1 within the group of short maturities. For the medium and long-term maturities, on the other hand, we see an increase over time until they are close to 1 as well. This suggests that spillovers among neighboring maturities, that go beyond what is captured by common factors, have become more and more important. The finding is confirmed when we estimate a pooled version of the model with homogeneous coefficients, which is shown in panel (b) of Figure IV.

Finally, we find that the two macroeconomic variables are de-selected from the model in all sub-samples up to the end of 2009. After the financial crisis, IP growth enters the model occasionally with a positive sign. The coefficient corresponding to the fed funds rate, on the other hand, is set to zero in all but five subsamples, in which it shows a negative sign. From these results, we conclude that, after accounting for complex cross-sectional dependencies, macroeconomic predictors play only a minor role for modeling the yield curve, confirming the observations made in Bauer and Hamilton (2018).

5.3 ‖ Out-of-Sample Performance

To assess the out-of-sample performance of our method, we again use rolling windows of 15 years (180 observations), and re-estimate our high-dimensional dynamic factor model with spatial errors and produce 1-month ahead forecasts in each of the resulting 283 subsamples, starting in June 1996. Due to the computational complexity of the grid search, the tuning parameters are only updated once a year, resulting in 24 different sets of tuning parameters.

The forecasting performance of our model is compared to the widely used dynamic Nelson-Siegel model, which is estimated using the two-step approach of Diebold and Li

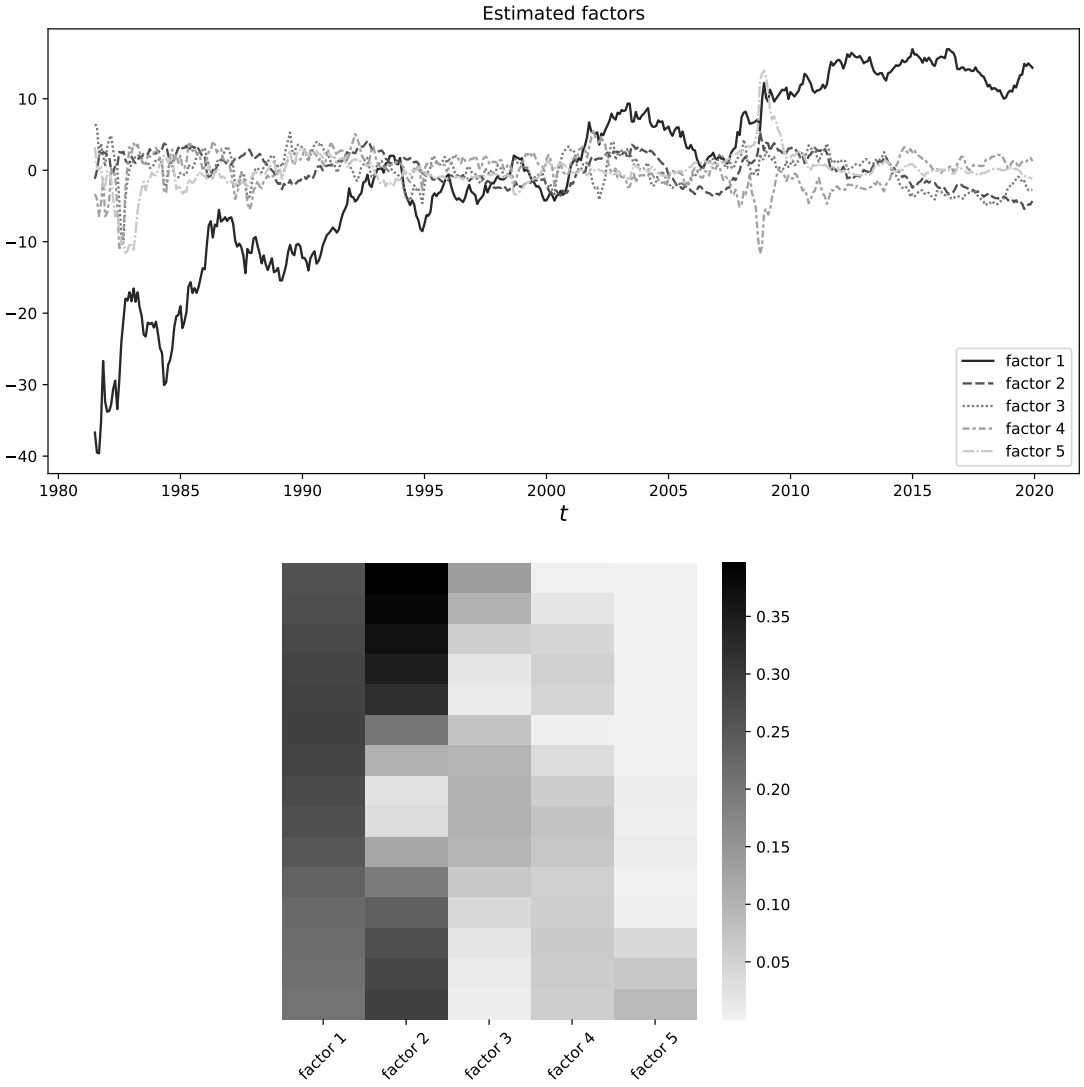
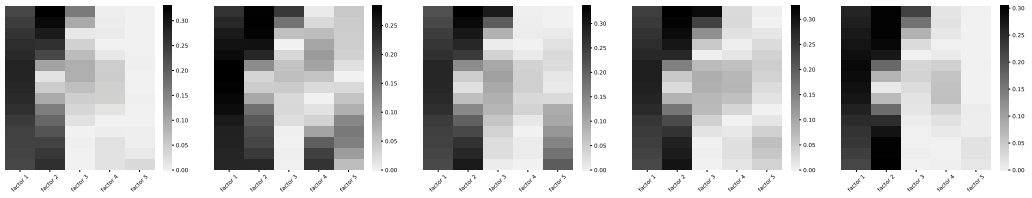
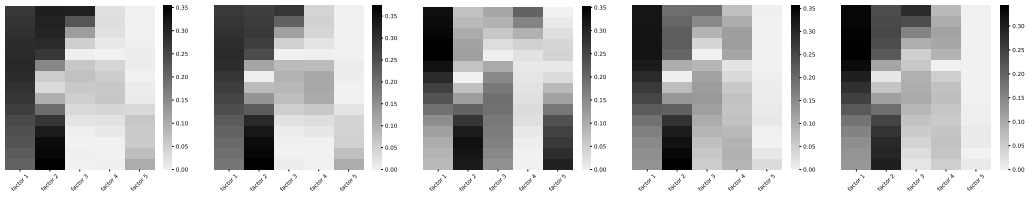


FIGURE II

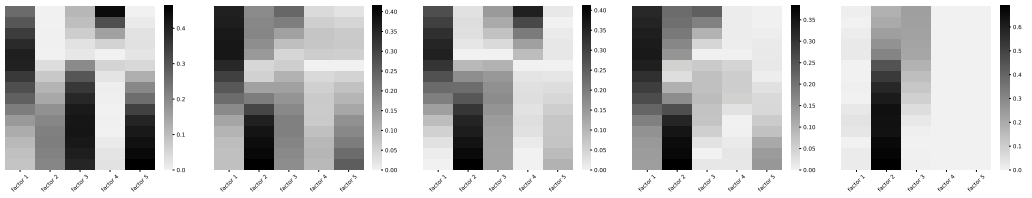
FULL SAMPLE SMOOTHED FACTORS AND CORRESPONDING HEATMAP OF LOADINGS



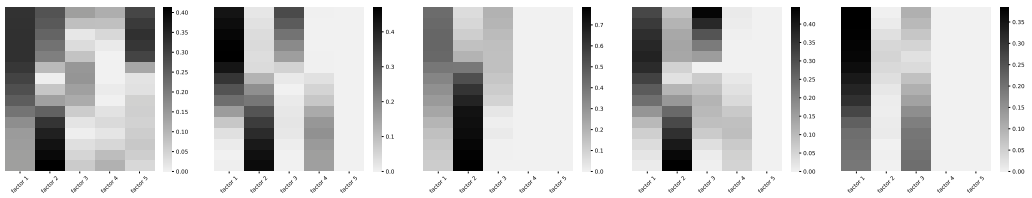
(A) $T = 1996$ (B) $T = 1997$ (C) $T = 1998$ (D) $T = 1999$ (E) $T = 2000$



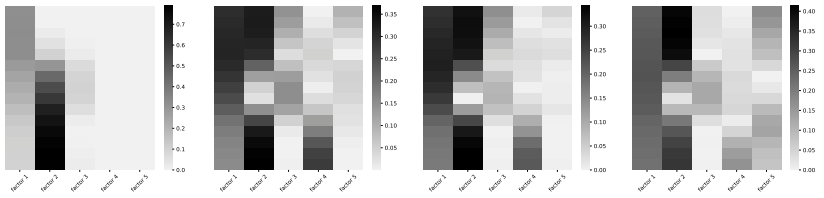
(F) $T = 2001$ (G) $T = 2002$ (H) $T = 2003$ (I) $T = 2004$ (J) $T = 2005$



(K) $T = 2006$ (L) $T = 2007$ (M) $T = 2008$ (N) $T = 2009$ (O) $T = 2010$



(P) $T = 2011$ (Q) $T = 2012$ (R) $T = 2013$ (S) $T = 2014$ (T) $T = 2015$

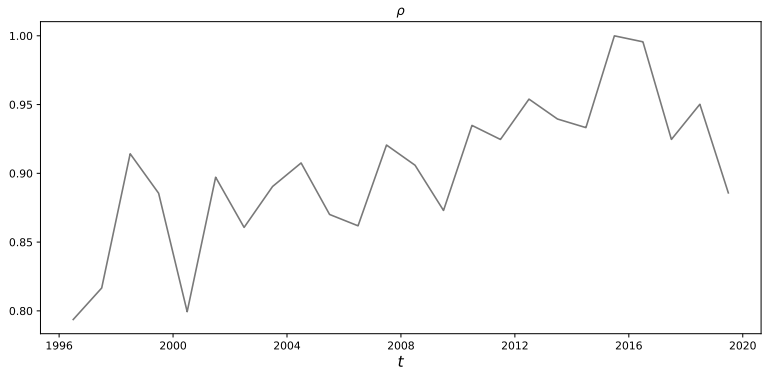


(U) $T = 2016$ (V) $T = 2017$ (W) $T = 2018$ (X) $T = 2019$

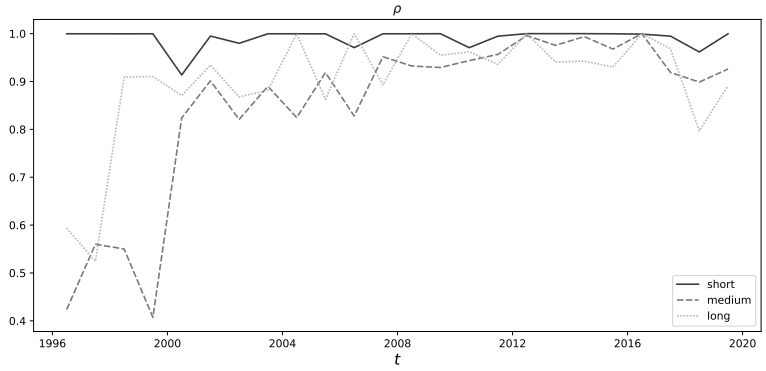
FIGURE III

HEATMAPS OF ROLLING WINDOW ESTIMATION RESULTS FOR LOADING MATRIX Λ USING ADAPTIVE LASSO

On the vertical axis the different maturities are shown, from top to bottom increasing in maturity. The rolling window estimation is performed every year is based on 180 observations. T denotes the end year of each window. Heterogeneous spatial dependence is taken into account during estmation.



(A) Homogeneous ρ



(B) Heterogeneous ρ

FIGURE IV

ROLLING WINDOW ESTIMATION RESULTS FOR SPATIAL DEPENDENCE PARAMETER ρ USING ADAPTIVE LASSO

(2006). We consider both the case of homogeneous and heterogeneous spatial dependence. As before, in the case of heterogeneous spatial dependence, we allow for three different groups: short-end with maturities of 1, 3, 6, and 9 months, medium range with maturities of 12, 24, 36, 48, and 60 months, and long-term with maturities of 84, 120, 150, 180, 210, and 240 months. In order to produce forecasts in the presence of the exogeneous regressors we assume a VAR(1) structure for industrial production and federal funds rate jointly.

TABLE IV

RELATIVE RMSFE

	Homogeneous ρ				Heterogeneous ρ			
	Full	1996-2006	2007-2009	2010-2019	Full	1996-2006	2007-2009	2010-2019
1	0.9933	1.2684*	1.6752*	0.7382	0.6481*	0.6961*	0.8433	0.5976*
3	1.0915	1.2699*	1.8398*	0.8833	0.6103*	0.6120*	0.8247	0.5728*
6	1.2441*	1.3362*	1.9740*	1.0863	0.6143*	0.5854*	0.7749	0.5973*
9	1.4147*	1.4263*	2.0167*	1.3109	0.6601*	0.6231*	0.7512	0.6563*
12	1.6148*	1.5171*	2.1016*	1.5633*	0.7519*	0.6843*	0.8005	0.7621*
24	2.0056*	1.6681*	2.2739*	2.0610*	1.0421	0.8433*	0.9966	1.1109
36	2.3439*	1.6199*	2.3089*	2.6370*	1.1532	0.8742*	1.0527	1.2899
48	2.5167*	1.5829*	2.3286*	2.9828*	1.2486	0.8939*	1.1216	1.4496
60	2.4563*	1.5715*	2.3620*	2.8828*	1.2941	0.9146*	1.1290	1.5116
84	2.3862*	1.5082*	1.8739*	2.8688*	1.3599	0.9197*	0.9245	1.6388*
120	2.1178*	1.4040*	1.0997	2.5641*	1.3229	0.8857*	0.6824*	1.6003*
150	1.9718*	1.3000*	0.8151	2.4392*	1.3400*	0.8565*	0.5877*	1.6588*
180	1.9234*	1.2195*	1.1102	2.2692*	1.4120*	0.8483*	0.7450	1.6831*
210	1.8383*	1.1488	1.2402	2.1281*	1.4430*	0.8555*	0.7750*	1.7003*
240	1.7289*	1.1263	1.2832	1.9593*	1.4376*	0.8785*	0.7883*	1.6683*

RMSFE for the homogeneous and heterogeneous ρ estimations are shown relative to the RMSFE for the dynamic Nelson-Siegel model for different maturities. The stars indicate significant difference in forecasting performance according to the Diebold-Mariano test at the 10% level.

Table IV shows the root mean squared forecast errors (RMSFE) of the homogeneous and heterogeneoeus high-dimensional dynamic factor model with spatial errors, relative to the dynamic Nelson-Siegel model. Besides the full forecasting period, we also consider three distinct sub-periods, including the time before the global financial crisis (1996–2006), the crisis years (2007–2009) and remaining years (2010–2019). We observe that restricting the spatial dependence to the scalar case leads to the dynamic Nelson-Siegel model of Diebold and Li (2006) outperforming our model for almost all maturities, independent of the sub-period. This is also confirmed by the Diebold-Mariano test, as an $*$ in Table IV indicates a significant difference in forecasting performance at the 10%

level.

In contrast, for the heterogeneous version of our model, we observe that, at the short end of the yield curve, we outperform the dynamic Nelson-Siegel model over the entire forecasting period. When it comes to the pre-crisis observations, our heterogeneous model even outperforms the dynamic Nelson-Siegel model for all maturities. During the financial crisis, on the other hand, our model seems to produce more accurate 1-month ahead forecasts for the long-end of the yield curve. In terms of the medium maturities, we seem to be mostly at par with the dynamic Nelson-Siegel model or outperform it slightly. Only in the post-crisis period, at the long-end of the yield curve, our model significantly underperforms. Looking at time series plots of the forecast errors shown in Figure V, it appears that this post-crisis underperformance of our heterogeneous model for the longer maturities seems to be related to a few erratic spikes in the forecast errors.[5]

Overall, the forecasting exercise confirms our in-sample finding, that it is not only important to allow for flexibility in the number of factors and structure of the loading matrices, but that it is also beneficial to account for heterogeneous local spillovers among neighboring maturities.

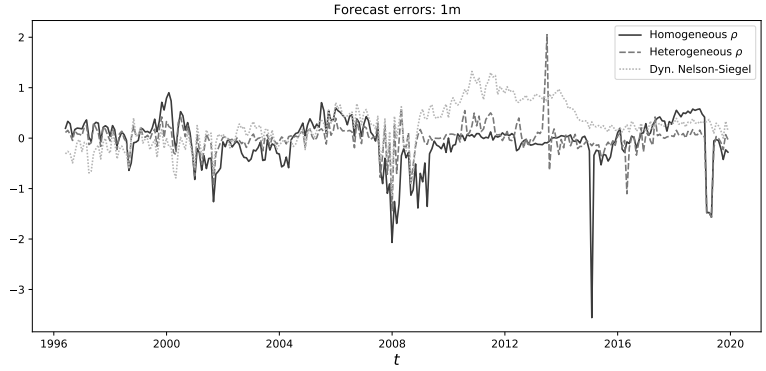
6 ‖ Conclusion

The paper introduced a unified method to estimate high-dimensional factor models with exogenous regressors and spatial error dependence. We provide an Expectation-Maximization type algorithm that allows us to maximize a penalized state space likelihood. Simulations show the good performance of the method in terms of estimation and model selection. In an empirical application, we estimate a factor model for the term structure of interest rates. A rolling window analysis suggests substantial variation in the number of factors and the magnitude of factor loadings. Beyond the common factors, we find evidence for maturity group-specific local dependence in the error terms, which becomes even stronger towards the end of our sample. In terms of out-of-sample performance, our method performs very well and shows improvements over the widely used dynamic Nelson-Siegel model.

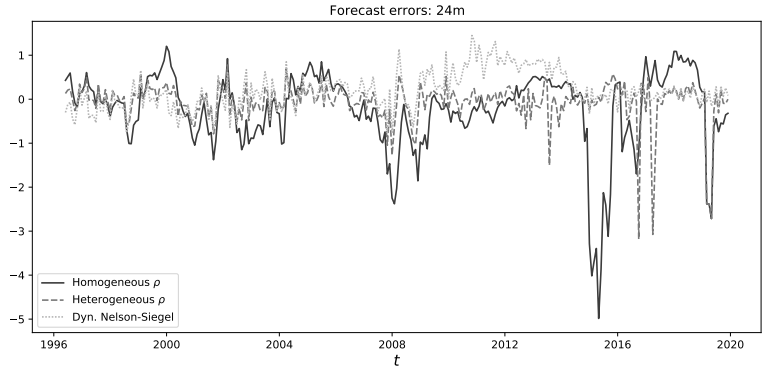
References

Anselin, L. (1988): *Spatial Econometrics: Methods and Models*, Springer.

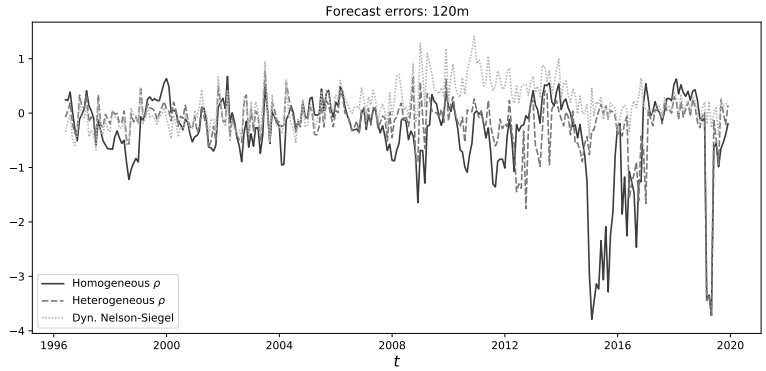
[5]These erratic spikes in the forecast errors may be due to the procedure with which we select the tuning parameters. The re-estimating of the tuning parameters each year is most likely not sufficient to capture the dynamics of the data well enough at the end of the sample.



(A) 1 month maturity



(B) 2 years maturity



(C) 10 years maturity

FIGURE V

FORECAST ERRORS FOR THREE DIFFERENT MATURITIES OVER THE OUT-OF-SAMPLE WINDOW 1996-2019

Aquaro, M., N. Bailey, and M. H. Pesaran (2020): "Estimation and inference for spatial models with heterogeneous coefficients: an application to US house prices", *Forthcoming Journal of Applied Econometrics*.

— (2019): "Estimation and inference for spatial models with heterogeneous coefficients: an application to U.S. house prices", eng, CESifo Working Paper 7542, Munich.

Bai, J. and K. Li (2015): "Dynamic spatial panel data models with common shocks", *Manuscript: Columbia University*.

Bailey, N., S. Holly, and M. H. Pesaran (2016): "A Two-Stage Approach to Spatio-Temporal Analysis with Strong and Weak Cross-Sectional Dependence", *Journal of Applied Econometrics*, 31(1), 249–280.

Bauer, M. D. and J. D. Hamilton (2018): "Robust bond risk premia", *The Review of Financial Studies*, 31(2), 399–448.

Bianchi, D., M. Büchner, and A. Tamoni (2020): "Bond risk premiums with machine learning", *The Review of Financial Studies*.

Chudik, A. and M. H. Pesaran (2013): "Large Panel Data Models with Cross-Sectional Dependence: A Survey", eng, CESifo Working Paper 4371, Munich.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011): "Weak and strong cross-section dependence and estimation of large panels", *The Econometrics Journal*, 14(1), C45–c90.

Coroneo, L., D. Giannone, and M. Modugno (2016): "Unspanned macroeconomic factors in the yield curve", *Journal of Business & Economic Statistics*, 34(3), 472–485.

Crump, R. K. and N. Gospodinov (2019): "Deconstructing the yield curve", *FRB of New York Staff Report*, (884).

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood from Incomplete Data Via the EM Algorithm", *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.

Diebold, F. X. and C. Li (2006): "Forecasting the term structure of government bond yields", *Journal of econometrics*, 130(2), 337–364.

Durbin, J. and S. J. Koopman (2012): *Time series analysis by state space methods*, Oxford university press.

Elhorst, J. P. (2014): *Spatial econometrics: from cross-sectional data to spatial panels*, vol. 479, Springer.

Eo, Y. and K. H. Kang (2020): "The effects of conventional and unconventional monetary policy on forecasting the yield curve", *Journal of Economic Dynamics and Control*, 111, 103812.

Fama, E. F. and R. R. Bliss (1987): "The information in long-maturity forward rates", *The American Economic Review*, 680–692.

Fan, Y. and C. Y. Tang (2013): "Tuning parameter selection in high dimensional penalized likelihood", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3), 531–552.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007): "Pathwise coordinate optimization", *The Annals of Applied Statistics*, 1(2), 302–332.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010): "Regularization Paths for Generalized Linear Models via Coordinate Descent", *Journal of Statistical Software*, 33(1).

Frühwirth-Schnatter, S. and H. F. Lopes (2018): "Sparse Bayesian factor analysis when the number of factors is unknown", *arXiv preprint arXiv:1804.04231*.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2007): "The US Treasury yield curve: 1961 to the present", *Journal of monetary Economics*, 54(8), 2291–2304.

Härdle, W. K. and P. Majer (2016): "Yield curve modeling and forecasting using semiparametric factor dynamics", *The European Journal of Finance*, 22(12), 1109–1129.

Harvey, A. C. (1989): *Forecasting, structural time series models and the Kalman filter*, Cambridge university press.

Hirose, K. and S. Konishi (2012): "Variable selection via the weighted group lasso for factor analysis models", *Canadian Journal of Statistics*, 40(2), 345–361.

Kaufmann, S. and C. Schumacher (2017): "Identifying relevant and irrelevant variables in sparse factor models", *Journal of Applied Econometrics*, 32(6), 1123–1144.

— (2019): "Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification", *Journal of Econometrics*, 210(1), Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective", 116–134.

Kooij, A. J. van der (2007): "Prediction accuracy and stability of regression with optimal scaling transformations", PhD thesis, Leiden University.

Kuersteiner, G. M. and I. R. Prucha (2020): "Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity", *Econometrica*, 88(5), 2109–2146.

Lam, C. and P. C. Souza (2019): "Estimation and Selection of Spatial Weight Matrix in a Spatial Lag Model", *Journal of Business & Economic Statistics*, 0(0), 1–41.

Lee, W. and Y. Liu (2012): "Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood", *Journal of Multivariate Analysis*, 111, 241–255.

LeSage, J. P. and R. K. Pace (2008): *Introduction to Spatial Econometrics*, CRC Press.

Linton, O., E. Mammen, J. P. Nielsen, and C. Tanggaard (2001): "Yield curve estimation by kernel smoothing methods", *Journal of Econometrics*, 105(1), 185–223.

Liu, Y. and J. C. Wu (2020): "Reconstructing the yield curve", tech. rep., National Bureau of Economic Research.

Lu, X. and L. Su (2016): "Shrinkage estimation of dynamic panel data models with interactive fixed effects", *Journal of Econometrics*, 190(1), 148–175.

Ludvigson, S. C. and S. Ng (2009): "Macro factors in bond risk premia", *The Review of Financial Studies*, 22(12), 5027–5067.

McCracken, M. W. and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research", *Journal of Business & Economic Statistics*, 34(4), 574–589.

Meng, X.-L. and D. B. Rubin (1993): "Maximum likelihood estimation via the ECM algorithm: A general framework", *Biometrika*, 80(2), 267–278.

Parikh, N. and S. Boyd (2014): "Proximal Algorithms", *Found. Trends Optim.*, 1(3), 127–239.

Schnücker, A. (2017): "Penalized Estimation of Panel Vector Autoregressive Models: A Lasso Approach", mimeo.

Shumway, R. H. and D. S. Stoffer (1982): "An Approach To Time Series Smoothing And Forecasting Using The Em Algorithm", *Journal of Time Series Analysis*, 3(4), 253–264.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013): "A Sparse-Group Lasso", *Journal of Computational and Graphical Statistics*, 22(2), 231–245.

Vayanos, D. and J.-L. Vila (2009): "A preferred-habitat model of the term structure of interest rates", tech. rep., National Bureau of Economic Research.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties", *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H., T. Hastie, and R. Tibshirani (2006): "Sparse Principal Component Analysis", *Journal of Computational and Graphical Statistics*, 15(2), 265–286.

Appendix

A ‖ Rotational Invariance

We can solve rotational invariance without restricting the factor loading matrix in our dynamic factor model with spatial errors. In that case we restrict the covariance of the idiosyncratic term of the dynamic equation for the factors to be identity and the autoregressive coefficient matrix to be diagonal. To prove that these restrictions are indeed enough to solve rotational invariance we will show that a $r \times r$ full rank rotation matrix A can only be equal to $\pm I$ under the proposed restrictions. Using the rotation matrix A we are able to rewrite Eq. (1) in terms of the rotated factors

$$\begin{aligned} y_t &= \Lambda A^{-1} A f_t + \xi_t \\ &= \Lambda^* f_t^* + \xi_t \,, \end{aligned}$$

with $\Lambda^* = \Lambda A^{-1}$ and $f_t^* = A f_t$. Hence, writing the dynamic equation of the factors in terms of the rotated factors f_t^* yields

$$f_t^* = A \phi A^{-1} f_t^* + A \eta_t \,.$$

The normalization that $\mathrm{Var}(A\eta_t) = I$ implies that

$$A A' = I \,,$$

hence A is an orthonormal matrix. Next, we know that the unconditional variance of the factors is equal to a diagonal matrix, denoted by Σ_f, as the factors are independent. From this restriction we know that $A \Sigma_f A'$ should be equal to a diagonal matrix. This implies that as Σ_f is diagonal A has to be triangular in order for $A \Sigma_f A'$ to be diagonal. Finally, an orthonormal matrix A that is also triangular has to be a diagonal matrix with ± 1 on the diagonal. To see this observe that A is lower triangular implies that (similar reasoning for when A is upper triangular)

$$A A' = \begin{bmatrix} a_{11}^2 & 0 & 0 & \cdots & 0 \\ a_{21} a_{11} & a_{21}^2 + a_{22}^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{r1} a_{11} & a_{21} a_{r1} + a_{22} a_{r2} & \cdots & \cdots & \sum_{i=1}^{r} a_{ri}^2 \end{bmatrix} .$$

Hence, $a_{11}^2 = 1$ implies that $a_{11} = \pm 1$. Moreover, $a_{21} a_{11} = 0$ implies that $a_{21} = 0$ as $a_{11} \neq 0$. This in turn leads to the conclusion that $a_{22} = \pm 1$ from $a_{21}^2 + a_{22}^2 = 1$, etc. Concluding, we have shown that under the stated restrictions A must be equal to $\pm I$, hence, the rotational invariance is solved.

B ‖ Covariance Smoothing Equation-by-Equation

In this section the covariance smoothing algorithm is derived when performing the Kalman Smoother equation-by-equation in the case of multivariate time-series. These covariance smoothing recursions are used in the E-step (4) of the CDPEM algorithm (Algorithm 1). When deriving the covariance smoothing recursions we do so for the general linear Gaussian state space model

$$\begin{aligned} y_t &= c_t + Z_t\alpha_t + \varepsilon_t\,, & \varepsilon_t &\sim \mathcal{N}(0, H_t)\,, \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t\,, & \eta_t &\sim \mathcal{N}(0, Q_t)\,, \qquad t = 1, \ldots, T\,, \\ & & \alpha_1 &\sim \mathcal{N}(a_1, P_1)\,, \end{aligned}$$

for more details see Durbin and Koopman (2012).

The state smoothing recursions (see Durbin and Koopman, 2012) equation-by-equation are

$$\begin{aligned} \hat{\alpha}_{t,1} &= a_{t,1} + P_{t,1}r_{t,0} \\ r_{t,i-1} &= Z_{t,i}'F_{t,i}^{-1}v_{t,i} + L_{t,i}'r_{t,i} \\ r_{t-1,N} &= T_{t-1}'r_{t,0}\,, \end{aligned}$$

where $a_{t,1}$ is provided by the equation-by-equation Kalman filter, $L_{t,i} = I - K_{t,i}Z_{t,i}$, $\hat{\alpha}_{t,1} = \hat{\alpha}_t$, $P_{t,1} = P_t$, and $r_{t,0} = r_{t-1}$ (see Durbin and Koopman (2012) for details).

Hence, we obtain

$$\begin{aligned} \mathrm{Cov}(\alpha_{t,1} - \hat{\alpha}_{t,1}, \alpha_{t+1,1} - \hat{\alpha}_{t+1,1}) &= \mathbb{E}\left[\alpha_{t,1}(\alpha_{t+1,1} - \hat{\alpha}_{t+1,1})'\right] \\ &= \mathbb{E}\left[\alpha_{t,1}x_{t+1,1}'\right] - \mathbb{E}\left[\alpha_{t,1}r_{t+1,0}'\right]P_{t+1,1}\,. \end{aligned}$$

We have for $i \neq 1$

$$\begin{aligned} x_{t+1,i} &= \alpha_{t+1,i} - a_{t+1,i} \\ &= \alpha_{t+1,i-1} - a_{t+1,i-1} - K_{t+1,i-1}v_{t+1,i-1} \\ &= x_{t+1,i+1} - K_{t+1,i-1}Z_{t+1,i-1}x_{t+1,i+1} - K_{t+1,i+1}\varepsilon_{t+1,i+1} \\ &= L_{t+1,i-1}x_{t+1,i-1} - K_{t+1,i+1}\varepsilon_{t+1,i+1} \end{aligned}$$

and when $i = 1$

$$\begin{aligned} x_{t+1,1} &= \alpha_{t+1,1} - a_{t+1,1} \\ &= T_t\alpha_{t,N} + R_t\eta_t - T_ta_{t,N} - T_tK_{t,N}v_{t,N} \\ &= T_tx_{t,N} + R_t\eta_t - T_tK_{t,N}Z_{t,N}x_{t,N} - T_tK_{t,N}\varepsilon_{t,N} \\ &= T_tL_{t,N}x_{t,N} + R_t\eta_t - T_tK_{t,N}\varepsilon_{t,N}\,. \end{aligned}$$

Hence,

$$\mathbb{E}\left[\alpha_{t,1}x'_{t+1,1}\right] = P_{t,1}L'_{t,1}\cdots L'_{t,N}T'_t\,.$$

For the second term we obtain

$$\begin{aligned}
\mathbb{E}\left[\alpha_{t,1}r'_{t+1,0}\right] &= \mathbb{E}\left[\alpha_{t,1}v'_{t+1,1}\right]F^{-1}_{t+1,1}Z_{t+1,1} + \mathbb{E}\left[\alpha_{t,1}v'_{t+1,2}\right]F^{-1}_{t+1,2}Z_{t+1,2}L_{t+1,1}\\
&\quad + \cdots + \mathbb{E}\left[\alpha_{t,1}v'_{t+1,N}\right]F^{-1}_{t+1,N}Z_{t+1,N}L_{t+1,N-1}\cdots L_{t+1,1}\\
&\quad + \mathbb{E}\left[\alpha_{t,1}r'_{t+2,0}\right]T_{t+1}L_{t+1,N}\\
&=\\
&\vdots\\
&= \mathbb{E}\left[\alpha_{t,1}v'_{t+1,1}\right]F^{-1}_{t+1,1}Z_{t+1,1} + \mathbb{E}\left[\alpha_{t,1}v'_{t+1,2}\right]F^{-1}_{t+1,2}Z_{t+1,2}L_{t+1,1}\\
&\quad + \cdots + \mathbb{E}\left[\alpha_{t,1}v'_{t+1,N}\right]F^{-1}_{t+1,N}Z_{t+1,N}L_{t+1,N-1}\cdots L_{t+1,1}\\
&\quad + \mathbb{E}\left[\alpha_{t,1}v'_{t+2,1}\right]F^{-1}_{t+2,1}Z_{t+2,1}T_{t+1}L_{t+1,N}\cdots L_{t+1,1}\\
&\quad + \cdots + \mathbb{E}\left[\alpha_{t,1}v'_{t+2,N}\right]F^{-1}_{t+2,N}Z_{t+2,N}L_{t+2,N-1}\cdots L_{t+2,1}T_{t+1}L_{t+1,N}\cdots L_{t+1,1}\\
&\quad + \cdots + \mathbb{E}\left[\alpha_{t,1}v'_{T,1}\right]F^{-1}_{T,1}Z_{T,1}T_{T-1}L_{T-1,N}\cdots L_{t+1,1}\\
&\quad + \cdots + \mathbb{E}\left[\alpha_{t,1}v'_{T,N}\right]F^{-1}_{T,N}Z_{T,N}L_{T,N-1}\cdots L_{T,1}T_{T-1}L_{T-1,N}\cdots L_{t+1,1}\,,
\end{aligned}$$

as $r_{T,N} = 0\,.$

Moreover, we have that

$$\begin{aligned}
\mathbb{E}\left[\alpha_{t,1}v'_{s,j}\right] &= \mathbb{E}\left[\alpha_{t,1}x'_{s,j}\right]Z'_{s,j}\\
&= P_{t,1}L'_{t,1}\cdots L'_{t,N}T_t\cdots T_{s-1}L'_{s,1}\cdots L'_{s,j-1}Z'_{s,j}\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\alpha_{t,1}r'_{t+1,0}\right] &= P_{t,1}L'_{t,1}\cdots L'_{t,N}T'_tZ'_{t+1,1}F^{-1}_{t+1,1}Z_{t+1,1}\\
&\quad + P_{t,1}L'_{t,1}\cdots L'_{t,N}T'_tL'_{t+1,1}Z'_{t+1,2}F^{-1}_{t+1,2}Z_{t+1,2}L_{t+1,1}\\
&\quad \vdots\\
&\quad + P_{t,1}L'_{t,1}\cdots L'_{t,N}T'_tL'_{t+1,1}\cdots L'_{t+1,N-1}Z'_{t+1,N}F^{-1}_{t+1,N}Z_{t+1,N}L_{t+1,N-1}\cdots L_{t+1,1}\\
&\quad \vdots\\
&\quad + P_{t,1}L'_{t,1}\cdots L'_{T-1,N}T'_{T-1}Z'_{T,1}F^{-1}_{T,1}Z_{T,1}T_{T-1}L_{T-1,N}\cdots L_{t+1,1}\\
&\quad \vdots\\
&\quad + P_{t,1}L'_{t,1}\cdots L'_{T,N-1}Z'_{T,1}F^{-1}_{T,1}Z_{T,1}T_{T-1}L_{T,N-1}\cdots L_{t+1,1}\,.
\end{aligned}$$

Using the following backward recursions

$$\begin{aligned}
N_{t,i-1} &= Z'_{t,i}F^{-1}_{t,i}Z_{t,i} + L'_{t,i}N_{t,i}L_{t,i}\\
N_{t-1,N} &= T'_{t-1}N_{t,0}T_{t-1}
\end{aligned}$$

we can rewrite the previous found expression for $\mathbb{E}\left[\alpha_{t,1} r'_{t+1,0}\right]$ as

$$\mathbb{E}\left[\alpha_{t,1} r'_{t+1,0}\right] = P_{t,1} L'_{t,1} \cdots L'_{t,N} T'_t N_{t+1,0} \,.$$

Hence, combining we obtain for the first order autocovariance of the smoothed estimates

$$\mathrm{Cov}(\alpha_{t,1} - \hat{\alpha}_{t,1}, \alpha_{t+1,1} - \hat{\alpha}_{t+1,1}) = P_{t,1} L'_{t,1} \cdots L'_{t,N} T'_t \left(I - N_{t+1,0} P_{t+1,1}\right) \,,$$

where $N_{t,0} = N_{t-1}$.