Journal of Econometrics 239 (2024) 105520

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Sparse generalized Yule–Walker estimation for large spatio-temporal autoregressions with an application to NO_2 satellite data

Hanno Reuvers [a,b], Etienne Wijler [c,*]

[a] Department of Econometrics, Erasmus University Rotterdam, 3062 PA Rotterdam, The Netherlands
[b] Iconsulting S.p.A., 00185 Rome, Italy
[c] Department of Econometrics and Data Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

JEL classification:
C33
C53
C55

Keywords:
Spatio-temporal models
SPLASH
Satellite data
Yule–Walker
High-dimensional

ABSTRACT

We consider a high-dimensional model in which variables are observed over time and space. The model consists of a spatio-temporal regression containing a time lag and a spatial lag of the dependent variable. Unlike classical spatial autoregressive models, we do not rely on a predetermined spatial interaction matrix, but infer all spatial interactions from the data. Assuming sparsity, we estimate the spatial and temporal dependence fully data-driven by penalizing a set of Yule–Walker equations. This regularization can be left unstructured, but we also propose customized shrinkage procedures when observations originate from spatial grids (e.g. satellite images). Finite sample error bounds are derived and estimation consistency is established in an asymptotic framework wherein the sample size and the number of spatial units diverge jointly. Exogenous variables can be included as well. A simulation exercise shows strong finite sample performance compared to competing procedures. As an empirical application, we model satellite measured nitrogen dioxide (NO_2) concentrations in London. Our approach delivers forecast improvements over a competitive benchmark and we discover evidence for strong spatial interactions.

1. Introduction

In this paper, we propose the SPatial LAsso-type SHrinkage (SPLASH) estimator: a fully data-driven, sparse estimator for large spatio-temporal models. A unique feature of our estimator is its capability to provide interpretable estimates of spatial interactions between many spatial units in a fully data-driven way, while simultaneously tackling important econometric challenges, such as inherent endogeneity problems and overparameterization of the model. SPLASH specializes in, without being limited to, modelling outcomes of spatial units that are ordered on a regularly spaced spatial grid. Relevant examples of such settings include the dynamic modelling of political preferences across voting districts, crime statistics across municipalities, and satellite-measured air pollutants. Illustrating the latter, we apply our estimator to predict NO_2 concentrations in Greater London based on daily satellite images.

Spatio-temporal models are powerful tools to explain and exploit dependencies between variables that are observed over both time and space, but they come with a number of challenges. In particular, endogeneity issues arise because the contemporaneous observations occur on both sides of the model equation. Furthermore, the inclusion of both spatial and temporal lags quickly results

* Correspondence to: Department of Econometrics and Data Science, School of Business and Economics, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands.
E-mail address: e.j.j.wijler@vu.nl (E. Wijler).

https://doi.org/10.1016/j.jeconom.2023.105520
Received 14 December 2021; Received in revised form 27 June 2023; Accepted 9 August 2023
Available online 21 October 2023

in heavily parameterized models. To circumvent these issues, a large part of the literature incorporates predetermined spatial weight matrices that govern the contemporaneous interactions between spatial units. Examples of this modelling strategy are: the spatial autoregressive model with a Gaussian quasi-maximum likelihood estimator (QMLE) (Lee, 2004); the QMLE estimation of stationary spatial panels with fixed effects detailed (Yu et al., 2008); an extension of these spatial panels to include spatially autoregressive disturbances (Lee and Yu, 2010); a further extension to a non-stationary setting in which units can be spatially cointegrated (Yu et al., 2012); and a computationally beneficial generalized method of moments (GMM) estimator (Lee and Yu, 2014). In some settings, however, the specification of a pre-determined spatial weight matrix is a non-trivial, and potentially error-prone, step in the model building process. Accordingly, recent literature focuses on either incorporating multiple weight matrices (e.g. Debarsy and LeSage, 2018; Zhang and Yu, 2018) or, at the expense of estimating many parameters, directly inferring all spatial interactions from the data (e.g. Lam and Souza, 2019; Gao et al., 2019; Ma et al., 2023). Our model is a member of the latter category.

SPLASH offers the flexibility of a data-driven spatial weight matrix, without suffering from endogeneity or overparameterization. Apart from the assumption of a generous bandwidth, SPLASH leaves the spatial weight matrix and autoregressive matrix unspecified, while employing regularization to estimate sparse solutions. We show how dependencies between neighbouring units on a spatial grid translate to diagonally structured patterns of sparsity in the spatial weight matrix. Accordingly, SPLASH incorporates a tailored regularization component to exploit such structured sparsity when present, but without imposing it from the outset. Furthermore, SPLASH does not require the use of external instruments to correct for endogeneity. Building upon previous works by Dou et al. (2016) and Gao et al. (2019), we utilize the generalized Yule–Walker equations to resolve the endogeneity problem. Further generalizing the latter contributions, we extend the system of Yule–Walker equations to allow for the inclusion of exogenous variables.

Our work is related to the literature on the estimation of (nonparametric) spatial weight matrices and banded spatio-temporal vector autoregressions. We highlight a number of key contributions along with differences and similarities with respect to the current literature. Lam and Souza (2014) consider a model specification where the spatial units depend linearly on a spatial lag and exogenous regressors and show that the adaptive lasso can consistently selects the correct sparsity pattern. To solve the endogeneity issue, they require the error variance to decay to zero as the time dimension grows large. Ahrens and Bhattacharjee (2015) solve the endogeneity problem using external instruments. Their two-step lasso estimation procedure selects the relevant instruments in the first step and the relevant spatial interactions in the second step. This approach, however, requires the instruments and the idiosyncratic component to be serially independent, resulting in limited applicability to spatio-temporal models. Finally, Lam and Souza (2019) augment a spatial lag model with a set of potentially endogenous variables. They decompose the spatial weight matrix into a pre-determined component based on expert knowledge and a sparse adjustment matrix that represents specification errors. The adjustment matrix is sparsely estimated based on a penalized version of instrumental variables (IV) regression. In contrast to our approach, Lam and Souza (2019) do not regularize the interactions between the dependent variables and the variables in the augmenting set, and they assume the number of such interactions to be fixed. A fixed number of interactions is inappropriate in high-dimensional settings in which the number of spatial units diverges.

Most closely related to our work are several recent contributions on banded estimation of spatio-temporal VARs. First, Gao et al. (2019), and Ma et al. (2023) consider the same model that appears in this work, and they also rely on the generalized Yule–Walker equations for estimation. The key difference with our paper lies in the method by which the model complexity is controlled during estimation. Gao et al. (2019) assume the coefficient matrices to be banded with a bandwidth that is small compared to the number of spatial units. The bandwidth is determined from the data and all parameters within the selected bandwidth are left unregularized. Our SPLASH estimator, however, has the ability to exploit (structured) sparsity within the bandwidth and thereby improve estimation and forecasting performance. In addition, apart from a generous upper bound on the bandwidth to ensure identification, SPLASH does not require an a priori choice regarding the bandwidth. The recently developed bagging approach by Ma et al. (2023) does allow for sparsity within the bands, yet it also requires the calculation of so-called solution paths. That is, a forward addition and backward deletion stage are needed to determine the variables that enter the final model specification. In contrast, the SPLASH estimator provides this solution at once. Furthermore, their approach is not designed to detect diagonally structured forms of sparsity, while the ability to do so results in clear performance improvements of SPLASH in both the simulations and the empirical application documented below. Finally, Wang and Tsay (2023) consider constrained estimation of the Yule–Walker equations for, potentially misspecified, high-dimensional VAR models. As part of their general framework, they consider banded VARs as a special case. Their model, however, does not include a contemporaneous (spatial) lag, nor is their estimator designed to uncover the type of (structured) sparsity that is essential to our application.

Several theoretical contributions are put forth in this paper. First, since SPLASH is based on the generalized Yule–Walker equations, which require estimates of the population autocovariance matrices, we derive a novel convergence result for banded estimates of the autocovariance matrices. Second, we derive finite-sample performance bounds for the estimation and prediction error of our estimator and utilize these bounds to derive asymptotic consistency of SPLASH in a variety of settings. For example, in the case of a finitely bounded bandwidth and unstructured sparsity, it follows that the number of spatial units N may grow at *any* polynomial rate of the number of temporal observations T. Finally, we develop a strategy to extend the system of Yule–Walker equations to accommodate additional exogenous regressors and we derive similar performance bounds of SPLASH in this broader framework as well.

We document strong performance of SPLASH in terms of estimation and forecast accuracy, both in simulated settings and in an empirical application. The simulation results highlight that SPLASH obtains highly competitive predictive accuracy even when compared to correctly-specified lattice-based methods. Furthermore, despite the heavier parametrization, the average estimation errors attained by SPLASH are among the lowest and the estimates are demonstrated to allow for straightforward interpretation

upon visualization. In our empirical application, we collect daily NO_2 column densities from 1 August 2018 to 31 March 2023, recorded by the TROPOspheric Monitoring Instrument (TROPOMI) on board of the Corpernicus Sentinel-5 Precursor satellite. Each spatial unit represents an aggregation of a small number of pixels on the satellite image. We again find that SPLASH is a competitive alternative to popular lattice-based methods in terms of forecasting performance, and that spatial–temporal methods in general deliver significant forecast improvements over regularized estimation of a reduced form VAR. In addition, we find evidence for spatial interactions between first-order and second-order neighbours (i.e. neighbours of neighbours). By visualizing these spatial interactions, we shed light on the dominant flow vectors of NO_2 which can aid the development of regional NO_2 containment policies.

This paper is organized as follows. In Section 2 we introduce the spatio-temporal vector autoregressive model and discuss its stability and sparsity properties. Next, we develop the sparse estimation strategy of this model in Section 3, which includes banded covariance matrix estimation (3.1) and the SPLASH estimator (3.2). The simulation results are described in Section 4, followed by the empirical application in Section 5. We conclude in Section 6.

Notation

The indicator function $\mathbb{1}_{\{A\}}$ equals 1 if A is true and zero otherwise. For a vector $\boldsymbol{x} \in \mathbb{R}^N$, the L_p-norm of $\boldsymbol{x}$ is denoted $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$, with $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$ as an important special case. The total number elements in $\boldsymbol{x}$ is denoted by $|\boldsymbol{x}|$ and the number of non-zero elements in $\boldsymbol{x}$ is denoted by $\mathscr{M}(\boldsymbol{x}) = \sum_{i=1}^N \mathbb{1}\{x_i \neq 0\}$. The Orlicz norm is defined as $\|\cdot\|_\psi = \inf\left\{c > 0 : \mathbb{E}\left[\psi\left(|\cdot|/c\right)\right] \leq 1\right\}$ for any $\psi(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ being a convex, increasing function with $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$. In addition, we rely on several matrix norms. For a matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, the matrix norms induced by the vector L_p-norms are given by $\|\boldsymbol{A}\|_p = \sup_{\boldsymbol{x}\in\mathbb{R}^M}\left(\|\boldsymbol{A}\boldsymbol{x}\|_p / \|\boldsymbol{x}\|_p\right)$. Noteworthy examples are: $\|\boldsymbol{A}\|_1 = \max_{1\leq j\leq N}\sum_{i=1}^M \left|a_{ij}\right|$, the spectral norm $\|\boldsymbol{A}\|_2 = \left[\lambda_{\max}(\boldsymbol{A}'\boldsymbol{A})\right]^{1/2}$ where $\lambda_{\max}(\cdot)$ stands for the maximum eigenvalue, and $\|\boldsymbol{A}\|_\infty = \max_{1\leq i\leq M}\sum_{j=1}^N \left|a_{ij}\right|$. Finally, we define $\|\boldsymbol{A}\|_\vdash = \max\left\{\|\boldsymbol{A}\|_1, \|\boldsymbol{A}\|_\infty\right\}$. Let $S \subseteq \{1,\dots,N\}$ denote an index set with cardinality $|S|$. Then, $\boldsymbol{x}_S$ denotes the $|S|$-dimensional vector with the elements of $\boldsymbol{x}$ indexed by S, whereas $\boldsymbol{A}_S$ denotes the $(M\times|S|)$-dimensional matrix containing the columns of $\boldsymbol{A}$ indexed by S. In addition, we define $\mathscr{D}_A(k) = \left\{a_{ij} \mid |i-j| = k\right\}$ as the collections of elements lying on (pairs of) the diagonals in the matrix $\boldsymbol{A}$, and $\mathscr{B}_h\left(\boldsymbol{A}\right) = \left(a_{ij}\mathbb{1}_{\{|i-j|\leq h\}}\right)$ as the banded counterpart of $\boldsymbol{A}$ with bandwidth h.

2. The spatio-temporal vector autoregression

As in the recent paper by Gao et al. (2019), we consider the spatio-temporal vector autoregression

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_t + \boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \qquad t = 1,\dots,T, \tag{1}$$

where $\boldsymbol{y}_t = (y_{1t},\dots,y_{Nt})'$ stacks the observations at time t over a collection of N spatial units. The contemporaneous spatial dependence between these spatial units is governed by the matrix $\boldsymbol{A} = (a_{ij})_{i,j=1}^N$ with $a_{ii} = 0$ for $i = 1,\dots,N$. The matrix $\boldsymbol{B} = (b_{ij})_{i,j=1}^N$ incorporates dependence on past realizations. Finally, $\boldsymbol{\epsilon}_t$ is the innovation vector driving the variation in $\boldsymbol{y}_t$.

A possible application of model (1) is the use of satellite data to study the concentration of an air pollutant over time (see the empirical application, Section 5, for more details). Overlaying each satellite image with a spatial grid of N grid cells, we record the pixel value of each grid cell at time t in the vector $\boldsymbol{y}_t = (y_{1t},\dots,y_{Nt})'$. From a physics viewpoint, the spatial lag $\boldsymbol{A}\boldsymbol{y}_t$ and the temporal lag $\boldsymbol{B}\boldsymbol{y}_{t-1}$ can capture diffusion processes in the atmosphere. That is, the concentration of the pollutant in a specific grid cell is determined by its past concentration and the inflow(outflow) of the pollutant from(to) neighbouring cells. These interactions may reasonably be expected to be localized in the sense that grid cells should be mainly influenced by the concentrations nearby, an observation that we return to in Section 2.2. The main notation and theoretical results are developed on the basis of model (1). In Section 3.3, we extend the model by including exogenous variables. In the context of air pollutants, this extension enables the explicit modelling of mechanisms that influence the dynamic evolution of pollutant concentrations beyond dispersion, such as atmospheric transport via air currents or new emissions.

2.1. Stability and dependence

To ensure that the spatio-temporal VAR defines a stable dynamic system, we restrict the parameter space in the following way.

Assumption 1 (*Stability*)**.** We require: (a) $\|\boldsymbol{A}\|_\vdash = \max\left\{\|\boldsymbol{A}\|_1, \|\boldsymbol{A}\|_\infty\right\} \leq \delta_A < 1$, and (b) $\|\boldsymbol{B}\|_\vdash \leq \delta_B$ and $\delta_C := \frac{\delta_B}{1-\delta_A} < 1$.

Assumption 1 ensures that (1) has a stable reduced form VAR(1) specification. This follows from the following two observations. First, Assumption 1(a) bounds the maximum row and column sums of $\boldsymbol{A}$ and thereby constraints the contemporaneous dependence between the time series. Invertibility of $\boldsymbol{I}_N - \boldsymbol{A}$ is guaranteed since $\|\boldsymbol{A}\|_2 \leq \sqrt{\|\boldsymbol{A}\|_1\,\|\boldsymbol{A}\|_\infty} \leq \delta_A < 1$. Hence, we can define the reduced-form representation as

$$\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{y}_{t-1} + \boldsymbol{D}\boldsymbol{\epsilon}_t, \tag{2}$$

with $\boldsymbol{C} = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}\boldsymbol{B}$ and $\boldsymbol{D} = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}$. It follows from $\|\boldsymbol{D}\|_\vdash \leq \sum_{j=0}^\infty \|\boldsymbol{A}\|_\vdash^j = \frac{1}{1-\delta_A}$ that the absolute row and column sums of $\left(\boldsymbol{I}_N - \boldsymbol{A}\right)^{-1}$ are bounded, which is the counterpart of assumption B2 in Dou et al. (2016). Second, Assumption 1(b) controls the serial dependence of $\boldsymbol{y}_t$. Indeed, we conclude from $\|\boldsymbol{C}\|_2 \leq \|\boldsymbol{C}\|_\vdash \leq \delta_C < 1$ that both unit roots and explosive behaviour of the reduced form specification are ruled out.

Remark 1. Assumption 1 is defined in terms of $\|\cdot\|_{\vdash}$. Since $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}'\|_\infty \leq \|\boldsymbol{A}\|_{\vdash}$ for any matrix $\boldsymbol{A}$, the norm $\|\cdot\|_{\vdash}$ is convenient when bounding products of matrices containing transposes.

Remark 2. Assumption 1 echoes the spatial econometrics literature in which the spatial parameter λ is bounded from above and the prespecified spatial weight matrix, say $\boldsymbol{W}_N$, is standardized (see, e.g. Lee, 2004 and Lee and Yu, 2010). Typically, the product $\lambda\boldsymbol{W}_N$ – the natural counterpart of the matrix $\boldsymbol{A}$ – is required to fulfil conditions similar to $\|\boldsymbol{A}\|_{\vdash} \leq \delta_A < 1$. For instance, it is not uncommon to row-normalize $\boldsymbol{W}_N$ (each absolute row sum equal to 1) and restrict $\lambda < 1$, see pages 1903–1904 of Lee (2004). If $\boldsymbol{W}_N$ is symmetric, then $\|\lambda\boldsymbol{W}_N\|_{\vdash} < 1$ holds.

Naturally, the dependence structure of $\{\boldsymbol{y}_t\}_{t\in\mathbb{Z}}$ is also determined by the innovation process, on which we impose the following assumption.

Assumption 2 (*Innovations*)**.**

(a) The sequence $\{\boldsymbol{\epsilon}_t\}_{t\in\mathbb{Z}}$ is a covariance stationary, martingale difference (m.d.) process with respect to the filtration $\mathscr{F}_{t-1} = \sigma\left(\boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \ldots\right)$, and geometrically strong mixing (α-mixing). That is, the mixing coefficients $\{\alpha_m\}$ satisfy $\alpha_m \leq c_2 e^{-\gamma_\alpha m}$ for all m and some constants $c_2, \gamma_\alpha > 0$.

(b) The smallest eigenvalue of $\boldsymbol{\Sigma}_\epsilon = \mathbb{E}(\boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_1') = (\sigma_{ij})_{i,j=1}^N$ has a positive lower bound, and $\|\boldsymbol{\Sigma}_\epsilon\|_{\vdash} \leq C_\epsilon < \infty$.

(c) Either one of the following assumptions holds:

(c1) For $\psi(x) = x^d$, we have $\sup_{i,t} \|\epsilon_{it}\|_\psi = \left(\mathbb{E}|\epsilon_{it}|^d\right)^{1/d} \leq \mu_d < \infty$ for $d \geq 4$.

(c2) For $\psi(x) = \exp(x) - 1$, we have $\sup_{i,t} \|\epsilon_{it}\|_\psi \leq \mu_\infty < \infty$.

Similarly to Masini et al. (2022), Assumption 2 restricts the stochastic properties of the innovation process $\{\boldsymbol{\epsilon}_t\}$. First, Assumption 2(a) ensures that the memory of the innovation process fades sufficiently fast and limits the cross-sectional dependence between the elements in $\boldsymbol{\epsilon}_t$. The m.d. assumption implies that $\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{y}_{t-j}') = \boldsymbol{0}$ while the mixing assumption controls the serial correlation in the data. Assumption 2(b) restricts the contemporaneous dependence in $\boldsymbol{\epsilon}_t$. The lower bound on the minimum eigenvalue of $\boldsymbol{\Sigma}_\epsilon$ implies that each innovation contains unique information, while the upper bound on the maximum eigenvalue of $\boldsymbol{\Sigma}_\epsilon$ bounds the degree of cross-sectional dependence, for example by excluding common factor structures. Polynomial or exponential tail decay of the distribution of the innovations is imposed through either Assumption 2(c1) or Assumption 2(c2), respectively. The type of tail decay will directly influence the growth rates we can allow for N and T. The discussions in Masini et al. (2022) demonstrate that Assumption 2 allows for a wide range of innovation models.

While these assumptions are sufficient to estimate the reduced form VAR in (2), there is a potential issue with endogeneity. In particular, since $\boldsymbol{y}_t$ appears on both sides of (1), equation-by-equation OLS estimation is inconsistent for general $\boldsymbol{A}$ and $\boldsymbol{B}$.[1] This issue can be addressed through the use of instrumental variables (e.g. Ahrens and Bhattacharjee, 2015; Lam and Souza, 2019) or via the generalized Yule–Walker equations, as considered in Gao et al. (2019) and this paper. Under Assumptions 1 and 2, we can post-multiply (1) by $\boldsymbol{y}_{t-1}'$ and take expectations to obtain

$$\boldsymbol{\Sigma}_1 = \boldsymbol{A}\boldsymbol{\Sigma}_1 + \boldsymbol{B}\boldsymbol{\Sigma}_0. \tag{3}$$

By plugging in estimates of the population covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_0$ in (3), we obtain a system of equations from which the parameters in $\boldsymbol{A}$ and $\boldsymbol{B}$ can be estimated. However, with each row in (3) defining N equations with $2N-1$ unknowns, additional structure is required to identify and estimate the parameters.

2.2. Structured sparsity

There are several possibilities to introduce structure into $\boldsymbol{A}$ and $\boldsymbol{B}$. Early spatial econometrics models, e.g. the spatial autoregressive (SAR) model or spatial Durbin model (SDM), incorporate spatial effects through the product $\lambda\boldsymbol{W}_N$ (with $\boldsymbol{W}_N$ pre-specified). The specification $\boldsymbol{A} = \lambda\boldsymbol{W}_N$ imposes substantial structure on $\boldsymbol{A}$ and leaves only the single parameter λ to estimate. Dou et al. (2016) consider a more general setting in which each row of $\boldsymbol{W}_N$ receives its own spatial autoregressive parameter. Specifically, they set $\boldsymbol{A} = \mathrm{diag}(\lambda_0)\boldsymbol{W}_N$ and $\boldsymbol{B} = \mathrm{diag}(\lambda_1) + \mathrm{diag}(\lambda_2)\boldsymbol{W}_N$, and estimate the $3N$ coefficients in $(\lambda_0', \lambda_1', \lambda_2')'$. Gao et al. (2019) require $\boldsymbol{A}$ and $\boldsymbol{B}$ to be banded matrices. We also employ a bandedness assumption, although our bandwidth may be substantially wider than in Gao et al. (2019).

Assumption 3 (*Banded Matrices*)**.** Recall $\boldsymbol{A} = (a_{ij})_{i,j=1}^N$, $\boldsymbol{B} = (b_{ij})_{i,j=1}^N$, and $\boldsymbol{\Sigma}_\epsilon = (\sigma_{ij})_{i,j=1}^N$. We have: (a) $a_{ij} = b_{ij} = 0$ for all $|i-j| > k_0$ with $k_0 \leq \lfloor (N-1)/4 \rfloor$, and (b) $\sigma_{ij} = 0$ for all $|i-j| > l_0$.

[1] As pointed out by an anonymous referee, notable exceptions occur when $\boldsymbol{A}$ is diagonal, which is ruled out in the current setting, or when $\boldsymbol{A}$ is triangular. The latter setting may be justifiable for certain structural VARs, but seems unrealistic in most spatial settings.

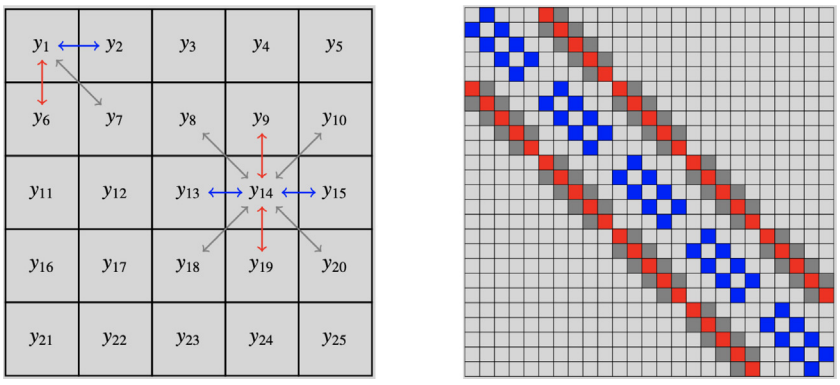

Fig. 1. Sparsity patterns in the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ due to localized interactions. **(a)** A (5×5) grid of spatial units with arrows depicting the nearest horizontal (blue), vertical (red), and diagonal (grey) interactions for y_1 and y_{14}. **(b)** The sparsity pattern resulting from nearest horizontal, nearest vertical and nearest diagonal interactions. The first row in the (25×25) matrix has three non-zero elements corresponding to the horizontal interaction $y_1 \to y_2$ causing a non-zero (1,2) element (blue), the vertical interaction $y_1 \to y_6$ causing a non-zero (1,6) element (red), and $y_1 \to y_7$ causing a non-zero (1, 7) element (grey). The full colour pattern is recovered by applying this reasoning to all remaining cells in the grid. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Assumption 3 serves two purposes. First, for each spatial unit $i = 1, \dots, N$, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are banded to have no more than N unknown parameters per equation. With N moment conditions for each i, Assumption 3(a) is the widest bandwidth that enables identification of the parameters. In many applications, including the one considered in this paper, the actual bandwidth may be smaller than the upper bound imposed in Assumption 3. While a smaller bandwidth is neither required nor imposed, it will result in additional sparsity that our estimator is designed to exploit. Second, we demonstrate in Theorem 1 that the combination of Assumptions 3(a)–(b) implies that the $(N \times N)$ autocovariance matrices $\boldsymbol{\Sigma}_j = \mathbb{E}(\boldsymbol{y}_t \boldsymbol{y}_{t-j}')$ are *approximately* banded. This in turn is exploited in the Yule–Walker estimation approach, in which we incorporate banded estimation of the autocovariance matrices to improve the convergence rates of our estimator.

Even under Assumption 3, the number of unknown parameters in $\boldsymbol{A}$ and $\boldsymbol{B}$ may grow quadratically in N. Accurate estimation of all these parameters becomes infeasible even for moderately sized systems. To alleviate this curse of dimensionality, we rely on sparsity, which in the present context occurs when two spatial units do not interact with each other (given their interactions with other units within the system). We demonstrate how a particular, and exploitable, sparsity pattern arises when the spatial units are ordered in a structured way.

As an illustrative example, let us consider repeated measurements on the (5×5) spatial grid shown in Fig. 1(a). The $N = 25$ spatial units are labelled y_1 up to y_{25} and enumerated row-wise. This ordering of the spatial entities creates an implicit notion of proximity and we intuitively expect economic/physical interactions to be most pronounced at short length scales. For instance, returning to our running example of air pollution, we expect elevated concentrations of pollutants (e.g. due to high traffic or polluting industries) to diffuse through space and mostly affect neighbouring locations. Fig. 1(a) shows the short-range interactions of y_1 and y_{14} towards their horizontal (blue), vertical (red) and diagonal (grey) nearest neighbour. If these are the only possible interactions, then the resulting sparsity pattern in the (25×25) matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ would correspond to the coloured elements in Fig. 1(b) (see its caption for further explanations). The nonzero elements in $\boldsymbol{A}$ and $\boldsymbol{B}$ are thus seen to cluster in specific, dense diagonals with an occasional zero whenever neighbours are absent at the grid boundary. It is easy to verify that more distant neighbours (in either direction) are still found on diagonals but located further away from the main diagonal. Not knowing a priori in which direction and at which distance these interactions become unimportant, we develop a penalized estimator that is able to detect and efficiently estimate such diagonal sparsity patterns *without imposing* this sparsity structure from the outset.

3. Sparse estimation

In this section, we develop the estimation and statistical theory for a sparse estimator of the spatio-temporal VAR in (1). As mentioned in Section 2.2, we exploit the banded structure in $\boldsymbol{A}$ and $\boldsymbol{B}$ via banded estimation of the sample autocovariance matrices. Accordingly, we start off by deriving finite-sample error bounds attained by banded sample auto-covariance matrices as estimators of their population counterparts. Afterwards, we fully develop our estimator and investigate its theoretical properties.

3.1. Banded autocovariance matrix estimation

Assumption 3 plays a crucial role in identifying the parameters in $\boldsymbol{A}$ and $\boldsymbol{B}$, while simultaneously opening up the door for efficiency gains in the estimation of the autocovariance matrices in (3). Since the inverse of a banded matrix is in general not

banded itself, the matrix $\boldsymbol{C} = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}\boldsymbol{B}$ containing the autoregressive coefficients of the reduced-form representation of the spatio-temporal VAR in (2) cannot be expected to be banded. However, given its structure, it may be expected to be well-approximated by a banded matrix. Consequently, combining the *approximate* bandedness of $\boldsymbol{C}$ with the bandedness of $\boldsymbol{\Sigma}_\epsilon$, a reasonable conjecture would be that the autocovariance matrices of $\boldsymbol{y}_t$ are approximately banded as well. Indeed, we formalize this conjecture in the following theorem, along with the expected error bounds one may expect from utilizing a banded estimation approach.

Theorem 1 (*Convergence Rates for Banded Sample Autocovariance Matrices*). *For any matrix $\boldsymbol{M} = (m_{ij})$, its h-banded counterpart is defined as $\mathscr{B}_h(\boldsymbol{M}) = (m_{ij}\mathbb{1}_{\{|i-j|\leq h\}})$. Define the $(N \times 2N)$ matrix $\hat{\boldsymbol{V}}_h = \left[\mathscr{B}_h\left(\hat{\boldsymbol{\Sigma}}_1\right)' \; \mathscr{B}_h\left(\hat{\boldsymbol{\Sigma}}_0\right)\right]$ with $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{T}\sum_{t=2}^{T}\boldsymbol{y}_t\boldsymbol{y}_{t-1}'$ and $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{T}\sum_{t=2}^{T}\boldsymbol{y}_t\boldsymbol{y}_t'$, and choose*

$$h := h(\epsilon) = \left(\frac{\max\left\{\log\left(\frac{K-\delta_c}{\delta_c}\right), \log\left(\frac{C_1}{\epsilon}\right)\right\}}{\left|\log(\delta_A)\right|} + 1\right)\left(\frac{\log\left(\frac{C_2}{\epsilon}\right)}{\left|\log(\delta_C)\right|} + 2\right)(k_0 - 1) + 2l_0 + 1, \tag{4}$$

for some $K \in (\delta_c, 1)$, with $C_1 = \frac{34C_\epsilon}{(1-\delta_A)^2(1-K^2)^2}$ and $C_2 = \frac{4C_\epsilon}{(1-\delta_A)^2(1-\delta_C^2)}$. Then, $\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_{\vdash} \leq 4\epsilon$ with a probability of at least

(a) $1 - 2\mathscr{P}_1(\epsilon, N, T)$ under Assumptions 1–3 using Assumption 2(b1) (polynomial tails),
(b) $1 - 2\mathscr{P}_2(\epsilon, N, T)$ under Assumptions 1–3 using Assumption 2(b2) (exponential tails),

where

$$\mathscr{P}_1(\epsilon, N, T) = (2h+1)N\left[\left(b_1 T^{(1-\delta)/3} + \frac{[2h+1]b_3}{\epsilon}\right)\exp\left(-\frac{T^{(1-\delta)/3}}{2b_1^2}\right) + \frac{b_2[2h+1]^d}{\epsilon^d T^{\frac{\delta}{2}(d-1)}}\right],$$

for some $0 < \delta < 1$, and

$$\mathscr{P}_2(\epsilon, N, T) = (2h+1)N\left[\frac{\kappa_1[2h+1]}{\epsilon} + \frac{2}{\kappa_2}\left(\frac{T\epsilon^2}{[2h+1]^2}\right)^{\frac{1}{7}}\right]\exp\left(-\frac{1}{\kappa_3}\left(\frac{T\epsilon^2}{[2h+1]^2}\right)^{\frac{1}{7}}\right).$$

The constants b_i, κ_i, for $i = 1, 2, 3$, are positive and independent of N and T.

Theorem 1 shows that banded estimators for $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ provide an accurate approximation to $\boldsymbol{V} = \left[\boldsymbol{\Sigma}_1' \;\; \boldsymbol{\Sigma}_0\right]$. Each of these banded matrices has at most $2h(\epsilon)+1$ nonzero elements in their columns/rows. In other words, given ϵ, l_0 and k_0, Assumptions 1–3 guarantee that $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ can be well-approximated by matrices with bandwidths smaller than N. In the following section, we develop our sparse estimator for $\boldsymbol{A}$ and $\boldsymbol{B}$ and show that the established approximability of the population autocovariance matrices improves the convergence rate of the estimator.

3.2. The SPLASH (α, λ) estimator

In this section, we introduce a sparse estimator for large spatio-temporal models generated by (1). Our estimator essentially combines generalized Yule–Walker estimation with a sparse group penalty (e.g. Simon et al., 2013). Estimation via the Yule–Walker equations is adopted to control for endogeneity, while the added penalization shrinks the estimates towards the diagonally structured sparsity displayed in Fig. 1. Compared to Gao et al. (2019), we hereby gain the ability exploit sparsity within the imposed bandwidths of $\boldsymbol{A}$ and $\boldsymbol{B}$. In practice, this means that we can assume a much larger bandwidth without substantially sacrificing estimation accuracy.

3.2.1. Definition

The formal definition of our estimator requires further notation. Part of this notation comes naturally if we briefly review the generalized Yule–Walker estimator. First, we rewrite the generalized Yule–Walker conditions in (3) more compactly as

$$\boldsymbol{\Sigma}_1' = \left[\boldsymbol{\Sigma}_1' \quad \boldsymbol{\Sigma}_0\right]\left[\boldsymbol{A} \quad \boldsymbol{B}\right]' =: \boldsymbol{V}\boldsymbol{C}'. \tag{5}$$

The ith column of $\boldsymbol{C}'$ contains all coefficients that belong to the ith equation in (1). Assumption 3 requires several of these coefficients to be zero, so we exclude these from the outset. We collect all remaining (possibly) nonzero coefficients of the ith equation in the vector $\boldsymbol{c}_i$, and define $\boldsymbol{V}_i$ as the matrix containing the corresponding columns from $\boldsymbol{V}$. In the population, we have $\boldsymbol{V}_i\boldsymbol{c}_i = \boldsymbol{\Sigma}_1'\boldsymbol{e}_i =: \boldsymbol{\sigma}_i$ for $i = 1, \ldots, N$, with $\boldsymbol{e}_i$ being the ith column of the $(N \times N)$ identity matrix. Sample counterparts of $\boldsymbol{V}_i$ and $\boldsymbol{\sigma}_i$ are readily available from the sample autocovariance matrices. More explicitly, Gao et al. (2019) set $\hat{\boldsymbol{\sigma}}_i = \frac{1}{T}\sum_{t=2}^{T}\boldsymbol{y}_{t-1}y_{it}$ and construct $\hat{\boldsymbol{V}}_i$ from the appropriate columns of $\hat{\boldsymbol{V}} = \left[\hat{\boldsymbol{\Sigma}}_1' \; \hat{\boldsymbol{\Sigma}}_0\right]$. Motivated by $\boldsymbol{\sigma}_i - \boldsymbol{V}_i\boldsymbol{c}_i = \boldsymbol{0}$, they define their estimator $\hat{\boldsymbol{c}}_i^{GMWY}$ as the following minimizer:

$$\hat{\boldsymbol{c}}_i^{GMWY} = \arg\min_{\boldsymbol{c}} \left\|\hat{\boldsymbol{\sigma}}_i - \hat{\boldsymbol{V}}_i\boldsymbol{c}\right\|_2^2. \tag{6}$$

We will adjust this objective function in three ways. First, we define our estimator in terms of banded estimated covariance matrices. Second, we add a group penalty to sparsely estimate the parameters in $\boldsymbol{A}$ and $\boldsymbol{B}$. To exploit the diagonally structured sparsity in Fig. 1, if it is present, we allow for a penalization of complete diagonals. This form of penalization, however, renders

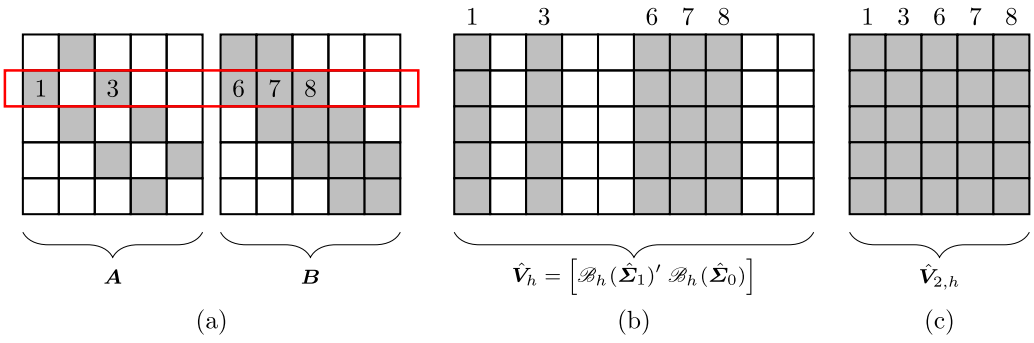

Fig. 2. A visualization on the construction of $\hat{\boldsymbol{V}}_{2,h}$ for $N = 5$. **(a)** If $h = 1$, then grey elements in $\boldsymbol{A}$ and $\boldsymbol{B}$ are (potentially) nonzero whereas white elements are zero by construction. Enumerating along the second row, the active elements are in the set $\{1, 3, 6, 7, 8\}$. **(b)** We select the columns from $\hat{\boldsymbol{V}}_h$ corresponding to the active set. **(c)** The matrix $\hat{\boldsymbol{V}}_{2,h}$ is the submatrix of $\hat{\boldsymbol{V}}_h$ with only active columns.

equation-by-equation estimation of the parameters infeasible. Therefore, recalling the definitions of $\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)$ and $\hat{\boldsymbol{V}}_h$ in Theorem 1, we define $\hat{\boldsymbol{\sigma}}_h = \mathrm{vec}\big(\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)'\big)$ and $\hat{\boldsymbol{V}}_h^{(d)} = \mathrm{diag}\big(\hat{\boldsymbol{V}}_{i,h}, \dots, \hat{\boldsymbol{V}}_{N,h}\big)$, where $\hat{\boldsymbol{V}}_{i,h}$ contains the columns of $\hat{\boldsymbol{V}}_h$ that multiply the elements in $\boldsymbol{c}_i$ (see Fig. 2 for an illustration). Finally, we construct the penalty function. We define an index set that partitions the vector $\boldsymbol{c}$ into sub-vectors, denoted $\{\boldsymbol{c}_g\}$, that contain the non-zero diagonals of $\boldsymbol{A}$ and $\boldsymbol{B}$ that are admissible under Assumption 3 as

$$\begin{aligned} \mathscr{G}_A &:= \big\{g \subset \mathbb{N} : \boldsymbol{c}_g = \mathscr{D}_A(k), k \in \{1, \dots, \lfloor (N-1)/4 \rfloor\}\big\}, \\ \mathscr{G}_B &:= \big\{g \subset \mathbb{N} : \boldsymbol{c}_g = \mathscr{D}_B(k), k \in \{0, \dots, \lfloor (N-1)/4 \rfloor\}\big\}, \end{aligned} \tag{7}$$

respectively, and let $\mathscr{G} = \mathscr{G}_A \cup \mathscr{G}_B$. Based on this notation, we define our objective function as

$$\mathscr{L}_\alpha(\boldsymbol{c}; \lambda) = \left\|\hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c}\right\|_2^2 + \lambda \underbrace{\left((1-\alpha)\sum_{g \in \mathscr{G}} \sqrt{|g|}\left\|\boldsymbol{c}_g\right\|_2 + \alpha\|\boldsymbol{c}\|_1\right)}_{=: P_\alpha(\boldsymbol{c})} = \left\|\hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c}\right\|_2^2 + \lambda P_\alpha(\boldsymbol{c}). \tag{8}$$

The spatial lasso-type shrinkage estimator, abbreviated SPLASH(α, λ) or SPLASH for short, is defined as the minimizer of (8), i.e. $\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} \mathscr{L}_\alpha(\boldsymbol{c}; \lambda)$. The influence of the penalty function $P_\alpha(\boldsymbol{c})$ is governed by the penalty parameter λ and the second hyperparameter α balances group-structured sparsity versus individual sparsity. At the extremities of $\alpha \in [0, 1]$ we find the group lasso ($\alpha = 0$) and the lasso ($\alpha = 1$). Intermediate values of α will shrink groups of diagonal coefficients in $\boldsymbol{A}$ and $\boldsymbol{B}$, and individual parameters. The SPLASH solution promotes completely sparse diagonals and sparse elements within nonzero diagonals, and thus shrinks towards the sparsity patterns displayed in Fig. 1(b).

3.2.2. *Implementation*

Algorithm 1 describes how to apply the SPLASH estimator. There are three key parts: (1) the selection of the bandwidth parameter h, (2) the determination of the hyperparameters α and λ, and (3) the minimization of the objective function $\mathscr{L}_\alpha(\boldsymbol{c}; \lambda)$. Further details on parts (1) and (3) can be found in the articles that are referenced in Algorithm 1. The hyperparameters are determined using time series cross-validation. The grid values being outlined in the algorithm are mere recommendations. Finer grids can yield better solutions at the cost of higher computational demands. If computational resources are scarce, then hyperparameters might also be pre-specified (e.g. fixing α to say 0.5 in order to obtain a one-dimensional grid search). Finally, we mention that there is no need to manually implement the steps in Algorithm 1. An efficient implementation of this procedure is readily available as an R package on the website of one of the authors.[2]

3.2.3. *Theoretical properties*

We require an additional assumption on the DGP to ensure that $\boldsymbol{A}$ and $\boldsymbol{B}$ in (1) are uniquely identified. To this end, we leverage on Assumption 3, which enables unique identification of $\boldsymbol{A}$ and $\boldsymbol{B}$ via a straightforward full-rank condition on sub-matrices of the autocovariance matrices that appear in the generalized Yule–Walker equations.

Assumption 4 (*Restricted Minimum Eigenvalue*)**.** Assume that

$$\phi_{\min}(\boldsymbol{x}) := \min_{\boldsymbol{x} \in \mathbb{R}^{2N} : \mathscr{M}(\boldsymbol{x}) \leq N} \frac{\|\boldsymbol{V}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \geq \phi_0 > 0.$$

[2] https://sites.google.com/view/etiennewijler/code

Algorithm 1 SPLASH(α, λ) implementation

1: Compute the $(N \times N)$ matrices $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{T}\sum_{t=2}^{T} \boldsymbol{y}_t \boldsymbol{y}_t'$ and $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{T}\sum_{t=2}^{T} \boldsymbol{y}_t \boldsymbol{y}_{t-1}'$.
2: Determine the bandwidth h using the bootstrap procedure in Guo et al. (2016, p. 7).
3: Using the selected bandwidth of step 2, calculate $\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_0)$, $\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)$, and construct $\hat{\boldsymbol{\sigma}}_h$, $\hat{\boldsymbol{V}}^{(d)}$ and $\mathscr{G}$ as in Section 3.2.1.
4: Define a grid of hyperparameter values. Depending on computational power, we suggest $\alpha_{\text{grid}} = \{0, 0.25, 0.5, 0.75, 1\}$ and a vector of K logarithmically spaced points for λ. That is, with $\lambda_{\max} = \max\left(\max_{g\in\mathscr{G}} \frac{1}{T}\left\|\hat{\boldsymbol{V}}_{h,g}^{(d)\prime}\hat{\boldsymbol{\sigma}}_h\right\|_2 / \sqrt{|g|}, \max_{1\le i\le N_c}\left|\hat{\boldsymbol{V}}_{h,i}^{(d)\prime}\hat{\boldsymbol{\sigma}}_h\right|\right)$ and $\lambda_{\min} = \eta\lambda_{\max}$ for some small $\eta > 0$, take $\lambda_{\text{grid}} = \{\lambda_1, \dots, \lambda_K\}$ where

$$\lambda_i = \exp\left(\ln\left(\lambda_{\max}\right) - \frac{i-1}{K-1}\left[\ln(\lambda_{\max}) - \ln(\lambda_{\min})\right]\right) \qquad i = 1, \dots, K.$$

5: For each combination (α, λ), use the SGL algorithm as in Simon et al. (2013) to compute the SPLASH solution $\hat{\boldsymbol{c}}(\alpha, \lambda)$. Select the solution with the hyperparameters that perform best via time series cross-validation (see, e.g. Hyndman and Athanasopoulos, 2018).

Assumption 4 states that every sub-matrix containing at most N columns from $\boldsymbol{V}$ has full column-rank and a minimum singular value bounded away from zero. Related assumptions appear in Bickel et al. (2009, Section 4), who refer to $\phi_{\min}(\boldsymbol{x})$ as a *restricted eigenvalue* and use this quantity to construct sufficient conditions for their restricted eigenvalue condition. Assumption 4 fits our framework particularly well, as the assumed maximum bandwidth of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in Assumption 3 imply that the diagonal blocks of the matrix $\boldsymbol{V}^{(d)}$ never contain more than N unique columns of $\boldsymbol{V}$. Using this property, we show in Lemma 1 of Appendix A that a Sparse Group Lasso compatibility condition is implied by Assumption 4.

Equipped with Assumption 4, we find the following finite-sample bounds on the prediction and estimation error of SPLASH.

Theorem 2. *Define*

$$\bar{\omega}_\alpha = \max\left\{(1-\alpha)\sum_{g\in\mathscr{G}_S}\sqrt{|g|}, \alpha\sqrt{|S|}\right\},$$

where $\mathscr{G}_s = \{g \in \mathscr{G} : \boldsymbol{c}_g \neq \boldsymbol{0}\}$ *and* $S = \{j : c_j \neq 0\}$. *Under Assumptions 3–4 and* $\|\boldsymbol{V}\|_{\vdash} \le C_V$, *it holds that*

$$\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}} - \boldsymbol{c})\right\|_2^2 + \lambda\left((1-\alpha)\sum_{g\in\mathscr{G}}\sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g - \boldsymbol{c}_g\right\|_2 + \alpha\|\hat{\boldsymbol{c}} - \boldsymbol{c}\|_1\right) \le \frac{64\bar{\omega}_\alpha^2\lambda^2}{\phi_0^2} \tag{9}$$

with a probability of at least

(a) $1 - 10\mathscr{P}_1\left(f(\lambda, \phi_0), N, T\right)$ *when Assumption 2(b1) (polynomial tail decay) is valid, or*
(b) $1 - 10\mathscr{P}_2\left(f(\lambda, \phi_0), N, T\right)$ *when Assumption 2(b2) (exponential tail decay) is valid,*

where $\mathscr{P}_1(x, N, T)$ *and* $\mathscr{P}_2(x, N, T)$ *are defined in Theorem 1 and* $f(\lambda, \phi_0) = \min\left(\frac{\lambda^{1/2}}{24}, \frac{\lambda}{96C_v}, \frac{\phi_0}{12}\right)$.

Theorem 2 contains a finite-sample bound on the prediction and estimation error for the SPLASH(α, λ) estimator. It offers some interesting insights. First, we focus on the probability with which inequality (9) holds. For VAR estimation with a penalized least-squares objective function, such probabilities are governed by tail probabilities of the process $\{\frac{1}{T}\sum_{t=1}^{T} y_{it}\epsilon_{jt}\}$ (see, e.g. Lemma 4 in Kock and Callot, 2015, or Lemmas 5–6 in Medeiros and Mendes, 2016). Because Yule–Walker estimation relies primarily on autocovariance matrix estimation, our probability depends on the tail decay of the distribution of $\{\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_{\vdash}\}$. Overall, the probability of (9) improves through faster tail decay of the innovation distribution (compare cases *(a)* and *(b)*) and banded autocovariance matrix estimation (Theorem 1). Second, we look closer at the performance upper bound itself. The right-hand side of (9) demonstrates that the upper bound on the prediction and estimation error is increasing in $\bar{\omega}_\alpha$, which in turn is increasing in the bandwidths k_0 and l_0, increasing in the group sizes ($\alpha < 1$), and increasing in the number of relevant interactions $|S|$ ($\alpha > 0$). Furthermore, the prediction and estimation error increases in the degree of penalization. Whereas this seemingly suggests to minimize λ as to improve performance bounds, we emphasize that the effect of regularization in Theorem 2 is two-fold: increasing regularization deteriorates the performance bound, but increases the probability of the set on which the performance bound holds. Intuitively, shrinkage induces finite-sample bias which worsens accuracy, but simultaneously reduces sensitivity to noise, thereby enabling performance guarantees at higher degrees of certainty.

The aforementioned effects can also be demonstrated by means of an asymptotic analysis. Based on Theorem 2, we derive the conditions for convergence of the prediction and estimation errors in the following corollary. The exact convergence rates are also provided.

Corollary 1. *Let* $\lambda \in O(T^{-q_\lambda})$, $N \in O(T^{q_N})$, $|\mathscr{G}_S| \in O\left(T^{q_g}\right)$, $|S| \in O(T^{q_s})$, $k_0, l_0 \in O(T^{q_k})$, *where* q_λ, q_N, q_s, *and* q_k *are fixed and positive constants. Maintain Assumptions 1–4 and assume that either (i)* $q_\lambda < \frac{\delta(d-1)}{2d} - \frac{(d+1)q_k}{d} - \frac{q_N}{d}$ *for some* $0 < \delta < 1$ *and Assumption 2(b1) holds, or (ii)* $q_\lambda < \frac{1}{2} - q_k$ *and Assumption 2(b2) holds. Then,*

(a) $\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2^2 = O_p\left((1-\alpha)T^{2q_g+q_N-2q_\lambda}+\alpha T^{q_s-2q_\lambda}\right)$,

(b) $P_\alpha(\hat{\boldsymbol{c}}-\boldsymbol{c}) = (1-\alpha)\sum_{g\in\mathscr{G}}\sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g-\boldsymbol{c}_g\right\|_2+\alpha\left\|\hat{\boldsymbol{c}}-\boldsymbol{c}\right\|_1 = O_p\left((1-\alpha)T^{2q_g+q_N-q_\lambda}+\alpha T^{q_s-q_\lambda}\right)$.

Corollary 1 provides insights into the determinants of the convergence rate. In particular, the result confirms that the convergence rate decreases in the bandwidths k_0 and l_0, the number of spatial units N, the number of interactions $|S|$ and the degree of penalization λ. To ensure that the set on which the performance bound in Theorem 2 holds occurs with probability converging to one, conditions (i) and (ii) impose that the degree of penalization does not decay too fast in T. The optimal convergence rate is obtained by choosing q_λ as large as possible without violating these conditions. Some concrete examples are provided in Remark 3.

Remark 3. Insightful special cases can be examined based on Corollary 1. For the sake of brevity, we consider two cases while focusing on the estimation error $P_\alpha(\hat{\boldsymbol{c}}-\boldsymbol{c})$ and assuming errors with at least d finite moments (Assumption 2(b1)). In the absence of within-group shrinkage ($\alpha = 0$), Corollary 1 demonstrates that $P_0(\hat{\boldsymbol{c}}-\boldsymbol{c}) = O_p\left(T^{2q_g+q_N-q_\lambda}\right)$, with $q_\lambda < \frac{1}{2}-\frac{1}{2d}-\frac{(d+1)q_k}{d}-\frac{q_N}{d}$. The estimator now converges almost at rate $\frac{T^{1/2-1/2d}}{N^{1+1/d}k_0^{(d+1)/d}|\mathscr{G}_S|}$. For fixed N and large d, this is close to the common $\sqrt{T}$-rate of fixed-dimensional settings without regularization. If shrinkage is imposed at the individual interaction level only ($\alpha = 1$), then $P_1(\hat{\boldsymbol{c}}-\boldsymbol{c}) = O_p(T^{q_s-q_\lambda})$ and the estimation error converges almost at the rate $\frac{T^{1/2-1/2d}}{|S|N^{1/d}k_0^{(d+1)/d}}$. Noting that $N^{1+1/d}|\mathscr{G}_S| > |S|N^{1/d}$, we see that SPLASH(1, λ) attains a convergence rate at least as fast SPLASH(0, λ), and possibly faster when the sparsity is unstructured or the diagonals are highly sparse.

3.2.4. Future work: Selection and inference

In this section, we propose two prospective avenues of future research opened up by the development of SPLASH. In particular, we discuss how the estimator may be extended to provide consistent variable selection or to enable uniformly valid statistical inference.

We conjecture that a weighted penalty scheme, similar in spirit to the adaptive lasso in Zou (2006), will be able to estimate the zero coefficients as exactly zero with probability converging to one, while maintaining the ability to consistently estimate the non-zero parameters in (1). Such an estimator could be defined as be the minimizer of

$$\mathscr{L}_\alpha(\boldsymbol{c};\lambda,\boldsymbol{\omega}) = \left\|\hat{\boldsymbol{\sigma}}_h-\hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c}\right\|_2^2+\lambda\left((1-\alpha)\sum_{g\in\mathscr{G}}\sqrt{|g|}\left\|\boldsymbol{c}_g\right\|_2+\alpha\sum_{j=1}^{M}\frac{|c_i|}{\omega_i}\right),$$

where M denotes number of coefficients in the model. Compared to the objective function in (8), the vector $\boldsymbol{\omega} = (\omega_1,\dots,\omega_M)'$ is a vector of positive weights that should be large for parameters that require little shrinkage and small for parameters that are small or zero. A natural choice that satisfies this requirement is $\omega_i = |\hat{c}_i|^\gamma$ for some $\gamma \geq 1$, where $\hat{c}_i$ is the ith element of the SPLASH estimator $\hat{\boldsymbol{c}}$ defined in (8). We leave the investigation of the theoretical properties and empirical performance of this *"adaptive SPLASH estimator"* for future research.

Another prospective extension of our method concerns the ability to conduct honest statistical inference. The main challenges for inference based on sparse estimators are: (1) the intractability of the limiting distribution and/or its dependence on unknown parameters (e.g. Knight and Fu, 2000), and (2) convergence results not holding uniformly over the parameter space (e.g. Leeb and Pötscher, 2005, 2008). To circumvent these challenges, a *"debiased SPLASH estimator"* for uniformly valid inference might be derived by inverting the Karush–Kuhn–Tucker (KKT) conditions, analogous to the desparsified lasso in Van de Geer et al. (2014). Compared to Van de Geer et al. (2014), however, additional care should be taken to account for the banded autocovariance matrix and the group penalty. A full derivation is outside the scope of this paper and subject to further investigation.

3.3. Exogenous variables

We generalize model specification (1) by accommodating K exogenous variables, i.e.

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_t+\boldsymbol{B}\boldsymbol{y}_{t-1}+\sum_{k=1}^{K}\operatorname{diag}(\boldsymbol{\beta}_k)\boldsymbol{x}_{t,k}+\boldsymbol{\epsilon}_t, \qquad t = 1,\dots,T. \tag{10}$$

Each vector $\boldsymbol{x}_{t,k} = (x_{1t,k},\dots,x_{Nt,k})'$ augments the spatio-temporal vector autoregression with an extra regressor. This regressor may vary over time and is assumed exogenous, i.e. we have $\mathbb{E}(\boldsymbol{x}_{t,k}\boldsymbol{\epsilon}_t') = \mathbf{O}$ for $k = 1,\dots,K$. For notational brevity, we consider the situation in which the exogenous regressors $x_{it,1}\dots,x_{it,K}$ can only directly influence spatial unit i. This explains the diagonal structure in $\operatorname{diag}(\boldsymbol{\beta}_k)$. In Remark 5 we argue that this simplification does not greatly hinder generality. In contrast to Ma et al. (2023), we allow $\boldsymbol{\beta}_k = (\beta_{1k},\dots,\beta_{Nk})'$ to vary with location. We keep K fixed.

To account for the exogenous variables, we modify the generalized Yule–Walker estimator of Section 3.2. We recall $\boldsymbol{\Sigma}_j = \mathbb{E}(\boldsymbol{y}_t\boldsymbol{y}_{t-j}')$, and define the matrices $\boldsymbol{\Sigma}_j^{x_k y} = \mathbb{E}(\boldsymbol{x}_{t,k}\boldsymbol{y}_{t-j}')$ and $\boldsymbol{\Sigma}_j^{x_k x_\ell} = \mathbb{E}(\boldsymbol{x}_{t,k}\boldsymbol{x}_{t-j,\ell}')$. Two sets of Yule–Walker equations, namely

$$\boldsymbol{\Sigma}_1 = \boldsymbol{A}\boldsymbol{\Sigma}_1+\boldsymbol{B}\boldsymbol{\Sigma}_0+\sum_{k=1}^{K}\operatorname{diag}(\boldsymbol{\beta}_k)\boldsymbol{\Sigma}_1^{x_k y} \tag{11a}$$

and

$$(\boldsymbol{\Sigma}_0^{x_j y})' = \boldsymbol{A}(\boldsymbol{\Sigma}_0^{x_j y})' + \boldsymbol{B}(\boldsymbol{\Sigma}_1^{x_j y})' + \sum_{k=1}^{K} \operatorname{diag}(\boldsymbol{\beta}_k)\boldsymbol{\Sigma}_0^{x_k x_j}, \qquad \text{for } j = 1, \dots, K, \tag{11b}$$

are derived by post-multiplying the model by respectively $\boldsymbol{y}'_{t-1}$ and $\boldsymbol{x}'_{t,k}$, and taking expectations. Compared to (5), the Yule–Walker equations in (11a) contain the additional term $\sum_{k=1}^{K} \operatorname{diag}(\boldsymbol{\beta}_k)\boldsymbol{\Sigma}_1^{x_k y}$ to provide information on $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$. However, if $\boldsymbol{\Sigma}_1^{x_k y} = \mathbf{O}$ (e.g. when $\{\boldsymbol{x}_{t,k}\}$ and $\{\boldsymbol{y}_t\}$ are independent and $\boldsymbol{\beta}_k = \mathbf{0}$), then (11a) alone will not identify $\boldsymbol{\beta}_k$. We therefore add the additional Yule–Walker equations in (11b). To develop the estimator, we combine (11a) and (11b) into

$$\begin{bmatrix} \boldsymbol{\Sigma}_1' \\ \boldsymbol{\Sigma}_0^{x_1 y} \\ \vdots \\ \boldsymbol{\Sigma}_0^{x_K y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_1' & \boldsymbol{\Sigma}_0 \\ \boldsymbol{\Sigma}_0^{x_1 y} & \boldsymbol{\Sigma}_1^{x_1 y} \\ \vdots & \vdots \\ \boldsymbol{\Sigma}_0^{x_K y} & \boldsymbol{\Sigma}_1^{x_K y} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix}' + \sum_{k=1}^{K} \begin{bmatrix} (\boldsymbol{\Sigma}_1^{x_k y})' \\ \boldsymbol{\Sigma}_0^{x_1 x_k} \\ \vdots \\ \boldsymbol{\Sigma}_0^{x_K x_k} \end{bmatrix} \operatorname{diag}(\boldsymbol{\beta}_k) := \boldsymbol{V}^* \boldsymbol{C}' + \sum_{k=1}^{K} \boldsymbol{W}_k^* \operatorname{diag}(\boldsymbol{\beta}_k). \tag{12}$$

From this point onward, the development of the SPLASHX(α, λ) estimator closely mimics the reasoning from Section 3.2.1. First, we focus on the ith spatial unit and collect all the nonzero coefficients of $\boldsymbol{A}$ and $\boldsymbol{B}$ (as stipulated by Assumption 3) in $\boldsymbol{c}_i$. Letting $\boldsymbol{V}_i^*$ denote the columns in $\boldsymbol{V}^*$ related to $\boldsymbol{c}_i$ and defining both $\boldsymbol{\sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_1 & (\boldsymbol{\Sigma}_0^{x_1 y})' & \cdots & (\boldsymbol{\Sigma}_0^{x_K y})' \end{bmatrix}' \boldsymbol{e}_i$ and $\boldsymbol{w}_{ik}^* = \boldsymbol{W}_k^* \boldsymbol{e}_i$, result (12) implies $\boldsymbol{V}_i^* \boldsymbol{c}_i + \sum_{k=1}^{K} \boldsymbol{w}_{ik}^* \beta_{ik} = \boldsymbol{\sigma}_i^*$. Second, we define (a) the sample counterparts of $\boldsymbol{\Sigma}_j$, $\boldsymbol{\Sigma}_j^{x_k y}$ and $\boldsymbol{\Sigma}_j^{x_k x_l}$ as respectively $\hat{\boldsymbol{\Sigma}}_j = \frac{1}{T}\sum_{t=j+1}^{T} \boldsymbol{y}_t \boldsymbol{y}'_{t-j}$, $\hat{\boldsymbol{\Sigma}}_j^{x_k y} = \frac{1}{T}\sum_{t=j+1}^{T} \boldsymbol{x}_{t,k} \boldsymbol{y}'_{t-j}$ and $\hat{\boldsymbol{\Sigma}}_j^{x_k x_\ell} = \frac{1}{T}\sum_{t=j+1}^{T} \boldsymbol{x}_{t,k} \boldsymbol{x}'_{t-j,\ell}$, and (b) define the quantities $\hat{\boldsymbol{\sigma}}_i^*$, $\hat{\boldsymbol{w}}_{ik}^*$ and $\hat{\boldsymbol{V}}_i^*$ based on their underlying sample covariance matrix estimators. Finally, set $\hat{\boldsymbol{\sigma}}^* = (\hat{\boldsymbol{\sigma}}_1^{*\prime}, \dots, \hat{\boldsymbol{\sigma}}_N^{*\prime})'$, $\hat{\boldsymbol{V}}^{*(d)} = \operatorname{diag}(\hat{\boldsymbol{V}}_1^*, \dots, \hat{\boldsymbol{V}}_N^*)$, and $\hat{\boldsymbol{W}}_k^{*(d)} = \operatorname{diag}(\hat{\boldsymbol{w}}_{1k}^*, \dots, \hat{\boldsymbol{w}}_{Nk}^*)$. The SPLASHX($\alpha$, λ) objective function is

$$\begin{aligned} \mathscr{L}_\alpha^*(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K, \boldsymbol{c}; \lambda) = &\left\| \hat{\boldsymbol{\sigma}}^* - \hat{\boldsymbol{V}}^{*(d)} \boldsymbol{c} - \sum_{k=1}^{K} \hat{\boldsymbol{W}}_k^{*(d)} \boldsymbol{\beta}_k \right\|_2^2 \\ &+ \lambda \left(P_\alpha(\boldsymbol{c}) + \sum_{k=1}^{K} (1-\alpha)\sqrt{N} \left\| \boldsymbol{\beta}_k \right\|_2 + \alpha \left\| \boldsymbol{\beta}_k \right\|_1 \right). \end{aligned} \tag{13}$$

This objective function allows for the estimation of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$, which may contain sparse coefficients or completely sparse vectors $\boldsymbol{\beta}_k$, as well as sparse diagonals in the coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. There is a clear mathematical resemblance between the SPLASH and SPLASHX estimators. Accordingly, under appropriate modifications to Assumptions 1–4, a finding similar to Theorem 2 is attainable. For the reader's convenience, we make the correspondence between these assumptions explicit by adhering to the original numbering while adding a "*". A short discussion of these assumption is found at the end of this section.

Assumption 1*. We require: (a) $\|\boldsymbol{A}\|_{\vdash} = \max\left\{\|\boldsymbol{A}\|_1, \|\boldsymbol{A}\|_\infty\right\} \leq \delta_A < 1$, (b) $\|\boldsymbol{B}\|_{\vdash} \leq C_B$ and $\frac{C_B}{1-\delta_A} < 1$, and (c) $\left\|\boldsymbol{\beta}^*\right\|_1 = \max\{\left\|\boldsymbol{\beta}_1\right\|_1, \dots, \left\|\boldsymbol{\beta}_K\right\|_1\} \leq C_\beta$.

Assumption 2*.

(a) The regressor $x_{i_1 t_1, k}$ and innovation ϵ_{i_2, t_2} are independent of each other for all $1 \leq i_1 < i_2 \leq N$, all $1 \leq t_1, t_2 \leq T$, and all $k = 1, \dots, K$.

(b) The sequence $\{\boldsymbol{\epsilon}_t\}$ is a covariance stationary, martingale difference process with respect to the filtration $\mathscr{F}_{t-1} = \sigma\left(\boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \dots\right)$, and geometrically strong mixing (α-mixing). That is, its mixing coefficients $\{\alpha_m\}$ satisfy $\alpha_m \leq c_2 e^{-\gamma_\alpha m}$ for all m and some constants $c_2, \gamma_\alpha > 0$. The largest and smallest eigenvalues of $\boldsymbol{\Sigma}_\epsilon = \mathbb{E}(\boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}'_1) = (\sigma_{ij})_{i,j=1}^N$ are bounded away from 0 and ∞.

(c) For each $k = 1, \dots, K$, the sequence $\{\boldsymbol{x}_{t,k}\}$ is covariance stationary and geometrically strong mixing (α-mixing). That is, its mixing coefficients $\{\alpha_m^*\}$ satisfy $\alpha_m^* \leq c_2^* e^{-\gamma_\alpha^* m}$ for all m and some constants $c_2^*, \gamma_\alpha^* > 0$.

(d) Either one of the following assumptions holds:

(d1) For $\psi(x) = x^d$ and $d \geq 4$, we require $\sup_{i,t} \left\|\epsilon_{it}\right\|_\psi = \left(\mathbb{E}|\epsilon_{it}|^d\right)^{1/d} \leq \mu_d < \infty$ and $\sup_{i,t} \left\|x_{it,k}\right\|_\psi = \left(\mathbb{E}|x_{it,k}|^d\right)^{1/d} \leq \mu_d^* < \infty$ ($k = 1, \dots, K$).

(d2) For $\psi(x) = \exp(x) - 1$, we have $\sup_{i,t} \left\|\epsilon_{it}\right\|_\psi \leq \mu_\infty < \infty$ and $\sup_{i,t} \left\|x_{it,k}\right\|_\psi \leq \mu_\infty < \infty$ ($k = 1, \dots, K$).

Assumption 3*. Recall $\boldsymbol{A} = (a_{ij})_{i,j=1}^N$, $\boldsymbol{B} = (b_{ij})_{i,j=1}^N$, and $\boldsymbol{\Sigma}_\epsilon = (\sigma_{ij})_{i,j=1}^N$. We have: (a) $a_{ij} = b_{ij} = 0$ for all $|i-j| > k_0$ with $k_0 \leq \lfloor (N-1)/4 \rfloor$, and (b) $\sigma_{ij} = 0$ for all $|i-j| > l_0$.

Assumption 4* (*Restricted Minimum Eigenvalue*)**.** Define the $((K+1)N \times (K+2)N)$ matrix

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{\Sigma}_1' & \boldsymbol{\Sigma}_0 & (\boldsymbol{\Sigma}_1^{x_1 y})' & \dots & (\boldsymbol{\Sigma}_1^{x_K y})' \\ \boldsymbol{\Sigma}_0^{x_1 y} & \boldsymbol{\Sigma}_1^{x_1 y} & \boldsymbol{\Sigma}_0^{x_1 x_1} & \dots & \boldsymbol{\Sigma}_0^{x_1 x_K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_0^{x_K y} & \boldsymbol{\Sigma}_1^{x_K y} & \boldsymbol{\Sigma}_0^{x_K x_1} & \dots & \boldsymbol{\Sigma}_0^{x_K x_K} \end{bmatrix}.$$

We assume that

$$\phi^*_{\min}(\boldsymbol{x}) := \min_{\boldsymbol{x}\in\mathbb{R}^{(K+2)N}:\mathscr{M}(\boldsymbol{x})\leq N+K} \frac{\|\boldsymbol{Q}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \geq \phi^*_0.$$

Assumption 1*(a)–(b) are as before. Assumption 1*(c) merely assumes an upper bound on the magnitude of the exogenous regressor coefficients. Assumption 2* controls dependencies over time, in the cross-section, and with the exogenous regressors. Similar to assumption A8(i) in Ma et al. (2023), we enforce exogeneity through Assumption 2*(a). Assumption 2*(b) is a simple repetition of Assumption 2(b) and its counterpart for the $\{\boldsymbol{x}_{t,k}\}$'s is encountered as Assumption 2*(c). All original moment conditions are also transferred to the exogenous regressors (Assumption 3*). Clearly, Assumptions 1*–4* allow for an easy analogy with the earlier assumptions in this paper (at the cost of possibly being more restrictive than strictly necessary). That is, define $\epsilon^*_t = \epsilon_t + \sum_{k=1}^K \text{diag}(\boldsymbol{\beta}_k)\boldsymbol{x}_{t,k}$ and note the linearity of ϵ^*_t in ϵ_t and $\{\boldsymbol{x}_{t,1},\dots,\boldsymbol{x}_{t,K}\}$. With K being fixed, all mixing properties and moments conditions simply carry over to ϵ^*_t. Having modified the assumptions to accommodate the setting with exogenous variables, we are able extend Theorem 2 and derive an applicable finite-sample error bound.

Theorem 3. *Define* $\boldsymbol{q} = (\boldsymbol{c}', \boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_K')'$, $S^* = \{j : q_j \neq 0\}$ *and*

$$\bar{\omega}^*_\alpha = \max\left\{(1-\alpha)\left(\sum_{g\in\mathscr{G}_S}\sqrt{|g|} + \sqrt{N}\sum_{k=1}^K \mathbb{1}_{\{\boldsymbol{\beta}_k\neq\boldsymbol{0}\}}\right), \alpha\sqrt{|S|^*}\right\}.$$

Under Assumptions 1–4* and* $\|\boldsymbol{Q}\|_{\vdash} \leq C_Q$, *it holds that*

$$\left\|\hat{\boldsymbol{V}}^{*(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c}) + \sum_{k=1}^K \hat{\boldsymbol{W}}^{*(d)}_k(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)\right\|_2^2 + \lambda\left[(1-\alpha)\left(\sum_{g\in\mathscr{G}}\sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g - \boldsymbol{c}_g\right\|_2 + \sum_{k=1}^K\sqrt{N}\left\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\right\|_2\right)\right.$$
$$\left. + \alpha\left(\|\hat{\boldsymbol{c}}-\boldsymbol{c}\|_1 + \sum_{k=1}^K\left\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1\right)\right] \leq \frac{64\bar{\omega}^{*2}_\alpha\lambda^2}{\phi^{*2}_0}$$

with a probability of at least

(a) $1 - 7(K+1)(K+2)\mathscr{P}^*_1\left(f^*(\lambda,\phi^*_0), N, T\right)$ *under Assumption 2(b1) (polynomial tail decay), or*
(b) $1 - 7(K+1)(K+2)\mathscr{P}^*_2\left(f^*(\lambda,\phi^*_0), N, T\right)$ *under Assumption 2(b2) (exponential tail decay),*

where $f^*(\lambda,\phi^*_0) = \min\left\{\frac{\lambda^{1/2}}{12\sqrt{6}}, \frac{\lambda}{144C_Q}, \frac{\lambda^{1/2}}{12\sqrt{6}C_\beta}, \frac{\lambda}{144C_QC_\beta}, \frac{\phi^*_0}{12}\right\}$,

$$\mathscr{P}^*_1(\epsilon,N,T) = N^2\left[\left(b_1T^{(1-\delta)/3} + \frac{(K+2)Nb_3}{\epsilon}\right)\exp\left(-\frac{T^{(1-\delta)/3}}{2b_1^2}\right) + \frac{b_2(K+2)^dN^d}{\epsilon^dT^{\frac{\delta}{2}(d-1)}}\right]$$

for some $0<\delta<1$, *and*

$$\mathscr{P}^*_2(\epsilon,N,T) = N^2\left[\frac{\kappa_1(K+2)N}{\epsilon} + \frac{2}{\kappa_2}\left(\frac{T\epsilon^2}{(K+2)N}\right)^{1/7}\right]\exp\left(-\frac{1}{\kappa_3}\left(\frac{T\epsilon^2}{(K+2)^2N^2}\right)^{1/7}\right)$$

All constants (b_1, b_2, κ_1, *etc.*) *are positive and independent of* N *and* T, *see Theorem 1.*

Remark 4. The inclusion of exogenous variables affects the autocovariance structure of the data. For example, if $\boldsymbol{B} = \boldsymbol{O}$, then $\boldsymbol{y}_t = (\boldsymbol{I}_n - \boldsymbol{A})^{-1}\left[\sum_{k=1}^K \text{diag}(\boldsymbol{\beta}_k)\boldsymbol{x}_{t,k} + \boldsymbol{\epsilon}_t\right]$ and

$$\mathbb{E}(\boldsymbol{y}_t\boldsymbol{y}_t') = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}\left[\sum_{k,\kappa=1}^K \text{diag}(\boldsymbol{\beta}_k)\mathbb{E}(\boldsymbol{x}_{t,k}\boldsymbol{x}_{t,\kappa})\,\text{diag}(\boldsymbol{\beta}_\kappa) + \boldsymbol{\Sigma}_\epsilon\right](\boldsymbol{I}_N - \boldsymbol{A}')^{-1}.$$

Clearly, $\mathbb{E}(\boldsymbol{y}_t\boldsymbol{y}_t')$ now also depends on the various second moments of the exogenous covariates. We do not make any a priori assumptions on $\mathbb{E}(\boldsymbol{x}_{t,k}\boldsymbol{x}_{t,\kappa}')$ and thus define SPLASHX(α, λ) in terms of the *unbanded* autocovariance matrix estimators.

Remark 5. Defining the coefficient matrix in front of $\boldsymbol{x}_{t,k}$ as diagonal is not restrictive. That is, by letting $\boldsymbol{x}_{t,k+1}$ be a reordered version of $\boldsymbol{x}_{t,k}$, the former's addition to the model can accommodate for the situation in which the dependent variable is influenced by the exogenous variable $\boldsymbol{x}_{t,k}$ from multiple locations.

4. Simulations

4.1. Simulation setting

In this section, we explore the finite sample performance of our estimator by Monte Carlo simulation. The data generating process underlying the simulations is the spatio-temporal VAR in (1). We study $T \in \{500, 1000, 2000\}$ and draw all errors ϵ_{it} independently

and $N(0,1)$ distributed. The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and the cross-sectional dimension N are specified in the two designs below. All simulation results are based on $N_{sim} = 500$ Monte Carlo replications.

Design A (Spatial grid with neighbour interactions): As in Fig. 1, we consider an $(m \times m)$ grid of spatial units. For $m = 5$ ($m = 10$), this results in a cross-sectional dimension of $N = 25$ ($N = 100$). The matrix $\boldsymbol{A}$ contains interactions between first horizontal and first vertical neighbours while all other coefficients are zero. The magnitude of these nonzero interactions are 0.2. For $m = 5$ ($m = 10$), the temporal matrix $\boldsymbol{B}$ is a diagonal matrix with elements 0.25 (0.21) on the diagonal. The reduced form VAR matrix $\boldsymbol{C} = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}\boldsymbol{B}$ has a maximum eigenvalue of 0.814 (0.904).

Design B (Banded specification): We revisit simulation Case 1 in Gao et al. (2019). The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are banded with a bandwidth of $k_0 = 3$. Specifically, the elements in the matrices $(\boldsymbol{A})_{i,j=1}^{N}$ and $(\boldsymbol{B})_{i,j}^{N}$ are generated according to the following three steps:

Step 1: If $|i-j| = k_0$, then a_{ij} and b_{ij} are drawn independently from a uniform distribution on the two points $\{-2, 2\}$. All remaining elements within the bandwidth are drawn from the mixture distribution $\omega I_{\{0\}} + (1-\omega)N(0,1)$ with $\mathbb{P}(\omega = 1) = 0.4$ and $\mathbb{P}(\omega = 0) = 0.6$.

Step 2: Rescale the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ from Step 1 to $\eta_1 \times \boldsymbol{A}/\|\boldsymbol{A}\|_2$ and $\eta_2 \times \boldsymbol{B}/\|\boldsymbol{B}\|_2$, where η_1 and η_2 are drawn independently from $U[0.4, 0.8]$.

Step 3: To avoid unstable systems, check if $\left\|(\boldsymbol{I}_N - \boldsymbol{A})^{-1}\boldsymbol{B}\right\|_2 < 0.95$. If this is not the case, then we discard the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and return to Step 1.

We vary the cross-sectional dimension over $N \in \{25, 100\}$.

The two simulation designs are chosen to provide a balanced comparison between modelling approaches with a predetermined spatial weight matrix and modelling approaches that estimate the spatial interactions in a data-driven way.

The two methods with pre-determined spatial weight matrix are taken from Yu et al. (2008) and Lam and Souza (2019). The estimator in Yu et al. (2008) departs from a spatial–temporal autoregression (hence abbreviated as ST-AR) in which three scalar parameters and a single spatial weight matrix parametrize the spatial and temporal lag. The parameters are estimated by Gaussian quasi-maximum likelihood. As mentioned in the introduction, the method in Lam and Souza (2019), LaSo for short, allows for sparse adjustments from a predetermined spatial weight matrix. To match our simulation designs, we include $\boldsymbol{y}_{t-1}$ in the matrix of covariates (see their Section 3.3.4). For both methods, we let the spatial weight matrix coincide with the "first-nearest neighbour" interactions implied by Design A. This choice clearly favours these methods in Design A while causing a misspecified model in Design B.

The second group of methods determines the spatial interactions in a data-driven way. First, we consider three implementations of our SPLASH estimator: (1) SPLASH(0, λ) promotes non-sparse groups only, (2) SPLASH(0.5, λ) gives equal weight to sparsity at the group and individual level, and (3) SPLASH(1, λ) encourages unstructured sparsity only. In congruence with Theorems 1 and 2, we rely on banded autocovariance matrices $\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_0)$ and $\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)$ with bandwidths selected by the bootstrap procedure in Guo et al. (2016, p. 7). The included estimators from Gao et al. (2019) are references by GMWY(k) and GMWY(k_0). The first of these estimators implements generalized Yule–Walker estimation for banded $\boldsymbol{A}$ and $\boldsymbol{B}$ using the selection rule in (2.17) of their paper. Having observed a rather poor performance of this method during some exploratory simulations for Design B, we provide the true bandwidths of $\boldsymbol{A}$ and $\boldsymbol{B}$ to the (infeasible) estimator denoted GMWY(k_0). To allow for a comparison with the original simulation results, neither of these GMWY estimators applies banding to the covariance matrix estimators $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{T}\sum_{t=2}^{T}\boldsymbol{y}_t\boldsymbol{y}_t'$ and $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{T}\sum_{t=2}^{T}\boldsymbol{y}_t\boldsymbol{y}_{t-1}'$. During additional exploratory simulations, we have seen that the performance of the GMWY estimator does not change much if banded covariance matrices would be used.

Finally, we also include an L_1-penalized reduced form VAR(1) estimator (abbreviated PVAR). In detail, we consider the reduced form VAR(1) specification $\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{y}_{t-1} + \boldsymbol{u}_t$ and estimate $\boldsymbol{C}$ by minimizing $\mathscr{L}_{pvar}(\boldsymbol{C}) = \sum_{t=2}^{T}\|\boldsymbol{y}_t - \boldsymbol{C}\boldsymbol{y}_{t-1}\|_2^2 + \lambda\sum_{i,j=1}^{N}|c_{ij}|$. This estimator is well-researched in the literature (see, e.g. Kock and Callot, 2015; Gelper et al., 2016; Masini et al., 2022), albeit in different settings. The inclusion of this estimator serves to uncover any difference in predictive performance stemming from the estimation of a spatio-temporal representation, as opposed to estimating a purely temporal specification.

The forecasting performance of each estimator will be assessed using the *Relative Mean-Squared Forecast Error (RMSFE)*. Using a superscript j to index a specific Monte Carlo replication, the RMSFE is calculated as

$$\text{RMSFE} = \frac{\sum_{j=1}^{N_{sim}}\left\|\boldsymbol{y}_{T+1}^{j} - \hat{\boldsymbol{C}}^{j}\boldsymbol{y}_{T}^{j}\right\|_2^2}{\sum_{j=1}^{N_{sim}}\left\|\boldsymbol{y}_{T+1}^{j} - \boldsymbol{C}\boldsymbol{y}_{T}^{j}\right\|_2^2}. \tag{14}$$

The estimation accuracy of the spatio-temporal methods is analysed via the mean estimation error (MEE). Using the superscript j as before, the (MEE) is defined as

$$\text{MEE} = \frac{1}{N_{sim}}\sum_{j=1}^{N_{sim}}\left\|\left(\hat{\boldsymbol{A}}^{j}, \hat{\boldsymbol{B}}^{j}\right) - \left(\boldsymbol{A}^{j}, \boldsymbol{B}^{j}\right)\right\|_2. \tag{15}$$

To provide some insights into the variability in estimation accuracy across simulation trials, we additionally report the standard errors of the MEEs, which we compute as

$$\text{SE(MEE)} = \sqrt{\frac{1}{N_{sim}-1}\sum_{j=1}^{N_{sim}}\left(\left\|\left(\hat{\boldsymbol{A}}^{j}, \hat{\boldsymbol{B}}^{j}\right) - \left(\boldsymbol{A}^{j}, \boldsymbol{B}^{j}\right)\right\|_2 - EE\right)^2}. \tag{16}$$

Next, a word on the determination of hyperparameters. For each regularized estimator, we construct a unique grid containing $K = 20$ penalty values that are evenly spaced on a logarithmic scale. The maximum value in each grid, $\lambda_{\max}$, is the smallest value that produces the zero solution (see step 4 in Algorithm 1 for its expression for SPLASH). Given $\lambda_{\max}$, we define the smallest penalty as $\lambda_{\min} = \eta\lambda_{\max}$ with $\eta = 10^{-4}$ ($\eta = 10^{-6}$) for Design A (B). Estimating SPLASH solutions for each $\lambda \in \{\lambda_1, \dots, \lambda_{20}\}$, each $\alpha \in \{0, 0.5, 1\}$, and each individual simulation trial remains too computationally expensive (especially for large N). We instead perform a small-scale preliminary analysis in which we draw a small set of simulations from Designs A and B on which we estimate all solutions for a given value of T. Then, we choose the order $i_T \in \{1, \dots, 20\}$ that minimizes the RMSFE in this preliminary set of simulations. This process of choosing the order i_T on a log-equidistant grid for each value of T, is equivalent to setting $\lambda = m_T\lambda_{\max}$ with $m_T = 10^{-4(i_T-1)/20}$ or $m_T = 10^{-6(i_T-1)/20}$ for designs A and B, respectively. For Design A (B), our selected orders for $T = \{500, 1000, 2000\}$ are $i_T = \{9, 10, 11\}$ ($i_T = \{10, 11, 12\}$), corresponding to $m_T \approx 0.025, 0.015, 0.01$ ($m_T \approx 0.002, 0.001, 0.0005$), respectively. Having fixed the preferred order or multiplier, it suffices to compute the SPLASH solution for each $\alpha \in \{0, 0.5, 1\}$. The penalized VAR method is computationally less expensive. Accordingly, we choose its penalty parameter based on a time series cross-validation (TSCV) scheme (e.g. Hyndman and Athanasopoulos, 2018). Since a preliminary analysis showed that the penalty selection approach used for SPLASH resulted in sub-par performance for the LaSo method, we choose the latter's penalty via TSCV as well. In our implementation of TSCV, the first 80% of the data is used to fit multiple solutions on, which are then evaluated based on the MSFE obtained on the latter 20% of the data. The preferred penalty is chosen as the solution that attains the smallest MSFE.[3]

4.2. Simulation results

The simulation outcomes for Design A and B are reported in Table 1. For each design we report the predictive performance in terms of the RMSFE and the estimation accuracy as measured by the mean estimation error (MEE). Whenever applicable, the standard errors of the latter quantity is written in parentheses.

First, we consider the predictive performance under Design A. For all methods, we observe a monotonic decrease in RMSFE when T increases. Unsurprisingly, the best performance is obtained by the ST-AR method which only needs to estimate two non-zero parameters to recover the DGP. More impressively, with a maximum RMSFE of 1.013, all three SPLASH implementations are close to oracle performance. This demonstrates that the cost incurred by estimating the DGP in a fully data-driven way, as opposed to relying on a correctly specified pre-determined weight matrix, is small. There is little difference across the three SPLASH implementations, with the effect of group-level regularization providing noteworthy improvements only in the most high-dimensional setting ($N = 100$ and $T = 500$). LaSo obtains a strong predictive accuracy as well, albeit slightly worse than SPLASH when T is smaller. This is somewhat surprising as the LaSo specification has all the correct spatial interactions readily included in its spatial weight matrix and simply needs to estimate the redundant adjustment matrix as zero. However, this requires a sufficient high penalty to be selected via time series cross-validation which was not the case across simulation trials. Additionally, the LaSo estimator does not always convergence and remains rather sensitive to its initialization. The RMSFEs for the GMWY estimators are strikingly high. In the setting $N = 25$ and $T = 500$, the GMWY estimator frequently underestimates the bandwidth by choosing it to equal 1. This causes inferior performance across all metrics. The GMWY(k_0) estimator, on the other hand, is based on the correct bandwidth, but its forecast performance is worse. Upon closer inspection, we find that these high RMSFEs are caused by several extreme prediction errors. These prediction outliers occur during simulation trials in which the smallest eigenvalue of the estimated matrix $\boldsymbol{I} - \hat{\boldsymbol{A}}$ is close to zero. As the forecasts are based on the reduced form representation and hence $(\boldsymbol{I} - \hat{\boldsymbol{A}})^{-1}$, the GMWY estimator is prone to stability issues when the bandwidth is large relative to the dimension. For $N = 25$ and $T = 2000$, the bandwidth selection in GMWY improves, while its forecast performance ironically worsens as a result of increasing stability issues. Finally, the penalized VAR produces stable forecasts but is outperformed by ST-AR and SPLASH across all settings. A correctly specified spatio-temporal model can thus outperform a reduced form VAR in terms of forecast performance.

Parallelling the results on the predictive performance, the ST-AR attains the lowest (variability in) MEE, followed by the SPLASH estimator. The MEE shows the added benefit of group-level regularization in SPLASH as the implementations for $\alpha = 0$ and $\alpha = 0.5$ attain substantially lower estimation errors. For the SPLASH(0.5, λ) estimator, Fig. 3 illustrates the similarity between the true $\boldsymbol{A}$ and its estimated counterpart. Indeed, owing to the imposed group-level regularization, the true sparsity pattern in $\boldsymbol{A}$ is clearly present in the (averaged) SPLASH estimates. Continuing the comparison, the LaSo method displays mixed performance, varying from very large ($N = 25$ and $T = 500$) to very small ($N = 100$ and $T = 200$) MEEs. This again reflects the inability to consistently select a sufficiently large penalty parameter, combined with the aforementioned issues with numerical convergence. Finally, GMWY(k) typically obtains a lower MEE than GMWY(k_0) as a result of choosing a lower bandwidth. Apparently, estimating a misspecified model by choosing $k < k_0$ results in a favourable bias–variance trade-off, highlighting the importance of regularization in high-dimensional settings.

Finally, we look at Design B. The methods that rely on a prespecified spatial weight matrix no longer have a competitive advantage as its interactions are misspecified. Consequently, the SPLASH estimators and the infeasible GMWY(k_0) estimator now exhibit the best overall forecast performance. The feasible GMWY(k) estimator shows good performance for $N = 100$ and $T \in \{1000, 2000\}$ because the increase in spatial units and time points improves bandwidth selection and diminishes invertibility issues. Performance differences between the SPLASH estimators are generally small as no specific type of sparsity dominates. That is, the group penalty will help in setting to zero the elements outside the bandwidth k_0 and the individual coefficient penalties are

[3] We also tried to select the penalty as the sparsest solution whose prediction error lies within one standard error of the minimum prediction error. This selection rule, however, did not lead to an improvement in forecast or estimation accuracy.

Table 1
Simulation results for Design A and B.

Metric	N	T	SPLASH			ST-AR	LaSo	GMWY		PVAR
			$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$			k	k_0	
Design A (Spatial grid with neighbour interactions)										
RMSFE	25	500	1.012	1.011	1.012	1.001	1.109	9.924	539.490	1.108
		1000	1.004	1.004	1.005	1.000	1.040	29.538	66.157	1.067
		2000	1.005	1.005	1.004	1.000	1.003	894.224	527.024	1.042
	100	500	1.013	1.014	1.022	1.000	1.526	1.151	463.714	1.158
		1000	1.010	1.011	1.012	1.000	1.000	1.104	137.901	1.109
		2000	1.005	1.004	1.004	1.000	1.000	1.087	2.300	1.077
MEE	25	500	0.334	0.342	0.477	0.019	3.381	1.162	4.448	
			(0.015)	(0.015)	(0.022)	(0.001)	(0.164)	(0.052)	(0.212)	
		1000	0.283	0.281	0.385	0.013	1.416	1.173	4.062	
			(0.013)	(0.013)	(0.017)	(0.001)	(0.081)	(0.053)	(0.187)	
		2000	0.243	0.232	0.295	0.009	0.207	1.212	3.956	
			(0.011)	(0.010)	(0.013)	(0.000)	(0.022)	(0.054)	(0.183)	
	100	500	0.393	0.411	0.573	0.009	0.768	1.119	2.174	
			(0.018)	(0.019)	(0.026)	(0.000)	(0.082)	(0.050)	(0.098)	
		1000	0.367	0.379	0.537	0.007	0.027	1.066	2.082	
			(0.017)	(0.017)	(0.024)	(0.000)	(0.002)	(0.048)	(0.094)	
		2000	0.373	0.362	0.504	0.005	0.019	1.046	2.059	
			(0.017)	(0.016)	(0.023)	(0.000)	(0.001)	(0.047)	(0.093)	
Design B (Banded specification)										
RMSFE	25	500	1.025	1.026	1.029	1.154	1.495	5.485	1.048	1.125
		1000	1.011	1.011	1.012	1.159	1.444	1.387	1.016	1.116
		2000	1.007	1.007	1.008	1.156	2.699	1.008	1.008	1.109
	100	500	1.052	1.058	1.074	1.140	1.316	1.051	1.037	1.110
		1000	1.025	1.028	1.035	1.136	14.308	1.017	1.016	1.104
		2000	1.017	1.018	1.021	1.135	6.929	1.009	1.009	1.102
MEE	25	500	0.723	0.761	0.883	0.750	1.338	24.662	1.222	
			(0.033)	(0.035)	(0.040)	(0.034)	(0.071)	(5.656)	(0.056)	
		1000	0.656	0.687	0.798	0.744	1.386	4.272	1.134	
			(0.030)	(0.032)	(0.037)	(0.034)	(0.073)	(0.923)	(0.052)	
		2000	0.600	0.617	0.709	0.750	1.426	1.829	1.064	
			(0.028)	(0.029)	(0.033)	(0.034)	(0.076)	(0.651)	(0.049)	
	100	500	0.831	0.894	1.075	0.735	1.203	1.027	0.799	
			(0.038)	(0.041)	(0.049)	(0.033)	(0.064)	(0.049)	(0.036)	
		1000	0.830	0.889	1.051	0.733	1.143	0.755	0.707	
			(0.038)	(0.041)	(0.048)	(0.033)	(0.060)	(0.035)	(0.032)	
		2000	0.795	0.849	0.994	0.735	1.203	0.621	0.617	
			(0.037)	(0.039)	(0.046)	(0.033)	(0.064)	(0.028)	(0.028)	

Note: The relative mean-squared forecast error (RMSFE) and mean estimation errors (MEE) are defined in (14) and (16), respectively. In general, lower numbers indicate better performance. As PVAR estimates a reduced form VAR, there are no model errors for $\boldsymbol{A}$ and $\boldsymbol{B}$ to report for this method. Standard errors of the model errors are listed in parentheses.

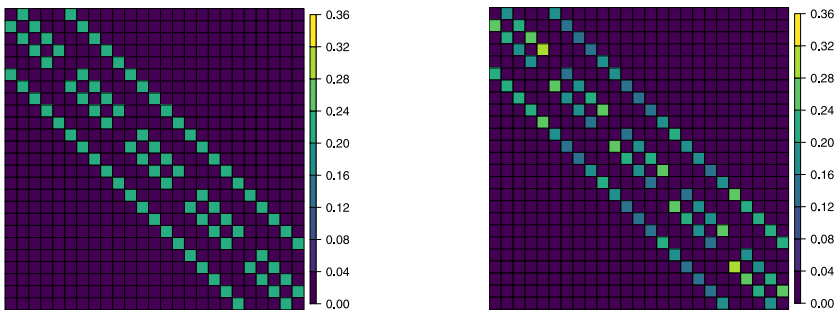

Fig. 3. Visualizations of the true and estimated spatial weight matrix $\boldsymbol{A}$ for Design B. **(a)** The true spatial weight matrix $\boldsymbol{A}$ implied by the (5×5) spatial grid design (Design B with $m = 5$). **(b)** The average absolute values of the entries in $\hat{\boldsymbol{A}}$ as computed by SPLASH(0.5, λ) for $N = 25$ and $T = 1000$. That is, the (i, j)th entry in the matrix on the right equals $\frac{1}{N_{sim}} \sum_{k=1}^{N_{sim}} |\hat{a}_{ij}^k|$ with $\hat{a}_{ij}^k$ being the estimated (i, j)th entry of $\boldsymbol{A}$ in the kth Monte Carlo replication.

useful to reveal the random zero pattern that occurs close to the main diagonals of $\boldsymbol{A}$ and $\boldsymbol{B}$. Only being able to indirectly incorporate contemporaneous interactions, the reduced form PVAR estimator performs neither badly nor among the top contenders. Finally, in terms of estimation accuracy, SPLASH appears to obtain the lowest MEE for $N = 25$, but is surpassed by GMWY(k_0) for $N = 100$. Surprisingly, ST-AR performs very competitive in terms of estimation accuracy, reflecting a beneficial bias–variance trade-off by estimating a misspecified low-dimensional approximation to the DGP.

5. Empirical application

In this section we consider the application of SPLASH to satellite-measured data. Section 5.1 serves as a general guide to practitioners. It explains the important steps in the analysis and provides additional details on the satellite data collection and processing. A concrete illustration is presented in Section 5.2 where we predict daily nitrogen dioxide (NO_2) concentrations over Greater London.

5.1. Applying SPLASH: A guide to practitioners

To facilitate the adoption of our SPLASH modelling strategy, we explain in detail how to download, process and model data from the Copernicus satellite. A readily available implementation of each step, together with the raw data used in the preceding application, is available at https://sites.google.com/view/etiennewijler/code in the form of separate code scripts and an R package.

Step 1: Data collection Satellite images are downloadable from the Copernicus Open Access Hub.[4] First-time users are required to register a free (at the time of writing) account. Single files can be downloaded interactively via a user interface, whereas repeated requests are best handled via the API hub. After defining the area of interest (AOI) based on GPS coordinates, all satellite images that overlap with this AOI can be downloaded. The AOI is subsequently cropped out and the original image is deleted to free storage space. Sentinel-5P Copernicus data products include ozone, sulphur dioxide, nitrogen dioxide, carbon monoxide, formaldehyde and methane. The user might also use satellite data from a different data source.

Step 2: Data processing The cropped raw satellite data is mapped onto the spatial grid. All pixels within a given grid cell are averaged into a single value. Enumerating the cells consecutively (by rows or by columns), the vector $\boldsymbol{y}_t$ with daily observations is constructed. This vector is likely to have some missing values as the satellite collects data while orbiting the earth. These missing observations should be imputed, e.g. using `Multivariate Time Series Data Imputation (mtsdi)` R package. This imputation method in Junger and Ponce De Leon (2015) was designed to impute missing values in air pollution data. Finally, one should assess/model the low-frequency seasonality in the data. That is, applying model (1) or (10) on daily data does not capture the slow concentration variation with the seasons due to the difference in frequency. We provide a possible approach in the next section.

Step 3: Data modelling The SPLASH R package fully automates the model-building process by estimating the bandwidth and banded covariance matrices, calculating the quantities of interest, and selecting the optimal penalty parameter by time series cross-validation (all steps in Algorithm 1). The sparsely estimated coefficients give an idea of the important interactions and the implied reduced form VAR can be used for out-of-sample predictions.

5.2. Predicting NO_2 concentrations based on satellite data

NO_2 is emitted during combustion of fossil fuels (e.g. by motor vehicles) and it has been associated with adverse effects on the respiratory system.[5] The Air Quality Standards Regulations 2010 requires a regular monitoring of NO_2 concentration levels in the UK.[6] Using satellite data, we examine the empirical performance of the SPLASH estimator when predicting daily NO_2 concentrations in Greater London. The data is available at the Copernicus Open Access Hub and we consider the time span from 1 August 2018 to 31 March 2023. The original NO_2 concentrations are reported in mol/m^2, which we convert to mol/cm^2 to avoid numerical instabilities caused by small-scale numbers. The far majority of measurements are captured between 11:00 and 14:00 UTC. As displayed in Fig. 5(c), the area of interest is mapped onto a (6×7) grid with even straight-line distances between the centres of horizontally (8.57 km) and vertically (8.53 km) neighbouring grid cells. We average pixels and impute missing observations as in Step 2 of our guide to practitioners.

A first inspection of the data reveals several noteworthy features. First, the average NO_2 concentrations (averaged over space and time) is 1.1689 mol/cm^2. Furthermore, there is substantial regional variation in the average daily NO_2 concentrations. The lowest daily average concentration of 0.9671 mol/cm^2 is observed over Chipping Ongar (a small market town North-East of the Greater London Area) and the highest concentration of 1.3479 mol/cm^2 is measured at Westminster (the bustling government area near Buckingham Palace).

[4] https://www.copernicus.eu/en/access-data/conventional-data-access-hubs

[5] The direct health effect of nitrogen dioxide is difficult to determine because its emission process is typically accompanied with the emission of other air pollutants (see, e.g. Brunekreef and Holgate, 2002).

[6] Source: https://www.legislation.gov.uk/uksi/2010/1001/contents/made.

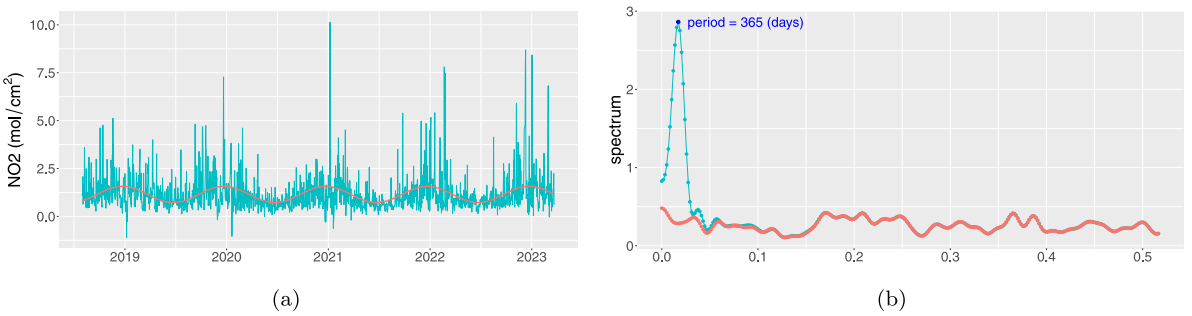

Fig. 4. The seasonality in the data. **(a)** The daily spatially averaged NO_2 concentration (green) and its first-order Fourier approximation at the angular frequency of $\omega_{yearly} = \frac{2\pi}{355} \approx 0.0172$ (red). **(b)** The spectral density estimate of the spatially averaged NO_2 concentration before (green) and after (red) subtracting the first-order Fourier approximation. The spectral densities are computed using the Bartlett kernel (e.g. (6.2.15) in Hamilton, 1994) with a bandwidth of $q = 400$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We proceed to the final part of Step 2 in our guide to practitioners and investigate the seasonality in the data. First, in Fig. 4(a) we visualize the spatially averaged NO_2 concentration over time. The plot reveals that this concentration is typically highest in December and lowest in June thereby pointing towards seasonality. To get an idea about the relevant frequencies, we look at the (estimated) spectral density. The clear peak at the angular frequency of $\omega_{yearly} = \frac{2\pi}{355} \approx 0.0172$ motivates us to include sine and cosine signals with this baseline period of 365 days. As the residuals after the inclusion of a single Fourier term already produce a rather flat spectrum (Fig. 4(b)) and further Fourier terms hardly change the spectrum, we decide to detrend the data with this first-order Fourier approximation throughout this empirical application. The Fourier approximation itself is displayed as the red line in Fig. 4(a).

A rolling-window approach is used to assess the predictive power of the SPLASH estimator. Motivated by our seasonality analysis, we assume that the data is generated by the following model:

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{d}_t + \boldsymbol{y}_t, \\ \boldsymbol{d}_t &= \boldsymbol{\mu} + \boldsymbol{\iota}_N \left(\sum_{j=1}^{M} a_j \cos(j\omega_{yearly} t) + b_j \sin(j\omega_{yearly} t) \right), \\ \boldsymbol{y}_t &= \boldsymbol{A}\boldsymbol{y}_t + \boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \end{aligned}$$

where $\boldsymbol{\iota}_N$ in an N-dimensional vector of ones. As mentioned in our spectral analysis, we select $M = 1$ (but also remind the reader that a different data set might require additional Fourier terms). Hence, each spatial unit is assigned its own mean and a common seasonal term. Each window contains 2 years of the data (730 days) allowing 964 one-step ahead forecasts to be made. For each window, we proceed along the following four steps: (i) regress out the deterministic component d_t to obtain the de-seasonalized data $\hat{y}_t = z_t - \hat{d}_t$, (ii) estimate the (hyper)parameters of each model based on $\hat{y}_t$, (iii) produce separate one-step ahead forecasts for the de-seasonalized data and the deterministic components, and (iv) add the forecasts together to obtain a forecast for the NO_2 concentration observed directly after the window.

We consider the same type of estimators as previously seen in the simulations. As before, we include SPLASH(0, λ), SPLASH(0.5, λ), and SPLASH(1, λ). With the current spatial grid of $N = 6{\times}7 = 42$ spatial units, these methods potentially require $2N^2 - N = 3486$ parameters. However, for identifiability (see Assumption 3), we band the spatial matrix $\boldsymbol{A}$ and autoregressive matrix $\boldsymbol{B}$ such that $a_{ij} = b_{ij} = 0$ for $|i - j| > \lfloor (N-1)/4 \rfloor = 10$. Similar to the simulation section, we select the penalty parameter as $\lambda = m\lambda_{max}$ with the multiplicative factor m being determined based on a preliminary set of out-of-sample forecasts. Having no a priori information concerning the spatial interactions, we implement the ST-AR and LaSo methods with two spatial weight matrices: the "first-nearest neighbour" specification is identical to the spatial grid (SG) in Design A of Section 4, and the (normalized) inverse distance (ID) spatial weight matrix $\boldsymbol{W} = (w_{ij})_{i,j=1}^{N}$ with each element w_{ij} representing the inverse distance in kilometres between spatial units i and j. The GMWY estimator is implemented with bandwidth $k = 10$, as this provides the best (i.e. least unstable) predictions. Penalty selection for LaSo and PVAR follow the same TSCV scheme discussed in Section 4.1.

The forecast performance is measured along three metrics. We report: (i) the average loss relative to PVAR, (ii) the number of spatial units that are predicted more accurately than the PVAR benchmark (#wins), and (iii) the inclusion in the Model Confidence Set (MCS) by Hansen et al. (2011) at a 10% significance level. These three metrics are calculated using two loss functions for the forecast errors: the mean squared forecast error (MSFE) and the mean robust forecast error (MRFE). That is, letting $\hat{e}_{i,T_0+h}^{(j)}$ denote the one-step ahead forecast error of estimator j for the outcome of the ith spatial unit at period $T_0 + h$, we calculate

$$\text{MSFE}_j = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \left(\hat{e}_{i,T_0+h}^{(j)} \right)^2, \text{ and } \text{MRFE}_j = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \frac{\left(\hat{e}_{i,T_0+h}^{(j)} \right)^2}{1 + \left(\hat{e}_{i,T_0+h}^{(j)} \right)^2}, \tag{17}$$

Table 2
Forecast performance of various methods for NO_2 satellite data on a (6×7) grid of observations.

	MSFE			MRFE		
	RMSFE	#wins	MCS	RMRFE	#wins	MCS
SPLASH(0, λ)	**0.448**	41	Yes	0.885	41	Yes
SPLASH(0.5, λ)	0.448	41	Yes	0.886	41	Yes
SPLASH(1, λ)	0.450	41	Yes	0.889	40	Yes
ST-AR(SG)	0.452	40	Yes	0.885	41	Yes
ST-AR(ID)	0.450	42	Yes	**0.885**	39	Yes
LaSo(SG)	0.503	27	Yes	0.892	36	Yes
LaSo(ID)	0.464	33	Yes	0.890	37	Yes
GMWY	36.056	0	No	1.759	0	No
PVAR	1.000	–	No	1.000	–	No

Note: Number of grid points (out of $N = 45$) with lower prediction errors (#wins) and significantly lower prediction errors (#sign. wins) compared to the L_1-penalized reduced form VAR(1) estimator (PVAR). Relative MSFE and MRFE are abbreviated by RMSFE and RMRFE, respectively. Values below (above) 1 indicate superior (inferior) performance compared to PVAR.

with $H = T - T_0$ the maximum forecast horizon. The MRFE is included to reduce the importance of incidental outliers caused by several abrupt spikes in the NO_2 column densities and/or the stability issues that continued to plague the GMWY method (see Section 4 for explanations). In Table 2 we report the *relative* MSFE (RMSFE) and *relative* MRFE (RMRFE) as the performance in comparison to the PVAR estimator.

We first look at the columns under MSFE. The RMSFEs show that all spatio-temporal methods, with the exception of GMWY, improve substantially over the purely penalized VAR benchmark. Incorporating spatial effects is thus important while modelling NO_2 concentrations empirically. The lowest MSFE is obtained by SPLASH(0, λ) but the differences in forecasting performance in comparison to ST-AR and LaSo are small. For GMWY, we again observed significant stability issues when inverting the estimated spatial weight matrix to obtain the reduced form representation that is necessary to construct the forecast, resulting in inferior forecasts. We were not able to resolve this by choosing a smaller bandwidth, possibly because such specifications may start to omit important spatial interactions. The SPLASH implementations predict 41 out of 42 spatial units more accurately than the PVAR benchmark, second only to the ST-AR (ID) approach which manages to improve the predictions for all spatial units on the grid. The LaSo approach performs less favourable here, beating the benchmark for only 27 or 33 spatial units based on the SG and ID spatial weight matrix, respectively. Finally, the MCS finds statistically significant evidence that the GMWY and PVAR methods obtain inferior forecasts, but fails to find any significant performance differences among the remaining methods.

Next, we focus on the mean robust forecast error (MRFE). The results are qualitatively similar. The RMRFEs for SPLASH, ST-AR and LaSo lie even closer together and are practically indistinguishable. The GMWY still displays the worst forecast performance, although the difference appears less dramatic as a result of the robust loss metric. SPLASH(α, λ) with $\alpha = 0, 0.5$ and the ST-AR(SG) method are able to beat the PVAR benchmark most often by obtaining lower MRFES for 41 out of 42 spatial units. The MCS again includes all methods except GMWY and PVAR.

We end this application with a visual analysis of the estimated spatial interactions between neighbouring regions in Greater London in Fig. 5. To have both individual and group sparsity, we focus on the SPLASH(0.5, λ) estimates. First, the window-averaged absolute magnitude of the spatial interactions is displayed in Fig. 5(a). A clear diagonal pattern emerges with the largest interactions clustering on the two diagonals closest to the main diagonal and the two outer diagonals that start at the (1, 7) and (7, 1) elements. These four diagonals correspond to first-order vertical and horizontal interactions, respectively. For each element in $\boldsymbol{A}$, the heat map in Fig. 5(b) indicates the fraction of rolling windows with a nonzero estimated coefficient. The aforementioned interactions between first-order horizontal and vertical neighbours are selected in all windows. Additional interactions between diagonal neighbours are relevant in approximately half of the rolling windows. To facilitate the interpretation of this sparsity pattern, we provide a spatial plot of our region of interest with the spatial grid overlaid in Fig. 5(c). For spatial unit 21 specifically, we visualize the spatial interactions that are estimate to be nonzero in at least 20% of the rolling windows and vary the thickness of the arrows to indicate their (average) estimated magnitudes. The strongest interactions observed are between first-order vertical neighbours, followed by those between first-order horizontal neighbours. Interactions between diagonal and second-order vertical neighbours are observed as well, although their average effects are substantially weaker. The fact that the diagonal interactions only seem to occur along the diagonal from South-West to North-East may be a consequence of the predominant South-West winds over London. Overall, the intuitive sparsity patterns that arise, in combination with the improvement in forecast performance, are encouraging and provide empirical validation for the use of SPLASH on spatial data, especially when the spatial units follow a natural ordering on a spatial grid. The knowledge about the relevant interactions as well as their magnitudes might be exploited by policymakers to study the effects of for example stricter emission regulations and/or traffic decrease on regional air quality. However, the analysis of these spatial impulse response would require additional results on statistical inference (see Section 3.2.4).

6. Conclusion

In this paper, we develop the Spatial Lasso-type Shrinkage (SPLASH) estimator, a novel estimation procedure for high-dimensional spatio-temporal models. The SPLASH estimator is designed to promote the recovery of structured forms of sparsity

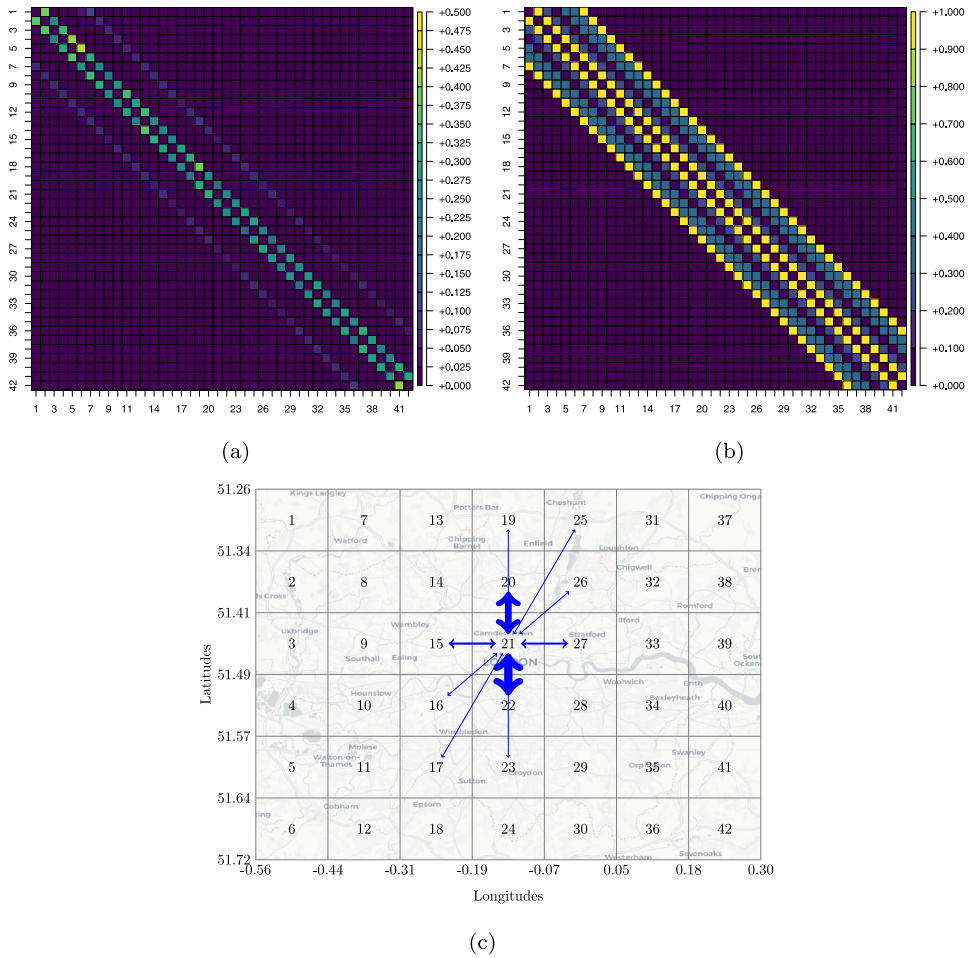

Fig. 5. Illustrations of the sparsity patterns in the estimated (42×42) matrix $\boldsymbol{A}$ based on rolling window samples. **(a)** The average absolute value of the entries in $\boldsymbol{A}$ as averaged across each one-step ahead forecast. **(b)** The proportion of rolling window forecast with an estimated coefficient being unequal to zero. **(c)** The spatial interaction for spatial unit 21 that are non-zero in at least 20% of the forecasts.

without imposing such structure a priori. We derive consistency of our estimator in an asymptotic framework in which the number of both spatial units and temporal observations diverge jointly. To solve the identifiability issue, we rely on a relatively non-restrictive assumption that the coefficient matrices in the spatio-temporal model are sufficiently banded. Based on this assumption, we consider banded estimation of high-dimensional spatio-temporal autocovariance matrices, for which we derive novel convergence rates that are likely to be of independent interest. The SPLASHX extension explains how to include exogenous variables. As an application, we use SPLASH to predict satellite-measured NO_2 concentrations in London. We find evidence for spatial interactions between neighbouring regions. In addition, our estimator obtains superior forecast accuracy compared to a number of competitive benchmarks, including the recently introduced spatio-temporal estimator by Gao et al. (2019) that inspired the development of SPLASH.

Acknowledgements

This paper (or earlier versions hereof) has been presented during the internal seminar of the Quantitative Economics department of Maastricht University, the Econometrics Internal Seminar (EIS) at Erasmus University Rotterdam, the Bernoulli-IMS One World Symposium, the workshop on Dimensionality Reduction and Inference in High-dimensional Time Series at Maastricht University, the 2021 Annual Conference of the International Association for Applied Econometrics (IAAE), the 5th Conference on Econometric Models of Climate Change, the internal seminar at Tor Vergata University of Rome, and the 2021 $(EC)^2$ Conference. We gratefully acknowledge comments and feedback from the participants. Suggestions by Stephan Smeekes and Ines Wilms were particularly helpful, so we thank them explicitly. In addition, we thank the two anonymous referees for their constructive feedback that has resulted in major improvements in our work. All remaining errors are our own.

Appendix A. Lemmas

The proofs of the lemmas in this section are available in the supplementary material.

Lemma 1. *Define the quantities* $N_c = |c|$, $S = \{j : c_j \neq 0\}$, $\mathscr{G}_S = \{g \in \mathscr{G} : \boldsymbol{c}_g \neq \boldsymbol{0}\}$, $\mathscr{G}_S^c = \{g \in \mathscr{G} : \boldsymbol{c}_g = \boldsymbol{0}\}$, $\bar{\omega}_\alpha = \max\left\{(1-\alpha)\sum_{g\in\mathscr{G}_S}\sqrt{|g|}, \alpha\sqrt{|S|}\right\}$ *and consider*

$$\boldsymbol{\Delta} \in \mathscr{C}_{N_c}(\mathscr{G}, S) := \left\{\boldsymbol{\Delta} \in \mathbb{R}^{N_c} : P_{\alpha,S^c}(\boldsymbol{\Delta}) \leq 3P_{\alpha,S}(\boldsymbol{\Delta})\right\},$$

where

$$P_{\alpha,S^c}(\boldsymbol{\Delta}) = (1-\alpha)\sum_{g\in\mathscr{G}_S^c}\sqrt{|g|}\left\|\Delta_g\right\|_2 + \alpha\left\|\boldsymbol{\Delta}_{S^c}\right\|_1, \text{ and}$$

$$P_{\alpha,S}(\boldsymbol{\Delta}) = (1-\alpha)\sum_{g\in\mathscr{G}_S}\sqrt{|g|}\left\|\Delta_g\right\|_2 + \alpha\left\|\boldsymbol{\Delta}_S\right\|_1.$$

Then, under Assumption 4, it holds that

$$\min_{\boldsymbol{\Delta}\in\mathscr{C}_{N_c}(\mathscr{G},S)} \frac{\bar{\omega}_\alpha\left\|\boldsymbol{V}^{(d)}\boldsymbol{\Delta}\right\|_2}{P_{\alpha,S}(\boldsymbol{\Delta})} \geq \frac{\phi_0}{2}. \tag{A.1}$$

Lemma 2. *Define the set* $\mathscr{V}(x) := \left\{\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_2 \leq x\right\}$. *Then, under Assumption 4, it holds on* $\mathscr{V}\left(\frac{\phi_0}{4}\right)$ *that*

$$\min_{\boldsymbol{x}\in\mathscr{C}_{N_c}(\mathscr{G},S)} \frac{\bar{\omega}_\alpha\left\|\hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{x}\right\|_2}{P_{\alpha,S}(\boldsymbol{x})} \geq \frac{\phi_0}{4}.$$

Lemma 3. *Define the* $N(p+1)$ *vector* $\boldsymbol{\xi}_t = (\boldsymbol{y}_t', \boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-p}')'$. *For all* $j = 1, \ldots, p$, *the elements of* $\hat{\boldsymbol{\Sigma}}_j = \frac{1}{T}\sum_{t=p+1}^T \boldsymbol{y}_t\boldsymbol{y}_{t-j}'$ *can be expressed as* $\frac{1}{T}\sum_{t=p+1}^T \xi_{it}\xi_{jt}$ *after an appropriate choice of* (i,j).

(a) If Assumptions 1 and 2(b1) (polynomial tails) hold, then we have:

$$\mathbb{P}\left(\left|\sum_{t=p+1}^T \xi_{it}\xi_{jt} - \mathbb{E}(\xi_{it}\xi_{jt})\right| > T\epsilon\right) \leq \left[b_1 T^{(1-\delta)/3} + \frac{b_3}{\epsilon}\right]\exp\left(-\frac{T^{(1-\delta)/3}}{2b_1^2}\right) + \frac{b_2}{\epsilon^d T^{\frac{\delta}{2}(d-1)}},$$

for some $0 < \delta < 1$.

(b) If Assumptions 1 and 2(b2) (subexponential tails) hold, then we have:

$$\mathbb{P}\left(\left|\sum_{t=p+1}^T \xi_{it}\xi_{jt} - \mathbb{E}(\xi_{it}\xi_{jt})\right| > T\epsilon\right) \leq \left[\frac{\kappa_1}{\epsilon} + \frac{2(T\epsilon^2)^{1/7}}{\kappa_2}\right]\exp\left(-\frac{(T\epsilon^2)^{1/7}}{\kappa_3}\right).$$

Explicit expressions for the constants b_1, b_2, b_3, κ_1, κ_2 *and* κ_3 *are provided in the proof.*

Appendix B. Proofs of main results

Proof of Theorem 1. We first prove various intermediate results, see (a)–(d) below. We afterwards combine these results and recover Theorem 1.

(a) The matrix $\widetilde{\boldsymbol{C}}_s = \left(\sum_{j=0}^{s-1}\boldsymbol{A}^j\right)\boldsymbol{B} =: \widetilde{\boldsymbol{D}}_s\boldsymbol{B}$ has a maximum bandwidth of $(s+1)(k_0-1)+1$ and satisfies

$$\left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash} \leq \delta_C\delta_A^s.$$

(b) Define $\boldsymbol{\Sigma}_0^{r,s} = \sum_{j=0}^r \widetilde{\boldsymbol{C}}_s^j\widetilde{\boldsymbol{D}}_s\boldsymbol{\Sigma}_\epsilon\widetilde{\boldsymbol{D}}_s'(\widetilde{\boldsymbol{C}}_s')^j$ with $\widetilde{\boldsymbol{C}}_s$ as in Theorem 1(a). The matrix $\boldsymbol{\Sigma}_0^{r,s}$ is a banded matrix with bandwidth no larger than $2(rs+r+s)(k_0-1)+2l_0+1$. Moreover,

$$\left\|\boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0\right\|_{\vdash} \leq \frac{8C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)},$$

whenever s is large enough such that $\delta_C(1+\delta_A^s) < 1$.

(c) Define $\boldsymbol{\Sigma}_1^{r,s} = \widetilde{\boldsymbol{C}}_s\boldsymbol{\Sigma}_0^{r,s}$ with $\widetilde{\boldsymbol{C}}_s$ and $\boldsymbol{\Sigma}_0^{r,s}$ as in Theorems 1(a) and 1(b), respectively. The matrix $\boldsymbol{\Sigma}_1^{r,s}$ is a banded matrix with bandwidth no larger than $(2rs+2r+3s+1)(k_0-1)+2l_0+1$. Moreover,

$$\left\|\boldsymbol{\Sigma}_1^{r,s} - \boldsymbol{\Sigma}_1\right\|_{\vdash} \leq \frac{9C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)},$$

whenever s is large enough such that $\delta_C(1+\delta_A^s) < 1$.

(d) Take any $h_1 \geq 2(rs + r + s)(k_0 - 1) + 2l_0 + 1$, then

$$\left\| \mathscr{B}_{h_1}(\widehat{\boldsymbol{\Sigma}}_0) - \boldsymbol{\Sigma}_0 \right\|_2 \leq \epsilon + \frac{16C_\epsilon \delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{2C_\epsilon \delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)},$$

with a probability of at least $1 - \mathscr{P}_1(\epsilon, N, T)$ (for polynomial tail decay) or $1 - \mathscr{P}_2(\epsilon, N, T)$ (for exponential tail decay).

(e) Take any $h_2 \geq (2rs + 2r + 3s + 1)(k_0 - 1) + 2l_0 + 1$, then

$$\left\| \mathscr{B}_{h_2}(\widehat{\boldsymbol{\Sigma}}_1) - \boldsymbol{\Sigma}_1 \right\|_2 \leq \epsilon + \frac{18C_\epsilon \delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{2C_\epsilon \delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)},$$

with a probability of at least $1 - \mathscr{P}_1(\epsilon, N, T)$ (for polynomial tail decay) or $1 - \mathscr{P}_2(\epsilon, N, T)$ (for exponential tail decay).

Explicit expressions for C_1, C_2, C_3, and $0 \leq \delta_C < 1$ are provided in the proofs below.

(a) The proof builds upon results from Guo et al. (2016) on banded vector autoregressions. Recall that $\boldsymbol{C} = \boldsymbol{D}\boldsymbol{B}$ with $\boldsymbol{D} = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}$. $\widetilde{\boldsymbol{D}}_s = \sum_{j=0}^{s-1} \boldsymbol{A}^j$ has a bandwidth of at most $s(k_0 - 1) + 1$ and satisfies[7]

$$\left\| \widetilde{\boldsymbol{D}}_s - \boldsymbol{D} \right\|_\vdash = \left\| \sum_{j=0}^{s-1} \boldsymbol{A}^j - (\boldsymbol{I}_N - \boldsymbol{A})^{-1} \right\|_\vdash = \left\| -\sum_{j=s}^{\infty} \boldsymbol{A}^j \right\|_\vdash \leq \sum_{j=s}^{\infty} \|\boldsymbol{A}\|_\vdash^j \leq \frac{\delta_A^s}{1-\delta_A}.$$

The product $\widetilde{\boldsymbol{C}}_s = \widetilde{\boldsymbol{D}}_s \boldsymbol{B}$ has a maximal bandwidth of $(s+1)(k_0 - 1) + 1$. Since $\|\boldsymbol{B}\|_\vdash \leq \delta_B$ and $\frac{\delta_B}{1-\delta_A} =: \delta_C < 1$ (Assumption 3(b)), we also have

$$\left\| \widetilde{\boldsymbol{C}}_s - \boldsymbol{C} \right\|_\vdash = \left\| \left(\widetilde{\boldsymbol{D}}_s - \boldsymbol{D} \right) \boldsymbol{B} \right\|_\vdash \leq \left\| \widetilde{\boldsymbol{D}}_s - \boldsymbol{D} \right\|_\vdash \|\boldsymbol{B}\|_\vdash \leq \delta_C \delta_A^s.$$

(b) Iterating on the observation in footnote 7, we conclude that the bandwidth of $\widetilde{\boldsymbol{C}}_s^r$ is at most $r\left[(s+1)(k_0-1)+1\right] - (r-1) = r(s+1)(k_0-1) + 1$. The bandwidth of $\boldsymbol{\Sigma}_0^{r,s}$ therefore does not exceed

$$2\left[r(s+1)(k_0-1)+1\right] + 2(s(k_0-1)+1) + (2l_0+1) - 4 = 2(rs + r + s)(k_0-1) + 2l_0 + 1.$$

We now bound $\left\| \boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0 \right\|_1$. Because $\boldsymbol{\Sigma}_0 = \sum_{j=0}^{\infty} \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \boldsymbol{D}' (\boldsymbol{C}')^j$, it holds that

$$\left\| \boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0 \right\|_\vdash \leq \left\| \sum_{j=0}^{r} \left[\widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s \boldsymbol{\Sigma}_\epsilon \widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j - \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \boldsymbol{D}' (\boldsymbol{C}')^j \right] \right\|_\vdash + \left\| \sum_{j=r+1}^{\infty} \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \boldsymbol{D}' (\boldsymbol{C}')^j \right\|_\vdash \tag{B.1}$$

Decomposing the first RHS term in (B.1) as

$$\begin{aligned}
&\sum_{j=0}^{r} \left[\widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s \boldsymbol{\Sigma}_\epsilon \widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j - \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \boldsymbol{D}' (\boldsymbol{C}')^j \right] \\
&= \sum_{j=0}^{r} \left[\left(\widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j \boldsymbol{D} \right) \boldsymbol{\Sigma}_\epsilon \widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j - \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \left(\boldsymbol{D}' (\boldsymbol{C}')^j - \widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j \right) \right] \\
&= \sum_{j=0}^{r} \left(\widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j \boldsymbol{D} \right) \boldsymbol{\Sigma}_\epsilon \left(\widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j - \boldsymbol{D}' (\boldsymbol{C}')^j \right) \\
&\quad + \sum_{j=0}^{r} \left[\left(\widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j \boldsymbol{D} \right) \boldsymbol{\Sigma}_\epsilon \boldsymbol{D}' (\boldsymbol{C}')^j - \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \left(\boldsymbol{D}' (\boldsymbol{C}')^j - \widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j \right) \right],
\end{aligned}$$

it follows that

$$\begin{aligned}
&\left\| \sum_{j=0}^{r} \left[\widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s \boldsymbol{\Sigma}_\epsilon \widetilde{\boldsymbol{D}}_s' (\widetilde{\boldsymbol{C}}_s')^j - \boldsymbol{C}^j \boldsymbol{D} \boldsymbol{\Sigma}_\epsilon \boldsymbol{D}' (\boldsymbol{C}')^j \right] \right\|_\vdash \\
&\leq C_\epsilon \sum_{j=0}^{r} \left\| \widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j \boldsymbol{D} \right\|_\vdash^2 + 2C_\epsilon \sum_{j=0}^{r} \left\| \widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j \boldsymbol{D} \right\|_\vdash \left\| \boldsymbol{C}^j \boldsymbol{D} \right\|_\vdash.
\end{aligned} \tag{B.2}$$

Furthermore, noting that

$$\left\| \widetilde{\boldsymbol{C}}_s^j \widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j \boldsymbol{D} \right\|_\vdash \leq \left\| \widetilde{\boldsymbol{C}}_s^j - \boldsymbol{C}^j \right\|_\vdash \left\| \widetilde{\boldsymbol{D}}_s \right\|_\vdash + \left\| \boldsymbol{C}^j \right\|_\vdash \left\| \boldsymbol{D} - \widetilde{\boldsymbol{D}}_s \right\|_\vdash, \tag{B.3}$$

[7] If matrices $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are banded matrices with bandwidths k_1 and k_2, respectively, then the product $\boldsymbol{F}_1 \boldsymbol{F}_2$ is again a banded matrix with a bandwidth of at most $k_1 + k_2 - 1$.

we proceed by bounding each norm in (B.3). First, expanding the matrix powers provides

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{C}}_s^j - \boldsymbol{C}^j\right\|_{\vdash} = \left\|\left(\widetilde{\boldsymbol{C}}_s - \boldsymbol{C} + \boldsymbol{C}\right)^j - \boldsymbol{C}^j\right\|_{\vdash} \\
&= \left\|\left((\widetilde{\boldsymbol{C}}_s - \boldsymbol{C})^j + (\widetilde{\boldsymbol{C}}_s - \boldsymbol{C})^{j-1}\boldsymbol{C} + (\widetilde{\boldsymbol{C}}_s - \boldsymbol{C})^{j-2}\boldsymbol{C}(\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}) + \cdots + \boldsymbol{C}(\widetilde{\boldsymbol{C}}_s - \boldsymbol{C})^{j-1} \ldots + \boldsymbol{C}^j\right) - \boldsymbol{C}^j\right\|_{\vdash} \\
&\leq \sum_{k=1}^{j}\binom{j}{k}\|\boldsymbol{C}\|_{\vdash}^{j-k}\left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash}^{k} = \sum_{k=0}^{j-1}\binom{j}{k+1}\|\boldsymbol{C}\|_{\vdash}^{(j-1)-k}\left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash}^{k+1} \\
&= \left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash}\sum_{k=0}^{j-1}\frac{j}{k+1}\binom{j-1}{k}\|\boldsymbol{C}\|_{\vdash}^{(j-1)-k}\left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash}^{k} \leq \delta_C\delta_A^s j\left[\left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash} + \|\boldsymbol{C}\|_{\vdash}\right]^{j-1} \\
&\leq \delta_C\delta_A^s j\left[\delta_C(1+\delta_A^s)\right]^{j-1} = \delta_C^j j\delta_A^s(1+\delta_A^s)^{j-1}.
\end{aligned}$$

Next, it holds that $\left\|\boldsymbol{D} - \widetilde{\boldsymbol{D}}_s\right\|_{\vdash} = \left\|\sum_{j=s}^{\infty}\boldsymbol{A}^j\right\|_{\vdash} \leq \frac{\delta_A^s}{1-\delta_A}$, $\left\|\widetilde{\boldsymbol{D}}_s\right\|_{\vdash} \leq \frac{1}{1-\delta_A}$ and $\|\boldsymbol{C}^j\|_{\vdash} \leq \delta_C^j$, such that we may bound (B.3) as

$$\left\|\widetilde{\boldsymbol{C}}_s^j\widetilde{\boldsymbol{D}}_s - \boldsymbol{C}^j\boldsymbol{D}\right\|_{\vdash} \leq \frac{\delta_A^s j\delta_C^j(1+\delta_A^s)^j}{(1+\delta_A^s)(1-\delta_A)} + \frac{\delta_A^s\delta_C^j}{1-\delta_A} \leq \frac{2\delta_A^s j\delta_C^j(1+\delta_A^s)^j}{(1-\delta_A)}. \tag{B.4}$$

Plugging (B.4) into (B.2), we obtain

$$\begin{aligned}
&\left\|\sum_{j=0}^{r}\left[\widetilde{\boldsymbol{C}}_s^j\widetilde{\boldsymbol{D}}_s\boldsymbol{\Sigma}_\epsilon\widetilde{\boldsymbol{D}}_s'(\widetilde{\boldsymbol{C}}_s')^j - \boldsymbol{C}^j\boldsymbol{D}\boldsymbol{\Sigma}_\epsilon\boldsymbol{D}'(\boldsymbol{C}')^j\right]\right\|_{\vdash} \\
&\leq \frac{4C_\epsilon\delta_A^{2s}}{(1-\delta_A)^2}\sum_{j=0}^{r}j^2\delta_C^{2j}(1+\delta_A^s)^{2j} + \frac{4C_\epsilon\delta_A^s}{(1-\delta_A)^2}\sum_{j=0}^{r}j\delta_C^{2j}(1+\delta_A^s)^j \\
&\leq \frac{4C_\epsilon\delta_A^{2s}\delta_C^2(1+\delta_A^s)^2}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{4C_\epsilon\delta_A^s\delta_C^2(1+\delta_A^s)}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} \leq \frac{8C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2},
\end{aligned} \tag{B.5}$$

where we have assumed that s is sufficiently large, such that $\delta_C(1+\delta_A^s) < 1$, and applied standard results for geometric series, i.e. $\sum_{j=1}^{\infty}jz^j = \frac{z}{(1-z)^2}$ and $\sum_{j=1}^{\infty}j^2z^j = \frac{z}{(1-z)^3}$ for $|z| < 1$. Next, we move to the second RHS term of (B.1), which is bounded by

$$\left\|\sum_{j=r+1}^{\infty}\boldsymbol{C}^j\boldsymbol{D}\boldsymbol{\Sigma}_\epsilon\boldsymbol{D}'(\boldsymbol{C}')^j\right\|_{\vdash} \leq \|\boldsymbol{\Sigma}_\epsilon\|_{\vdash}\|\boldsymbol{D}\|_{\vdash}^2\sum_{j=r+1}^{\infty}\|\boldsymbol{C}\|_{\vdash}^{2j} \leq \frac{C_\epsilon}{(1-\delta_A)^2}\sum_{j=r+1}^{\infty}\delta_C^{2j} = \frac{C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}. \tag{B.6}$$

Plugging both (B.5) and (B.6) into (B.1), results in the final bound

$$\left\|\boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0\right\|_{\vdash} \leq \frac{8C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}$$

thereby proving the intermediate result in part (b).

(c) We have $\boldsymbol{\Sigma}_1 = (\boldsymbol{I}_N - \boldsymbol{A})^{-1}\boldsymbol{B}\boldsymbol{\Sigma}_0 = \boldsymbol{C}\boldsymbol{\Sigma}_0$, and hence

$$\begin{aligned}
\left\|\boldsymbol{\Sigma}_1^{r,s} - \boldsymbol{\Sigma}_1\right\|_{\vdash} &= \left\|\widetilde{\boldsymbol{C}}_s(\boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0) + (\widetilde{\boldsymbol{C}}_s - \boldsymbol{C})\boldsymbol{\Sigma}_0\right\|_{\vdash} \leq \left\|\widetilde{\boldsymbol{C}}_s\right\|_{\vdash}\left\|\boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0\right\|_{\vdash} + \left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash}\|\boldsymbol{\Sigma}_0\|_{\vdash} \\
&\leq \frac{8C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)} + \frac{C_\epsilon\delta_C\delta_A^s}{(1-\delta_A^2)(1-\delta_C^2)} \\
&\leq \frac{9C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2} + \frac{C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}
\end{aligned} \tag{B.7}$$

where we have used that $\left\|\widetilde{\boldsymbol{C}}_s - \boldsymbol{C}\right\|_{\vdash} \leq \delta_C\delta_A^s$ for the intermediate result (a), $\left\|\widetilde{\boldsymbol{C}}_s\right\|_{\vdash} \leq \delta_C(1+\delta_A^s) < 1$ since s is taken sufficiently large, and $\|\boldsymbol{\Sigma}_0\|_{\vdash} \leq \sum_{j=0}^{\infty}\|\boldsymbol{C}^j\boldsymbol{D}\boldsymbol{\Sigma}_\epsilon\boldsymbol{D}'(\boldsymbol{C}')^j\|_{\vdash} \leq \frac{C_\epsilon}{(1-\delta_A^2)(1-\delta_C^2)}$.

(d) We have

$$\begin{aligned}
\left\|\mathscr{B}_{h_1}(\widehat{\boldsymbol{\Sigma}}_0) - \boldsymbol{\Sigma}_0\right\|_{\vdash} &\leq \left\|\mathscr{B}_{h_1}(\widehat{\boldsymbol{\Sigma}}_0) - \mathscr{B}_{h_1}(\boldsymbol{\Sigma}_0)\right\|_{\vdash} + \left\|\mathscr{B}_{h_1}(\boldsymbol{\Sigma}_0) - \mathscr{B}_{h_1}(\boldsymbol{\Sigma}_0^{r,s})\right\|_{\vdash} + \left\|\mathscr{B}_{h_1}(\boldsymbol{\Sigma}_0^{r,s}) - \boldsymbol{\Sigma}_0\right\|_{\vdash} \\
&= \left\|\mathscr{B}_{h_1}(\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\right\|_{\vdash} + \left\|\mathscr{B}_{h_1}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0^{r,s})\right\|_{\vdash} + \left\|\boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0\right\|_{\vdash} \\
&\leq \left\|\mathscr{B}_{h_1}(\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\right\|_{\vdash} + 2\left\|\boldsymbol{\Sigma}_0^{r,s} - \boldsymbol{\Sigma}_0\right\|_{\vdash}
\end{aligned} \tag{B.8}$$

because $\boldsymbol{\Sigma}_0^{r,s}$ is a banded matrix already and banding can only decrease the norm difference between $\boldsymbol{\Sigma}_0^{r,s}$ and $\boldsymbol{\Sigma}_0$. We consider the two terms in the RHS of (B.8) separately. Note that $\left\|\hat{\Sigma} - \sigma_0\right\|_{\vdash} = \left\|\hat{\Sigma} - \sigma_0\right\|_1$. by symmetry. Then, letting ξ_t be as defined in Lemma 3,

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\mathscr{B}_{h_1}\left(\widehat{\boldsymbol{\Sigma}}_0-\boldsymbol{\Sigma}_0\right)\right\|_{\vdash}\leq x\right)=\mathbb{P}\left(\max_{i\leq j\leq N}\sum_{i=1\wedge(j-h_1)}^{N\vee(j+h_1)}\left|\sum_{t=2}^{T}\xi_{it}\xi_{jt}-\mathbb{E}(\xi_{it}\xi_{jt})\right|\leq Tx\right)\\
&\geq 1-\sum_{j=1}^{N}\mathbb{P}\left(\sum_{i=1\wedge(j-h_1)}^{N\vee(j+h_1)}\left|\sum_{t=2}^{T}\xi_{it}\xi_{jt}-\mathbb{E}(\xi_{it}\xi_{jt})\right|>Tx\right)\\
&\geq 1-\sum_{j=1}^{N}\sum_{i=1\wedge(j-h_1)}^{N\vee(j+h_1)}\mathbb{P}\left(\left|\sum_{t=2}^{T}\xi_{it}\xi_{jt}-\mathbb{E}(\xi_{it}\xi_{jt})\right|>\frac{Tx}{2h_1+1}\right)\\
&\geq\begin{cases}1-(2h_1+1)N\left[\left(b_1T^{(1-\delta)/3}+\frac{(2h_1+1)b_3}{x}\right)\exp\left(-\frac{T^{(1-\delta)/3}}{2b_1^2}\right)+\frac{b_2(2h_1+1)^d}{x^dT^{\frac{\delta}{2}(d-1)}}\right] & \text{(polynomial tails)},\\ 1-(2h_1+1)N\left[\frac{\kappa_1(2h_1+1)}{x}+\frac{2}{\kappa_2}\left(\frac{Tx^2}{(2h_1+1)^2}\right)^{1/7}\right]\exp\left(-\frac{1}{\kappa_3}\left(\frac{Tx^2}{(2h_1+1)^2}\right)^{1/7}\right) & \text{(exponential tails)},\end{cases}
\end{aligned}
\tag{B.9}
$$

where the last inequality exploits Lemma 3. Note that the probabilities in (B.9) coincide with the probabilities defined as $1-\mathscr{P}_1(x,N,T)$ and $1-\mathscr{P}_2(x,N,T)$ in Theorem 1. Overall, if $\left\|\mathscr{B}_{h_1}\left(\widehat{\boldsymbol{\Sigma}}_0-\boldsymbol{\Sigma}_0\right)\right\|_{\vdash}\leq\epsilon$ holds, then, applying intermediate result (b) to (B.8), we obtain the bound

$$
\left\|\mathscr{B}_{h_1}\left(\widehat{\boldsymbol{\Sigma}}_0\right)-\boldsymbol{\Sigma}_0\right\|_{\vdash}\leq\epsilon+2\left\|\boldsymbol{\Sigma}_0^{r,s}-\boldsymbol{\Sigma}_0\right\|_{\vdash}\leq\epsilon+\frac{16C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2}+\frac{2C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}.
$$

(e) Mimicking the steps from (d), we find

$$
\left\|\mathscr{B}_{h_2}\left(\widehat{\boldsymbol{\Sigma}}_1\right)-\boldsymbol{\Sigma}_1\right\|_{\vdash}\leq\epsilon+2\left\|\boldsymbol{\Sigma}_1^{r,s}-\boldsymbol{\Sigma}_1\right\|_{\vdash}\leq\epsilon+\frac{18C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2}+\frac{2C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}
$$

if we use part (c) and if $\left\{\left\|\mathscr{B}_{h_2}\left(\widehat{\boldsymbol{\Sigma}}_1-\boldsymbol{\Sigma}_1\right)\right\|_{\vdash}\leq\epsilon\right\}$ holds. The probability of the latter event is $1-\mathscr{P}_1(\epsilon,N,T)$ (polynomial tails) or $1-\mathscr{P}_2(\epsilon,N,T)$ (exponential tails).

We now combine all these intermediate results to recover the result from the theorem. Parts (d) and (e) are both applicable since

$$
\begin{aligned}
&h=\max\{h_1,h_2\}=h_2=(2rs+2r+3s+1)(k_0-1)+2l_0+1\\
&\leq\Big(2r(s+1)+3(s+1)\Big)(k_0-1)+2l_0+1\leq(s+1)(2r+3)(k_0-1)+2l_0+1.
\end{aligned}
\tag{B.10}
$$

Then, by intermediate results (d) and (e), it holds that

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{V}}_h-\boldsymbol{V}\right\|_{\vdash}&\leq\left\|\mathscr{B}_{h_2}\left(\widehat{\boldsymbol{\Sigma}}_1\right)-\boldsymbol{\Sigma}_1\right\|_{\vdash}+\left\|\mathscr{B}_{h_1}\left(\widehat{\boldsymbol{\Sigma}}_0\right)-\boldsymbol{\Sigma}_0\right\|_{\vdash}\\
&\leq 2\epsilon+\frac{34C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2}+\frac{4C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}
\end{aligned}
$$

with probabilities $1-2\mathscr{P}_1(\epsilon,N,T)$ and $1-2\mathscr{P}_2(\epsilon,N,T)$ in the cases of polynomial and exponential tail decay, respectively.

It remains to determine s and r such that $\delta_C(1+\delta_A^s)<1$,

$$
\frac{34C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2}=\epsilon,\quad\text{and}\quad\frac{4C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}=\epsilon.
$$

First, note that we can ensure $\delta_C(1+\delta_A^s)=K<1$, by choosing $s>\log\left(\frac{\delta_C}{K-\delta_C}\right)/\left|\log(\delta_A)\right|$. Then,

$$
\frac{34C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-\delta_C^2(1+\delta_A^s)^2\right)^2}=\frac{34C_\epsilon\delta_A^s}{(1-\delta_A)^2\left(1-K^2\right)^2}=\epsilon\Leftarrow s=\frac{\log\left(\frac{34C_\epsilon}{(1-\delta_A)^2(1-K^2)^2\epsilon}\right)}{\left|\log(\delta_A)\right|}.
\tag{B.11}
$$

Furthermore,

$$
\frac{4C_\epsilon\delta_C^{2(r+1)}}{(1-\delta_A)^2(1-\delta_C^2)}=\epsilon\Leftarrow r=\frac{\log\left(\frac{4C_\epsilon}{(1-\delta_A)^2(1-\delta_C^2)\epsilon}\right)}{2\left|\log(\delta_C)\right|}-\frac{1}{2}.
\tag{B.12}
$$

Define $C_1=\frac{34C_\epsilon}{(1-\delta_A)^2(1-K^2)^2}$ and $C_2=\frac{4C_\epsilon}{(1-\delta_A)^2(1-\delta_C^2)}$. Plugging (B.11)–(B.12) into (B.10), we obtain the result that when

$$
h=\left(\frac{\max\left\{\log\left(\frac{K-\delta_c}{\delta_c}\right),\log\left(\frac{C_1}{\epsilon}\right)\right\}}{\left|\log(\delta_A)\right|}+1\right)\left(\frac{\log\left(\frac{C_2}{\epsilon}\right)}{\left|\log(\delta_C)\right|}+2\right)(k_0-1)+2l_0+1,
$$

for some $K \in (\delta_c, 1)$, it holds that $\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_{\vdash} \leq 4\epsilon$ with probability at least $1 - 2\mathscr{P}_1(\epsilon, N, T)$ and $1 - 2\mathscr{P}_2(\epsilon, N, T)$ in the cases of polynomial and exponential tail decay, respectively. □

Proof of Theorem 2. The proof of the theorem relies on the properties of dual norms. Recall $P_\alpha(\boldsymbol{c}) = \alpha \sum_{g\in\mathscr{G}} \sqrt{|g|} \left\|\boldsymbol{c}_g\right\|_2 + (1-\alpha)\|\boldsymbol{c}\|_1$. Exploiting the properties of $\|\cdot\|_1$ and $\|\cdot\|_2$, it is straightforward to verify that $P_\alpha(\cdot)$ is a norm for any $0 \leq \alpha \leq 1$. For any norm $\|\cdot\|$, we define its dual norm $\|\cdot\|^*$ through $\|\boldsymbol{c}\|^* = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{|\boldsymbol{c}'\boldsymbol{x}|}{\|\boldsymbol{x}\|}$. The dual-norm inequality states that

$$\boldsymbol{c}'\boldsymbol{x} \leq \|\boldsymbol{c}\|^* \, \|\boldsymbol{x}\| \qquad \text{for all conformable vectors } \boldsymbol{c} \text{ and } \boldsymbol{x}. \tag{B.13}$$

For the norm $P_\alpha(\boldsymbol{c})$, its dual norm $P_\alpha^*(\boldsymbol{c})$ is bounded by

$$\begin{aligned} P_\alpha^*(\boldsymbol{c}) &= \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{|\boldsymbol{c}'\boldsymbol{x}|}{P_\alpha(\boldsymbol{x})} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{|\boldsymbol{c}'\boldsymbol{x}|}{\alpha \sum_{g\in\mathscr{G}} \sqrt{|g|} \left\|\boldsymbol{x}_g\right\|_2 + (1-\alpha)\|\boldsymbol{x}\|_1} \\ &\overset{(i)}{\leq} \alpha \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{|\boldsymbol{c}'\boldsymbol{x}|}{\sum_{g\in\mathscr{G}} \sqrt{|g|} \left\|\boldsymbol{x}_g\right\|_2} + (1-\alpha) \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{|\boldsymbol{c}'\boldsymbol{x}|}{\|\boldsymbol{x}\|_1} \overset{(ii)}{\leq} \alpha \max_{g\in\mathscr{G}} \frac{\left\|\boldsymbol{c}_g\right\|_2}{\sqrt{|g|}} + (1-\alpha)\|\boldsymbol{c}\|_\infty, \end{aligned} \tag{B.14}$$

by convexity of the function $f(x) = x^{-1}$ in step (i), and using for step (ii) both $\|\boldsymbol{c}\|_1^* = \|\boldsymbol{c}\|_\infty$ and

$$\begin{aligned} \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{|\boldsymbol{c}'\boldsymbol{x}|}{\sum_{g\in\mathscr{G}} \sqrt{|g|} \left\|\boldsymbol{x}_g\right\|_2} &= \sup_{\boldsymbol{x}\neq\boldsymbol{0}, \sum_{g\in\mathscr{G}} \sqrt{|g|}\left\|\boldsymbol{x}_g\right\|_2 = 1} \left| \sum_{g\in\mathscr{G}} \boldsymbol{c}_g'\boldsymbol{x}_g \right| \\ &\leq \sup_{\boldsymbol{x}\neq\boldsymbol{0}, \sum_{g\in\mathscr{G}} \sqrt{|g|}\left\|\boldsymbol{x}_g\right\|_2 = 1} \sum_{g\in\mathscr{G}} \left\| \frac{\boldsymbol{c}_g}{\sqrt{|g|}} \right\|_2 \left\|\sqrt{|g|}\boldsymbol{x}_g\right\|_2 \leq \max_{g\in\mathscr{G}} \frac{\left\|\boldsymbol{c}_g\right\|_2}{\sqrt{|g|}}. \end{aligned}$$

We now start the actual proof. Recall $\hat{\boldsymbol{\sigma}}_h = \operatorname{vec}\left(\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)'\right)$, $\hat{\boldsymbol{V}}_h = \left[\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)' \quad \mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_0)\right]$, $\boldsymbol{V} = \left[\boldsymbol{\Sigma}_1' \quad \boldsymbol{\Sigma}_0\right]$, and $\boldsymbol{C} = \left[\boldsymbol{A} \quad \boldsymbol{B}\right]$. Exploiting standard properties of $\operatorname{vec}(\cdot)$, we find

$$\begin{aligned} \hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c} &= \operatorname{vec}\left(\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)' - \hat{\boldsymbol{V}}_h\boldsymbol{C}'\right) = \operatorname{vec}\left(\left[\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1) - \boldsymbol{\Sigma}_1\right]' - \left[\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right]\boldsymbol{C}' + \underbrace{\left[\boldsymbol{\Sigma}_1' - \boldsymbol{V}\boldsymbol{C}'\right]}_{=\boldsymbol{O}, \text{ see (5)}} \right) \\ &= \underbrace{\operatorname{vec}\left(\mathscr{B}_h(\hat{\boldsymbol{\Sigma}}_1)' - \boldsymbol{\Sigma}_1'\right)}_{\hat{\boldsymbol{\Delta}}_\Sigma} - \underbrace{\left[\hat{\boldsymbol{V}}_h^{(d)} - \boldsymbol{V}^{(d)}\right]}_{\hat{\boldsymbol{\Delta}}_V}\boldsymbol{c} = \hat{\boldsymbol{\Delta}}_\Sigma - \hat{\boldsymbol{\Delta}}_V\boldsymbol{c}. \end{aligned} \tag{B.15}$$

Using (B.15), we rewrite

$$\begin{aligned} \left\|\hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\hat{\boldsymbol{c}}\right\|_2^2 &= \left\|\left[\hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c}\right] - \left[\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}} - \boldsymbol{c})\right]\right\|_2^2 \\ &= \left\|\hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c}\right\|_2^2 + \left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}} - \boldsymbol{c})\right\|_2^2 - 2(\hat{\boldsymbol{c}} - \boldsymbol{c})'\hat{\boldsymbol{V}}_h^{(d)\prime}\left(\hat{\boldsymbol{\Delta}}_\Sigma - \hat{\boldsymbol{\Delta}}_V\boldsymbol{c}\right). \end{aligned}$$

Recalling the objective function $\mathscr{L}_\alpha(\boldsymbol{c};\lambda) = \left\|\hat{\boldsymbol{\sigma}}_h - \hat{\boldsymbol{V}}_h^{(d)}\boldsymbol{c}\right\|_2^2 + \lambda P_\alpha(\boldsymbol{c})$ and noting that $\mathscr{L}_\alpha(\hat{\boldsymbol{c}};\lambda) \leq \mathscr{L}_\alpha(\boldsymbol{c};\lambda)$ by construction, it follows that

$$\begin{aligned} \left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}} - \boldsymbol{c})\right\|_2^2 + \lambda P_\alpha(\hat{\boldsymbol{c}}) &\leq 2(\hat{\boldsymbol{c}} - \boldsymbol{c})'\hat{\boldsymbol{V}}_h^{(d)\prime}\left(\hat{\boldsymbol{\Delta}}_\Sigma - \hat{\boldsymbol{\Delta}}_V\boldsymbol{c}\right) + \lambda P_\alpha(\boldsymbol{c}) \\ &\leq 2P_\alpha(\hat{\boldsymbol{c}} - \boldsymbol{c})\, P_\alpha^*\left(\hat{\boldsymbol{V}}_h^{(d)\prime}\left(\hat{\boldsymbol{\Delta}}_\Sigma - \hat{\boldsymbol{\Delta}}_V\boldsymbol{c}\right)\right) + \lambda P_\alpha(\boldsymbol{c}) \\ &\leq P_\alpha(\hat{\boldsymbol{c}} - \boldsymbol{c})\left[2P_\alpha^*\left(\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_\Sigma\right) + 2P_\alpha^*\left(\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_V\boldsymbol{c}\right)\right] + \lambda P_\alpha(\boldsymbol{c}), \end{aligned} \tag{B.16}$$

where we used the dual-norm inequality (see (B.13)) and the triangle property of (dual) norms in the second and third inequality, respectively. Define the sets

$$\mathscr{H}_1(x) = \left\{2P_\alpha^*\left(\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_\Sigma\right) \leq x\right\} \qquad \text{and} \qquad \mathscr{H}_2(x) = \left\{2P_\alpha^*\left(\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_V\boldsymbol{c}\right) \leq x\right\}. \tag{B.17}$$

On the set $\mathscr{H}_1(\frac{\lambda}{4}) \cap \mathscr{H}_2(\frac{\lambda}{4})$, we can scale (B.16) by a factor 2 to obtain

$$2\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}} - \boldsymbol{c})\right\|_2^2 + 2\lambda P_\alpha(\hat{\boldsymbol{c}}) \leq \lambda P_\alpha(\hat{\boldsymbol{c}} - \boldsymbol{c}) + 2\lambda P_\alpha(\boldsymbol{c}). \tag{B.18}$$

We subsequently manipulate $P_\alpha(\hat{\boldsymbol{c}})$ and $P_\alpha(\hat{\boldsymbol{c}}-\boldsymbol{c})$. Using the reverse triangle inequality, we have

$$\begin{aligned} P_\alpha(\hat{\boldsymbol{c}}) &= \alpha \sum_{g\in\mathscr{G}} \sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g\right\|_2 + (1-\alpha)\left\|\hat{\boldsymbol{c}}\right\|_1 \\ &\geq \alpha \sum_{g\in\mathscr{G}_S} \sqrt{|g|}\left[\left\|\boldsymbol{c}_g\right\|_2 - \left\|\hat{\boldsymbol{c}}_g - \boldsymbol{c}_g\right\|_2\right] + \alpha \sum_{g\in\mathscr{G}_{S^c}} \sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g\right\|_2 + (1-\alpha)\left[\left\|\boldsymbol{c}_S\right\|_1 - \left\|\hat{\boldsymbol{c}}_S - \boldsymbol{c}_S\right\|_1 + \left\|\hat{\boldsymbol{c}}_{S^c}\right\|_1\right] \\ &= P_{\alpha,S}(\boldsymbol{c}) + P_{\alpha,S^c}(\hat{\boldsymbol{c}}-\boldsymbol{c}) - P_{\alpha,S}(\hat{\boldsymbol{c}}-\boldsymbol{c}), \end{aligned} \tag{B.19}$$

where $\mathscr{G}_S$ and $\mathscr{G}_{S^c}$ are defined in Lemma 1. Simple rewriting provides

$$\begin{aligned} P_\alpha(\hat{\boldsymbol{c}}-\boldsymbol{c}) &= \alpha \sum_{g\in\mathscr{G}} \sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g - \boldsymbol{c}_g\right\|_2 + (1-\alpha)\left\|\hat{\boldsymbol{c}}_S - \boldsymbol{c}_S\right\|_1 + \alpha \sum_{g\in\mathscr{G}_{S^c}} \sqrt{|g|}\left\|\hat{\boldsymbol{c}}_g\right\|_2 + (1-\alpha)\left\|\hat{\boldsymbol{c}}_{S^c}\right\|_1 \\ &= P_{\alpha,S}(\hat{\boldsymbol{c}}-\boldsymbol{c}) + P_{\alpha,S^c}(\hat{\boldsymbol{c}}-\boldsymbol{c}). \end{aligned} \tag{B.20}$$

Combining results (B.18)–(B.20) yields

$$2\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2^2 + \lambda P_{\alpha,S^c}(\hat{\boldsymbol{c}}-\boldsymbol{c}) \leq 3\lambda P_{\alpha,S}(\hat{\boldsymbol{c}}-\boldsymbol{c}), \tag{B.21}$$

and $\hat{\boldsymbol{c}}-\boldsymbol{c}$ is thus a member of the set $\mathscr{C}_{N_c}(\mathscr{G},S)$ as defined in Lemma 1.

In combination with Lemma 2, thus requiring $\mathscr{H}_1(\frac{\lambda}{4}) \cap \mathscr{H}_2(\frac{\lambda}{4}) \cap \mathscr{V}\left(\frac{\phi_0}{2}\right)$ to hold, we conclude

$$\begin{aligned} 2\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2^2 + \lambda P_\alpha(\hat{\boldsymbol{c}}-\boldsymbol{c}) &= 2\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2^2 + \lambda P_{\alpha,S}(\hat{\boldsymbol{c}}-\boldsymbol{c}) + \lambda P_{\alpha,S^c}(\hat{\boldsymbol{c}}-\boldsymbol{c}) \\ &\overset{(i)}{\leq} 4\lambda P_{\alpha,S}(\hat{\boldsymbol{c}}-\boldsymbol{c}) \overset{(ii)}{\leq} 16\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2 \left(\frac{\bar{\omega}_\alpha \lambda}{\phi_0}\right) \overset{(iii)}{\leq} \left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2^2 + \frac{64\bar{\omega}_\alpha^2\lambda^2}{\phi_0^2}, \end{aligned} \tag{B.22}$$

where step (i) follows from (B.21), step (ii) is implied by $P_{\alpha,S}(\hat{\boldsymbol{c}}-\boldsymbol{c}) \leq \frac{4\bar{\omega}_0\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}}-\boldsymbol{c})\right\|_2}{\phi_0}$ for $\hat{\boldsymbol{c}}-\boldsymbol{c} \in \mathscr{C}_{N_c}(\mathscr{G},S)$ (Lemma 2), and step (iii) uses the elementary inequality $16uv \leq u^2 + 64v^2$ (i.e. manipulating $(u-8v)^2 \geq 0$). A straightforward rearrangement of (B.22) provides the inequality of Theorem 2.

It remains to determine a lower bound on the probability of $\mathscr{H}_1(\frac{\lambda}{4}) \cap \mathscr{H}_2(\frac{\lambda}{4}) \cap \mathscr{V}\left(\frac{\phi_0}{2}\right)$. We rely on the elementary inequality $\mathbb{P}\left(\mathscr{H}_1(\frac{\lambda}{4}) \cap \mathscr{H}_2(\frac{\lambda}{4}) \cap \mathscr{V}\left(\frac{\phi_0}{2}\right)\right) \geq 1 - \mathbb{P}\left(\mathscr{H}_1(\frac{\lambda}{4})^c\right) - \mathbb{P}\left(\mathscr{H}_2(\frac{\lambda}{4})^c\right) - \mathbb{P}\left(\mathscr{V}\left(\frac{\phi_0}{2}\right)^c\right)$ to bound the individual probabilities.

We start with $\mathbb{P}\left(\mathscr{H}_1(\frac{\lambda}{4})^c\right) = \mathbb{P}\left(2P_\alpha^*\left(\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_\Sigma\right) > \frac{\lambda}{4}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_\Sigma\right\|_\infty > \frac{\lambda}{8}\right)$. The last inequality is true because continuing from (B.14), we have

$$P_\alpha^*(\boldsymbol{c}) \leq \alpha \max_{g\in\mathscr{G}} \frac{\left\|\boldsymbol{c}_g\right\|_2}{\sqrt{|g|}} + (1-\alpha)\left\|\boldsymbol{c}\right\|_\infty \leq \alpha \max_{g\in\mathscr{G}} \frac{\sqrt{|g|}\left\|\boldsymbol{c}_g\right\|_\infty}{\sqrt{|g|}} + (1-\alpha)\left\|\boldsymbol{c}\right\|_\infty = \left\|\boldsymbol{c}\right\|_\infty \tag{B.23}$$

for any vector $\boldsymbol{c}$. Subsequently, we have

$$\begin{aligned} \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_\Sigma\right\|_\infty > \frac{\lambda}{8}\right) &= \mathbb{P}\left(\left\|\left[(\hat{\boldsymbol{V}}_h^{(d)} - \boldsymbol{V}^{(d)}) + \boldsymbol{V}^{(d)}\right]'\hat{\boldsymbol{\Delta}}_\Sigma\right\|_\infty > \frac{\lambda}{8}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_1\left\|\hat{\boldsymbol{\Delta}}_\Sigma\right\|_\infty + C_V\left\|\hat{\boldsymbol{\Delta}}_\Sigma\right\|_\infty > \frac{\lambda}{8}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash^2 + C_V\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda}{8}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda^{1/2}}{4}\right) + \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda}{16C_V}\right), \end{aligned} \tag{B.24}$$

exploiting block-diagonality of $\hat{\boldsymbol{V}}_h^{(d)} - \boldsymbol{V}^{(d)}$ and $\boldsymbol{V}^{(d)}$ such that $\left\|\hat{\boldsymbol{V}}_h^{(d)} - \boldsymbol{V}^{(d)}\right\|_1 = \max_{1\leq i\leq N}\left\|\hat{\boldsymbol{V}}_{i,h} - \boldsymbol{V}_i\right\|_1 \leq \left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash$ and $\left\|\boldsymbol{V}^{(d)}\right\|_1 \leq \left\|\boldsymbol{V}\right\|_\vdash \leq C_V$ (explicitly assumed in Theorem 2). Bounds for the final RHS terms in (B.24) are available from Theorem 1.

Second, we have

$$\begin{aligned} \mathbb{P}\left(\mathscr{H}_2\left(\frac{\lambda}{4}\right)^c\right) &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h^{(d)\prime}\hat{\boldsymbol{\Delta}}_V\boldsymbol{c}\right\|_\infty > \frac{\lambda}{8}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h^{(d)\prime}\left[\hat{\boldsymbol{V}}_h^{(d)} - \boldsymbol{V}^{(d)}\right]\right\|_\infty > \frac{\lambda}{8\left\|\boldsymbol{c}\right\|_\infty}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_1\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\infty + C_V\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\infty > \frac{\lambda}{8\left\|\boldsymbol{c}\right\|_\infty}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda^{1/2}}{4}\right) + \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda}{16C_V}\right), \end{aligned} \tag{B.25}$$

where the last line relies on the union bound and the fact that $\left\|\boldsymbol{c}\right\|_\infty < 1$ (implied by Assumption 1). Hence, the sets $\mathscr{H}_1\left(\frac{\lambda}{4}\right)^c$ and $\mathscr{H}_2\left(\frac{\lambda}{4}\right)^c$ admit the same probability bound.

Finally, $\mathbb{P}\left(\mathscr{V}\left(\frac{\phi_0}{2}\right)^c\right) = \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_1 > \frac{\phi_0}{2}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\phi_0}{2}\right)$. Combining all previous results, we conclude

$$\begin{aligned}
&\mathbb{P}\left(\mathscr{H}_1\left(\frac{\lambda}{4}\right) \cap \mathscr{H}_2\left(\frac{\lambda}{4}\right) \cap \mathscr{V}\left(\frac{\phi_0}{2}\right)\right) \\
&\geq 1 - 2\mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda^{1/2}}{4}\right) - 2\mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\lambda}{16C_V}\right) - \mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > \frac{\phi_0}{2}\right) \\
&\geq 1 - 5\mathbb{P}\left(\left\|\hat{\boldsymbol{V}}_h - \boldsymbol{V}\right\|_\vdash > 6f(\lambda, \phi_0)\right),
\end{aligned} \tag{B.26}$$

where $f(\lambda, \phi_0) = \min\left(\frac{\lambda^{1/2}}{24}, \frac{\lambda}{96C_V}, \frac{\phi_0}{12}\right)$. The proof is completed by evaluating the final probability in (B.26) using Theorem 1. □

Proof of Corollary 1. First, we derive the conditions under which the set on which the performance bound in Theorem 2 holds occurs with probability converging to one. Under Assumption 2(b1), along with the remaining assumptions in Theorem 2, this probability is given by $1 - \mathscr{P}_1(f(\lambda, \phi_0), N, T)$, where we recall from Theorem 1 that

$$\begin{aligned}
\mathscr{P}_1(f(\lambda, \phi_0), N, T) &= (2h(f(\lambda, \phi_0)) + 1)N\left(b_1 T^{(1-\delta)/3} + \frac{[2h(f(\lambda, \phi_0)) + 1]b_3}{f(\lambda, \phi_0)}\right)\exp\left(-\frac{T^{(1-\delta)/3}}{2b_1^2}\right) \\
&\quad + (2h(f(\lambda, \phi_0)) + 1)N\frac{b_2[2h(f(\lambda, \phi_0)) + 1]^d}{f(\lambda, \phi_0)^d T^{\frac{\delta}{2}(d-1)}},
\end{aligned} \tag{B.27}$$

for some $0 < \delta < 1$ and $f(\lambda, \phi_0) = \min\left(\frac{\lambda^{1/2}}{24}, \frac{\lambda}{96C_v}, \frac{\phi_0}{12}\right)$. Given that $\lambda = O(T^{-q_\lambda})$ with $q_\lambda > 0$, it follows immediately that

$$f(\lambda, \phi_0) = \min\left(\frac{\lambda^{1/2}}{24}, \frac{\lambda}{96C_V}, \frac{\phi_0}{12}\right) = O\left(T^{-q_\lambda}\right) \tag{B.28}$$

and, recalling the definition of $h\left(f(\lambda, \phi_0)\right)$ in (4),

$$h\left(f(\lambda, \phi_0)\right) = O\left(\log(T)^2 T^{q_k}\right). \tag{B.29}$$

Since these are polynomial rates in T, it follows that the first RHS-term in (B.27) converges to zero exponentially in T for any $\delta < 1$. The second RHS-term, however, converges to zero at most at a polynomial rate, such that

$$\begin{aligned}
\mathscr{P}_1(f(\lambda, \phi_0), N, T) &= O\left(\frac{(2h(f(\lambda, \phi_0)) + 1)N[h\left(f(\lambda, \phi_0)\right) + 1]^d}{\lambda^d T^{\frac{\delta}{2}(d-1)}}\right) \\
&= O\left(\log(T)^{2(d+1)} T^{q_N + (d+1)q_k + dq_\lambda - \frac{\delta(d-1)}{2}}\right).
\end{aligned} \tag{B.30}$$

From (B.30), it follows that $\mathscr{P}_1(f(\lambda, \phi_0), N, T) \to 0$ as $T \to \infty$ if

$$q_N + (d+1)q_k + dq_\lambda - \frac{\delta(d-1)}{2} < 0 \Rightarrow q_\lambda < \frac{\delta(d-1)}{2d} - \frac{(d+1)q_k}{d} - \frac{q_N}{d}$$

In a similar fashion, we derive the conditions under which $\mathscr{P}_2(f(\lambda, \phi_0), N, T) \to 0$, by noting that

$$\begin{aligned}
\mathscr{P}_2(f(\lambda, \phi_0), N, T) &= (2h(f(\lambda, \phi_0)) + 1)N\left[\frac{\kappa_1[2h(f(\lambda, \phi_0)) + 1]}{f(\lambda, \phi_0)} + \frac{2}{\kappa_2}\left(\frac{T f(\lambda, \phi_0)^2}{[2h + 1]^2}\right)^{\frac{1}{7}}\right] \\
&\quad \times \exp\left(-\frac{1}{\kappa_3}\left(\frac{T f(\lambda, \phi_0)^2}{[2h(f(\lambda, \phi_0)) + 1]^2}\right)^{\frac{1}{7}}\right)
\end{aligned} \tag{B.31}$$

converges to zero exponentially fast in T if $\frac{T f(\lambda, \phi_0)^2}{[2h(f(\lambda, \phi_0)) + 1]^2}$ diverges at a polynomial rate in T. Making use of (B.28) and (B.29), it follows that

$$\frac{T f(\lambda, \phi_0)^2}{[2h(f(\lambda, \phi_0)) + 1]^2} = O\left(\log(T)^{-4} T^{1 - 2q_\lambda - 2q_k}\right),$$

which translates to the condition $1 - 2q_\lambda - 2q_k > 0$, or $q_\lambda < \frac{1}{2} - q_k$. This establishes conditions (i) and (ii) in Corollary 1.

We proceed by deriving the order of the performance bound in Theorem 2. Noting that

$$\sum_{g \in \mathscr{G}_S} \sqrt{|g|} \leq |\mathscr{G}_S| \max_{g \in \mathscr{G}_S} \sqrt{|g|} = O\left(T^{q_g + q_N/2}\right),$$

it follows that

$$\bar{\omega} = O\left((1 - \alpha)T^{q_g + q_N/2} + \alpha T^{q_S/2}\right).$$

Then, by Theorem 2,

$$\left\|\hat{\boldsymbol{V}}_h^{(d)}(\hat{\boldsymbol{c}} - \boldsymbol{c})\right\|_2^2 \leq \frac{4\bar{\omega}_\alpha^2 \lambda^2}{\phi_0^2} = O\left((1 - \alpha)T^{2q_g + q_N - 2q_\lambda} + \alpha T^{q_S - 2q_\lambda}\right)$$

and

$$(1-\alpha)\sum_{g\in\mathscr{G}}\sqrt{|g|}\left\|\hat{c}_g-c_g\right\|_2+\alpha\left\|\hat{c}-c\right\|_1\leq\frac{4\bar{\omega}_\alpha^2\lambda}{\phi_0^2}=O\left((1-\alpha)T^{2q_g+q_N-q_\lambda}+\alpha T^{q_S-q_\lambda}\right),$$

on a set with probability $1-\mathscr{P}_1(f(\lambda,\phi_0),N,T)$ or $1-\mathscr{P}_2(f(\lambda,\phi_0),N,T)$, depending on whether Assumption 2(b1) or 2(b2) applies, respectively. Since we have shown that both $\mathscr{P}_1(f(\lambda,\phi_0),N,T)\to 0$ and $\mathscr{P}_2(f(\lambda,\phi_0),N,T)\to 0$ under the conditions imposed in Corollary 1, the proof is complete. □

Proof of Theorem 3. Set $\boldsymbol{\beta}^*=(\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_K')'$, $\boldsymbol{U}^{(d)}=\begin{bmatrix}\boldsymbol{W}_1^{*(d)} & \cdots & \boldsymbol{W}_K^{*(d)}\end{bmatrix}$, and $\boldsymbol{Q}^{(d)}=\begin{bmatrix}\boldsymbol{V}^{*(d)} & \boldsymbol{U}^{(d)}\end{bmatrix}$ (and similar quantities with "hats"). The residuals entering the L_2 component of the SPLASHX objective function are

$$\begin{aligned}\hat{\boldsymbol{\sigma}}^*-\hat{\boldsymbol{V}}^{*(d)}\boldsymbol{c}-\sum_{k=1}^K\hat{\boldsymbol{W}}_k^{*(d)}\boldsymbol{\beta}_k&=\hat{\boldsymbol{\sigma}}^*-\begin{bmatrix}\hat{\boldsymbol{V}}^{*(d)} & \hat{\boldsymbol{U}}^{(d)}\end{bmatrix}\begin{bmatrix}\boldsymbol{c}\\ \boldsymbol{\beta}^*\end{bmatrix}\\&=\underbrace{\boldsymbol{\sigma}^*-\begin{bmatrix}\boldsymbol{V}^{*(d)} & \boldsymbol{U}^{(d)}\end{bmatrix}\boldsymbol{q}}_{=\mathbf{O}}+\underbrace{\left[\hat{\boldsymbol{\sigma}}^*-\boldsymbol{\sigma}^*\right]}_{:=\hat{\boldsymbol{\Delta}}_1}-\underbrace{\left[\hat{\boldsymbol{V}}^{*(d)}-\boldsymbol{V}^{*(d)}\right]}_{:=\hat{\boldsymbol{\Delta}}_2}\boldsymbol{c}-\underbrace{\left[\hat{\boldsymbol{U}}^{(d)}-\boldsymbol{U}^{(d)}\right]}_{:=\hat{\boldsymbol{\Delta}}_3}\boldsymbol{\beta}^*\\&=\hat{\boldsymbol{\Delta}}_1-\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}-\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}^*.\end{aligned}\tag{B.32}$$

Exploiting the previously defined notation and (B.32), we have

$$\begin{aligned}\left\|\hat{\boldsymbol{\sigma}}^*-\hat{\boldsymbol{V}}^{*(d)}\hat{\boldsymbol{c}}-\sum_{k=1}^K\hat{\boldsymbol{W}}_k^{*(d)}\hat{\boldsymbol{\beta}}_k\right\|_2^2&=\left\|\hat{\boldsymbol{\sigma}}^*-\hat{\boldsymbol{Q}}^{(d)}\hat{\boldsymbol{q}}\right\|_2^2=\left\|\left\{\hat{\boldsymbol{\sigma}}^*-\hat{\boldsymbol{Q}}^{(d)}\boldsymbol{q}\right\}-\left\{\hat{\boldsymbol{Q}}^{(d)}(\hat{\boldsymbol{q}}-\boldsymbol{q})\right\}\right\|_2^2\\&=\left\|\hat{\boldsymbol{\sigma}}^*-\hat{\boldsymbol{Q}}^{(d)}\boldsymbol{q}\right\|_2^2+\left\|\hat{\boldsymbol{Q}}^{(d)}(\hat{\boldsymbol{q}}-\boldsymbol{q})\right\|_2^2-2(\hat{\boldsymbol{q}}-\boldsymbol{q})'\hat{\boldsymbol{Q}}^{(d)\prime}\left(\hat{\boldsymbol{\Delta}}_1-\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}-\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}^*\right).\end{aligned}\tag{B.33}$$

We subsequently adjust the penalty function $P_\alpha(\boldsymbol{c})$ to incorporate the penalty on the coefficients in front of the exogenous variables. Define the index set g_k such that $\boldsymbol{q}_{g_k}=\boldsymbol{\beta}_k$ $(k=1\ldots,K)$ and enlarge $\mathscr{G}$ to $\mathscr{G}^*=\mathscr{G}\cup\bigcup_{k=1}^K g_k$. As $|\boldsymbol{\beta}_k|=|g_k|=N$, we have

$$P_\alpha(\boldsymbol{c})+\sum_{k=1}^K(1-\alpha)\sqrt{N}\left\|\boldsymbol{\beta}_k\right\|_2+\alpha\left\|\boldsymbol{\beta}_k\right\|_1=(1-\alpha)\sum_{g\in\mathscr{G}}^*\sqrt{|g|}\left\|\boldsymbol{q}_g\right\|_2+\alpha\left\|\boldsymbol{q}\right\|_1=:\widetilde{P}_\alpha(\boldsymbol{q}).$$

Using this newly defined norm $\widetilde{P}_\alpha(\boldsymbol{q})$, a more concise notation for the SPLASHX objective function is $\mathscr{L}_\alpha^*(\boldsymbol{q};\lambda)=\left\|\hat{\boldsymbol{\sigma}}^*-\hat{\boldsymbol{Q}}^{(d)}\boldsymbol{q}\right\|_2^2+\lambda\widetilde{P}_\alpha(\boldsymbol{q})$ and its dual norm $\widetilde{P}_\alpha^*(\boldsymbol{q})$ satisfies $\widetilde{P}_\alpha^*(\boldsymbol{q})\leq\|\boldsymbol{q}\|_\infty$ (see (B.23) in the main text). From $\mathscr{L}_\alpha^*(\hat{\boldsymbol{q}};\lambda)\leq\mathscr{L}_\alpha^*(\boldsymbol{q};\lambda)$, it follows that

$$\begin{aligned}\left\|\hat{\boldsymbol{Q}}^{(d)}(\hat{\boldsymbol{q}}-\boldsymbol{q})\right\|_2^2+\lambda\widetilde{P}_\alpha(\hat{\boldsymbol{q}})&\leq 2(\hat{\boldsymbol{q}}-\boldsymbol{q})'\hat{\boldsymbol{Q}}^{(d)\prime}\left(\hat{\boldsymbol{\Delta}}_1-\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}-\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}^*\right)+\lambda\widetilde{P}_\alpha(\boldsymbol{q})\\&\leq\widetilde{P}_\alpha(\hat{\boldsymbol{q}}-\boldsymbol{q})\left[2\widetilde{P}_\alpha^*\left(\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_1\right)+2\widetilde{P}_\alpha^*\left(\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}\right)+2\widetilde{P}_\alpha^*\left(\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}^*\right)\right]+\lambda\widetilde{P}_\alpha(\boldsymbol{q})\\&\leq\widetilde{P}_\alpha(\hat{\boldsymbol{q}}-\boldsymbol{q})\left[2\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_1\right\|_\infty+2\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}\right\|_\infty+2\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}^*\right\|_\infty\right]+\lambda\widetilde{P}_\alpha(\boldsymbol{q}).\end{aligned}\tag{B.34}$$

We subsequently define the following three sets:

$$\mathscr{H}_1^*(x)=\left\{2\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_1\right\|_\infty\leq x\right\},\qquad\mathscr{H}_2^*(x)=\left\{2\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}\right\|_\infty\leq x\right\}$$

and

$$\mathscr{H}_3^*(x)=\left\{2\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}\right\|_\infty^*\leq x\right\}.$$

Rescaling (B.34) by a factor 2 and on $\mathscr{H}_1^*(\frac{\lambda}{6})\cap\mathscr{H}_2^*(\frac{\lambda}{6})\cap\mathscr{H}_3^*(\frac{\lambda}{6})$, we get

$$2\left\|\hat{\boldsymbol{Q}}^{(d)}(\hat{\boldsymbol{q}}-\boldsymbol{q})\right\|_2^2+2\lambda\widetilde{P}_\alpha(\hat{\boldsymbol{q}})\leq\lambda\widetilde{P}_\alpha(\hat{\boldsymbol{q}}-\boldsymbol{q})+2\lambda\widetilde{P}_\alpha(\boldsymbol{q}).\tag{B.35}$$

We first manipulate the term $\widetilde{P}_\alpha(\hat{\boldsymbol{q}})$ in the LHS of (B.35). As in (B.19) of the main paper, we get

$$\widetilde{P}_\alpha(\hat{\boldsymbol{q}})\geq\widetilde{P}_{\alpha,S^*}(\boldsymbol{q})+\widetilde{P}_{\alpha,S^{*c}}(\hat{\boldsymbol{q}}-\boldsymbol{q})-\widetilde{P}_{\alpha,S^*}(\hat{\boldsymbol{q}}-\boldsymbol{q}),$$

where S^* is the index sets of all nonzero coefficients in $\boldsymbol{q}$. For the term $\widetilde{P}_\alpha(\hat{\boldsymbol{q}}-\boldsymbol{q})$ in the RHS of (B.35), if follows from (B.20) that $\widetilde{P}_\alpha(\hat{\boldsymbol{q}}-\boldsymbol{q})=\widetilde{P}_{\alpha,S^*}(\hat{\boldsymbol{q}}-\boldsymbol{q})+\widetilde{P}_{\alpha,S^{*c}}(\hat{\boldsymbol{q}}-\boldsymbol{q})$. We obtain

$$2\left\|\hat{\boldsymbol{Q}}^{(d)}(\hat{\boldsymbol{q}}-\boldsymbol{q})\right\|_2^2+\lambda\widetilde{P}_{\alpha,S^{*c}}(\hat{\boldsymbol{q}}-\boldsymbol{q})\leq 3\lambda\widetilde{P}_{\alpha,S^*}(\hat{\boldsymbol{q}}-\boldsymbol{q}).\tag{B.36}$$

Accordingly, with $N^*=N_c+NK$,

$$(\hat{\boldsymbol{q}}-\boldsymbol{q})\in\mathscr{C}_{N^*}^*(\mathscr{G},S):=\left\{\boldsymbol{\Delta}\in\mathbb{R}^{N^*}:\widetilde{P}_{\alpha,S^{*c}}(\boldsymbol{\Delta})\leq 3\widetilde{P}_{\alpha,S^*}(\boldsymbol{\Delta})\right\}.$$

Lemma 1 now goes through with $\bar{\omega}_\alpha^*$ as in Theorem 3. For applicability of Lemma 2, we define $\mathscr{V}^*(x) = \left\{\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_2 \leq x\right\}$ and assume $\mathscr{V}^*(\frac{\phi_0^*}{2})$. Derivations identical to those in (B.22) prove the main inequality of Theorem 3.

To derive the probability of the inequality being true, we need the probability of the occurrence of the event $\mathscr{H}_1^*(\frac{\lambda}{6}) \cap \mathscr{H}_2^*(\frac{\lambda}{6}) \cap \mathscr{H}_3^*(\frac{\lambda}{6}) \cap \mathscr{V}^*(\frac{\phi_0^*}{2})$. This probability is no smaller than

$$1 - \mathbb{P}\left(\mathscr{H}_1^*(\tfrac{\lambda}{6})^c\right) - \mathbb{P}\left(\mathscr{H}_2^*(\tfrac{\lambda}{6})^c\right) - \mathbb{P}\left(\mathscr{H}_3^*(\tfrac{\lambda}{6})^c\right) - \mathbb{P}\left(\mathscr{V}^*(\tfrac{\phi_0^*}{2})^c\right), \tag{B.37}$$

and all probabilities in (B.37) can be retraced to probabilities involving $\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\|_\vdash$. That is, bounding terms as in (B.24)–(B.25), we find

$$\begin{aligned}\mathbb{P}\left(\mathscr{H}_1^*(\tfrac{\lambda}{6})^c\right) &= \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_1\right\|_\infty > \frac{\lambda}{12}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash^2 + C_Q\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda}{12}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda^{1/2}}{\sqrt{24}}\right) + \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda}{24C_Q}\right),\end{aligned} \tag{B.38}$$

$$\begin{aligned}\mathbb{P}\left(\mathscr{H}_2^*(\tfrac{\lambda}{6})^c\right) &= \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_2\boldsymbol{c}\right\|_\infty > \frac{\lambda}{12}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\left[\hat{\boldsymbol{Q}}^{(d)} - \boldsymbol{Q}^{(d)}\right]\right\|_\infty > \frac{\lambda}{12}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda^{1/2}}{\sqrt{24}}\right) + \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda}{24C_Q}\right),\end{aligned} \tag{B.39}$$

and

$$\begin{aligned}\mathbb{P}\left(\mathscr{H}_3^*(\tfrac{\lambda}{6})^c\right) &= \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}}^{(d)\prime}\hat{\boldsymbol{\Delta}}_3\boldsymbol{\beta}^*\right\|_\infty > \frac{\lambda}{12}\right) \\ &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda^{1/2}}{\sqrt{24}C_\beta}\right) + \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\lambda}{24C_QC_\beta}\right)\end{aligned} \tag{B.40}$$

For (B.39) and (B.40), Assumption 1*(a)–(b) and Assumption 1*(c) are needed to guarantee $\|\boldsymbol{c}\|_1 \leq 1$ and $\|\boldsymbol{\beta}^*\|_1 \leq C_\beta$, respectively. Also, $\mathbb{P}\left(\mathscr{V}^*(\frac{\phi_0^*}{4})^c\right) = \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_2 > \frac{\phi_0^*}{2}\right) \leq \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > \frac{\phi_0^*}{2}\right)$. Given these bounds, (B.37) translates to

$$\mathbb{P}\left(\mathscr{H}_1^*(\tfrac{\lambda}{6}) \cap \mathscr{H}_2^*(\tfrac{\lambda}{6}) \cap \mathscr{H}_3^*(\tfrac{\lambda}{6}) \cap \mathscr{V}^*(\tfrac{\phi_0^*}{2})\right) \geq 1 - 7\mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > 6f^*(\lambda, \phi_0^*)\right),$$

where $f^*(\lambda, \phi_0^*) = \min\left\{\frac{\lambda^{1/2}}{12\sqrt{6}}, \frac{\lambda}{144C_Q}, \frac{\lambda^{1/2}}{12\sqrt{6C_\beta}}, \frac{\lambda}{144C_QC_\beta}, \frac{\phi_0^*}{12}\right\}$.

All that remains is a lower bound for $\mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash > x\right)$. We instead derive an upper bound for $\mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash \leq x\right)$ as follows

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\vdash \leq x\right) &= \mathbb{P}\left(\max\left\{\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_1, \left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_\infty\right\} \leq x\right) \geq \mathbb{P}\left((K+2)N\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_{max} \leq x\right) \\ &= 1 - \mathbb{P}\left(\left\|\hat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_{max} > \frac{x}{(K+2)N}\right) \\ &\geq 1 - \mathbb{P}\left(\bigcup_{1\leq i\leq (K+1)N, 1\leq j\leq (K+2)N}\left|[\hat{\boldsymbol{Q}} - \boldsymbol{Q}]_{ij}\right| > \frac{x}{(K+2)N}\right) \\ &\geq 1 - \sum_{i=1}^{(K+1)N}\sum_{j=1}^{(K+2)N}\mathbb{P}\left(\left|\sum_{t=2}^{T}\xi_{it}\xi_{jt} - \mathbb{E}(\xi_{it}\xi_{jt})\right| > \frac{Tx}{(K+2)N}\right), \\ &\geq \begin{cases} 1 - (K+1)(K+2)N^2\left[\left(b_1T^{(1-\delta)/3} + \frac{(K+2)Nb_3}{x}\right)\exp\left(-\frac{T^{(1-\delta)/3}}{2b_1^2}\right) + \frac{b_2(K+2)^dN^d}{x^dT^{\frac{\delta}{2}(d-1)}}\right] \\ \qquad\qquad \text{(polynomial tails)}, \\ 1 - (K+1)(K+2)N^2\left[\frac{\kappa_1(K+2)N}{x} + \frac{2}{\kappa_2}\left(\frac{Tx^2}{(K+2)N}\right)^{1/7}\right]\exp\left(-\frac{1}{\kappa_3}\left(\frac{Tx^2}{(K+2)^2N^2}\right)^{1/7}\right) \\ \qquad\qquad \text{(exponential tails)}, \end{cases}\end{aligned}$$

where ξ_{it} denotes a generic element of an autocovariance matrix (as in Lemma 3). In accordance with Theorem 3, these RHS probabilities are equivalent to $1 - \mathscr{P}_1^*(x, N, T)$ (polynomial tails) and $1 - \mathscr{P}_2^*(x, N, T)$ (exponential tails). □

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.105520.

References

Ahrens, A., Bhattacharjee, A., 2015. Two-step lasso estimation of the spatial weights matrix. Econometrics 3, 128–155.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37, 1705–1732.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. Lancet 360 (9341), 1233–1242.

Debarsy, N., LeSage, J., 2018. Flexible dependence modeling using convex combinations of different types of connectivity structures. Reg. Sci. Urban Econ. 69, 48–68.

Dou, B., Parrell, M.L., Yao, Q., 2016. Generalized Yule-Walker estimation for spatio-temporal models with unknown diagonal coefficients. J. Econometrics 194, 369–382.

Gao, Z., Ma, Y., Wang, H., Yao, Q., 2019. Banded spatio-temporal autoregressions. J. Econometrics 208, 211–230.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42, 1166–1202.

Gelper, S., Wilms, I., Croux, C., 2016. Identifying demand effects in a large network of product categories. J. Retail. 92, 25–39.

Guo, S., Wang, Y., Yao, Q., 2016. High-dimensional and banded vector autoregressions. Biometrika 103, 889–903.

Hamilton, J.D., 1994. Time Series Analysis, first ed. Princeton University Press.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: Principles and Practice. OTexts.

Junger, W.L., Ponce De Leon, A., 2015. Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96–104.

Knight, K., Fu, W., 2000. Asymptotics for LASSO-type estimators. Ann. Statist. 28, 1356–1378.

Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. J. Econometrics 186, 325–344.

Lam, C., Souza, P.C.L., 2014. Regularization for Spatial Panel Time Series using the Adaptive Lasso. Working paper. London School of Economics and Political Science.

Lam, C., Souza, P.C.L., 2019. Estimation and selection of spatial weight matrix in a spatial lag model. J. Bus. Econom. Statist. 38, 1–41.

Lee, L.-F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica 72, 1899–1925.

Lee, L.-F., Yu, J., 2010. Estimation of spatial autoregressive panel data models with fixed effects. J. Econometrics 154, 165–185.

Lee, L.-f., Yu, J., 2014. Efficient GMM estimation of spatial dynamic panel data models with fixed effects. J. Econometrics 180, 174–197.

Leeb, H., Pötscher, B.M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21, 21–59.

Leeb, H., Pötscher, B.M., 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. J. Econometrics 142 (1), 201–211.

Ma, Y., Guo, S., Wang, H., 2023. Sparse spatio-temporal autoregressions by profiling and bagging. J. Econometrics 232 (1), 132–147.

Masini, R.P., Medeiros, M.C., Mendes, E.F., 2022. Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. J. Time Series Anal. 43 (4), 532–557.

Medeiros, M.C., Mendes, E.F., 2016. ℓ_1-regularization of high-dimensional time series models with non-Gaussian and heteroskedastic errors. J. Econometrics 191, 255–271.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Statist. 22, 231–245.

Wang, D., Tsay, R.S., 2023. Rate-optimal robust estimation of high-dimensional vector autoregressive models. Ann. Statist. 51 (2), 846–877.

Yu, J., de Jong, R.M., Lee, L.-F., 2008. Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. J. Econometrics 146, 118–134.

Yu, J., de Jong, R., Lee, L.-f., 2012. Estimation for spatial dynamic panel data with fixed effects: The case of spatial cointegration. J. Econometrics 167, 16–37.

Zhang, X., Yu, J., 2018. Spatial weight matrix selection and model averaging for spatial autoregressive models. J. Econometrics 203, 1–18.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.