Vector Autoregressive Models with Spatially Structured Coefficients for Time Series on a Spatial Grid

Yuan YAN, Hsin-Cheng HUANG, and Marc G. GENTON

Motivated by the need to analyze readily available data collected in space and time, especially in environmental sciences, we propose a parsimonious spatiotemporal model for time series data on a spatial grid. In essence, our model is a vector autoregressive model that utilizes the spatial structure to achieve parsimony of autoregressive matrices at two levels. The first level ensures the sparsity of the autoregressive matrices using a lagged-neighborhood scheme. The second level performs a spatial clustering of the nonzero autoregressive coefficients such that within some subregions, nearby locations share the same autoregressive coefficients while across different subregions the coefficients may have distinct values. The model parameters are estimated using the penalized maximum likelihood with an adaptive fused Lasso penalty. The estimation procedure can be tailored to accommodate the need and prior knowledge of a modeler. Performance of the proposed estimation algorithm is examined in a simulation study. Our method gives reliable estimation results that are interpretable and especially useful to identify geographical subregions, within each of which, the time series have similar dynamical behavior with homogeneous autoregressive coefficients. We apply our model to a wind speed time series dataset generated from a climate model over Saudi Arabia to illustrate its power in explaining the dynamics by the spatially structured coefficients. Moreover, the estimated model can be useful for building stochastic weather generators as an approximation of the computationally expensive climate model.

Key Words: Adaptive fused Lasso; Coefficients homogeneity; Penalized maximum likelihood; Regularization; Spatial clusters; Spatiotemporal model.

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No: OSR-2018-CRG7-3742 and Academia Sinica Investigator Award AS-IA-109-M05.

Y. Yan, Department of Mathematics & Statistics, Dalhousie University, Halifax, NS B3H 4R2, Canada (E-mail: *yuan.yan@dal.ca*). H.-C. Huang, Institute of Statistical Science Academia Sinica, Taipei 115, Taiwan (E-mail: *hchuang@stat.sinica.edu.tw*). M. G. Genton, Statistics Program, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia (E-mail: *marc.genton@kaust.edu.sa*).

© 2021 International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 3, Pages 387–408
https://doi.org/10.1007/s13253-021-00444-4

1. INTRODUCTION

As spatiotemporal datasets are readily available in many scientific fields, various models have been proposed for modeling purposes. In this paper, we consider, in particular, time series data collected on a spatial grid, which are frequently encountered in the environmental, climate, and marine sciences. These data may be calculated from a satellite-borne instrument signal (e.g., scatterometer) or simulated from a physical model. Flexible and interpretable spatiotemporal models that are also parsimonious and computationally feasible are needed in order to understand the space-time dynamics underlying the data, make better forecasts, or generate fast simulation (stochastic weather generator) of such spatiotemporal data.

Traditional covariance-based space-time models are descriptive and are suitable for modeling data in continuous space and time, and allow interpolation (kriging) between locations and time points. However, these models consider no data generating mechanism. Alternatively, Cressie and Wikle (2011) recommend dynamical spatiotemporal models (DSTM) over the descriptive spatiotemporal covariance modeling. One commonly adopted linear DSTM for describing temporal dynamics is the vector autoregressive (VAR) model, where each variable corresponds to the process at one spatial location. Here, we give a brief review of a VAR(p) model in the DSTM framework. Throughout this paper, we assume $\mathbf{Z}_t = (Z_t(\mathbf{s}_1), \ldots, Z_t(\mathbf{s}_n))'$ for $t = 1, \ldots, T$ is a time series of a zero-mean spatial process at n locations. A Gaussian VAR(p) model of dimension n is defined as

$$\mathbf{Z}_t = \mathbf{A}_1 \mathbf{Z}_{t-1} + \cdots + \mathbf{A}_p \mathbf{Z}_{t-p} + \boldsymbol{\epsilon}_t, \quad t = 2, \ldots, T, \tag{1}$$

where $\mathbf{A}_i$ are the $n \times n$ matrices of autoregressive coefficients (i.e., transition/propagator /redistribution matrices) of order i, and $\boldsymbol{\epsilon}_t = (\epsilon_t(\mathbf{s}_1), \ldots, \epsilon_t(\mathbf{s}_n))'$ is a spatially correlated innovation process that follows an independent and identically distributed (i.i.d.) multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$, $\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Psi})$, and $\boldsymbol{\Psi}$ could be completely unknown or modeled by some commonly used parametric spatial covariance function.

The spatial setting for a VAR model can be both a curse and a blessing. On the one hand, for time series obtained at n gridded locations, the number of parameters, n^2, for each transition matrix $\mathbf{A}_i$ can be enormous, which makes estimation difficult and unstable, if not impossible. On the other hand, by taking advantage of the spatial structure, parsimonious assumptions can be made for the transition matrices in order to get sensible estimation and thus gain insights of the underlying dynamics of the phenomenon over space and time.

The main idea of our spatial VAR models are to enforce parsimony at two levels based on the gridded spatial structure. First, the sparsity of each $\mathbf{A}_i$ is obtained by using a lagged-neighborhood scheme for a given set of K_i neighbors. The nonzero coefficients in $\mathbf{A}_i$ can be interpreted as the influence or flow from a neighbor in a certain direction at time lag i. Next, we assume the nonzero coefficients are homogeneous in space such that within some subregions, nearby locations share the same autoregressive coefficients while across different subregions the coefficients may have distinct values. For parameter estimation, the usual least squares or maximum likelihood method suffer from high variances for time series data rich in space but not long in time, which ultimately lead to low interpretability.

In order to impose coefficients homogeneity, we propose a penalized likelihood method for parameter estimation, with an adaptive fused Lasso penalty on the coefficients.

The originality of our model truly lies in the combination of sparsity, coefficients homogeneity, and the estimation procedure for flexible and interpretable spatial VAR models. Since coefficients homogeneity can be regarded as an extension of the concept of sparsity, the two-level specifications of our spatial VAR models achieve further dimension reduction for the coefficients than the sparse VAR models and therefore are capable of dealing with time series data rich in space (hundreds of locations, which is generally considered high-dimensional for a VAR model) but not long in time ($T \ll n^2$), and is especially useful to identify spatial subregions from the clustered coefficients.

In Sect. 2, we put the two-level parsimony assumptions of our spatial VAR models under the context of related and alternative modeling strategies in the literature. In Sect. 3, we formalize our model for the VAR(1) case and develop as well as implement the parameter-estimation method. In Sect. 4, we demonstrate the performance of our estimation algorithm in a simulation study. In Sect. 5, we apply our model to a daily wind speed dataset on a spatial grid covering Saudi Arabia generated from a climate model in order to illustrate its usefulness. Limitations and possible extensions of our model are discussed in Sect. 6. Proofs and asymptotic properties of the penalized estimation are found in Appendix.

2. LITERATURE REVIEW

For the first level, our spatial VAR models adopt a lagged-neighborhood scheme for a number of pre-specified neighbors (not necessarily nearest), which endows each $\mathbf{A}_i$ with some given sparsity structure. The assumption of fixed sparsity structure for the transition matrices is simple while sensible and was adopted by multiple authors. The hierarchical model in Wikle et al. (1998) included a VAR(1) process with a lagged-nearest-neighbor scheme for the transition matrix and was applied to the analysis of monthly averaged maximum-temperature data. Rao (2008) introduced a VAR(p) model with a diagonal structure for each $\mathbf{A}_i$. Tagle et al. (2019) adopted a VAR(2) process with a lagged-nearest-neighbor scheme for $\mathbf{A}_1$ and a diagonal structure for $\mathbf{A}_2$ to analyze wind speed data.

Rather than a fixed sparsity structure, a more flexible sparsity structure on the VAR transition matrices based on the data can be achieved by a variable selection/model building scheme or a regularized estimation method. In econometrics, there is a vast literature on Bayesian VAR with different shrinkage priors on the vectorized transition matrices, e.g., Minnesota prior (Bańbura et al. 2010), adaptive hierarchical priors (Korobilis and Pettenuzzo 2019), global-local shrinkage priors, i.e., Dirichlet–Laplace prior (Kastner and Huber 2020). However, these Bayesian VAR models only considered sparsity in the transition matrices and did not adapt to spatially referenced data. For spatial data, de Luna and Genton (2005) proposed a variable selection procedure carried out at each location separately, in which the variables (locations) were ordered based on their relative distances (or importance, when additional information is available) from the location under study. First, they defined a spatial partial correlation function with respect to the order. Then they selected the model from the sample partial correlation functions, similarly to the order selection by checking the partial

autocorrelation function (PACF) in the time series context. Hsu et al. (2008) adopted the Lasso penalty for subset selection in a VAR model and estimated the nonzero coefficients simultaneously for a better forecasting performance. Basu and Michailidis (2015) provided a theoretical justification for estimating the VAR models with the Lasso penalty. Ngueyep and Serban (2015) used a weighted L_1 penalty as a function of the spatial distance between two locations for multilayer spatial VAR. Schweinberger et al. (2017) developed a two-step L_1 penalized least squares method to capture the sparsity structure of the transition matrices based on distance, while Monbet and Ailliot (2017) applied the smoothly clipped absolute deviation (SCAD) penalty to facilitate a sparse structure on both the transition matrices and the precision matrix ($\mathbf{\Psi}^{-1}$) for the innovation process.

Other approaches to reduce the number of parameters involved in a spatial VAR include parametric modeling of the transition matrices and possibly combined with dimension reduction. Ailliot et al. (2006) employed a VAR(1) model for wind fields with a latent process that described the motion of the air masses, where $\mathbf{A}_1$ and $\mathbf{\Psi}$ were both parameterized as a function of the latent translation of the wind fields. Wikle et al. (2001) and Katzfuss and Cressie (2012) implemented a dimension reduction technique with spatial basis functions, and modeled the resulting lower-dimensional process using a VAR(1) model, where the transition matrix had a parametric block structure related to the spatial resolution of the spatial bases. Bessac et al. (2015) built a spatiotemporal model in the state-space form, where the process at all points depended on a common univariate state variable, which was an AR(1) time series that contained large-scale spatial variation (regional information).

For the second level, our model assumes the nonzero coefficients that represent the influence from a neighbor in a certain direction are region-wise constant. Consequently, the dynamics of the process are the same within each subregions and are allowed to vary discontinuously across different subregions that are possibly with distinct geographical features. Previously, for linear regression, in order to identify homogeneous groups of predictors, Shen and Huang (2010) used a nonconvex and overcomplete penalty on pairwise differences of the coefficients. Ke et al. (2015) proposed a two-step procedure to first have preliminary ordered segments of the coefficients and then penalize the pairwise differences of the coefficients between adjacent segments and within individual segments. When the predictors are naturally ordered, it becomes the fused Lasso method proposed by Tibshirani et al. (2005). The method penalizes the successive differences among the coefficients to achieve local constancy of the coefficients. Zou (2006) introduced the adaptive Lasso, by giving different weights adaptively for different penalty terms given the data. Sun et al. (2016) used adaptive fused Lasso to estimate spatial and temporal quantile functions with spatially and temporally homogeneous groups, respectively. Under the spatial regression setting, Huang et al. (2010) proposed the spatial Lasso with homogeneous coefficients given known subregion structure. Recently, to identify the spatially clustered patterns in the coefficients, Li and Sang (2019) used the fused Lasso penalty on edges of the minimum spanning tree formed by the spatial locations.

Instead of the homogeneity approach, VAR models with smoothly varying diagonal coefficients were considered in Wikle et al. (1998) and Rao (2008). Wikle et al. (1998) let the diagonal coefficients varied spatially according to a conditionally autoregressive (CAR) model for the spatial lattice while coefficients related to the four nearest neighbors are kept

constant over space, while in Rao (2008), the diagonal coefficients vary smoothly in space, which are estimated by localized least squares, in order to accommodate the spatial non-stationarity of the process and allow spatial interpolation. The above two models are the VAR extensions of the spatially varying coefficient processes (Gelfand et al. 2003; Mu et al. 2018), which themselves are spatial extensions of the varying coefficient regression models (Fan and Zhang 1999).

Compared to the existing models, the novelty of our spatial VAR models is essentially to extend the pursuit of coefficients homogeneity to the spatial VAR setting, with the help of a pre-specified sparsity structure on the transition matrices that makes the nonzero coefficients physically meaningful with respect to the temporal dynamics.

3. METHODOLOGY

3.1. THE SPATIAL VAR(1) MODEL

We formalize our model specifications for the simple VAR(1) case first to elucidate the essence of our model. We let $D_s \subset \mathbb{R}^2$ be gridded locations, $D_s = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \{(i_1, i_2) : i_1 = 1, \ldots, n_x,\ i_2 = 1, \ldots, n_y\}$. We assume that for location $\mathbf{s}_i$, $Z_t(\mathbf{s}_i)$ is directly affected only by the lag-1 process at the K neighboring locations (including the location itself), $Z_{t-1}(\mathbf{s}_i + \mathbf{u}_k)$, $k = 1, \ldots, K$, for some translation vector $\mathbf{u}_k$. For example, if $\mathbf{u}_k = (-1, 0)'$, then $\mathbf{s}_i + \mathbf{u}_k$ is the left neighbor of $\mathbf{s}_i$, and we fix $\mathbf{u}_1 = \mathbf{0}$ so that the location $\mathbf{s}_i$ itself is always included. For a finite grid, boundary conditions are required when $\mathbf{s}_i + \mathbf{u}_k \notin D_s$. We are only interested in coefficients and dynamics for the n_I inner grid points, for each of which all K neighbors are within D_s. Since the dynamics of the boundary locations are not of interest and are modeled only to assist estimation of the coefficients related to the inner grid points, a simple model can be assumed for the $n_B = n - n_I$ boundary points in order to simulate data on the boundary. For easy notation, we order the grid points such that the first n_I points $\mathbf{s}_i$ for $i = 1, \ldots, n_I$ are the inner grid points. We adopt the lagged-neighborhood scheme only for these inner grid points; for the boundary grid points, we set $K = 1$ so that their temporal evolution depends only on themselves, i.e., AR(1) model. This model can be expressed as

$$\begin{aligned} Z_t(\mathbf{s}_i) &= \sum_{k=1}^{K} \alpha_k(\mathbf{s}_i) Z_{t-1}(\mathbf{s}_i + \mathbf{u}_k) + \epsilon_t(\mathbf{s}_i), && i = 1, \ldots, n_I,\ t = 2, \ldots, T, \\ Z_t(\mathbf{s}_i) &= \alpha_1(\mathbf{s}_i) Z_{t-1}(\mathbf{s}_i) + \epsilon_t(\mathbf{s}_i), && i = n_I + 1, \ldots, n,\ t = 2, \ldots, T, \end{aligned} \tag{2}$$

where $\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Psi})$ and $\alpha_k(\mathbf{s}_i)$ corresponds to the influence on the process at $\mathbf{s}_i$ by its kth neighbor. The lagged-nearest-neighbor scheme in Wikle et al. (1998) and Tagle et al. (2019) corresponds to the case when $K = 5$, which takes account of the influence by the location itself and its four nearest neighbors. These five locations on a grid may be referred to as the five-point stencil in the numerical analysis literature, the first-order stencil neighborhood scheme in high-performance computing, and the rook neighbors or the first-order neighborhood elsewhere. It can be easily extended to the second-order neighborhood

with $K = 9$ (queen neighbors), which includes the nearest eight neighbors and the location itself. More flexibly, $\mathbf{u}_k$ in (2) can be selected based on prior knowledge of the spatiotemporal process and the choices are not limited to the nearest neighbors. For example, if a river flows to the east, then it is sensible to model the chemical concentration at a location in the river to depend on only its western neighbors.

For each $k = 1, \ldots, K$, we let $\boldsymbol{\alpha}_k = (\alpha_k(\mathbf{s}_1), \ldots, \alpha_k(\mathbf{s}_{n_I}))'$ be the vector of the unknown parameters of interest, which are the coefficients related to the kth neighbor for each of the n_I inner grid points. Let $\boldsymbol{\alpha}_{\mathrm{B}} = (\alpha_1(\mathbf{s}_{n_I+1}), \ldots, \alpha_1(\mathbf{s}_n))'$ be the vector of the unknown autoregressive coefficients of the boundary points. Time series of a spatial process often have some regional structure such that locations within one region share the same spatiotemporal dynamics while the dynamics could vary discontinuously across different subregions. Therefore, a spatial coefficient-grouping structure such that $\boldsymbol{\alpha}_k$ is composed of spatial clusters of values for each $k = 1, \ldots, K$ is desirable. We assume no spatial structure for $\boldsymbol{\alpha}_{\mathrm{B}}$; again, these coefficients are of little practical importance and are only modeled to allow simulation and forecasting several steps ahead as well as to simplify notations. Figure 1 of our simulation design gives an intuitive visual illustration of the idea of spatial homogeneity in the autoregressive coefficients.

The model defined by (2) can be written as the VAR model in (1) with order $p = 1$ and $\mathbf{A}_1 = \mathbf{A}$. We let a_{ij} be the (i, j)th entry of $\mathbf{A}$. With the lagged-neighborhood scheme, $\mathbf{A}$ has a sparse form such that $a_{ij} \neq 0$, only if $\mathbf{s}_j = \mathbf{s}_i + \mathbf{u}_k$ for some k and there are at most a total of $m = Kn_I + n_B$ nonzero elements in $\mathbf{A}$. Since the sparsity of $\mathbf{A}$ is combined with the spatial clustered structure of each $\boldsymbol{\alpha}_k$, the reduction in the number of parameters is drastic compared to the n^2 parameters required by the full VAR(1) model. We let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K', \boldsymbol{\alpha}_{\mathrm{B}}')'$ be the vector of all the unknown coefficients of length m; then, the sparse transition matrix $\mathbf{A}$ depends only on $\boldsymbol{\alpha}$ through $\mathrm{vec}(\mathbf{A}) = \mathbf{P}\boldsymbol{\alpha}$, where $\mathbf{P}$ is a $n^2 \times m$ matrix of zeros and ones that indicates the positions of the corresponding nonzero elements in $\mathrm{vec}(\mathbf{A})$ for $\boldsymbol{\alpha}$.

3.2. Parameter Estimation by the Adaptive Fused Lasso

For our model, the dynamical spatiotemporal coefficients $\boldsymbol{\alpha}$ and the matrix $\boldsymbol{\Psi}$ that models the spatial covariance of the residuals need to be estimated. To capture the spatially homogeneous groups of $\boldsymbol{\alpha}$, we consider a penalized likelihood method with the adaptive fused Lasso penalty. Tibshirani and Taylor (2011) developed a path algorithm for a 'generalized' Lasso problem that included the Lasso, fused Lasso, and adaptive Lasso as special cases. This generalized Lasso problem can be solved efficiently and is easy to implement using the 'genlasso' package (Arnold and Tibshirani 2020) in R (R Development Core Team 2020).

We consider the log-likelihood of $\boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_T$ conditioning on $\boldsymbol{Z}_1$ for our model with $\boldsymbol{\Psi}$ given. For any matrix $\boldsymbol{X}$ and vector $\boldsymbol{y}$ that satisfy

$$\boldsymbol{X}'\boldsymbol{X} = \mathbf{P}'\left\{\left(\sum_{t=2}^{T} \boldsymbol{Z}_{t-1}\boldsymbol{Z}_{t-1}'\right) \otimes \boldsymbol{\Psi}^{-1}\right\}\mathbf{P}, \tag{3}$$

$$\boldsymbol{X}'\boldsymbol{y} = \mathbf{P}'\mathrm{vec}\left(\boldsymbol{\Psi}^{-1}\sum_{t=2}^{T} \boldsymbol{Z}_t\boldsymbol{Z}_{t-1}'\right), \tag{4}$$

the conditional log-likelihood can be written as a second-order polynomial of $\boldsymbol{\alpha}$ (the proof is found in Appendix):

$$l(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{t=2}^{T}(\mathbf{Z}_t - \mathbf{A}\mathbf{Z}_{t-1})'\boldsymbol{\Psi}^{-1}(\mathbf{Z}_t - \mathbf{A}\mathbf{Z}_{t-1}) = -\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\right\|_2^2 + \text{constant}, \quad (5)$$

where $\|\cdot\|_2$ is the ℓ_2-norm. One way to choose $\boldsymbol{X}$ is to compute the Cholesky decomposition of the matrix on the right-hand side of (3) and, then, solve for $\boldsymbol{y}$ from (4).

We consider a penalized maximum likelihood (PML) method with the following objective function:

$$F(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda\sum_{k=1}^{K}\sum_{i\sim j} w_{k,i,j}|\alpha_k(\mathbf{s}_i) - \alpha_k(\mathbf{s}_j)|, \quad (6)$$

where the last term is an adaptive 2D fusion problem penalty that facilitates the grouping of neighboring coefficients, $i \sim j$ denotes that $\mathbf{s}_i$ is a direct neighbor of $\mathbf{s}_j$ (there is an edge between $\mathbf{s}_i$ and $\mathbf{s}_j$ on the grid), for $i, j = 1, \ldots, n_I$, λ is a tuning parameter that controls the degree of grouping, and $\{w_{k,i,j}\}$ are known weight values set to $w_{k,i,j} = |\tilde{\alpha}_k(\mathbf{s}_i) - \tilde{\alpha}_k(\mathbf{s}_j)|^{-\gamma}$, where $\tilde{\boldsymbol{\alpha}}$ is a root-T consistent estimator of $\boldsymbol{\alpha}$ and γ is some fixed positive constant. We penalize the pairwise difference of $\alpha_k(\cdot)$ between direct neighbors within the inner grid points. For each $k = 1, \ldots, K$, the number of penalty terms equals to the number of edges in the inner grid. Instead of the extensive pairwise difference penalty used by Shen and Huang (2010) or the difference between the coefficient of a location and the average of its neighboring coefficients used by Sun et al. (2016), we take advantage of the spatial configuration and penalize on 'just enough' pairs for coefficients clustering. Even though additional penalty terms could be put on individual coefficients to encourage some coefficients to be exactly zeros, with penalty terms only on the pairwise differences estimation of true zero coefficients can be close enough to zero and avoid extra computational difficulty.

The maximization problem (6), given $\boldsymbol{\Psi}$, is a generalized Lasso problem,

$$\min_{\boldsymbol{\alpha}}\left\{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\mathbf{D}\boldsymbol{\alpha}\|_1\right\}, \quad (7)$$

where $\|\cdot\|_1$ is the ℓ_1-norm for some penalty matrix $\mathbf{D}$, which is determined by the grid structure as well as $w_{k,i,j}$. The neighborhood structure of $\boldsymbol{\alpha}$ given by (6) and (7) corresponds to a graph formed by K disjoint subgraphs with the same grid structure; K determines the level of sparsity for $\mathbf{A}$, while the penalty on pairwise difference between direct neighbors assists in capturing the spatial homogeneity of the VAR coefficients.

When $\lambda = 0$, the solution of (7) corresponds to the restricted generalized least squares (GLS) estimator of $\mathbf{A}$ with the constraints that $a_{ij} = 0$ except for the m parameters corresponding to $\boldsymbol{\alpha}$. As $\lambda \to \infty$, $\alpha_k(\cdot)$ reduces to a constant function because the network corresponding to $\{\mathbf{s}_i : i = 1, \ldots, n_I\}$ is connected. A proper tuning parameter λ can be selected by fivefold cross-validation (Bergmeir and Benítez 2012) or by the Bayesian information criterion (BIC), $\text{BIC}(\lambda) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}_\lambda\|_2^2 + (T-1)\log(|\boldsymbol{\Psi}_\lambda|) + \log(T-1)\{\text{df}(\boldsymbol{X}\boldsymbol{\alpha}_\lambda) + \text{df}(\boldsymbol{\Psi})\}$,

where $\boldsymbol{\alpha}_\lambda$ is the solution to the generalized Lasso problem in (7) for a given λ, $\boldsymbol{\Psi}_\lambda$ is an estimator of $\boldsymbol{\Psi}$ given $\boldsymbol{\alpha}_\lambda$, $\mathrm{df}(\boldsymbol{X}\boldsymbol{\alpha}_\lambda)$ can be estimated from the number of distinct values in $\boldsymbol{\alpha}_\lambda$ (Tibshirani and Taylor 2011), and $\mathrm{df}(\boldsymbol{\Psi})$ is the number of parameters in estimating $\boldsymbol{\Psi}$. We can estimate $\boldsymbol{\Psi}$, given $\boldsymbol{\alpha}$ and the corresponding $\mathbf{A}$, using the maximum likelihood estimator:

$$\hat{\boldsymbol{\Psi}} = \frac{1}{T-1}\sum_{t=2}^{T}(\mathbf{Z}_t - \mathbf{A}\mathbf{Z}_{t-1})(\mathbf{Z}_t - \mathbf{A}\mathbf{Z}_{t-1})'. \tag{8}$$

However, there is a trade-off in estimating $\mathbf{A}$ and $\boldsymbol{\Psi}$, namely they compete to account for the spatial correlation. When $\boldsymbol{\Psi}$ is difficult to estimate, it could be assumed to be diagonal and estimated accordingly. For very short time series when $T < n$, the nonparametric estimation of $\boldsymbol{\Psi}$ in (8) is singular. In this case, further regularization can be imposed on $\boldsymbol{\Psi}$ or a parametric model can be adopted to estimate $\boldsymbol{\Psi}$, such as the Matérn covariance family.

We propose the following algorithm for estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\Psi}$ iteratively in two steps to get the adaptive fused Lasso estimator. We start with $\boldsymbol{\Psi} = \mathbf{I}_n$ and solve the PML (7) with uniform weights and λ selected by the BIC to get the fused Lasso estimator. The restricted OLS estimator could be obtained by setting $\lambda = 0$ in step I. Then in step II, we compute weights from the fused Lasso estimator and solve the PML (7) with $\boldsymbol{\Psi}$ estimated from the residuals $\boldsymbol{Z}_t - \hat{\mathbf{A}}^{(1)}\boldsymbol{Z}_{t-1}$ by (8) to obtain the adaptive fused Lasso estimator.

ALGORITHM:

1. Step I: Fused Lasso and restricted OLS
   (a) Fix $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}^{(0)} = \mathbf{I}_n$ and set $w_{k,i,j} \equiv 1$, solve (7) with a sequence of λ to obtain the fused Lasso estimators $\hat{\boldsymbol{\alpha}}_\lambda^{(1)}$ (where $\lambda = 0$ corresponds to the restricted OLS estimator with the constraints that $a_{ij} = 0$ except for the m parameters corresponding to $\boldsymbol{\alpha}$);
   (b) For each λ, find and fix the corresponding $\mathbf{A} = \hat{\mathbf{A}}_\lambda^{(1)}$ and get $\hat{\boldsymbol{\Psi}}_\lambda^{(1)}$ from (8), then calculate BIC with $\hat{\boldsymbol{\alpha}}_\lambda^{(1)}$ and $\hat{\boldsymbol{\Psi}}_\lambda^{(1)}$;
   (c) Select λ by BIC and get the fused Lasso estimation $\hat{\boldsymbol{\alpha}}^{(1)}$, the corresponding $\hat{\mathbf{A}}^{(1)}$ and $\hat{\boldsymbol{\Psi}}^{(1)}$;
2. Step II: Adaptive fused Lasso and restricted GLS
   (a) Fix $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}^{(1)}$ and update $w_{k,i,j} = |\hat{\alpha}_k^{(1)}(\mathbf{s}_i) - \hat{\alpha}_k^{(1)}(\mathbf{s}_j)|^{-1}$, to obtain the adaptive fused Lasso estimators $\hat{\boldsymbol{\alpha}}_\lambda$ by solving (7) with a sequence of λ (where $\lambda = 0$ corresponds to the restricted GLS estimator);
   (b) For each λ, find and fix the corresponding $\mathbf{A} = \hat{\mathbf{A}}_\lambda$ and get $\hat{\boldsymbol{\Psi}}_\lambda$ from (8), then calculate BIC with $\hat{\boldsymbol{\alpha}}_\lambda$ and $\hat{\boldsymbol{\Psi}}_\lambda$;
   (c) Select λ by BIC and get the adaptive fused Lasso estimation $\hat{\boldsymbol{\alpha}}$, the corresponding $\hat{\mathbf{A}}$ and $\hat{\boldsymbol{\Psi}}$.

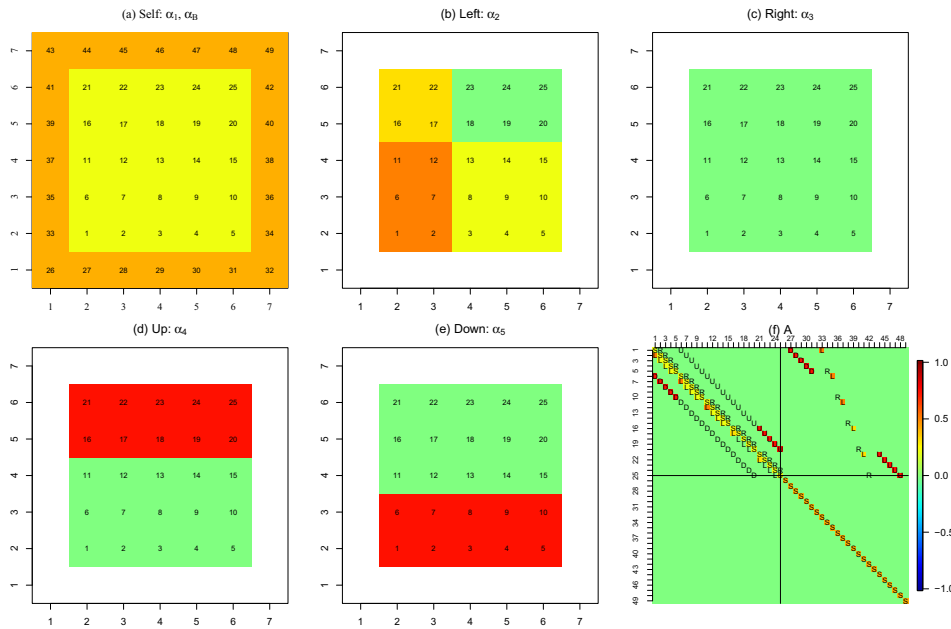

Figure 1. **a** Values of $\boldsymbol{\alpha}_1$ (yellow) and $\boldsymbol{\alpha}_B$ (orange), and the indices of the 7×7 grid; **b** values of $\boldsymbol{\alpha}_2$; **c** values of $\boldsymbol{\alpha}_3$; **d** values of $\boldsymbol{\alpha}_4$; **e** values of $\boldsymbol{\alpha}_5$; **f** values of **A** with its elements corresponding to the indices in (**a**) (Color figure online) .

This adaptive fused Lasso estimator possesses the asymptotic properties of consistency in grouping and asymptotic normality. Details and proof are provided in Appendix.

3.3. EXTENSIONS

Extending to the general VAR(p) case, our model assumes for $j = 1, \ldots, p$, the process at each of the $i = 1, \ldots, n_I$ inner grid points is directly affected only by the lag-j process at the prescribed K_j neighboring locations. For boundary grid points, an AR(p) process is assumed. The model can be written as

$$Z_t(\mathbf{s}_i) = \sum_{j=1}^{p} \sum_{k=1}^{K_j} \alpha_{jk}(\mathbf{s}_i) Z_{t-j}(\mathbf{s}_i + \mathbf{u}_k) + \epsilon_t(\mathbf{s}_i), \qquad i = 1, \ldots, n_I,\ t = p+1, \ldots, T,$$

$$Z_t(\mathbf{s}_i) = \sum_{j=1}^{p} \alpha_j(\mathbf{s}_i) Z_{t-j}(\mathbf{s}_i) + \epsilon_t(\mathbf{s}_i), \qquad i = n_I + 1, \ldots, n,\ t = p+1, \ldots, T.$$

Again, the lagged-neighborhood scheme ensures sparsity of each of the transition matrices. Then the homogeneity assumption is on $\boldsymbol{\alpha}_{jk} = (\alpha_{jk}(\mathbf{s}_1), \ldots, \alpha_{jk}(\mathbf{s}_{n_I}))'$ for each $j = 1, \ldots, p$ and $k = 1, \ldots, K_j$. Detailed equations and estimation algorithm can be derived similarly as the VAR(1) case.

Moreover, instead of using a fixed order p and pre-defined neighboring scheme, order and neighbor selections can be made simultaneously using an additional Lasso step or various variable selection methods mentioned in Sect. 2.

4. SIMULATION STUDY

4.1. SIMULATION DESIGN

We examine the performance of our estimation algorithm in a simulation study. We simulate time series data on a 7×7 grid ($n = 49$) in a unit square from our model (2) with $T \in \{100, 500\}$ and $K = 5$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_5', \boldsymbol{\alpha}_{\mathrm{B}}')'$ and the corresponding $\mathbf{A}$ are shown in Fig. 1. The spatially correlated error term $\boldsymbol{\epsilon}_t$ is generated from a multivariate Gaussian distribution with mean $\mathbf{0}$, and the covariance matrix $\boldsymbol{\Psi}$ is determined by an exponential covariance function with unit variance and a range parameter of 0.25. For this 7×7 square grid, we have $n_I = 25, n_B = 24$, and thus, $m = 149$ for the length of $\boldsymbol{\alpha}$. We design the structure of $\boldsymbol{\alpha}$ such that $\{\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_5'\}$ have different numbers of clusters (i.e., distinct elements). For example, $\boldsymbol{\alpha}_4$ and $\boldsymbol{\alpha}_5$, corresponding to the neighbors above and below, have two clusters which mimic effects by a mountainous region in the middle separating the two subregions above and below; $\boldsymbol{\alpha}_2$ corresponding to the left neighbor has four clusters; $\boldsymbol{\alpha}_1$ corresponding to the location itself is a constant vector; and $\boldsymbol{\alpha}_3$ corresponding to the right neighbor is constant $\mathbf{0}$.

We estimate $\mathbf{A}$ and $\boldsymbol{\Psi}$ using the proposed algorithm to get the adaptive fused Lasso estimator. For comparison, we also consider the restricted OLS estimator (with $\lambda = 0$ in step I(a)), the restricted GLS estimator (with $\lambda = 0$ in step II(a)), and the fused Lasso estimator $\hat{\boldsymbol{\alpha}}^{(1)}$ (in step I(c)) of our algorithm.

4.2. ESTIMATION RESULTS

Even though the structure of the coefficients is best illustrated two-dimensionally as shown in Fig. 1, we can 'stretch' (vectorize) them into a single vector ordered by the numbers labeled in each panel of Fig. 1. Then, we could use the functional boxplot (Sun and Genton 2011) to visualize the performance of the estimators based on 500 simulation replicates. Figure 2 shows the functional boxplots for the estimated coefficients of the five neighbors (in five rows) by the four methods (in four columns).

From Fig. 2, we see that both the fused Lasso and the adaptive fused Lasso estimators have much smaller variances and are, therefore, more stable than the restricted OLS and GLS estimators. The functional boxplot is designed to capture the structure of the estimated curve as a whole instead of estimation at individual points. Comparing the median curves of each plot, we see that the non-regularized methods do not reproduce the structure of the coefficients, while the fused Lasso and the adaptive fused Lasso both capture the grouping of the coefficients very well. There is a small departure of the median curves from the green lines for the fused Lasso, and the adaptive fused Lasso fixes this bias problem.

Although $\boldsymbol{\alpha}_3 = \mathbf{0}$, the estimates of $\boldsymbol{\alpha}_3$ from our adaptive fused Lasso are only very close to but not exactly zeros. Tibshirani et al. (2005) used a sparse fused Lasso that penalized both the coefficients and the pairwise differences of the coefficients to enforce both sparsity and grouping. However, in our algorithm, we only penalize the pairwise differences, not the coefficients themselves, i.e., the 2D fusion problem. In an extra simulation (results not shown), we used the sparse fused Lasso estimator, which can also be written in the form

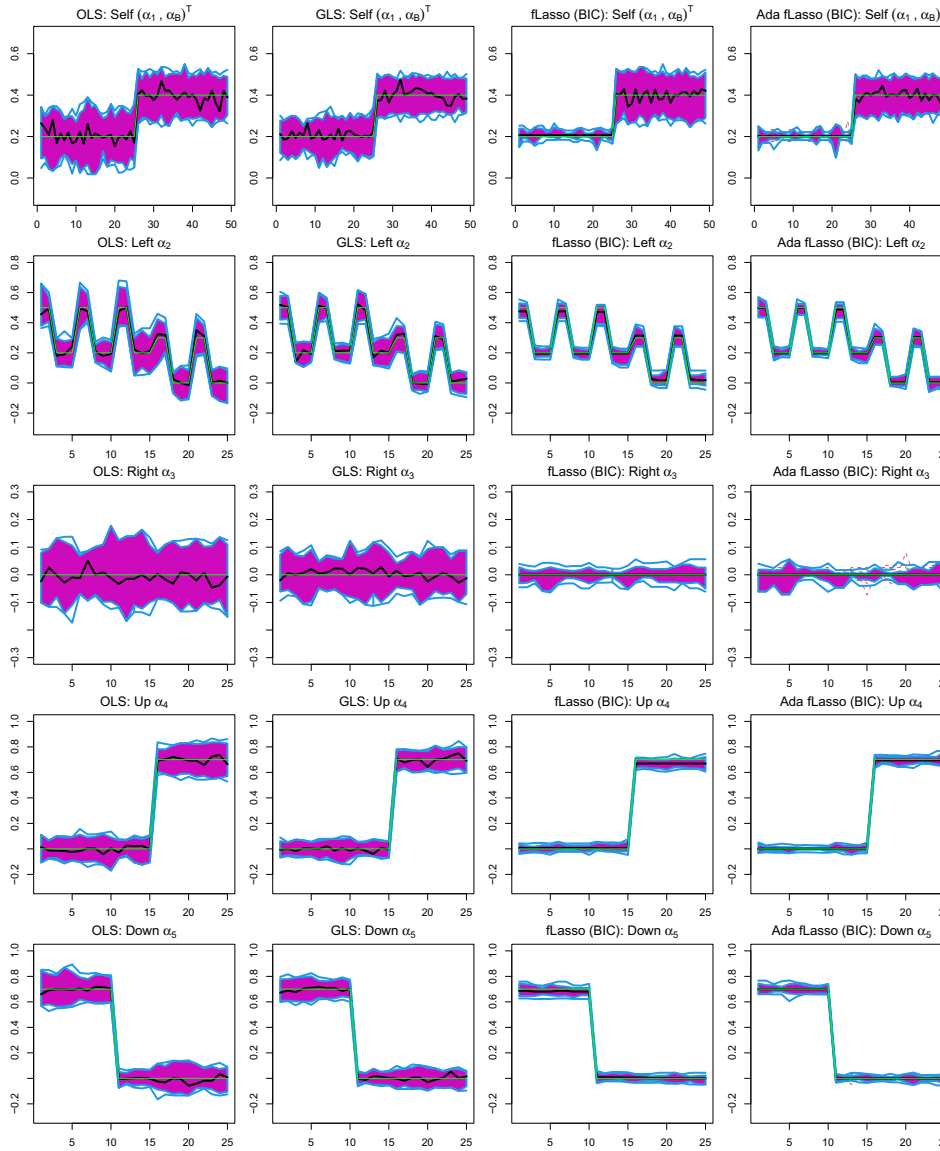

Figure 2. Functional boxplots of the estimated $\boldsymbol{\alpha}_k, k = 1, \ldots, 5$ using restricted OLS, restricted GLS, fused Lasso, and adaptive fused Lasso methods with $K = 5$ and $T = 500$ from 500 simulations. True values for $\boldsymbol{\alpha}_k$ are indicated by the green lines, where the order of the coefficients $\boldsymbol{\alpha}_k$ is indicated in the corresponding panel of Figure 1 .

of (7), where $\mathbf{D}$ has Kn_I additional rows added to penalize the coefficients themselves. We found that our adaptive fused Lasso method without the Lasso sparsity penalization not only performs as satisfactorily as the sparse fused Lasso method, but it is also much more efficient in computation.

Though the choice of K might seem subjective, we can always specify a larger K until the residuals pass the standard model validation procedure for a VAR model. Using a larger

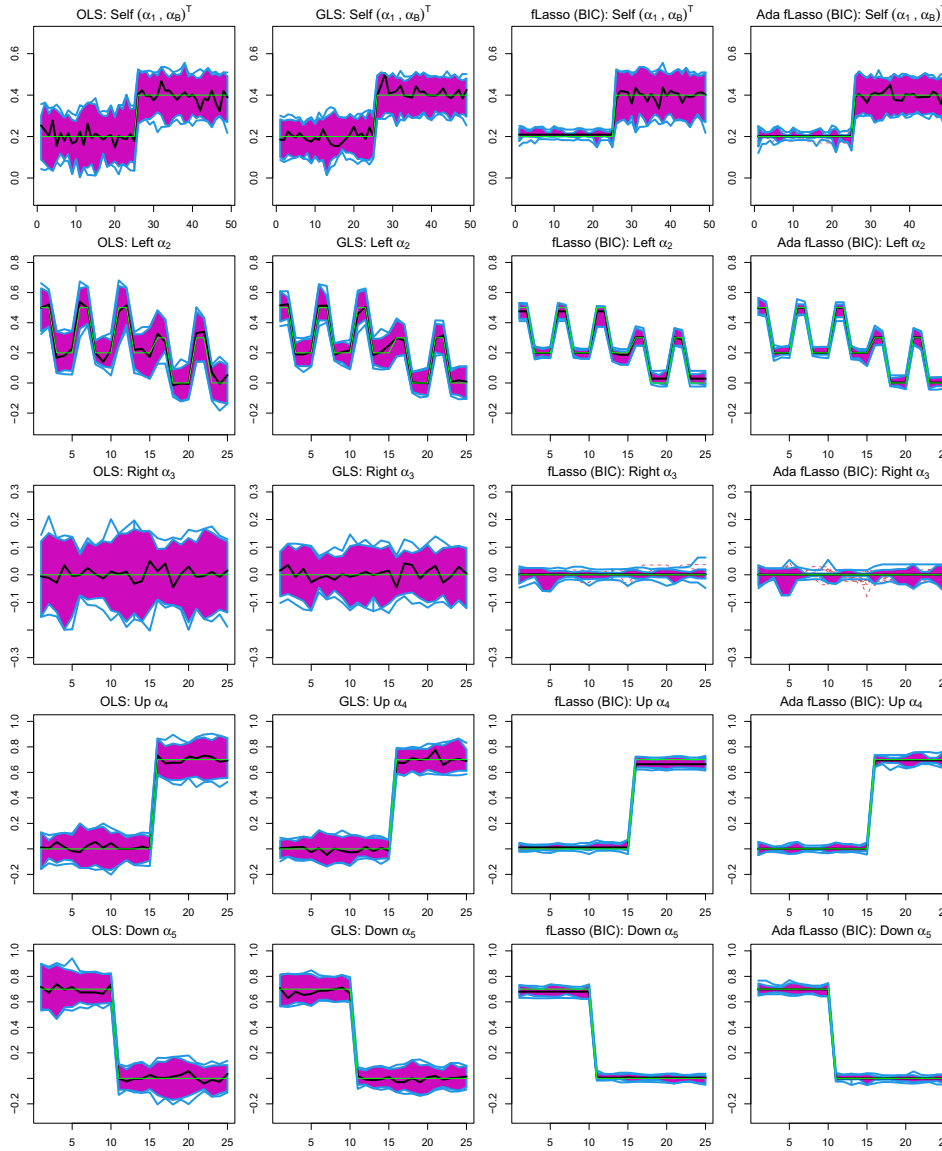

Figure 3. Functional boxplots of the estimated $\boldsymbol{\alpha}_k, k = 1, \ldots, 5$ using restricted OLS, restricted GLS, fused Lasso, and adaptive fused Lasso methods with $K = 9$ and $T = 500$ from 500 simulations. True values for $\boldsymbol{\alpha}_k$ are indicated by the green lines, where the order of the coefficients $\boldsymbol{\alpha}_k$ is indicated in the corresponding panel of Figure 1 .

K increases the computational effort, but does not significantly affect the performance of the nonzero coefficients since the adaptive fused Lasso already enforces sparsity in the coefficients. Figure 3 shows estimation results with $K = 9$ in the same simulation design as described before. Comparing Figs. 2, 3, we see that the estimates of $\boldsymbol{\alpha}_k, k = 1, \ldots, 5$, by the fused Lasso and the adaptive fused Lasso with $K = 9$ are similar to those with $K = 5$, whereas the variances of the unregularized estimators by the restricted OLS and the

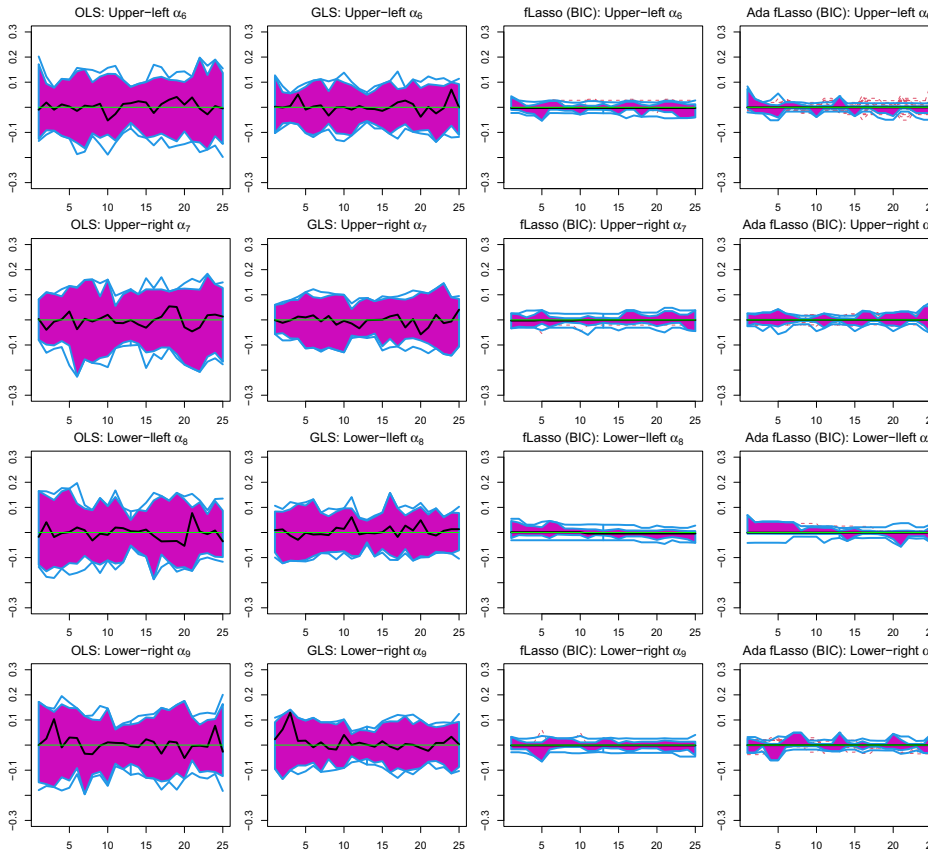

Figure 4. Functional boxplots of the estimated $\boldsymbol{\alpha}_k, k = 6, \ldots, 9$ using restricted OLS, restricted GLS, fused Lasso, and adaptive fused Lasso methods with $K = 9$ and $T = 500$ from 500 simulations. True values for $\boldsymbol{\alpha}_k$ are indicated by the green lines, which are all zeros.

restricted GLS are slightly expanded. Functional boxplots of the estimated $\boldsymbol{\alpha}_k$ (where the true $\boldsymbol{\alpha}_k$ is $\mathbf{0}$) for $k = 6, \ldots, 9$ are shown in Fig. 4. Again, we see the restricted OLS estimates and the restricted GLS estimates have high variances, whereas both the fused Lasso and the adaptive fused Lasso can capture the sparsity in $\boldsymbol{\alpha}$ much better.

The pattern of the estimated $\boldsymbol{\Psi}$ is intrinsically two-dimensional; therefore, we resort to the surface boxplot (Genton et al. 2014), which is the two-dimensional version of the functional boxplot, to demonstrate the estimation results. Figure 5 shows the surface boxplot for the proposed estimator of $\boldsymbol{\Psi}$ based on 500 simulation replicates. We see that the median image is very close to the true structure, even though our method is somewhat nonparametric.

This simulation study shows that the structure of the coefficients in our model can be captured and estimated quite well by the proposed estimation algorithm, despite the fairly small number of time points T relative to the number of parameters $n^2 + n(n+1)/2$ to be estimated for a conventional VAR(1) model with dimension n.

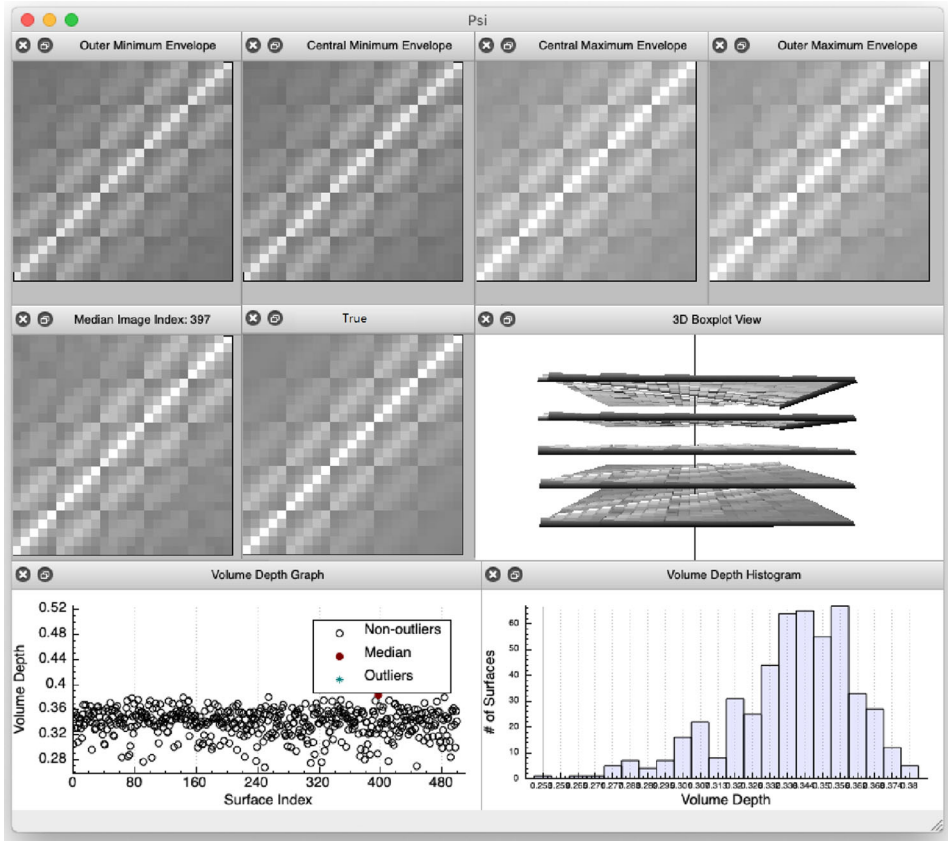

Figure 5. Surface boxplot of the estimated $\mathbf{\Psi}$ using our algorithm with $T = 500$ from 500 simulations.

Table 1. PMSE of h-step ahead forecasts for $h = 1, 2, 3$ based on four estimates of $\mathbf{A}$ from 500 simulations for both the 49 locations and the inner 25 points; the standard errors are given in parentheses

	h	Restricted OLS	Restricted GLS	Fused Lasso	Adaptive fused Lasso
49 locations	1	1.0060 (0.00012)	1.0033 (0.00006)	1.0025 (0.00006)	1.0012 (0.00003)
	2	1.0035 (0.00008)	1.0020 (0.00004)	1.0018 (0.00004)	1.0008 (0.00002)
	3	1.0019 (0.00005)	1.0011 (0.00003)	1.0010 (0.00003)	1.0005 (0.00001)
25 inner points	1	1.0098 (0.00021)	1.0052 (0.00010)	1.0030 (0.00008)	1.0010 (0.00004)
	2	1.0058 (0.00013)	1.0032 (0.00007)	1.0025 (0.00007)	1.0010 (0.00002)
	3	1.0033 (0.00009)	1.0018 (0.00005)	1.0017 (0.00005)	1.0007 (0.00002)

4.3. FORECAST RESULTS

We compare the h-step-ahead forecasts of a VAR(1) model for $h = 1, 2, 3$ at all grid points when using the estimated $\mathbf{A}$ by the four methods under consideration (i.e., restricted OLS, restricted GLS, fused Lasso, and adaptive fused Lasso). To measure the forecasting performance, we consider the normalized h-step prediction mean squared error (PMSE),

defined as (modified from Hsu et al. 2008):

$$\begin{aligned}\text{PMSE} &= \frac{1}{n}\mathbb{E}\left[(\boldsymbol{Z}_{t+h} - \hat{\boldsymbol{Z}}_{t+h})'\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{Z}_{t+h} - \hat{\boldsymbol{Z}}_{t+h})\right] \\ &= 1 + \frac{1}{49}\mathbb{E}\left[(\hat{\boldsymbol{Z}}^*_{t+h} - \hat{\boldsymbol{Z}}_{t+h})'\boldsymbol{\Sigma}_h^{-1}(\hat{\boldsymbol{Z}}^*_{t+h} - \hat{\boldsymbol{Z}}_{t+h})\right], \end{aligned} \tag{9}$$

where $\hat{\boldsymbol{Z}}^*_{t+h}$ and $\hat{\boldsymbol{Z}}_{t+h}$ are the h-step best linear predictors based on the true and estimated models, respectively, and $\boldsymbol{\Sigma}_h$ is the theoretical h-step prediction variance with the true model. The PMSE compares forecast performance relative to forecasting with the true parameters and thus has the advantage of resistance to unusual data at the forecasting time points. The PMSE for only the inner grid points can be obtained by dividing by 25 instead of 49, and using the sub-vector of $\hat{\boldsymbol{Z}}^{*(i)}_{t+h}$ and $\hat{\boldsymbol{Z}}^{(i)}_{t+h}$, and sub-matrix $\boldsymbol{\Sigma}_h$ corresponding to the inner grid points in (9).

Table 1 reports the empirical PMSE based on the four estimation methods for all 49 locations as well as the inner 25 points. We see that forecasts with $\mathbf{A}$ estimated using the adaptive fused Lasso always have the smallest PMSE.

5. APPLICATION TO DAILY WIND SPEED DATA

We illustrate the usefulness of our model by a real wind speed data example. The Large Ensemble Project (LENS) dataset is publicly available and consists of 30 ensembles of daily wind speeds over the globe at a spatial resolution of 1.25° longitude and 0.94° latitude from the year 1920 to 2100 (Kay et al. 2015) generated from a climate model. We randomly select one ensemble from the dataset and use the historical 86 years (1920-2005) of data at $n = 195$ points that covering the domain of Saudi Arabia, plus one additional outer layer of the domain for boundary conditions. First, we estimate the seasonality of the daily wind speeds at each location by taking the average wind speed across each day of all 86 years at each location. We remove the seasonal effect from each time point and get the residual wind speed for each location. Since the residual daily wind speed is right-skewed, we use the Tukey g-and-h transformation, as in Yan and Genton (2019), to Gaussianize the residuals. We estimate the transformation at each grid point and use the transformed residuals to fit our spatially structured VAR(1) model. Maps of the estimated parameters related to the transformation can be found in Yan and Genton (2019). We use the Gaussianized residual daily wind speeds of January and July in the 86 years ($T = (31-1)\times 86-3 = 2577$, ignoring points with year jumps and the last 3 time points for forecasting) to fit our model with $K = 5$, and use our proposed adaptive fused Lasso algorithm to estimate the coefficients. For this grid configuration, we have $n_I = 149$, $n_B = 46$, and hence $m = 791$ for the length of $\boldsymbol{\alpha}$. Here $\boldsymbol{\Psi}$ is assumed to be a diagonal matrix and estimated accordingly for better forecasting performance.

Figure 6 shows the estimation results for the $n_I = 149$ inner grid locations. Comparing maps of these adaptive fused Lasso estimators for January and July, we see distinct patterns of the coefficients. While our method is designed to identify spatially clustered patterns, in this example, it captures mainly smoothly varying coefficients. Overall, the autoregressive

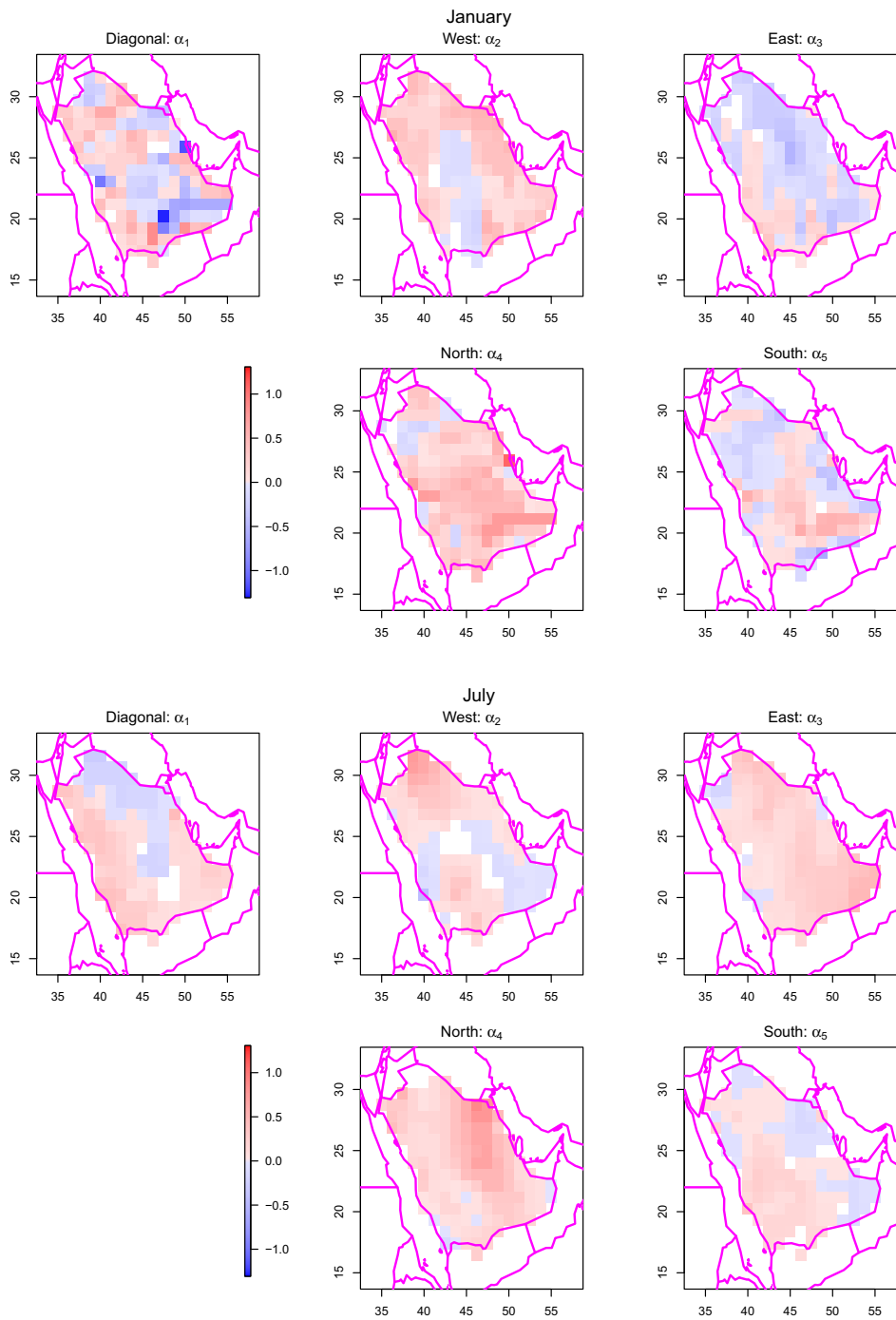

Figure 6. Maps of the estimated autoregressive coefficients using our algorithm (adaptive fused Lasso) with $K = 5$ for the Tukey g-and-h transformed residual daily wind speeds on January and July of 86 years (1920–2005) on a grid covering Saudi Arabia .

Table 2. Mean squared prediction error of h-step ahead forecasts with $h = 1, 2, 3$ for the 149 inner locations by different estimation methods in January and July

	h	Persistent	Unconstrained VAR(1)	Restricted OLS	Adaptive fused Lasso
Jan	1	1.1510	0.7608	**0.6951**	0.6971
	2	**1.3327**	1.6851	1.3483	1.3369
	3	1.5031	1.4495	1.3039	**1.2921**
July	1	0.7998	0.4852	0.3839	**0.3838**
	2	1.5367	1.0933	0.4934	**0.4932**
	3	2.8240	1.8587	1.0515	**1.0432**

Bold font indicates smallest values

coefficients in July have larger magnitudes and are smoother than those in January, which indicate stronger temporal dependencies. The difference can be seen eminently by comparing the coefficients for the northern neighbor between January and July. The stronger dependencies by the northern neighbors in July might be explained by the intensified northerly winds during the monsoon circulation in summer. In January, coefficients for the eastern and southern neighbors are negative in the northern part of Saudi Arabia, while in July the patterns are more homogeneous. We also notice that coefficients for coastal locations are not clustered together and exhibit complex patterns. These findings are insightful for explaining the spatiotemporal mechanisms of the wind speed data. With the estimated parameters, the model can be then used for building stochastic weather generators as an approximation of the computationally expensive climate model.

Although forecasting is not the main concern for this dataset, for completeness we present the forecast performance by different estimation methods in Table 2. Our adaptive fused Lasso outperforms the persistent predictor and the unconstrained VAR(1) estimator. Although occasionally slightly worse than the restricted OLS, the adaptive fused Lasso is more illuminating when visualizing the coefficients.

6. DISCUSSION

In this paper, we introduced VAR models for stationary time series on gridded locations with sparse and spatially coherent autoregressive coefficients to achieve dimension reduction. Our model is interpretable, flexible, and captures the spatial non-stationarity and essential dynamics of spatiotemporal processes. We detailed our model for the VAR(1) case, developed an adaptive fused Lasso estimation algorithm and derived its asymptotic properties. By a simulation study, we showed the satisfying performance of our estimation algorithm as well as the improvement in forecasting gained by the stable estimation. As illustrated by the data example, our model can be useful to find patterns in the estimated coefficients and thus shed light on the driving mechanisms of the phenomenon of interest over space and time. Furthermore, our model designed for gridded spatial data is particularly suitable for building stochastic weather generators.

For the boundary points, we assumed a simple AR(1) model in order to provide a dynamic mechanism on the boundary for forecasting. This assumption might seem over-simplified;

however, estimation of the dynamics for the inner grid points should not be affected much by this assumption as it is done conditional on the boundary variables. If several steps ahead forecast is not of interest, then there is no need to specify a model for the boundary points and the likelihood in (5) can be replaced by the likelihood conditioning on the boundary points. In this case the transition matrix is of size $n_I \times n$. Overall, the effect of edge sites specification is a thorny problem in spatial statistics and deserves further investigation.

Our model is suitable for data rich in space with n up to hundreds of locations, which is typically considered as high-dimensional for a VAR model. For the data example, with $n = 195$, our estimation algorithm using the 'genlasso' package took about 1 hour on a Dell desktop with a 3.30 GHz Intel CPU. However, with higher dimension, the estimation will take much longer time and become unstable. Zhao and Bondell (2020) proposed an algorithm to transform any generalized Lasso problem into a standard Lasso problem, which can be solved using any existing efficient Lasso solver. We implemented their algorithm and found that although faster for large n, the solution path is very unstable for λ small. Therefore, a more efficient and stable algorithm solving the generalized Lasso problem is much welcome to be developed and implemented.

Our model can be extended to various directions. Currently, we only apply the adaptive fused Lasso penalty to coefficients corresponding to each of the K neighbors separately in the penalized likelihood (6). We could further add an adaptive Lasso penalty across the K neighbors, e.g., coefficients corresponding to the left neighbor is expected to be similar to those corresponding to the bottom left and upper left neighbor. For the non-Gaussian spatiotemporal wind speed data, we estimated the transformation at each location separately. This method could be extended to estimate the parameters related to the transformation and $\boldsymbol{\alpha}$ for the underlying spatially structured Gaussian VAR process simultaneously; in this situation, a fused Lasso penalty would also be imposed on the transformation parameters. It is also possible to extend our method to irregularly spaced data by putting the data on a fine grid with missing observations. For uncertainty quantification, our simulation study and the functional boxplots can be seen as empirical measures of the uncertainty in the generalized Lasso estimates. Recently, Zhao and Bondell (2020) proposed the residual bootstrap method to construct confidence intervals for the generalized Lasso estimates. At last, observations with measurement error are rarely dealt with for regularized VAR models and could be an interesting topic for further research.

[Received May 2020. Revised January 2021. Accepted February 2021. Published Online February 2021.]

A APPENDIX

Derivation of (5):

$$l(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{t=2}^{T} \mathbf{Z}_{t-1}'\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\mathbf{Z}_{t-1} + \sum_{t=2}^{T} \mathbf{Z}_t'\boldsymbol{\Psi}^{-1}\mathbf{A}\mathbf{Z}_{t-1} + \text{constant}.$$

For the two terms involved in the above equation,

$$\sum_{t=2}^{T} \mathbf{Z}_t'\boldsymbol{\Psi}^{-1}\mathbf{A}\mathbf{Z}_{t-1} = \text{vec}\left(\boldsymbol{\Psi}^{-1}\sum_{t=2}^{T}\mathbf{Z}_t\mathbf{Z}_{t-1}'\right)'\text{vec}(\mathbf{A}) = \text{vec}\left(\boldsymbol{\Psi}^{-1}\sum_{t=2}^{T}\mathbf{Z}_t\mathbf{Z}_{t-1}'\right)'\mathbf{P}\boldsymbol{\alpha}, \text{ and}$$

$$\sum_{t=2}^{T} \mathbf{Z}_{t-1}'\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\mathbf{Z}_{t-1} = \text{vec}(\mathbf{A})'\left\{\left(\sum_{t=1}^{T-1}\mathbf{Z}_t\mathbf{Z}_t'\right)\otimes\boldsymbol{\Psi}^{-1}\right\}\text{vec}(\mathbf{A}) = \boldsymbol{\alpha}'\mathbf{P}'\left\{\left(\sum_{t=1}^{T-1}\mathbf{Z}_t\mathbf{Z}_t'\right)\otimes\boldsymbol{\Psi}^{-1}\right\}\mathbf{P}\boldsymbol{\alpha}.$$

Then, it is easy to see that the log-likelihood is a quadratic form of $\boldsymbol{\alpha}$, i.e.,

$$l(\boldsymbol{\alpha}) = -\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\right\|_2^2 + \text{constant},$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ satisfy (3) and (4). □

Asymptotic Properties:

Some notations are needed to state the asymptotic properties. We consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, m\}$ is a set of vertices corresponding to the $m = Kn_I + n_B$ parameters in $\boldsymbol{\alpha}$ to be estimated, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of undirected edges that depends on the structure of the spatial grid. Here $\mathcal{V}$ can be partitioned into $n_B + K$ disjoint components, including n_B isolated nodes (corresponding to the non-regularized coefficients in $\boldsymbol{\alpha}_{\mathrm{B}}$) and K identical disjoint components each with n_I vertices and the same inner grid structure (corresponding to the penalized coefficients in $\boldsymbol{\alpha}_k$, for $k = 1, \ldots, K$). With these notations, the penalty term in (6) can be rewritten as

$$\lambda\sum_{k=1}^{K}\sum_{i\sim j} w_{k,i,j}|\alpha_k(\mathbf{s}_i) - \alpha_k(\mathbf{s}_j)| = \lambda\sum_{(i,j)\in\mathcal{E}} w_{i,j}|\alpha_i - \alpha_j|,$$

where $w_{i,j} = |\tilde{\alpha}_i - \tilde{\alpha}_j|^{-\gamma}$.

We let $\boldsymbol{\alpha}^0$ be the true parameter vector and define $\boldsymbol{\xi} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}^0$. We introduce $\mathcal{A} = \left\{(i, j) \in \mathcal{E} : \alpha_i^0 = \alpha_j^0, i, j = 1, \ldots, n_I\right\}$ and consider the subgraph $\mathcal{G}_{\mathcal{A}} = (\mathcal{V}, \mathcal{A})$ of $\mathcal{G}$. We denote by m_0 the number of its connected components, that is, the number of distinct values in $\boldsymbol{\alpha}^0$ supported by $\mathcal{G}$. We further denote by $\mathcal{V}_1, \ldots, \mathcal{V}_{m_0}$ the sets of nodes of each connected components of $\mathcal{G}_{\mathcal{A}}$ and set $l_i = \min \mathcal{V}_i$ for $i = 1, \ldots, m_0$. We define $\boldsymbol{\alpha}_{\mathcal{A}}^0 = (\alpha_{l_1}^0, \ldots, \alpha_{l_{m_0}}^0)'$ and $\boldsymbol{X}_{\mathcal{A}}$ a matrix whose i-th column, for $i = 1, \ldots, m_0$, is $\boldsymbol{X}_{\mathcal{A}_i} = \sum_{j\in\mathcal{V}_i}\boldsymbol{X}_j$, where $\boldsymbol{X}_j$ is the j-th column of $\boldsymbol{X}$. We assume that $\boldsymbol{X}'\boldsymbol{X}/(T-1) \to \mathbf{C}$ as $T \to \infty$ for some positive-definite $m \times m$ matrix $\mathbf{C}$, which depends on $\boldsymbol{\alpha}^0$ and $\boldsymbol{\Psi}$. We denote by $\mathbf{C}_{\mathcal{A}}$ the limiting $m_0 \times m_0$ matrix of $\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{X}_{\mathcal{A}}/(T-1)$ as $T \to \infty$.

We derive the asymptotic properties of the adaptive Lasso estimator $\hat{\boldsymbol{\alpha}}$ from solving (7). We let $\mathcal{A}_n = \left\{(i, j) \in \mathcal{E} : \hat{\alpha}_i = \hat{\alpha}_j, i, j = 1, \ldots, n_I\right\}$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{A}} = (\hat{\alpha}_{l_1}, \ldots, \hat{\alpha}_{l_{m_0}})'$.

Theorem 1. *Suppose that* $\lambda/\sqrt{T} \to 0$, $\lambda T^{(\gamma-1)/2} \to \infty$, *and* $\mathbf{C} = \lim_{\mathbf{T}\to\infty} \boldsymbol{X}'\boldsymbol{X}/(T-1)$ *is positive-definite. Then as* $T \to \infty$, *the following are satisfied by the adaptive Lasso estimator* $\hat{\boldsymbol{\alpha}}$*:*

1. *Consistency in grouping:* $\Pr(\mathcal{A}_n = \mathcal{A}) \to 1$.
2. *Asymptotic normality:* $\sqrt{T}\left(\hat{\boldsymbol{\alpha}}_{\mathcal{A}} - \boldsymbol{\alpha}^0_{\mathcal{A}}\right) \xrightarrow{d} \mathcal{N}_{m_0}(\mathbf{0}, \mathbf{C}^{-1}_{\mathcal{A}})$.

Proof. The following proof is obtained by adapting the Proof of Theorem 2 in Zou (2006) and Theorem 3 in Viallon et al. (2013).

First, it is known that $\hat{\boldsymbol{\Psi}}$ in (8) is a consistent estimator of $\boldsymbol{\Psi}$, as T goes to infinity; see Chapter 5 of Lütkepohl (2005). We define $V_T(\mathbf{u}) = F(\boldsymbol{\alpha}^0) - F(\boldsymbol{\alpha}^0 + \mathbf{u}/\sqrt{T})$, with F defined in (6). It is obvious that $V_T(\mathbf{u})$ is minimized at $\sqrt{T}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)$ and

$$V_T(\mathbf{u}) = \mathbf{u}'\left(\frac{1}{2T}\boldsymbol{X}'\boldsymbol{X}\right)\mathbf{u} - \frac{\boldsymbol{\xi}'\boldsymbol{X}}{\sqrt{T}}\mathbf{u} + \frac{\lambda}{\sqrt{T}}\sum_{(i,j)\in\mathcal{A}} w_{i,j}\sqrt{T}\left(\left|\alpha^0_i - \alpha^0_j + \frac{u_i - u_j}{\sqrt{T}}\right| - |\alpha^0_i - \alpha^0_j|\right).$$

We have

$$\lambda\, w_{i,j}\left|\alpha^0_i - \alpha^0_j + \frac{u_i - u_j}{\sqrt{T}}\right| - |\alpha^0_i - \alpha^0_j| \xrightarrow{p} \begin{cases} 0, & \text{if } \alpha^0_i \neq \alpha^0_j \text{ or } (\alpha^0_i = \alpha^0_j \text{ and } u_i = u_j), \\ \infty, & \text{otherwise.} \end{cases}$$

We denote $\mathbf{u}_{\mathcal{A}} = (\mathbf{u}_{l_1}, \ldots, \mathbf{u}_{l_{m_0}})'$ and then, with an application of the Martingale difference central limit theory to $\boldsymbol{\xi}'\boldsymbol{X}$, we obtain

$$V_T(\mathbf{u}) \xrightarrow{d} V(\mathbf{u}) = \begin{cases} \frac{1}{2}\mathbf{u}'_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} - \mathbf{u}'_{\mathcal{A}}\mathbf{W}_{\mathcal{A}}, & \text{if } u_i = u_j \text{ for } (i, j) \in \mathcal{A}, \\ \infty, & \text{otherwise,} \end{cases}$$

for $\mathbf{u} \in \mathbb{R}^m$, where $\mathbf{W}_{\mathcal{A}} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{C}_{\mathcal{A}})$; V is convex and has a unique minimum satisfying $u_i = u_j$ for all $(i, j) \in \mathcal{A}$ and $\mathbf{u}_{\mathcal{A}} = \mathbf{C}^{-1}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}}$. The asymptotic normality part can be derived as in Zou (2006) by using the epi-convergence results.

For consistency in grouping, we need to show that for all $(i, j) \notin \mathcal{A}$, $\Pr((i, j) \in \mathcal{A}^c_n) \to 1$ and for all $(i, j) \in \mathcal{A}$, $\Pr((i, j) \in \mathcal{A}^c_n) \to 0$. The first part is implied by the asymptotic normality result. To prove the second part, we apply the subgradient equations for the optimality condition, for $i = 1, \ldots, m$:

$$\boldsymbol{X}'_i(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}) = \lambda \sum_{i:(i,j)\in\mathcal{E}} w_{i,j} t_{ij},$$

where $t_{ij} = \operatorname{sign}(\hat{\alpha}_i - \hat{\alpha}_j)$ if $\hat{\alpha}_i \neq \hat{\alpha}_j$ and t_{ij} is some real number in $[-1, 1]$ if $\hat{\alpha}_i = \hat{\alpha}_j$. We prove by contradiction. Suppose that for $\mathcal{V}_k$ that contains at least two vertices, there exist $i, j \in \mathcal{V}_k$ such that $\hat{\alpha}_i \neq \hat{\alpha}_j$. We define $a^{\min} = \min_{i\in\mathcal{V}_k} \hat{\alpha}_i$ and

$\mathcal{V}^{\min} = \left\{ i : i \in \mathcal{V}_k \text{ and } \hat{\alpha}_i = a^{\min} \right\}$. Summing up the optimality conditions over the indices in $\mathcal{V}^{\min}$, we get:

$$\sum_{i \in \mathcal{V}^{\min}} \frac{\boldsymbol{X}_i'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}})}{\sqrt{T}} = \frac{\lambda}{\sqrt{T}} \sum_{i \in \mathcal{V}^{\min}} \sum_{\substack{i:(i,j)\in\mathcal{E} \\ \alpha_i^0 \neq \alpha_j^0}} \frac{t_{ij}}{|\tilde{\alpha}_i - \tilde{\alpha}_j|^{\gamma}} + \lambda T^{(\gamma-1)/2} \sum_{i \in \mathcal{V}^{\min}} \sum_{\substack{i:(i,j)\in\mathcal{E} \\ \alpha_i^0 = \alpha_j^0 \\ \hat{\alpha}_j > a^{\min}}} \frac{t_{ij}}{|\sqrt{T}(\tilde{\alpha}_i - \tilde{\alpha}_j)|^{\gamma}},$$

where in the right-hand side, the first sum converges to 0 in probability, while the second sum tends to $-\infty$. However, the left-hand side is $O_p(1)$, since it can be decomposed as the sum of two asymptotically normal variables as in Zou (2006) with an application of the martingale central limit theorem. Therefore, $\Pr((i, j) \in \mathcal{A}_n^c) \to 0$. □

REFERENCES

Ailliot P, Monbet V, Prevosto M (2006) An autoregressive model with time-varying coefficients for wind fields. Environmetrics 17(2):107–117

Arnold TB, Tibshirani RJ (2020) genlasso: path algorithm for generalized Lasso problems. R Package Vers 1:5

Bańbura M, Giannone D, Reichlin L (2010) Large Bayesian vector auto regressions. J Appl Econ 25(1):71–92

Basu S, Michailidis G (2015) Regularized estimation in sparse high-dimensional time series models. Ann Stat 43(4):1535–1567

Bergmeir C, Benítez JM (2012) On the use of cross-validation for time series predictor evaluation. Inf Sci 191:192–213

Bessac J, Ailliot P, Monbet V (2015) Gaussian linear state-space model for wind fields in the North-East Atlantic. Environmetrics 26(1):29–38

Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, Hoboken

de Luna X, Genton MG (2005) Predictive spatio-temporal models for spatially sparse environmental data. Statistica Sinica 15:547–568

Fan J, Zhang W (1999) Statistical estimation in varying coefficient models. Ann Stat 27(5):1491–1518

Gelfand AE, Kim H-J, Sirmans CF, Banerjee S (2003) Spatial modeling with spatially varying coefficient processes. J Am Stat Assoc 98(462):387–396

Genton M, Johnson C, Potter K, Stenchikov G, Sun Y (2014) Surface boxplots. Stat 3(1):1–11

Hsu N-J, Hung H-L, Chang Y-M (2008) Subset selection for vector autoregressive processes using Lasso. Comput Stat Data Anal 52(7):3645–3657

Huang H-C, Hsu N-J, Theobald DM, Breidt FJ (2010) Spatial Lasso with applications to GIS model selection. J Comput Gr Stat 19(4):963–983

Kastner G, Huber F (2020) Sparse Bayesian vector autoregressions in huge dimensions. J Forecast 39:1142–1165

Katzfuss M, Cressie N (2012) Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics 23(1):94–107

Kay JE, Deser C, Phillips A, Mai A, Hannay C, Strand G, Arblaster JM, Bates SC, Danabasoglu G, Edwards J, Holland M, Kushner P, Lamarque J-F, Lawrence D, Lindsay K, Middleton A, Munoz E, Neale R, Oleson K, Polvani L, Vertenstein M (2015) The community earth system model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. Bull Am Meteorol Soc 96(8):1333–1349

Ke ZT, Fan J, Wu Y (2015) Homogeneity pursuit. J Am Stat Assoc 110(509):175–194

Korobilis D, Pettenuzzo D (2019) Adaptive hierarchical priors for high-dimensional vector autoregressions. J Econ 212(1):241–271

Li F, Sang H (2019) Spatial homogeneity pursuit of regression coefficients for large datasets. J Am Stat Assoc 114(527):1050–1062

Lütkepohl H (2005) New introduction to multiple time series analysis. Springer, Berlin

Monbet V, Ailliot P (2017) Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature. Comput Stat Data Anal 108:40–51

Mu J, Wang G, Wang L (2018) Estimation and inference in spatially varying coefficient models. Environmetrics 29(1):e2485

Ngueyep R, Serban N (2015) Large-vector autoregression for multilayer spatially correlated time series. Technometrics 57(2):207–216

R Development Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna

Rao SS (2008) Statistical analysis of a spatio-temporal model with location-dependent parameters and a test for spatial stationarity. J Time Ser Anal 29(4):673–694

Schweinberger M, Babkin S, Ensor KB (2017) High-dimensional multivariate time series with additional structure. J Comput Gr Stat 26(3):610–622

Shen X, Huang H-C (2010) Grouping pursuit through a regularization solution surface. J Am Stat Assoc 105(490):727–739

Sun Y, Genton MG (2011) Functional boxplots. J Comput Gr Stat 20(2):316–334

Sun Y, Wang HJ, Fuentes M (2016) Fused adaptive Lasso for spatial and temporal quantile function estimation. Technometrics 58(1):127–137

Tagle F, Castruccio S, Crippa P, Genton MG (2019) A non-Gaussian spatio-temporal model for daily wind speeds based on a multivariate skew-t distribution. J Time Ser Anal 40:312–326

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J Roy Stat Soc B 67(1):91–108

Tibshirani RJ, Taylor J (2011) The solution path of the generalized lasso. Ann Stat 39(3):1335–1371

Viallon V, Lambert-Lacroix S, Höfling H, Picard F (2013) Adaptive generalized fused-Lasso: asymptotic properties and applications. In: HAL preprint, hal-00813281

Wikle CK, Berliner LM, Cressie N (1998) Hierarchical Bayesian space-time models. Environ Ecol Stat 5(2):117–154

Wikle CK, Milliff RF, Nychka D, Berliner LM (2001) Spatiotemporal hierarchical bayesian modeling: tropical ocean surface winds. J Am Stat Assoc 96(454):382–397

Yan Y, Genton MG (2019) Non-Gaussian autoregressive processes with Tukey g-and-h transformations. Environmetrics 30:e2503

Zhao Y, Bondell H (2020) Solution paths for the generalized lasso with applications to spatially varying coefficients regression. Comput Stat Data Anal 142:106821

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.