High-Dimensional Multivariate Time Series With Local Dependence

Michael Schweinberger Sergii Babkin Katherine B. Ensor

Abstract

We consider high-dimensional multivariate time series with additional structure. The additional structure takes the form of a metric space endowed with local dependence, i.e., time series that are close in space may be dependent whereas distant time series are independent. Such additional structure is available in a wide range of applications, e.g., in studies of air pollution and climate change. We introduce a simple two-step estimation approach that takes advantage of local dependence. The two-step estimation approach first estimates the range of dependence and then exploits the estimated range of dependence to estimate local dependencies among time series. We shed light on the theoretical properties of the two-step estimation approach under high-dimensional scaling and provide non-asymptotic error bounds that hold with high probability. The usefulness of the two-step estimation approach is demonstrated by an application to air pollution in the U.S. The two-step estimation approach can be extended to other high-dimensional models, such as high-dimensional graphical models, as long as additional structure is available and consistent model selection in high dimensions is possible.

Keywords: multivariate time series, vector-autoregressive processes, graphical models, local dependence, spatial dependence.

1 Introduction

Multivariate time series (e.g., Lütkepohl, 2007; Wilson et al., 2015) arise in a wide range of applications, from economics and genomics to studies of air pollution and climate change (e.g., Ensor et al., 2013; Hoek et al., 2013; Chen et al., 2015). The age of computing has made it possible to collect data sets with large numbers of time series, where the number of predictors may exceed the number of observations. A common approach to dealing with high-dimensional data is to endow models with additional structure in the form of sparsity (e.g., Bühlmann and van de Geer, 2011). In the case of high-dimensional multivariate time series, an additional challenge is the complex dependence within and between time series. Loh and Wainwright (2012) considered high-dimensional dependent data, including high-dimensional vector autoregressive (VAR) processes, and studied the theoretical properties of ℓ_1-penalized M-estimators in high-dimensional settings. Chudik and Pesaran (2011), Song and Bickel (2011), and Basu and Michailidis (2015) studied high-dimensional VAR processes in more depth. In particular, Basu and Michailidis (2015) established consistency of ℓ_1-penalized least squares (LS) and maximum likelihood (ML) estimators in high-dimensional settings and related the estimation and prediction error to the complex dependence structure of VAR processes. Some other estimation approaches are discussed by, e.g., Davis et al. (2012) and Nguyen et al. (2014), but the theoretical properties of most of these estimators are unknown.

We consider high-dimensional multivariate time series with additional structure. In most applications of multivariate time series, there is additional structure: e.g., time series may have spatial locations and dependence may be local in the sense that time series that are close in space may be strongly dependent whereas distant time series are weakly dependent or independent. An example is air pollution. Air pollution in metropolitan areas in the U.S. may influence air pollution in surrounding areas, but air pollution in metropolitan areas on the East Coast does not influence air pollution in metropolitan areas on the West Coast. Indeed, in many areas in which dependent data arise, such as physics (e.g., Ising, 1925), spatial statistics (e.g., Besag, 1974), and random graphs (e.g., Schweinberger and Handcock, 2015), dependence is local in the sense that units that are close in a well-defined sense may be strongly dependent whereas distant units are weakly dependent or independent.

We take advantage of additional structure in the form of a metric space endowed with local dependence to reduce computing time along with statistical error. The proposed two-step estimation approach first estimates the range of dependence and then exploits the estimated range of dependence to estimate local dependencies among time series. If the range of dependence is short and the number of time series is large, the two-step estimation approach reduces computing time and the statistical error of the parameters of primary interest, i.e., the parameters governing short-distance edges. We establish the theoretical properties of the two-step estimation approach under high-dimensional scaling and provide non-asymptotic error bounds that hold with high probability. The two-step estimation approach can be extended to other high-dimensional models, such as high-dimensional graphical models (e.g., Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010), as long as additional structure is available and consistent model selection in high dimensions is possible.

The paper is structured as follows. We introduce VAR processes in Section 2 and local dependence in Section 3. A simple two-step estimation approach is introduced in Section 4. We establish the theoretical properties of the two-step estimation approach in Section 5 and present simulation results in Section 6. An application to air pollution in the U.S. is presented in Section 7.

2 Model

We assume that $\boldsymbol{X}(t) = (X_1(t), \dots, X_k(t))_{t \in \mathbb{Z}}$ is generated by a k-dimensional VAR(L) process of the form

$$\boldsymbol{X}(t) \;\; = \;\; \sum_{l=1}^{L} \boldsymbol{A}_l \, \boldsymbol{X}(t-l) + \boldsymbol{e}(t), \tag{1}$$

where $\boldsymbol{A}_1, \dots, \boldsymbol{A}_L$ are $k \times k$ transition matrices and $\boldsymbol{e}(t) \overset{\text{iid}}{\sim} \text{MVN}_k(\boldsymbol{0}_k, \boldsymbol{\Sigma})$ is k-variate Gaussian with mean $\boldsymbol{0}_k$ and positive-definite variance-covariance matrix $\boldsymbol{\Sigma}$. We assume that the order L of the VAR process is known and that the VAR process is stable and thus stationary (Lütkepohl, 2007).

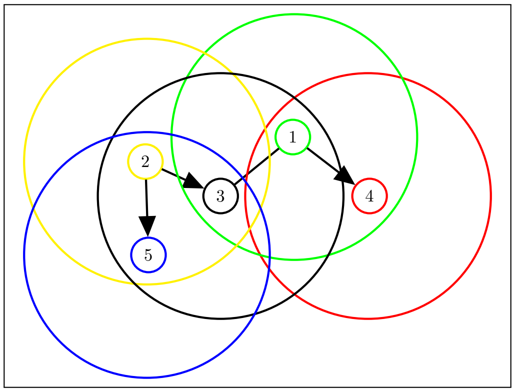
A convenient representation of the dependence structure of VAR processes is given by graphs (e.g., Eichler, 2012). A VAR process can be represented by a mixed graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ with a set of nodes $\mathcal{N} = \{1, \dots, k\}$ and a set of directed and undirected edges $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ representing past-present dependencies (directed edges) and present-present dependencies (undirected edges) between the nodes corresponding to the components of the VAR process. The mixed graph is determined as follows:

$$\begin{aligned} (i \longrightarrow j) \;\; &\notin \;\; \mathcal{E} \;\; \Longleftrightarrow \;\; A_{1,i,j} = \cdots = A_{L,i,j} = 0 \\ (i \text{ —— } j) \;\; &\notin \;\; \mathcal{E} \;\; \Longleftrightarrow \;\; \Sigma_{i,j}^{-1} = 0, \end{aligned} \tag{2}$$

where $A_{h,i,j}$ denotes element (i, j) of matrix $\boldsymbol{A}_h$ and $\Sigma_{i,j}^{-1}$ denotes element (i, j) of matrix $\boldsymbol{\Sigma}^{-1}$. The Markov properties of mixed graphs representing multivariate time series are discussed by Eichler (2012). An example of a graphical representation of a VAR(1) process with 5 components is shown in Figure 1, where the nodes represent the components of the VAR process and the lines indicate dependencies between components.

Figure 1: Local dependence in VAR(1) processes: nodes represent components of the VAR process and edges represent dependencies. The range of dependence is restricted to the closed balls with radius ρ centered at the positions of the components. The elements $\star$ of matrices indicate non-zero elements.

$$\boldsymbol{A}_1 = \begin{pmatrix} \star & 0 & 0 & \star & 0 \\ 0 & \star & \star & 0 & \star \\ 0 & 0 & \star & 0 & 0 \\ 0 & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & \star \end{pmatrix} \quad \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \star & 0 & \star & 0 & 0 \\ 0 & \star & 0 & 0 & 0 \\ \star & 0 & \star & 0 & 0 \\ 0 & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & \star \end{pmatrix}$$



3 Local dependence

We consider multivariate time series with additional structure in the form of a metric space endowed with local dependence. These assumptions can be stated as follows.

(A.1) Metric space. *The components of the multivariate time series have positions in a compact metric space* $(\mathbb{A}, d)$.

In most applications, $\mathbb{A}$ is a compact subset of $\mathbb{R}^2$ or $\mathbb{R}^3$. Examples are air pollution monitors on two-dimensional maps of the U.S. and brain cells on three-dimensional maps of the human brain.

(A.2) Local dependence. *The dependence between the components of the multivariate time series is local in the sense that there exists* $0 < \rho < \infty$ *such that there are no edges between pairs of components* $(i, j) \in \mathcal{N} \times \mathcal{N}$ *at distances* $d(i, j) > \rho$.

An example is shown in Figure 1. The dependence between the components of the

Table 1: Two-step estimation approach.

```
1. Estimate the range of dependence $\rho$:
    1.1 Sample a subset of nodes $\mathcal{S}$ from the set of nodes $\mathcal{N}$.
    1.2 Estimate edges by regressing nodes $i \in \mathcal{S}$ on $\{j \mid j \in \mathcal{N}\}$.
    1.3 Estimate the range of dependence $\rho$ by $\widehat{\rho}$ , the maximum distance that
        separates a pair of nodes with an estimated edge.
2. Estimate the graph by regressing nodes $i \in \mathcal{N}$ on $\{j \mid j \in \mathcal{N},\, d(i,j) \leq \widehat{\rho}\,\}$.
```

VAR process is local in the sense that the dependence of the components is restricted to the closed balls with radius ρ centered at the positions of the components.

4 Two-step estimation approach

We introduce a simple two-step estimation approach that takes advantage of the additional structure considered here.

The two-step estimation approach is sketched in Table 1. It can be motivated as follows. Suppose that we want to estimate the unknown edges in the graph, i.e., the dependence structure of the VAR process. If there is additional structure in the sense that there are no edges at distances $d > \rho$, ignoring the additional structure may generate false-positive edges at large distances, which contradicts scientific knowledge and may lead consumers of statistical results to question the usefulness of statistics. A striking example is presented in Section 7. By restricting the estimation of edges to distances $d \leq \rho$, the local nature of multivariate time series is respected and computing time along with statistical error can be reduced.

In practice, the problem is that the range of dependence ρ is unknown. If the structure of the graph was known, one could take ρ to be the maximum distance that separates a pair of nodes with an edge. If the structure of the graph is unknown, one needs to estimate the graph. An appealing alternative to estimating the whole graph—which is time-consuming when the set of nodes $\mathcal{N}$ is large—is to estimate a subgraph by sampling a subset of nodes $\mathcal{S}$, estimating the edges of nodes $i \in \mathcal{S}$, and then estimating ρ by $\widehat{\rho}$, defined as the maximum distance that separates a pair of nodes with an estimated edge. Step 1 estimates the range of dependence ρ by $\widehat{\rho}$ along these lines. Step 2 estimates the graph by restricting the estimation of edges to distances $d \leq \widehat{\rho}$. If the sample in Step 1 is small but well-chosen and the range of dependence ρ is short, the two-step estimation approach can reduce computing time along with statistical error.

We discuss the implementation of the two-step estimation approach in Sections 4.1 and 4.2, shed light on its theoretical properties in Section 5, and demonstrate its usefulness by simulation results and an application in Sections 6 and 7, respectively. Throughout, we assume that $\boldsymbol{\Sigma}$ is diagonal, so that all edges are directed; extensions to undirected edges are possible, though less attractive on computational grounds, as explained in Section 4.2. We denote by $\|.\|_1$, $\|.\|_2$, and $\|.\|_\infty$ the ℓ_1, ℓ_2, and ℓ_∞-norm of vectors, respectively. The total number of observations is denoted by M and the effective number of observations by $N = M - L + 1$.

4.1 Step 1

In Step 1.1, a sample of nodes $\mathcal{S}$ from the set of nodes $\mathcal{N}$ is generated. In Step 1.2, edges are estimated by regressing nodes $i \in \mathcal{S}$ on $\{j \mid j \in \mathcal{N}\}$ by using ℓ_1-penalized LS, which is attractive on both computational and theoretical grounds (Basu and Michailidis, 2015),

To introduce the ℓ_1-penalized LS approach used in Step 1.2, note that the conventional ℓ_1-penalized LS approach estimates the $p = k^2\, L$-dimensional parameter vector $\boldsymbol{\beta}_{\mathcal{N}} = (\boldsymbol{\beta}_i)_{i \in \mathcal{N}}$ corresponding to the vectorized transition matrices $\text{vec}(\boldsymbol{A}_1^\top, \dots, \boldsymbol{A}_L^\top)$ by

$$\widehat{\boldsymbol{\beta}}_{\mathcal{N}} \;\; \in \;\; \underset{\boldsymbol{\beta}_i,\, i \in \mathcal{N}}{\arg\min} \sum_{i \in \mathcal{N}} \left[\frac{1}{N} \|\boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}\, \boldsymbol{\beta}_i\|_2^2 + \lambda_1\, \|\boldsymbol{\beta}_i\|_1 \right], \tag{3}$$

where $\boldsymbol{\beta}_i$ denotes the $p_i = k\, L$-dimensional parameter vectors governing possible incoming edges of nodes i; $\boldsymbol{\mathcal{Y}}_i$ denotes the i-th column of the matrix of observations $\boldsymbol{\mathcal{Y}} = (\boldsymbol{X}(M)^\top, \dots, \boldsymbol{X}(L)^\top)$; $\boldsymbol{\mathcal{X}}$ denotes the predictors $((\boldsymbol{X}(M-1)^\top, \dots, \boldsymbol{X}(L-1)^\top)$, $\dots$, $(\boldsymbol{X}(M-L)^\top, \dots, \boldsymbol{X}(0)^\top))$; and $\lambda_1 > 0$ denotes a regularization parameter.

The ℓ_1-penalized LS approach used in Step 1.2 applies the same procedure to the subset of nodes $\mathcal{S}$ and estimates the parameter vector $\boldsymbol{\beta}_{\mathcal{S}} = (\boldsymbol{\beta}_i)_{i \in \mathcal{S}}$ by

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}} \;\; \in \;\; \underset{\boldsymbol{\beta}_i,\, i \in \mathcal{S}}{\arg\min} \sum_{i \in \mathcal{S}} \left[\frac{1}{N} \|\boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}\, \boldsymbol{\beta}_i\|_2^2 + \lambda_1\, \|\boldsymbol{\beta}_i\|_1 \right]. \tag{4}$$

The incoming edges of nodes $i \in \mathcal{S}$ can be inferred from the sparsity pattern of $\widehat{\boldsymbol{\beta}}_{\mathcal{S}} =$

$(\widehat{\boldsymbol{\beta}}_i)_{i\in\mathcal{S}}$ by using (2). The range of dependence ρ can be estimated by $\widehat{\rho}$, the maximum distance that separates a pair of nodes $(j,i) \in \mathcal{N} \times \mathcal{S}$ with an estimated edge.

4.2 Step 2

In Step 2, given the estimator $\widehat{\rho}$ of the range of dependence ρ, nodes $i \in \mathcal{N}$ are regressed on $\{j \mid j \in \mathcal{N},\, d(i,j) \leq \widehat{\rho}\,\}$, i.e., the parameter vector $\boldsymbol{\beta}$ is estimated by

$$\widehat{\boldsymbol{\beta}} \;\in\; \underset{\boldsymbol{\beta}_i,\, i\in\mathcal{N}}{\arg\min} \sum_{i\in\mathcal{N}} \left[\frac{1}{N}\|\boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}\,\boldsymbol{\beta}_i\|_2^2 + \lambda_2\,\|\boldsymbol{\beta}_i\|_1\right] \tag{5}$$

subject to the constraint that all parameters governing possible edges at distances $d > \widehat{\rho}$ are 0, where $\lambda_2 > 0$ is a regularization parameter.

Remark 1. An important observation is that the parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are variation-independent in the sense that the parameter space of $\boldsymbol{\beta}$ is a product space of the form $\mathbb{R}^{k^2L} = \mathbb{R}^{kL} \times \cdots \times \mathbb{R}^{k\,L}$. As a result, optimization problems (3), (4), and (5) can be decomposed into k separate optimization problems that can be solved in parallel, thus reducing computing time.

Remark 2. The variance-covariance matrix $\boldsymbol{\Sigma}$ can be estimated by using the ℓ_1-penalized ML approach of Basu and Michailidis (2015). However, the ℓ_1-penalized ML approach is more time-consuming than the ℓ_1-penalized LS approach, because it cannot be executed in parallel—in contrast to the ℓ_1-penalized LS approach. We therefore focus on diagonal variance-covariance matrices $\boldsymbol{\Sigma}$, though both the methods and theoretical results discussed here can be extended to non-diagonal $\boldsymbol{\Sigma}$.

5 Theoretical properties

We shed light on the theoretical properties of the two-step estimation approach.

To facilitate the discussion, we follow Loh and Wainwright (2012) and Basu and Michailidis (2015) by expressing optimization problems (3), (4), and (5) as M-estimation problems of the form

$$\widehat{\boldsymbol{\beta}} \;\in\; \underset{\boldsymbol{\beta}\,\in\,\mathcal{C}}{\arg\min} \left[-2\,\boldsymbol{\beta}^\top\widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^\top\,\widehat{\boldsymbol{\Gamma}}\,\boldsymbol{\beta} + \lambda\,\|\boldsymbol{\beta}\|_1\right], \tag{6}$$

where $\mathcal{C}$ is a subset of $\mathbb{R}^p$ that depends on the constraints imposed by optimization problems (3), (4), and (5), $\widehat{\boldsymbol{\gamma}} = (\boldsymbol{I} \otimes \boldsymbol{\mathcal{X}}^\top)\,\mathrm{vec}(\boldsymbol{\mathcal{Y}})/N$, and $\widehat{\boldsymbol{\Gamma}} = (\boldsymbol{I} \otimes \boldsymbol{\mathcal{X}}^\top\boldsymbol{\mathcal{X}})/N$, where $\boldsymbol{I}$ denotes the identity matrix of suitable order and $\otimes$ denotes the Kronecker product.

Notation. Throughout, we assume that the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are ordered according to distance and denote by $\boldsymbol{\beta}_{[d_1,d_2]}$ and $\boldsymbol{\gamma}_{[d_1,d_2]}$ the subvectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ corresponding to parameters governing possible edges at distances $d \in [d_1, d_2]$, respectively. The rows and columns of $\boldsymbol{\Gamma}$ are ordered in accordance. Denote by $p(d_1, d_2)$ the total number of parameters governing possible edges at distances $d \in [0, d_2]$ if $d_1 = 0$ and the total number of parameters governing possible edges at distances $d \in (d_1, d_2]$ otherwise. Let $\delta > 0$ and denote by $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}, \mathbf{0}_{p(\rho-\delta,\rho)}, \mathbf{0}_{p-p(0,\rho-\delta)-p(\rho-\delta,\rho)})$ the estimator of the true parameter vector $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}^\star_{[0,\rho-\delta]}, \boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}, \mathbf{0}_{p-p(0,\rho-\delta)-p(\rho-\delta,\rho)})$ obtained by the two-step estimation approach, where $\boldsymbol{\beta}^\star_{[0,\rho-\delta]}$ denotes the $p(0, \rho - \delta)$-vector of parameters governing possible edges at distances $d \in [0, \rho - \delta]$, $\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}$ denotes the $p(\rho - \delta, \rho)$-vector of parameters governing possible edges at distances $d \in (\rho - \delta, \rho]$, and $\mathbf{0}_{p-p(0,\rho-\delta)-p(\rho-\delta,\rho)}$ denotes a $p - p(0, \rho - \delta) - p(\rho - \delta, \rho)$-vector of 0's. Denote by $\mathbb{S}$ the support of $\boldsymbol{\beta}^\star$ and by $\widehat{\mathbb{S}}$ the estimated support and let s be the size of support $\mathbb{S}$. Assume that nodes i are sampled independently with probabilities $0 < \theta_i < 1$ and let $\mathcal{S}(\delta)$ be the subset of nodes with incoming edges at distances $d \in [\rho - \delta, \rho]$. Denote by $c_0, c_1, c_2 > 0$ constants.

We assume that the following conditions hold. The first two conditions are conventional and hold with high probability (Loh and Wainwright, 2012; Basu and Michailidis, 2015), whereas the third condition is deterministic.

(C.1) Restricted eigenvalue condition. $\widehat{\boldsymbol{\Gamma}}$ *satisfies the restricted eigenvalue condition with curvature* $\alpha > 0$ *and tolerance* $\tau > 0$ *provided* $s\,\tau \leq \alpha/32$ *and*

$$\boldsymbol{b}^\top \widehat{\boldsymbol{\Gamma}}\, \boldsymbol{b} \;\; \geq \;\; \alpha\, \|\boldsymbol{b}\|_2^2 - \tau\, \|\boldsymbol{b}\|_1^2 \quad \textit{for all} \quad \boldsymbol{b} \in \mathbb{R}^p. \tag{7}$$

(C.2) Deviation condition. *There exists a deterministic function* $\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) > 0$ *such that* $\widehat{\boldsymbol{\gamma}}$ *and* $\widehat{\boldsymbol{\Gamma}}$ *satisfy*

$$\|\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\, \boldsymbol{\beta}^\star\|_\infty \;\; \leq \;\; \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) \sqrt{\frac{\log p}{N}}. \tag{8}$$

(C.3) Boundary condition. *For all* $\delta > 0$ *and* $\epsilon > 0$, *there exists a constant* $k_{\delta,\epsilon} > 0$ *such that, for all* $k > k_{\delta,\epsilon}$, *the fraction of edges at distances* $d \in [\rho - \delta, \rho]$ *is at most* ϵ.

Condition (C.3) is needed to ensure that in the event $\widehat{\rho} < \rho$ most edges are recovered with high probability. It states that the fraction of edges close to the boundary of the closed balls centered at the positions of nodes is small provided that the number of time series k is large. If the condition were violated, the two-step estimation approach would not be able to recover most edges—regardless of how close the estimated range of dependence is to the true range of dependence ρ, because a large fraction of edges could be arbitrarily close to ρ. An example of a graph violating the condition would be a bicycle-wheel-without-rim graph with a hub and spokes but without a rim, i.e., a graph with long-distance edges between hub and nodes but without short-distance edges between neighboring nodes. The fact that the boundary condition rules out such graphs does not restrict the range of applications much, because such graphs are unrealistic.

In practice, the goal of the two-step estimation approach is to recover all edges and parameters while taking advantage of local dependence. The following result sheds light on the conditions under which the two-step estimation approach achieves its goals.

Theorem 1. Consider $N \geq c_0\, s \log p$ ($c_0 > 1$) observations from a stable VAR(L) process with range of dependence $\rho = \max_{(i,j) \in \mathcal{E}} d(i,j) > 0$ and minimum signal strength $\beta^\star_{\min} = \min_{i \in \mathbb{S}} |\beta^\star_i| \geq 32\sqrt{s}\,\lambda / \alpha > 0$.

1.1 Assume that λ_1 satisfies

$$\lambda_1 \;\;\geq\;\; 4\,\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\,\sqrt{\frac{\log p}{N}}. \tag{9}$$

Then, for all $\delta > 0$, with at least probability

$$1 - 2\,\exp(-c_1\,N) - 6\,\exp(-c_2 \log p) - \exp\left(-\sum_{i \in \mathcal{S}(\delta)} \theta_i\right), \tag{10}$$

the estimator $\widehat{\rho}$ of ρ satisfies

$$\widehat{\rho} - \rho \;\;\geq\;\; -\delta. \tag{11}$$

1.2 Suppose $\widehat{\rho} - \rho \geq -\delta$, where $\delta > 0$. Assume that λ_2 satisfies

$$\lambda_2 \;\;\geq\;\; 4\,\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\,\sqrt{\frac{\log p(0, \widehat{\rho}\,)}{N}}. \tag{12}$$

Then, for all $\delta > 0$, with at least probability

$$1 - 2\,\exp(-c_1\,N) - 6\,\exp(-c_2 \log p(0, \rho - \delta)), \tag{13}$$

the ℓ_2-error of the estimator $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$ of $\boldsymbol{\beta}^\star_{[0,\rho-\delta]}$ is bounded above by

$$||\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}||_2 \;\;\leq\;\; \frac{16\sqrt{s}\,\lambda_2}{\alpha}, \tag{14}$$

whereas the ℓ_2-error of the estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^\star$ is bounded above by

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \;\leq\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha} + \mathbb{1}_{\widehat{\rho}\,<\rho}\,\|\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}\|_2, \tag{15}$$

where $\mathbb{1}_{\widehat{\rho}\,<\rho}$ is an indicator of the event $\widehat{\rho}\, < \rho$.

1.3 Suppose $\widehat{\rho}\, - \rho \geq -\delta$, where $\delta > 0$. Assume that condition (C.3) is satisfied and that λ_2 satisfies (12). Then, for all $\delta > 0$, with at least probability (13), the fraction of false-negative edges is small, i.e., for all $\epsilon > 0$, however small, there exists $k_0 > 0$ such that, for all $k > k_0$,

$$\frac{|\mathbb{S} \setminus \widehat{\mathbb{S}}|}{|\mathbb{S}|} \;\leq\; \mathbb{1}_{\widehat{\rho}\,<\rho}\,\epsilon. \tag{16}$$

We compare the statistical error and computing time of the two-step estimation approach to existing approaches and then discuss the importance of sampling.

Remark 3. **Comparison in terms of statistical error.** Among the existing approaches, the most attractive approach is the ℓ_1-penalized LS approach of Basu and Michailidis (2015), because it has computational advantages and its theoretical properties are well-understood. We refer to the ℓ_1-penalized LS estimation approach as LS-ℓ_1 and to the two-step estimation approach as local LS-ℓ_1. Let $\delta > 0$ be small and suppose that $\widehat{\rho}\, - \rho \geq -\delta$, an event that occurs with high probability. Suppose that $\boldsymbol{\beta}^\star$ is estimated by LS-ℓ_1 with $\lambda_1 = 4\,\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\,\sqrt{\log p/N}$ and by local LS-ℓ_1 with $\lambda_2 = 4\,\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \widehat{\rho}\,)/N}$. Then, with high probability,

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \leq \underbrace{\frac{16}{\alpha}\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\sqrt{\frac{s\log p(0,\widehat{\rho}\,)}{N}}}_{\text{local LS-}\ell_1} \;\leq\; \underbrace{\frac{16}{\alpha}\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\sqrt{\frac{s\log p}{N}}}_{\text{LS-}\ell_1}, \tag{17}$$

because $p(0, \widehat{\rho}\,) = \sum_{i=1}^{k} n_i(\widehat{\rho}\,)\, L \leq p = k^2\, L$, where $n_i(\widehat{\rho}\,)$ is the number of possible edges of node i at distances $d \in [0, \widehat{\rho}\,]$. Observe that the error bounds hold regardless of whether ρ is underestimated or overestimated. The error bounds suggest that the parameter vector $\boldsymbol{\beta}^\star_{[0,\rho-\delta]}$ governing possible edges in the interior of the closed balls—which, in most applications, are the parameters of primary interest—should be estimated by local LS-ℓ_1 rather than LS-ℓ_1, and more so when ρ is believed to be small, such as in studies of air pollution and climate change. It is worth noting that both error bounds are small when the number of observations N is large relative to the size of the support s and the total number of parameters p.

Concerning $\boldsymbol{\beta}^\star$, it is important to distinguish between over- and underestimation of ρ. In the case of overestimation, $\mathbb{1}_{\widehat{\rho}\, < \rho}\, \|\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}\|_2 = 0$, which implies that local LS-ℓ_1 tends to outperform LS-ℓ_1. In the case of underestimation, the reverse is true. However, as long as the ℓ_2-norm of the long-distance parameter vector $\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}$ is small—as one would expect when dependence is primarily short-range—the bounds on the error of $\widehat{\boldsymbol{\beta}}$ can be expected to be small even when ρ is underestimated.

Remark 4. **Comparison in terms of computing time.** In terms of computing time, local LS-ℓ_1 tends to be superior to LS-ℓ_1: while LS-ℓ_1 amounts to running k regressions with $k\, L$ predictors, local LS-ℓ_1 amounts to running $|\mathcal{S}|$ regressions with $k\, L$ predictors in Step 1 and k regressions with $\max_{1 \leq i \leq k} n_i(\widehat{\rho}\,)\, L$ predictors in Step 2, where $n_i(\widehat{\rho}\,)$ is the number of possible edges of node i at distances $d \in [0, \widehat{\rho}\,]$. Therefore, as long as the sample is small but well-chosen and the range of dependence is short, local LS-ℓ_1 outperforms LS-ℓ_1.

Remark 5. **Importance of sampling.** Theorem 1 shows that, for any $\delta > 0$, the probability of the event $\widehat{\rho}\, - \rho \geq -\delta$ depends on (a) the size of $\mathcal{S}(\delta)$ and (b) the sample inclusion probabilities of nodes in $\mathcal{S}(\delta)$. The first factor is outside of the control of investigators, whereas the second factor is under the control of investigators. The fact that the probability of $\widehat{\rho}\, - \rho \geq -\delta$ depends on the sample inclusion probabilities shows that prior knowledge about which nodes may have long-distance edges could and should be used to improve the sampling process and thus the estimator $\widehat{\rho}$ of ρ. Such prior knowledge may be available in a wide range of applications, such as in studies of air pollution: e.g., it may be known that some areas have geographical features and wind conditions that facilitate the spread of air pollution, suggesting that investigators should sample time series in such areas with high probability.

6 Simulation results

To compare LS-ℓ_1 and local LS-ℓ_1, we consider three high-dimensional scenarios with k = 100, 200, and 300 time series and N = 150, 300, and 450 observations, respectively; note that $p = k^2\, L \gg N$ in all cases. For each scenario, we generate data from a VAR(1) process with transition matrix $\boldsymbol{A} \equiv \boldsymbol{A}_1$ with 2% sparsity and 5–10 overlapping neighborhoods, with most edges connecting nodes belonging to the same neighborhood. We compare LS-ℓ_1 and local LS-ℓ_1 as well as an oracle version of local LS-ℓ_1 with known ρ in terms of (a) model selection error: the area under the receiving operator characteristic curve (AUROC); the fraction of false-positive (FP)

Table 2: Comparison of LS-ℓ_1, local LS-ℓ_1, and oracle local LS-ℓ_1 with known ρ. Monte Carlo standard deviations are given in parentheses.

		$k = 100$	$k = 200$	$k = 300$
AUROC	LS-ℓ_1	.994 (.005)	.968 (.013)	.867 (.033)
	Local LS-ℓ_1	.987 (.016)	.988 (.011)	.960 (.021)
	Oracle local LS-ℓ_1	.999 (.001)	.996 (.003)	.969 (.019)
Estimation error	LS-ℓ_1	.374 (.026)	.525 (.032)	.714 (.043)
	Local LS-ℓ_1	.343 (.028)	.492 (.037)	.666 (.052)
	Oracle local LS-ℓ_1	.324 (.019)	.479 (.032)	.655 (.052)
Fraction of FP	LS-ℓ_1	.003 (.000)	.003 (.001)	.005 (.000)
	Local LS-ℓ_1	.001 (.000)	.002 (.000)	.004 (.001)
	Oracle local LS-ℓ_1	.001 (.000)	.002 (.000)	.004 (.001)
Fraction of FN	LS-ℓ_1	.054 (.016)	.105 (.028)	.291 (.058)
	Local LS-ℓ_1	.034 (.022)	.052 (.036)	.156 (.068)
	Oracle local LS-ℓ_1	.018 (.013)	.033 (.018)	.133 (.062)

and false-negative (FN) edges; and (b) model estimation error: the relative estimation accuracy measured by $||\boldsymbol{A} - \widehat{\boldsymbol{A}}||_F/||\boldsymbol{A}||_F$, where $||\boldsymbol{A}||_F = \sqrt{\text{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$. In all cases, we use stability selection (Meinshausen and Bühlmann, 2010) to sidestep the problem that the choice of λ depends on the unknown values of $\boldsymbol{\beta}^\star$ and $\boldsymbol{\Sigma}$.

In Table 2, we report the results based on 1,000 Monte Carlo simulations along with Monte Carlo standard deviations. It is not surprising that the oracle version of local LS-ℓ_1 seems to perform best, but local LS-ℓ_1 seems to be close. Both seem to outperform LS-ℓ_1. In Figure 2, we assess the impact of the number of observations N on edge recovery in terms of AUROC using $k = 200$ time series. It is evident that local LS-ℓ_1 outperforms LS-ℓ_1 even when N is fairly small.

7 Application to air pollution in the U.S.

Air pollution is an important health concern. The American Lung Association (2015) states that in the U.S. alone almost 138.5 million people live in areas where air pollution makes breathing dangerous. Air pollution has been associated with cardiac arrest (Ensor et al., 2013), lung disease (Hoek et al., 2013), and cancer (Chen et al., 2015), and the World Health Organization (2014) attributed more than 7 million deaths in 2012 alone to air pollution.

Figure 2: The area under the AUROC curve plotted against N. The blue and red line correspond to local LS-ℓ_1 and LS-ℓ_1, respectively.

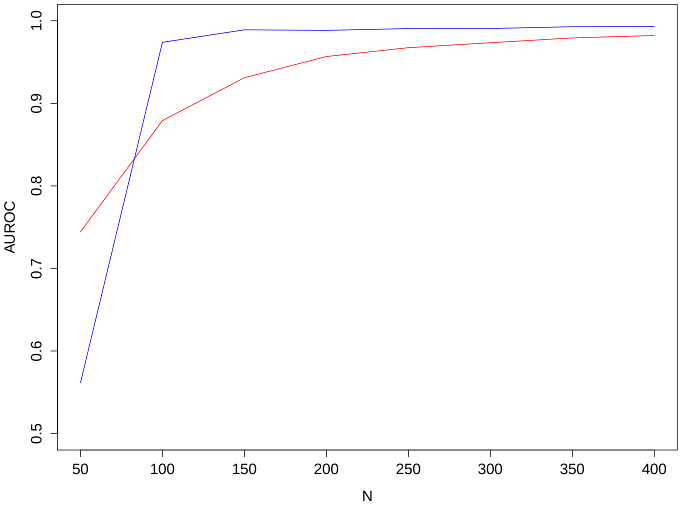


Table 3: Comparison of LS-ℓ_1 and local LS-ℓ_1 in terms of predictive power.

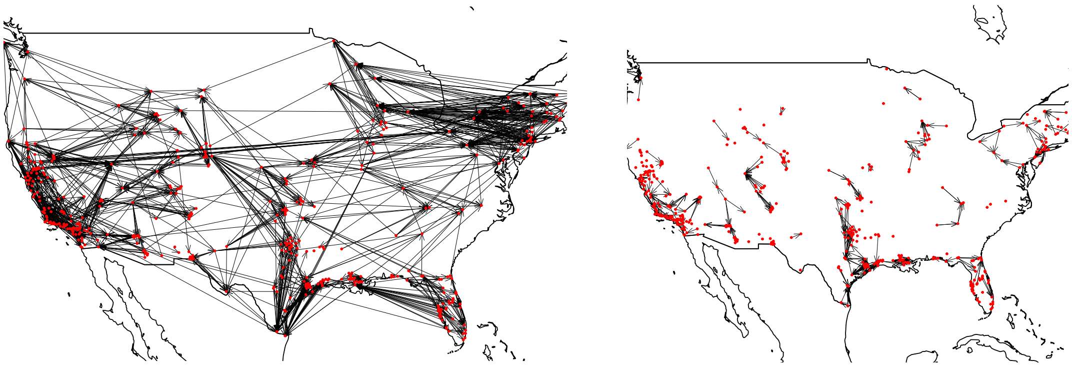
	Method	
	LS-ℓ_1	Local LS-ℓ_1
RPMSE	1.1748	.9980
Time in seconds	938.619	123.365

We exploit local LS-ℓ_1 to contribute to the understanding of the one-day transport of air pollution across space by using data from the U.S. Environmental Protection Agency obtained from http://www.epa.gov/airdata. We first take a bird's eye view at air pollution in the U.S. (Section 7.1) and then zoom in on the Gulf region (Section 7.2), one of the most monitored regions in the U.S.

7.1 A bird's eye view: air pollution in the U.S.

We consider daily measurements of 8-hour maximum concentration of Ozone (O_3) recorded at monitors across the U.S. from January 2010 to December 2014. We use all time series with less than 10% of missing values. The data set contains 444 time series with 1,826 observations. We deal with missing values by univariate linear interpolation and address the strong seasonality issue by applying seasonal AR(1) model to each series separately with the span of seasonality equal to one year and using the scaled

Figure 3: U.S.: estimated graphs based on (a) LS-ℓ_1 and (b) local LS-ℓ_1. While (a) contradicts scientific knowledge by reporting long-range dependence, (b) is in line with scientific knowledge.



(a) LS-ℓ_1

(b) Local LS-ℓ_1

residuals as observations.

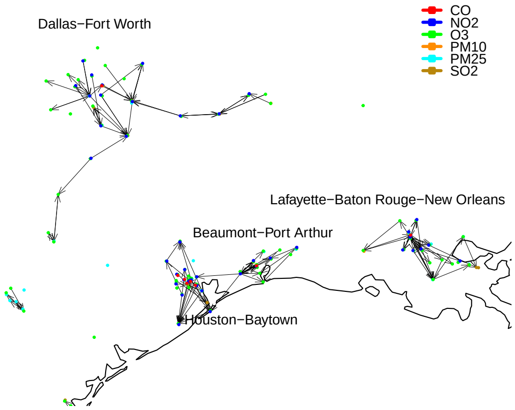
We compare the predictive power of LS-ℓ_1 and local LS-ℓ_1 by using out-of-sample 24-hour ahead root prediction mean squared error (RPMSE) defined as:

$$RPMSE \;=\; \sqrt{\frac{1}{k\,T_{test}} \sum_{l}^{k} \sum_{t}^{T_{test}} (X_l(t) - \widehat{X}_l(t))^2}, \tag{18}$$

where the summation is over T_{test} test observations. We use the first 4 years of the data as training data and the rest as testing data.

Table 3 shows that local LS-ℓ_1 reduces the prediction error by 15% and is almost 8 times faster than LS-ℓ_1. The graphs estimated by LS-ℓ_1 and local LS-ℓ_1 are shown in Figure 3. It is striking that LS-ℓ_1 reports a number of long-distance edges—some of them between monitors separated by more than 2,280 miles. The long-distance edges conflict with scientific knowledge, which suggests that dependence is primarily short-range (e.g., Rao et al., 1997). Local LS-ℓ_1 reports that the estimated range of 24-hour dependence is 182 miles, though almost 2/3 of the edges are between pairs of time series separated by less than 28 miles. These results are consistent with scientific knowledge: it is believed that dependence is primarily short-range, but some extreme examples of 24–72-hour long-range dependence with up to 250 miles have been documented (e.g., Rao et al., 1997).

Figure 4: Gulf region: estimated graph based on local LS-ℓ_1.



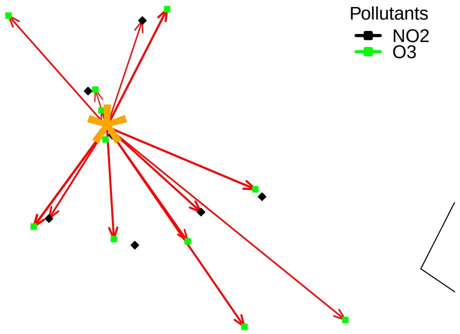
7.2 Zooming in: air pollution in the Gulf region

We zoom in on the Gulf region and consider the one-day transport of 6 pollutants: Ozone (O_3), particle matter ($PM10$ and $PM2.5$), Carbon monoxide (CO), Nitrogen dioxide (NO_2), and Sulfur dioxide (SO_2). The data set consists of 199 time series each having 1,826 observations between January 2010 and December 2014.

The graphical structure estimated by local LS-ℓ_1 is presented in Figure 4. The figure shows 4 clusters: Dallas—Fort Worth, Houston—Baytown, Beaumont—Port Arthur, and Lafayette—Baton Rouge—New Orleans. The estimated range of dependence is less than 59 miles and the median distance of edges is close to 10 miles, which confirms that dependence is short-range.

One interesting observation is that, while dependence between the center and the surrounding areas of the clusters seems to work in both directions, Baton Rouge stands out in that it has more outgoing than incoming edges. The subset of outgoing edges of the largest hub in the Baton Rouge area is presented in Figure 5. Most edges pointing out of Baton Rouge carry positive weights (i.e., the corresponding parameters are positive) and connect NO_2 monitors in the center with O_3 monitors in the surrounding locations. That suggests that an increase in the NO_2 levels in Baton Rouge leads to an increase in O_3 readings at other monitors on the following day. A possible explanation is that Baton Rouge is home to a large industrial complex with one of the largest refineries in the U.S., which may increase air pollution in surrounding areas.

Figure 5: Subset of outgoing edges of the largest hub in the Baton Rouge area, which is close to one of the largest refineries in the U.S. indicated by the orange asterisk. The edges of the monitor are positive (red) or negative (blue) depending on whether the corresponding autoregressive coefficients are positive or negative, respectively.



8 Discussion

We have introduced a simple two-step estimation approach that takes advantage of additional structure in the form of a metric space endowed with local dependence. The two-step estimation approach respects the local nature of multivariate time series and tends to reduce computing time and the statistical error of the parameters of primary interest, i.e., the parameters governing short-distance edges. We have clarified under which conditions the two-step estimation approach results in small estimation errors and have provided non-asymptotic error bounds that hold with high probability. An application to air pollution in the U.S. demonstrated the usefulness of the approach compared with existing approaches.

It is worth noting that the two-step estimation approach is not restricted to high-dimensional multivariate time series: all that is needed to implement the two-step estimation approach in practice and derive its theoretical properties are consistent model selection methods. Therefore, the two-step estimation approach can be extended to other high-dimensional models, such as high-dimensional graphical models (e.g., Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010), as long as additional structure of the form considered here is available and consistent model selection in high dimensions is possible.

Supplementary material All results are proved in the appendix, but the supplement provides more detailed proofs of some auxiliary results.

Acknowledgements The research of the first two authors was supported by the National Science Foundation (NSF award DMS-1513644). The authors are grateful to Sumanta Basu for helping to reproduce results of Basu and Michailidis (2015).

A Proofs

A.1 Proof of Theorem 1.1

We need two lemmas to prove Theorem 1.1.

Lemma 1. For all $\delta > 0$, the probability that none of the nodes $i \in \mathcal{S}(\delta)$ is sampled is bounded above by

$$\exp\left(-\sum_{i \in \mathcal{S}(\delta)} \theta_i\right). \tag{19}$$

Proof of Lemma 1. By definition of $\rho > 0$, for all $\delta > 0$, there exists at least one node with incoming edges at distances $d \in [\rho - \delta, \rho]$, thus $\mathcal{S}(\delta)$ is non-empty. Since nodes i are sampled independently with probabilities $0 < \theta_i < 1$, the probability that none of the nodes $i \in \mathcal{S}(\delta)$ is sampled is bounded above by

$$\exp\left(\sum_{i \in \mathcal{S}(\delta)} \log(1 - \theta_i)\right) \;\; \leq \;\; \exp\left(-\sum_{i \in \mathcal{S}(\delta)} \theta_i\right). \tag{20}$$

Lemma 2. Let $\beta^\star_{\min} \geq 32\sqrt{s}\,\lambda_1/\alpha$, where λ_1 satisfies (9). Then, for any $\delta > 0$ and any non-empty subset $\mathcal{A} \subseteq \mathcal{S}(\delta)$, the probability that none of the incoming edges of nodes $i \in \mathcal{A}$ at distances $d \in [\rho - \delta, \rho]$ is detected is bounded above by

$$2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p). \tag{21}$$

Proof of Lemma 2. By definition of $\rho > 0$, for all $\delta > 0$, there exists at least one node with incoming edges at distances $d \in [\rho - \delta, \rho]$, thus $\mathcal{S}(\delta)$ is non-empty. Consider any non-empty subset $\mathcal{A} \subseteq \mathcal{S}(\delta)$. Let $\mathcal{G}$ be the event that all incoming edges of all nodes $i \in \mathcal{A}$ are detected and $\mathcal{B}$ be its complement. Then the event that none of the incoming edges of nodes $i \in \mathcal{A}$ at distances $d \in [\rho - \delta, \rho]$ is detected is contained in $\mathcal{B}$ and the probability of the event of interest is bounded above by the probability of $\mathcal{B}$.

To bound the probability of $\mathcal{B}$, let $\widehat{\boldsymbol{\beta}}_{\mathcal{N}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ be solutions of optimization problems (3) and (4), respectively, and observe that $\mathcal{G}$ is implied by

$$\frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_\infty \;\leq\; 1. \tag{22}$$

By the variation-independence of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k$,

$$\frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_\infty \;\leq\; \frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}\|_\infty, \tag{23}$$

and by (C.1) and (C.2) and $\beta^\star_{\min} \geq 32\sqrt{s}\,\lambda_1/\alpha$,

$$\frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}\|_\infty \;\leq\; \frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}\|_2 \;\leq\; \frac{2}{\beta^\star_{\min}}\, \frac{16\sqrt{s}\,\lambda_1}{\alpha} \;\leq\; 1. \tag{24}$$

The bound $||\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}||_2 \leq 16\sqrt{s}\,\lambda_1/\alpha$ used in (24) follows from Proposition 4.1 of Basu and Michailidis (2015) and holds as long as (C.1) and (C.2) hold. Therefore, $\mathcal{G}$ occurs as long as both (C.1) and (C.2) hold, whereas $\mathcal{B}$ occurs when either (C.1) or (C.2) or both are violated. A union bound along with $N \geq c_0\, s \log p \geq \log p$ ($c_0 > 1$) shows that the probability of $\mathcal{B}$, and thus the event of interest, is bounded above by

$$2\, \exp(-c_1\, N) + 6\, \exp(-c_2 \log p), \tag{25}$$

where the two terms in (25) are upper bounds on the probabilities that (C.1) or (C.2) are violated, which follow from Propositions 4.2 and 4.3 of Basu and Michailidis (2015), respectively.

Proof of Theorem 1.1. By definition of $\rho > 0$, for all $\delta > 0$, there exists at least one node with incoming edges at distances $d \in [\rho - \delta, \rho]$, thus $\mathcal{S}(\delta)$ is non-empty. Let $\mathcal{G}_1$ be the event that at least one node $i \in \mathcal{S}(\delta)$ with incoming edges at distances $d \in [\rho - \delta, \rho]$ is sampled and that at least one of its incoming edges at distances $d \in [\rho - \delta, \rho]$ is detected and $\mathcal{G}_2$ be the event that at least one false-positive incoming edge of nodes $i \in \mathcal{S}$ at distances $d \in [\rho - \delta, \infty)$ is reported. Then the event $\widehat{\rho} - \rho \geq -\delta$ is equivalent to the event $\mathcal{G}_1 \cup \mathcal{G}_2 \supseteq \mathcal{G}_1$. By Lemmas 1 and 2,

$$\begin{aligned} \mathbb{P}(\widehat{\rho} - \rho \geq -\delta) \;&\geq\; \left[1 - \exp\left(-\sum_{i \in \mathcal{S}(\delta)} \theta_i\right)\right] [1 - 2\, \exp(-c_1\, N) - 6\, \exp(-c_2 \log p)] \\ &\geq\; 1 - 2\, \exp(-c_1\, N) - 6\, \exp(-c_2 \log p) - \exp\left(-\sum_{i \in \mathcal{S}(\delta)} \theta_i\right). \end{aligned}$$

A.2 Proof of Theorem 1.2

Let $\delta > 0$ and $\widehat{\rho} - \rho \geq -\delta$. It is convenient to express the estimator $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$ of $\boldsymbol{\beta}^\star_{[0,\rho-\delta]}$ obtained in Step 2 as the solution of the M-estimation problem

$$\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} \in \underset{\boldsymbol{\beta}_{[0,\rho-\delta]}}{\arg\min} \left[-2\,\boldsymbol{\beta}_{[0,\rho-\delta]}^\top\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \boldsymbol{\beta}_{[0,\rho-\delta]}^\top\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}_{[0,\rho-\delta]} + \lambda_2\,\|\boldsymbol{\beta}_{[0,\rho-\delta]}\|_1\right].$$

We need three additional lemmas to prove Theorem 1.2.

Lemma 3. Assume $N \geq c_0\,s\log p$ ($c_0 > 1$). Then, for all $\delta \geq 0$, with at least probability $1 - 2\,\exp(-c_1\,N)$,

$$\boldsymbol{b}^\top\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{b} \;\;\geq\;\; \alpha\,\|\boldsymbol{b}\|_2^2 - \tau\,\|\boldsymbol{b}\|_1^2 \quad \text{for all } \;\boldsymbol{b}\;\in\;\mathbb{R}^{p(0,\rho-\delta)}. \tag{26}$$

Proof of Lemma 3. Observe that $\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}$ can be written as $\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} = \boldsymbol{E}^\top\,\widehat{\boldsymbol{\Gamma}}\,\boldsymbol{E}$, where $\boldsymbol{E}$ is a 0-1 elimination matrix of suitable order that eliminates the elements of $\widehat{\boldsymbol{\Gamma}}$ that are not elements of $\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}$. By (C.1), for all $\boldsymbol{b} \in \mathbb{R}^{p(0,\rho-\delta)}$,

$$\boldsymbol{b}^\top\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{b} \;\;\geq\;\; \alpha\,\|\boldsymbol{E}\,\boldsymbol{b}\|_2^2 - \tau\,\|\boldsymbol{E}\,\boldsymbol{b}\|_1^2 \;\;=\;\; \alpha\,\|\boldsymbol{b}\|_2^2 - \tau\,\|\boldsymbol{b}\|_1^2, \tag{27}$$

where $||\boldsymbol{E}\,\boldsymbol{b}||_i = ||\boldsymbol{b}||_i$, $i = 1, 2$, because the p-vector $\boldsymbol{E}\,\boldsymbol{b}$ consists of the $p(0, \rho-\delta)$ elements of $\boldsymbol{b}$ and $p - p(0,\rho-\delta)$ 0's. The lower bound (27) holds as long as (C.1) holds. By Proposition 4.2 of Basu and Michailidis (2015), the probability that (C.1) is violated is bounded above by $2\,\exp(-c_1\,N)$ provided $N \geq c_0\,s\log p$ ($c_0 > 1$).

Lemma 4. Assume $N \geq \log p(0, \rho-\delta)$. Then, for all $\delta \geq 0$, with at least probability $1 - 6\,\exp\left(-c_2\log p(0,\rho-\delta)\right)$,

$$||\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^\star_{[0,\rho-\delta]}||_\infty \;\;\leq\;\; \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\sqrt{\frac{\log p(0,\rho-\delta)}{N}}. \tag{28}$$

Proof of Lemma 4. The proof proceeds along the lines of Proposition 4.3 of Basu and Michailidis (2015, supplement, pp. 6–7) by applying concentration inequality (2.11) of Basu and Michailidis (2015) to bound the probability of

$$||\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^\star_{[0,\rho-\delta]}||_\infty \;\;>\;\; 2\,\pi\,\frac{\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})}{a}\,\eta, \tag{29}$$

where $a > 0$ and $\eta > 0$. Choosing $\eta = [a/(2\,\pi)]\,\sqrt{\log p(0,\rho-\delta)/N}$ gives

$$||\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^\star_{[0,\rho-\delta]}||_\infty \;\;>\;\; \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\sqrt{\frac{\log p(0,\rho-\delta)}{N}}. \tag{30}$$

The concentration inequality (2.11) of Basu and Michailidis (2015) and a union bound show that, provided $N \geq \log p(0, \rho - \delta)$, the probability of (30) is bounded above by

$$6 \, \exp\left(-c\, N \, \min\left(\eta, \eta^2\right) + \log p(0, \rho - \delta)\right) \;\;\leq\;\; 6 \, \exp\left(-c_2 \log p(0, \rho - \delta)\right). \tag{31}$$

Lemma 5. Assume that conditions (26) and (28) are satisfied. Then, for all $\delta \geq 0$,

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}. \tag{32}$$

Proof of Lemma 5. By definition of $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$, for all $\boldsymbol{\beta}_{[0,\rho-\delta]} \in \mathbb{R}^{p(0,\rho-\delta)}$,

$$\begin{aligned} & -2\,\widehat{\boldsymbol{\beta}}^\top_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \widehat{\boldsymbol{\beta}}^\top_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} + \lambda_2\,\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}\|_1 \\ \leq\;\; & -2\,\boldsymbol{\beta}^\top_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \boldsymbol{\beta}^\top_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}_{[0,\rho-\delta]} + \lambda_2\,\|\boldsymbol{\beta}_{[0,\rho-\delta]}\|_1. \end{aligned} \tag{33}$$

Set $\boldsymbol{\beta}_{[0,\rho-\delta]} = \boldsymbol{\beta}^\star_{[0,\rho-\delta]}$ and $\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}$. Then (33) reduces to

$$\begin{aligned} & \widehat{\boldsymbol{v}}^\top\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{v}} \\ & \leq 2\,\widehat{\boldsymbol{v}}^\top(\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\boldsymbol{\beta}^\star_{[0,\rho-\delta]}) + \lambda_2\,(\|\boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_1 - \|\boldsymbol{\beta}^\star_{[0,\rho-\delta]} - \widehat{\boldsymbol{v}}\|_1). \end{aligned} \tag{34}$$

The supplement shows that, by using conditions (26) and (28) along with $\lambda_2 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \widehat{\rho}\,)/N}$, $\widehat{\boldsymbol{v}}^\top\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{v}}$ can be bounded as follows:

$$\frac{\alpha}{4}\,\|\widehat{\boldsymbol{v}}\|_2^2 \;\;\leq\;\; \frac{1}{2}\,\widehat{\boldsymbol{v}}^\top\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{v}} \;\;\leq\;\; 4\,\sqrt{s}\,\lambda_2\,\|\widehat{\boldsymbol{v}}\|_2, \tag{35}$$

implying

$$\|\widehat{\boldsymbol{v}}\|_2 \;\;=\;\; \|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}. \tag{36}$$

Proof of Theorem 1.2. Consider $\widehat{\rho}\, < \rho$. By Lemma 5, as long as conditions (26) and (28) are satisfied,

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}. \tag{37}$$

By the triangle inequality,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha} + \|\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}\|_2. \tag{38}$$

The upper bounds (37) and (38) hold as long as conditions (26) and (28) hold. By Lemmas 3 and 4 along with $N \geq c_0\, s \log p \geq \log p(0, \rho - \delta)$ ($c_0 > 1$) and a union bound, the probability that (26) or (28) are violated is bounded above by

$$2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p(0, \rho - \delta)). \tag{39}$$

Consider $\widehat{\rho} \geq \rho$. Then the probability that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \ \leq \ \frac{16\sqrt{s}\,\lambda_2}{\alpha} \tag{40}$$

is violated is bounded above by (39).

A.3 Proof of Theorem 1.3

Proof of Theorem 1.3. Consider $\widehat{\rho} < \rho$. By (C.3), for all $\delta > 0$ and all $\epsilon/2 > 0$, there exists $k_{\delta,\epsilon/2} > 0$ such that, for all $k > k_{\delta,\epsilon/2}$, the fraction of edges at distances $d \in [\rho - \delta, \rho]$ is at most $\epsilon/2$. Assume $k > k_{\delta,\epsilon/2}$. Let g be the fraction of false-negative edges. There are two possible sources of error: (a) some of the edges at distances $d \in [0, \rho - \delta]$ are not be detected by local LS-ℓ_1 and (b) none of the edges at distances $d \in (\rho - \delta, \rho]$ is detected by design. Let g_1 be the fraction of edges at distances $d \in [0, \rho - \delta]$ and g_2 be the fraction of edges at distances $d \in (\rho - \delta, \rho]$. Then

$$\mathbb{P}(g > \epsilon) \ \leq \ \mathbb{P}\left(g_1 > \frac{\epsilon}{2}\right) + \mathbb{P}\left(g_2 > \frac{\epsilon}{2}\right). \tag{41}$$

(a) Term $\mathbb{P}(g_1 > \epsilon/2)$: Since $k > k_{\delta,\epsilon/2}$, there exist at least $(1 - \epsilon/2)\, n > 0$ edges at distances $d \in [0, \rho - \delta]$, where $n > 0$ is the total number of edges in the graph. Denote by $\mathcal{G}$ the event that all edges at distances $d \in [0, \rho - \delta]$ are detected and by $\mathcal{B}$ its complement. Then the event that the fraction of undetected edges at distances $d \in [0, \rho - \delta]$ exceeds $\epsilon/2$ is contained in $\mathcal{B}$ and the probability of the event of interest is bounded above by the probability of $\mathcal{B}$. To bound the probability of $\mathcal{B}$, observe that $\mathcal{G}$ is implied by

$$\frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_\infty \ \leq \ 1. \tag{42}$$

By (26) and (28) in combination with Lemma 5 and $\beta^\star_{\min} \geq 32\sqrt{s}\,\lambda_2/\alpha$,

$$\frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_\infty \leq \frac{2}{\beta^\star_{\min}}\, \|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \leq \frac{2}{\beta^\star_{\min}}\, \frac{16\sqrt{s}\,\lambda_2}{\alpha} \leq 1. \tag{43}$$

The upper bound (43) holds as long as (26) and (28) hold. By Lemmas 3 and 4 with $N \geq c_0\, s \log p \geq \log p(0, \rho - \delta)$ ($c_0 > 1$), the probability that (26) or (28) are violated is bounded above by

$$\mathbb{P}\left(g_1 > \frac{\epsilon}{2}\right) \ \leq \ 2\, \exp(-c_1\, N) + 6\, \exp(-c_2 \log p(0, \rho - \delta)). \tag{44}$$

(b) Term $\mathbb{P}(g_2 > \epsilon/2)$: Since $k > k_{\delta,\epsilon/2}$,

$$\mathbb{P}\left(g_2 > \frac{\epsilon}{2}\right) \ = \ 0. \tag{45}$$

Combining (41) with (44) and (45) gives (13).

Consider $\widehat{\rho} \geq \rho$. An argument along the lines of (a) shows that, with at least probability (13), $g = 0$.

References

American Lung Association (2015), "State of the Air: 2015," Tech. rep., American Lung Association.

Basu, S., and Michailidis, G. (2015), "Regularized estimation in sparse high-dimensional time series models," *The Annals of Statistics*, 43, 1535–1567.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 36, 192–225.

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York: Springer.

Chen, G., Wan, X., Yang, G., and Zou, X. (2015), "Traffic-related air pollution and lung cancer: A meta-analysis," *Thoracic Cancer*, 6, 307–318.

Chudik, A., and Pesaran, M. H. (2011), "Infinite-dimensional VARs and factor models," *Journal of Econometrics*, 163, 4–22.

Davis, R., Zang, P., and Zheng, T. (2012), "Sparse vector autoregressive modeling," *arXiv preprint arXiv:1207.0520.*

Eichler, M. (2012), "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, 153, 233–268.

Ensor, K. B., Raun, L., and Persse, D. (2013), "A case-crossover analysis of out-of-hospital cardiac arrest and air pollution," *Circulation*, 1192–1199.

Hoek, G., Krishnan, R., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. (2013), "Long-term air pollution exposure and cardio-respiratory mortality: a review," *Environmental Health*, 12, 43.

Ising, E. (1925), "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift für Physik A*, 31, 253–258.

Loh, P. L., and Wainwright, M. J. (2012), "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity," *The Annals of Statistics*, 40, 1637–1664.

Lütkepohl, H. (2007), *New introduction to multiple time series analysis*, Springer Science & Business Media.

Meinshausen, N., and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the LASSO," *The Annals of Statistics*, 34, 1436–1462.

— (2010), "Stability selection," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72, 417–473.

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014), "Spatio-temporal data fusion for very large remote sensing datasets," *Technometrics*, 56, 174–185.

Rao, S. T., Zurbenko, I. G., Neagu, R., Porter, P. S., Ku, J. Y., and Henry, R. F. (1997), "Space and time scales for ambient ozone data," *Bulletin of the American Meteorological Society*, 78, 2153–2166.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, 38, 1287–1319.

Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society B*, 77, 647–676.

Song, S., and Bickel, P. J. (2011), "Large vector auto regressions," *arXiv preprint arXiv:1106.3915.*

Wilson, G. T., Reale, M., and Haywood, J. (2015), *Models for Dependent Time Series*, CRC Press.

World Health Organization (2014), "7 million premature deaths annually linked to air pollution," Tech. rep., World Health Organization.

Supplementary Material: High-Dimensional Multivariate Time Series With Local Dependence

Michael Schweinberger   Sergii Babkin   Katherine B. Ensor

We provide a more detailed proof of Lemma 5 stated in the appendix of the main manuscript.

Proof of Lemma 5. By definition of $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$, for all $\boldsymbol{\beta}_{[0,\rho-\delta]} \in \mathbb{R}^{p(0,\rho-\delta)}$,

$$\begin{aligned} & -2\,\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}^{\top}\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} + \lambda_2\,\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}\|_1 \\ \leq\ & -2\,\boldsymbol{\beta}_{[0,\rho-\delta]}^{\top}\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \boldsymbol{\beta}_{[0,\rho-\delta]}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}_{[0,\rho-\delta]} + \lambda_2\,\|\boldsymbol{\beta}_{[0,\rho-\delta]}\|_1. \end{aligned} \tag{46}$$

Set $\boldsymbol{\beta}_{[0,\rho-\delta]} = \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ and $\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$. Then (46) reduces to

$$\begin{aligned} & \widehat{\boldsymbol{v}}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{v}} \\ & \leq 2\,\widehat{\boldsymbol{v}}^{\top}(\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}) + \lambda_2\,(\|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_1 - \|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]} - \widehat{\boldsymbol{v}}\|_1). \end{aligned} \tag{47}$$

The first term on the right-hand side of (47) can be bounded by using condition (28) stated in the appendix of the main manuscript and $\lambda_2 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \widehat{\rho}\ )/N} \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \rho-\delta)/N}$ provided $\widehat{\rho}\ \geq \rho - \delta$:

$$\begin{aligned} & 2\,\widehat{\boldsymbol{v}}^{\top}(\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}) \\ \leq\ & 2\,\|\widehat{\boldsymbol{v}}\|_1\,\|\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_{\infty} \\ \leq\ & 2\,\|\widehat{\boldsymbol{v}}\|_1\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\frac{\log p(0,\rho-\delta)}{N}} \\ \leq\ & \frac{\lambda_2}{2}\,\|\widehat{\boldsymbol{v}}\|_1 \ = \ \frac{\lambda_2}{2}\,(\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1), \end{aligned} \tag{48}$$

where $\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}$ and $\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}$ are the subvectors of $\widehat{\boldsymbol{v}}$ corresponding to the support $\mathbb{S}[0,\rho-\delta]$ of $\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ and its complement $\overline{\mathbb{S}}[0,\rho-\delta]$, respectively. The second term on the right-hand side of (47) can be bounded as follows:

$$\lambda_2\,(\|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_1 - \|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]} - \widehat{\boldsymbol{v}}\|_1) \ \leq\ \lambda_2\,(\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 - \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1) \tag{49}$$

using the triangle inequality

$$\|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_1 \ = \ \|\boldsymbol{\beta}^{\star}_{\mathbb{S}[0,\rho-\delta]}\|_1 \ \leq \ \|\boldsymbol{\beta}^{\star}_{\mathbb{S}[0,\rho-\delta]} - \widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 \tag{50}$$

along with

$$\|\boldsymbol{\beta}^\star_{[0,\rho-\delta]} - \widehat{\boldsymbol{v}}\|_1 \;\; = \;\; \|\boldsymbol{\beta}^\star_{\mathbb{S}[0,\rho-\delta]} - \widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1. \tag{51}$$

Therefore, combining (47) with (48) and (49),

$$0 \;\; \leq \;\; \widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}} \;\; \leq \;\; \frac{3\,\lambda_2}{2} \|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 - \frac{\lambda_2}{2} \, \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1. \tag{52}$$

Thus, $\|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1 \leq 3\,\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1$, implying

$$\|\widehat{\boldsymbol{v}}\|_1 \;\; = \;\; \|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1 \;\; \leq \;\; 4\,\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 \;\; \leq \;\; 4\sqrt{s}\,\|\widehat{\boldsymbol{v}}\|_2. \tag{53}$$

An upper bound on $\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}}$ can therefore be obtained by using (52) and (53):

$$\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}} \;\; \leq \;\; \frac{3\,\lambda_2}{2} \|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 - \frac{\lambda_2}{2} \, \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1 \;\; \leq \;\; 2\,\lambda_2\,\|\widehat{\boldsymbol{v}}\|_1, \tag{54}$$

implying

$$\frac{1}{2}\,\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}} \;\; \leq \;\; \lambda_2\,\|\widehat{\boldsymbol{v}}\|_1 \;\; \leq \;\; 4\sqrt{s}\,\lambda_2\,\|\widehat{\boldsymbol{v}}\|_2. \tag{55}$$

A lower bound on $\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}}$ can be derived by using Lemma 3 stated in the appendix of the main manuscript and (53) along with $s\,\tau \leq \alpha/32$, giving

$$\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}} \;\geq\; \alpha\,\|\widehat{\boldsymbol{v}}\|_2^2 - \tau\,\|\widehat{\boldsymbol{v}}\|_1^2 \;\geq\; \alpha\,\|\widehat{\boldsymbol{v}}\|_2^2 - \tau\,16\,s\,\|\widehat{\boldsymbol{v}}\|_2^2 \;\geq\; \frac{\alpha}{2}\,\|\widehat{\boldsymbol{v}}\|_2^2. \tag{56}$$

Combining the upper and lower bounds on $\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}}$ gives

$$\frac{\alpha}{4}\,\|\widehat{\boldsymbol{v}}\|_2^2 \;\; \leq \;\; \frac{1}{2}\,\widehat{\boldsymbol{v}}^\top \, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} \, \widehat{\boldsymbol{v}} \;\; \leq \;\; 4\sqrt{s}\,\lambda_2\,\|\widehat{\boldsymbol{v}}\|_2, \tag{57}$$

implying

$$\|\widehat{\boldsymbol{v}}\|_2 \;\; = \;\; \|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \;\; \leq \;\; \frac{16\sqrt{s}\,\lambda_2}{\alpha}. \tag{58}$$