On the Sample Complexity of Graphical Model Selection from Non-Stationary Samples

Nguyen Tran FIRST.LAST@AALTO.FI
Department of Computer Science
Aalto University
Espoo, FI

Oleksii Abramenko FIRST.LAST@AALTO.FI
Department of Computer Science
Aalto University
Espoo, FI

Alexander Jung FIRST.LAST@AALTO.FI
Department of Computer Science
Aalto University
Espoo, FI

Editor: TBD

Abstract

We characterize the sample size required for accurate graphical model selection from non-stationary samples. The observed data is modelled as a vector-valued zero-mean Gaussian random process whose samples are uncorrelated but have different covariance matrices. This model contains as special cases the standard setting of i.i.d. samples as well as the case of samples forming a stationary or underspread (non-stationary) processes. More generally, our model applies to any process model for which an efficient decorrelation can be obtained. By analyzing a particular model selection method, we derive a sufficient condition on the required sample size for accurate graphical model selection based on non-stationary data.

1. Introduction

A powerful approach to managing massive datasets (big data) is based on network or graph representations of the datasets (Sandryhaila and Moura, 2014; Hallac et al., 2015; Koller et al., 2009; Cui et al., 2016). Examples of networked data are found in signal processing where signal samples can be arranged as a chain, in image processing with pixels arranged on a grid, in wireless sensor networks where measurements conform to sensor proximity (Sandryhaila and Moura, 2014). Organising data using networks is also used in knowledge bases (graphs) whose items are linked by relations (Vrandečić and Krötzsch, 2014; Sadeghi et al., 2017).

Using network models is beneficial from a computational and statistical perspective. Indeed, network models for data lend naturally to highly scalable learning algorithms in the form of message passing on the data network (Boyd et al., 2010). Moreover, the network structure allows to borrow statistical strength across different localized high-dimensional statistical models which are associated with individual data points (nodes) (Hallac et al., 2015; Ambos et al., 2018). Finally, network models provide a high level of flexibility in order to cope with heterogeneous datasets composed of different data types (e.g., mixtures of audio, video and text data).

In some applications, the network structure underlying the data is not known explicitly but has to be learned in a data-driven fashion. This task can be accomplished in a principled way by using probabilistic graphical models (PGM) (Koller et al., 2009; Wainwright and Jordan, 2008). Within a PGM we interpret data points as realizations of random variables. A particular type of PGM is based on representing the conditional independence relations between individual data points using a network structure (graph) (Lauritzen, 1996; Wainwright and Jordan, 2008). The problem of estimating the network structure of a PGM from observed data is known as graphical model selection (GMS).

Many efficient methods have been proposed for GMS for data which is modelled as sequences of i.i.d. realizations of some underlying random vector (Ravikumar et al., 2010; Friedmann et al., 2008; Tan et al., 2014). The extension of GMS from the i.i.d. setting to cope with correlations between vector samples using stationary process models has been studied in (Bach and Jordan, 2004; Jung, 2015; Hannak et al., 2014; Jung et al., 2014; Jung et al., 2015). A robust GMS method which is able to cope with outliers is proposed in (Yang and Lozano, 2015).

It is of practical relevance for the usage of GMS methods to understand the fundamental requirements on the available data such that accurate GMS is possible. For data which can be modelled as i.i.d. realizations of a Gaussian random vector (Gaussian Markov random field), the required sample size is well understood. A lower bound on the sample size has been obtained by (Wang et al., 2010), which does not place any computational constraints on the GMS method. Remarkably, this lower bound can be achieved by computationally tractable convex optimizaton methods (Ravikumar et al., 2011) proving them as optimal in terms of sample size requirement. By adapting the information-theoretic approach of (Wang et al., 2010), a lower bound on the sample size required for accurate GMS from data conforming to a stationary random process model is presented in (Hannak et al., 2014).

Contribution. Our focus is on the required sample (data) size which allows for accurate GMS. In contrast to most existing work, we study GMS for data which cannot be well modelled as a stationary random process. To this end, we propose a simple but useful probabilistic model for non-stationary data whose statistical properties vary over time or space (see Section 2). This model requires that samples can be grouped into blocks (of known size) within which the samples can be considered as i.i.d. Our model includes, as important special cases, the case of i.i.d. data as well as data forming a stationary time series. Moreover, the model also applies to data which can be represented as either cyclostationary (Kipnis et al., 2018), locally stationary (Mallat et al., 1998) or underspread random processes Boashash (2003). Thus, the applicability of our model is quite broad.

In general, the process model used in this paper is applicable whenever an efficient decorrelation transformation, which allows one to turn the raw data into blocks of i.i.d. random vectors, is available. Our model has been used in (Danaher et al., 2014) in the context of a bioinformatics application. The main contribution of (Danaher et al., 2014) is the formulation of a tractable GMS method based on convex optimization. Instead, our aim is not the design of a computationally tractable ("polynomial time") GMS method but rather a characterization of the required amount of data (sample size) for reliable GMS. To this end we provide a careful analysis of a computationally intractable neighborhood regression method which amounts to an exhaustive search for conditional dependencies between two particular data points (represented by two nodes in the PGM), when conditioning on all remaining data points.

Our conceptual approach to GMS extends the sparse neighbourhood regression approach put forward in (Meinshausen and Bühlmann, 2006) for GMS from i.i.d. samples to the non-stationary setting. However, while (Meinshausen and Bühlmann, 2006) proposes a computationally attractive convex relaxation of sparse neighbourhood regression using a Lasso-based estimator, we are mainly interested in the fundamental limits on the required sample size without constraining the computational complexity of the GMS method.

The main contribution of this work is a precise characterization of the sample size required for accurate GMS from non-stationary data. In particular, we show that the required sample size depends crucially on the minimum average squared partial correlation between the individual process components. If this quantity is sufficiently large, accurate GMS is possible even in the high-dimensional regime, where the length of the vector samples might (drastically) exceed the number of available training samples (data points).

Outline. After formalizing the problem setup in Section 2, we analyze a simple GMS method, which we term sparse neighbourhood regression, in Section 3. In particular, for a given sample size and sparsity level of the network structure, we derive an upper bound on the probability that sparse neighbourhood regression fails in recovering the correct network structure of the PGM. This upper bound on the error probability implies an upper bound on the required sample size such that GMS is feasible. We verify our theoretical findings by means of numerical experiments in Section 4.

Notation For a vector $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, the Euclidean and ∞-norm are $\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^T\mathbf{x}}$ and $\|\mathbf{x}\|_\infty := \max_i |x_i|$, respectively. The m-th largest eigenvalue of a positive semidefinite (psd) matrix $\mathbf{C}$ is $\lambda_m(\mathbf{C})$. Given a matrix $\mathbf{Q}$, we denote its transpose, trace, rank, spectral norm and Frobenius norm by $\mathbf{Q}^T$, $\mathrm{tr}\{\mathbf{Q}\}$, $\mathrm{rank}\{\mathbf{Q}\}$, $\|\mathbf{Q}\|_2$ and $\|\mathbf{Q}\|_{\mathrm{F}}$, respectively. For a finite sequence of matrices $\mathbf{Q}_l \in \mathbb{R}^{d\times d}$, with $l = 1, \ldots B$, we denote by $\mathrm{blkdiag}\{\mathbf{Q}_l\}$ the block diagonal matrix of size $Bd \times Bd$ with the lth diagonal block given by $\mathbf{Q}_l$. The identity matrix of size $d \times d$ is $\mathbf{I}_d$. The minimum (maximum) of two numbers a and b is denoted $a \wedge b$ ($a \vee b$). The set of non-negative real (integer) numbers is denoted $\mathbb{R}_+$ ($\mathbb{Z}_+$). The probability of an event $\mathcal{E}$ is $\mathrm{P}\{\mathcal{E}\}$. The complement of an event $\mathcal{A}$ is denoted $\mathcal{A}^c$. The expectation of a random variable y is $\mathrm{E}\{y\}$.

2. Problem Formulation

We consider a system which is constituted by p components $\mathbf{x}_i$, for $i = 1, \ldots, p$. In a bioinformatics application, such a system might be a gene regulatory network with the components $\mathbf{x}_i$ representing concentrations of particular genes (Davidson and Levin, 2005). The system is observed by acquiring N vector-valued samples $\{\mathbf{x}[n]\}_{n=1}^N$, each sample

$$\mathbf{x}[n] = \big(x_1[n], \ldots, x_p[n]\big)^T \in \mathbb{R}^p$$

constituted by p scalar "measurements" $x_i[n]$ for $i = 1, \ldots, p$.

The samples $\mathbf{x}[n]$ are modelled as realizations of zero-mean Gaussian random vectors, which are uncorrelated such that

$$\mathrm{E}\big\{\mathbf{x}[n]\big(\mathbf{x}[n']\big)^{T}\big\}=\mathbf{0} \text{ for } n\neq n'.$$

The probability distribution of the samples $\mathbf{x}[n]$ is fully specified by the covariance matrices

$$\mathbf{C}[n] := \mathrm{E}\big\{\mathbf{x}[n]\big(\mathbf{x}[n]\big)^{T}\big\}. \tag{1}$$

In general, the covariance matrix $\mathbf{C}[n]$ varies with sample index n, i.e., $\mathbf{C}[n] \neq \mathbf{C}[n']$ for $n\neq n'$ in general. However, we do not allow for arbitrary variation of the covariance matrix but require it to be constant over blocks of L consecutive samples $\mathbf{x}[n],\ldots,\mathbf{x}[n+L-1]$. We model the observed samples as blocks of i.i.d. Gaussian random vectors,

$$\underbrace{\overbrace{\begin{pmatrix}x_1[1]\\ \vdots \\ x_p[1]\end{pmatrix}}^{\mathbf{x}[1]},\ldots,\overbrace{\begin{pmatrix}x_1[L]\\ \vdots \\ x_p[L]\end{pmatrix}}^{\mathbf{x}[L]}}_{\text{i.i.d.}\sim\mathcal{N}(\mathbf{0},\mathbf{C}^{(b=1)})},\ldots,\ldots,\ldots\underbrace{\overbrace{\begin{pmatrix}x_1[(B-1)L+1]\\ \vdots \\ x_p[(B-1)L+1]\end{pmatrix}}^{\mathbf{x}[(B–1)L+1]},\ldots,\overbrace{\begin{pmatrix}x_1[N]\\ \vdots \\ x_p[N]\end{pmatrix}}^{\mathbf{x}[N]}}_{\text{i.i.d.}\sim\mathcal{N}(\mathbf{0},\mathbf{C}^{(b=B)})} \tag{2}$$

Our goal is to estimate the conditional dependencies between the components $\mathbf{x}_i$ which are represented by the sequences $x_i[1],\ldots,x_i[N]$ in (2). Such a global dependence structure between entire sequences has also been considered in (Bach and Jordan, 2004). However, (Bach and Jordan, 2004) considered stationary time series data, we consider global dependence structure between quantities that are modelled as a non-stationary process of the form (2).

The vector samples $\mathbf{x}[n]$ in (2) are uncorrelated (independent) zero-mean Gaussian vectors with covariance matrix

$$\mathbf{C}[n]=\mathbf{C}^{(b)} \text{ for } n \in \{(b-1)L+1,\ldots,bL\}. \tag{3}$$

For ease of exposition and without essential loss of generality, we henceforth assume the sample size N to be a integer multiple of the block length L (which is assumed fixed and known), such that $N=BL$, with the number B of data blocks. Moreover, we tacitly assume the covariance matrices $\mathbf{C}[n]$ to be non-singular (invertible) with inverse $\big(\mathbf{C}[n]\big)^{-1}$ (see Assumption 3 below).

The model (2) reduces to the i.i.d. setting for $B = 1$ and block length $L = N$. In this paper, we study the fundamental limits of accurate GMS based on non-stationary data which conforms to the model (2) with $B > 1$ (the non-stationary setting).

At first glance, the process model (2) might seem overly restrictive as it still requires blocks of consecutive samples to be i.i.d. However, as we will now discuss, the model (2) can be used as an approximation at least for some large and practically relevant classes of random processes. Moreover, for each of these process classes we are able to identify useful choices for the block length L in (2).

Stationary Processes. The model (2) covers the case where the observed samples form a stationary process (Jung, 2015; Dahlhaus, 2000; Bach and Jordan, 2004; Jung et al., 2015).

Indeed, consider a zero-mean Gaussian stationary process $\mathbf{x}[n]$ with auto-covariance function

$$\mathbf{R}_x[m] := \mathrm{E}\big\{\mathbf{x}[n]\big(\mathbf{x}[n-m]\big)^T\big\} \tag{4}$$

and spectral density matrix (SDM) (Dahlhaus, 2000)

$$\mathbf{S}_x(\theta) := \sum_{m=-\infty}^{\infty} \mathbf{R}_x[m] \exp(-j2\pi\theta m). \tag{5}$$

Let

$$\hat{\mathbf{x}}[k] := (1/\sqrt{N}) \sum_{n=1}^{N} \mathbf{x}[n] \exp(-j2\pi(n-1)(k-1)/N)$$

denote the discrete Fourier transform (DFT) of the stationary process $\mathbf{x}[n]$. Then, by well-known properties of the DFT (see, e.g., (Brockwell and Davis, 1991)), the vectors $\hat{\mathbf{x}}[k]$, for $k = 1, \ldots, N$, are approximately uncorrelated Gaussian random vectors with zero mean and covariance matrix $\mathbf{C}[k] \approx \mathbf{S}_z(k/N)$. For a stationary process $\mathbf{x}[n]$ with (effective) correlation width W, the SDM is approximately constant (flat) over a frequency interval of length $1/W$. Thus, the DFT vectors $\hat{\mathbf{x}}[k]$ approximately conform to the process model (2) with block length $L = N/W$ (since the DFT vectors correspond to SDM samples at evenly spaced frequency points separated by $1/N$).

Cyclostationary Processes. As detailed in (Kipnis et al., 2018), (discrete-time) cyclostationary processes can be transformed to vector-valued (or multivariate) stationary processes which can then, in turn, be transformed to a process of the form (2) via a DFT.

Locally Stationary Processes. The process model (2) applies to locally stationary processes (Starica and Granger, 2005; Wahlberg and Hansson, 2007; Mallat et al., 1998). The i.i.d. blocks of L consecutive vector samples in (2) correspond to the homogeneity intervals defined in (Starica and Granger, 2005). Particular approaches for optimally chosing the block length L for the model (2) are studied in (Starica and Granger, 2005; Dahlhaus and Giraitis, 1998). One important example of locally stationary processes are time-varying autoregressive processes which extend traditional autoregressive process models by allowing time-varying regression coefficients (Dahlhaus, 2009; Brockwell and Davis, 1991).

Underspread Processes. The process model (2) is also useful for the important class of underspread non-stationary processes (Jung et al., 2013; Boashash, 2003). A continuous-time random process $\mathbf{x}(t)$ is called underspread if its expected ambiguity function (EAF)

$$\bar{\mathbf{A}}_x(\tau, \nu) := \int_{t=-\infty}^{\infty} \mathrm{E}\big\{\mathbf{x}(t+\tau/2)\big(\mathbf{x}(t-\tau/2)\big)^T\big\} \exp(-j2\pi t\nu) dt$$

is well-concentrated around the origin in the (τ, ν) plane. In particular, if the EAF $\bar{\mathbf{A}}_x(\tau, \nu)$ of $\mathbf{x}(t)$ is (effectively) supported on the rectangle $[-\tau_0/2, \tau_0/2] \times [-\nu_0/2, \nu_0/2]$, then the process $\mathbf{x}(t)$ is underspread if $\tau_0\nu_0 \ll 1$.

It can be shown that for a suitably chosen prototype function $g(t)$ (e.g., a Gaussian pulse) and grid constants T and F, the Weyl-Heisenberg set $\big\{g^{(n,k)}(t) := g(t-nT)e^{-2\pi kFt}\big\}_{n,k\in\mathbb{Z}}$, yields zero-mean analysis coefficients $\hat{\mathbf{x}}[n,k] = \int_{t=-\infty}^{\infty} \mathbf{x}(t) g^{(n,k)}(t) dt$ which are approximately uncorrelated and provide a complete representation of the process $\mathbf{x}(t)$. Moreover, the

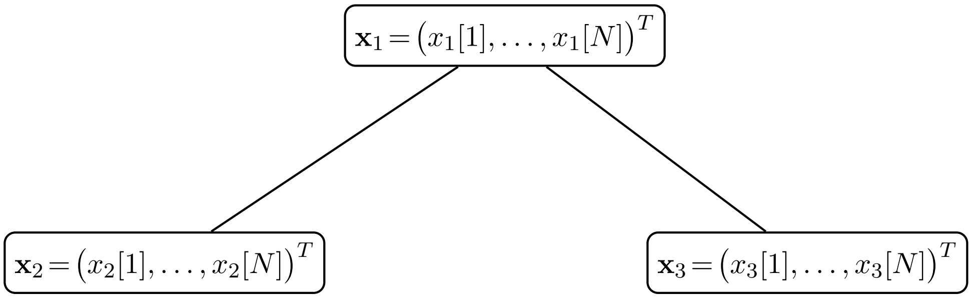

$\mathbf{x}_1 = \big(x_1[1], \ldots, x_1[N]\big)^T$
$\mathbf{x}_2 = \big(x_2[1], \ldots, x_2[N]\big)^T$
$\mathbf{x}_3 = \big(x_3[1], \ldots, x_3[N]\big)^T$

Figure 1: Example of a CIG underlying some vector-valued samples $\mathbf{x}[n] = (x_1[n], x_1[n], x_3[n])^T$, for $n = 1, \ldots, N$ (see (2)). Our aim is to learn the CIG solely from observing the samples $\mathbf{x}[n]$.

covariance matrix of $\hat{\mathbf{x}}[(n,k)]$ is approximately equal to the value $\overline{\mathbf{W}}_x(nT, kF)$ of the Wigner-Ville spectrum (WVS) (Velez and Absher, 1990)

$$\overline{\mathbf{W}}_x(t,f) := \int\limits_{\tau=-\infty}^{\infty} \int\limits_{\nu=-\infty}^{\infty} \bar{\mathbf{A}}_x(\tau,\nu) \exp(-2\pi(f\tau - \nu t)) d\tau d\nu$$

which can be loosely interpreted as a time-varying power spectral density. For an underspread process whose EAF is effectively supported on $[-\tau_0/2, \tau_0/2] \times [-\nu_0/2, \nu_0/2]$, the WVS $\overline{\mathbf{W}}_x(nT, kF)$ is approximately constant over a rectangle of area $\approx 1/(\tau_0\nu_0)$. Thus, the vectors $\hat{\mathbf{x}}[(n,k)]$ approximately conform to the process model (2) with block length $L \approx \frac{1}{TF\tau_0\nu_0}$.

Conditional Independence Graph. We now define a PGM for the observed samples $\{\mathbf{x}[n]\}_{n=1}^{N}$ (cf. (2)) by identifying the individual components

$$\mathbf{x}_i = (x_i[1], \ldots, x_i[N])^T \tag{6}$$

with the nodes $\mathcal{V} = \{1, \ldots, p\}$ of an undirected simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (see Figure 1). This graph encodes conditional independence relations between the components $\mathbf{x}_i$ and is hence called the conditional independence graph (CIG) of the process $\mathbf{x}[n]$. In particular, an edge is absent between nodes $i, j \in \mathcal{V}$, i.e., $\{i, j\} \notin \mathcal{E}$, if the corresponding process components $\mathbf{x}_i$ and $\mathbf{x}_j$ are conditionally independent, given the remaining components $\{\mathbf{x}_r\}_{r \in \mathcal{V} \setminus \{i,j\}}$.

We highlight that the CIG $\mathcal{G}$ represents stochastic dependencies between the components $\big\{\mathbf{x}_i\big\}_{i=1}^{p}$ (see (6)) of the vector samples $\mathbf{x}[1], \ldots, \mathbf{x}[N]$ in a global fashion, i.e., jointly for all $n = 1, \ldots, N$. In particular, the edge set $\mathcal{E}$ does not depend on the sample index n since we define the CIG for the entire sample process for $n = 1, \ldots, N$.[1] Using a global CIG between data points which are modelled as non-stationary vector samples is useful for many applications (see (Bach and Jordan, 2004; Danaher et al., 2014; Dahlhaus, 2000; Eichler et al., 2003) and references therein).

[1] In principle, it is also possible to define a CIG $\mathcal{G}^{(n)}$ separately for each individual sample $\mathbf{x}[n] = \big(x_1, \ldots, x_p[n]\big)^T$, which can be interpreted as a single realization of a Gaussian random Markov field. The edge set of the global CIG we are considering in this paper is the union of the edge sets in the sample-wise CIGs $\mathcal{G}^{(n)}$.

Since we model the observed samples $\mathbf{x}[n]$ as realizations of a Gaussian process (see (2)), the edges of the CIG can be read off conveniently from the inverse covariance (precision) matrices $\mathbf{K}[n] := \big(\mathbf{C}[n]\big)^{-1}$ (see (1)). In particular, $\mathbf{x}_i$ are $\mathbf{x}_j$ are conditionally independent, given $\{\mathbf{x}_r\}_{r \in \mathcal{V} \setminus \{i,j\}}$, if and only if $K_{i,j}[n] = 0$ for all $n \in \{1, \dots, N\}$ (Brockwell and Davis, 1991, Prop. 1.6.6). Thus, we have the following characterization of the CIG $\mathcal{G}$ associated with the process $\mathbf{x}[n]$:

$$\{i, j\} \notin \mathcal{E} \text{ if and only if } K_{i,j}[n] = 0 \text{ for all } n = 1, \dots, N. \tag{7}$$

Note that the CIG characterization (7) involves a coupling over all samples $\{\mathbf{x}[n]\}_{n \in \{1,\dots,N\}}$. Indeed, an edge is absent between two different nodes $i, j \in \mathcal{V}$ in the CIG, i.e., $\{i, j\} \notin \mathcal{E}$, if and only if the precision matrix entry $K_{i,j}[n]$ is zero *for all* $n \in \{1, \dots, N\}$.

The strength of a connection $\{i, j\} \in \mathcal{E}$ between process components $\mathbf{x}_i$ and $\mathbf{x}_j$ is measured by the *average squared partial correlation*

$$\begin{aligned} \rho^2_{i,j} &:= (1/N) \sum_{n=1}^{N} K^2_{i,j}[n] / K^2_{i,i}[n] \\ &\overset{(3)}{=} (1/B) \sum_{b=1}^{B} \big(K^{(b)}_{i,j}\big)^2 / \big(K^{(b)}_{i,i}\big)^2. \end{aligned} \tag{8}$$

We highlight that the quantity $\rho^2_{i,j}$ is determined by the conditional distribution of $\mathbf{x}_i$ and $\mathbf{x}_j$ given all the remaining process components $\mathbf{x}_m$ with $m \in \{1, \dots, p\} \setminus \{i, j\}$. Note that the squared partial correlation (8) is not symmetric since $\rho^2_{i,j} \neq \rho^2_{j,i}$ This is different from the widely used definition of partial correlation whose square would be (see (Meinshausen and Bühlmann, 2006; Wang et al., 2010; Ravikumar et al., 2011))

$$(1/B) \sum_{b=1}^{B} \big(K^{(b)}_{i,j}[n]\big)^2 / (K^{(b)}_{i,i}[n] K^{(b)}_{j,j}[n]).$$

However, we find the non-symmetric definition more natural for our analysis of the simple GMS method proposed in Section 3.

By (7) and (8), two nodes $i, j \in \mathcal{V}$ are connected by an edge $\{i, j\} \in \mathcal{E}$ if and only if $\rho_{i,j} \neq 0$. Note that $\rho_{i,j}$ is an average measure, i.e., even if the marginal squared partial correlation $K^2_{i,j}[n]/K^2_{i,i}[n]$ is very small for some n, the average squared partial correlation $\rho^2_{i,j}$ might still be sufficiently large. The definition (8) is a natural extension of (a non-symmetric version of) (Wang et al., 2010, Eq. (3)), which considers i.i.d. samples, to the non-stationary model (2) considered in this paper.

Accurate estimation of the CIG $\mathcal{G}$ based on a finite number N of samples (incurring unavoidable estimation errors) is only possible for sufficiently large squared partial correlations $\rho^2_{i,j}$ for all edges $\{i, j\} \in \mathcal{E}$ in the CIG $\mathcal{G}$.

Assumption 1 *The average squared partial correlation $\rho^2_{i,j}$ (see (8)) between any two connected components $\mathbf{x}_i$ and $\mathbf{x}_j$ with $\{i, j\} \in \mathcal{E}$ is lower bounded*

$$\rho^2_{i,j} \geq \rho^2_{\min} \quad \textit{for every } \{i, j\} \in \mathcal{E}, \tag{9}$$

with some known lower bound $\rho^2_{\min} > 0$.

Given a node $i \in \mathcal{V}$ in the CIG $\mathcal{G}$, we denote its neighbourhood and degree as $\mathcal{N}(i) := \{j \in \mathcal{V} \setminus \{i\} : \{i,j\} \in \mathcal{E}\}$ and $s_i = |\mathcal{N}(i)|$, respectively. While in principle, our analysis of GMS applies to processes with arbitrary CIG structure, our results will be most useful if the underlying CIG $\mathcal{G}$ is sparse in the sense of having a small (bounded) maximum node degree.

Assumption 2 *The node degrees in the CIG are bounded by some sparsity level s as*

$$\max_{i \in \mathcal{V}} s_i \leq s, \text{ with } s < (p/3) \wedge (L/3). \tag{10}$$

We highlight that our approach to GMS for the process (2) requires a known sparsity level s for the upper bound (10). However, in contrast to (Wainwright, 2009a), the sparsity s is only required to form an upper bound on the node degrees s_i. In particular, our approach is able to handle nodes $i \in \mathcal{G}$ which have smaller degrees $s_i < s$.

The requirement (10) implies a trade-off between the block length L of consecutive i.i.d. samples in (2) and the sparsity s of the underlying CIG. In particular, for a given sample size N, we can tolerate less smoothness (smaller block length L in (2)), if the underlying CIG is more sparse (having smaller maximum degree s).

It will be notationally convenient to assume the samples $\mathbf{x}[n]$ suitably scaled such that the eigenvalues of the covariance matrices $\mathbf{C}[n]$ are bounded with known constants.

Assumption 3 *The eigenvalues of the covariance matrices $\mathbf{C}[n]$ are bounded as*

$$1 \leq \lambda_l(\mathbf{C}[n]) \leq \beta \text{ for all } l \in \{1, \ldots, p\} \text{ and } n \in \{1, \ldots, N\}, \tag{11}$$

with some known upper bound $\beta \geq 1$.

Fixing the lower bound in Assumption 3 to be equal to 1 is not restrictive since we assume the covariance matrices $\mathbf{C}[n]$ to be invertible. Indeed, the eigenvalues $\lambda_l(\mathbf{C}[n])$ can be scaled suitably by multiplying the samples $\mathbf{x}[n]$ with a constant.

3. Sparse Neighborhood Regression

The CIG $\mathcal{G}$ of the process $\mathbf{x}[n]$ in (2) is fully specified by the neighbourhoods of the nodes in the CIG. Indeed, rather trivially, we can determine the CIG by determining the neighbourhoods $\mathcal{N}(i)$ separately for each node $i \in \mathcal{V}$. Thus, without loss of generality, we will focus on the sub-problem of determining the neighbourhood $\mathcal{N}(i)$ of an arbitrary but fixed node $i \in \mathcal{V}$.

In view of the process model (2) we define, for an arbitrary but fixed block $b \in \{1, \ldots, B\}$, the ith process component as

$$\mathbf{x}_i^{(b)} := \big(x_i[(b-1)L+1], \ldots, x_i[bL]\big)^T \in \mathbb{R}^L.$$

The process components of different blocks are uncorrelated, i.e.,

$$\mathrm{E}\big\{\mathbf{x}_i^{(b)} \big(\mathbf{x}_j^{(b')}\big)^T\big\} = \mathbf{0} \text{ for } b \neq b' \text{ and any } i, j \in \mathcal{V}.$$

Elementary properties of multivariate normal distributions (see, e.g., (Gallager, 2013, Thm. 3.5.1)) and the fact $K_{i,j}[n] = 0$ for $j \notin \mathcal{N}(i)$ (cf. (7)), yield

$$\mathbf{x}_i^{(b)} = \sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}_j^{(b)} + \boldsymbol{\varepsilon}_i^{(b)}, \tag{12}$$

with the coefficients $a_j = -K^{(b)}_{i,j}/K^{(b)}_{i,i}$. The error vector $\boldsymbol{\varepsilon}^{(b)}_i \sim \mathcal{N}(\mathbf{0}, (1/K^{(b)}_{i,i})\mathbf{I}_L)$ is uncorrelated with the vectors $\{\mathbf{x}^{(b)}_j\}_{j \in \mathcal{N}(i)}$. Note that the random vector $\sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}^{(b)}_j$ in (12) is the minimum mean squared error (MMSE) estimator of $\mathbf{x}^{(b)}_i$ using the random vectors $\{\mathbf{x}^{(b)}_j\}_{j \in \mathcal{N}(i)}$ as observations (see (Papoulis and Pillai, 2002)).

Given some index set $\mathcal{T} \subseteq \{1, \ldots, p\}$ with $\mathcal{N}(i) \setminus \mathcal{T} \neq \emptyset$, another application of (Gallager, 2013, Thm. 3.5.1) to the component $\sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}^{(b)}_j$ in the decomposition (12) yields

$$\mathbf{x}^{(b)}_i = \underbrace{\sum_{j \in \mathcal{T}} c_j \mathbf{x}^{(b)}_j + \tilde{\mathbf{x}}^{(b)}_i}_{= \sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}^{(b)}_j \text{ see (12)}} + \boldsymbol{\varepsilon}^{(b)}_i, \tag{13}$$

with the random vectors $\tilde{\mathbf{x}}^{(b)}_i$, $\{\mathbf{x}^{(b)}_j\}_{j \in \mathcal{T}}$ and $\boldsymbol{\varepsilon}^{(b)}_i$ being jointly Gaussian. Moreover, the random vectors $\tilde{\mathbf{x}}^{(b)}_i$ are uncorrelated with the random vectors $\{\mathbf{x}^{(b)}_j\}_{j \in \mathcal{T}}$, $\boldsymbol{\varepsilon}^{(b)}_i$ and distributed as

$$\tilde{\mathbf{x}}^{(b)}_i \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2_b \mathbf{I}_L). \tag{14}$$

Note that the vector $\tilde{\mathbf{x}}^{(b)}_i$ in (13) is the estimation error incurred by the MMSE estimator of the random vector $\sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}^{(b)}_j$ using $\{\mathbf{x}^{(b)}_j\}_{j \in \mathcal{T}}$ as observations.

Using (Gallager, 2013, Thm. 3.5.1), the variance $\tilde{\sigma}^2_b$ of (the i.i.d. entries of) $\tilde{\mathbf{x}}^{(b)}_i$ can be obtained as

$$\tilde{\sigma}^2_b = \mathbf{a}^T \big(\widetilde{\mathbf{K}}^{(b)}\big)^{-1} \mathbf{a} \tag{15}$$

with the matrix $\widetilde{\mathbf{K}}^{(b)} = \big(\big(\mathbf{C}^{(b)}_{\mathcal{N}(i) \cup \mathcal{T}}\big)^{-1}\big)_{\mathcal{N}(i) \setminus \mathcal{T}}$ and the vector $\mathbf{a} \in \mathbb{R}^{|\mathcal{N}(i) \setminus \mathcal{T}|}$ whose entries are given by $a_j = -K^{(b)}_{i,j}/K^{(b)}_{i,i}$, for $j \in \mathcal{N}(i) \setminus \mathcal{T}$. In what follows we will make use of a lower bound on the variance $\tilde{\sigma}^2_b$ which is due to Assumption 3. Indeed, by Assumption 3 we have $\lambda_l\big(\big(\widetilde{\mathbf{K}}^{(b)}\big)^{-1}\big) \geq 1$, for all $l = 1, \ldots, |\mathcal{N}(i) \setminus \mathcal{T}|$, which implies (see (15)) the lower bound

$$\tilde{\sigma}^2_b \geq \sum_{j \in \mathcal{N}(i) \setminus \mathcal{T}} (K^{(b)}_{i,j}/K^{(b)}_{i,i})^2. \tag{16}$$

On the other hand, we can use Assumption 3 to obtain (via (13) and (12)) the upper bound[2]

$$\tilde{\sigma}^2_b \leq \beta. \tag{17}$$

It will be convenient to stack the vectors $\tilde{\mathbf{x}}^{(b)}_i$ (cf. (14)) into a single Gaussian random vector

$$\tilde{\mathbf{x}}_i := \big(\big(\tilde{\mathbf{x}}^{(1)}_i\big)^T, \ldots, \big(\tilde{\mathbf{x}}^{(B)}_i\big)^T\big) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\tilde{x}_i}), \text{ with } \mathbf{C}_{\tilde{x}_i} = \text{blkdiag}\{\tilde{\sigma}^2_b \mathbf{I}_L\}_{b=1}^B. \tag{18}$$

[2] The variance $\tilde{\sigma}^2_b$ of (the i.i.d. entries of) the random vector $\tilde{\mathbf{x}}^{(b)}_i$ does not exceed the variance of (the i.i.d. entries of) the random vector $\sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}^{(b)}_j$ due to the (orthogonal) decomposition (13). The variance of $\sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}^{(b)}_j$ is, in turn, upper bounded by the variance of the random vector $\mathbf{x}^{(b)}_i$ due to the (orthogonal) decomposition (12).

The decompositions (12) and (13) naturally suggest a simple strategy for estimating (selecting) the neighbourhoods $\mathcal{N}(i)$ of the nodes $i \in \mathcal{V}$ in the CIG $\mathcal{G}$. To this end, let $\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \in \mathbb{R}^{L \times L}$ denote the orthogonal projection matrix for the complement of the subspace $\mathcal{X}^{(b)}_{\mathcal{T}} := \operatorname{span}\{\mathbf{x}^{(b)}_{j}\}_{j \in \mathcal{T}} \subseteq \mathbb{R}^{L}$, i.e.,

$$\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} := \mathbf{I} - \mathbf{P}^{(b)}_{\mathcal{T}}, \text{ with } \mathbf{P}^{(b)}_{\mathcal{T}} := \sum_{j=1}^{\dim \mathcal{X}^{(b)}_{\mathcal{T}}} \mathbf{u}^{(b)}_{j} \big(\mathbf{u}^{(b)}_{j}\big)^{T}, \tag{19}$$

with $\{\mathbf{u}^{(b)}_{j}\}_{j=1}^{\dim \mathcal{X}_{\mathcal{T}}}$ being an orthonormal basis for the subspace $\mathcal{X}^{(b)}_{\mathcal{T}} \subseteq \mathbb{R}^{L}$. The matrix $\mathbf{P}^{(b)}_{\mathcal{T}}$ in (19) is an orthogonal projection matrix on the subspace $\mathcal{X}^{(b)}_{\mathcal{T}}$.

According to (12), for any index set $\mathcal{T} \supseteq \mathcal{N}(i)$ (such that $\mathcal{N}(i) \setminus \mathcal{T} = \emptyset$),

$$\|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_{i}\|_{2}^{2} = \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \boldsymbol{\varepsilon}^{(b)}_{i}\|_{2}^{2} \text{ for all } b \in \{1, \ldots, B\}. \tag{20}$$

On the other hand, for any index set $\mathcal{T}$ with $\mathcal{N}(i) \setminus \mathcal{T} \neq \emptyset$, (13) entails

$$\|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_{i}\|_{2}^{2} = \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} (\tilde{\mathbf{x}}^{(b)}_{i} + \boldsymbol{\varepsilon}^{(b)}_{i})\|_{2}^{2} \text{ for all } b \in \{1, \ldots, B\}, \tag{21}$$

with some random vector $\tilde{\mathbf{x}}^{(b)}_{i}$. Some of our efforts go into showing that

$$\|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} (\tilde{\mathbf{x}}^{(b)}_{i} + \boldsymbol{\varepsilon}^{(b)}_{i})\|_{2}^{2} \approx \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \tilde{\mathbf{x}}^{(b)}_{i}\|_{2}^{2} + \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \boldsymbol{\varepsilon}^{(b)}_{i}\|_{2}^{2},$$

for all blocks $b \in \{1, \ldots, B\}$. Thus, according to (20) and (21), if the component $\tilde{\mathbf{x}}^{(b)}_{i}$ in (13) is not too small, the estimator

$$\widehat{\mathcal{N}}(i) := \arg \min_{|\mathcal{T}| \leq s} (1/N) \sum_{b=1}^{B} \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_{i}\|_{2}^{2} + \lambda |\mathcal{T}|, \tag{22}$$

delivers the true neighbourhood, i.e., $\widehat{\mathcal{N}}(i) = \mathcal{N}(i)$, with high probability. The penalty term $\lambda|\mathcal{T}|$ in (22) is required since we allow nodes $i \in \mathcal{V}$ in the CIG to potentially have fewer than s neighbours ($|\mathcal{N}(i)| < s$).[3] Indeed, the statistic $\|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_{i}\|_{2}^{2}$ does not allow to distinguish between different sets $\mathcal{T}$ which contain the neighborhood $\mathcal{N}(i)$. Therefore, we need to add the penalty term $\lambda|\mathcal{T}|$ in (22) in order to prefer smaller sets $\mathcal{T}$ as an estimate for $\mathcal{N}(i)$.

The estimator (22) performs sparse block-wise least squares regression by approximating the ith component $\mathbf{x}_{i}$ (cf. (6)) in a sparse manner (by allowing only s active components) using the remaining process components. Indeed, the summands $\|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_{i}\|_{2}^{2}$ in (22) are the errors obtained from the block-wise regression problems

$$\|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_{i}\|_{2}^{2} = \min_{\mathbf{w}^{(b)} \in \mathbb{R}^{p}, w^{(b)}_{i} = 0} \|\mathbf{x}^{(b)}_{i} - \sum_{j \in \mathcal{T}} w^{(b)}_{j} \mathbf{x}^{(b)}_{j}\|_{2}^{2}.$$

[3] In contrast to our approach (22), the analysis of GMS presented in (Wainwright, 2009a), for the special case of i.i.d. samples, requires all neighbourhoods $\mathcal{N}(i)$ to have exactly the same size s, i.e., $\mathcal{N}(i) = s$ for all $i \in \mathcal{V}$.

We highlight that the estimator (22) is mainly useful as a theoretical device which allows for a simple performance analysis and, in turn, a characterization of the required sample size for accurate GMS. Using a naive implementation of (22), by searching over all subsets of $\{1, \ldots, p\}$ with size at most s, has a complexity which grows exponentially in the sparsity level s. Thus, the estimator (22) is typically intractable except for very small sparsity levels s (corresponding to a very sparse CIG). More tractable methods for GMS can be obtained by using convex relaxations of (22) which result in Lasso-type methods (see (Danaher et al., 2014) and Section 4.1).

The estimator (22) itself only delivers an estimate for the neighbourhood of some node $i \in \mathcal{V}$ in the CIG $\mathcal{G}$ underlying the process (2). In order to obtain an estimate of the entire CIG, we have to repeatedly apply the estimator (22) to each node $i \in \mathcal{V}$. It might then happen that due to estimation errors, we obtain $i \in \widehat{\mathcal{N}}(j)$ but $j \notin \widehat{\mathcal{N}}(i)$ for two different nodes $i, j \in \mathcal{V}$. There are different options how to handle such a situation such as insisting in consistency between the neighbourhoods when declaring the presence of an edge (see (Meinshausen and Bühlmann, 2006, Eq. (7))). However, the implementation details for handling such cases are not relevant to our analysis, which aims at sufficient conditions such that (22) delivers the correct neighborhood for all nodes simultaneously (with high probability).

Our main result is an upper bound on the probability of the sparse neighbourhood regression (22) to fail in delivering the correct neighbourhood $\mathcal{N}(i)$, i.e., the error event

$$\mathcal{E}_i := \{\mathcal{N}(i) \neq \widehat{\mathcal{N}}(i)\}. \tag{23}$$

Theorem 1 *Consider the vector samples* $\mathbf{x}[n]$*, for* $n = 1, \ldots, N$*, conforming to the process model* (2) *and such that Assumption 1, 2 and 3 are valid. We estimate the neighbourhood* $\mathcal{N}(i)$ *of an arbitrary but fixed node* $i \in \mathcal{V}$ *in the CIG via sparse regression* (22) *with* $\lambda = \rho^2_{\rm min}/6$*. Then, if the average partial squared correlations between connected components are sufficiently large such that (see* (9)*)*

$$\rho^2_{\rm min} \geq 24\beta/L \tag{24}$$

for any sample size

$$N \geq 864(\beta/\rho^2_{\rm min}) \log(6ps^2/\eta), \tag{25}$$

the probability of the error event (23) *is bounded as* $\mathrm{P}\{\mathcal{E}_i\} \leq \eta$*.*

By Theorem 1, the true neighbourhood $\mathcal{N}(i)$ of a node $i \in \mathcal{V}$ can be recovered via (22) with high probability if the samples size N is on the order of $(\beta/\rho^2_{\rm min}) \log(ps^2)$ (for a fixed error tolerance η). Therefore, given sufficiently large computational power, GMS via sparse neighbourhood regression (22) is feasible in the high dimensional regime where $N \ll p$.

Since a CIG $\mathcal{G}$ is entirely determined by the neighbourhoods $\mathcal{N}(i)$ of all nodes $i \in \mathcal{V}$, we obtain the following result on GMS as a direct consequence of Theorem 1.

Corollary 2 *Consider a process* (2) *with underlying CIG $\mathcal{G}$ and satisfying all the assumptions in Theorem 1. Then, for any sample size*

$$N \geq 864(\beta/\rho_{\min}^2)\log(6p^2s^2/\eta), \tag{26}$$

there is a GMS method delivering a CIG estimate $\widehat{\mathcal{G}}$ with $\mathrm{P}\{\widehat{\mathcal{G}} \neq \mathcal{G}\} \leq \eta$.

Proof Using (22), we compute an estimate $\widehat{\mathcal{N}}(i)$ for each node $i \in \mathcal{V}$. Then, we construct a CIG estimate $\widehat{\mathcal{G}}$ having an edge $\{i,j\}$ between nodes $i,j \in \mathcal{V}$ when $j \in \widehat{N}(i)$ and $i \in \mathcal{N}(j)$. The estimate $\widehat{\mathcal{G}}$ is correct, i.e., $\widehat{\mathcal{G}} = \mathcal{G}$ whenever all of the estimates $\widehat{\mathcal{N}}(i)$ are correct, i.e., $\widehat{\mathcal{N}}(i) = \mathcal{N}(i)$ for each node $i \in \mathcal{V}$. The result then follows by combining Theorem 1 with a union bound (over all nodes $i \in \mathcal{V}$). ■

It turns out that the bound (26) is sharp since it matches a fundamental lower bound on the required sample size for any GMS method which performs uniformly well for any process of the form (2) and satisfying Assumption 1, (2) and 3. This lower bound follows directly from the results in (Hannak et al., 2014).

Lemma 3 *(Hannak et al., 2014, Theorem 3.1) Consider a GMS method which reads in vector samples* $\mathbf{x}[1], \ldots, \mathbf{x}[N]$ *(see* (2)*) and delivers an estimate $\widehat{\mathcal{G}}$ for the CIG $\mathcal{G}$ between the components $\mathbf{x}_i$, for $i = 1, \ldots, p$ (see* (6)*). If the method achieves an error probability* $\mathrm{P}\{\mathcal{E}_i\}$ *uniformly bounded by some prescribed error level η for any process of the form* (2) *satisfying Assumption 1, 2 and 3 with $\rho_{\min}^2 \leq 1/4$, then the sample size must necessarily satisfy $N > \frac{\log\binom{p}{2}-1}{4\rho_{\min}^2}$.*

Combining Theorem 1 with Lemma 3, we conclude that the bound (26) characterizes, up to a constant factor, the minimum required sample size for accurate GMS based on processes of the form (2).

It is instructive to compare the sufficient condition (26) on the sample size N with the results obtained in (Ravikumar et al., 2011; Wainwright, 2009a; Wang et al., 2010) for the special case of i.i.d. samples, which coincides with the model (2) for $B=1$ and $N=L$. We note that for this special case, the bound (26) matches the necessary condition on sample size derived in (Wang et al., 2010), which confirms the sparse regression method (22) to be optimal in terms of sample size requirement. However, this is already certified by Lemma 3, which is extends the results of (Wang et al., 2010) to non-stationary processes (2).

At first sight it appears that the bound (26) suggests a smaller required sample size compared to the bound $N \propto s^2 \log p$ obtained in (Ravikumar et al., 2011, Corollary 1). However, it is important to note that the lower bound $\rho_{\min}^2$ (see (9)) on the minimum partial squared correlations $\rho_{i,j}^2$ (between connected components) cannot be chosen arbitrarily in order to have at least one process (2) satisfying Assumption 1. In particular the off-diagonal entries $K_{i,j}[n]$ of the precision matrices cannot take on arbitrary (large) values, since the precision matrix $\mathbf{K}[n] = \big(\mathbf{C}[n]\big)^{-1}$ (see (1)) must be positive definite.

A practically relevant regime for the minimum squared partial correlation is $\rho_{\min}^2 \leq c/s$ with some constant c which may depend on β (see (11)). For this regime, which is also considered in (Wang et al., 2010), the bound (26) becomes $N \propto s^2 \log p$ which closely resembles the sample size requirement for the convex GMS method in (Ravikumar et al., 2011).

Finally, we note that Theorem 1 does not involve some incoherence condition, which requires sub-matrices of the covariance matrices $\mathbf{C}^{(b)}$ (see (3)) to be well-conditioned. Such incoherence conditions are typically required by convex relaxations of the sparse regression estimator (22). While convex (Lasso-based) methods are computationally more tractable than non-convex estimators such as (22), convex methods place more stringent conditions (such as some incoherence condition) on the process (2) in order to guarantee accurate estimation of the underlying CIG (Wainwright, 2009b; Loh and Wainwright, 2017).

4. Numerical Results

We verify the predictions of Theorem 1 by means of numerical experiments involving synthetic data (see Section 4.1) and data collected by pedestrian count devices located in the city of Turku in Finland (see Section 4.2). We also compare our results with the empirical performance obtained from computationally efficient convex optimization methods (see Section 4.1). In order to support reproducible research, we have made the source code for our experiments available under `https://github.com/alexjungaalto/ResearchPublic/tree/master/GMSNonStat`.

4.1 Chain

Our first experiment revolves around a synthetic dataset $\mathbf{x}[n]$ which is generated according to the process model (2) such that the true underlying CIG $\mathcal{G}$ is a chain graph as depicted in Figure 2.

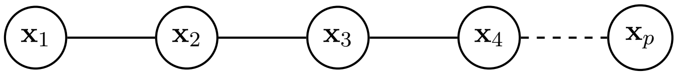


Figure 2: The CIG of a process (2) with a chain structure.

In particular, we generated Gaussian random vectors conforming to the process model (2) with $B = 4$ blocks. The b-th block consists of L i.i.d. random vectors $\mathbf{x}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(b)})$ with $\mathbf{C}^{(b)}$ being chosen such that the marginal CIG $\mathcal{G}^{(b)}$ is a chain (see Figure 2) with the edge $\{b, b+1\}$ missing (see Figure 3).

In order to estimate the neighbourhood $\mathcal{N}(i)$ of a given node $i \in \mathcal{V}$ in the CIG from the generated vector samples, we use the sparse regression estimator (22) with $s = 2$. For such a small sparsity, It is still feasible to compute the estimator (22) by exhaustive search over all subsets of size at most $s = 2$. However, for larger values of s the estimator (22) becomes intractable and one has to use computationally cheaper methods such as convex optimization methods (Danaher et al., 2014; Boyd et al., 2010).

We estimate the error probability (23) by an empirical average $\widehat{\mathrm{P}}\{\mathcal{E}_i\}$ over $K = 100$ i.i.d. simulation runs. In particular, using the j-th realization of the process (2) as input to the sparse regression estimator (22) in order to obtain the estimate $\widehat{\mathcal{N}}^{(j)}(i)$, we compute the

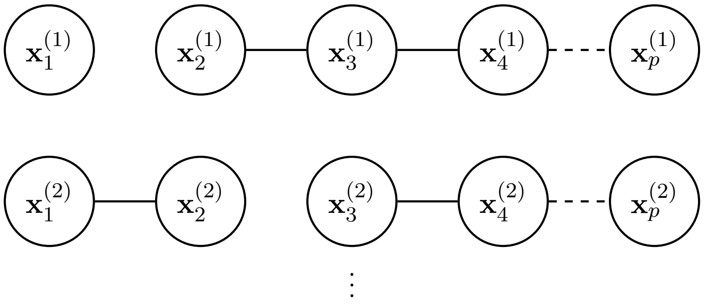
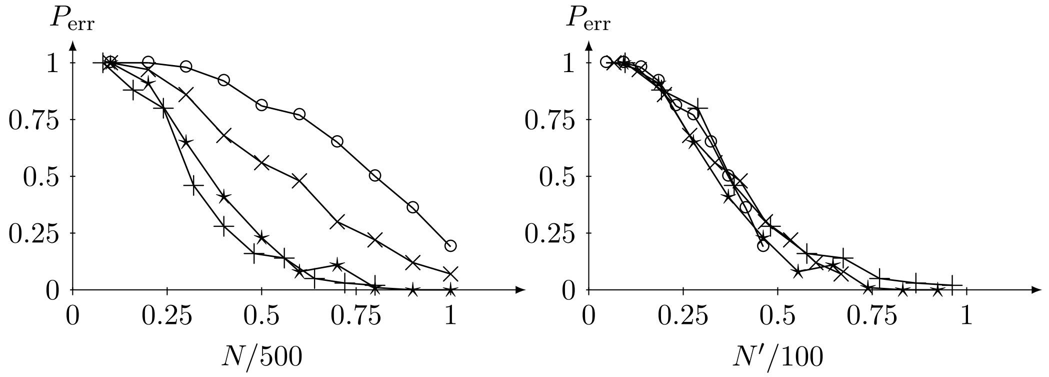


Figure 3: The marginal CIG $\mathcal{G}^{(b)}$ underlying the b-th block, constituted by the i.i.d. samples $\mathbf{x}[(b-1)L+1], \ldots, \mathbf{x}[bL]$, is a chain with the edge $\{b, b+1\}$ removed.

empirical error rate

$$\widehat{\mathrm{P}}\{\mathcal{E}_i\} := (1/K) \sum_{j=1}^{K} \mathcal{I}(\widehat{\mathcal{N}}^{(j)}(i) \neq \mathcal{N}(i)) (\approx \mathrm{P}\{\mathcal{E}_i\}). \tag{27}$$

Here, $\hat{\mathcal{N}}^{(j)}(i)$ is the estimated neighbourhood of node $i \in \mathcal{V}$ during the j-th simulation run.



(a)

(b)

Figure 4: Empirical error rate $P_{\text{err}} = \widehat{\mathrm{P}}\{\widehat{\mathcal{N}}(2) \neq \mathcal{N}(2)\}$ (see (27)) incurred by (22) when estimating the neighbourhood $\mathcal{N}(2)$ of node $i = 2$ in a chain CIG with $p = 64$ ("+"), $p = 128$ ("×"), $p = 256$ ("○") and $p = 512$ ("⋆"). (a) Error rate as a function of the sample size N. (b) Error rate as a function of scaled sample size $N' = N/\log p$. Error rate has been obtained using $K = 100$ simulation runs.

In Figure 4-(a) we depict the error rate $\widehat{\mathrm{P}}\{\mathcal{E}_i\}$, achieved by the estimator (2) when estimating the neighbourhood or node $i = 2$ (see Figure 2), as a function of the sample size N. The three curves in Figure 4-(a) corresponds to three different processes. Each process is of the form (2) with CIG being a chain (see Figure 2), but with different p and $\rho^2_{\min}$ (see (9)).

As indicated by the upper bound (25) of Theorem 1, the error rate $\widehat{\mathrm{P}}\{\mathcal{E}_i\}$ (see (27)) crucially depends on the scaled sample size $N' := N\rho^2_{\min}/\log p$. Therefore, we plot in Figure 4-(b) the error rate $\widehat{\mathrm{P}}\{\mathcal{E}_i\}$ as a function of the scaled sample size N'. In agreement with our theoretical findings, we observe that the curves in Figure 4-(b) are almost lying on top of each other.

The sparse regression estimator (22) implements a form of pooling of the samples $\mathbf{x}[1], \ldots, \mathbf{x}[N]$ across different blocks. Indeed, the objective function in (22) sums up the contributions from all blocks such that the required sample size depends on the average squared partial correlations (8). A simple alternative approach would be to consider the samples of each block in (2) as i.i.d. samples from a marginal CIG $\mathcal{G}^{(b)}$ and apply existing GMS methods for i.i.d. samples to obtain estimates for the marginal CIGs. We can then obtain an estimate for the global CIG $\mathcal{G}$ by using the union of the edge sets in each marginal CIG $\mathcal{G}^{(b)}$.

Since the samples $\mathbf{x}[n]$ of the process (2) are i.i.d. within each block (constituted by L consecutive samples), we can estimate the marginal (block-wise) CIGs $\mathcal{G}^{(b)}$ separately using any of the established GMS methods for i.i.d. samples. An estimate for the (global) CIG $\mathcal{G}$ underlying the process (2) can then be obtained via the union of the (edges in the) individual CIG estimates $\widehat{\mathcal{G}}^{(b)}$. More precisely, a “naive” estimate $\widehat{\mathcal{N}}^{(\mathrm{naive})}(i)$ for the neighborhood $\mathcal{N}(i)$ of some node $i \in \mathcal{V}$ can be obtained from the union of the block-wise neighborhood estimates $\widehat{\mathcal{N}}^{(b)}(i)$, for $b=1, \ldots, B$.

In Figure 5, we compare the error rate achieved by our pooled approach to this naive approach. In particular, we use a constrained ℓ_1 minimization approach (referred to as “CLIME”) to estimate the support of the sparse precision matrix (Cai et al., 2011) within each block. From Figure 5 we obtain that the pooled approach (22) clearly outperforms the naive approach. This result should not come as a surprise since the pooled estimator (22) allows to cope with few blocks with very small partial correlations (of connected nodes in the CIG) as long as the average squared partial correlation (see (8)) is large enough. In contrast, the naive approach is likely to fail if there is at least one block of samples which does not allow accurate GMS.

As pointed out in Section 3, the sparse regression method (22) becomes intractable except for very small number p of process components in (2) and sparsity level s of the underlying CIG (see Assumption 2). A computationally more tractable GMS method can be obtained by replacing (relaxing) the non-convex penalty term $\lambda|\mathcal{T}|$ in (22) by a convex approximation. The group Lasso is obtained by a particular choice for this convex approximation as (Bach, 2008)

$$\hat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w}^{(b)} \in \mathbb{R}^p, w_i^{(b)}=0} \sum_{b=1}^{B} \|\mathbf{x}_i^{(b)} - \sum_{j=1}^{p} w_j^{(b)} \mathbf{x}_j^{(b)}\|_2^2 + \lambda \sum_{j=1}^{p} \|\mathbf{w}_j\|_2. \tag{28}$$

In order to obtain an estimate for the neighbourhood $\mathcal{N}(i)$ from the estimator (28), we threshold the squared block norms $\|\hat{\mathbf{w}}_j\|_2^2 = \sum_{b=1}^{B} \big(\hat{w}_j^{(b)}\big)^2$ at the level $\eta = \rho^2_{\min}/2$ to obtain

$$\widehat{\mathcal{N}}^{(\mathrm{gLasso})}(i) := \{j \in \{1, \ldots, p\} \setminus \{i\} : \|\hat{\mathbf{w}}_j\|_2^2 \geq \eta\}. \tag{29}$$

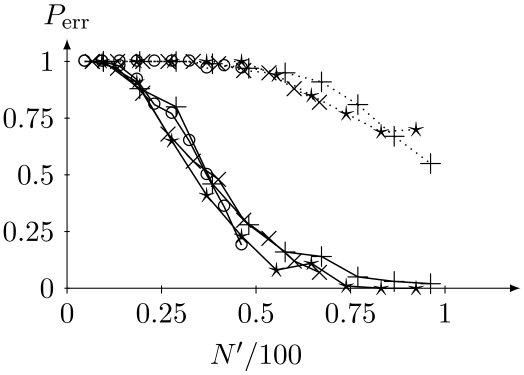


Figure 5: Error rate (27) achieved by estimating the neighborhood $\mathcal{N}(2)$ using the pooled estimator (22) (solid curves) and by the union of the neighbourhoods obtained by applying CLIME (Cai et al., 2011) to each block in (2) separately and then forming a union (over all blocks $b = 1, \ldots, B$) of all block-wise neighbourhood estimates (dotted curves),

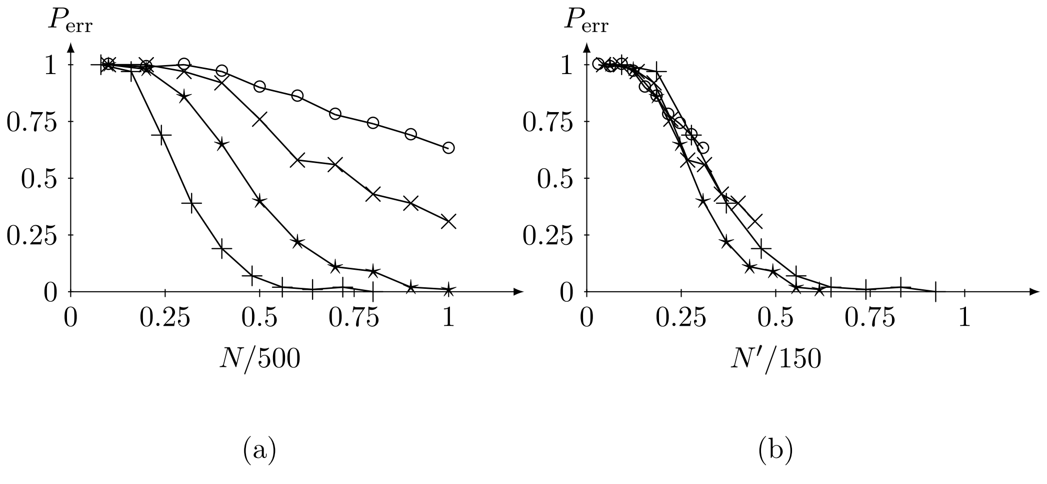


Figure 6: Empirical error rate $P_{\text{err}} = \widehat{\mathrm{P}}\{\widehat{\mathcal{N}}^{(\text{gLasso})}(2) \neq \mathcal{N}(2)\}$ (see (27)) incurred by the neighborhood estimate (29) applied to a process with chain structured CIG of size $p = 64$ ("+"), $p = 128$ ("×"), $p = 256$ ("∘") and $p = 512$ ("⋆"). (a) Error rate as a function of the sample size N. (b) Error rate as a function of scaled sample size $N' = N/\log p$. Error rate has been obtained using $K = 100$ simulation runs.

In Figure 6, we depict the error rate incurred by (29) for a process with chain-structured CIG (see Figure 2). While Figure 6-(a) shows the error rate as a function of the original sample size N, Figure 6-(b) displays the error rate as a function of the scaled sampled size $N' = N\rho^2_{\rm min}/\log p$. In agreement with our theoretical findings (see Theorem 1), the error rate of the estimator (29) seems to be mainly determined by the scaled sample size N' as indicated in Figure 6-(b).

4.2 Pedestrian Counts

In this experiment we applied the sparse regression estimator (22) to hourly pedestrian counts collected in the city of Turku (Finland). The city operates pedestrian counting devices at certain locations in the city center (see Figure 7). The counting devices (based on cameras) measure the number of pedestrians which pass one of two counting lines in a certain direction (see Figure 7).

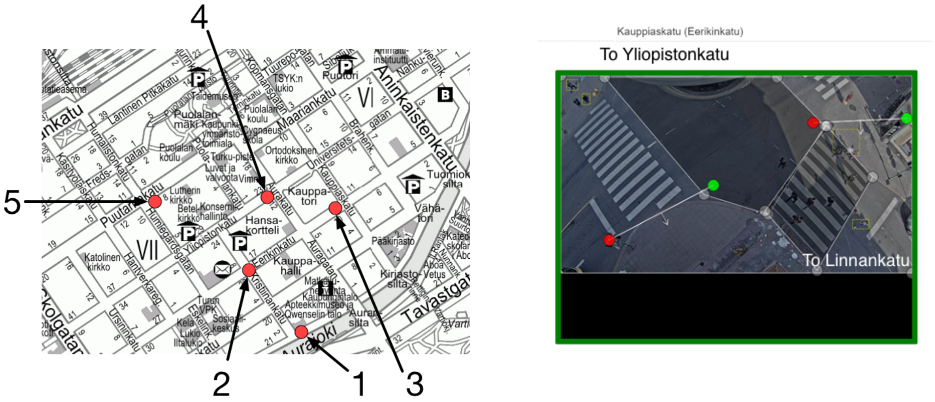


Figure 7: Left: Map of Turku city including the locations of pedestrian count devices (depicted as red dots). Right: Snapshot generated by counting device "3" in order to count the number of pedestrians crossing each of two (virtual) counting lines in a particular direction.

We have been provided with hourly count data obtained from $p = 5$ different counting devices located in the city center of Turku (see Figure 7) and collected since 23rd of July, 2018. For each counting device, we compute the average count $z^{(i)}[n]$, for $i=1,\ldots,p$ at time n. We depict the average count $z^{(1)}[n]$ in Figure 8, which indicates a seasonal component with period 24. This is not too surprising as we expect the pedestrian movements for different days to be similar for the same daytime.

In order to remove the seasonal component we difference the time series $z^{(j)}[n]$ at lag 24 to obtain the time series (see (Brockwell and Davis, 1991, Chapter 1.4))

$$\tilde{z}^{(i)}[n] := z^{(i)}[n+24] - z^{(i)}[n] \text{ for } i = 1, \ldots, p. \tag{30}$$

We depict the time series $\tilde{z}^{(i)}[n]$ in Figure 9, which suggests that is is reasonable to model $\tilde{z}^{(i)}[n]$ as a stationary timer series (or discrete time process).

As discussed in Section 2, we can transform a stationary process into a process conforming to our non-stationary model (2) by applying a DFT. We compute the DFT of the difference

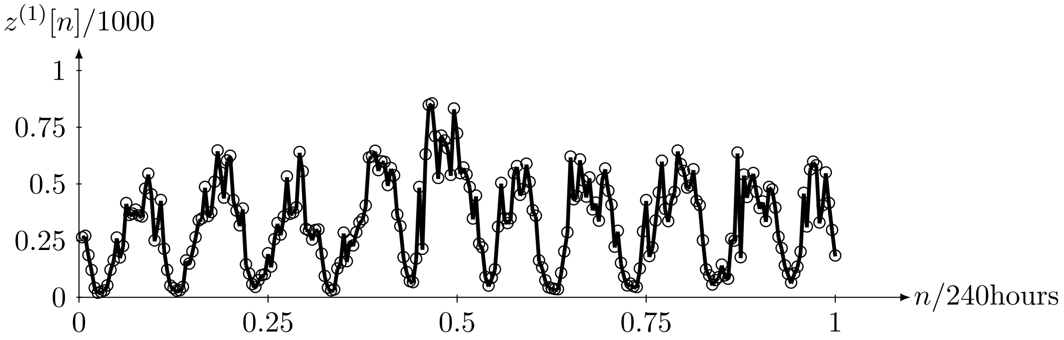


Figure 8: Hourly pedestrian counts (averaged over two counting lines) at location $j = 1$ as indicated in Figure 7.

time series $\tilde{z}^{(i)}[n]$ (see (30)) using a period $N = 3072$ to obtain the vector-valued samples

$$\mathbf{x}[n] = \big(x^{(1)}[n], \ldots, x^{(p)}[n]\big)^T, \text{ with } x^{(i)}[n] := \sum_{n'=1}^{N} \tilde{z}^{(i)}[n'] \exp\big(-2\pi\iota(n'-1)(n-1)/N\big) \quad (31)$$

for $n = 1, \ldots, N$. Here, $\iota := \sqrt{-1}$ denotes the imaginary unit.

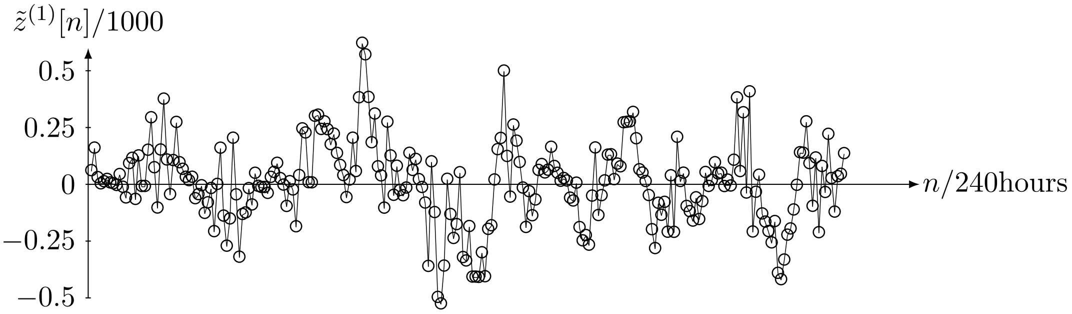


Figure 9: Differenced (at lag 24) hourly pedestrian counts $\tilde{z}^{(1)}[n]$ (see (30)) at location $j = 1$ which is indicated in Figure 7.

We model the samples $\mathbf{x}[n]$ using (2) with a block-length $L = 12$ which has been chosen based on the empirical autocorrelation functions of the differenced time series $\tilde{z}^{(i)}$ (see (30)). In order to infer the neighbourhoods $\mathcal{N}(i)$ in the CIG underlying the count measurements, we compute the test statistic

$$Z(\mathcal{T}) := (1/N) \sum_{b=1}^{B} \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}} \mathbf{x}^{(b)}_i\|_2^2$$

$$= (1/N) \sum_{b=1}^{B} \min_{w_j^{(b)} \in \mathbb{R}} \big\| \mathbf{x}_i^{(b)} - \sum_{j \in \mathcal{T}} w_j^{(b)} \mathbf{x}_j^{(b)} \big\|_2^2, \tag{32}$$

with the DFT samples (31) and varying candidate sets $\mathcal{T} \subseteq \{1, \ldots, p\} \setminus \{i\}$. [4]

Since we neither know the maximum node degree (sparsity) s, nor a lower bound $\rho^2_{\rm min}$ on the average squared partial correlations, we cannot directly implement the sparse regression estimator (22). Instead, we try to estimate the neighborhood $\mathcal{N}(i)$ of node $i \in \mathcal{V}$ by evaluating the decay of the score $\mathcal{E}(s) := \min_{|\mathcal{T}|=s} Z(\mathcal{T})$ using the statistic (32) (which is the first component in the objective function of the sparse regression estimator (22)).

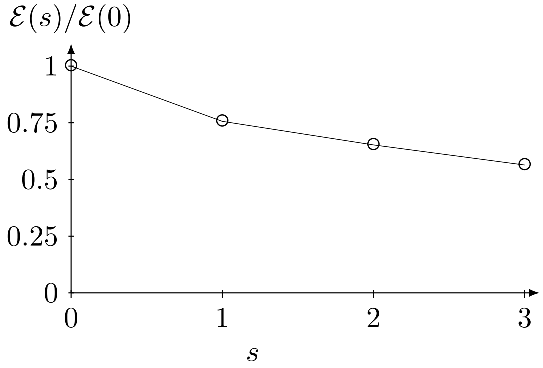


Figure 10: Score $\mathcal{E}(s) = \min_{|\mathcal{T}|=s} Z(\mathcal{T})$ achieved by minimizing the statistic (32) (for node $i=1$) over all candidate sets $\mathcal{T}$ with a prescribed size s.

In Figure 10, we depict the score $\mathcal{E}(s)$ obtained for node $i = 2$. We then choose the neighborhood size s as the smallest number such that $\mathcal{E}(s) - \mathcal{E}(s+1) < 2(\mathcal{E}(s) - \widetilde{\mathcal{E}}(s))$ with the "auxiliary score"

$$\widetilde{\mathcal{E}}(s) = (1/N) \sum_{b=1}^{B} \min_{w_j^{(b)}} \big\| \mathbf{x}_i^{(b)} - \big(\sum_{j \in \mathcal{T}'} w_j^{(b)} \mathbf{x}_j^{(b)} + c\mathbf{f}^{(b)} \big) \big\|_2^2 \tag{33}$$

Here, the index set $\mathcal{T}'$ is chosen as $\mathcal{T}' = \arg\min_{|\mathcal{T}|=s} Z(\mathcal{T})$.

The idea behind comparing $\mathcal{E}(s)$ with $\widetilde{\mathcal{E}}(s)$ is to test if adding another process component to the components in $\mathcal{T}'$ yields a reduction in the statistic $\mathcal{E}(s+1)$ which is at least twice as large as the reduction of $\mathcal{E}(s)$ achieved by adding a "fake" pedestrian count signal obtained by i.i.d. uniformly distributed random variables $f[n] \sim \mathcal{U}[0, U]$. The interval size U is chosen in order to match the empirical variance of the pedestrian counts $z^{(i)}[n]$.

[4] While our analysis applies only to real-valued vector samples $\mathbf{x}[n]$ in (2), the vector samples (31) obtained from a DFT are typically complex-valued. However, we expect our analysis to also apply to complex-valued samples in (31) by applying straightforward modifications of our methods. In particular, we believe that the fundamental dependencies (see (26)) between required sample size N on number p of process components, sparsity s and average partial squared correlations $\rho^2_{\rm min}$ to remain valid when allowing the samples in (2) to be complex-valued Gaussian vectors.

We have obtained the following estimates for the neighbourhoods in the CIG underlying the pedestrian count data:

$$\widehat{\mathcal{N}}(1)=\{2\},\widehat{\mathcal{N}}(2)=\{3,5\},\widehat{\mathcal{N}}(3)=\{2,4\},\widehat{\mathcal{N}}(4)=\{3,5\},\widehat{\mathcal{N}}(5)=\{2,4\}. \tag{34}$$

In Figure 11, we depicted the CIG estimate obtained by placing an edge between nodes $i, j \in \mathcal{V}$ if either $j \in \widehat{\mathcal{N}}(i)$ or $i \in \widehat{\mathcal{N}}(j)$. The estimated graph structure seems well-aligned with the local road network.

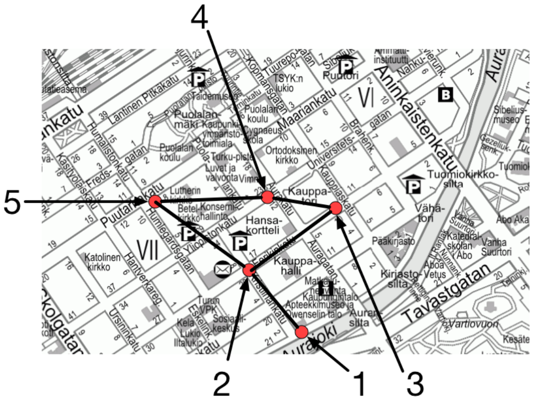


Figure 11: Map of Turku city including the locations of pedestrian count devices (depicted as red dots). The links between the count devices indicate the presence of an edge in the estimated CIG underlying the count data.

Acknowledgement

The authors are grateful for the comments received from the reviewers and the action editor. We also thank Tuomas Piippo and Arvi Leino from the city of Turku for help with the pedestrian count data.

5. Proof of the Main Result

We now verify Theorem 1 by analyzing the probability $\mathrm{P}\{\mathcal{E}_i\}$ of the error event $\mathcal{E}_i$ (see (23)) when (22) fails to deliver the correct neighbourhood $\mathcal{N}(i)$ of a particular node $i \in \mathcal{V}$ of the CIG $\mathcal{G}$. Let us introduce the shorthands

$$\mathcal{E}_{\mathcal{T}} := \{Z(\mathcal{N}(i)) + \lambda s_i > Z(\mathcal{T}) + \lambda|\mathcal{T}|\}, \text{ with } Z(\mathcal{T}) := \frac{1}{N}\sum_{b=1}^{B} \|\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}}\mathbf{x}^{(b)}_i\|_2^2. \tag{35}$$

It will be convenient to denote the set of all subsets of $\{1, \ldots, p\}$ of size at most s but different from the true neighbourhood $\mathcal{N}(i)$ by

$$\Sigma_s^p := \{\mathcal{T} \subseteq \{1, \ldots, p\} : |\mathcal{T}| \leq s, \mathcal{T} \neq \mathcal{N}(i)\}.$$

Moreover, for given $\ell_1, t \leq s$, denote

$$\mathcal{N}(\ell_1, t) := \{\mathcal{T} \in \Sigma_s^p : |\mathcal{T}| = t, |\mathcal{N}(i) \setminus \mathcal{T}| = \ell_1\}. \tag{36}$$

Thus, the set $\mathcal{N}(\ell_1, t) \subseteq \Sigma_s^p$ collects all the index sets in Σ_s^p with a prescribed size $t = |\mathcal{T}|$ and overlap $\ell_1 = |\mathcal{N}(i) \setminus \mathcal{T}|$ with the true neighbourhood $\mathcal{N}(i)$.

An elementary combinatorial argument (see (Wainwright, 2009a, Sec. IV)) reveals that the number of these index sets is

$$N(\ell_1, t) := |\mathcal{N}(\ell_1, t)| = \binom{s_i}{\ell_1}\binom{p - s_i}{\ell_2}. \tag{37}$$

with

$$\ell_2 := \ell_1 + (t - s_i). \tag{38}$$

Given a particular node $i \in \mathcal{V}$ with neighbourhood $\mathcal{N}(i)$, the quantities ℓ_1 and ℓ_2 are fully determined by the index set $\mathcal{T}$. For notational convenience we will not make this dependence on $\mathcal{T}$ explicit, i.e., we write ℓ_1 and ℓ_2 instead of $\ell_1(\mathcal{T})$ and $\ell_2(\mathcal{T})$. Note that

$$\ell_2 = |\mathcal{T} \setminus \mathcal{N}(i)| \text{ and } \ell_1 + \ell_2 > 0 \text{ for every index set } \mathcal{T} \in \mathcal{N}(\ell_1, t). \tag{39}$$

Using the index set

$$\mathcal{I} := \{(\ell_1, t) \in \mathbb{Z}_+^2 : \ell_1 \leq s_i, t \leq s\} \setminus \{(0, s_i)\} \text{ with cardinality } |\mathcal{I}| \leq s^2, \tag{40}$$

we can write

$$\Sigma_s^p \subseteq \bigcup_{(\ell_1, t) \in \mathcal{I}} \mathcal{N}(\ell_1, t). \tag{41}$$

Since the error event $\mathcal{E}_i$ (see (23)) can only occur if at least one of the events $\mathcal{E}_{\mathcal{T}}$, for some $\mathcal{T} \in \Sigma_s^p$, occurs,

$$\mathcal{E}_i \subseteq \bigcup_{\mathcal{T} \in \Sigma_s^p} \mathcal{E}_{\mathcal{T}}, \tag{42}$$

implying, in turn via a union bound,

$$\mathrm{P}\{\mathcal{E}_i\} \overset{(42)}{\leq} \sum_{\mathcal{T} \in \Sigma_s^p} \mathrm{P}\{\mathcal{E}_{\mathcal{T}}\} \overset{(41)}{\leq} \sum_{(\ell_1,t)\in\mathcal{I}} \sum_{\mathcal{T}\in\mathcal{N}(\ell_1,t)} \mathrm{P}\{\mathcal{E}_{\mathcal{T}}\}. \tag{43}$$

We now derive an upper bound $M(\ell_1,t)$ on the individual probabilities $\mathrm{P}\{\mathcal{E}_{\mathcal{T}}\}$ such that

$$\mathrm{P}\{\mathcal{E}_{\mathcal{T}}\} \leq M(\ell_1,t) \text{ for any } \mathcal{T} \in \mathcal{N}(\ell_1,t). \tag{44}$$

As the notation already indicates, the upper bound $M(\ell_1,t)$ depends on the index set $\mathcal{T}$ only via the overlap $\ell_1 = |\mathcal{N}(i) \setminus \mathcal{T}|$ and the size $t = |\mathcal{T}|$.

Combining (41) with (43) implies, via a union bound,

$$\begin{aligned}
\log \mathrm{P}\{\mathcal{E}_i\} &\overset{(43)}{\leq} \log \sum_{(\ell_1,t)\in\mathcal{I}} \sum_{\mathcal{T}\in\mathcal{N}(\ell_1,t)} \mathrm{P}\{\mathcal{E}_{\mathcal{T}}\} \\
&\overset{(44)}{\leq} \log \sum_{(\ell_1,t)\in\mathcal{I}} \sum_{\mathcal{T}\in\mathcal{N}(\ell_1,t)} M(\ell_1,t) \\
&\leq \log|\mathcal{I}| + \max_{(\ell_1,t)\in\mathcal{I}} \big[\log N(\ell_1,t) + \log M(\ell_1,t)\big] \\
&\overset{(37),(40)}{\leq} 2\log s + \max_{(\ell_1,t)\in\mathcal{I}} \big[\ell_1 \log s_i + \ell_2 \log(p-s_i) + \log M(\ell_1,t)\big] \\
&\leq 2\log s + \max_{(\ell_1,t)\in\mathcal{I}} \big[(\ell_1+\ell_2)\log p + \log M(\ell_1,t)\big].
\end{aligned} \tag{45}$$

Our next goal is to find a sufficiently tight upper bound $M(\ell_1,t)$ on the probabilities of the events $\mathrm{P}\{\mathcal{E}_{\mathcal{T}}\}$ (see (35)) with some index set $\mathcal{T} \in \mathcal{N}(\ell_1,t)$. To this end, we make (12) more handy by stacking the (block-wise) noise vectors $\boldsymbol{\varepsilon}_i^{(b)} \in \mathbb{R}^L$ into the single noise vector

$$\boldsymbol{\varepsilon}_i = \big((\boldsymbol{\varepsilon}_i^{(1)})^T, \ldots, (\boldsymbol{\varepsilon}_i^{(B)})^T\big)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varepsilon}_i}) \text{ , with } \mathbf{C}_{\boldsymbol{\varepsilon}_i} = \mathrm{blkdiag}\{(1/K_{i,i}^{(b)})\mathbf{I}_L\}_{b=1}^B. \tag{46}$$

By introducing the projection matrix

$$\mathbf{P}_{\mathcal{T}^\perp} := \mathrm{blkdiag}\{\mathbf{P}_{\mathcal{T}^\perp}^{(b)}\}_{b=1}^B, \tag{47}$$

we can characterize the error event $\mathcal{E}_{\mathcal{T}}$ in (35), for any $\mathcal{T} \in \mathcal{N}(\ell_1,t)$ (see (36)), as

$$\mathcal{E}_{\mathcal{T}} = \big\{Z(\mathcal{N}(i)) - (1/N)\|\mathbf{P}_{\mathcal{T}^\perp}\boldsymbol{\varepsilon}_i\|_2^2 > Z(\mathcal{T}) - (1/N)\|\mathbf{P}_{\mathcal{T}^\perp}\boldsymbol{\varepsilon}_i\|_2^2 + \lambda(t-s_i)\big\}. \tag{48}$$

In order to derive the upper bound $M(\ell_1, t)$ let us, for some number $\delta > 0$ whose precise value to be chosen in what follows, define the two error events

$$\mathcal{E}_1(\delta) := \big\{ Z(\mathcal{N}(i)) - (1/N)\|\mathbf{P}_{\mathcal{T}^\perp}\boldsymbol{\varepsilon}_i\|_2^2 \geq \delta + (\lambda/2)(t - s_i) \big\}, \tag{49a}$$

$$\mathcal{E}_2(\delta) := \big\{ Z(\mathcal{T}) - (1/N)\|\mathbf{P}_{\mathcal{T}^\perp}\boldsymbol{\varepsilon}_i\|_2^2 + (\lambda/2)(t - s_i) \leq 2\delta \big\}. \tag{49b}$$

By (48), an error $\mathcal{E}_{\mathcal{T}}$ can only occur if either $\mathcal{E}_1(\delta)$ or $\mathcal{E}_2(\delta)$ occurs, i.e., $\mathcal{E}_{\mathcal{T}} \subseteq \mathcal{E}_1(\delta) \cup \mathcal{E}_2(\delta)$. Therefore, by a union bound,

$$\begin{aligned} \mathrm{P}\{\mathcal{E}_{\mathcal{T}}\} &\leq \mathrm{P}\{\mathcal{E}_1(\delta)\} + \mathrm{P}\{\mathcal{E}_2(\delta)\} \\ &= \mathrm{E}\big\{\mathrm{P}\{\mathcal{E}_1(\delta)|\mathbf{x}_{\mathcal{T}}\}\big\} + \mathrm{E}\big\{\mathrm{P}\{\mathcal{E}_2(\delta)|\mathbf{x}_{\mathcal{T}}\}\big\}, \end{aligned} \tag{50}$$

where we condition on the components $\mathbf{x}_{\mathcal{T}} = \{\mathbf{x}_i\}_{i \in \mathcal{T}}$ (cf. (6)).

We will now bound each of the two summands in (50) separately. To this end, we will use the singular value decomposition (SVD)

$$\mathbf{P}_{\mathcal{T}^\perp}\mathbf{C}_{\tilde{x}_i}^{1/2} = \mathbf{U}\mathrm{diag}\{d_j\}_{j=1}^{N}\mathbf{V}^T \tag{51}$$

with the singular values $d_j \in \mathbb{R}_+$ and the singular vectors in the columns of the orthonormal matrices $\mathbf{U} \in \mathbb{R}^{N \times N}$ and $\mathbf{V} \in \mathbb{R}^{N \times N}$ (i.e., $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I}$). The singular values d_j, which satisfy

$$d_j \overset{(14),(17)}{\leq} \sqrt{\beta} \tag{52}$$

will play a prominent role in controlling the probabilities of the error events $\mathcal{E}_1(\delta)$ and $\mathcal{E}_2(\delta)$ (see (49a), (49b)). In particular, we will analyze the probabilities of those events for the choice $\delta = m_3/4$ with

$$\begin{aligned} m_3 &:= \mathrm{E}\{(1/N)\|\mathbf{P}_{\mathcal{T}^\perp}\tilde{\mathbf{x}}_i\|_2^2 \mid \mathbf{x}_{\mathcal{T}}\} \\ &\overset{(a)}{=} (1/N)\,\mathrm{tr}\{\mathbf{C}_{\tilde{x}_i}^{1/2}\mathbf{P}_{\mathcal{T}^\perp}\mathbf{C}_{\tilde{x}_i}^{1/2}\} \\ &\overset{(51)}{=} (1/N)\sum_{j=1}^{N} d_j^2, \end{aligned} \tag{53}$$

where in step (a) we used the statistical independence of $\tilde{\mathbf{x}}_i$ and $\mathbf{x}_{\mathcal{T}}$ (cf. (13)).

The quantity m_3 measures the minimum achievable error when approximating the process component $\mathbf{x}_i$ (see (6)) using a linear combination of the process components $\mathbf{x}_{\mathcal{T}} = \{\mathbf{x}_j\}_{j \in \mathcal{T}}$. A lower bound on m_3 can be obtained via the minimum average squared partial correlation ρ^2_{min} (see Assumption 1). Indeed,

$$\begin{aligned} m_3 &\overset{(53)}{=} (1/N)\,\mathrm{tr}\{\mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2}\mathbf{P}_{\mathcal{T}^\perp}\mathbf{C}_{\tilde{\mathbf{x}}_i}^{1/2}\} \\ &= (1/N)\,\mathrm{tr}\{\mathbf{C}_{\tilde{\mathbf{x}}_i}\mathbf{P}_{\mathcal{T}^\perp}\} \end{aligned}$$

$$\overset{(18)}{=} (1/N)\sum_{b=1}^{B}\operatorname{tr}\left\{\mathbf{P}^{(b)}_{\mathcal{T}^{\perp}}\tilde{\sigma}_b^2\mathbf{I}\right\}$$

$$\overset{(19)}{=} (1/N)\sum_{b=1}^{B}\tilde{\sigma}_b^2(L-|\mathcal{T}|). \tag{54}$$

This can be further developed by using the lower bound (16) for the variance $\tilde{\sigma}_b^2$,

$$\begin{aligned} m_3 &\overset{(54),(16)}{\geq} \sum_{j\in\mathcal{N}(i)\setminus\mathcal{T}}\sum_{b=1}^{B}(K^{(b)}_{i,j}/K^{(b)}_{i,i})^2(L-|\mathcal{T}|)/N \\ &\overset{(8),(9)}{\geq} \ell_1 B\rho^2_{\mathrm{min}}(L-|\mathcal{T}|)/N \\ &\overset{(10)}{\geq} (2/3)\ell_1\rho^2_{\mathrm{min}}. \end{aligned} \tag{55}$$

For the choice $\delta = m_3/4$ this implies, in turn,

$$\delta \geq (1/6)\ell_1\rho^2_{\mathrm{min}}. \tag{56}$$

In order to upper bound the probability of the event $\mathcal{E}_1(\delta)$, observe

$$\begin{aligned} Z(\mathcal{N}(i)) &\overset{(35)}{=} (1/N)\sum_{b=1}^{B}\|\mathbf{P}^{(b)}_{\mathcal{N}(i)^{\perp}}\mathbf{x}^{(b)}_i\|_2^2 \\ &\overset{(12)}{=} (1/N)\sum_{b=1}^{B}\left\|\mathbf{P}^{(b)}_{\mathcal{N}(i)^{\perp}}\big(\sum_{i\in\mathcal{N}(i)}a_j\mathbf{x}^{(b)}_j+\boldsymbol{\varepsilon}^{(b)}_i\big)\right\|_2^2 \\ &\overset{(47),(46)}{=} (1/N)\|\mathbf{P}_{\mathcal{N}(i)^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2. \end{aligned} \tag{57}$$

Hence,

$$\begin{aligned} \mathrm{P}\{\mathcal{E}_1(\delta)\mid\mathbf{x}_{\mathcal{T}}\} &\overset{(49a)}{=} \mathrm{P}\big\{Z(\mathcal{N}(i))-(1/N)\|\mathbf{P}_{\mathcal{T}^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2\geq\delta+(\lambda/2)(t-s_i)\,|\,\mathbf{x}_{\mathcal{T}}\big\} \\ &\overset{(57)}{=} \mathrm{P}\big\{(1/N)\|\mathbf{P}_{\mathcal{N}(i)^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2-(1/N)\|\mathbf{P}_{\mathcal{T}^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2\geq\delta+(\lambda/2)(t-s_i)\mid\mathbf{x}_{\mathcal{T}}\big\} \\ &\overset{\lambda=\rho^2_{\mathrm{min}}/6}{=} \mathrm{P}\big\{(1/N)\|\mathbf{P}_{\mathcal{N}(i)^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2-(1/N)\|\mathbf{P}_{\mathcal{T}^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2\geq\delta+(\rho^2_{\mathrm{min}}/12)(t-s_i)\mid\mathbf{x}_{\mathcal{T}}\big\} \\ &\overset{(56),(38)}{\leq} \mathrm{P}\big\{(1/N)\|\mathbf{P}_{\mathcal{N}(i)^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2-(1/N)\|\mathbf{P}_{\mathcal{T}^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2\geq\rho^2_{\mathrm{min}}(\ell_1+\ell_2)/12\mid\mathbf{x}_{\mathcal{T}}\big\}. \end{aligned} \tag{58}$$

By elementary properties of projections in Euclidean spaces (Wainwright, 2009a, Appx. A)

$$\begin{aligned} \|\mathbf{P}_{\mathcal{N}(i)^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2-\|\mathbf{P}_{\mathcal{T}^{\perp}}\boldsymbol{\varepsilon}_i\|_2^2 &= \|\mathbf{P}_{\mathcal{T}}\boldsymbol{\varepsilon}_i\|_2^2-\|\mathbf{P}_{\mathcal{N}(i)}\boldsymbol{\varepsilon}_i\|_2^2 \\ &= \|\big(\mathbf{P}_{\mathcal{T}}-\mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}\big)\boldsymbol{\varepsilon}_i\|_2^2-\|\big(\mathbf{P}_{\mathcal{N}(i)}-\mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}\big)\boldsymbol{\varepsilon}_i\|_2^2, \end{aligned} \tag{59}$$

with

$$\mathbf{P}_{\mathcal{T}} := \text{blkdiag}\big\{\mathbf{P}_{\mathcal{T}}^{(b)}\big\}_{b=1}^{B}.$$

Combining (59) with (58),

$$\mathrm{P}\{\mathcal{E}_1(\delta) \mid \mathbf{x}_{\mathcal{T}}\} \leq \mathrm{P}\big\{(1/N)\big\|\big(\mathbf{P}_{\mathcal{T}} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}\big)\boldsymbol{\varepsilon}_i\big\|_2^2 \geq \rho_{\min}^2(\ell_1+\ell_2)/12 \mid \mathbf{x}_{\mathcal{T}}\big\}. \tag{60}$$

with

$$\big(\mathbf{P}_{\mathcal{T}} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}\big) = \text{blkdiag}\big\{\mathbf{P}_{\mathcal{T}}^{(b)} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}^{(b)}\big\}_{b=1}^{B}.$$

The matrix $\mathbf{P}_{\mathcal{T}}^{(b)} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}^{(b)} \in \mathbb{R}^{L\times L}$ is a random (since it depends on $\mathbf{x}_{\mathcal{T}} = \{\mathbf{x}_j\}_{j\in\mathcal{T}}$) orthogonal projection matrix on a subspace of dimension at most $\ell_2 = |\mathcal{T}\setminus\mathcal{N}(i)|$ (cf. (39)), i.e.,

$$\mathbf{P}_{\mathcal{T}}^{(b)} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}^{(b)} = \sum_{j=1}^{\ell_2} \tilde{a}_j^{(b)} \mathbf{u}_j^{(b)} \big(\mathbf{u}_j^{(b)}\big)^T \tag{61}$$

with some coefficients $\tilde{a}_j^{(b)} \in \{0,1\}$ and orthonormal vectors $\{\mathbf{u}_j^{(b)} \in \mathbb{R}^L\}_{j=1,\ldots,\ell_2}$. Inserting (61) into (60),

$$\begin{aligned}
\mathrm{P}\{\mathcal{E}_1(\delta) \mid \mathbf{x}_{\mathcal{T}}\} &\leq \mathrm{P}\big\{(1/N)\big\|\big(\mathbf{P}_{\mathcal{T}} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}\big)\boldsymbol{\varepsilon}_i\big\|_2^2 \geq \rho_{\min}^2(\ell_1+\ell_2)/12 \mid \mathbf{x}_{\mathcal{T}}\big\} \\
&= \mathrm{P}\big\{(1/N)\sum_{b=1}^{B}\big\|\big(\mathbf{P}_{\mathcal{T}}^{(b)} - \mathbf{P}_{\mathcal{T}\cap\mathcal{N}(i)}^{(b)}\big)\boldsymbol{\varepsilon}_i^{(b)}\big\|_2^2 \geq \rho_{\min}^2(\ell_1+\ell_2)/12 \mid \mathbf{x}_{\mathcal{T}}\big\} \\
&\overset{(61)}{=} \mathrm{P}\big\{(1/N)\sum_{b=1}^{B}\sum_{j=1}^{\ell_2} \tilde{a}_j^{(b)}\big(z_j^{(b)}\big)^2 \geq \rho_{\min}^2(\ell_1+\ell_2)/12 \mid \mathbf{x}_{\mathcal{T}}\big\}
\end{aligned} \tag{62}$$

with $z_j^{(b)} = \big(\mathbf{u}_j^{(b)}\big)^T \boldsymbol{\varepsilon}_i^{(b)} \sim \mathcal{N}(0, 1/K_{i,i}^{(b)})$ (conditioned on $\mathbf{x}_{\mathcal{T}}$). Then, as can be verified easily,

$$(1/N)\sum_{b=1}^{B}\sum_{j=1}^{\ell_2} \tilde{a}_j^{(b)}\big(z_j^{(b)}\big)^2 = \sum_{n=1}^{N} \tilde{a}_n z_n^2, \text{ with } z_n \sim \mathcal{N}(0,1) \tag{63}$$

and coefficients $\tilde{a}_n \in [0, \beta/N]$ (cf. (11)) satisfying

$$\sum_{n=1}^{N} \tilde{a}_n \overset{(a)}{\leq} \ell_2\beta B/N \overset{(24)}{\leq} \ell_2\rho_{\min}^2/24. \tag{64}$$

Here, step (a) can be verified by taking (conditional, w.r.t. $\mathbf{x}_{\mathcal{T}} = \{\mathbf{x}_j\}_{j\in\mathcal{T}}$) expectations of (63) and using $|a_j^{(b)}| \leq 1$, $1/K_{i,i}^{(b)} \overset{(11)}{\leq} \beta$.

Inserting (63) into (62),

$$\mathrm{P}\{\mathcal{E}_1(\delta) \mid \mathbf{x}_{\mathcal{T}}\} \leq \mathrm{P}\big\{\sum_{j=1}^{N} \tilde{a}_j z_j^2 \geq \rho_{\min}^2(\ell_1+\ell_2)/12 \mid \mathbf{x}_{\mathcal{T}}\big\}$$

$$= \mathrm{P}\big\{ \sum_{j=1}^{N} \tilde{a}_j z_j^2 - \sum_{j=1}^{N} \tilde{a}_j \geq \rho_{\min}^2(\ell_1+\ell_2)/12 - \sum_{j=1}^{N} \tilde{a}_j \mid \mathbf{x}_{\mathcal{T}} \big\}$$

$$\begin{aligned} &\overset{(64)}{\leq} \mathrm{P}\big\{ \sum_{j=1}^{N} \tilde{a}_j z_j^2 - \sum_{j=1}^{N} \tilde{a}_j \geq \rho_{\min}^2(\ell_1+\ell_2)/24 \mid \mathbf{x}_{\mathcal{T}} \big\} \\ &\overset{z_j|\mathbf{x}_{\mathcal{T}} \sim \mathcal{N}(0,1)}{\leq} \mathrm{P}\big\{ \sum_{j=1}^{N} \tilde{a}_j z_j^2 - \mathrm{E}\{\sum_{j=1}^{N} \tilde{a}_j z_j^2 | \mathbf{x}_{\mathcal{T}}\} \geq \rho_{\min}^2(\ell_1+\ell_2)/24 \mid \mathbf{x}_{\mathcal{T}} \big\}. \end{aligned} \tag{65}$$

We now apply Lemma 4 to (65) using the choice

$$\eta := \rho_{\min}^2(\ell_1+\ell_2)/24, \tag{66}$$

$a_j := \tilde{a}_j$ and $b_j := 0$ (cf. (84)). This yields

$$\begin{aligned} \mathrm{P}\{\mathcal{E}_1(\delta) \mid \mathbf{x}_{\mathcal{T}}\} &\overset{(65),(85)}{\leq} 2 \exp\bigg(- \frac{\eta^2/8}{\sum_{j=1}^{N} \tilde{a}_j^2 + \eta \max\limits_{j=1,\ldots,N} \tilde{a}_j} \bigg) \\ &\leq 2 \exp\bigg(- \frac{N\eta^2/(8\beta)}{\sum_{j=1}^{N} \tilde{a}_j + \eta} \bigg), \end{aligned} \tag{67}$$

where the second inequality uses $\max\limits_{j=1,\ldots,N} \tilde{a}_j \overset{(63)}{\leq} \beta/N$. Combining

$$\sum_{j=1}^{N} \tilde{a}_j \overset{(64)}{\leq} \ell_2 \rho_{\min}/24 \overset{(66)}{\leq} \eta \tag{68}$$

with (67), we arrive at

$$\begin{aligned} \mathrm{P}\{\mathcal{E}_1(\delta)\} &= \mathrm{E}\big\{\mathrm{P}\big\{\mathcal{E}_1(\delta) \mid \mathbf{x}_{\mathcal{T}}\big\}\big\} \\ &\overset{(67),(68)}{\leq} 2 \exp\Big(- N\eta/(16\beta) \Big) \\ &\overset{(66)}{=} 2 \exp\Big(- N\rho_{\min}(\ell_1+\ell_2)/(24 \cdot 16\beta) \Big). \end{aligned} \tag{69}$$

To upper bound the probability of $\mathcal{E}_2(\delta)$ (cf. (49b)), consider

$$\begin{aligned} \mathrm{P}\big\{\mathcal{E}_2(\delta) \mid \mathbf{x}_{\mathcal{T}}\big\} &\overset{(49b)}{:=} \mathrm{P}\Big\{ Z(\mathcal{T}) - (1/N)\|\mathbf{P}_{\mathcal{T}^{\perp}} \boldsymbol{\varepsilon}_i\|_2^2 + (\lambda/2)(t-s_i) \leq 2\delta \mid \mathbf{x}_{\mathcal{T}} \Big\} \\ &\overset{(35)}{=} \mathrm{P}\Big\{ (1/N)\mathbf{x}_i^T \mathbf{P}_{\mathcal{T}^{\perp}} \mathbf{x}_i - (1/N)\boldsymbol{\varepsilon}_i^T \mathbf{P}_{\mathcal{T}^{\perp}} \boldsymbol{\varepsilon}_i + (\lambda/2)(t-s_i) \leq 2\delta \mid \mathbf{x}_{\mathcal{T}} \Big\} \end{aligned}$$

$$\overset{(13)}{=} \mathrm{P}\Big\{(1/N)\tilde{\mathbf{x}}_i^T\mathbf{P}_{\mathcal{T}^\perp}\tilde{\mathbf{x}}_i+(2/N)\tilde{\mathbf{x}}_i^T\mathbf{P}_{\mathcal{T}^\perp}\boldsymbol{\varepsilon}_i+(\lambda/2)(t-s_i)\le 2\delta \mid \mathbf{x}_{\mathcal{T}}\Big\} \tag{70}$$

with $\boldsymbol{\varepsilon}_i=(\varepsilon_1,\ldots,\varepsilon_N)^T$ (cf. (46)) and $\tilde{\mathbf{x}}_i$ (cf. (18)).

By defining the random vector

$$\mathbf{v}=(v_1,\ldots,v_N)^T:=\mathbf{V}^T\mathbf{C}_{\tilde{x}_i}^{-1/2}\tilde{\mathbf{x}}_i,$$

using the (random) orthonormal matrix $\mathbf{V}\in\mathbb{R}^{N\times N}$ constituted by the singular vectors of the matrix $\mathbf{P}_{\mathcal{T}^\perp}\mathbf{C}_{\tilde{x}_i}^{1/2}$ (cf. (51)), we can rewrite (70) as

$$\begin{aligned}\mathrm{P}\{\mathcal{E}_2(\delta)\mid\mathbf{x}_{\mathcal{T}}\}&=\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j\le 2\delta-(\lambda/2)(t-s_i)\mid\mathbf{x}_{\mathcal{T}}\bigg\}\\ &\overset{\delta=m_3/4}{=}\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j-m_3\le -m_3/2-(\lambda/2)(t-s_i)\mid\mathbf{x}_{\mathcal{T}}\bigg\}.\end{aligned} \tag{71}$$

Note that, conditioned on $\mathbf{x}_{\mathcal{T}}$, the vector $\mathbf{v}$ is standard Gaussian, i.e., $\mathbf{v}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_N)$.We now consider (71) for the particular choice $\lambda=\rho_{\min}/6$ which yields, using (55),

$$\begin{aligned}\mathrm{P}\big\{\mathcal{E}_2(\delta)\mid\mathbf{x}_{\mathcal{T}}\big\}&\overset{\delta=m_3/4}{=}\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j-m_3\le -m_3/2-(\lambda/2)(t-s_i)\mid\mathbf{x}_{\mathcal{T}}\bigg\}\\ &=\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j-m_3\le -(3/8)m_3-((\lambda/2)(t-s_i)+m_3/8)\mid\mathbf{x}_{\mathcal{T}}\bigg\}\\ &\overset{(55)}{\le}\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j-m_3\le -(3/8)m_3-((\lambda/2)(t-s_i)+\ell_1\rho_{\min}/12)\mid\mathbf{x}_{\mathcal{T}}\bigg\}\\ &\overset{\lambda=\rho_{\min}/6}{=}\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j-m_3\le -(3/8)m_3-(\rho_{\min}/12)((t-s_i)+\ell_1)\mid\mathbf{x}_{\mathcal{T}}\bigg\}\\ &\overset{(39)}{=}\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^{N}v_j^2d_j^2+\frac{2}{N}\sum_{j=1}^{N}v_jd_j\varepsilon_j-m_3\le -(3/8)m_3-(\rho_{\min}/12)\ell_2\mid\mathbf{x}_{\mathcal{T}}\bigg\}.\end{aligned} \tag{72}$$

We will invoke Lemma 4 to obtain an upper bound for $\mathrm{P}\big\{\mathcal{E}_2(\delta=m_3/4)\mid\mathbf{x}_{\mathcal{T}}\big\}$. To this end, in order to control the term $(2/N)\sum\limits_{j=1}^{N}v_jd_j\varepsilon_j$ in (71), we condition on the event

$$\mathcal{A}:=\bigg\{(1/N)\sum_{j=1}^{N}d_j^2\varepsilon_j^2\le\underbrace{2(\beta/N)\sum_{j=1}^{N}d_j^2}_{\overset{(53)}{=}2\beta m_3}+\beta\ell_2(\rho_{\min}/12)\bigg\} \tag{73}$$

with the constant β of Assumption 3. The event $\mathcal{A}$ is, conditioned on $\mathbf{x}_{\mathcal{T}}$, statistically independent of $\tilde{\mathbf{x}}_i$ (cf. (13)) since, loosely speaking, its definition (73) involves only the random variables $\{\varepsilon_j\}_{j=1,\dots,N}$ which are statistically independent of $\tilde{\mathbf{x}}_i$ (cf. (13)) and quantities (e.g., the singular values d_j) which are constant when conditioning on $\mathbf{x}_{\mathcal{T}}$.

We can upper bound the probability $\mathrm{P}\big\{\mathcal{E}_2(\delta = m_3/4) \mid \mathbf{x}_{\mathcal{T}}\big\}$ as

$$\begin{aligned}\mathrm{P}\{\mathcal{E}_2(\delta) \mid \mathbf{x}_{\mathcal{T}}\} &= \mathrm{P}\{\mathcal{E}_2(\delta)|\mathcal{A},\mathbf{x}_{\mathcal{T}}\}\mathrm{P}\{\mathcal{A} \mid \mathbf{x}_{\mathcal{T}}\} + \underbrace{\mathrm{P}\{\mathcal{E}_2(\delta)|\mathcal{A}^c,\mathbf{x}_{\mathcal{T}}\}}_{\leq 1}\mathrm{P}\{\mathcal{A}^c \mid \mathbf{x}_{\mathcal{T}}\} \\ &\leq \mathrm{P}\{\mathcal{E}_2(\delta)|\mathcal{A},\mathbf{x}_{\mathcal{T}}\} + \mathrm{P}\{\mathcal{A}^c \mid \mathbf{x}_{\mathcal{T}}\}. \end{aligned} \tag{74}$$

In order to control the probability $\mathrm{P}\{\mathcal{A}^c \mid \mathbf{x}_{\mathcal{T}}\}$ in (74), we will invoke Lemma 4. To this end, observe

$$\begin{aligned}\mathrm{P}\{\mathcal{A}^c \mid \mathbf{x}_{\mathcal{T}}\} &= \mathrm{P}\{(1/N)\sum_{j=1}^{N} d_j^2\varepsilon_j^2 \geq 2(\beta/N)\sum_{j=1}^{N} d_j^2 + \beta\ell_2(\rho_{\min}/12) \mid \mathbf{x}_{\mathcal{T}}\} \\ &\overset{(a)}{\leq} \mathrm{P}\{(1/N)\sum_{j=1}^{N} d_j^2\varepsilon_j^2 - \mathrm{E}\{(1/N)\sum_{j=1}^{N} d_j^2\varepsilon_j^2 \mid \mathbf{x}_{\mathcal{T}}\} \geq (\beta/N)\sum_{j=1}^{N} d_j^2 + \beta\ell_2(\rho_{\min}/12) \mid \mathbf{x}_{\mathcal{T}}\}, \end{aligned} \tag{75}$$

where (a) is due to

$$\mathrm{E}\bigg\{(1/N)\sum_{j=1}^{N} d_j^2\varepsilon_j^2 \mid \mathbf{x}_{\mathcal{T}}\bigg\} = (1/N)\sum_{j=1}^{N} d_j^2\mathrm{E}\{\varepsilon_j^2 \mid \mathbf{x}_{\mathcal{T}}\} \overset{(11),(46)}{\leq} (\beta/N)\sum_{j=1}^{N} d_j^2.$$

The random variables $\{\varepsilon_j\}_{j=1,\dots,N}$ are, conditioned on $\mathbf{x}_{\mathcal{T}}$, i.i.d. zero-mean Gaussian variables with variance $\sigma_\varepsilon^2 \leq \beta$ (cf. (46)). Therefore, we can use the innovation representation

$$\varepsilon_j = \tilde{b}_j z_j \tag{76}$$

with i.i.d. standard Gaussian random variables $z_j \sim \mathcal{N}(0,1)$ and some coefficients $\tilde{b}_j \in [0, \sqrt{\beta}]$. Inserting (76) into (75),

$$\begin{aligned}&\mathrm{P}\{\mathcal{A}^c \mid \mathbf{x}_{\mathcal{T}}\} \leq \\ &\mathrm{P}\Big\{\sum_{j=1}^{N}\tilde{b}_j^2 d_j^2 z_j^2 - \mathrm{E}\Big\{\sum_{j=1}^{N}\tilde{b}_j d_j^2 z_j^2 \mid \mathbf{x}_{\mathcal{T}}\Big\} \geq \beta\Big(\sum_{j=1}^{N} d_j^2 + N\ell_2(\rho_{\min}/12)\Big) \mid \mathbf{x}_{\mathcal{T}}\Big\}. \end{aligned} \tag{77}$$

Applying (85), using the choice $\eta := \beta\big(\sum_{j=1}^{N} d_j^2 + N\ell_2(\rho_{\min}/12)\big)$, $a_j := \tilde{b}_j^2 d_j^2$ and $b_j = 0$ (cf. (84)) to (77), yields

$$\mathrm{P}\{\mathcal{A}^c \mid \mathbf{x}_{\mathcal{T}}\} \overset{(85)}{\leq} \exp\Bigg(-\frac{\beta^2\big(\sum_{j=1}^{N} d_j^2 + N\ell_2(\rho_{\min}/12)\big)^2/8}{\sum_{j=1}^{N}\tilde{b}_j^4 d_j^4 + \beta\big(\sum_{j=1}^{N} d_j^2 + N\ell_2(\rho_{\min}/12)\big)\max\limits_{j=1,\dots,N}\tilde{b}_j^2 d_j^2}\Bigg)$$

$$\overset{\tilde{b}_j^2\leq\beta,(52)}{\leq} \exp\bigg(-\frac{\big(\sum_{j=1}^N d_j^2+N\ell_2(\rho_{\min}/12)\big)^2}{16\beta\big(\sum_{j=1}^N d_j^2+N\ell_2(\rho_{\min}/12)\big)}\bigg)$$

$$\overset{(53)}{\leq} \exp\bigg(-\frac{N\big(m_3+\ell_2(\rho_{\min}/12)\big)}{16\beta}\bigg)$$

$$\overset{(55)}{\leq} \exp\bigg(-\frac{N\big((2/3)\ell_1\rho_{\min}+\ell_2(\rho_{\min}/12)\big)}{16\beta}\bigg)$$

$$\leq \exp\bigg(-\frac{N\rho_{\min}\big(\ell_1+\ell_2\big)}{192\beta}\bigg). \tag{78}$$

In order to control the probability $\mathrm{P}\{\mathcal{E}_2(\delta)|\mathcal{A},\mathbf{x}_{\mathcal{T}}\}$ appearing in (74), we will again use Lemma 4. To this end, note that

$$\mathrm{E}\bigg\{\frac{1}{N}\sum_{j=1}^N v_j^2 d_j^2+\frac{2}{N}\sum_{j=1}^N v_j d_j\varepsilon_j\bigg|\mathcal{A},\mathbf{x}_{\mathcal{T}}\bigg\}\overset{(a)}{=}\frac{1}{N}\sum_{j=1}^N d_j^2\overset{(53)}{=}m_3, \tag{79}$$

with (a) due to the random variables $\{v_j\}_{j=1,\dots,N}$ being i.i.d standard Gaussian $\mathcal{N}(0,1)\}$, conditioned on $\mathbf{x}_{\mathcal{T}}$ and $\mathcal{A}$ (see (73)). Then,

$$\mathrm{P}\{\mathcal{E}_2(\delta)|\mathcal{A},\mathbf{x}_{\mathcal{T}}\}\overset{(72),(79)}{\leq}\mathrm{P}\bigg\{\frac{1}{N}\sum_{j=1}^N v_j^2 d_j^2+\frac{2}{N}\sum_{j=1}^N v_j d_j\varepsilon_j-m_3\leq -(3/8)m_3-(\rho_{\min}/12)\ell_2\mid\mathcal{A},\mathbf{x}_{\mathcal{T}}\bigg\}$$

$$\leq\mathrm{P}\bigg\{\big|\frac{1}{N}\sum_{j=1}^N v_j^2 d_j^2+\frac{2}{N}\sum_{j=1}^N v_j d_j\varepsilon_j-m_3\big|\geq(3/8)m_3+\rho_{\min}\ell_2/12\mid\mathcal{A},\mathbf{x}_{\mathcal{T}}\bigg\}$$

$$\leq\mathrm{P}\bigg\{\big|\sum_{j=1}^N v_j^2 d_j^2+2\sum_{j=1}^N v_j d_j\varepsilon_j-Nm_3\big|\geq N((3/8)m_3+\rho_{\min}\ell_2/12)\mid\mathcal{A},\mathbf{x}_{\mathcal{T}}\bigg\}. \tag{80}$$

Applying Lemma 4 to (80), using $\eta:=N(3m_3/8+\rho_{\min}\ell_2/12)$, $a_j:=d_j^2\overset{(52)}{\leq}\beta$, $b_j:=d_j\varepsilon_j$ yields

$$\mathrm{P}\{\mathcal{E}_2(\delta)|\mathcal{A},\mathbf{x}_{\mathcal{T}}\}\overset{(85)}{\leq}$$

$$2\exp\bigg(-\frac{N^2(3m_3/8+\rho_{\min}\ell_2/12)^2/8}{\beta(\sum_{j=1}^N d_j^2+(1/\beta)\sum_{j=1}^N d_j^2\varepsilon_j^2+N(3m_3/8+\rho_{\min}\ell_2/12))}\bigg)$$

$$\overset{(52)}{\leq} 2\exp\bigg(-\frac{N^2(3m_3/8+\rho_{\min}\ell_2/12)^2/8}{\beta(Nm_3+(1/\beta)\sum_{j=1}^N d_j^2\varepsilon_j^2+N(3m_3/8+\rho_{\min}\ell_2/12))}\bigg)$$

$$\overset{(73)}{\leq} 2\exp\bigg(-\frac{N^2(3m_3/8+\rho_{\min}\ell_2/12)^2/8}{\beta(3Nm_3+N\rho_{\min}\ell_2/12+N(3m_3/8+\rho_{\min}\ell_2/12))}\bigg)$$

$$\leq 2\exp\left(-\frac{N^2(3m_3/8+\rho_{\min}\ell_2/12)^2/8}{\beta 9N(3m_3/8+\rho_{\min}\ell_2/12)}\right)$$

$$\leq 2\exp\left(-\frac{N(3m_3/8+\rho_{\min}\ell_2/12)}{72\beta}\right)$$

$$\overset{(55)}{\leq} 2\exp\left(-\frac{N\rho_{\min}(\ell_1+\ell_2)/12}{72\beta}\right). \tag{81}$$

By combining (81) and (78) with (74),

$$\mathrm{P}\{\mathcal{E}_2(\delta)\} = \mathrm{E}\{\mathrm{P}\{\mathcal{E}_2(\delta) \mid \mathbf{x}_{\mathcal{T}}\}\} \leq 4\exp\left(-\frac{N\rho_{\min}(\ell_1+\ell_2)}{864\beta}\right) \tag{82}$$

Summing (69) and (82) yields (cf. (50))

$$\mathrm{P}\{\mathcal{E}_{\mathcal{T}}\} \leq M(\ell_1,t) := 6\exp\left(-\frac{N\rho_{\min}(\ell_1+\ell_2)}{864\beta}\right). \tag{83}$$

Inserting the upper bound (83) into (45),

$$\log \mathrm{P}\{\mathcal{E}_i\} \leq 2\log s + \max_{(\ell_1,t)\in\mathcal{I}} \big[(\ell_1+\ell_2)\log p + \log 6 - \frac{N\rho_{\min}(\ell_1+\ell_2)}{864\beta}\big].$$

Thus, $\mathrm{P}\{\mathcal{E}_i\} \leq \eta$ whenever $N \geq 864\log(p6s^2/\eta)(\beta/\rho_{\min})$.

Appendix

The main device underlying our analysis is the following large deviation property of a quadratic form involving Gaussian random variables.

Lemma 4 *Consider two vectors* $\mathbf{a} = (a_1,\ldots,a_N)^T \in \mathbb{R}^N$ *and* $\mathbf{b} = (b_1,\ldots,b_N)^T \in \mathbb{R}^N$*. For* N *i.i.d. random variables* $z_j \sim \mathcal{N}(0,1)$*, define*

$$y = \sum_{j=1}^{N} a_j z_j^2 + b_j z_j. \tag{84}$$

Then,

$$\mathrm{P}\{|y-\mathrm{E}\{y\}| \geq \eta\} \leq 2\exp\left(-\frac{\eta^2/8}{\|\mathbf{a}\|_2^2+\|\mathbf{b}\|_2^2+\|\mathbf{a}\|_\infty \eta}\right). \tag{85}$$

Proof An elementary calculation (see, e.g., (Foucart and Rauhut, 2012, Lemma 7.6)) reveals

$$\mathrm{E}\{\exp(\lambda(a_i z_i^2 + b_i z_i)\} = \exp\left(\frac{\lambda^2 b_i^2/2}{1-2\lambda a_i}\right)\sqrt{\frac{1}{1-2\lambda a_i}}, \tag{86}$$

which holds for any $\lambda \in [0, 1/(4\|\mathbf{a}\|_\infty)]$. Hence, for any $i \in \{1,\ldots,N\}$,

$$\log \mathrm{E}\{\exp(\lambda(a_i z_i^2 + b_i z_i - a_i))\}$$

$$\begin{aligned}&\overset{(86)}{=} \frac{\lambda^2 b_i^2/2}{1-2\lambda a_i} - (1/2)\log(1-2\lambda a_i)-\lambda a_i\\ &\overset{\lambda|a_i|\leq 1/4}{\leq} \lambda^2 b_i^2 - (1/2)\log(1-2\lambda|a_i|) - \lambda|a_i|. \end{aligned}\tag{87}$$

Since $-\log(1-u) \leq u + \frac{u^2}{2(1-u)}$, for $0 \leq u \leq 1$, the RHS of (87) yields, for every $i \in [N]$,

$$\begin{aligned}\log \mathrm{E}\{\exp(\lambda(a_i z_i^2 + b_i z_i - a_i))\} &\leq \lambda^2 b_i^2 + \frac{\lambda^2 a_i^2}{1-2\lambda|a_i|}\\ &\leq 2\lambda^2(a_i^2 + (1/2)b_i^2).\end{aligned}\tag{88}$$

Summing (88) for $i = 1, \ldots, N$ and inserting into (84),

$$\log \mathrm{E}\{\exp(\lambda(y - \mathrm{E}\{y\}))\} \leq 2\lambda^2(\|\mathbf{a}\|_2^2 + (1/2)\|\mathbf{b}\|_2^2).\tag{89}$$

Now, consider the tail bound (see, e.g., (Foucart and Rauhut, 2012, Remark 7.4))

$$\begin{aligned}\mathrm{P}\{y - \mathrm{E}\{y\} \geq \eta\} &\leq \exp(-\lambda\eta)\mathrm{E}\{\exp(\lambda(y - \mathrm{E}\{y\}))\}\\ &\overset{(89)}{\leq} \exp(-\lambda\eta + 2\lambda^2(\|\mathbf{a}\|_2^2 + (1/2)\|\mathbf{b}\|_2^2)).\end{aligned}\tag{90}$$

Minimizing the RHS of (90) over $\lambda \in [0, 1/(4\|\mathbf{a}\|_\infty)]$,

$$\begin{aligned}\mathrm{P}\{y-\mathrm{E}\{y\} \geq \eta\} &\leq \exp\bigg(-\frac{\eta^2/8}{(\|\mathbf{a}\|_2^2 + (1/2)\|\mathbf{b}\|_2^2) \vee (\eta\|\mathbf{a}\|_\infty)}\bigg)\\ &\overset{(a)}{\leq} \exp\bigg(-\frac{\eta^2/8}{(\|\mathbf{a}\|_2^2 + (1/2)\|\mathbf{b}\|_2^2) + \|\mathbf{a}\|_\infty\eta}\bigg),\end{aligned}\tag{91}$$

where (a) is due to $x \vee y \leq x+y$ for $x, y \in \mathbb{R}_+$. Similar to (91), one can also verify

$$\mathrm{P}\{y - \mathrm{E}\{y\} \leq -\eta\} \leq \exp\bigg(-\frac{\eta^2/8}{(\|\mathbf{a}\|_2^2 + (1/2)\|\mathbf{b}\|_2^2) + \|\mathbf{a}\|_\infty\eta}\bigg).\tag{92}$$

Adding (91) and (92) yields (85) by union bound. ■

References

H. Ambos, N. Tran, and A. Jung. Classifying big data over networks via the logistic network lasso. In *Proc. 52nd Asilomar Conference on Signals, Systems, and Computers*. 10.1109/ACSSC.2018.8645260, 2018.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Lear. Research*, 9:1179–1225, 2008.

F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. *IEEE Trans. Signal Processing*, 52(8):2189–2199, Aug. 2004.

B. Boashash, editor. *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*. Elsevier, Amsterdam, The Netherlands, 2003.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, volume 3. Now Publishers, Hanover, MA, 2010.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer New York, 1991.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011. doi: 10.1198/jasa.2011.tm10155.

S. Cui, A. Hero, Z.-Q. Luo, and J.M.F. Moura, editors. *Big Data over Networks*. Cambridge Univ. Press, 2016.

R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:151–172, 2000.

R. Dahlhaus. Local inference for locally stationary time series based on the empirical spectral measure. *Journal of Econometrics*, 2009.

R. Dahlhaus and L. Giraitis. On the optimal segment length for parameter estimates for locally stationary time series. *Journal of Time Series Analysis*, 19(6), 1998.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. B*, 76:373–397, 2014.

E. Davidson and M. Levin. Gene regulatory networks. *Proc. Natl. Acad. Sci.*, 102(14), Apr. 2005.

M. Eichler, R. Dahlhaus, and J. Sandkühler. Partial correlation analysis for the identification of synaptic connections. *Biol Cybern.*, 89(4), 2003.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, New York, 2012.

J. H. Friedmann, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, Jul. 2008.

R. G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.

D. Hallac, J. Leskovec, and S. Boyd. Network lasso: Clustering and optimization in large graphs. In *Proc. SIGKDD*, pages 387–396, 2015.

G. Hannak, A. Jung, and N. Görtz. On the information-theoretic limits of graphical model selection for Gaussian time series. In *Proc. EUSIPCO 2014*, Lisbon, Portugal, 2014.

A. Jung. Learning the conditional independence structure of stationary time series: A multitask learning approach. *IEEE Trans. Signal Processing*, 63(21), Nov. 2015.

A. Jung, G. Tauböck, and F. Hlawatsch. Compressive spectral estimation for nonstationary random processes. *IEEE Trans. Inf. Theory*, 59(5):3117–3138, May 2013.

A. Jung, R. Heckel, H. Bölcskei, and F. Hlawatsch. Compressive nonparametric graphical model selection for time series. In *Proc. IEEE ICASSP-2014*, Florence, Italy, May 2014.

A. Jung, G. Hannak, and N. Görtz. Graphical LASSO Based Model Selection for Time Series. *IEEE Sig. Proc. Letters*, 22(10):1781–1785, Oct. 2015.

A. Kipnis, A.J. Goldsmith, and Y.C. Eldar. The distortion rate function of cyclostationary gaussian processes. *IEEE Trans. Inform. Theory*, 64(5):3810–3824, 2018.

D. Koller, N., and Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.

P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.*, 45(6):2455–2482, 2017.

S. Mallat, G. Papanicolaou, and Z. Zhang. Adaptive covariance estimation of locally stationary processes. *Ann. Statist.*, 26(1):1–47, 1998.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, 34(3):1436–1462, 2006.

A. Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables, and Stochastic Processes*. Mc-Graw Hill, New York, 4 edition, 2002.

P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319, 2010.

P. Ravikumar, M. J. Wainwright, and B. Raskutti, G. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

A. Sadeghi, C. Lange, M.E. Vidal, and S. Auer. Communication metadata using knowledge graphs. In *Lecture Notes in Computer Science*. Springer, 2017.

A. Sandryhaila and J. M. F. Moura. Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure. *IEEE Signal Processing Magazine*, 31(5):80–90, Sept 2014.

C. Starica and C. Granger. Nonstationarities in stock returns. *The Review of Economics and Statistics*, 87 (3):495–502, 2005.

K. M. Tan, P. London, K. Mohan, S.-I. Lee, M. Fazel, and D. Witten. Learning graphical models with hubs. *Jour. Mach. Learning Res.*, 15(10):3297–3331, Oct. 2014.

E. F. Velez and R. G. Absher. Spectral estimation based on the wigner-ville representation. *Signal Processing*, 20, 1990.

D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, Sep. 2014.

P. Wahlberg and M. Hansson. Kernels and multiple windows for estimation of the wigner-ville spectrum of gaussian locally stationary processes. *IEEE Transactions on Signal Processing*, 55(10), 2007.

M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory*, 55(12):5728–5741, Dec. 2009a.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202, May 2009b.

M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*, volume 1 of *Foundations and Trends in Machine Learning*. Now Publishers, Hanover, MA, 2008.

W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic bounds on model selection for Gaussian Markov random fields. In *Proc. IEEE ISIT-2010*, pages 1373–1377, Austin, TX, Jun. 2010.

E. Yang and A. Lozano. Robust gaussian graphical modeling with the trimmed graphical lasso. In *Advances in Neural Information Processing Systems 28*, pages 2602–2610, 2015.