ON THE SAMPLE COMPLEXITY OF GRAPHICAL MODEL SELECTION FROM NON-STATIONARY SAMPLES

Nguyen Q. Tran and Alexander Jung

Dept. of Computer Science, Aalto University, Finland; firstname.lastname@aalto.fi

ABSTRACT

We characterize the sample size required for accurate graphical model selection from non-stationary samples. The observed samples are modeled as a zero-mean Gaussian random process whose samples are uncorrelated but have different covariance matrices. This includes the case where observations form stationary or underspread processes. We derive a sufficient condition on the required sample size by analyzing a simple sparse neighborhood regression method.

***Index Terms*—sparsity, graphical model selection, neighborhood regression, high-dimensional statistics.**

1. INTRODUCTION

One of the most successful approaches to manage massive high-speed datasets (big data) is based on graph or network models [1], [2]. However, in many application domains the network or graph structure has to be learned in a data-driven fashion from training samples. Most existing methods for graphical model selection (GMS) model the training samples to be i.i.d. or samples of a stationary random process [3]–[6].

In contrast, we consider samples forming a non-stationary uncorrelated process. This covers the case where the samples form a stationary process or a locally stationary (underspread) process, for which efficient decorrelation can be achieved by discrete Fourier or local cosine transforms [7].

Our main conceptual contribution is the formulation of a sparse neighborhood regression GMS method for high-dimensional non-stationary processes. As our main analytical contribution, we derive upper bounds on the sample size such that accurate GMS is possible. In particular, our analysis reveals that the crucial parameter determining the required sample size is the minimum average partial correlation between the process components. If this quantity is not too small, accurate GMS is feasible even in the high-dimensional regime, where the system size might exceed (drastically) the number of available training samples.

After formalizing the problem setup in Section 2, we analyze a simple GMS method in Section 3. This analysis results in upper bounds on the sample size required for GMS.

Notation: For a vector $\mathbf{x} = (x_1, \ldots, x_d)^T$, 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 sequence of matrices $\mathbf{Q}_l$, we denote by $\mathrm{blkdiag}\{\mathbf{Q}_l\}$ the block diagonal matrix with lth diagonal block $\mathbf{Q}_l$. 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 some event $\mathcal{A}$ is denoted $\mathcal{A}^c$. The expectation of a random variable y is $\mathrm{E}\{y\}$. For a natural number n, we denote $[n] = \{1, \ldots, n\}$.

2. PROBLEM FORMULATION

We consider observing N vector-valued samples $\{\mathbf{x}[n]\}_{n=1}^N$, each sample $\mathbf{x}[n] \in \mathbb{R}^p$ containing p scalars $\{x_i[n]\}_{i\in[p]}$. The samples are modelled as realizations of zero-mean Gaussian random vectors, which are uncorrelated, i.e., $\mathrm{E}\{\mathbf{x}[n]\mathbf{x}^T[n']\} = \mathbf{0}$ for $n \neq n'$. Thus, the probability distribution of the observed samples is fully specified by the covariance matrices $\mathbf{C}[n] := \mathrm{E}\{\mathbf{x}[n]\mathbf{x}^T[n]\}$.

By contrast to the widely used i.i.d. assumption (where $\mathbf{C}[n]$ is the same for all n), we allow the covariance matrix $\mathbf{C}[n]$ to vary with sample index n. However, we impose a piecewise smoothness constraint on the variation of the covariance matrix $\mathbf{C}[n]$ over sample index n. In particular, the covariance matrix $\mathbf{C}[n]$ is constant over blocks of L consecutive vector samples. Thus, our model for the samples is

$$\underbrace{\mathbf{x}[1], \ldots, \mathbf{x}[L]}_{\text{i.i.d.}\sim\mathcal{N}(\mathbf{0},\mathbf{C}^{(b=1)})}, \underbrace{\mathbf{x}[L+1], \ldots, \mathbf{x}[2L]}_{\text{i.i.d.}\sim\mathcal{N}(\mathbf{0},\mathbf{C}^{(b=2)})}, \ldots, \tag{1}$$

i.e., samples are independent zero-mean Gaussian vectors with covariance

$$\mathbf{C}[n] = \mathbf{C}^{(b)} \text{ for } n \in \{(b-1)L+1, \ldots, bL\}. \tag{2}$$

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, i.e., $N = BL$, with the number B of data blocks. The model (1) reduces to the well-studied i.i.d. setting for $B = 1$ and block length $L = N$. In this paper, we study limits of accurate GMS using model (1) with $B > 1$.

Stationary Processes. The model (1) covers the case where the observed samples form a stationary process [5], [6], [8]. Indeed, consider a zero-mean Gaussian stationary process $\mathbf{z}[n]$ with auto-covariance function

$$\mathbf{R}_z[m] := \mathrm{E}\{\mathbf{z}[n]\mathbf{z}^T[n-m]\} \tag{3}$$

and spectral density matrix (SDM) [9]

$$\mathbf{S}_z(\theta) := \sum_{m=-\infty}^{\infty} \mathbf{R}_z[m] \exp(-j2\pi\theta m). \tag{4}$$

Let $\mathbf{x}[k] := \sum_{n=0}^{N-1} \mathbf{z}[n] \exp(-j2\pi nk/N)$ denote the discrete Fourier transform (DFT) of the stationary process $\mathbf{z}[n]$. Then, by well-known properties of the DFT [10], the vectors $\mathbf{x}[k]$, for $k = 0, \ldots, N-1$, are approximately uncorrelated Gaussian random vectors with zero mean and covariance matrix $\mathbf{C}[k] \approx$

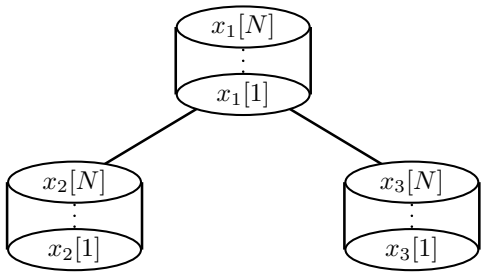

Fig. 1. The CIG of a vector process $\mathbf{x}[n] = (x_1[n], x_1[n], x_3[n])^T$ (cf. (1)).

$\mathbf{S}_z(k/N)$. Moreover, if the effective correlation width W of the process $\mathbf{z}[n]$ is small, i.e., $W \ll N$, the SDM $\mathbf{S}_z(\theta)$ is nearly constant over a frequency interval of length $1/W$. Thus, the DFT vectors $\mathbf{x}[k]$ approximately conform to the process model (1) with block length $L = N/W$.

Underspread Processes. The process model (1) is also useful for the important class of underspread non-stationary processes [7]. A continuous-time random process $\mathbf{z}(t)$ is called underspread if its expected ambiguity function (EAF) $\bar{\mathbf{A}}(\tau,\nu) := \int_{t=-\infty}^{\infty} \mathrm{E}\{\mathbf{z}(t+\tau/2)\mathbf{z}^T(t-\tau/2)\}\exp(-j2\pi t\nu)dt$ is well concentrated around the origin in the (τ,ν) plane. In particular, if the EAF of $\mathbf{z}(t)$ is supported on the rectangle $[-\tau_0/2,\tau_0/2]\times[-\nu_0/2,\nu_0/2]$, then the process $\mathbf{z}(t)$ is underspread if $\tau_0\nu_0 \ll 1$.

One of the most striking properties of an underspread process is that its Wigner-Ville spectrum (which can be loosely interpreted as a time-varying power spectral density) $\overline{W}(t,f) := \int_{\tau,\nu} \bar{\mathbf{A}}(\tau,\nu)\exp(-2\pi(f\tau-\nu t))d\tau d\nu$ is approximately constant over a rectangle of area $1/(\tau_0\nu_0)$. Moreover, 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 $\{g^{(n,k)}(t) := g(t-nT)e^{-2\pi kFt}\}_{n,k\in\mathbb{Z}}$ [7], yields zero-mean expansion coefficients $\mathbf{x}[n,k] = \int_t \mathbf{z}(t)g^{(n,k)}(t)dt$ which are approximately uncorrelated. Moreover, the covariance matrix of $\mathbf{x}[(n,k)]$ is approximately $\overline{W}(nT,kF)$. Thus, the vectors $\mathbf{x}[(n,k)]$ conform to the process model (1), with block length $L \approx \frac{1}{TF\tau_0\nu_0}$.

Conditional Independence Graph. We now define a graphical model for the observed samples $\{\mathbf{x}[n]\}_{n=1}^N$ by identifying the individual process components

$$\mathbf{x}_i = (x_i[1],\ldots,x_i[N])^T \tag{5}$$

with the nodes $\mathcal{V} = [p]$ of an undirected simple graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ (cf. Fig. 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, i.e., $\{i,j\} \notin \mathcal{E}$, if the corresponding 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 the fact that the CIG $\mathcal{G}$ represents stochastic dependencies between components of $\mathbf{x}[n]$ globally for all n. In particular, the edge set $\mathcal{E}$ does not depend on the sample index n. Our setting is similar to the one of [11], which considers samples grouped into different classes.

Since we model the process $\mathbf{x}[n]$ as Gaussian (cf. (1)), the CIG structure can be read off conveniently from the inverse covariance (precision) matrices $\mathbf{K}[n] := \mathbf{C}[n]^{-1}$. 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 $n \in [N]$ [10, Prop. 1.6.6]. Thus, we have the following characterization of the CIG $\mathcal{G}$ associated with the process $\mathbf{x}[n]$:

$$\{i,j\} \notin \mathcal{E} \text{ if and only if } K_{i,j}[n] = 0 \text{ for all } n. \tag{6}$$

We highlight the coupling over samples in the CIG characterization (6): An edge is absent, i.e., $\{i,j\} \notin \mathcal{E}$, if and only if the precision matrix entry $K_{i,j}[n]$ is zero *for all* $n \in [N]$.

In order to measure the strength of a connection between process components $\mathbf{x}_i$ and $\mathbf{x}_j$ for $\{i,j\} \in \mathcal{E}$, we define the *average partial correlation*

$$\begin{aligned}\rho_{i,j} &:= (1/N)\sum_{n=1}^{N} K_{i,j}^2[n]/K_{i,i}^2[n] \\ &\overset{(2)}{=} (1/B)\sum_{b=1}^{B} \big(K_{i,j}^{(b)}\big)^2/\big(K_{i,i}^{(b)}\big)^2. \end{aligned}\tag{7}$$

By (6) and (7), $\{i,j\} \in \mathcal{E}$ if and only if $\rho_{i,j} \neq 0$. Note that $\rho_{i,j}$ is an average measure, i.e., even if the marginal partial correlation is small for some n, $\rho_{i,j}$ might still be large.

Accurate estimation of the CIG for finite sample size N (incurring unavoidable sampling noise) is only possible for sufficiently large partial correlations $\rho_{i,j}$ for $\{i,j\} \in \mathcal{E}$.

Assumption 1. *There is a constant* $\rho_{\min} > 0$ *such that*

$$\rho_{i,j} \geq \rho_{\min} \text{ for any } \{i,j\} \in \mathcal{E}. \tag{8}$$

Moreover, we assume the CIG underlying $\mathbf{x}[n]$ to be sparse in the sense of each node having small neighborhood. In what follows, we denote the neighbourhood and degree of node $i \in \mathcal{V}$ by $\mathcal{N}(i) := \{j \in \mathcal{V}\setminus\{i\} : \{i,j\} \in \mathcal{E}\}$ and $s_i = |\mathcal{N}(i)|$, respectively.

Assumption 2. *For some* $s < (p/3)\wedge(L/3)$,

$$s_i \leq s, \text{ for any node } i \in \mathcal{V}. \tag{9}$$

3. SPARSE NEIGHBORHOOD REGRESSION

The CIG $\mathcal{G}$ of the process $\mathbf{x}[n]$ in (1) is fully specified by the neighborhoods, i.e., once we have found all neighborhoods, we can reconstruct the full CIG. In what follows we focus on the sub-problem of learning the neighborhood $\mathcal{N}(i)$ of an arbitrary but fixed node $i \in \mathcal{V}$.

In view of the process model (1), we denote for block $b \in [B]$ the ith process component as

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

By basic properties of multivariate normal distributions [12, Thm. 3.5.1] and the fact that $K_{i,j}[n] = 0$ for $j \notin \mathcal{N}(i)$, $\mathbf{x}_i^{(b)}$ can be decomposed as

$$\mathbf{x}_i^{(b)} = \sum_{j\in\mathcal{N}(i)} a_j \mathbf{x}_j^{(b)} + \boldsymbol{\varepsilon}_i^{(b)}, \tag{10}$$

with coefficients $a_j = -K_{i,j}^{(b)}/K_{i,i}^{(b)}$. The error term $\boldsymbol{\varepsilon}_i^{(b)} \sim \mathcal{N}(\mathbf{0},(1/K_{i,i}^{(b)})\mathbf{I}_L)$ is uncorrelated with $\mathbf{x}_j^{(b)}$, for $j \in \mathcal{N}(i)$.

Consider now an index set $\mathcal{T} \subseteq [p]$ with $\mathcal{N}(i) \setminus \mathcal{T} \neq \emptyset$. Another application of [12, Thm. 3.5.1] to the component $\sum_{j \in \mathcal{N}(i)} a_j \mathbf{x}_j^{(b)}$ yields

$$\mathbf{x}_i^{(b)} = \sum_{j \in \mathcal{T}} c_j \mathbf{x}_j^{(b)} + \tilde{\mathbf{x}}_i^{(b)} + \boldsymbol{\varepsilon}_i^{(b)}, \tag{11}$$

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

$$\tilde{\mathbf{x}}_i^{(b)} \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}_b^2 \mathbf{I}_L). \tag{12}$$

The variance $\tilde{\sigma}_b^2$ of $\tilde{\mathbf{x}}_i^{(b)}$ is obtained as

$$\tilde{\sigma}_b^2 = \mathbf{a}^T \tilde{\mathbf{K}}^{-1} \mathbf{a} \tag{13}$$

with $\tilde{\mathbf{K}} = \left((\mathbf{C}^{(b)}_{\mathcal{N}(i) \cup \mathcal{T}})^{-1} \right)_{\mathcal{N}(i) \setminus \mathcal{T}}$ and vector $\mathbf{a} \in \mathbb{R}^{|\mathcal{N}(i) \setminus \mathcal{T}|}$ whose entries are given by $\{a_j = -K^{(b)}_{i,j} / K^{(b)}_{i,i}\}_{j \in \mathcal{N}(i) \setminus \mathcal{T}}$.

The decompositions (10) and (11) naturally suggest a simple strategy for estimating (selecting) the neighborhood $\mathcal{N}(i)$. Let $\mathbf{P}_{\mathcal{T}}^{(b)}$ be the orthogonal projection on the complement of $\mathrm{span}\{\mathbf{x}_j^{(b)}\}_{j \in \mathcal{T}} \subseteq \mathbb{R}^L$. According to (10), for any index set $\mathcal{T}$ with $\mathcal{N}(i) \setminus \mathcal{T} = \emptyset$,

$$\|\mathbf{P}_{\mathcal{T}}^{(b)} \mathbf{x}_i^{(b)}\|_2^2 = \|\mathbf{P}_{\mathcal{T}}^{(b)} \boldsymbol{\varepsilon}_i^{(b)}\|_2^2 \text{ for all } b \in [B], \tag{14}$$

while for any index set $\mathcal{T}$ with $\mathcal{N}(i) \setminus \mathcal{T} \neq \emptyset$, (11) entails

$$\|\mathbf{P}_{\mathcal{T}}^{(b)} \mathbf{x}_i^{(b)}\|_2^2 = \|\mathbf{P}_{\mathcal{T}}^{(b)} (\tilde{\mathbf{x}}_i^{(b)} + \boldsymbol{\varepsilon}_i^{(b)})\|_2^2 \text{ for all } b \in [B], \tag{15}$$

with non-zero $\tilde{\mathbf{x}}_i^{(b)}$. Some of our efforts go into showing that

$$\|\mathbf{P}_{\mathcal{T}}^{(b)} (\tilde{\mathbf{x}}_i^{(b)} + \boldsymbol{\varepsilon}_i^{(b)})\|_2^2 \approx \|\mathbf{P}_{\mathcal{T}}^{(b)} \tilde{\mathbf{x}}_i^{(b)}\|_2^2 + \|\mathbf{P}_{\mathcal{T}}^{(b)} \boldsymbol{\varepsilon}_i^{(b)}\|_2^2, \tag{16}$$

for all blocks $b \in [B]$. Thus, if the component $\tilde{\mathbf{x}}_i^{(b)}$ in (11) is not too small, the estimator

$$\widehat{\mathcal{N}}(i) := \operatorname*{arg\,min}_{|\mathcal{T}| \leq s} (1/N) \sum_{b=1}^{B} \|\mathbf{P}_{\mathcal{T}}^{(b)} \mathbf{x}_i^{(b)}\|_2^2 + \lambda |\mathcal{T}|, \tag{17}$$

will deliver the true neighbourhood, i.e., $\mathcal{N}(i) = \widehat{\mathcal{N}}(i)$ with high probability. Note that the penalty term $\lambda|\mathcal{T}|$ in (17) is required to cope with the case of the degree of node i being smaller than s, i.e., with the case $|\mathcal{N}(i)| < s$.

The estimator (17) performs sparse neighborhood regression by approximating the ith component $\mathbf{x}_i$ (cf. (5)) in a sparse manner (by allowing s active components) using the remaining process components. We highlight that the estimator (17) is only useful for deriving achievability results since it allows for a simple performance analysis. However, a naive implementation of (17) would be intractable since it involves a combinatorial search over all subsets of size at most s. A tractable convex optimization method for learning the CIG for the process model (1) has been presented in [11].

For the analysis of the estimator (17) we require a bound on the eigenvalues of the covariance matrices $\mathbf{C}[n]$.

Assumption 3. *For some $\beta \geq 1$, $1 \leq \lambda_l(\mathbf{C}[n]) \leq \beta$ for all i, n.*

As can be verified easily, Asspt. 3 implies (cf. (11))

$$\tilde{\sigma}_b^2 \leq \beta. \tag{18}$$

Our main analytical result is an upper bound on the probability of the sparse neighborhood regression (17) to fail. To this end, we define the error event

$$\mathcal{E}_i := \{\mathcal{N}(i) \neq \widehat{\mathcal{N}}(i)\}. \tag{19}$$

Theorem 3.1. *Consider observed samples $\{\mathbf{x}[n]\}_{n \in [N]}$ conforming to the process model* (1) *such that Asspt.* 1, 2 *and* 3 *are valid. Then, if*

$$\rho_{\min} \geq 12 s \beta / L \tag{20}$$

and moreover

$$N \geq \frac{24 \beta (10 + 3L/s)}{\rho_{\min}} \big(4s \log(pe) + \log(4/\eta)\big), \tag{21}$$

the probability of (17) *to fail is bounded as $\mathrm{P}\{\mathcal{E}_i\} \leq \eta$, for the choice $\lambda \leq \rho_{\min}/(6s)$.*

By Theorem 3.1, the true neighborhood $\mathcal{N}(i)$ can be recovered via (17) with high probability if the samples size N is on the order of $\log p$ when the other parameters are held fixed. Therefore, GMS via sparse neighborhood regression (17) is feasible in the high dimensional regime where $N \ll p$. Moreover, the bound (21) indicates that the required sample size N depends on the ratio s/L and therefore reveals an interesting trade-off between block length L (of consecutive samples which are approximately i.i.d.) and the sparsity s of the underlying CIG. In particular, for a given sample size N, we can tolerate less smoothness, i.e., smaller block length L (cf. (1)), if the underlying CIG is more sparse, i.e., has a smaller maximum degree s (cf. (9)).

4. PROOF OF THE MAIN RESULT

We now verify Thm. 3.1 by analyzing the probability $\mathrm{P}\{\mathcal{E}_i\}$ of the event $\mathcal{E}_i$ (cf. (19)) when (17) fails to deliver the correct neighborhood $\mathcal{N}(i)$. Let us introduce the shorthands

$$\mathcal{E}_{\mathcal{T}} := \{Z(\mathcal{N}(i)) + \lambda s_i > Z(\mathcal{T}) + \lambda |\mathcal{T}|\},$$
$$Z(\mathcal{T}) := \frac{1}{N} \sum_{b=1}^{B} \|\mathbf{P}_{\mathcal{T}}^{(b)} \mathbf{x}_i^{(b)}\|_2^2. \tag{22}$$

The error event $\mathcal{E}_i$ (cf. (19)) can only occur if at least one $\mathcal{E}_{\mathcal{T}}$, for some $\mathcal{T} \neq \mathcal{N}(i)$ with $|\mathcal{T}| \leq s$, occurs, i.e.,

$$\mathcal{E}_i \subseteq \bigcup_{|\mathcal{T}| \leq s, \mathcal{N}(i) \neq \mathcal{T}} \mathcal{E}_{\mathcal{T}}, \tag{23}$$

and, in turn via a union bound,

$$\mathrm{P}\{\mathcal{E}_i\} \leq \sum_{|\mathcal{T}| \leq s, \mathcal{T} \neq \mathcal{N}(i)} \mathrm{P}\{\mathcal{E}_{\mathcal{T}}\}. \tag{24}$$

We now derive an upper bound $M(\ell_1, t)$ on $\mathrm{P}\{\mathcal{E}_{\mathcal{T}}\}$ which depends on the index set $\mathcal{T}$ only via the overlap $\ell_1 = |\mathcal{N}(i) \setminus \mathcal{T}|$ and the size $t = |\mathcal{T}|$. Let the set $\mathcal{N}(\ell_1, t)$ collect all those index sets with prescribed size $t = |\mathcal{T}|$ and overlap ℓ_1, i.e.,

$$\mathrm{P}\{\mathcal{E}_{\mathcal{T}}\} \leq M(\ell_1, t) \text{ for any } \mathcal{T} \in \mathcal{N}(\ell_1, t). \tag{25}$$

A basic combinatorial argument (see, e.g., [13, Sec. IV]) reveals that the number of these index sets is

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

It will be convenient to introduce the index set

$$\mathcal{I} := \{(\ell_1, t) \in \mathbb{Z}_+^2 : \ell_1 \leq s_i, t \leq s\} \setminus \{(0, s_i)\} \tag{26}$$

with cardinality $|\mathcal{I}| \leq s^2$. Combining

$$\{|\mathcal{T}| \leq s, \mathcal{T} \neq \mathcal{N}(i)\} \subseteq \bigcup_{(\ell_1,t)\in\mathcal{I}} \mathcal{N}(\ell_1,t), \tag{27}$$

with (24) implies, via a union bound,

$$\log \mathrm{P}\{\mathcal{E}_i\} \leq \max_{(\ell_1,t)\in\mathcal{I}} \log|\mathcal{I}| N(\ell_1,t) + \log M(\ell_1,t). \tag{28}$$

Our next goal is to determine a sufficiently tight upper bound $M(\ell_1,t)$ on the probability $\mathrm{P}\{\mathcal{E}_\mathcal{T}\}$ (cf. (24)) for an index set $\mathcal{T} \in \mathcal{N}(\ell_1,t)$. To this end, we make (10) more handy by stacking $\boldsymbol{\varepsilon}_i^{(b)}$ into $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_i^{(1)T}, \ldots, \boldsymbol{\varepsilon}_i^{(B)T})^T \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varepsilon}_i})$ with

$$\mathbf{C}_{\boldsymbol{\varepsilon}_i} = \mathrm{blkdiag}\{(1/K_{i,i}^{(b)})\mathbf{I}_L\}_{b=1}^B. \tag{29}$$

Using $\mathbf{P}_\mathcal{T} := \mathrm{blkdiag}\{\mathbf{P}_\mathcal{T}^{(b)}\}_{b=1}^B$, we can characterize the error event $\mathcal{E}_\mathcal{T}$ as (cf. (22))

$$\begin{aligned}\mathcal{E}_\mathcal{T} = \big\{ & Z(\mathcal{N}(i)) - (1/N)\|\mathbf{P}_\mathcal{T}\boldsymbol{\varepsilon}_i\|_2^2 \\ & > Z(\mathcal{T}) - (1/N)\|\mathbf{P}_\mathcal{T}\boldsymbol{\varepsilon}_i\|_2^2 + \lambda(t-s_i)\big\}.\end{aligned} \tag{30}$$

For some number $\delta > 0$, whose precise value to be chosen later, we define the two events

$$\mathcal{E}_1(\delta) := \big\{ \big| Z(\mathcal{N}(i)) - (1/N)\|\mathbf{P}_\mathcal{T}\boldsymbol{\varepsilon}_i\|_2^2 \big| \geq \delta \big\}, \tag{31a}$$

$$\mathcal{E}_2(\delta) := \big\{ Z(\mathcal{T}) - (1/N)\|\mathbf{P}_\mathcal{T}\boldsymbol{\varepsilon}_i\|_2^2 + \lambda(t-s_i) \leq 2\delta \big\}. \tag{31b}$$

By (30), the event $\mathcal{E}_\mathcal{T}$ can only occur if either the event $\mathcal{E}_1(\delta)$ or $\mathcal{E}_2(\delta)$ occurs. Therefore, by a union bound,

$$\mathrm{P}\{\mathcal{E}_\mathcal{T}\} \leq \mathrm{P}\{\mathcal{E}_1(\delta)\} + \mathrm{P}\{\mathcal{E}_2(\delta)\}. \tag{32}$$

We now bound each of the two summands in (32) separately. To this end, let us define

$$m_3 := \mathrm{E}\{(1/N)\|\mathbf{P}_\mathcal{T}\tilde{\mathbf{x}}_i\|_2^2\},$$

with

$$\tilde{\mathbf{x}}_i = (\tilde{\mathbf{x}}_i^{(1)T}, \ldots, \tilde{\mathbf{x}}_i^{(B)T})^T.$$

The bounds for $\mathrm{P}\{\mathcal{E}_1(\delta)\}$ and $\mathrm{P}\{\mathcal{E}_2(\delta)\}$ are stated in the following lemma.

Lemma 4.1. *For the choice of $\delta = m_3/4$, the following results are hold*

$$\mathrm{P}\{\mathcal{E}_1(\delta)\} \leq 2\exp\left(-\frac{\ell_1 N \rho_{\min}}{96\beta}\right), \tag{33}$$

and

$$\mathrm{P}\{\mathcal{E}_2(\delta)\} \leq 2\exp\left(-\frac{N\ell_1\rho_{\min}}{24\beta(10+3L/s)}\right). \tag{34}$$

Applying the results of Lemma 4.1 into (32) gets us to

$$\begin{aligned}\mathrm{P}\{\mathcal{E}_\mathcal{T}\} &\overset{(32)}{\leq} \mathrm{P}\{\mathcal{E}_1(\delta)\} + \mathrm{P}\{\mathcal{E}_2(\delta)\} \\ &\overset{(33),(34)}{\leq} 4\exp\left(-\frac{N\ell_1\rho_{\min}}{24\beta(10+3L/s)}\right).\end{aligned} \tag{35}$$

We finalize the proof of Theorem 3.1, by using the RHS of (35) as $M(\ell_1,t)$ in (28). Thus, $\mathrm{P}\{\mathcal{E}_i\} \leq \eta$ holds if

$$\max_{(\ell_1,t)\in\mathcal{I}} \left\{ \log\frac{4s^2 N(\ell_1,t)}{\eta} - \frac{N\ell_1\rho_{\min}}{24\beta(10+3L/s)} \right\} \leq 0. \tag{36}$$

The validity of (36), in turn, is guaranteed if

$$N \geq \frac{24\beta(10+3L/s)}{\rho_{\min}\ell_1}\big(\log s^2 N(\ell_1,t) + \log(4/\eta)\big), \tag{37}$$

for all $(\ell_1,t) \in \mathcal{I}$. Since $s \leq p/3$ (cf. (9)),

$$s^2 N(\ell_1,t) \leq \binom{p}{s}^4 \overset{(a)}{\leq} \left[\frac{pe}{s}\right]^{4s} \tag{38}$$

where (a) is due to $\binom{p}{q} \leq \left(\frac{pe}{q}\right)^q$ [13]. Combining (37) and (38), we finally obtain (21) of Theorem 3.1.

5. REFERENCES

[1] A. Jung, N. Q. Tran, and A. Mara, "When is network lasso accurate?" *arXiv:1704.02107*, 2017.

[2] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The M.I.T. Press, 2009.

[3] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Stat.*, vol. 34, no. 3, pp. 1436–1462, 2006.

[4] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Stat.*, vol. 38, no. 3, pp. 1287–1319, 2010.

[5] A. Jung, G. Hannak, and N. Görtz, "Graphical LASSO Based Model Selection for Time Series," *IEEE Sig. Proc. Letters*, vol. 22, no. 10, pp. 1781–1785, Oct. 2015.

[6] A. Jung, "Learning the conditional independence structure of stationary time series: A multitask learning approach," *IEEE Trans. Signal Processing*, vol. 63, no. 21, Nov. 2015.

[7] A. Jung, G. Tauböck, and F. Hlawatsch, "Compressive spectral estimation for nonstationary random processes," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3117–3138, May 2013.

[8] 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.

[9] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 151–172, 2000.

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

[11] P. Danaher, P. Wang, and D. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 76, no. 2, pp. 373–397, 3 2014.

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

[13] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge Univ. Press, 1985.