ON THE INFORMATION-THEORETIC LIMITS OF GRAPHICAL MODEL SELECTION FOR GAUSSIAN TIME SERIES

Gabor Hannak, Alexander Jung, Norbert Goertz

Institute of Telecommunications
Vienna University of Technology, Austria
Email: {ghannak, ajung, ngoertz}@nt.tuwien.ac.at

ABSTRACT

We consider the problem of inferring the conditional independence graph (CIG) of a multivariate stationary dicrete-time Gaussian random process based on a finite length observation. Using information-theoretic methods, we derive a lower bound on the error probability of any learning scheme for the underlying process CIG. This bound, in turn, yields a minimum required sample-size which is necessary for any algorithm regardless of its computational complexity, to reliably select the true underlying CIG. Furthermore, by analysis of a simple selection scheme, we show that the information-theoretic limits can be achieved for a subclass of processes having sparse CIG. We do not assume a parametric model for the observed process, but require it to have a sufficiently smooth spectral density matrix (SDM).

Index Terms— CIG, Fano-inequality, stationary time series

1. INTRODUCTION

We consider multivariate time series, i.e., vector-valued discrete time stationary random processes $\mathbf{x}[n] = (x_1[n], \ldots, x_p[n])^T, n \in \mathbb{Z}$. An important problem within multivariate time series analysis, e.g. in the context of medical data or environmental monitoring data, is the characterization of the interdependencies between the individual process components [1, 2]. A particular representation of the statistical relations governing the process components is obtained via the conditional independence graph (CIG) [3]. In particular, the CIG associated with the time series $\mathbf{x}[n]$ is an undirected graph with node set $V = \{1, \ldots, p\}$, where a specific node r represents the component process $x_r[n]$ and an edge (r, r') represents a dependency between process components $x_r[n]$ and $x_{r'}[n]$.

By graphical model selection, we refer to the problem of determining the underlying CIG based on a finite-length observation of the time series. The problem of graphical model selection for the case of Gaussian Markov random fields (GMRF), considered in [4], is contained in our setting as the special case of a white (i.i.d.) process. Thus, our work

can be regarded as a generalization of [4] in the sense that we allow for temporal correlation of the observed samples.

Numerous graphical model selection schemes for vector-valued stationary processes have been proposed [5–9]. Most of them are based on finite-dimensional parametric models. In particular, [6–8] model the observed process as a vector-valued autoregressive (VAR) process. Recently, the authors of [9] proposed a nonparametric selection scheme which does not rely on a parametric process model but requires only certain smoothness constraints to be satisfied by the process spectral density matrix (SDM). A performance analysis for the proposed scheme in [9] provides sufficient conditions on the observed sample-size that enables reliable model selection.

In this work we complement the sufficient conditions presented in [9] with lower-bounds on the sample-size required by any graphical model selection scheme to be reliable. We highlight that, similar to [4], these necessary conditions apply to any model selection scheme regardless of its computational complexity.

Our results apply to the high-dimensional regime, where the process dimension p and the sample-size N may diverge simultaneously. In particular, we allow for scenarios where the sample-size N is much smaller than the process dimension p. To allow for accurate selection schemes in the high-dimensional regime, one needs to require additional structural properties of the process. The structure induced by requiring the CIG to be sparse allows for reliable selection even in the high-dimensional regime. Our results can be used as a standard against which the performance (in terms of required sample-size) of a specific selection scheme can be compared. We begin in Section 2 with a discussion of the set-up and the problem formulation. In Section 3 we introduce the main results of this work, Theorem 3.1 and Theorem 3.2, and discuss their consequences. In Section 4 we give a high-level outline of the proofs for the main results.

Notation. The identity matrix of dimension $L \times L$ is denoted $\mathbf{I}_L$, without the subscript if the dimension is clear by context. We denote the kth column of the identity matrix $\mathbf{I}$ by $\mathbf{e}_k$. Given a matrix $\mathbf{A}$, we denote its entry in the ith row and jth column by $(\mathbf{A})_{ij}$. The set $\{1, \ldots, p\}$ is denoted by

$[p]$. The Kronecker product of the matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted $\mathbf{A} \otimes \mathbf{B}$. We define the Kronecker delta $\delta[n]$ as $\delta[0] = 1$ and $\delta[n] = 0$ otherwise. Given an index set $\mathcal{A}$, we denote $\mathbf{I}_\mathcal{A} = \sum_{r \in \mathcal{A}} \mathbf{e}_r \mathbf{e}_r^T$ and $\mathbf{1}_\mathcal{A} = \sum_{r \in \mathcal{A}} \mathbf{e}_r$, respectively.

2. PROBLEM FORMULATION

Consider a zero-mean multivariate Gaussian time series $\mathbf{x}[n] = (x_1[n], \ldots, x_p[n])^T$, $n \in \mathbb{Z}$. The temporal dependence between the samples $\mathbf{x}[n]$ is captured by the matrix-valued autocorrelation function (ACF)

$$\mathbf{R}_x[m] = \mathrm{E}\left\{ \mathbf{x}[m]\,\mathbf{x}^T[0] \right\},$$

assumed to be summable, i.e., $\sum_{k=-\infty}^{\infty} \|\mathbf{R}_x[m]\|_\infty < \infty$. The spectral density matrix (SDM),

$$\mathbf{S}_x(\theta) = \sum_{m=-\infty}^{\infty} \mathbf{R}_x[m] \exp(-j2\pi\theta m), \tag{1}$$

describes the correlation structure between the process components in the frequency domain. In what follows, we assume

$$1 \le \lambda_{\min}(\mathbf{S}_x(\theta)) \le \lambda_{\max}(\mathbf{S}_x(\theta)) \le \mathrm{B} < \infty, \tag{2}$$

where $\lambda_{\min}(\mathbf{S}_x(\theta))$ and $\lambda_{\max}(\mathbf{S}_x(\theta))$ denote the minimum and maximum eigenvalue of the SDM, respectively. The lower bound in (2) ensures certain Markov properties of the CIG [10] and the upper bound follows from the summability of the ACF.

Our analysis applies to processes with a smooth SDM, i.e., the entries of $\mathbf{S}_x(\theta)$ are smooth functions. Due to the Fourier relationship (1), these smoothness constraints can be expressed via the ACF moment

$$\mu_x = \sum_{m=-\infty}^{\infty} |m| \|\mathbf{R}_x[m]\|_\infty. \tag{3}$$

For a small moment μ_x, the ACF has to be well concentrated around $m = 0$.

Given a process $\mathbf{x}[n]$, we define its underlying conditional independence graph (CIG) as $\mathcal{G} = (V, E)$ with node set $V = [p]$ and edge set $E \subseteq V \times V$. The nodes of $\mathcal{G}$ represent the scalar component processes $x_r[n]$. The edge set is characterized by requiring $(r, r') \notin E$ if and only if the process components $x_r[n]$ and $x_{r'}[n]$ are conditionally independent given the remaining components $\{x_t[n] \mid t \in [p] \setminus \{r, r'\}\}$. For a Gaussian process with SDM satisfying (2), the CIG can be characterized conveniently in terms of the SDM. In particular [11],

$$(r, r') \notin E, \text{ if and only if } (\mathbf{S}_x^{-1}(\theta))_{rr'} = 0 \quad \forall \theta \in [0, 1). \tag{4}$$

To quantify the strength of the dependence between connected process components, we define the minimum partial spectral coherence of a process $\mathbf{x}[n]$ as

$$\rho_x = \min_{(r,r') \in E} \left(\int_0^1 \frac{\left(\mathbf{S}_x^{-1}(\theta)\right)^2_{rr'}}{\left(\mathbf{S}_x^{-1}(\theta)\right)_{rr} \left(\mathbf{S}_x^{-1}(\theta)\right)_{r'r'}} d\theta \right)^{\frac{1}{2}}. \tag{5}$$

The neighborhood and degree of a node r are defined as

$\mathcal{N}(r) = \{r' \mid (r, r') \in E\}$, and as $d_r = |\mathcal{N}(r)|$, respectively. We consider processes whose CIG have bounded degrees, i.e., for some (typically small) $d_{\max}$

$$d_r \le d_{\max} \quad \forall r \in [p]. \tag{6}$$

In the following $\mathcal{M}(p, d_{\max}, \rho_{\min})$ denotes the class of p dimensional vector processes whose underlying CIGs have maximum degree at most $d_{\max}$, whose SDMs fulfill (2) with some constant B and have minimum partial coherence not smaller than $\rho_{\min}$.

A graphical model selection scheme $\widehat{\mathcal{G}}$ maps the observed samples $\mathbf{X} = (\mathbf{x}[1], \ldots, \mathbf{x}[N]) \in \mathbb{R}^{p \times N}$ to an estimate $\widehat{\mathcal{G}}$ of the true CIG $\mathcal{G}$. We define the maximum selection error probability of a scheme $\widehat{\mathcal{G}}(\cdot)$ as

$$p_{err}(\widehat{\mathcal{G}}) = \max_{x \in \mathcal{M}} \Pr\left\{ \widehat{\mathcal{G}}(\mathbf{X}) \ne \mathcal{G} \right\}. \tag{7}$$

For a specific sample-size N, we define the minimax detection error as

$$p_{err}(N) = \min_{\widehat{\mathcal{G}}} p_{err}(\widehat{\mathcal{G}}).$$

We then study conditions on the sample-size $N(p, d_{\max}, \rho_{\min})$ as a function of the remaining problem parameters, such that asymptotically reliable selection is achievable, that is, $p_{err}(N) \to 0$ as $N(p, d_{\max}, \rho_{\min}) \to \infty$.

3. MAIN RESULTS

3.1. Necessary Conditions for Consistent Model Selection

For our first result, we closely follow the argument in [4] used to derive lower-bounds on the required sample-size for i.i.d. samples. The observation of i.i.d. samples is contained in our setup as the special case obtained for $\mu_x = 0$, implying that the SDM $\mathbf{S}_x(\theta)$ is flat, i.e., it does not depend on θ.

Theorem 3.1. *Consider a process in the class* $\mathcal{M}(p, d_{\max}, \rho_{\min})$ *with* $\rho_{\min} \in (0, \frac{1}{4}]$. *A necessary condition for asymptotically reliable graphical model selection is*

$$N > \frac{\log \binom{p}{2} - 1}{4\rho_{\min}^2}. \tag{8}$$

A proof is sketched in Section 4.

The lower bound (8) on the sample-size completely ignores the correlation width of the process, which is quantified by μ_x. Intuitively, we would expect that with increasing correlation width the required sample-size becomes larger. This is also reflected by the sufficient conditions on the sample-size in [9] for a novel nonparametric selection scheme. In particular the results of [9] suggest that the sample-size has to grow proportionally to μ_x^2.

Another argument supporting the intuition that that the required sample-size N has to be larger for increasing μ_x is that for larger values of μ_x the SDM has faster variations over θ. Therefore, in order to determine the zeros of the inverse SDM, we have to estimate the SDM values at more sampling points (placed denser). However, the values of the SDM at different frequencies θ are strongly coupled via the condition (4).

This implies via (6) (with a small $d_{\max}$) that the values of the inverse SDM $\mathbf{S}^{-1}(\theta)$ have a small joint support, which corresponds to the edge set E of the CIG.

3.2. Sufficient Condition for Consistent Model Selection

We next show that, at least for the special case of $d_{\max} = 1$, the coupling via (4) compensates the effect of increasing correlation width μ_x such that the required sample-size is independent of μ_x. This will be accomplished by analyzing a specific model selection scheme. This scheme is similar to the exhaustive search decoder used in [12] for the derivation of sufficient conditions for sparsity recovery in the high-dimensional sparse linear model.

In order to keep the argument as simple as possible, we assume that

- the ACF is exactly supported within $\{-\frac{N}{2} + 1, \ldots, 0, \ldots, \frac{N}{2} - 1\}$, i.e.,

$$\mathbf{R}_x[m] = 0 \text{ if } |m| \geq \frac{N}{2}, \qquad (9)$$

- the ACF is real-valued and symmetric (instead of being merely Hermitian symmetric), i.e.,

$$\mathbf{R}_x[-m] = \mathbf{R}_x[m]. \qquad (10)$$

A specific subclass of processes satisfying these assumptions is obtained by applying a real-valued scalar filter componentwise to a white noise vector process, i.e.,

$$\mathbf{x}[n] = \sum_{m=0}^{K-1} h[m]\mathbf{w}[n-m], \qquad (11)$$

with $h[m]$ being a length-K filter impulse response with $K < N/4$. We assume that the filter $h[n]$ is normalized such that $\sum_{m=-\infty}^{\infty} h^2[n] = 1$, which by Parseval's theorem implies $\sum_{f=0}^{2N-1} |H(\frac{f}{2N})|^2 = 2N$, where $H(\theta) = \sum_{n=-\infty}^{\infty} h[n] \exp(-j2\pi n\theta)$ denotes the discrete time Fourier transform of the impuse response $h[n]$.

Here, $\mathbf{w}[n]$ is an i.i.d. zero-mean Gaussian process with marginal covariance matrix $\mathbf{C}$, i.e., $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$. The associated precision matrix $\mathbf{C}^{-1}$ has at most one non-zero off-diagonal entry in each row. This implies that the CIG of $\mathbf{x}[n]$ satisfies (6) with $d_{\max} = 1$.

The constraint (9) is satisfied for processes with a smooth SDM or, equivalently, a small moment μ_x. Typically the ACF will not be exactly zero for $|m| \geq N/2$. We make this idealized assumption only to keep our argument as simple as possible. However, we expect that our main conclusions are also valid for any process with sufficiently small correlation width, i.e., small μ_x.

Following a method in [13], we define, given the observations $\mathbf{x}[1], \ldots, \mathbf{x}[N+1]$, the enlarged observation set $\tilde{\mathbf{x}}[1], \ldots, \tilde{\mathbf{x}}[2N]$ by

$$\tilde{\mathbf{x}}[n] = \begin{cases} \mathbf{x}[n] & \text{for } n \in [N+1] \\ \mathbf{x}[2N-n+2] & \text{for } n \in [2N] \setminus [N+1]. \end{cases}$$

Because of (9) and (10), the covariance matrix $\widetilde{\mathbf{C}}$ of the stacked vector $\tilde{\mathbf{x}} := \left(\tilde{\mathbf{x}}^T[1], \ldots, \tilde{\mathbf{x}}^T[2N] \right)^T$ can be shown to be a block circulant matrix [14] with first row given by

$$\left(\mathbf{R}_x[0], \ldots, \mathbf{R}_x[N], \mathbf{R}_x[N-1], \ldots, \mathbf{R}_x[2] \right). \qquad (12)$$

Let us define the DFT of the enlarged observation set as

$$\hat{\mathbf{x}}[f] := (1/\sqrt{2N}) \sum_{n \in [2N]} \tilde{\mathbf{x}}[n] \exp \left(-j2\pi \frac{(f-1)(n-1)}{2N} \right)$$

Some calculation reveals that the covariance matrix $\widehat{\mathbf{C}}$ of the stacked vector $\hat{\mathbf{x}} := \left(\hat{\mathbf{x}}^T[1], \ldots, \hat{\mathbf{x}}^T[2N] \right)^T$ satisfies

$$\widehat{\mathbf{C}} = \frac{1}{2N} \mathbf{P} \widetilde{\mathbf{C}} \mathbf{P}^H,$$

with the matrix $\mathbf{P}$ as defined in [14, p. 809]. Since $\widetilde{\mathbf{C}}$ is block circulant, it follows that $\widehat{\mathbf{C}}$ is a block diagonal matrix with the fth $p \times p$ diagonal block given by the fth bin of the DFT of the row (12), which can be shown to coincide with $\mathbf{S}_x((f-1)/2N)$. Therefore, the vectors $\hat{\mathbf{x}}[f]$ are independent (across f) zero-mean Gaussian vectors with covariance matrix $\mathbf{S}_x((f-1)/2N)$.

Let $\hat{\mathbf{x}}_r$ denote the rth row of $(\hat{\mathbf{x}}[1], \ldots, \hat{\mathbf{x}}[2N])$. The estimation of the neighborhood $\mathcal{N}(r)$, under the assumption that the true CIG satisfies (6) with $d_{max} = 1$, is carried out for every $r \in V$ by:

- for all $r' \in [p] \setminus \{r\}$ compute the statistic

$$Z(r') := \hat{\mathbf{x}}_r^T \hat{\mathbf{x}}_{r'}, \qquad (13)$$

- determine the maximizing index

$$\hat{r} := \underset{r' \in [p] \setminus \{r\}}{\operatorname{argmax}} |Z(r')|, \qquad (14)$$

- compare the maximum statistic with threshold η to obtain

$$\widehat{\mathcal{N}}(r) = \begin{cases} \{\hat{r}\} & \text{if } |Z(\hat{r})| \geq \eta \\ \emptyset & \text{otherwise.} \end{cases} \qquad (15)$$

The choice of η will be discussed in Section 4.2. Note that this algorithm does not necessarily produce a CIG satisfying (6) with $d_{\max} = 1$.

The following result shows that for the subclass of $\mathcal{M}(p, d_{\max} = 1, \rho_{\min})$ given by processes of the form (11), the simple selection scheme (13)-(15) achieves the information theoretic limit stated in Theorem 3.1.

Theorem 3.2. *Consider a process in the class* $\mathcal{M}(p, d_{\max} = 1, \rho_{\min})$ *generated according to* (11). *A sufficient condition on the sample-size* N *for achieving probability of incorrect selection not larger than* δ *using the selection scheme given by* (13)-(15) *is*

$$N > \frac{32B^4}{\rho_{\min}^2} \log \left(\frac{2p^2}{\delta} \right). \qquad (16)$$

A proof is sketched in Section 4.

Discussion. The bound (16) matches the scaling of the nec-

essary condition (8) in Theorem 3.1, in particular, it scales proportional to $\frac{1}{\rho_{\min}^2}$ and to $\log p$. We deduce that for the processes of the form (11) in the subclass $\mathcal{M}(p, d_{\max} = 1, \rho_{\min})$ the theoretic limit stated in Theorem 3.1 can be achieved. Thus, since (16) does not depend on μ_x, for this special case the temporal correlation of the process, quantified by μ_x, does not increase the required sample-size for reliable graphical model selection.

4. PROOF SKETCHES

4.1. Proof of Theorem 3.1

For the derivation of Theorem 3.1, we closely follow the method put forward in [4]. In particular, the result is based on a finite ensemble $\mathcal{M}_0$ containing $M \in \mathbb{N}$ different processes $\mathbf{x}^{(i)}[n] \in \mathcal{M}(p, d_{\max}, \rho_{\min})$ with associated SDM $\mathbf{S}_i(\theta)$ and CIG $\mathcal{G}(i)$, respectively.

Assuming that the observed process $\mathbf{x}[n]$ is taken uniformly at random out of $\mathcal{M}_0 \subseteq \mathcal{M}(p, d_{\max}, \rho_{\min})$, we may interpret the graphical model selection problem as a communication problem: using a random index i, distributed uniformly over the set $[M]$, we select the process $\mathbf{x}^{(i)}[n]$ as the observed process, i.e., $\mathbf{x}[n] = \mathbf{x}^{(i)}[n]$. Based on the samples $\mathbf{X} = (\mathbf{x}[1], \ldots, \mathbf{x}[N])$, the problem of selecting the true graphical model is now equivalent to detecting $\mathcal{G}(i)$. A selection rule $\widehat{\mathcal{G}}(\mathbf{X})$ can be interpreted as a decoder, mapping the observation $\mathbf{X}$ to an estimate of the true CIG $\mathcal{G}(i)$.

The maximum probability $p_{err}(\hat{G})$ (cf. (7)) probability (7) of any selection rule can be lower bounded via Fano's inequality [15] as

$$p_{err}(\widehat{\mathcal{G}}) \geq 1 - \frac{\mathrm{I}(\mathbf{X}; i) + 1}{H(\mathcal{G}(i))} . \tag{17}$$

Here, $H(\mathcal{G}(i))$ denotes the entropy of the random CIG $\mathcal{G}(i)$ associated with the process $\mathbf{x}^{(i)}[n] \in \mathcal{M}_0$, which is selected uniformly at random from $\mathcal{M}_0$. Since the bound in (17) applies to any selection rule, it is also a lower bound on the minimax error probability $p_{err}(N)$, i.e., $p_{err}(N) \geq 1 - \frac{\mathrm{I}(\mathbf{X};i)+1}{H(\mathcal{G}(i))}$. Therefore, for asymptotically reliable model selection, i.e., $\lim_{N \to \infty} p_{err}(N) = 0$, a necessary condition is

$$\mathrm{I}(\mathbf{X}; i) \geq H(\mathcal{G}(i)) - 1. \tag{18}$$

Based on (18), we will now derive necessary conditions on the sample-size N by using upper bounds on the mutual information $\mathrm{I}(\mathbf{X}; i)$ which depend explicitly on N. In particular, since given the index i, the observation $\mathrm{vec}(\mathbf{X})$ is a multivariate normal vector with zero-mean and covariance matrix $\mathbf{C}_i := \mathrm{E}\{\mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^T | i\}$, we can use the following *entropy-based* upper bound [4, 16]

$$\mathrm{I}(\mathbf{X}; i) \leq \log|\overline{\mathbf{C}}| - (1/M) \sum_{i \in [M]} \log|\mathbf{C}_i|, \tag{19}$$

where $\overline{\mathbf{C}} := (1/M) \sum_{i \in [M]} \mathbf{C}_i$.

Given p nodes, let $\mathcal{S}(i)$ denote an enumeration of the $\bar{p} = \binom{p}{2}$ different simple graphs (with node set $[p]$) containing a

single edge. Let us define a bijective map $\mathcal{S}(i)$ which assigns an edge (r_i, r_i') uniquely to an index $i \in [\bar{p}]$. Consider the ensemble $\mathcal{M}_0$ of size $M = \bar{p}$, constituted by the Gaussian processes $\mathbf{x}^{(i)}[n]$ with SDM

$$\mathbf{S}_i(\theta) = 2\mathbf{I} - \frac{2\rho_{\min}}{1 + 4\rho_{\min}} \mathbf{1}_{\mathcal{S}(i)} \mathbf{1}_{\mathcal{S}(i)}^T \tag{20}$$

for $i \in [\bar{p}]$. Note that the SDM in (20) does not depend on θ. Therefore, due to the Fourier relationship (1), the corresponding ACF is given by $\mathbf{R}_i[m] = \mathbf{S}_i(0)\delta[m]$, implying that $\mu_x = 0$ for all $\mathbf{x}[n] \in \mathcal{M}_0$ (cf. (3)).

The CIG $\mathcal{G}(i)$ associated with (20) contains a single edge between nodes r_i and $r_{i'}$. Therefore, (6) is satisfied for any $d_{\max} \geq 1$. Moreover, each process $\mathbf{x}^{(i)}[n] \in \mathcal{M}_0$ with SDM given by (20) has a unique CIG $\mathcal{G}(i)$, i.e., $\mathcal{G}(i) \neq \mathcal{G}(i')$ for $i \neq i'$ and, in turn,

$$H(\mathcal{G}(i)) = \log_2 |M| = \log_2 \bar{p}. \tag{21}$$

If $\rho_{\min} \leq 1/4$, the eigenvalues of the SDM in (20) satisfy (2) (with $B = 3$). Applying the matrix inversion lemma [17] to (20), we obtain

$$\mathbf{S}_i^{-1}(\theta) = \mathbf{I} + 2\rho_{\min} \mathbf{1}_{\mathcal{S}(i)} \mathbf{1}_{\mathcal{S}(i)}^T. \tag{22}$$

Using (22) and $\rho_{\min} \leq 1/4$, the minimum partial coherence ρ_x (cf. (5)) of the process $\mathbf{x}^{(i)}[n]$ can be shown to satisfy

$$\rho_x \geq \rho_{\min} .$$

Therefore, the process $\mathbf{x}^{(i)}[n]$ belongs to $\mathcal{M}(p, d_{\max}, \rho_{\min})$ for any $\rho_{\min} \leq 1/4$ and $d_{\max} \geq 1$. The covariance matrix $\mathbf{C}_i = \mathrm{E}\{\mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^T | i\}$ of the observation $\mathbf{X}$, given the index i satisfies

$$\mathbf{C}_i = \mathbf{I}_N \otimes \mathbf{S}_i(0) \tag{23}$$

and, in turn,

$$\overline{\mathbf{C}} = (1/M) \sum_{i \in [M]} \mathbf{C}_i = \mathbf{I}_N \otimes \overline{\mathbf{S}} \tag{24}$$

with $\overline{\mathbf{S}} := (1/M) \sum_{i \in [M]} \mathbf{S}_i(0)$. Inserting (23) and (24) into the bound (19),

$$\mathrm{I}(\mathbf{X}; i) \leq N \Big(\log|\overline{\mathbf{S}}| - (1/M) \sum_{i \in [M]} \log|\mathbf{S}_i(0)| \Big), \tag{25}$$

where we used the identity $|\mathbf{I}_N \otimes \mathbf{B}| = |\mathbf{B}|^N$ (cf. [18, Ch. 4]). Note that the matrix $\mathbf{S}_i(0)$ has one eigenvalue equal to $2(1 - 2a/(1 + 2a))$ and $p - 1$ eigenvalues equal to 2. Furthermore, setting $\gamma := 1 - \frac{2a}{(1+2a)p} + \frac{2a}{(1+2a)p(p-1)}$, the matrix $\overline{\mathbf{S}}$ has one eigenvalue equal to $2(\gamma - \frac{2a}{(p-1)(1+2a)})$ and $p - 1$ eigenvalues equal to 2γ. Using the inequality $\log(1 + x) \leq x$ for $x \geq 0$ and closely following the calculation in [16, Sec. 4.5.1.], one obtains from (25)

$$\mathrm{I}(\mathbf{X}; i) \leq N 16\rho_{\min}^2. \tag{26}$$

Inserting (26) and (21) into (18) yields the bound (8).

4.2. Proof of Theorem 3.2

Consider the selection scheme $\widehat{\mathcal{N}}(r)$ described through (13)-(15) for the neighborhood of a specific node r. We aim

at bounding the probability of failing to recover the correct neighborhood $\mathcal{N}(r)$, $p_{err,r} = \Pr\left\{\widehat{\mathcal{N}}(r) \neq \mathcal{N}(r)\right\}$. We consider separately the two cases: $\mathcal{N}(r) = \emptyset$ and $\mathcal{N}(r) \neq \emptyset$.

I. Node r has no neighbor

In this case, a selection error of the exhaustive search decoder can only occur if for some node $r' \in V \setminus \{r\}$ the statistic exceeds the threshold, i.e., $|Z(r')| \geq \eta$. Thus, $p_a = \Pr\left\{\exists r' \in V \setminus \{r\} : |Z(r')| \geq \eta\right\}$ which, via a union bound, can be further bounded as

$$p_a \leq (p-1) \max_{r' \in V \setminus \{r\}} \Pr\left\{|Z(r')| \geq \eta\right\}.$$

II. Node r has a single neighbor

Suppose $(r,c) \in E$. The probability of erroneous detection of $\mathcal{N}(r)$ can be written as

$$p_b = \Pr\left\{\exists r' \in V \setminus \{r,c\} : |Z(r')| \geq \eta \text{ and } |Z(c)| \leq \eta\right\},$$

for some positive η. Using a union bound argument, this probability can be bounded as

$$p_b \leq (p-2)\Pr\left\{|Z(r')| \geq \eta\right\} + \Pr\left\{|Z(c)| \leq \eta\right\}.$$

Since $Z(\cdot)$ is the inner product of two zero-mean Gaussian vectors (cf. (13)), we can make direct use of [19, Lemma E.2] with $\mathbf{Q} = \mathbf{I}$ so separately bound p_a and p_b and in turn $p_{err,r}$ since $p_{err,r} \leq \max\{p_a, p_b\}$. In order to apply it to above derivations, we need to characterize $\mathrm{E}\left\{Z(c)\right\}$ and $\mathrm{E}\left\{Z(r')\right\}$ for $r' \notin \mathcal{N}(r)$. Elementary calculations reveal that

$$\mathrm{E}\left\{\hat{\mathbf{x}}_r^T \hat{\mathbf{x}}_c\right\} = (\mathbf{C}_0)_{rc} \sum_{f=1}^{2N} \left| H\left(\frac{f}{2N}\right) \right|^2 = 2N\,(\mathbf{C}_0)_{rc}$$

and $\frac{\rho_{\min}}{B} \leq |(\mathbf{C}_0)_{rc}|$. Further, $\mathrm{E}\left\{\hat{\mathbf{x}}_r^T \hat{\mathbf{x}}_{r'}\right\} = 0$ for any $r' \notin \mathcal{N}(r)$. Based on (2), we can bound the spectral norm of the covariance matrix $\mathbf{C}_{\hat{x}_r} = E\left\{\hat{\mathbf{x}}_r \hat{\mathbf{x}}_r^T\right\}$ as $\|\mathbf{C}_{\hat{\mathbf{x}}_r}\|_2 \leq B \quad \forall r \in V$. Choosing the threshold in (15) as $\nu = \rho_{\min}/(2B)$, we obtain via [19, Lemma E.2] the bound

$$p_{err,r} \leq 2p \exp\left(\frac{-N\rho_{\min}^2}{32B^4}\right).$$

Using again a union bound argument, we obtain that the probability p_{err} of inferring an incorrect graph, that is, the probability of selecting at least one out of p neighbor sets incorrectly, is bounded as

$$p_{err} \leq p \cdot p_{err,r} \leq 2p^2 \exp\left(\frac{-N\rho_{\min}^2}{32B^4}\right).$$

Thus, if we require that the probability of a selection error does not exceed a small number δ, we obtain the sufficient condition (16) on the sample-size N.

5. CONCLUSIONS

We characterized the information theoretic limits of graphical model selection for Gaussian time series by deriving a necessary condition on the sample-size such that reliable selection may be possible. For a specific subclass of time series with extremely sparse CIGs we showed that the necessary condi-

tion is sharp. In particular, we verified that a simple selection scheme is successful for a sample-size close to the information theoretic limit. Somewhat unexpected, our analysis reveals that in general the required sample-size is independent of the correlation width, i.e., it does not depend on the amount of smoothness of the SDM. This suggests, in turn, that the sufficient condition presented in [9] for a novel selection scheme is far from optimal.

REFERENCES

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

[2] Helmut Lütkepohl, *New Introduction to Multiple Time Series Analysis*, Springer, New York, 2005.

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

[4] W. Wang, M.J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for gaussian markov random field," in *Proc. IEEE ISIT-2010*, Austin, TX, Jun. 2010, pp. 1373–1377.

[5] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on time series using structural equation models," *ArXiv e-prints*, Jul. 2012.

[6] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Advances in Neural Information Processing Systems 23*, Vancouver, CN, 2010, pp. 172–180.

[7] J. Songsiri, J. Dahl, and L. Vandenberghe, "Graphical models of autoregressive processes," in *Convex Optimization in Signal Processing and Communications*, Y. Eldar and D. Palomar, Eds., pp. 89–116. Cambridge Univ. Press, Cambridge, UK, 2010.

[8] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *Journal of Machine Learning Research*, vol. 11, pp. 2671–2705, 2010.

[9] A. Jung, R. Heckel, H. Bölcskei, and F. Hlawatsch, "Compressive nonparametric graphical model selection for stationary time series," in *Proc. IEEE ICASSP 2014*, Florence, Italy, May 2014.

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

[11] R. Brillinger, "Remarks concerning graphical models for time series and point processes," *Revista de Econometria*, vol. 16, pp. 1–23, 1996.

[12] 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, Jun. 2009.

[13] T. T. Cai, Z. Ren, and H. H. Zhou, "Optimal rates of convergence for estimating toeplitz covariance matrices," *Probability Theory and Related Fields*, vol. 156, no. 1-2, pp. 101–143, 2013.

[14] T. De Mazancourt and D. Gerlic, "The inverse of a block-circulant matrix," *IEEE Trans. Antennas and Propagation*, vol. AP-31, no. 5, Sep. 1983.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New Jersey, 2 edition, 2006.

[16] W. Wang, *Sparse signal recovery using sparse random projections*, Ph.D. thesis, EECS Department, University of California, Berkeley, Dec 2009.

[17] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, Jan. 1981.

[18] Roger A. Horn and Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1991.

[19] A. Jung, R. Heckel, H. Bölcskei, and F. Hlawatsch, "Compressive nonparametric graphical model selection for stationary time series: A multitask learning approach," Tech. Rep., Vienna University of Technology, Institute of Telefommunications, March 2014.