LEARNING CONDITIONAL INDEPENDENCE STRUCTURE FOR HIGH-DIMENSIONAL UNCORRELATED VECTOR PROCESSES

Nguyen Tran Quang and Alexander Jung

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

ABSTRACT

We formulate and analyze a graphical model selection method for inferring the conditional independence graph of a high-dimensional nonstationary Gaussian random process (time series) from a finite-length observation. The observed process samples are assumed uncorrelated over time but having a time-varying marginal distribution. The selection method is based on testing conditional variances obtained for small subsets of process components. This allows to cope with the high-dimensional regime, where the sample size can be (much) smaller than the process dimension. We characterize the required sample size such that the proposed selection method is successful with high probability.

Index Terms— Sparsity, graphical model selection, conditional variance testing, high-dimensional statistics.

1. INTRODUCTION

Consider a zero-mean, d-dimensional Gaussian discrete-time random process (time series)

$$\mathbf{x}[n] := \big(x_1[n], \ldots, x_d[n]\big)^T \in \mathbb{R}^d, \text{ for } n = 1, \ldots, N.$$

Based on the observation of a single process realization of length N, we are interested in learning the conditional independence graph (CIG) [1–4] of $\mathbf{x}[n]$. The learning method shall cope with the *high-dimensional regime*, where the number d of process components is (much) larger than the number N of observed vector samples [5–11]. In this regime, accurate estimation of the CIG is only possible under structural assumptions on the process $\mathbf{x}[n]$. In this work, we will consider processes whose CIGs are *sparse* in the sense of containing relatively few edges. This problem is relevant, e.g., in the analysis of medical diagnostic data (EEG) [8], climatology [12] and genetics [13].

Most of the existing approaches to graphical model selection (GMS) for Gaussian vector processes are based on modelling the observed data either as i.i.d. samples of a single random vector, or as samples of a stationary random process. For nonstationary processes, the problem of inferring time-varying graphical models has been considered [14, 15]. By contrast, we assume one single CIG representing the correlation structure for all samples $\mathbf{x}[n]$, which are assumed uncorrelated but having different marginal distributions, which are determined by the covariance matrix $\mathbf{C}[n]$.

Contributions: Our main conceptual contribution resides in the formulation of a simple GMS method for uncorrelated

nonstationary Gaussian processes, which is based on conditional variance tests. For processes having a sparse CIG, these tests involve only small subsets of process components. We provide a lower bound on the sample size which guarantees that the correct CIG is selected by our GMS method with high probability. This lower bound depends only logarithmically on the process dimension and polynomially on the maximum degree of the true CIG. Moreover, our analysis reveals that the crucial parameter determining the required sample size is the minimum partial correlation of the process.

Outline: The remainder of this paper is organized as follows. In Section 2, we formalize the considered process model and the notion of a CIG. In particular, we will state four assumptions on the class of processes that will be considered in the following. Section 3 presents a GMS method based on conditional variance testing. There, we also state and discuss a lower bound on the sample size guaranteeing success of our GMS method with high probability.

Notation: Given a d-dimensional process $\mathbf{x}[1], \ldots, \mathbf{x}[N]$ of length N , we denote a scalar component process as $\mathbf{x}_i[\cdot] := \big(x_i[1], \ldots, x_i[N]\big)^T \in \mathbb{R}^N$ for $i \in \{1, \ldots, d\}$. The Kronecker-delta is denoted $\delta_{n,n'}$ with $\delta_{n,n'} = 1$ if $n = n'$ and $\delta_{n,n'} = 0$ else. By $\mathfrak{S}^r_{s_{\max}}$, we denote all subsets of $\{1, \ldots, d\}$ of size at most $s_{\max}$ and which do not contain r. We denote by $\mathbf{A}_{\{\mathcal{A}, \mathcal{B}\}}$ the submatrix with rows indexed by $\mathcal{A}$ and columns indexed by $\mathcal{B}$. Given a matrix $\mathbf{A}$, we define its infinity norm as $\|\mathbf{A}\|_\infty := \max_i \sum_j |A_{i,j}|$. The minimum and maximum eigenvalues of a positive semidefinite (psd) matrix $\mathbf{C}$ are denoted $\lambda_{\min}(\mathbf{C})$ and $\lambda_{\max}(\mathbf{C})$, respectively.

2. PROBLEM FORMULATION

Let $\mathbf{x}[n]$, for $n \in \{1, \ldots, N\}$, be a zero-mean d-dimensional, real-valued Gaussian random process of length N. We model the time samples $\mathbf{x}[n]$ as uncorrelated, and therefore independent due to Gaussianity. The probability distribution of the Gaussian process $\mathbf{x}[n]$ is fully specified by the covariance matrices $\mathbf{C}[n]$ which might vary with n. To summarize, in what follows we only consider processes conforming to the model

$$\{\mathbf{x}[n]\}_{n=1}^N \text{ jointly Gaussian zero-mean with}$$
$$\mathrm{E}\{\mathbf{x}[n]\mathbf{x}^T[n']\} = \delta_{n,n'}\mathbf{C}[n]. \tag{1}$$

The process model (1) is relevant for applications facing weakly dependent time series, so that samples which are sufficiently separated in time can be effectively considered as

uncorrelated [16]. Moreover, the process model (1) can be used as an approximation for the discrete Fourier transform of stationary processes with limited correlation width or fast decay of the autocovariance function [17–20]. Another setting where the model (1) is useful are vector-valued locally stationary processes, where a suitable local cosine basis yields approximately uncorrelated vector processes [21].

For our analysis we assume a known range within which the eigenvalues of the covariance matrices $\mathbf{C}[n]$ are guaranteed to fall.

Assumption 1. *The eigenvalues of the psd covariance matrices $\mathbf{C}[n]$ are bounded as*
$$0 < \alpha[n] \le \lambda_{\min}(\mathbf{C}[n]) \le \lambda_{\max}(\mathbf{C}[n]) \le \beta[n] \quad (2)$$
with known bounds $\beta[n] \ge \alpha[n] > 0$.

We highlight that for the actual implementation of our GMS method in Alg. 1, we do not require knowledge of the bounds $\alpha[n]$ and $\beta[n]$. Those are only required for the performance analysis of Alg. 1. It will be notationally convenient to associate with the observed samples $\mathbf{x}[1], \ldots, \mathbf{x}[N]$ the "time-wise" stacked vector
$$\mathbf{x} = (\mathbf{x}[1]^T, \ldots, \mathbf{x}[N]^T)^T \in \mathbb{R}^{Nd}$$
and the "component-wise" stacked vector
$$\tilde{\mathbf{x}} := (\mathbf{x}_1[\cdot]^T, \ldots, \mathbf{x}_d[\cdot]^T)^T \in \mathbb{R}^{Nd}.$$
We have, for some permutation matrix $\mathbf{P} \in \{0,1\}^{Nd \times Nd}$,
$$\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x}. \quad (3)$$

For data samples $\mathbf{x}[n]$ conforming to (1), the associated vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are zero-mean Gaussian random vectors, with covariance matrices
$$\mathbf{C}_x = \mathrm{E}\{\mathbf{x}\mathbf{x}^T\} = \begin{pmatrix} \mathbf{C}[1] & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{C}[N] \end{pmatrix}, \quad (4)$$
and
$$\mathbf{C}_{\tilde{x}} = \mathrm{E}\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\} = \begin{pmatrix} \mathbf{C}_{\tilde{x}}[1,1] & \cdots & \mathbf{C}_{\tilde{x}}[1,d] \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{\tilde{x}}[d,1] & \cdots & \mathbf{C}_{\tilde{x}}[d,d] \end{pmatrix}, \quad (5)$$
respectively. Due to (3), we have
$$\mathbf{C}_{\tilde{x}} = \mathbf{P}\mathbf{C}_x\mathbf{P}^T.$$
Since the permutation matrix $\mathbf{P}$ is orthogonal ($\mathbf{P}^T = \mathbf{P}^{-1}$), the precision matrix $\mathbf{K}_x := \mathbf{C}_x^{-1}$ is also block diagonal with diagonal blocks $\mathbf{K}[n] = (\mathbf{C}[n])^{-1}$. As can be verified easily, the (a,b)th $N \times N$ block $\mathbf{K}_{\tilde{x}}[a,b]$ of the matrix $\mathbf{K}_{\tilde{x}} = \mathbf{P}\mathbf{K}_x\mathbf{P}^T$ is diagonal:
$$\mathbf{K}_{\tilde{x}}[a,b] = \begin{pmatrix} (\mathbf{K}[1])_{\{a,b\}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & (\mathbf{K}[N])_{\{a,b\}} \end{pmatrix}. \quad (6)$$

2.1. Conditional Independence Graph

We now define the CIG of a d-dimensional Gaussian process $\mathbf{x}[n] \in \mathbb{R}^d$ as an undirected simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = \{1, 2, \ldots, d\}$. Node $j \in \mathcal{V}$ represents the process component $\mathbf{x}_j[\cdot] = (x_j[1], \ldots, x_j[N])^T$. An edge is absent between nodes a and b, i.e., $(a,b) \notin \mathcal{E}$ if the corresponding process components $\mathbf{x}_a[\cdot]$ and $\mathbf{x}_b[\cdot]$ are conditionally independent, given the remaining components $\{\mathbf{x}_r[\cdot]\}_{r \in \mathcal{V} \setminus \{a,b\}}$. Since we model the process $\mathbf{x}[n]$ as Gaussian (cf. (1)), this conditional independence can be read off conveniently from the inverse covariance (precision) matrix $\mathbf{K}_{\tilde{x}} := \mathbf{C}_{\tilde{x}}^{-1}$.

In particular, $\mathbf{x}_a[\cdot]$ are $\mathbf{x}_b[\cdot]$ are conditionally independent, given $\{\mathbf{x}_r[\cdot]\}_{r \in \mathcal{V} \setminus \{a,b\}}$ if and only if $\mathbf{K}_{\tilde{x}}[a,b] = \mathbf{0}$ [22, Prop. 1.6.6.]. Thus, we have the following characterization of the CIG $\mathcal{G}$ associated with the process $\mathbf{x}[n]$:
$$(a,b) \notin \mathcal{E} \text{ if and only if } \mathbf{K}_{\tilde{x}}[a,b] = \mathbf{0}. \quad (7)$$
Inserting (6) into (7) yields, in turn,
$$(a,b) \notin \mathcal{E} \text{ if and only if } (\mathbf{K}[n])_{\{a,b\}} = 0 \text{ for all } n \in [N]. \quad (8)$$

We highlight the coupling in the CIG characterization (8): An edge is absent, i.e., $(a,b) \notin \mathcal{E}$ only if the precision matrix entry $(\mathbf{K}[n])_{\{a,b\}}$ is zero *for all* $n \in \{1, \ldots, N\}$.

We will also need a measure for the strength of a connection between process components $\mathbf{x}_a[\cdot]$ and $\mathbf{x}_b[\cdot]$ for $(a,b) \in \mathcal{E}$. To this end, we define the *partial correlation* between $\mathbf{x}_a[\cdot]$ and $\mathbf{x}_b[\cdot]$ as
$$\rho_{a,b} := (1/N) \sum_{n=1}^{N} \alpha[n] \left[(\mathbf{K}[n])_{a,b} / (\mathbf{K}[n])_{a,a} \right]^2. \quad (9)$$
Inserting (8) into (9) shows that $(a,b) \notin \mathcal{E}$ implies $\rho_{a,b} = 0$.

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

Assumption 2. *For any edge $(a,b) \in \mathcal{E}$, the partial correlation $\rho_{a,b}$ (cf. (9))is lower bounded by a constant $\rho_{\min}$, i.e.,*
$$(a,b) \in \mathcal{E} \Rightarrow \rho_{a,b} \ge \rho_{\min}.$$

The CIG $\mathcal{G}$ of a vector-process $\mathbf{x}[n]$ is fully characterized by the neighborhoods $\mathcal{N}(r) = \{t \in \mathcal{V} : (r,t) \in \mathcal{E}\}$ of all nodes $r \in \mathcal{V}$. Many applications involve processes with these neighborhoods being small compared to the overall process dimension d. The CIG is then called *sparse* since it contains few edges compared to the complete graph.

Assumption 3. *The size of any neighborhood $\mathcal{N}(r)$, i.e., the degree of node r is upper bounded as*
$$|\mathcal{N}(r)| \le s_{\max},$$
where typically $s_{\max} \ll d$.

2.2. Slowly Varying Covariance

For several practically relevant settings, such as stationary processes with limited correlation width [17–20] or underspread nonstationary processes [23], the observed processes can be well approximated by the model (1) with the additional property of a *slowly varying* covariance matrix $\mathbf{C}[n]$ [14, 15].

Assumption 4. *For a (small) positive constant κ,*
$$\|\mathbf{C}[n_1] - \mathbf{C}[n_2]\|_{\infty} \le \kappa(|n_2 - n_1|/N). \quad (10)$$

In view of (10), for some $n_0 \in \{1, \ldots, N - L\}$ and block-length L such that $\kappa(L/N) \ll 1$, we may approximately model L consecutive samples $\mathbf{x}[n_0], \mathbf{x}[n_0 + 1], \ldots, \mathbf{x}[n_0 + L - 1]$ as i.i.d. zero-mean Gaussian vectors with covariance matrix $\mathbf{C} = (1/L) \sum_{n=n_0}^{n_0+L-1} \mathbf{C}[n]$. This suggests to partition the observed samples evenly into length-L blocks $\mathcal{B}_b = \{(b-1)L + 1, \ldots, bL\}$ for $b = 1, \ldots, B = N/L$.[1] We can estimate the covariance matrix of the samples within block $\mathcal{B}_b$ using the sample covariance matrix

$$\widehat{\mathbf{C}}[b] = (1/L) \sum_{n \in \mathcal{B}_b} \mathbf{x}[n]\mathbf{x}^T[n]. \tag{11}$$

3. GMS VIA CONDITIONAL VARIANCE TESTING

We will now formulate and analyze a GMS method for a non-stationary process $\mathbf{x}[n]$ conforming to the model (1). To this end, we will first show how the CIG of $\mathbf{x}[n]$ can be characterized in terms of conditional variance tests. The GMS method implements these conditional variance tests using the covariance matrix estimate $\widehat{\mathbf{C}}[b]$ (cf. (11)).

3.1. Conditional Variance Testing
The characterization (8) for the CIG $\mathcal{G}$ of the process $\mathbf{x}[n]$ (1) seems convenient: We just have to determine the non-zero pattern of the precision matrices $\mathbf{K}[n]$ and immediately can estimate the edge set of the CIG $\mathcal{G}$. However, the problem is in estimating the precision matrix $\mathbf{K}[n]$, in the high-dimensional regime where $N \ll d$. In particular, in the high-dimensional regime, any reasonable estimator $\widehat{\mathbf{C}}$ for the covariance matrix $\mathbf{C}[n]$ is singular, preventing the use of the inverse $\widehat{\mathbf{C}}^{-1}$ as an estimate for $\mathbf{K}[n]$.

In order to cope with the high-dimensional regime, we will now present an approach to GMS via determining the neighborhoods $\mathcal{N}(r)$ for all nodes r which exploits the sparsity of the CIG (cf. Assumption 3). Our strategy for determining the neighborhoods $\mathcal{N}(r)$ will be based on evaluating the conditional variance

$$V_x^{(r,\mathcal{T})} := (1/N)\mathrm{Tr}\{\mathbf{V}_x^{(r,\mathcal{T})}\}, \tag{12}$$

with the conditional covariance matrix

$$\mathbf{V}_x^{(r,\mathcal{T})} := \mathrm{cov}\{\mathbf{x}_r[\cdot] \big| \{\mathbf{x}_t[\cdot]\}_{t \in \mathcal{T}}\}.$$

Here, $\mathcal{T} \subseteq \mathcal{V} \setminus \{r\}$ is a subset of at most $s_{\max}$ nodes, i.e., $|\mathcal{T}| \leq s_{\max}$. We can express the conditional covariance matrix $\mathbf{V}_x^{(r,\mathcal{T})}$ in terms of the covariance matrix $\mathbf{C}_{\tilde{\mathbf{x}}}$ (cf. (5)) as [24, Thm. 23.7.4.]

$$\mathbf{V}_x^{(r,\mathcal{T})} = \mathbf{C}_{\tilde{x}}[r,r] - \mathbf{C}_{\tilde{x}}[r,\mathcal{T}]\big(\mathbf{C}_{\tilde{x}}[\mathcal{T},\mathcal{T}]\big)^{-1}\mathbf{C}_{\tilde{x}}[\mathcal{T},r]. \tag{13}$$

Note that the conditional covariance matrix $\mathbf{V}_x^{(r,\mathcal{T})}$ depends only on a (small) submatrix of $\mathbf{C}_{\tilde{\mathbf{x}}}$ constituted by the $N \times N$ blocks $\mathbf{C}_{\tilde{x}}[i,j]$ for $i, j \in \mathcal{T} \cup \{r\}$.

Using the block diagonal structure of $\mathbf{C}_x$ (cf. (4)), we can simplify (13) to obtain the following representation for the conditional variance:

---

[1]For ease of notation and without essential loss of generality, we assume the sample size N to be a multiple of the blocklength L.

Lemma 3.1. *The conditional variance* $V_x^{(r,\mathcal{T})}$ *(12) satisfies*

$$V_x^{(r,\mathcal{T})} = (1/N) \sum_{n=1}^{N} \frac{1}{((\mathbf{C}[n])_{\{\mathcal{T}',\mathcal{T}'\}})^{-1}{}_{\{r,r\}}}, \tag{14}$$

with $\mathcal{T}' := \{r\} \cup \mathcal{T}$.

Proof. Consider the subset $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$, let $\mathbf{x}_{\mathcal{T}}[n] = (x_{t_1}[n], \ldots, x_{t_k}[n])^T$ and $\mathbf{P}_{\mathcal{T}}$ be the permutation matrix transforming $\mathbf{x}_{\mathcal{T}} := ((\mathbf{x}_{\mathcal{T}}[1])^T, \ldots, (\mathbf{x}_{\mathcal{T}}[N])^T)^T$ into $\tilde{\mathbf{x}}_{\mathcal{T}} := (\mathbf{x}_{t_1}[\cdot]^T, \ldots, \mathbf{x}_{t_k}[\cdot]^T)^T$, i.e., $\tilde{\mathbf{x}}_{\mathcal{T}} = \mathbf{P}_{\mathcal{T}}\mathbf{x}_{\mathcal{T}}$. The covariance matrix for $\tilde{\mathbf{x}}_{\mathcal{T}}$ is obtained as $\mathbf{C}_{\tilde{x}}[\mathcal{T},\mathcal{T}] = \mathbf{P}_{\mathcal{T}}\mathbf{C}_{\mathcal{T}}\mathbf{P}_{\mathcal{T}}^T$, and, in turn since $\mathbf{P}_{\mathcal{T}}^{-1} = \mathbf{P}_{\mathcal{T}}^T$, $\big(\mathbf{C}_{\tilde{x}}[\mathcal{T},\mathcal{T}]\big)^{-1} = \mathbf{P}_{\mathcal{T}}(\mathbf{C}_{\mathcal{T}})^{-1}\mathbf{P}_{\mathcal{T}}^T$.

The conditional variance $V_x^{(r,\mathcal{T})}$ is then given as

$$(1/N)\,\mathrm{Tr}\big\{\mathbf{C}_{\tilde{x}}[r,r] - \mathbf{C}_{\tilde{x}}[r,\mathcal{T}]\big(\mathbf{C}_{\tilde{x}}[\mathcal{T},\mathcal{T}]\big)^{-1}\mathbf{C}_{\tilde{x}}[\mathcal{T},r]\big\} \tag{15}$$
$$= (1/N)\,\mathrm{Tr}\big\{\mathbf{C}_{\tilde{x}}[r,r] - \mathbf{C}_{\tilde{x}}[r,\mathcal{T}]\mathbf{P}_{\mathcal{T}}(\mathbf{C}_{\mathcal{T}})^{-1}\mathbf{P}_{\mathcal{T}}^T\mathbf{C}_{\tilde{x}}[\mathcal{T},r]\big\}.$$

Due to the block-diagonal structure of $\mathbf{C}_{\tilde{x}}$ (cf. (5)),

$$\mathbf{C}_{\tilde{x}}[r,\mathcal{T}]\mathbf{P}_{\mathcal{T}} = \begin{pmatrix} (\mathbf{C}[1])_{\{r,\mathcal{T}\}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & (\mathbf{C}[N])_{\{r,\mathcal{T}\}} \end{pmatrix}. \tag{16}$$

Inserting (16) into (15), yields further

$$V_x^{(r,\mathcal{T})} = (1/N) \sum_{n=1}^{N} \Big((\mathbf{C}[n])_{\{r,r\}} - (\mathbf{C}[n])_{\{r,\mathcal{T}\}}\big((\mathbf{C}[n])_{\{\mathcal{T},\mathcal{T}\}}\big)^{-1}(\mathbf{C}[n])_{\{\mathcal{T},r\}}\Big). \tag{17}$$

The expression (14) for the conditional variance follows then from (17) using the matrix inversion lemma for block matrices [25, Ex. 2.2.4.]. $\square$

Using the conditional variance $V_x^{(r,\mathcal{T})}$, we can characterize the neighborhoods $\mathcal{N}(r)$ in the CIG $\mathcal{G}$ as:

Theorem 3.2. *For any set* $\mathcal{T} \in \mathfrak{S}_{s_{\max}}^r$:

- *If* $\mathcal{N}(r) \setminus \mathcal{T} \neq \emptyset$, *we have*
$$V_x^{(r,\mathcal{T})} \geq \rho_{\min} + (1/N)\mathrm{Tr}\big\{\big(\mathbf{K}_{\tilde{x}}[r,r]\big)^{-1}\big\}. \tag{18}$$

- *For* $\mathcal{N}(r) \subseteq \mathcal{T}$, *we obtain*
$$V_x^{(r,\mathcal{T})} = (1/N)\mathrm{Tr}\big\{\big(\mathbf{K}_{\tilde{x}}[r,r]\big)^{-1}\big\}. \tag{19}$$

Proof. see Appendix. $\square$

As an immediate consequence of Theorem 3.2, we can determine the neighborhood $\mathcal{N}(r)$ by a simple conditional variance test procedure:

$$\mathcal{N}(r) = \underset{\mathcal{T} \in \mathfrak{S}_{s_{\max}}^r}{\arg\min}\, V_x^{(r,\mathcal{T})} + \rho_{\min}|\mathcal{T}|. \tag{20}$$

3.2. The GMS method

We now turn the procedure (20) into a practical GMS method by replacing $V_x^{(r,\mathcal{T})}$ in (20) with the estimate

$$\widehat{V_x^{(r,\mathcal{T})}} = (1/B) \sum_{b=1}^{B} \frac{1}{((\widehat{\mathbf{C}}[b])_{\{\mathcal{T}',\mathcal{T}'\}})^{-1}{}_{\{r,r\}}} \tag{21}$$

using the sample covariance matrix $\widehat{\mathbf{C}}[b]$ (cf. (11)) and $\mathcal{T}' := \{r\} \cup \mathcal{T}$.

Algorithm 1 GMS for uncorrelated nonstationary processes

Input: $\mathbf{x}[1], \ldots, \mathbf{x}[N]$, $\rho_{\min}$, $s_{\max}$, blocklength L

1: **for** each node r in $\mathcal{V}$ **do**
2: $\quad \widehat{\mathcal{N}}(r) := \arg\min_{\mathcal{T} \in \mathfrak{S}^r_{s_{\max}}} \widehat{V_x^{(r,\mathcal{T})}} + |\mathcal{T}|\rho_{\min}$ (cf. (21))
3: **end for**
4: combine estimates $\widehat{\mathcal{N}}(r)$ by "OR-" or "AND rule"

  • OR: $(i,j) \in \widehat{\mathcal{E}}$ if either $(i,j) \in \widehat{\mathcal{N}}(i)$ or $(i,j) \in \widehat{\mathcal{N}}(j)$
  • AND: $(i,j) \in \widehat{\mathcal{E}}$ if $(i,j) \in \widehat{\mathcal{N}}(i)$ and $(i,j) \in \widehat{\mathcal{N}}(j)$

Output: CIG estimate $\widehat{\mathcal{G}} = (\mathcal{V}, \widehat{\mathcal{E}})$

For a sufficiently large sample size N, the CIG estimate $\widehat{\mathcal{G}}$ delivered by Alg. 1 coincides with the true CIG $\mathcal{G}$ with high probability.

Theorem 3.3. *There are constants c_1, c_2 depending only on $\{\alpha[n], \beta[n]\}_{n \in \{1,\ldots,N\}}$ such that for a sample size*

$$N \geq c_1 \frac{s_{\max}^{5/2}}{\rho_{\min}^3}(\log \frac{\kappa s_{\max}^{7/2}}{\delta \rho_{\min}^3} + s_{\max}\log d) \qquad (22)$$

Alg. 1 used with blocklength

$$L = c_2 \frac{s_{\max}^2}{\rho_{\min}^2}(\log \frac{\kappa s_{\max}^{7/2}}{\delta \rho_{\min}^3} + s_{\max}\log d),$$

delivers the correct CIG with prob. at least $1-\delta$, i.e., $\mathrm{P}\{\widehat{\mathcal{G}} = \mathcal{G}\} \geq 1-\delta$

A detailed proof of Theorem 3.3 is omitted for space restrictions and will be provided in a follow up journal publication. However, the high-level idea is straightforward: If the maximum deviation

$$E = \max_{r \in \mathcal{V}, \mathcal{T} \in \mathfrak{S}^r_{s_{\max}}} |\widehat{V_x^{(r,\mathcal{T})}} - V_x^{(r,\mathcal{T})}|$$

is less than $\rho_{\min}/2$, Alg. 1 is guaranteed to select the correct neighborhoods, i.e., $\widehat{\mathcal{N}}(r) = \mathcal{N}(r)$ for all nodes $r \in \mathcal{V}$, implying the selection of the correct CIG, i.e., $\mathcal{G} = \widehat{\mathcal{G}}$. For controlling the probability of the event $E \geq \rho_{\min}/2$, we apply a large deviation characterization for Gaussian quadratic forms [20, Lemma F.1].

The lower bound (22) on sample size N stated by Theorem 3.3, depends only logarithmically on the process dimension d and polynomially on the maximum node degree $s_{\max}$. Thus, for processes having a sufficiently sparse CIG (small $s_{\max}$), the GMS method in Alg. 1 delivers the correct CIG even in scenarios where the process dimension is exponentially larger than the available sample size. Moreover, the bound (22) depends inversely on the minimum partial correlation $\rho_{\min}$, which is reasonable as a smaller partial correlation is more difficult to detect. Note that the quantity $\rho_{\min}$ occurring in (22) represents the average (over n) of the marginal conditional correlations between the process components.

Appendix: Proof of Theorem 3.2

We detail the proof only for the neighborhood $\mathcal{N}(1)$ of the particular node 1. The generalization to an arbitrary node is then straightforward.

Let us introduce the weight matrices $\mathbf{L}_{1,r} := (\mathbf{K}_{\tilde{x}}[1,1])^{-1}\mathbf{K}_{\tilde{x}}[1,r]$. According to (7) we have $\mathbf{L}_{1,r} = 0$ for $r \notin \mathcal{N}(1)$. Using elementary properties of multivariate normal distributions (cf. [22, Prop. 1.6.6.]), we have the decomposition

$$\mathbf{x}_1[\cdot] = \sum_{r \in \mathcal{N}(1)} \mathbf{L}_{1,r}\mathbf{x}_r[\cdot] + \boldsymbol{\varepsilon}_1 \qquad (23)$$

with the zero-mean "error term" $\boldsymbol{\varepsilon}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_x^{(1,\mathcal{N}(1))})$ whose covariance matrix is $\mathbf{V}_x^{(1,\mathcal{N}(1))} = (\mathbf{K}_{\tilde{x}}[1,1])^{-1}$. The identity (19) is then obtained as

$$\mathbf{V}_x^{(1,\mathcal{T})} \stackrel{(23),\mathcal{N}(1) \subseteq \mathcal{T}}{=} \mathbf{V}_x^{(1,\mathcal{N}(1))} = (\mathbf{K}_{\tilde{x}}[1,1])^{-1}.$$

Moreover, by the projection property of conditional expectations (cf. [22, Sec. 2.7]), the error term $\boldsymbol{\varepsilon}_1$ in (23) is uncorrelated (and hence independent) with (of) the process components $\{\mathbf{x}_r[\cdot]\}_{r \in \{2,\ldots,d\}}$, i.e.,

$$\mathrm{E}\{\mathbf{x}_r[\cdot]\boldsymbol{\varepsilon}_1^T\} = \mathbf{0} \text{ for all } r \in \{2,\ldots,d\}. \qquad (24)$$

Let us now focus on the conditional variance $V_x^{(1,\mathcal{T})}$ for a subset $\mathcal{T} \in \mathfrak{S}^1_{s_{\max}}$ with $\mathcal{N}(1) \setminus \mathcal{T} \neq \emptyset$, i.e., there is an index $j \in \mathcal{N}(1) \setminus \mathcal{T}$. We use the shorthands $\mathcal{P} := \mathcal{T} \cup \mathcal{N}(1)$ and $\mathcal{Q} := \mathcal{P} \setminus \{j\}$. Note that $\mathcal{T} \subseteq \mathcal{Q}$. For the conditional mean $\widehat{\mathbf{x}_j[\cdot]} := \mathrm{E}\{\mathbf{x}_1[\cdot] | \{\mathbf{x}_r[\cdot]\}_{r \in \mathcal{Q}}\}$, we have the decomposition

$$\mathbf{x}_j[\cdot] = \widehat{\mathbf{x}_j[\cdot]} + \boldsymbol{\varepsilon}_j. \qquad (25)$$

with the zero-mean "error term" $\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{e,j})$ being uncorrelated with the components $\{\mathbf{x}_r[\cdot]\}_{r \in \mathcal{Q}}$, i.e.,

$$\mathrm{E}\{\mathbf{x}_r[\cdot]\boldsymbol{\varepsilon}_j^T\} = \mathbf{0} \text{ for all } r \in \mathcal{Q}. \qquad (26)$$

Moreover, the inverse covariance of $\boldsymbol{\varepsilon}_j$ satisfies

$$\mathbf{C}_{e,j}^{-1} = \mathbf{K}[j,j],$$

with $\mathbf{K} = (\mathbf{C}_{\tilde{x}}[\mathcal{T}', \mathcal{T}'])^{-1}$, where $\mathcal{T}' = \mathcal{T} \cup \{j\}$. Since the blocks $\mathbf{C}_{\tilde{x}}[a,b]$ of the matrix $\mathbf{C}_{\tilde{x}}$ (cf. (6)), the matrix $\mathbf{K}[j,j]$ is diagonal with main-diagonal given by the values $\frac{1}{((\mathbf{C}[n])_{\{\mathcal{T}', \mathcal{T}'\}})^{-1}_{\{1,1\}}}$ which, together with Assumption 2, yields

$$\mathbf{C}_{e,j} \succeq \mathrm{diag}\{\alpha[n]\}_{n=1,\ldots,N}. \qquad (27)$$

Inserting (25) into (23) yields

$$\mathbf{x}_1[\cdot] = \sum_{r \in \mathcal{N}(1)\setminus\{j\}} \mathbf{L}_{1,r}\mathbf{x}_r[\cdot] + \mathbf{L}_{1,j}\widehat{\mathbf{x}_j[\cdot]} + \mathbf{L}_{1,j}\boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_1$$

$$= \sum_{r \in \mathcal{Q}} \mathbf{M}_r\mathbf{x}_r[\cdot] + \mathbf{L}_{1,j}\boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_1. \qquad (28)$$

Due to (24) and (26), the terms $\mathbf{L}_{1,j}\boldsymbol{\varepsilon}_j$ and $\boldsymbol{\varepsilon}_1$ are both uncorrelated (and therefore independent due to Gaussianity) to all the components $\{\mathbf{x}_r[\cdot]\}_{r \in \mathcal{Q}}$ and moreover are also mutually uncorrelated, i.e., $\mathrm{E}\{\boldsymbol{\varepsilon}_r(\boldsymbol{\varepsilon}_1^T, \boldsymbol{\varepsilon}_j^T)\} = \mathbf{0}$ for all $r \in \mathcal{Q}$ and $\mathrm{E}\{\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_1^T\} = \mathbf{0}$. According to the law of total variance [26] and since $\mathcal{T} \subseteq \mathcal{Q}$, we have $V_x^{(1,\mathcal{T})} \geq V_x^{(1,\mathcal{Q})}$. Therefore, we obtain the lower bound:

$$V_x^{(1,\mathcal{T})} \geq V_x^{(1,\mathcal{Q})} \stackrel{(12)}{=} (1/N)\mathrm{Tr}\{\mathrm{cov}\{\mathbf{x}_1[\cdot] | \{\mathbf{x}_r[\cdot]\}_{r \in \mathcal{Q}}\}$$

$$\stackrel{(28)}{=} (1/N)\mathrm{Tr}\{\mathbf{L}_{1,j}\mathbf{C}_{e,j}\mathbf{L}_{1,j}^T + \mathbf{V}_x^{(1,\mathcal{N}(1))}\} \qquad (29)$$

$$\stackrel{(27),(6)}{\geq} (1/N)\sum_{n=1}^{N} \alpha[n]\left[(\mathbf{K}[n])_{a,b}/(\mathbf{K}[n])_{a,a}\right]^2 + V_x^{(1,\mathcal{N}(1))}$$

valid for any $\mathcal{T} \in \mathfrak{S}^1_{s_{\max}}$ with $\mathcal{T} \neq \mathcal{N}(1)$. We obtain (18) by combining (29) with Asspt. 2.

4. REFERENCES

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

[2] R. Dahlhaus and M. Eichler, "Causality and graphical models for time series," in *Highly Structured Stochastic Systems*, P. Green, N. Hjort, and S. Richardson, Eds. Oxford, UK: Oxford Univ. Press, 2003, pp. 115–137.

[3] F. R. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2189–2199, Aug. 2004.

[4] M. Eichler, "Graphical models in time series analysis," Ph.D. dissertation, Universität Heidelberg, 1999.

[5] N. E. Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2717–2756, 2008.

[6] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, Jul. 2012.

[7] 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.

[8] A. Bolstad, B. D. van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2628–2641, Jun. 2011.

[9] 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.

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

[11] J. H. Friedmann, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.

[12] I. Ebert-Uphoff and Y. Deng, "A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer," *Geophysical Research Letters*, vol. 39, no. 19. [Online]. Available: http://dx.doi.org/10.1029/2012GL053269

[13] E. Davidson and M. Levin, "Gene regulatory networks," *Proc. Natl. Acad. Sci.*, vol. 102, no. 14, Apr. 2005.

[14] M. Kolar and E. P. Xing, "On time varying undirected graphs," *In Proc. the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Ft. Lauderdale, FL, pp. 407-415, 2011.

[15] S. Zhou, J. Lafferty, and L. Wasserman, "Time varying undirected graphs," *In 21st Annual Conference on Learning Theory (COLT)*, Helsinki, Finland, July 2008.

[16] E. Hwang and D. W. Shin, "A note on exponential inequalities of ψ-weakly dependent sequences," *Comm. for Stat. Applications and Methods*, 2014.

[17] 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.

[18] 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.

[19] A. Jung, G. Hannak, and N. Görtz, "Graphical LASSO Based Model Selection for Time Series," *to appear in IEEE Sig. Proc. Letters*, 2015.

[20] 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.

[21] D. L. Donoho, S. Mallat, and R. von Sachs, "Estimating covariances of locally stationary processes: Consistency of best basis methods," in *Proc. IEEE-SP Int. Sympos. Time-Frequency Time-Scale Analysis*, Paris, France, June 1996, pp. 337–340.

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

[23] G. Matz and F. Hlawatsch, "Nonstationary spectral analysis based on time-frequency operator symbols and underspread approximations," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1067–1086, March 2006.

[24] A. Lapidoth, *A Foundation in Digital Communication*. New York: Cambridge University Press, 2009.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[26] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.