Graphical LASSO based Model Selection for Time Series

Alexander Jung, Gabor Hannak, and Norbert Goertz

***Abstract*—We propose a novel graphical model selection scheme for high-dimensional stationary time series or discrete time processes. The method is based on a natural generalization of the graphical LASSO algorithm, introduced originally for the case of i.i.d. samples, and estimates the conditional independence graph of a time series from a finite length observation. The graphical LASSO for time series is defined as the solution of an ℓ_1-regularized maximum (approximate) likelihood problem. We solve this optimization problem using the alternating direction method of multipliers. Our approach is nonparametric as we do not assume a finite dimensional parametric model, but only require the process to be sufficiently smooth in the spectral domain. For Gaussian processes, we characterize the performance of our method theoretically by deriving an upper bound on the probability that our algorithm fails. Numerical experiments demonstrate the ability of our method to recover the correct conditional independence graph from a limited amount of samples.**

***Index Terms*—ADMM, graphical LASSO, graphical model selection, nonparametric time series, sparsity.**

I. Introduction

We consider the problem of inferring the conditional independence graph (CIG) of a stationary high-dimensional discrete time process or time series $\mathbf{x}[n]$ from observing N samples $\mathbf{x}[1],\ldots,\mathbf{x}[N]$. This problem is referred to as graphical model selection (GMS) and of great practical interest, e.g., for gene analysis, econometrics, environmental monitoring and medical diagnosis [1]–[5].

Most of existing work on GMS for time series is based on finite dimensional parametric models [6], [7]. By contrast, a first nonparametric GMS method for high-dimensional time series has been proposed recently [8]. This approach is based on performing neighborhood regression (cf. [9]) in the frequency domain.

In this paper, we present an alternative nonparametric GMS scheme for time series based on generalizing the graphical LASSO (gLASSO) [10]–[12] to stationary time series. The resulting algorithm is implemented using the alternating direction method of multipliers (ADMM) [13], for which we derive closed-form update rules. Typically, ADMM-based methods allow for straightforward distributed implementations, e.g., in wireless sensor networks [13], [14].

While algorithmically our approach is similar to the joint gLASSO proposed in [15], the deployment of a gLASSO type algorithm for GMS of time series is new.

Our main analytical contribution is a performance analysis which yields an upper bound on the probability that our scheme fails to correctly identify the CIG. The effectiveness of our GMS method is also verified numerically.

Notation. Given a natural number F, we define $[F] := \{1,\ldots,F\}$. For a square matrix $\mathbf{X} \in \mathbb{C}^{p\times p}$, we denote by $\bar{\mathbf{X}}$, $\mathbf{X}^H$, $\mathrm{tr}\{\mathbf{X}\}$ and $|\mathbf{X}|$ its elementwise complex conjugate, its Hermitian transpose, its trace and its determinant, respectively. We also need the matrix norm $\|\mathbf{X}\|_\infty := \max_{i,j}|X_{i,j}|$. By writing $\mathbf{X} \preceq \mathbf{Y}$ we mean that $\mathbf{Y}-\mathbf{X}$ is a positive-semidefinite (psd) matrix. A strictly positive definite matrix $\mathbf{X}$ is indicated as $\mathbf{0} \prec \mathbf{X}$.

We denote by $\mathcal{H}_p^{[F]}$ the set of all length-F sequences $\mathbf{X}[\cdot] := (\mathbf{X}[1],\ldots,\mathbf{X}[F])$ with Hermitian matrices $\mathbf{X}[f] \in \mathbb{C}^{p\times p}$. For a sequence $\mathbf{X}[\cdot] \in \mathcal{H}_p^{[F]}$, we define $\|X_{i,j}[\cdot]\|^2 := (1/F)\sum_{f\in[F]}|X_{i,j}[f]|^2$, its squared generalized Frobenius norm $\|\mathbf{X}[\cdot]\|_{\mathrm{F}}^2 := \sum_{i,j}\|X_{i,j}[\cdot]\|^2$ and its ℓ_1-norm as $\|\mathbf{X}\|_1 := \sum_{i,j}\|X_{i,j}[\cdot]\|$. We equip the set $\mathcal{H}_p^{[F]}$ with the inner product $\langle \mathbf{A}[\cdot],\mathbf{B}[\cdot]\rangle := (1/F)\sum_{f\in[F]}\mathrm{tr}\{\mathbf{A}[f]\mathbf{B}[f]\}$.

For a sequence $\mathbf{X}[\cdot] \in \mathcal{H}_p^{[F]}$ and some subset $\mathcal{S} \subseteq [p]\times[p]$, we denote by $\mathbf{X}_{\mathcal{S}}[\cdot]$ the matrix sequence which is obtained by zeroing separately for each index $f \in [F]$ all entries of the matrix $\mathbf{X}[f]$ except those in $\mathcal{S}$. The generalized support of a sequence $\mathbf{X}[\cdot] \in \mathcal{H}_p^{[F]}$ is defined as $\mathrm{gsupp}(\mathbf{X}[\cdot]) := \{(i,j) \in [p]\times[p] : (\mathbf{X}[f])_{i,j} \neq 0 \text{ for some } f \in [F]\}$. We also use $a_+ := \max\{a,0\}$.

II. Problem Formulation

Consider a p-dimensional real-valued zero-mean stationary time series $\mathbf{x}[n] = (x_1[n],\ldots,x_p[n])^T$, for $n \in \mathbb{Z}$. We assume its autocorrelation function (ACF) $\mathbf{R}[m] := \mathrm{E}\{\mathbf{x}[m]\mathbf{x}^T[0]\}$ to be absolutely summable, i.e., $\sum_{m=-\infty}^{\infty}\|\mathbf{R}[m]\| < \infty$, such that we can define the spectral density matrix (SDM) $\mathbf{S}(\theta)$ via a Fourier transform:

$$\mathbf{S}(\theta) := \sum_{m=-\infty}^{\infty}\mathbf{R}[m]\exp(-j2\pi m\theta), \quad \theta \in [0,1). \tag{1}$$

We require the eigenvalues of the SDM to be uniformly bounded as

$$L \leq \lambda_i\left(\mathbf{S}(\theta)\right) \leq U, \quad \text{for all } \theta \in [0,1), \tag{2}$$

where, without loss of generality, we will assume $L = 1$ in what follows. The upper bound in (2) is valid if the ACF $\mathbf{R}[m]$ is

Manuscript received October 22, 2014; revised March 13, 2015; accepted April 17, 2015. Date of publication April 22, 2015; date of current version June 01, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Glenn Easley.

The authors are with the Institute of Telecommunications, Vienna University of Technology, 1040 Vienna, Austria (e-mail: ajung@nt.tuwien.ac.at; ghannak@nt.tuwien.ac.at; norbert.goertz@nt.tuwien.ac.at).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2425434

summable; the lower bound ensures certain Markov properties of the CIG [3], [16], [17].

Our approach is based on the assumption that the SDM is a smooth function. Due to the Fourier relationship (1), this smoothness will be quantified via certain ACF moments

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

Here, $h[m]$ denotes a weight function which typically increases with $|m|$. For a process with sufficiently small moment $\mu_x^{(h)}$, thereby enforcing smoothness of the SDM, we are allowed to base our considerations on a discretized version of the SDM, given by $\mathbf{S}[f] = \mathbf{S}(\theta_f)$, with $\theta_f := (f-1)/F$, for $f \in [F]$. The number F of sampling points is a design parameter which has to be chosen suitably large, compared to the ACF moment $\mu_x^{(h)}$ (cf. [8, Lemma 2.1]).

The CIG of a process $\mathbf{x}[n]$ is a simple undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = [p]$. Each node $r \in \mathcal{V}$ represents a single scalar component process $x_r[n]$. An edge between nodes r and r' is absent, i.e., $(r, r') \notin \mathcal{E}$, if the component processes $x_r[n]$ and $x_{r'}[n]$ are conditionally independent given all remaining component processes [3].

If the process $\mathbf{x}[n]$ is Gaussian, the CIG can be conveniently characterized via the process inverse SDM $\mathbf{K}[f] := \mathbf{S}^{-1}[f]$. More specifically, it can be shown that, for sufficiently small $\mu_x^{(h)}$ with $h[m] = |m|$, [3], [8],

$$(i,j) \in \mathcal{E} \Leftrightarrow (i,j) \in \mathrm{gsupp}(\mathbf{K}[\cdot]). \tag{4}$$

Thus, the edge set $\mathcal{E}$ of the CIG is determined by the generalized support of the inverse SDM $\mathbf{K}[f]$, for $f \in [F]$.

Our goal is to robustly estimate the CIG from a finite length observation, incurring unavoidable estimation errors. Therefore, we have to require that, in addition to (4), the non-zero off-diagonal entries of $\mathbf{K}[f]$ are sufficiently large, such that a certain amount of estimation error is tolerable. To this end, we define the process (un-normalized) minimum global partial spectral coherence as

$$\rho_x := \min_{(i,j)\in\mathcal{E}} \|K_{i,j}[\cdot]\|.$$

For the analysis of our GMS scheme we require

Assumption II.1: We have $\rho_x \geq \rho_{\min}$ for a known $\rho_{\min} > 0$.

Our approach to GMS in the high-dimensional regime exploits a specific problem structure induced by the assumption that the true CIG is sparse.

Assumption II.2: The CIG of the observed process $\mathbf{x}[n]$ is sparse such that $|\mathcal{E}| \leq s$ for some small $s \ll p(p-1)/2$.

The performance analysis of the proposed GMS algorithm requires to quantify the conditioning of SDM sub-matrices. In particular, we will use the following assumption, which is a natural extension of the (multitask) compatibility condition, originally introduced in [11] to analyze LASSO for the ordinary sparse linear (multitask) model.

Assumption II.3: Given a process $\mathbf{x}[n]$ whose CIG contains no more than s edges indexed by $\mathcal{S} \subseteq [p] \times [p]$, we assume that there exists a positive constant ϕ such that

$$\frac{1}{F}\sum_{f\in[F]} \mathrm{vec}\{\mathbf{X}[f]\}^H (\bar{\mathbf{S}}[f] \otimes \mathbf{S}[f]) \mathrm{vec}\{\mathbf{X}[f]\} \geq \frac{\phi}{s}\|\mathbf{X}_{\mathcal{S}}\|_1^2 \tag{5}$$

holds for all $\mathbf{X}[\cdot] \in \mathcal{H}_p^{[F]}$ with $\|\mathbf{X}_{\mathcal{S}^c}\|_1 \leq 3\|\mathbf{X}_{\mathcal{S}}\|_1$.

The constant ϕ in (5) is essentially a lower bound on the eigenvalues of small sub-matrices of the SDM. As such, the Assumption II.3 is closely related to the concept of the restricted isometry property (RIP) [18].

As verified easily, Assumption II.3 is always valid with $\phi = L$ for a process satisfying (2). However, for processes having a sparse CIG, we typically expect $\phi \gg L$.

III. Graphical LASSO for Time Series

The *graphical least absolute shrinkage and selection operator* (gLASSO) [10]–[12], [19] is an algorithm for estimating the inverse covariance matrix $\mathbf{K} := \mathbf{C}^{-1}$ of a high-dimensional Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ based on i.i.d. samples. In particular, gLASSO is based on optimizing a ℓ_1-penalized log-likelihood function and can therefore be interpreted as regularized maximum likelihood estimation.

A. Extending gLASSO to Stationary Time Series

A natural extension of gLASSO to the case of stationary Gaussian time series is obtained by replacing the objective function for the i.i.d. case (which can be interpreted as a penalized log-likelihood) with the corresponding penalized log-likelihood function for a stationary Gaussian process. However, since the exact likelihood lacks a simple closed-form expression (in terms of the SDM), we will use the "Whittle-approximation" [20], [21], to arrive at the following gLASSO estimator for general stationary time series:

$$\widehat{\mathbf{K}}[\cdot] := \operatorname*{arg\,min}_{\mathbf{X}[\cdot]\in\mathcal{C}} -A\{\mathbf{X}\} + \langle \widehat{\mathbf{S}}[\cdot], \mathbf{X}[\cdot]\rangle + \lambda\|\mathbf{X}[\cdot]\|_1 \tag{6}$$

with $A\{\mathbf{X}\} := (1/F)\sum_{f\in[F]} \log|\mathbf{X}[f]|$ and constraint set

$$\mathcal{C} := \{\mathbf{X}[\cdot] \in \mathcal{H}_p^{[F]} : \mathbf{0} \prec \mathbf{X}[f] \preceq \mathbf{I} \text{ for all } f \in [F]\}. \tag{7}$$

This constraint set is reasonable since (i) the function $A\{\mathbf{X}\}$ is finite only if $\mathbf{0} \prec \mathbf{X}[f]$ and (ii) the true inverse SDM satisfies $\mathbf{K}[f] \preceq \mathbf{I}$, for all $f \in [F]$, due to (2) (with $L = 1$).

The formulation (6) involves an estimator $\widehat{\mathbf{S}}[f]$ of the SDM values $\mathbf{S}(\theta_f)$, for $f \in [F]$. While in principle any reasonable estimator could be used, we will restrict the choice to a multivariate Blackman-Tukey (BT) estimator [22]

$$\widehat{\mathbf{S}}[f] = \sum_{m=-N+1}^{N-1} w[m]\widehat{\mathbf{R}}[m]\exp(-j2\pi m\theta_f) \tag{8}$$

with the standard biased autocorrelation estimate $\widehat{\mathbf{R}}[m] = (1/N)\sum_{n=m+1}^{N} \mathbf{x}[n]\mathbf{x}^T[n-m]$ for $m = 0, \ldots, N-1$. Enforcing the symmetry $\widehat{\mathbf{R}}[-m] = \widehat{\mathbf{R}}^H[m]$, we can obtain the ACF estimate for $m = -N+1, \ldots, -1$. The window function $w[m]$ in (8) is a design parameter, which can be chosen freely as long as it yields a psd estimate $\widehat{\mathbf{S}}[f]$. A sufficient condition such that $\widehat{\mathbf{S}}[f]$ is guaranteed to be psd is non-negativeness of the Fourier transform $W(\theta) := \sum_{m=-\infty}^{\infty} w[m]\exp(-j2\pi\theta m)$ [22, Sec. 2.5.2].

The existence of a minimizer in (6) is guaranteed for any choice of $\lambda \geq 0$ as the optimization problem (6) is equivalent to the unconstrained problem $\min_{\mathbf{X}[\cdot]} -A\{\mathbf{X}\} + \langle \widehat{\mathbf{S}}[\cdot], \mathbf{X}[\cdot]\rangle + \lambda\|\mathbf{X}[\cdot]\|_1 + I_{\mathcal{C}}(\mathbf{X}[\cdot])$, where $I_{\mathcal{C}}(\mathbf{X}[\cdot])$ is the indicator function of the constraint set $\mathcal{C}$. Existence of a minimizer of this equivalent problem is guaranteed by [23, Theorem 27.2]: The objective function is a closed proper convex function and is finite only on

the bounded set $\mathcal{C}$ (cf. (7)), which trivially implies that the objective function has no direction of recession [23].

We will present in Section III-C a specific choice for the tuning parameter λ in (6), which ensures that the gLASSO estimator $\widehat{\mathbf{K}}[\cdot]$ is accurate, i.e., the estimation error $\mathbf{\Delta}[\cdot] := \widehat{\mathbf{K}}[\cdot] - \mathbf{K}[\cdot]$ is small. Based on the gLASSO (6), an estimate for the edge set of the CIG may then be obtained by thresholding:

$$\widehat{\mathcal{E}}(\eta) := \{(i,j) : \|\widehat{K}_{i,j}[\cdot]\| \geq \eta\}. \tag{9}$$

Obviously, under Assumption II.1, if

$$\|\mathbf{\Delta}\|_1 \leq \rho_{\min}/2, \tag{10}$$

we have $\widehat{\mathcal{E}}(\rho_{\min}/2) = \mathcal{E}$, i.e., the CIG is recovered perfectly.

B. ADMM Implementation

An efficient numerical method for solving convex optimization problems of the type (6) is the *alternating direction method of multipliers* (ADMM). Defining the augmented Lagrangian [13] of the problem (6) as

$$L_\rho(\mathbf{X}[\cdot], \mathbf{Z}[\cdot], \mathbf{U}[\cdot]) := -A\{\mathbf{X}\} + \langle \widehat{\mathbf{S}}[\cdot], \mathbf{X}[\cdot]\rangle + \lambda\|\mathbf{Z}\|_1 + (\rho/2)\|\mathbf{X}[\cdot] - \mathbf{Z}[\cdot] + \mathbf{U}[\cdot]\|_{\mathrm{F}}^2,$$

the (scaled) ADMM method iterates, starting with arbitrary initializations for $\mathbf{X}^{(0)}[\cdot]$, $\mathbf{Z}^{(0)}[\cdot]$ and $\mathbf{U}^{(0)}[\cdot]$, the following update rules

$$\mathbf{X}^{(k+1)}[\cdot] = \operatorname*{arg\,min}_{\mathbf{X}[\cdot]\in\mathcal{C}} L_\rho(\mathbf{X}[\cdot], \mathbf{Z}^{(k)}[\cdot], \mathbf{U}^{(k)}[\cdot]) \tag{11}$$

$$\mathbf{Z}^{(k+1)}[\cdot] = \operatorname*{arg\,min}_{\mathbf{Z}[\cdot]\in\mathcal{H}_p^{[F]}} L_\rho(\mathbf{X}^{(k+1)}[\cdot], \mathbf{Z}[\cdot], \mathbf{U}^{(k)}[\cdot]) \tag{12}$$

$$\mathbf{U}^{(k+1)}[\cdot] = \mathbf{U}^{(k)}[\cdot] + (\mathbf{X}^{(k+1)}[\cdot] - \mathbf{Z}^{(k+1)}[\cdot]). \tag{13}$$

It can be shown (cf. [13, Sec. 3.1]) that for any $\rho > 0$, the iterates $\mathbf{X}^{(k)}[\cdot]$ converge to a solution of (6) i.e., $\lim_{k\to\infty} \mathbf{X}^{(k)}[\cdot] = \widehat{\mathbf{K}}[\cdot]$. Thus, while the precise choice for ρ has some influence on the convergence speed of ADMM [13, Sec. 3.4.1], it is not very crucial. We used $\rho = 100$ in all of our experiments (cf. Section IV).

Closed forms for updates (11) and (12) are stated in

Proposition III.1: Let us denote the eigenvalue decomposition of the matrix $\widehat{\mathbf{S}}[f] + \rho(\mathbf{U}^{(k)}[f] - \mathbf{Z}^{(k)}[f])$ by $\mathbf{V}_f\mathbf{D}_f\mathbf{V}_f^H$ with the diagonal matrix $\mathbf{D}_f$ composed of the eigenvalues $d_{f,i}$, sorted non-increasingly. Then, the ADMM update rule (11) is accomplished by setting, separately for each $f \in [F]$,

$$\mathbf{X}^{(k+1)}[f] = \mathbf{V}_f\widetilde{\mathbf{D}}_f\mathbf{V}_f^H \tag{14}$$

with the diagonal matrix $\widetilde{\mathbf{D}}_f$ whose ith diagonal element is given by $\tilde{d}_{f,i} = \min\{(1/(2\rho))[-d_{f,i} + \sqrt{d_{f,i}^2 + 4\rho}], 1\}$.

If we define the block-thresholding operator $\mathbf{W}[\cdot] = \mathcal{S}_\kappa(\mathbf{Y}[\cdot])$ via $W_{i,j}[f] := (1 - \kappa/\|Y_{i,j}[\cdot]\|)_+ Y_{i,j}[f]$, the update rule (12) results in

$$\mathbf{Z}^{(k+1)}[\cdot] = \mathcal{S}_{\lambda/\rho}(\mathbf{X}^{(k+1)}[\cdot] + \mathbf{U}^{(k)}[\cdot]). \tag{15}$$

Proof: Since the minimization problem (12) is equivalent to the ADMM update for a group LASSO problem [13, Sec. 6.4.2], the explicit form (15) follows from the derivation in [13, Sec. 6.4.2].

For the verification of (14), note that the optimization problem (11) splits into F separate subproblems, one for each $f \in [F]$. The subproblem for a specific frequency bin f is (omitting the index f)

$$\min_{\mathbf{0}\prec\mathbf{X}\preceq\mathbf{I}} -\log|\mathbf{X}| + \langle \mathbf{X}, \widehat{\mathbf{S}} + \rho(\mathbf{U}^{(k)} - \mathbf{Z}^{(k)})\rangle + (\rho/2)\|\mathbf{X}\|_{\mathrm{F}}^2. \tag{16}$$

Let us denote the non-increasing eigenvalues of the Hermitian matrices $\mathbf{X}$ and $\mathbf{Y} := \widehat{\mathbf{S}} + \rho(\mathbf{U}^{(k)} - \mathbf{Z}^{(k)})$ by x_i and d_i, for $i \in [p]$, respectively. According to [24, Lemma II.1], we have the trace inequality $\langle \mathbf{X}, \mathbf{Y}\rangle \geq \sum_{i\in[p]} x_i d_{p-i-1}$ with equality if $\mathbf{X}$ is of the form $\mathbf{X} = \mathbf{V}\mathrm{diag}\{d_{p-i-1}\}\mathbf{V}^H$ with a unitary matrix $\mathbf{V}$ whose columns are eigenvectors of $\mathbf{Y}$. Due to this trace inequality, a lower bound on (16) is

$$\min_{0\leq x_i\leq 1} \sum_{i\in[p]} -\log x_i + x_i d_{p-i+1} + (\rho/2)x_i^2. \tag{17}$$

The minimum in (17) is achieved by the choice $\tilde{x}_i = h(d_{p-i+1})$ with $h(z) := \min\{(-z + \sqrt{z^2 + 4\rho})/2\rho, 1\}$. However, for the choice $\mathbf{X} = \mathbf{V}\mathrm{diag}\{\tilde{x}_i\}\mathbf{V}^H$ (which is (14)), the objective function in (16) achieves the lower bound (17), certifying optimality. □

We summarize our GMS method in

Algorithm 1

Given samples $\mathbf{x}[1], \ldots, \mathbf{x}[N]$, parameters T, F, η, λ and window function $w[m]$ perform the steps:

Step 1: For each $f \in [F]$, compute the SDM estimate $\widehat{\mathbf{S}}[f]$ according to (8).

Step 2: Approximate the gLASSO $\widehat{\mathbf{K}}[\cdot]$ (cf. (6)) by iterating (14), (15) and (13) for a fixed number T yielding $\mathbf{X}^{(T)}[\cdot]$.

Step 3: Estimate the edge set via $\widehat{\mathcal{E}}(\eta) = \{(i,j) : \|X_{i,j}^{(T)}[\cdot]\| \geq \eta\}$.

C. Performance Analysis

Let us for simplicity assume that the ADMM iterates $\mathbf{X}^{(k)}[\cdot]$ converged perfectly to the gLASSO estimate $\widehat{\mathbf{K}}[\cdot]$ given by (6), i.e., $\mathbf{X}^{(T)}[\cdot] = \widehat{\mathbf{K}}[\cdot]$. Then, under Assumption II.1, a sufficient condition for our GMS method to recover the edge set of the CIG $\mathcal{G}$ is (10).

We will now derive an upper bound on the probability that (10) fails to hold. This will be accomplished by (i) showing that the norm $\|\mathbf{\Delta}\|_1$ can be bounded in terms of the SDM estimation error $\mathbf{E}[f] := \mathbf{S}[f] - \widehat{\mathbf{S}}[f]$ and (ii) controlling the probability that the error $\mathbf{E}[f]$ is sufficiently small.

Upper bounding $\|\mathbf{\Delta}\|_1$. By definition of $\widehat{\mathbf{K}}[\cdot]$ (cf. (6)),

$$-A\{\widehat{\mathbf{K}}\} + \langle \mathbf{\Delta}[\cdot], \widehat{\mathbf{S}}[\cdot]\rangle + \lambda(\|\widehat{\mathbf{K}}\|_1 - \|\mathbf{K}\|_1) \leq -A\{\mathbf{K}\}. \tag{18}$$

Combining with $\arg\min_{\mathbf{X}[\cdot]\in\mathcal{C}} -A\{\mathbf{X}\} + \langle \mathbf{X}[\cdot], \mathbf{S}[\cdot]\rangle = \mathbf{K}[\cdot]$,

$$\lambda\|\widehat{\mathbf{K}}\|_1 \leq \lambda\|\mathbf{K}\|_1 + \langle \mathbf{\Delta}[\cdot], \mathbf{E}[\cdot]\rangle. \tag{19}$$

Let us, for the moment, assume validity of the condition

$$E := \max_{f\in[F]} \|\mathbf{E}[f]\|_\infty \leq \lambda/2, \tag{20}$$

implying $|\langle \mathbf{\Delta}[\cdot], \mathbf{E}[\cdot]\rangle| \leq (\lambda/2)\|\mathbf{\Delta}\|_1$ and, in turn via (19),

$$\lambda\|\widehat{\mathbf{K}}\|_1 \leq \lambda\|\mathbf{K}\|_1 + (\lambda/2)\|\mathbf{\Delta}\|_1. \tag{21}$$

Below, we present a specific choice for the tuning parameter λ in (6) such that (20) holds with high probability. Using $\|\widehat{\mathbf{K}}\|_1 \overset{(\star)}{=} \|\widehat{\mathbf{K}}_{\mathcal{S}^c}\|_1 + \|\mathbf{\Delta}_{\mathcal{S}} + \mathbf{K}_{\mathcal{S}}\|_1$ and $\|\mathbf{\Delta}\|_1 = \|\mathbf{\Delta}_{\mathcal{S}}\|_1 + \|\mathbf{\Delta}_{\mathcal{S}^c}\|_1$ and the (reverse) triangle inequality, (21) yields

$$\|\mathbf{\Delta}_{\mathcal{S}^c}\|_1 \overset{(\star)}{=} \|\widehat{\mathbf{K}}_{\mathcal{S}^c}\|_1 \leq 3\|\mathbf{\Delta}_{\mathcal{S}}\|_1, \tag{22}$$

where $(\star)$ is due to $\mathcal{S} = \mathrm{gsupp}(\mathbf{K}[\cdot])$. Thus, the estimation error $\mathbf{\Delta}[\cdot]$ tends to be sparse.

As a next step, we rewrite (18) as

$$-(A\{\widehat{\mathbf{K}}\} - A\{\mathbf{K}\}) + \langle \mathbf{\Delta}[\cdot], \mathbf{S}[\cdot]\rangle \leq (3\lambda/2)\|\mathbf{\Delta}\|_1 \tag{23}$$

where we used (20). Let us denote the eigenvalues of the psd matrix $\mathbf{\Gamma}[f] := \mathbf{S}^{1/2}[f]\mathbf{\Delta}[f]\mathbf{S}^{1/2}[f]$ by $\gamma_{f,i}$. We can then rewrite the LHS of (23) as

$$-(A\{\widehat{\mathbf{K}}\} - A\{\mathbf{K}\}) + \langle \mathbf{\Delta}[\cdot], \mathbf{S}[\cdot]\rangle = \frac{1}{F}\sum_{f,i} -\log(1+\gamma_{f,i}) + \gamma_{f,i}.$$

Since $\widehat{\mathbf{K}}[f], \mathbf{K}[f] \preceq \mathbf{I}$ and $\mathbf{S}[f] \preceq U\mathbf{I}$ (cf. (7), (2)), we have $\gamma_{f,i} \leq 2U$. Using $\log(1+x) = x - \frac{x^2}{2(1+\varepsilon x)^2}$, with some $\varepsilon \in [0,1]$, we further obtain

$$-(A\{\widehat{\mathbf{K}}\} - A\{\mathbf{K}\}) + \langle \mathbf{\Delta}[\cdot], \mathbf{S}[\cdot]\rangle \geq \|\mathbf{\Gamma}[\cdot]\|_{\mathrm{F}}^2/(4(2U+1)^2).$$

Applying [25, Lemma 4.3.1(b)] to the RHS,

$$\begin{aligned} &-(A\{\widehat{\mathbf{K}}\} - A\{\mathbf{K}\}) + \langle \mathbf{\Delta}[\cdot], \mathbf{S}[\cdot]\rangle \qquad (24)\\ &\geq \frac{1}{4F(2U+1)^2}\sum_{f\in[F]} \mathrm{vec}\{\mathbf{\Delta}[f]\}^H(\bar{\mathbf{S}}[f]\otimes\mathbf{S}[f])\mathrm{vec}\{\mathbf{\Delta}[f]\}. \end{aligned}$$

Combining (24) with (23) and Assumption II.3 (which can be invoked due to (22)), we arrive at

$$\|\mathbf{\Delta}\|_1 \leq 96(2U+1)^2(s/\phi)\lambda. \tag{25}$$

Controlling the SDM estimation error. It remains to control the probability that condition (20) is valid, i.e., the SDM estimation error $\mathbf{E}[f]$ incurred by the BT estimator (8) is sufficiently small. According to [26, Lemma IV.4], for any $\nu \in [0, 1/2)$,

$$P\{E \geq \nu + \mu_x^{(h_1)}\} \leq 2e^{-\frac{(1/32)N\nu^2}{\|w[\cdot]\|_1^2U^2} + 2\log(2pN)} \tag{26}$$

where $\mu_x^{(h_1)}$ is the ACF moment (3) with $h_1[m] := |1 - w[m](1-|m|/N)|$ for $|m| < N$ and $h_1[m] := 1$ else.

The main result. Recall that a sufficient condition for $\widehat{\mathcal{E}}(\rho_{\min}/2)$, given by (9), to coincide with the true edge set is (10). Under the condition (20), implying validity of (25), the inequality (10) will be satisfied if λ is chosen as

$$\lambda = \phi\rho_{\min}/(192s(2U+1)^2). \tag{27}$$

Using (26) to bound the probability for (20) to hold yields

Proposition III.2: Consider a stationary Gaussian zero-mean time series $\mathbf{x}[n]$ satisfying (2) and Assumption II.1-II.3. Then, using the choice (27) in (6) and if the conditions

$$\mu_x^{(h_1)} \leq \phi\rho_{\min}/(384s(2U+1)^2) \tag{28}$$

$$N/\log(4Np^2/\delta) \geq 10^8(2U+1)^6s^2\phi^{-2}\rho_{\min}^{-2}\|w[\cdot]\|_1^2 \tag{29}$$

are satisfied, we have $P\{\widehat{\mathcal{E}}(\rho_{\min}/2) \neq \mathcal{E}\} \leq \delta$.

In order to satisfy the condition (28), the window function $w[\cdot]$ in (8) has to be chosen as the indicator function for the effective support of the ACF $\mathbf{R}[m]$. Thus, the factor $\|w[\cdot]\|_1^2$ in (29) corresponds to a scaling of the sample size with the square of the effective ACF width. Moreover, the sufficient condition (29) scales inversely with the square of the minimum partial spectral coherence $\rho_{\min}$ which agrees with the scaling of the sufficient condition obtained for the neighborhood regression approach in [8]. Note that, while our performance analysis applies only to Gaussian time series, the GMS method in Algorithm I can also be applied to non-Gaussian time series. However, for a non-Gaussian time series the resulting edge estimate $\widehat{\mathcal{E}}(\eta)$ (cf. (9)) is then not related to a CIG anymore but to a partial correlation graph [3].

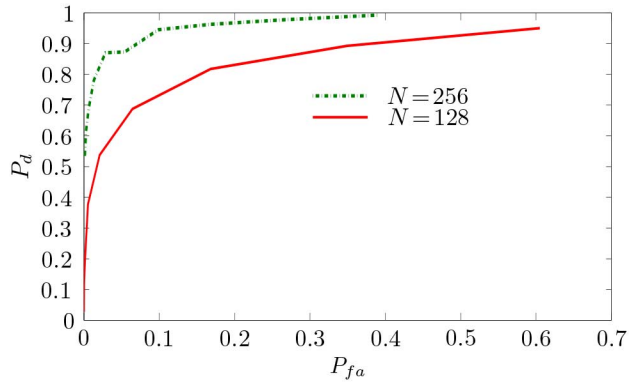

Fig. 1. ROC curves of gLASSO based GMS method.

IV. Numerical Results

We generated a Gaussian time series $\mathbf{x}[n]$ of dimension $p = 64$ by applying a finite impulse response filter $g[n]$ of length 2 to a zero-mean, stationary, white, Gaussian noise process $\mathbf{e}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_0)$. We choose the covariance matrix $\mathbf{C}_0$ such that the resulting CIG $\mathcal{G} = ([p], \mathcal{E})$ is a star graph containing a hub node with $|\mathcal{E}| = 4$ neighbors. The corresponding precision matrix $\mathbf{K}_0 = \mathbf{C}_0^{-1}$ has main diagonal entries equal to 0.5 and off diagonal entries equal to 0.1. The filter coefficients $g[n]$ are such that the magnitude of the associated transfer function is uniformly bounded from above and below by positive constants, thereby ensuring that conditions (2) and (4) are satisfied with $L = 1$, $U = 10$ and $F = 4$. Thus, the generated time series satisfies Assumption II.1 - II.3 with $\rho_{\min} = 0.1\sqrt{\sum_{m=-\infty}^{\infty} r_g^2[m]}$, $s = 4$ and $\phi = 1$. Here, $r_g[m] = \sum_{n=-\infty}^{\infty} g[n+m]g[n]$ denotes the autocorrelation sequence of $g[n]$.

Based on $N \in \{128, 256\}$ observed samples, we estimated the edge set of the CIG using Algorithm 1 with $L = 10$ ADMM iterations, $F = 4$ frequency points and window function $w[m] = \exp(-m^2)$. The gLASSO parameter λ (cf. (6) and (27)) was varied in the range $[c_1, c_2] \times \phi\rho_{\min}/(192s(2U+1)^2)$, where constants c_1 and c_2 have been tuned empirically.

In Fig. 1, we show receiver operating characteristic (ROC) curves with the empirical false alarm rate $P_{fa} := \frac{1}{M}\sum_{i\in[M]} \frac{|\widehat{\mathcal{E}}_i \setminus \mathcal{E}|}{p(p-1)/2 - |\mathcal{E}|}$ and the empirical detection probability $P_d := \frac{1}{M}\sum_{i\in[M]} \frac{|\widehat{\mathcal{E}}_i \cap \mathcal{E}|}{|\mathcal{E}|}$, both averaged over $M = 100$ independent simulation runs. Here, $\widehat{\mathcal{E}}_i$ denotes the edge estimate in the i-th simulation run.

REFERENCES

[1] J. Gohlke, O. Armant, F. Parham, M. Smith, C. Zimmer, D. Castro, L. Nguyen, J. Parker, G. Gradwohl, C. Portier, and F. Guillemot, "Characterization of the proneural gene regulatory network during mouse telencephalon development," *BMC Biology*, vol. 6, no. 1, p. 15, 2008.

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

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

[4] A. Abdelwahab, O. Amor, and T. Abdelwahed, "The analysis of the interdependence structure in international financial markets by graphical models," *Int. Res. J. Finance Econ.*, pp. 291–306, 2008.

[5] J. Dauwels, H. Yu, X. Wang, F. Vialatte, C. Latchoumane, J. Jeong, and A. Cichocki, "Inferring brain networks through graphical models with hidden variables," in *NIPS 2011 Workshop on Machine Learning and Interpretation in Neuroimaging*, Sierra Nevada, Spain, Dec. 2011.

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

[7] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *J. Mach. Learn. Res.*, vol. 11, pp. 2671–2705, 2010.

[8] 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.

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

[10] 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.

[11] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*. New York: Springer, 2011.

[12] D. M. Witten, J. H. Friedmann, and N. Simon, "New insights and faster computations for the graphical lasso," *J. Comput. Graph. Stat.*, vol. 20, no. 4, pp. 892–900, 2011.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[14] A. Wiesel and A. O. Hero, "Distributed covariance estimation in Gaussian graphical models," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 211–220, Jan. 2012.

[15] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *J. R. Statist. Soc. B*, vol. 76, pp. 373–397, 2014.

[16] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

[17] M. Eichler, "Graphical Models in Time Series Analysis," Ph.D. thesis, Univ. Heidelberg, Heidelberg, Germany, 1999.

[18] S. A. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the Lasso," *Electron. J. Statist.*, vol. 3, pp. 1360–1392, 2009.

[19] P. Ravikumar, M. J. Wainwright, B. Raskutti, and G. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.

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

[21] P. Whittle, "Estimation and information in time series," *Ark. Mater.*, vol. 2, pp. 423–434, 1953.

[22] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Englewood Cliffs, NJ, USA: Prentice Hall, 1997.

[23] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.

[24] J. B. Lasserre, "A trace inequality for matrix product," *IEEE Trans. Automat. Contr.*, vol. 40, no. 8, Aug. 1995.

[25] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[26] A. Jung, "Learning the conditional independence structure of stationary time series: A multitask learning approach," *IEEE Trans. Signal Process.* [Online]. Available: http://arxiv.org/pdf/1404.1361v3.pdf, submitted for publication