Compressive Nonparametric Graphical Model Selection for Time Series: A Multitask Learning Approach

Alexander Jung[1], *Reinhard Heckel*[2], *Helmut Bölcskei*[2], *and Franz Hlawatsch*[1]

[1]**Institute of Telecommunications, Vienna University of Technology, Austria**
[2]**Dept. IT & EE, ETH Zurich, Switzerland**

Introduction

- Consider a p-dimensional vector-valued Gaussian process
$$\mathbf{x}[n] = (x_1[n], \ldots, x_p[n])^T.$$
- We model $\mathbf{x}[n]$ as stationary with summable auto-covariance function (ACF)
$$\mathbf{R}[m] := \mathrm{E}\{\mathbf{x}[m]\mathbf{x}^T[0]\}.$$
- The spectral density matrix (SDM) is
$$\mathbf{S}(\theta) := \sum_{m=-\infty}^{\infty} \mathbf{R}[m] \exp(-j2\pi m\theta), \text{ for } \theta \in [0,1).$$
- We require the SDM to be sufficiently smooth, or equivalently we require small ACF moments
$$\mu^{(h)} := \sum_{m=-\infty}^{\infty} h[m] \|\mathbf{R}[m]\|_{\infty}. \qquad (1)$$
- Consider the conditional independence graph (CIG) $\mathcal{G} = ([p] := \{1, \ldots, p\}, E)$ of $\mathbf{x}[n]$.
- For a Gaussian process, $(k,l) \notin E \iff \left[\mathbf{S}^{-1}(\theta)\right]_{k,l} \equiv 0$.
- Our interest is in sparse CIGs, containing few edges.
- Goal: Estimate sparse CIG from a finite-length observation $\mathbf{x}[1], \ldots, \mathbf{x}[N]$, where typically $N \ll p$.
- We propose a nonparametric compressive selection scheme for the CIG of a stationary vector process.

Neighborhood Regression

- First, consider the special case of an i.i.d. sampling process, i.e., $\mathbf{R}[m] = \mathbf{C}\delta[m]$.
- This corresponds to model selection for a Gaussian Markov random field (GMRF) with covariance matrix $\mathbf{C}$.
- Here, the SDM is flat, i.e., $\mathbf{S}(\theta) = \mathbf{C}$ for all $\theta \in [0,1)$.
- Determine neighborhood $\mathcal{N}(r) := \{l : (r,l) \in E\}$ by regressing $x_r[n]$ on the remaining components.
- For the i.i.d. case, this regression becomes
$$x_r[n] = \sum_{l \in [p] \setminus \{r\}} \beta_l x_l[n] + w[n],$$
with $\beta_l = -\left[\mathbf{C}^{-1}\right]_{r,l} / \left[\mathbf{C}^{-1}\right]_{r,r}$.
- Since $\mathcal{N}(r)$ coincides with $\mathrm{supp}(\boldsymbol{\beta})$, with $\boldsymbol{\beta} := \mathrm{vec}\{\beta_l\}_{l \in [p]\setminus\{r\}}$ determining the neighborhood is reduced to sparse support recovery!
- LASSO based selection scheme proposed by [Meinshausen & Bühlmann, 2006].

Multitask Learning Formulation

- For a general stationary process, we perform neighborhood regression in the frequency domain.
- Let $\widehat{\mathbf{S}}(\theta_f)$ denote an estimate of the SDM $\mathbf{S}(\theta_f)$ for $\theta_f := (f-1)/F$, $f \in [F]$.
- For each frequency θ_f, $f \in [F]$, we define
$$\mathbf{y}^{(f)} = \mathbf{X}^{(f)} \boldsymbol{\beta}^{(f)} + \mathbf{w}^{(f)}, \qquad (2)$$
with
$$\left(\mathbf{y}^{(f)}\ \mathbf{X}^{(f)}\right) := \sqrt{\mathbf{P}_{1\leftrightarrow r} \widehat{\mathbf{S}}(\theta_f) \mathbf{P}_{1\leftrightarrow r}}, \qquad (3)$$
and parameter vectors
$$\boldsymbol{\beta}^{(f)} := \left[\mathbf{S}(\theta_f)\right]^{-1}_{[p]\setminus\{r\},[p]\setminus\{r\}} \left[\mathbf{S}(\theta_f)\right]_{[p]\setminus\{r\},r}. \qquad (4)$$
- Based on (4) it can be shown that
$$\mathcal{N}(r) = \mathrm{gsupp}(\boldsymbol{\beta}) := \bigcup_{f \in [F]} \mathrm{supp}(\boldsymbol{\beta}^{(f)}),$$
with stacked parameter vector $\boldsymbol{\beta} := \left(\left(\boldsymbol{\beta}^{(1)}\right)^T, \ldots, \left(\boldsymbol{\beta}^{(F)}\right)^T\right)^T$.
- If the CIG is sparse, i.e., $|\mathcal{N}(r)| \ll p$, it follows that $\boldsymbol{\beta}$ is a block-sparse vector.
- Recovering a block-sparse vector $\boldsymbol{\beta}$ from the measurements (2) is recognized as a multitask learning problem.
- A popular approach to this is multitask LASSO (mLASSO).

Novel Selection Scheme

- Let $w[m]$ denote a window function with non-negative discrete time Fourier transform.
- We propose the following selection scheme:

> - First, for each θ_f, we compute a multivariate Blackman-Tukey SDM estimate
> $$\widehat{\mathbf{S}}(\theta) := \sum_{m=-N+1}^{N-1} w[m] \widehat{\mathbf{R}}[m] \exp(-j2\pi\theta m).$$
> Here, $\widehat{\mathbf{R}}[m] := (1/N) \sum_{n=1}^{N-m} \mathbf{x}[n+m]\mathbf{x}^T[n]$ for $m \in \{0, \ldots, N-1\}$ and, by symmetry of the ACF, $\widehat{\mathbf{R}}[m] := \widehat{\mathbf{R}}^T[-m]$ for $m \in \{-N+1, \ldots, -1\}$.
> - Second, based on $\widehat{\mathbf{S}}(\theta_f)$, $f \in [F]$, we construct $\mathbf{y}^{(f)}$ and $\mathbf{X}^{(f)}$ according to (3) and compute mLASSO estimate
> $$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ \frac{1}{F} \sum_{f \in [F]} \|\mathbf{y}^{(f)} - \mathbf{X}^{(f)} \boldsymbol{\beta}^{(f)}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{2,1} \right\},$$
> where $\|\boldsymbol{\beta}\|_{2,1} := \sum_{r \in [q]} \|\boldsymbol{\beta}_r\|_2$ with $\boldsymbol{\beta}_r := \left(\left[\boldsymbol{\beta}^{(1)}\right]_r, \ldots, \left[\boldsymbol{\beta}^{(F)}\right]_r\right)^T \in \mathbb{C}^F$.
> - The neighborhood $\mathcal{N}(r)$ is finally estimated by
> $$\widehat{\mathcal{N}}(r) := \{l : \|\hat{\boldsymbol{\beta}}_l\|_2 > \tau\}.$$

- Scheme is modular: Different combinations of SDM estimators and sparse support recovery schemes possible.

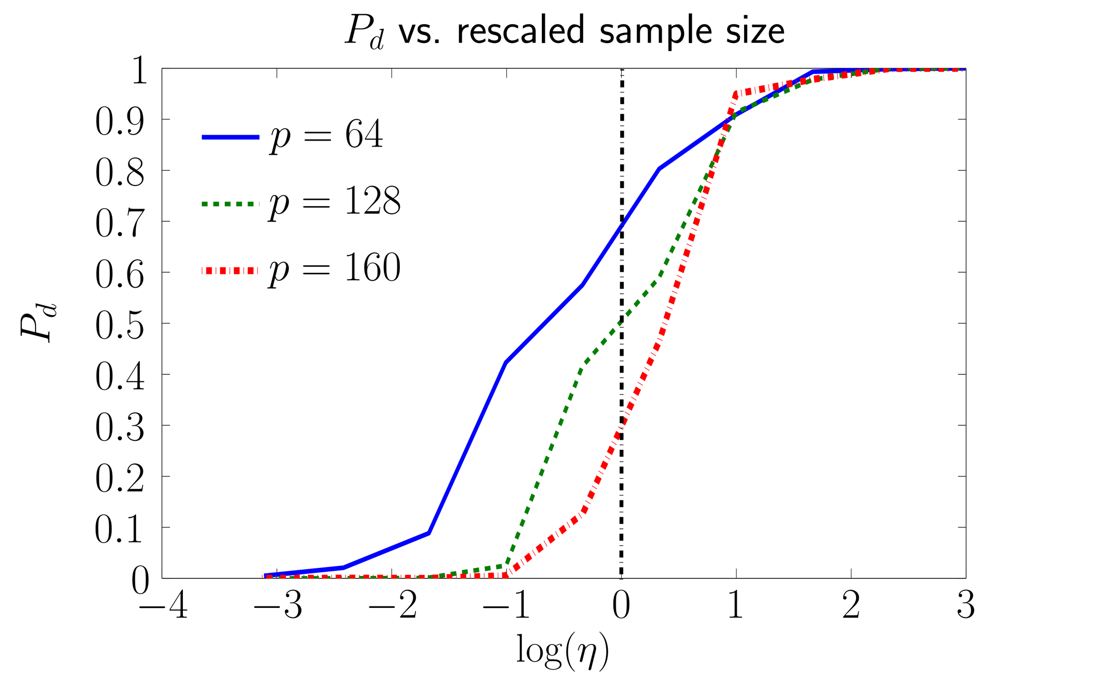
Performance Guarantees

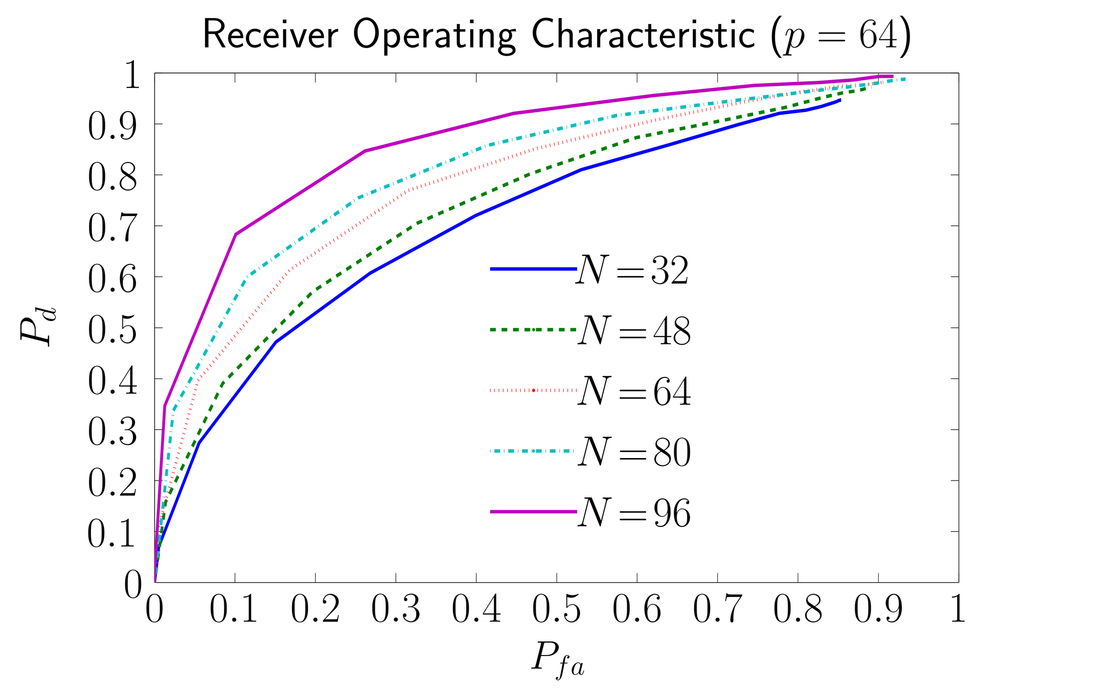
- Denote by $\mathcal{M}(A, B, s_{\max}, \rho_{\min}, \mu^{(h_1)}, \phi_{\min})$ the class of stationary Gaussian processes that satisfy the following conditions
  - Uniform boundedness of SDM eigenvalues:
  $$0 < A \leq \lambda_{\min}(\mathbf{S}(\theta)) \leq \lambda_{\max}(\mathbf{S}(\theta)) \leq B < \infty.$$
  This technical assumption ensures certain Markov properties of the CIG.
  - Maximum node degree $s_{\max}$: We consider sparse CIGs, whose maximum node degree is bounded as
  $$|\mathcal{N}(r)| \leq s_{\max} \ll p.$$
  - Minimum partial coherence $\rho_{\min} > 0$: This parameter quantifies the minimum partial correlation between the spectral components of the process. In particular, we require
  $$\sum_{f \in [F]} \left| \left[\mathbf{S}^{-1}(\theta_f)\right]_{r,r'} / \left[\mathbf{S}^{-1}(\theta_f)\right]_{r,r} \right|^2 \geq \rho_{\min}^2$$
  for all $r \in [p]$, $r' \in \mathcal{N}(r)$.
  - ACF moment $\mu^{(h_1)}$: We quantify the smoothness of the processes in $\mathcal{M}$ using the ACF moment (1) with weight function $h_1[m] := |1 - w[m](1 - |m|/N)|$.
  - Minimum multitask compatibility constant $\phi_{\min} > 0$: For every process in $\mathcal{M}$, we require
  $$s_{\max} \sum_{f \in [F]} \left(\boldsymbol{\beta}^{(f)}\right)^H \mathbf{S}(\theta_f) \boldsymbol{\beta}^{(f)} \geq \phi_{\min}^2 \|\boldsymbol{\beta}_{\mathcal{N}(r)}\|_{2,1}^2$$
  for all $r \in [p]$ and all vectors $\boldsymbol{\beta} \in \mathbb{C}^{qF}$ such that
  $$\|\boldsymbol{\beta}_{\mathcal{S}}\|_{2,1} > 0 \text{ and } \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_{2,1} \leq 3\|\boldsymbol{\beta}_{\mathcal{S}}\|_{2,1}.$$
- We choose mLASSO parameter $\lambda = \phi_{\min}^2 \rho_{\min} / (18 s_{\max} F)$ and threshold $\tau = \rho_{\min}/2$.
- Combining a deterministic analysis of mLASSO with a large-deviation characterization of the SDM estimator, we derived the following result:

Theorem 1 *Consider a p-dimensional stationary Gaussian time series $\mathbf{x}[n]$ belonging to $\mathcal{M}(A, B, s_{\max}, \rho_{\min}, \mu^{(h_1)}, \phi_{\min})$. Then, if for some $\delta > 0$, the rescaled sample size $\eta := N/(\log(p) s_{\max}^3)$ and the correlation moment $\mu^{(h_1)}$ satisfy*
$$\eta > 10^3 \log\left(\frac{4F}{\delta}\right) \|w\|_1^2 B^2 / \kappa^2 \text{ and } \mu^{(h_1)} \leq \frac{\kappa}{2 s_{\max}^{3/2}},$$
with $\kappa := (\phi_{\min}^2/174) \rho_{\min} \sqrt{\frac{AF}{B}}$, the probability of correctly selecting the edge set E is at least $1 - \delta$, i.e.,
$$\mathrm{P}\left\{ \bigcap_{r \in [p]} \{\widehat{\mathcal{N}}(r) = \mathcal{N}(r)\} \right\} \geq 1 - \delta.$$

Numerical Results

- We applied our method to a synthetic process obtained by filtering a p-dimensional white Gaussian process with a FIR filter of length 2.
- The filtered process has a sparse CIG with $s_{\max} = 3$.
- We computed empirical false alarm (P_{fa}) and detection ratios (P_d) based on 10 simulation runs.





- Our method yields reasonable performance even if $N = 32$ only for a 64-dimensional process.
- The rescaled sample size η seems to be a good performance indicator.

institute of telecommunications

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology