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

Alexander Jung
Institute of Telecommunications
Vienna University of Technology

joint work with Reinhard Heckel (ETH Zurich), Helmut Bölcskei (ETH Zurich), and Franz Hlawatsch (Vienna University of Technology)
What is it all about?

- Consider a $p$-dimensional stationary vector-valued Gaussian time series $\mathbf{x}[n]$.

- Based on observing a finite snapshot $\mathbf{x}[1], \ldots, \mathbf{x}[N]$, we want to infer the correlation structure behind the time series.

- In particular, we want to estimate the conditional independence graph (CIG) of the time series.

- We are interested in the high-dimensional regime, where $p$ may be much larger than the sample size $N$, i.e., $N \ll p$. 
• CIG of a stationary time series

• LASSO based graphical model selection for i.i.d. observations

• Multitask-LASSO based graphical model selection for time series

• Selection consistency

• Numerical results
Outline

- CIG of a stationary time series
- LASSO based graphical model selection for i.i.d. observations
- Multitask-LASSO based graphical model selection for time series
- Selection consistency
- Numerical results
The Spectral Density Matrix

- Consider a vector valued zero-mean stationary Gaussian time series

\[ \mathbf{x}[n] = \begin{pmatrix} x_1[n] \\ \vdots \\ x_p[n] \end{pmatrix}, \quad n \in \mathbb{Z}, \]

with autocovariance function (ACF)

\[ \mathbf{R}_x[m] \triangleq \mathbb{E}\{\mathbf{x}[m]\mathbf{x}^T[n]\}. \]

- We assume the ACF to be summable, i.e., \( \sum_{m=-\infty}^{\infty} \|\mathbf{R}_x[m]\| < \infty \), so that the spectral density matrix (SDM)

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

exists for all \( \theta \in [0, 1) \).

- We require uniform boundedness of the SDM eigenvalues, i.e.,

\[ L \leq \lambda_{\min}(\mathbf{S}_x(\theta)) \leq \lambda_{\max}(\mathbf{S}_x(\theta)) \leq U, \]

for some constants \( U, L > 0 \).
The CIG of a Time Series

- We can associate with the time series $x[n]$ a simple undirected graph, referred to as the conditional independence graph (CIG), $G_x = (V, E)$.

- The elements of the node set $V \triangleq \{1, \ldots, p\}$ represent the component time series $\{x_l[n]\}_{l \in V}$.

- Two nodes $i, j \in V$ are connected, i.e., $(i, j) \in E$ if and only if $x_i[\cdot]$ and $x_j[\cdot]$ are conditionally dependent given the remaining components $\{x_l[\cdot]\}_{l \in \{1, \ldots, p\}\setminus\{i,j\}}$.

- For a stationary Gaussian time series with SDM $S_x(\theta)$,

$$ (i, j) \notin E \iff (S_x^{-1}(\theta))_{i,j} \equiv 0. $$
The CIG of a Gaussian Markov random field (GMRF) may be interpreted as the CIG of an i.i.d. time series $x[n]$ with $x[n] \sim \mathcal{N}(0, \Sigma)$.

The SDM is then constant, i.e., $S_x(\theta) \equiv \Sigma$ and therefore,

$$(S_x^{-1}(\theta))_{i,j} \equiv 0 \iff (\Sigma^{-1})_{i,j} = 0.$$

LASSO based graphical model selection for a GMRF based on i.i.d. observations $x[n]$ was studied by Meinshausen & Bühlmann, 2006.

By contrast, we consider a correlated time series, i.e., we take sample memory into account.
• Consider CIG associated with $\mathbf{x}[n] = (x_1[n], x_2[n], x_3[n], x_4[n], x_5[n])^T$

- The node set $B$ separates $A$ from $C$.
- **Global Markov property:** $\{x_r[n]\}_{r \in A}$ conditionally independent of $\{x_r[n]\}_{r \in C}$ given $\{x_r[n]\}_{r \in B}$. 
Outline

- CIG of a stationary time series
- LASSO based graphical model selection for i.i.d. observations
- Multitask-LASSO based graphical model selection for time series
- Selection consistency
- Numerical results
Graphical Model Selection

- Because \((i, j) \notin E \iff (S_{x}^{-1}(\theta))_{i,j} \equiv 0\), graphical model selection is equivalent to finding the zero entries of the inverse SDM \(S_{x}^{-1}(\theta)\).

- Therefore, [Dahlhaus, 2000] proposes to perform graphical model selection by
  1. computing an SDM estimate \(\hat{S}_{x}(\theta)\) based on the observation \(x[1], \ldots, x[N]\);
  2. estimating the edge set \(E\) by detecting the zero (negligible) entries of \(\hat{S}_{x}^{-1}(\theta)\).

- However, in the high-dimensional regime where \(N \ll p\), the estimate \(\hat{S}_{x}(\theta)\) will be singular and thus cannot be inverted!

- To cope with this, step 2 is replaced by a compressed sensing (CS) recovery method.
Graphical Model Selection for GMRF

- Consider a GMRF $\mathbf{x} = (x_1, \ldots, x_p) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with CIG $\mathcal{G}_x = (V, E)$.

- There is an edge between $i, j \in V$ if and only if $K_{i,j} \neq 0$ with the precision matrix $\mathbf{K} \triangleq \Sigma^{-1}$.

- Note that the edge set $E$ is completely determined by the neighborhoods $\mathcal{N}(r) \triangleq \{i \in V | (i, r) \in E\}$.

- We require CIG to be sparse, i.e., $\max_{r \in V} |\mathcal{N}(r)| \leq s_{\max} \ll p$.

- Neighborhood regression estimates the neighborhood $\mathcal{N}(r)$ by regressing $x_r$ on the remaining components $\{x_l\}_{l \in V \setminus \{r\}}$.

- In what follows, without loss of generality, we focus on the neighborhood $\mathcal{N}(1)$ of node $r = 1$. 
Neighborhood Regression

- Regressing (in MMSE sense) $x_1$ on the remaining components yields

$$x_1 = \sum_{l=2}^{p} \beta_l x_l + w.$$ 

- The Gaussian variable $w$ is independent of, or orthogonal to, $\{x_l\}_{l \in V \setminus \{r\}}$.

- The (MMSE-)coefficients $\beta_l$ are given as $\beta_l = -K_{1,l}/K_{1,1}$.

- Note that $\text{supp}(\beta) = \mathcal{N}(1) - 1$.

- For a sparse CIG, i.e., $|\mathcal{N}(1)| \ll p$, we obtain a sparse linear model with parameter vector $\beta \triangleq (\beta_2, \ldots, \beta_p)^T \in \mathbb{R}^{p-1}$.

- Thus, in this framework, finding $\mathcal{N}(1)$ is equivalent to sparse support recovery!
Neighborhood Regression

- Many methods are available, e.g., LASSO based support recovery.

- The above sparse linear model lives in the abstract Hilbert space $\mathcal{L} \triangleq \text{span}\{x_l\}_{l \in V}$ with inner product $\langle a, b \rangle_{\mathcal{L}} \triangleq \mathbb{E}\{ab\}$.

- Crucial observation: The LASSO estimate $\hat{\beta}$ depends only on the inner products $\langle x_l, x_{l'} \rangle_{\mathcal{L}} = \Sigma_{l,l'}$.

- If we use the partition

$$
\Sigma = \begin{pmatrix}
\gamma & c \\
c^H & G
\end{pmatrix},
$$

the LASSO estimate is given as

$$
\hat{\beta} = \hat{\beta}(\Sigma) \triangleq \arg\min_{\beta \in \mathbb{R}^{p-1}} \{ \beta^H G \beta - 2\text{Re}\{c\beta\} + \lambda \|\beta\|_1 \}. 
$$
The Meinshausen & Bühlmann Approach

- We do not have access to the “vectors” $x_l \in \mathcal{L}$ but only to $N$ i.i.d. realizations $x[n]$.

- Estimate the covariance matrix $\Sigma$ by the empirical covariance matrix

$$\hat{\Sigma} = \begin{pmatrix} \hat{\gamma} & \hat{c} \\ \hat{c}^H & \hat{G} \end{pmatrix} \triangleq \frac{1}{N} \sum_{n=1}^{N} x[n]x^T[n].$$

- Compute a LASSO estimate $\hat{\beta}(\hat{\Sigma})$ based on $\hat{\Sigma}$ instead of $\Sigma$.

- The neighborhood $\mathcal{N}(1)$ is then estimated by $\hat{\mathcal{N}}(1) \triangleq \text{supp}(\hat{\beta}) + 1$.

- Conditions for asymptotic ($N \to \infty$) consistency are given in the seminal paper by Meinshausen & Bühlmann, 2006.
Outline

- CIG of a stationary time series
- LASSO-based graphical model selection for i.i.d. observations
- Multitask-LASSO based graphical model selection for time series
- Selection consistency
- Numerical results
Recap: Estimate neighborhood $\mathcal{N}(1)$ of the CIG of a GMRF by performing support detection for a sparse linear model associated with (a noisy version of) the covariance matrix $\Sigma$.

The SDM $S_x(\theta)$ of a stationary time series gives a “covariance matrix” for each frequency $\theta \in [0, 1)$.

Let us sample the SDM $S_x(\theta)$ regularly at $F$ frequency points $\theta_f \triangleq f/F$, for $f = 0, \ldots, F - 1$.

We may interpret $S_x(\theta_f)$ as the “covariance matrix” of a sparse linear model.
Multitask Learning Formulation

• For each $f \in \{1, \ldots, F\}$, let us construct a sparse linear model
  \[ y^{(f)} = X^{(f)} \beta^{(f)} + w^{(f)} \]
  such that
  \[ S_x(\theta_f) = 
  \begin{pmatrix}
  \|y^{(f)}\|_2^2 & (y^{(f)})^H X^{(f)} \\
  (X^{(f)})^H y^{(f)} & (X^{(f)})^H X^{(f)}
  \end{pmatrix} \]
  and the parameter vector $\beta \triangleq (\beta^{(1)}, \ldots, \beta^{(F)})^T$ satisfies
  \[ \text{gsupp}(\beta) \triangleq \bigcup_{f \in F} \text{supp}(\beta^{(f)}) = \mathcal{N}(1) - 1. \]

• Then, recovering the vector $\beta$ from $\{y^{(f)}\}_{f \in \{1, \ldots, F\}}$ is a multitask learning problem.

• A multitask learning problem is an instance of a group sparse recovery problem.
Multitask LASSO

- A popular approach to recover $\beta$ is the multitask LASSO (MLASSO):

$$
\hat{\beta} \triangleq \arg\min_{\beta'} \left\{ \frac{1}{F} \sum_{f=1}^{F} \| y^{(f)} - X^{(f)} \beta^{(f)} \|^2_2 + \lambda \| \beta' \|_{2,1} \right\}.
$$

- Partitioning the SDM $S_x(\theta)$ as

$$
S_x(\theta) = \begin{pmatrix}
\gamma(f) & c(f) \\
c^H(f) & G(f)
\end{pmatrix},
$$

we can reformulate the MLASSO as

$$
\hat{\beta} \triangleq \arg\min_{\beta'} \left\{ \frac{1}{F} \sum_{f=1}^{F} \left[ \left( \beta'(f) \right)^H G(f) \beta'(f) - 2 \Re \{ c^H(f) \beta'(f) \} \right] + \lambda \| \beta' \|_{2,1} \right\}.
$$
Since the (generalized) support of $\beta$ determines $\mathcal{N}(1)$, we can perform model selection by running MLASSO with the SDM $S_x(\theta)$.

However, since the SDM is unknown, we replace it with some estimate.

Here, we use a multivariate extension of the Blackman-Tukey estimator of the power spectral density (PSD) of a scalar time series.
The Novel Selection Scheme

- Two-step procedure:
  1. For every \( f \in \{1, \ldots, F\} \), compute the SDM estimate

\[
\hat{S}_x(\theta_f) = \begin{pmatrix}
\hat{\gamma}(f) & \hat{c}(f) \\
\hat{c}^H(f) & \hat{G}(f)
\end{pmatrix} \triangleq \sum_{m=-N+1}^{N-1} w[m] \hat{R}_x[m] \exp(-j2\pi \theta_f m),
\]

with the empirical ACF

\[
\hat{R}_x[m] \triangleq \begin{cases}
(1/N) \sum_{n=1}^{N-m} x[n + m] x^T[n], & \text{for } m \in \{0, \ldots, N - 1\} \\
\hat{R}_x^H[-m], & \text{for } m \in \{-N + 1, \ldots, -1\}
\end{cases}
\]

2. Estimate the neighborhood \( \mathcal{N}(1) \) via MLASSO based on \( \hat{S}_x(\theta_f) \) instead of \( S_x(\theta_f) \), i.e.,

\[
\hat{\beta}_{\text{LASSO}} \triangleq \arg \min_{\beta'} \left\{ \frac{1}{F} \sum_{f=1}^{F} \left[ \left( \beta'(f) \right)^H \hat{G}(f) \beta'(f) - 2 \Re \{ \hat{c}(f) \beta'(f) \} \right] + \lambda \|\beta'\|_{2,1} \right\}
\]
Outline

- CIG of a stationary time series
- LASSO based graphical model selection for i.i.d. observations
- Multitask-LASSO based graphical model selection for time series
- Selection consistency
- Numerical results
The Incoherence Condition

- Necessary and sufficient condition for \( gsupp(\hat{\beta}_{\text{LASSO}}) = gsupp(\beta) \) is the incoherence condition proved by Bach, 2008.

- For our setting, the incoherence conditions reads

\[
\left\| \left( S_x(\theta_f) \right)_{N(1), N(1)}^{-1} \left( S_x(\theta_f) \right)_{N(1), N(1)} \right\|_\infty < 1
\]

for all \( f \in \{1, \ldots, F\} \).

- This condition limits the correlation between neighbors \( \{x_l\}_{l \in N(1)} \) and non-neighbors \( \{x_l\}_{l \notin N(1)} \).

- Drawback: The incoherence condition (1) is a worst-case requirement: If it fails for a single frequency \( \theta_f \), the plain MLASSO yields a wrong support.
The Multitask Compatibility Condition

- However, we estimate $\mathcal{N}(1)$ by a thresholded LASSO.

- Then, a sufficient condition for $\hat{\mathcal{N}}(1) = \mathcal{N}(1)$ may be based on the multitask compatibility condition introduced by Bühlmann & Van De Geer, 2011.

- In our setting, the multitask compatibility condition is satisfied with constant $\phi_{\text{min}}$ if

$$
\left(\frac{s_{\text{max}}}{F}\right) \sum_{f=1}^{F} (\beta(f))^H S_x(\theta_f) \beta(f) \geq \phi_{\text{min}} \|\beta\|^2_{2,1}
$$

for all sufficiently sparse $\beta$, i.e., $\|\beta_{\mathcal{N}_c(1)}\|_{2,1} \leq 3 \|\beta_{\mathcal{N}(1)}\|_{2,1}$.

- The multitask compatibility condition is an average (over frequency $\theta$) requirement: If $S_x(\theta)$ is “bad” for some $\theta$, we can still be rescued by the behavior of the SDM at other frequencies.
We consider a class of time series described by the following four parameters:

- **Sparsity** $s_{\text{max}}$: The maximum node degree is bounded by a small number $s_{\text{max}} \ll p$, i.e., $|\mathcal{N}(r)| \leq s_{\text{max}}$ for all nodes $r \in \{1, \ldots, p\}$.

- **Minimum partial coherence** $\rho_{\text{min}}$: This parameter quantifies the influence of the neighbors $\{x_l[n]\}_{l \in \mathcal{N}(r)}$ on $x_r[n]$.

- **Multitask compatibility constant** $\phi_{\text{min}}$: We assume that the SDM satisfies the multitask compatibility condition with constant $\phi_{\text{min}}$.

- **Correlation moment** $\mu_{\mathcal{N}}^{(w)}$: This parameter measures the effective correlation width of the time series:

$$
\mu_{\mathcal{N}}^{(w)} \triangleq \sum_{m=-N+1}^{N-1} \| \mathbf{R}_x[m] \|_{\infty} |1 - w[m](1 - |m|/N)|.
$$
Combining a deterministic analysis of MLASSO with a large-deviation characterization of the SDM estimator, one can derive the following result:

**Theorem.** Consider a $p$-dimensional stationary Gaussian time series $x[n]$ with CIG $G_x = (V = \{1, \ldots, p\}, E)$ and SDM $S_x(\theta) \in \mathbb{C}^{p \times p}$. Then, if for some $\delta > 0$, the rescaled sample size $\eta \triangleq N/(\log(p)s_{\text{max}}^3)$ and the correlation moment $\mu^{(w)}_N$ satisfy

$$
\eta > 10^3 \log \left( \frac{4F}{\delta} \right) \|w[.]\|_1^2U^2/\kappa^2 \quad \text{and} \quad \mu^{(w)}_N \leq \frac{\kappa}{2s_{\text{max}}^{3/2}},
$$

with $\kappa \triangleq (\phi_{\text{min}}^2/174)\rho_{\text{min}}\sqrt{LF/U}$, the probability of correctly selecting the edge set $E$ is not smaller than $1 - \delta$, i.e.,

$$
P\left\{ \bigcap_{r \in V} \{\hat{N}(r) = N(r)\} \right\} \geq 1 - \delta.
$$

Based on this result, one can derive sufficient conditions on $N$, $s_{\text{max}}$ and $p$ such that asymptotic consistency is achievable.
Outline

- CIG of a stationary time series
- LASSO based graphical model selection for i.i.d. observations
- Multitask-LASSO based graphical model selection for time series
- Selection consistency
- Numerical results
Simulation Setup

• We randomly generated a symmetric matrix $\Lambda \in \{0, 1\}^{p \times p}$ with zero diagonal and at most $s_{\text{max}} = 3$ non-zeroes in each row and column.

• We generated i.i.d. realizations $z[n]$ of a Gaussian random vector $\mathcal{N}(0, \Sigma)$ with

$$\Sigma = ((s_{\text{max}} + 1)I + \Lambda)^{-1}.$$ 

• The observed time series $x[n]$ is obtained by applying a scalar FIR filter to $z[n]$ such that $R_x[m] = 0$ for all $|m| \geq 2$.

• The CIG $\mathcal{G}_x$ of $x[n]$ has maximum node degree not larger than $s_{\text{max}}$.

• The minimum partial coherence is obtained as $\rho_{\text{min}} = 1/s_{\text{max}}$.

• The multitask compatibility constant is given by $\phi_{\text{min}} = 1/(2s_{\text{max}} + 1)$.
We compared the proposed scheme with an AR-based approach by Bolstad et al., 2011.

We obtained receiver operating characteristic (ROC) curves by varying the MLASSO parameter $\lambda$. 

$\rho = 64, F = 4$
We evaluated the empirical detection probability $P_d$ for fixed MLASSO parameter and varying sample size $N$. 

$P_d$ vs. rescaled sample size

- $p = 64$
- $p = 128$
- $p = 160$
To Summarize ...

- We proposed a compressive nonparametric graphical model selection scheme for stationary time series.

- **Compressive:** Our method works even if we use far fewer samples than the number of time series components.

- **Nonparametric:** We do not rely on a parametric model for the observed time series but only require it to be sufficiently underspread.

- Our approach is **modular:** CS-based post-processing of an SDM estimator. Here: MLASSO and Blackman-Tukey SDM estimator.

- A theoretical performance analysis of our scheme revealed tradeoffs between sample size and model complexity such that consistency is achievable.

- Numerical experiments showed superior performance compared to AR-based scheme in the case of a model mismatch.
Future Directions

- Combine selection scheme with sketching methods for big-data applications.
- Investigate information-theoretic limits on consistent graphical model selection (certificate for optimality of algorithms).
- Extend approach to non-stationary time series.
- Consider directed graphical models (Bayesian networks) and Granger causality.
Thank you!