Sylvester Graphical Models for Complex Spatio-Temporal Processes

Yu (Wayne) Wang
Joint work with Byoungwook Jang and Alfred Hero

University of Michigan

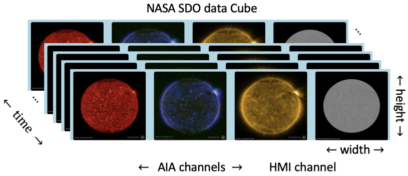
Outline

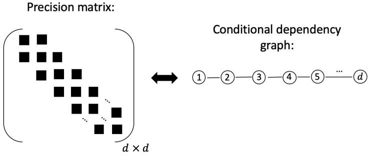
Outline

Ultra-high dimensional multi-indexed heterogeneous data

Soln: a structured multiway graphical model

- SyGlasso/SG-PALM sqrt. Kronecker sum decomp.
  - Generative: $\boldsymbol{AX} + \boldsymbol{XB} = \boldsymbol{Z}$
  - Physical interpretability: directly related to PDEs
  - Robustness, low runtime and high accuracy





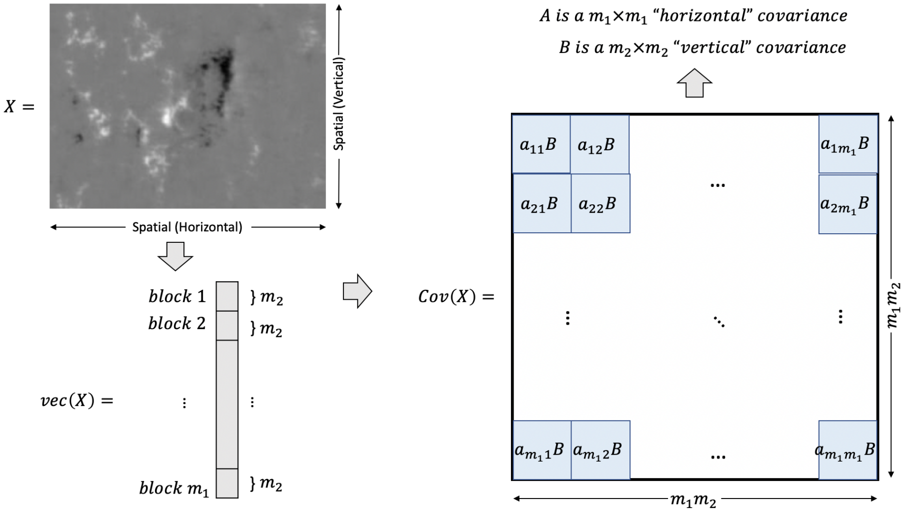
Applications

- Astrophysics: solar flare prediction
- Climatology: spatio-temporal weather forecasting
- Neuroscience: EEG analysis
- Radar imaging: STAP, SAR
- Computer vision: video sequence prediction

Challenges

- High dim. $d = \prod_{k=1}^{K} m_k$ and $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$
- Non-commutative
- Sample-starved learning and high computational complexity

Kronecker product representation for matrix-variate data



Sparse Kronecker-structured models

Sparsity models:

- Sparse covariance ($\mathbf{\Sigma}$) models for marginal dependencies:
  - Examples: M-dependent processes, moving average (MA) processes.
- Sparse **precision** ($\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$) models for conditional dependencies:
  - Examples: Markov random fields, autoregressive (AR) processes.

Kronecker-structured models:

- Kronecker product covariance/inverse covariance
  - Transposable regularized covariance[1]
  - KGlasso[2]
  - GEMINI[3]
- Kronecker sum covariance/inverse covariance
  - Kronecker sum covariance ($K = 2$) for error-in-variable models[4]
  - Bigraphical Lasso (BiGlasso)[5]
  - Tensor-graphical Lasso (TeraLasso)[6]

[1] Genevera I Allen and Robert Tibshirani. "Transposable regularized covariance models with an application to missing data imputation". In: *The Annals of Applied Statistics* 4.2 (2010), p. 764.

[2] Theodoros Tsiligkaridis, Alfred O Hero III, and Shuheng Zhou. "On convergence of kronecker graphical lasso algorithms". In: *IEEE transactions on signal processing* 61.7 (2013), pp. 1743–1755.

[3] Shuheng Zhou. "GEMINI: Graph Estimation with Matrix Variate Normal Instances". In: *The Annals of Statistics* 42.2 (2014), pp. 532–562.

[4] Mark Rudelson and Shuheng Zhou. "Errors-in-variables models with dependent measurements". In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1699–1797.

[5] Alfredo Kalaitzis et al. "The bigraphical lasso". In: *International Conference on Machine Learning*. 2013, pp. 1229–1237.

[6] Kristjan Greenewald, Shuheng Zhou, and Alfred Hero III. "Tensor graphical lasso (TeraLasso)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.5 (2019), pp. 901–931.

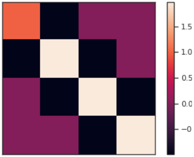
Kronecker sum vs Kronecker product

For $\mathbf{A} \in \mathbb{R}^{m_1 \times m_1}$, $\mathbf{B} \in \mathbb{R}^{m_2 \times m_2}$:

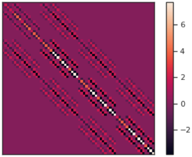
$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1m_1}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m_1 1}\mathbf{B} & \cdots & a_{m_1 m_1}\mathbf{B} \end{bmatrix},$$

and

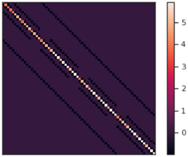
$$\mathbf{A} \oplus \mathbf{B} = \mathbf{I}_{m_2} \otimes \mathbf{A} + \mathbf{B} \otimes \mathbf{I}_{m_1}.$$



$\boldsymbol{\Psi}_k$



KP $\boldsymbol{\Omega} = \bigotimes_{k=1}^{3} \boldsymbol{\Psi}_k$



KS $\boldsymbol{\Omega} = \bigoplus_{k=1}^{3} \boldsymbol{\Psi}_k$

KS vs KP

KP models:

- Separable: $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ and $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^{m_2} (\det \mathbf{B})^{m_1}$.
- Biconvex negative log-likelihood: estimation using alternating graphical lasso, e.g., Tlasso[7].
- Generative representation: $\mathbf{X} = \mathbf{C}^{-1}\mathbf{Z}\mathbf{D}^{-1}$, where $\mathbf{A} = \mathbf{C}\mathbf{C}^T$, $\mathbf{B} = \mathbf{D}\mathbf{D}^T$, and $\mathbf{Z}$ white noise $\Rightarrow \mathrm{Cov}^{-1}(\mathbf{X}) = \mathbf{A} \otimes \mathbf{B}$.

KS models:

- Parsimonious: Cartesian product of graphs avoids explosion of edges[8].
- Convex negative log-likelihood: estimation using ISTA-type procedure, e.g., TeraLasso[9].
- No obvious generative representation.

Motivating question: KP + KS?

[7] Xiang Lyu et al. "Tensor Graphical Model: Non-convex Optimization and Statistical Inference". In: *IEEE transactions on pattern analysis and machine intelligence* (2019).

[8] Alfredo Kalaitzis et al. "The bigraphical lasso". In: *International Conference on Machine Learning*. 2013, pp. 1229–1237.

[9] Kristjan Greenewald, Shuheng Zhou, and Alfred Hero III. "Tensor graphical lasso (TeraLasso)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.5 (2019), pp. 901–931.

Outline

Sylvester Glasso: a generative Kronecker sum model

Let a random tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ be generated by the Sylvester tensor equation[10] [11]:

$$\boldsymbol{\mathcal{X}} \times_1 \boldsymbol{\Psi}_1 + \cdots + \boldsymbol{\mathcal{X}} \times_K \boldsymbol{\Psi}_K = \boldsymbol{\mathcal{T}} \quad \Leftrightarrow \quad (\boldsymbol{\Psi}_1 \oplus \cdots \oplus \boldsymbol{\Psi}_K)\,\mathrm{vec}(\boldsymbol{\mathcal{X}}) = \mathrm{vec}(\boldsymbol{\mathcal{T}}),$$

where $\boldsymbol{\Psi}_k \in \mathbb{R}^{m_k \times m_k}$, $k = 1, \ldots, K$ are sparse matrices, $\boldsymbol{\mathcal{T}}$ is a random tensor of the same dimension as $\boldsymbol{\mathcal{X}}$, and $\times_k$ a k-mode product.

- If $\mathrm{vec}(\boldsymbol{\mathcal{T}}) \sim \mathcal{N}(0, \mathbf{I}_p)$, then $\mathrm{Cov}^{-1}(\mathrm{vec}(\boldsymbol{\mathcal{X}})) = (\boldsymbol{\Psi}_1 \oplus \cdots \oplus \boldsymbol{\Psi}_K)^2$.
- The $\boldsymbol{\Psi}_k$'s can be estimated by minimizing the negative log-pseudolikelihood defined as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\lambda}}(\{\boldsymbol{\Psi}_k\}_{k=1}^K) = &-\frac{N}{2}\log\left|(\mathrm{diag}(\boldsymbol{\Psi}_1) \oplus \cdots \oplus \mathrm{diag}(\boldsymbol{\Psi}_K))^2\right| \\ &+ \frac{N}{2}\mathrm{tr}(\mathbf{S} \cdot (\boldsymbol{\Psi}_1 \oplus \cdots \oplus \boldsymbol{\Psi}_K)^2) + \sum_{k=1}^K \lambda_k \|\boldsymbol{\Psi}_k\|_{1,\mathrm{off}},\end{aligned}$$

  using a proximal alternating linearized minimization method, called SG-PALM.

[10] Lars Grasedyck. "Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure". In: *Computing* 72.3-4 (2004), pp. 247–265.

[11] Daniel Kressner and Christine Tobler. "Krylov subspace methods for linear systems with tensor product structure". In: *SIAM journal on matrix analysis and applications* 31.4 (2010), pp. 1688–1714.

SG-PALM: optimization convergence[12]

Lemma

The function H (the $\log\det$ plus tr terms) is convex and continuously differentiable on an open set containing domG (G the regularization term) and its gradient is block-wise Lipschitz continuous. Further, the objective function $\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi})$ satisfies the Kurdyka - Łojasiewicz (KL) property with a KL exponent of $\frac{1}{2}$.

Theorem

Let $\{\boldsymbol{\Psi}^{(t)}\}_{t\geq 0}$ be generated by SG-PALM and assume that $\boldsymbol{\Psi}^{(t)} \in \Omega \subset dom\partial\mathcal{L}_{\lambda}$. Then, the SG-PALM converges linearly in the sense that

$$\frac{\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}^{(t+1)}) - \min \mathcal{L}_{\boldsymbol{\lambda}}}{\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}^{(t)}) - \min \mathcal{L}_{\boldsymbol{\lambda}}} \leq \left(\frac{\alpha^2 L_{\min}}{4Kc^2(\sum_{j=1}^{K} L_j)^2 + 4c^2 L_{\max}} + 1 \right)^{-1},$$

where $L_{\min} = \min_j L_j$, $L_{\max} = \max_j L_j$, $\alpha > 0$, and $c \in (0,1)$.

[12] Yu Wang and Alfred Hero. "A Proximal Alternating Linearized Minimization Method for Tensor Graphical Models". In: *Submitted* (2020).

SyGlasso: statistical convergence[13]

Assume: sub-Gaussianity on $\boldsymbol{\mathcal{X}}$, bounded eigenvalues of $\boldsymbol{\Omega}$, and incoherence condition on the loss function. Further, let $q_k := |\{(i,j) : (\boldsymbol{\Psi}_k)_{i,j} \neq 0, i \neq j\}|$, and

$$\lambda_k = O\Big(\sqrt{\frac{m_k \log d}{N}}\Big), \quad N > O(\max_k q_k m_k \log d).$$

Theorem (Graph recovery consistency)

There exists a constant $C(\boldsymbol{\beta}) > 0$ [a] such that for any $c_0 > 0$ the following events hold with probability at least $1 - O(\exp(-c_0 \log p))$:

- *(Estimation consistency) Any minimizer $\hat{\boldsymbol{\beta}}$ satisfies:*

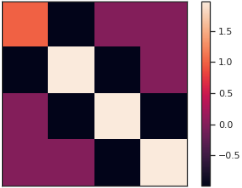
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq C(\boldsymbol{\beta})\sqrt{K} \max_k \sqrt{q_k}\lambda_k.$$

- *(Sign consistency) If minimal signal strength is satisfied for $\boldsymbol{\Psi}_k$ for each k, then sign($\hat{\boldsymbol{\beta}}$)=sign($\boldsymbol{\beta}$).*

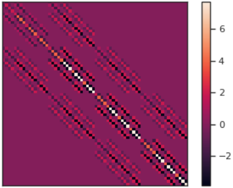
[a] $\boldsymbol{\beta}$ denotes all off-diagonal elements of $\boldsymbol{\Psi}_k$'s.

[13] Yu Wang, Byoungwook Jang, and Alfred Hero. "Sylvester Graphical Lasso (SyGlasso)". In: *Proceedings of The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* (2020).

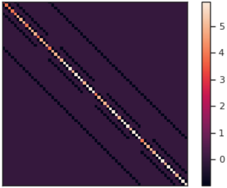
Comparison with KS and KP



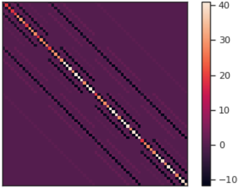
(a) $\mathbf{\Psi}_k$



(b) KP $\mathbf{\Omega}$



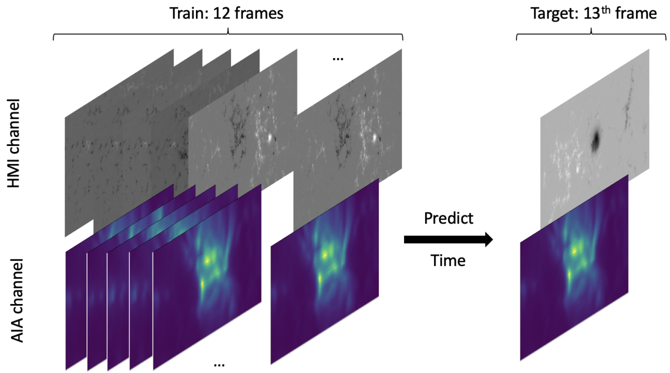
(c) KS $\mathbf{\Omega}$



(d) SyGlasso $\mathbf{\Omega}$

Outline

Multi-instrument solar imaging data



Raw Data: 319 videos before weak or strong flares. Each sequence is $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_{time} \times m_{width} \times m_{height} \times m_{channel}}$, where $m_{time} = 13$ (13-hour window with 1-hour cadence), $m_{width} = 100$, $m_{height} = 50$, and $m_{channel} = 7$ (3 HMI channels and 4 AIA channels).

Goal: Constructing linear forward time series predictors for the last frame (at or right before a flare) by using estimated precision matrix from all previous frames.

Multi-output sparse regression for time series prediction

Consider a spatio-temporal process observed at p time stamps and q locations, i.e., $\boldsymbol{X} \in \mathbb{R}^{p\times q}$. The estimated precision matrix $\boldsymbol{\Omega} \in \mathbb{R}^{pq\times pq}$ can be used to construct an optimal linear predictor, for example,

$$\boldsymbol{y}_t = \boldsymbol{\Omega}_{2,2}^{-1}\boldsymbol{\Omega}_{2,1}\boldsymbol{y}_{t-1:t-(p-1)},$$

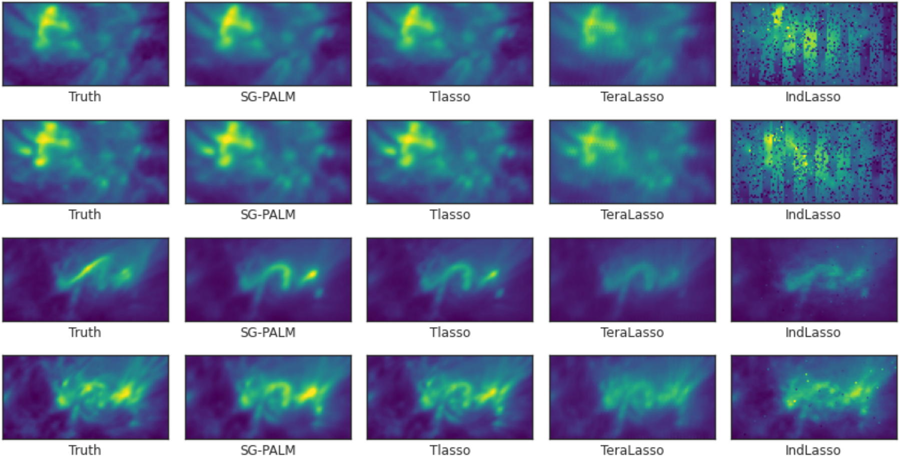
where $\boldsymbol{y}_t \in \mathbb{R}^q$, $\boldsymbol{y}_{t-1:t-(p-1)} \in \mathbb{R}^{(p-1)q}$, and $\boldsymbol{\Omega}_{2,2} \in \mathbb{R}^{q\times q}$, $\boldsymbol{\Omega}_{2,1} \in \mathbb{R}^{q\times(p-1)q}$ are appropriate submatrices of $\boldsymbol{\Omega}$.

For the solar flare data, we estimate

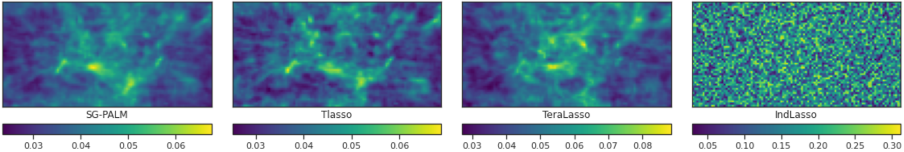
$$\boldsymbol{\Omega} = (\boldsymbol{\Psi}_{time} \oplus \boldsymbol{\Psi}_{height} \oplus \boldsymbol{\Psi}_{width} \oplus \boldsymbol{\Psi}_{channel})^2 \in \mathbb{R}^{455000\times 455000}$$

in training and predict $\boldsymbol{y}_{13} \in \mathbb{R}^{35000}$ in testing using estimated $\boldsymbol{\Omega}$.

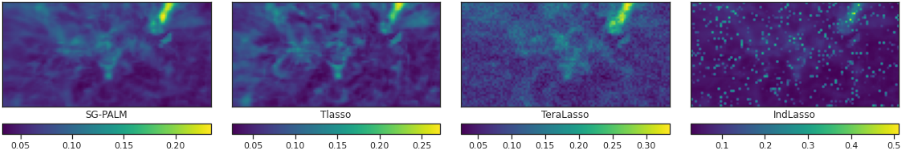
Real vs. predicted images



NRMSE



(a) B-class. Avg. NRMSE (left to right): 0.0379, 0.0386, 0.0579, 0.1628.



(b) MX-class. Avg. NRMSE (left to right): 0.0620, 0.0790, 0.0913, 0.1172.

Outline

Summary

Methodology:

- Kronecker-structured graphical modeling framework inspired by the Sylvester equations in physics.

Theory:

- Convergence (with geometric rate) of the optimization error.
- Convergence of the statistical error.

Applications:

- Multi-modal solar imaging data and flare prediction.

Thank you! Questions?