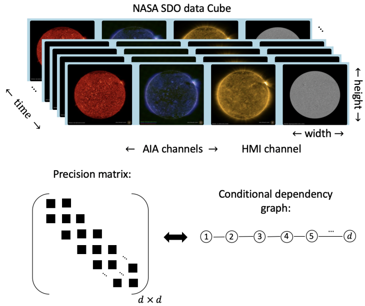
SG-PALM: a Fast Physically Interpretable Tensor Graphical Model

Yu Wang, Alfred Hero

University of Michigan

Overview

For a K-way tensor-valued Gaussian r.v. $\mathcal{\boldsymbol{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, the Sylvester graphical model proposed to model the precision matrix $\boldsymbol{\Omega} = \left(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k\right)^2 \in \mathbb{R}^{d \times d}$, $d = \prod_k d_k$, where $\boldsymbol{\Psi}_k \in \mathbb{R}^{d_k \times d_k}$'s can be obtained via min. of the penalized negative log-pseudolikelihood.



$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}) &= -\frac{N}{2}\log\left|\left(\bigoplus_{k=1}^{K}\operatorname{diag}(\boldsymbol{\Psi}_k)\right)^2\right| + \frac{N}{2}\operatorname{tr}\left(\mathbf{S}\cdot\left(\bigoplus_{k=1}^{K}\boldsymbol{\Psi}_k\right)^2\right) + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k) \\
&:= H(\boldsymbol{\Psi}_1,\ldots,\boldsymbol{\Psi}_K) + \sum_{k=1}^{K} G_k(\boldsymbol{\Psi}_k).
\end{aligned}
\tag{1}
$$

Optimization

$H(\cdot)$ has block-wise Lipschitz gradients and $G(\cdot)$ is proximal friendly $\Rightarrow$ a Proximal Alternating Linearized Minimizing (PALM) algorithm:

Algorithm 1 SG-PALM

Require: $\boldsymbol{\mathcal{X}}$, $\lambda_k > 0$, $c \in (0,1)$, $\eta_0 > 0$, initial iterates $\{\boldsymbol{\Psi}_k\}_{k=1}^K$.

  while not converged **do**

    for $k = 1, \ldots, K$ **do**

      Line search: Let η_k^t be the largest element of $\{c^j \eta_{k,0}^t\}_{j=1,\ldots}$ such that a sufficient descent condition is satisfied.

      Update: $\boldsymbol{\Psi}_k^{t+1} \leftarrow \mathrm{prox}_{G_k}^{\eta_k^t \lambda_k}\Big(\boldsymbol{\Psi}_k^t - \eta_k^t \nabla_k H(\boldsymbol{\Psi}_{i<k}^{t+1}, \boldsymbol{\Psi}_{i\geq k}^t)\Big)$.

    end for

    Update initial step size: Compute $\eta_0^{t+1} = \min_k \eta_{k,0}^{t+1}$, where $\eta_{k,0}^{t+1}$ is computed via the Barzilai-Borwein strategy.

  end while

Ensure: Final iterates $\{\boldsymbol{\Psi}_k\}_{k=1}^K$.

Iterative convergence

Pros:

- $O\Big(\sum_{k=1}^{K}(s_k d_k^2 + N\sum_{j\neq k} s_j d_j^2)\Big)$ operations per iteration $\Rightarrow$ lower than competing methods for similar models when $N \ll d$ and $s_k \ll d_k$.
- No matrix inversion/factorization & expensive storage $\Rightarrow$ comm.-efficient parallelism.
- Fast convergence:

Theorem (For convex objective♯)

The sequence $\{\mathbf{\Psi}^{(t)}\}_{t\geq 0}$ *generated by SG-PALM converges linearly in the sense that*

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

where $L_{\min} = \min_j L_j > 0$, $L_{\max} = \max_j L_j > 0$, $\alpha > 0$, *and* $c \in (0,1)$. ♯ *Nonconvex extensions available in the paper.*

Statistical convergence

Theorem (For ℓ_1-penalty functions)

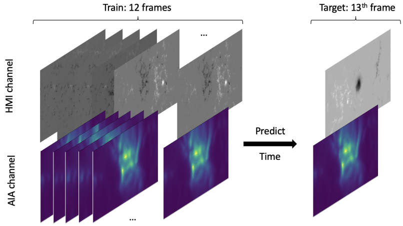
Let $\mathcal{A}_k := \{(i,j) : (\bar{\mathbf{\Psi}}_k)_{i,j} \neq 0, i \neq j\}$ *and* $q_k := |\mathcal{A}_k|$ *for* $k = 1, \ldots, K$. *If* $\lambda_k = O(\sqrt{\frac{d_k \log d}{N}})$ *for all* $k = 1, \ldots, K$*, then under regularity conditions specified in the paper,* $\exists C > 0$ *such that* $\forall \eta > 0$ *the following holds with probability at least* $1 - O(\exp(-\eta \log d))$*:*

$$\sum_{k=1}^{K} \|\mathit{offdiag}(\hat{\mathbf{\Psi}}_k) - \mathit{offdiag}(\bar{\mathbf{\Psi}}_k)\|_F \leq C\sqrt{K} \max_k \sqrt{q_k}\lambda_k. \tag{3}$$

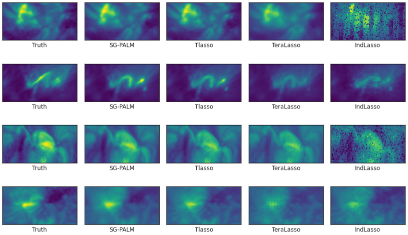
Application to solar flare prediction

Construct linear forward predictors for the last frame (at or right before a flare) by using estimated precision matrix from all previous frames, i.e., $\hat{\boldsymbol{\mathcal{X}}}_{t,:,:,:} = \hat{\boldsymbol{\Omega}}_{2,2}^{-1}\hat{\boldsymbol{\Omega}}_{2,1}\boldsymbol{\mathcal{X}}_{t-1:t-(p-1),:,:,:}$, where $q = d_{width} \cdot d_{height} \cdot d_{channel}$ and $p = d_{time}$, $\hat{\boldsymbol{\Omega}}_{2,2} \in \mathbb{R}^{q\times q}$, $\hat{\boldsymbol{\Omega}}_{2,1} \in \mathbb{R}^{q\times(p-1)q}$ are submatrices of $\hat{\boldsymbol{\Omega}}$.

Task illustration



Predicted examples



Physical interpretation

Consider the 2D spatio-temporal process $u(\mathbf{x}, t)$:

$$\partial u/\partial t = \theta \sum_{i=1}^{2} \partial^2 u/\partial x_i^2 + \epsilon \sum_{i=1}^{2} \partial u/\partial x_i, \tag{4}$$

where θ, ϵ are positive real (unknown) coefficients. This is the basic form of a class of parabolic and hyperbolic PDEs, the Convection-Diffusion equation.

After finite-difference discretization, Equation (4) is equivalent to the Sylvester matrix equation

$$\mathbf{A}_{\theta,\epsilon}\mathbf{U}_t + \mathbf{U}_t\mathbf{A}_{\theta,\epsilon} = \mathbf{U}_{t-1}, \tag{5}$$

where $\mathbf{U}_t = (u((i,j),t))_{ij}$ and $\mathbf{A}_{\theta,\epsilon}$ is a tridiagonal matrix with values that depend on the coefficients θ, ϵ and discretization step sizes. This is the same Sylvester equation used for defining the objective function of our graphical model!