Appendix

Section A provides detailed derivation of the log-pseudolikelihood function.

Section B provides justifications for the Barzilai-Borwein step sizes implemented in Algorithm 1.

Section C provides detailed proofs of Theorems 4.1 and 4.2.

Section D discusses extensions of Algorithm 1 and its convergence properties to non-convex cases.

Section E provides additional details of the solar flare experiments.

A. Derivation of the Log-Pseudolikelihood

By rewriting the Sylvester tensor equation defined in (2) element-wise, we first observe that

$$\begin{aligned}&\left(\sum_{k=1}^{K}(\mathbf{\Psi}_k)_{i_k,i_k}\right)\boldsymbol{\mathcal{X}}_{i_{[1:K]}}\\&=-\sum_{k=1}^{K}\sum_{j_k\neq i_k}(\mathbf{\Psi}_k)_{i_k,j_k}\boldsymbol{\mathcal{X}}_{i_{[1:k]},j_k,i_{[k+1:K]}}+\boldsymbol{\mathcal{T}}_{i_{[1:K]}}.\end{aligned} \tag{12}$$

Note that the left-hand side of (12) involves only the summation of the diagonals of the $\mathbf{\Psi}_k$'s and the right-hand side is composed of columns of $\mathbf{\Psi}_k$'s that exclude the diagonal terms. Equation (12) can be interpreted as an autogregressive model relating the $(i_1,\dots,i_K)$-th element of the data tensor (scaled by the sum of diagonals) to other elements in the fibers of the data tensor. The columns of $\mathbf{\Psi}_k$'s act as regression coefficients. The formulation in (12) naturally leads to a pseudolikelihood-based estimation procedure (Besag, 1977) for estimating $\mathbf{\Omega}$ (see also Khare et al. (2015) for how this procedure applied to vector-variate Gaussian graphical model estimation). It is known that inference using pseudo-likelihood is consistent and enjoys the same $\sqrt{N}$ convergence rate as the MLE in general (Varin et al., 2011). This procedure can also be more robust to model misspecification (e.g., non-Gaussianity) in the sense that it assumes *only that the sub-models/conditional distributions (i.e., $\boldsymbol{\mathcal{X}}_i|\boldsymbol{\mathcal{X}}_{-i}$) are Gaussian*. Therefore, in practice, even if the data is not Gaussian, the Maximum Pseudolikelihood Estimation procedure is able to perform reasonably well. Wang et al. (2020) also studied a different model misspecification scenario where the Kronecker product/sum and Sylvester structures are mismatched for SyGlasso.

From (12) we can define the sparse least-squares estimators for $\mathbf{\Psi}_k$'s as the solution of the following convex optimization problem:

$$\begin{aligned}\min_{\substack{\mathbf{\Psi}_k\in\mathbb{R}^{d_k\times d_k}\\k=1,\dots K}}\; & -N\sum_{i_1,\dots,i_K}\log\boldsymbol{\mathcal{W}}_{i_{[1:K]}}\\&+\frac{1}{2}\sum_{i_1,\dots,i_K}\|(I)+(II)\|_2^2+\sum_{k=1}^{K}P_{\lambda_k}(\mathbf{\Psi}_k).\end{aligned}$$

where $P_{\lambda_k}(\cdot)$ is a penalty function indexed by the tuning parameter λ_k and

$$\begin{aligned}(I)&=\boldsymbol{\mathcal{W}}_{i_{[1:K]}}\boldsymbol{\mathcal{X}}_{i_{[1:K]}}\\(II)&=\sum_{k=1}^{K}\sum_{j_k\neq i_k}(\mathbf{\Psi}_k)_{i_k,j_k}\boldsymbol{\mathcal{X}}_{i_{[1:k]},j_k,i_{[k+1:K]}},\end{aligned}$$

with $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}:=\sum_{k=1}^{K}(\mathbf{\Psi}_k)_{i_k,i_k}$.

The optimization problem above can be put into the following matrix form:

$$\begin{aligned}\min_{\substack{\mathbf{\Psi}_k\in\mathbb{R}^{d_k\times d_k}\\k=1,\dots K}}\; & -\frac{N}{2}\log|(\mathrm{diag}(\mathbf{\Psi}_1)\oplus\cdots\oplus\mathrm{diag}(\mathbf{\Psi}_K))^2|\\&+\frac{N}{2}\mathrm{tr}(\mathbf{S}(\mathbf{\Psi}_1\oplus\cdots\oplus\mathbf{\Psi}_K)^2)+\sum_{k=1}^{K}P_{\lambda_k}(\mathbf{\Psi}_k)\end{aligned}$$

where $\mathbf{S} \in \mathbb{R}^{d\times d}$ is the sample covariance matrix, i.e., $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \operatorname{vec}(\boldsymbol{\mathcal{X}}^i)\operatorname{vec}(\boldsymbol{\mathcal{X}}^i)^T$. Note that this is equivalent to the negative log-pseudolikelihood function that approximates the ℓ_1-penalized Gaussian negative log-likelihood in the log-determinant term by including only the Kronecker sum of the diagonal matrices instead of the Kronecker sum of the full matrices.

B. The Barzilai-Borwein Step Size

The BB method has been proven to be very successful in solving nonlinear optimization problems. In this section we outline the key ideas behind the BB method, which is motivated by quasi-Newton methods. Suppose we want to solve the unconstrained minimization problem

$$\min_x f(x),$$

where f is differentiable. A typical iteration of quasi-Newton methods for solving this problem is

$$x_{t+1} = x_t - B_t^{-1}\nabla f(x_t),$$

where B_t is an approximation of the Hessian matrix of f at the current iterate x_t. Here, B_t must satisfy the so-called secant equation: $B_t s_t = y_t$, where $s_t = x_t - x_{t-1}$ and $y_t = \nabla f(x_t) - \nabla f(x_{t-1})$ for $t \geq 1$. It is noted that in to get B_t^{-1} one needs to solve a linear system, which may be computationally expensive when B_t is large and dense.

One way to alleviate this burden is to use the BB method, which replaces B_t by a scalar matrix $(1/\eta_t)\mathbf{I}$. However, it is hard to choose a scalar η_t such that the secant equation holds with $B_t = (1/\eta_t)\mathbf{I}$. Instead, one can find η_t such that the residual of the secant equation, i.e., $\|(1/\eta_t)s_t - y_t\|_2^2$, is minimized, which leads to the following choice of η_t:

$$\eta_t = \frac{\|s_t\|_2^2}{s_t^T y_t}.$$

Therefore, a typical iteration of the BB method for solving the original problem is

$$x_{t+1} = x_t - \eta_t \nabla f(x_t),$$

where η_t is computed via the previous formula.

For convergence analysis, generalizations and variants of the BB method, we refer the interested readers to Raydan (1993; 1997); Dai & Liao (2002); Fletcher (2005) and references therein. BB method has been successfully applied for solving problems arising from emerging applications, such as compressed sensing (Wright et al., 2009), sparse reconstruction (Wen et al., 2010) and image processing (Wang & Ma, 2007).

C. Proofs of Theorems

C.1. Proof of Theorem 4.1

We first state the regularity conditions needed for establishing convergence of the SG-PALM estimators $\{\hat{\boldsymbol{\Psi}}_k\}_{k=1}^K$ to their true value $\{\bar{\boldsymbol{\Psi}}_k\}_{k=1}^K$.

(A1 - Subgaussianity) The data $\boldsymbol{\mathcal{X}}^1, \dots, \boldsymbol{\mathcal{X}}^N$ are i.i.d subgaussian random tensors, that is, $\operatorname{vec}(\boldsymbol{\mathcal{X}}^i) \sim \mathbf{x}$, where $\mathbf{x}$ is a subgaussian random vector in $\mathbb{R}^d$, i.e., there exist a constant $c > 0$, such that for every $\mathbf{a} \in \mathbb{R}^d$, $\mathbb{E}e^{\mathbf{a}^T x} \leq e^{c\mathbf{a}^T \bar{\boldsymbol{\Sigma}}\mathbf{a}}$, and there exist $\rho_j > 0$ such that $\mathbb{E}e^{tx_j^2} \leq +\infty$ whenever $|t| < \rho_j$, for $1 \leq j \leq d$.

(A2 - Bounded eigenvalues) There exist constants $0 < \Lambda_{\min} \leq \Lambda_{\max} < \infty$, such that the minimum and maximum eigenvalues of $\boldsymbol{\Omega}$ are bounded with $\lambda_{\min}(\bar{\boldsymbol{\Omega}}) = (\sum_{k=1}^K \lambda_{\max}(\boldsymbol{\Psi}_k))^{-2} \geq \Lambda_{\min}$ and $\lambda_{\max}(\bar{\boldsymbol{\Omega}}) = (\sum_{k=1}^K \lambda_{\min}(\boldsymbol{\Psi}_k))^{-2} \leq \Lambda_{\max}$.

(A3 - Incoherence condition) There exists a constant $\delta < 1$ such that for $k = 1, \dots, K$ and all $(i,j) \in \mathcal{A}_k$

$$|\bar{\mathcal{L}}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\Psi}})[\bar{\mathcal{L}}''_{\mathcal{A}_k,\mathcal{A}_k}(\bar{\boldsymbol{\Psi}})]^{-1}\text{sign}(\bar{\boldsymbol{\Psi}}_{\mathcal{A}_k,\mathcal{A}_k})| \leq \delta,$$

where for each k and $1 \leq i < j \leq d_k$, $1 \leq k < l \leq d_k$,

$$\bar{\mathcal{L}}''_{ij,kl}(\bar{\boldsymbol{\Psi}}) := E_{\bar{\boldsymbol{\Psi}}}\left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\Psi})}{\partial(\boldsymbol{\Psi}_k)_{i,j}\partial(\boldsymbol{\Psi}_k)_{k,l}}|_{\boldsymbol{\Psi}=\bar{\boldsymbol{\Psi}}}\right),$$

and

$$\mathcal{L}(\mathbf{\Psi}) = -\frac{N}{2}\log|(\bigoplus_{k=1}^{K}\text{diag}(\mathbf{\Psi}_k))^2| + \frac{N}{2}\,\text{tr}(\mathbf{S}\cdot(\bigoplus_{k=1}^{K}\mathbf{\Psi}_k)^2).$$

Given assumptions (A1-A3), the theorem follows from Theorem 3.3 in Wang et al. (2020).

C.2. Proof of Theorem 4.2

We next turn to convergence of the iterates $\{\mathbf{\Psi}^t\}$ from SG-PALM to a global optimum of (3). The proof leverages recent results in the convergence of alternating minimization algorithms for non-strongly convex objective (Bolte et al., 2014; Karimi et al., 2016; Li & Pong, 2018; Zhang, 2020). We outline the proof strategy:

1. We establish Lipschitz continuity of the blockwise gradient $\nabla_k H(\mathbf{\Psi})$ for $k = 1, \ldots, K$.
2. We show that the objective function $\mathcal{L}_\lambda$ satisfies the Kurdyka - Łojasiewicz (KL) property. Further, it has a KL exponent of $\frac{1}{2}$ (defined later in the proofs).
3. The KL property (with exponent $\frac{1}{2}$) is equivalent to a generalized Error Bound (EB) condition, which enables us to establish linear iterative convergence of the objective function (3) to its global optimum.

Definition C.1 (Subdifferentials). *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Its domain is defined by*

$$dom f := \{x \in \mathbb{R}^d : f(x) < +\infty\}.$$

If we further assume that f is convex, then the subdifferential of f at $x \in dom f$ can be defined by

$$\partial f(x) := \{v \in \mathbb{R}^d : f(z) \geq f(x) + < v, z - x >, \forall z \in \mathbb{R}^d\}.$$

The elements of $\partial f(x)$ are called subgradients of f at x.

Denote the domain of ∂f by $\text{dom}\partial f := \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\}$. Then, if f is proper, semicontinuous, convex, and $x \in \text{dom} f$, then $\partial f(x)$ is a nonempty closed convex set. In this case, we denote by $\partial^0 f(x)$ the unique least-norm element of $\partial f(x)$ for $x \in \text{dom}\partial f$, along with $\|\partial^0 f(x)\| = +\infty$ for $x \notin \text{dom}\partial f$. Points whose subdifferential contains 0 are critical points, denoted by **crit**f. For convex f, **crit**$f = \text{argmin}\, f$.

Definition C.2 (KL property). *Let Γ_{c_2} stands for the class of functions $\phi : [0, c_2] \to \mathbb{R}_+$ for $c_2 > 0$ with the properties:*

i. *ϕ is continuous on $[0, c_2]$;*

ii. *ϕ is smooth concave on $(0, c_2)$;*

iii. $\phi(0) = 0, \phi'(s) > 0, \forall s \in (0, c_2)$.

Further, for $x \in \mathbb{R}^d$ and any nonempty $Q \subset \mathbb{R}^d$, define the distance function $d(x, Q) := \inf_{y \in Q}\|x - y\|$. Then, a function f is said to have the Kurdyka - Łojasiewicz (KL) property at point x_0, if there exist $c_1 > 0$, a neighborhood B of x_0, and $\phi \in \Gamma_{c_2}$ such that for all

$$x \in B(x_0, c_1) \cap \{x : f(x_0) < f(x) < f(x_0) + c_2\},$$

the following inequality holds

$$\phi'\Big(f(x) - f(x_0)\Big)dist(0, \partial f(x)) \geq 1.$$

If f satisfies the KL property at each point of $dom\partial f$ then f is called a KL function.

We first present two lemmas that characterize key properties of the loss function.

Lemma C.1 (Blockwise Lipschitzness). *The function H is convex and continuously differentiable on an open set containing $domG$ and its gradient, is block-wise Lipschitz continuous with block Lipschitz constant $L_k > 0$ for each k, namely for all $k = 1, \ldots, K$ and all $\mathbf{\Psi}_k, \mathbf{\Psi}'_k \in \mathbb{R}^{d_k \times d_k}$*

$$\begin{aligned}&\|\nabla_k H(\mathbf{\Psi}_{i<k}, \mathbf{\Psi}_k, \mathbf{\Psi}_{i>k}) - \nabla_k H(\mathbf{\Psi}_{i<k}, \mathbf{\Psi}'_k, \mathbf{\Psi}_{i>k})\|\\ &\leq L_k\|\mathbf{\Psi}_k - \mathbf{\Psi}'_k\|,\end{aligned}$$

where $\nabla_k H$ denotes the gradient of H with respect to $\mathbf{\Psi}_k$ with all remaining $\mathbf{\Psi}_i$, $i \neq k$ fixed. Further, the function G_k for each $k = 1, \dots, K$ is a proper lower semicontinuous (lsc) convex function.

Proof. For simplicity of notation, in this and the following proofs we use $\mathbf{\Psi}$ (i.e., omitting the subscript) to denote the set $\{\mathbf{\Psi}_k\}_{k=1}^K$ or the K-tuple $(\mathbf{\Psi}_1, \dots, \mathbf{\Psi}_K)$ whenever there is no confusion. Recall the blockwise gradient of the smooth part of the objective function H with respect to $\mathbf{\Psi}_k$, for each $k = 1, \dots, K$, is given by

$$\begin{aligned}\nabla_k H(\mathbf{\Psi}) = \text{diag}\Big(\Big[\text{tr}\{(\text{diag}((\mathbf{\Psi}_k))_{ii} + \bigoplus_{j\neq k}\text{diag}(\mathbf{\Psi}_j))^{-1}\} \quad i = 1 : d_k\Big]\Big)\\ + \mathbf{S}_k\mathbf{\Psi}_k + \mathbf{\Psi}_k\mathbf{S}_k + 2\sum_{j\neq k}\mathbf{S}_{j,k}.\end{aligned}$$

Then for $\mathbf{\Psi}_k$, $\mathbf{\Psi}'_k$,

$$\begin{aligned}&\|\mathbf{S}_k\mathbf{\Psi}_k + \mathbf{\Psi}_k\mathbf{S}_k + 2\sum_{j\neq k}\mathbf{S}_{j,k} - (\mathbf{S}_k\mathbf{\Psi}'_k + \mathbf{\Psi}'_k\mathbf{S}_k + 2\sum_{j\neq k}\mathbf{S}_{j,k})\|\\ &= \|\mathbf{S}_k\mathbf{\Psi}_k + \mathbf{\Psi}_k\mathbf{S}_k - \mathbf{S}_k\mathbf{\Psi}'_k - \mathbf{\Psi}'_k\mathbf{S}_k\|\\ &\leq 2\|\mathbf{S}_k\|\|\mathbf{\Psi}_k - \mathbf{\Psi}'_k\|.\end{aligned}$$

To prove Lipschitzness of the remaining parts, we consider the case of $K = 2$ for simplicity of notations. The arguments easily carry over cases of $K > 2$. In this case, denote $\mathbf{A} = (a_{ij}) := \mathbf{\Psi}_1$ and $\mathbf{B} = (b_{kl}) := \mathbf{\Psi}_2$. Let $f(\mathbf{A}) := \frac{\partial}{\partial\mathbf{A}}\log|\text{diag}(\mathbf{A}\oplus\mathbf{B})|$, then

$$f(\mathbf{A}) - f(\mathbf{A}') = \text{diag}\Big(\Big[\sum_{i=1}^{m_2}(a_{jj} + b_{ii})^{-1} - \sum_{i=1}^{m_2}(a'_{jj} + b_{ii})^{-1} \quad j = 1, \dots, m_1\Big]\Big)$$

and

$$\begin{aligned}\|f(\mathbf{A}) - f(\mathbf{A}')\|_F &= \left(\sum_{j=1}^{m_1}\Big(\sum_{i=1}^{m_2}(a_{jj} + b_{ii})^{-1} - \sum_{i=1}^{m_2}(a'_{jj} + b_{ii})^{-1}\Big)^2\right)^{1/2}\\ &\leq \left(\sum_{j=1}^{m_1} m_2\sum_{i=1}^{m_2}\Big((a_{jj} + b_{ii})^{-1} - (a'_{jj} + b_{ii})^{-1}\Big)^2\right)^{1/2}\\ &= \Big(m_2\sum_{j=1}^{m_1}\sum_{i=1}^{m_2}(c_{ji}^{-1} - (c'_{ji})^{-1})^2\Big)^{1/2}\\ &= \Big(m_2\sum_{j=1}^{m_1}\sum_{i=1}^{m_2}(c'_{ji})^{-2}(c'_{ji} - c_{ji})^2 c_{ji}^{-2}\Big)^{1/2}\\ &= \Big(m_2\sum_{j=1}^{m_1}(a_{jj} - a'_{jj})^2\sum_{i=1}^{m_2}(c'_{ji}c_{ji})^{-2}\Big)^{1/2}\\ &\leq \Big(Cm_2\sum_{j=1}^{m_1}\sum_{i=1}^{m_2}(c'_{ji}c_{ji})^{-2}\Big)^{1/2}\|\mathbf{A} - \mathbf{A}'\|_F,\end{aligned}$$

where the first inequality is due to Cauchy-Schwartz inequality; the third line is due to $c_{ji} := a_{jj} + b_{ii}$; and in the last inequality we upper-bound each $(a_{jj} - a'_{jj})^2$ by its maximum over all j, which is absorbed in a constant C. Note that the first term in the last line above is finite as long as the summations of the diagonal elements of the factors $\mathbf{A}$ and $\mathbf{B}$ are finite, which is implied if the precision matrix $\mathbf{\Omega}$ defined by the Sylvester generating equation as $(\mathbf{A}\oplus\mathbf{B})^2$ has finite diagonal elements. This follows from Theorem 3.1 of Oh et al. (2014), who proved that if a symmetric matrix $\mathbf{\Omega}$ satisfying $\mathbf{\Omega} \in \mathcal{C}_0$, where

$$\mathcal{C}_0 = \Big\{\mathbf{\Omega}|\mathcal{L}_\lambda(\mathbf{\Omega}) \leq \mathcal{L}_\lambda(\mathbf{\Omega}^{(0)}) = M\Big\},$$

and $\mathbf{\Omega}^{(0)}$ is an arbitrary initial point with a finite function value $\mathcal{L}_\lambda(\mathbf{\Omega}^{(0)}) := M$, the diagonal elements of $\mathbf{\Omega}$ are bounded above and below by constants which depend only on M, the regularization parameter λ, and the sample covariance matrix

$\mathbf{S}$. Therefore, we have

$$\|f(\mathbf{A}) - f(\mathbf{A}')\|_F \leq \tilde{C}\|\mathbf{A} - \mathbf{A}'\|_F$$

for some constant $\tilde{C} \in (0, +\infty)$. Similarly, we can establish such an inequality for $\mathbf{B}$, proving that the first term in $\nabla_k H$ is Lipschitz continuous. □

As a consequence of Lemma C.1, the gradient of H, defined by $\nabla H = (\nabla_1 H, \dots, \nabla_K H)$ is Lipschitz continuous on bounded subsets $\mathbb{B}_1 \times \cdots \times \mathbb{B}_K$ of $\mathbb{R}^{d_1 \times d_1} \times \cdots \times \mathbb{R}^{d_K \times d_K}$ with some constant $L > 0$, such that for all $(\mathbf{\Psi}_k, \mathbf{\Psi}'_k) \in \mathbb{B}_k \times \mathbb{B}_k$,

$$\begin{aligned}&\|(\nabla_1 H(\mathbf{\Psi}_1, \mathbf{\Psi}_{i>1}) - \nabla_1 H(\mathbf{\Psi}'_1, \mathbf{\Psi}'_{i>1}), \dots,\\ &\nabla_K H(\mathbf{\Psi}'_{i<K}, \mathbf{\Psi}'_K) - \nabla_K H(\mathbf{\Psi}'_{i<K}, \mathbf{\Psi}'_K))\|\\ &\leq L\|(\mathbf{\Psi}_1 - \mathbf{\Psi}'_1, \dots, \mathbf{\Psi}_K - \mathbf{\Psi}'_K)\|,\end{aligned}$$

and we have $L \leq \sum_{k=1}^K L_k$.

Lemma C.2 (KL property of $\mathcal{L}_\lambda$)**.** *The objective function $\mathcal{L}_\lambda(\mathbf{\Psi})$ defined in (3) satisfies the KL property. Further, ϕ in this case can be chosen to have the form $\phi(s) = \alpha s^{1/2}$, where α is some positive real number. Functions satisfying the KL property with this particular choice of ϕ is said to have a KL exponent of $\frac{1}{2}$.*

Proof. This can be established in a few steps:

1. It can be shown that the function (of $\mathbf{X}$) $\operatorname{tr}(\mathbf{S}\mathbf{X}^2) + \|\mathbf{X}\|_{1,\text{off}}$ satisfies the KL property with exponent $\frac{1}{2}$ (Karimi et al., 2016). We then apply the calculus rules of the KL exponent (compositions and separable summations) studied in Li & Pong (2018) to prove that $\operatorname{tr}(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2)$ and $\sum_j \|\mathbf{\Psi}_j\|_{1,\text{off}}$ are also KL functions with exponent $\frac{1}{2}$.

2. The $-\log\det\left(\bigoplus_j \operatorname{diag}(\mathbf{\Psi}_j)\right)$ term can be shown to be KL with exponent $\frac{1}{2}$ using a transfer principle studied in Lourenço & Takeda (2019).

3. Finally, using the calculus rules of KL exponent one more time, we combine the first two results and establish that $\mathcal{L}_\lambda$ has KL exponent of $\frac{1}{2}$.

Karimi et al. (2016) proved that the following function, parameterized by some symmetric matrix $\mathbf{X}$, satisfies the KL property with KL exponent $\frac{1}{2}$:

$$\operatorname{tr}(\mathbf{S}\mathbf{X}^2) + \|\mathbf{X}\|_{1,\text{off}} = \|\mathbf{A}\mathbf{X}\|_F^2 + \|\mathbf{X}\|_{1,\text{off}}$$

for $\mathbf{S} = \mathbf{A}\mathbf{A}^T$, even when $\mathbf{A}$ is not of full rank.

We apply the calculus rules of the KL exponent studied in Li & Pong (2018) to prove that $\operatorname{tr}(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2)$ and $\sum_j \|\mathbf{\Psi}_j\|_{1,\text{off}}$ are KL functions with exponent $\frac{1}{2}$. Particularly, we observe that $\operatorname{tr}\left(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2\right)$ is the composition of functions $\mathbf{X} \to \operatorname{tr}(\mathbf{S}\mathbf{X})$ and $(\mathbf{X}_1, \dots, \mathbf{X}_K) \to \bigoplus_j \mathbf{X}_j$; and $\sum_j \|\mathbf{\Psi}_j\|_{1,\text{off}}$ is a separable block summation of functions $\mathbf{X}_j \to \|\mathbf{X}_j\|_{1,\text{off}}$.

Thus, by Theorem 3.2. (exponent for composition of KL functions) in Li & Pong (2018), since the Kronecker sum operation is linear and hence continuously differentiable, the trace function is KL with exponent $\frac{1}{2}$, and the mapping $(\mathbf{X}_1, \dots, \mathbf{X}_K) \to \bigoplus_j \mathbf{X}_j$ is clearly one to one, the function $\operatorname{tr}(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2)$ has the KL exponent of $\frac{1}{2}$. By Theorem 3.3. (exponent for block separable sums of KL functions) in Li & Pong (2018), since the function $\|\cdot\|_{1,\text{off}}$ is proper, closed, continuous on its domain, and is KL with exponent $\frac{1}{2}$, the function $\|\mathbf{X}_j\|_{1,\text{off}}$ is KL with an exponent of $\frac{1}{2}$.

It remains to prove that $-\log\det\left(\bigoplus_j \operatorname{diag}(\mathbf{\Psi}_j)\right)$ is also a KL function with an exponent of $\frac{1}{2}$. By Theorem 30 in Lourenço & Takeda (2019), if we have $f : \mathbb{R}^r \to \mathbb{R}$ a symmetric function and $F : \mathcal{E} \to \mathbb{R}$ the corresponding spectral function, the followings hold

(i). F satisfies the KL property at $\mathbf{X}$ iff f satisfies the KL property at $\lambda(\mathbf{X})$, i.e., the eigenvalues of $\mathbf{X}$.

(ii). F satisfies the KL property with exponent α iff f satisfies the KL property with exponent α at $\lambda(\mathbf{X})$.

Here, take $f(\lambda(\mathbf{X})) := -\sum_{i=1}^{r}\log(\lambda_i(\mathbf{X}))$, and $F(\mathbf{X}) := -\log\det(\mathbf{X})$ the corresponding spectral function. Then, the function f is symmetric since its value is invariant to permutations of its arguments, and it is a strictly convex function in its domain, so it satisfies the KL property with an exponent of $\frac{1}{2}$. Therefore, F satisfies the KL property with the same KL exponent of $\frac{1}{2}$. Now, we apply the calculus rules for KL functions again. As both the Kronecker sum and the diag operators are linear, we conclude that $-\log\det\left(\bigoplus_j \operatorname{diag}(\mathbf{\Psi}_j)\right)$ is a KL function with an exponent of $\frac{1}{2}$.

Overall, we have that the negative log-pseudolikelihood function $\mathcal{L}(\mathbf{\Psi})$ satisfies the KL property with an exponent of $\frac{1}{2}$. □

Now we are ready to prove Theorem 4.2. We follow Zhang (2020) and divide the proof into three steps.

Step 1. We obtain a sufficient decrease property for the loss function $\mathcal{L}$ in terms of the squared distance of two successive iterates:

$$\mathcal{L}(\mathbf{\Psi}^{(t)}) - \mathcal{L}(\mathbf{\Psi}^{(t+1)}) \geq \frac{L_{\min}}{2}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2. \tag{13}$$

Here and below, $\mathbf{\Psi}^{(t+1)} := (\mathbf{\Psi}_1^{(t+1)}, \ldots, \mathbf{\Psi}_K^{(t+1)})$ and $L_{\min} := \min_k L_k$. First note that at iteration t, the line search condition is satisfied for step size $\frac{1}{\eta_k^{(t)}} \geq L_k$, where L_k is the Lipschitz constant for $\nabla_k H$. Further, it follows that for SG-PALM with backtracking one has for every $t \geq 0$ and each $k = 1, \ldots, K$,

$$\frac{1}{\eta_k^{(0)}} \leq \frac{1}{\eta_k^{(t)}} \leq cL_k,$$

where $c > 0$ is the backtracking constant.

Then by Lemma 3.1 in Shefi & Teboulle (2016), we get

$$\begin{aligned}\mathcal{L}(\mathbf{\Psi}^{(t)}) - \mathcal{L}(\mathbf{\Psi}^{(t+1)}) &\geq \frac{1}{2\eta_{\min}^{(t+1)}}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2 \\ &\geq \frac{L_{\min}}{2}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2\end{aligned}$$

for $\eta_{\min}^{(t)} := \min_k \eta_k^{(t)}$.

Step 2. By Lemma C.2, $\mathcal{L}$ satisfies the KL property with an exponent of $\frac{1}{2}$. Then from Definition C.2, this suggests that at $x = \mathbf{\Psi}^{t+1}$ and $f(x_0) = \min \mathcal{L}$

$$\|\partial^0 \mathcal{L}(\mathbf{\Psi}^{t+1})\| \geq \alpha\sqrt{\mathcal{L}(\mathbf{\Psi}^{t+1}) - \min \mathcal{L}}, \tag{14}$$

where $\alpha > 0$ is a fixed constant defined in Lemma C.2. This property is equivalent to the error bound condition, $(\partial^0\mathcal{L}, \alpha, \Omega)$-(res-obj-EB), defined in Definition 5 in Zhang (2020), for $\Omega \subset \operatorname{dom}\partial\mathcal{L}$. This is strictly weaker than strong convexity (see Section 4 in Zhang (2020)).

At iteration $t+1$, there exists $\xi_k^{(t+1)} \in \partial G_k(\mathbf{\Psi}_k^{(t+1)})$ satisfying the optimality condition:

$$\nabla_k H(\mathbf{\Psi}_{i<k}^{(t+1)}, \mathbf{\Psi}_{i\geq k}^{(t)}) + \frac{1}{\eta_k^{(t+1)}}(\mathbf{\Psi}_k^{(t+1)} - \mathbf{\Psi}_k^{(t)}) + \xi_k^{(t+1)} = 0.$$

Let $\xi^{(t+1)} := (\xi_1^{(t+1)}, \ldots, \xi_K^{(t+1)})$. Then,

$$\nabla H(\mathbf{\Psi}^{(t+1)}) + \xi^{(t+1)} \in \partial\mathcal{L}(\mathbf{\Psi}^{(t+1)})$$

and hence the error bound condition becomes

$$\mathcal{L}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L} \leq \frac{\|\partial^0\mathcal{L}(\mathbf{\Psi}^{(t+1)})\|^2}{\alpha^2} \leq \frac{\|\nabla H(\mathbf{\Psi}^{(t+1)}) + \xi^{(t+1)}\|^2}{\alpha^2}.$$

It follows that

$$\begin{aligned}
\|\nabla H(\boldsymbol{\Psi}^{(t+1)}) + \xi^{(t+1)}\|^2 &= \sum_{k=1}^{K} \|\nabla_k H(\boldsymbol{\Psi}^{(t+1)}) - \nabla_k H(\boldsymbol{\Psi}_{i<k}^{(t+1)}, \boldsymbol{\Psi}_{i\geq k}^{(t)}) - \frac{1}{\eta_k^{(t+1)}}(\boldsymbol{\Psi}_k^{(t+1)} - \boldsymbol{\Psi}_k^{(t)})\|^2 \\
&\leq \sum_{k=1}^{K} 2\|\nabla_k H(\boldsymbol{\Psi}^{(t+1)}) - \nabla_k H(\boldsymbol{\Psi}_{i<k}^{(t+1)}, \boldsymbol{\Psi}_{i\geq k}^{(t)})\|^2 + \sum_{k=1}^{K} \frac{2}{(\eta_k^{(t+1)})^2}\|\boldsymbol{\Psi}_k^{(t+1)} - \boldsymbol{\Psi}_k^{(t)}\|^2 \\
&\leq \sum_{k=1}^{K} 2\|\nabla H(\boldsymbol{\Psi}^{(t+1)}) - \nabla H(\boldsymbol{\Psi}_{i<k}^{(t+1)}, \boldsymbol{\Psi}_{i\geq k}^{(t)})\|^2 + \sum_{k=1}^{K} \frac{2}{(\eta_k^{(t+1)})^2}\|\boldsymbol{\Psi}_k^{(t+1)} - \boldsymbol{\Psi}_k^{(t)}\|^2 \\
&\leq \sum_{k=1}^{K} 2\Big(\sum_{j=1}^{K} \frac{1}{\eta_j^{(t+1)}}\Big)^2 \|\boldsymbol{\Psi}_{i\geq k}^{(t+1)} - \boldsymbol{\Psi}_{i\geq k}^{(t)}\|^2 + \sum_{k=1}^{K} \frac{2}{(\eta_k^{(t+1)})^2}\|\boldsymbol{\Psi}_k^{(t+1)} - \boldsymbol{\Psi}_k^{(t)}\|^2 \\
&\leq \left(2Kc^2\Big(\sum_{j=1}^{K} L_j\Big)^2 + 2c^2 L_{\max}\right)\|\boldsymbol{\Psi}^{(t+1)} - \boldsymbol{\Psi}^{(t)}\|^2.
\end{aligned}$$

Therefore, we get

$$\mathcal{L}(\boldsymbol{\Psi}^{(t+1)}) - \min \mathcal{L} \leq \frac{\left(2Kc^2\Big(\sum_{j=1}^{K} L_j\Big)^2 + 2c^2 L_{\max}\right)}{\alpha^2}\|\boldsymbol{\Psi}^{(t+1)} - \boldsymbol{\Psi}^{(t)}\|^2. \tag{15}$$

Step 3. Combining (13) and (15), we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\Psi}^{(t)}) - \min \mathcal{L} &= \Big(\mathcal{L}(\boldsymbol{\Psi}^{(t)}) - \mathcal{L}(\boldsymbol{\Psi}^{(t+1)})\Big) + \Big(\mathcal{L}(\boldsymbol{\Psi}^{(t+1)}) - \min \mathcal{L}\Big) \\
&\geq \frac{L_{\min}}{2}\|\boldsymbol{\Psi}^{(t)} - \boldsymbol{\Psi}^{(t+1)}\|^2 + \Big(\mathcal{L}(\boldsymbol{\Psi}^{(t+1)}) - \min \mathcal{L}\Big) \\
&\geq \left(\frac{\alpha^2 L_{\min}}{4Kc^2(\sum_{j=1}^{K} L_j)^2 + 4c^2 L_{\max}} + 1\right)\Big(\mathcal{L}(\boldsymbol{\Psi}^{(t+1)}) - \min \mathcal{L}\Big).
\end{aligned}$$

This completes the proof.

D. SG-PALM with Non-Convex Regularizers

The estimation algorithm for non-convex regularizer is largely the same as Algorithm 1, except with an additional term added to the gradient term. Specifically, the updates are of the form

$$\boldsymbol{\Psi}_k^{(t+1)} = \text{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}}\Big(\boldsymbol{\Psi}_k^t - \eta_k^t \nabla_k \bar{H}(\boldsymbol{\Psi}_{i<k}^{t+1}, \boldsymbol{\Psi}_{i\geq k}^t)\Big),$$

where

$$\bar{H}(\boldsymbol{\Psi}) = H(\boldsymbol{\Psi}) + \sum_{k=1}^{K}\sum_{i\neq j}\Big(g_{\lambda_k}([\boldsymbol{\Psi}_k]_{i,j}) - \lambda_k|[\boldsymbol{\Psi}_k]_{i,j}|\Big).$$

Here, the formulation covers a range of non-convex regularizations. Particularly, the SCAD penalty (Fan & Li, 2001) with parameter $a > 2$ is given by

$$g_\lambda(t) = \begin{cases} \lambda|t|, \quad \text{if} \quad |t| < \lambda \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)}, \quad \text{if} \quad \lambda < |t| < a\lambda \\ \frac{(a+1)\lambda^2}{2}, \quad \text{if} \quad a\lambda < |t|, \end{cases}$$

which is linear for small $|t|$, constant for large $|t|$, and a transition between the two regimes for moderate $|t|$.

The MCP penalty (Zhang et al., 2010) with parameter $a > 0$ is given by

$$g_\lambda(t) = \text{sign}(t)\lambda \int_0^{|t|}\Big(1 - \frac{z}{a\lambda}\Big)_+ dz,$$

which gives a smoother transition between the approximately linear region and the constant region ($t > a\lambda$) as defined in SCAD.

The updates can also be written as

$$\mathbf{\Psi}_k^{(t+1)} = \text{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}} \left(\mathbf{\Psi}_k^t - \eta_k^t \nabla_k \Big(H(\mathbf{\Psi}_{i<k}^{t+1}, \mathbf{\Psi}_{i\geq k}^t) + Q'_{\lambda_k}(\mathbf{\Psi}_k) \Big) \right),$$

where $q'_\lambda(t) := \frac{d}{dt}(g_\lambda(t) - \lambda|t|)$ for $t \neq 0$ and $q'_\lambda(0) = 0$ and Q'_λ denotes q'_λ applied elementwise to a matrix argument. These updates can be inserted into the framework of Algorithm 1. The details are summarized in Algorithm 2.

Algorithm 2 SG-PALM with non-convex regularizer

```
Input: Data tensor $\boldsymbol{\mathcal{X}}$, mode-$k$ Gram matrix $\mathbf{S}_k$, regularizing parameter $\lambda_k$, backtracking constant $c \in (0,1)$, initial step size $\eta_0$, initial iterate $\mathbf{\Psi}_k$ for each $k = 1, \dots, K$.
  while not converged do
    for $k = 1, \dots, K$ do
      Line search: Let $\eta_k^t$ be the largest element of $\{c^j \eta_{k,0}^t\}_{j=1,\dots}$ such that condition (8) is satisfied for $\mathbf{\Psi}_k^{t+1} = \text{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}} \left( \mathbf{\Psi}_k^t - \eta_k^t \nabla_k \Big( H(\mathbf{\Psi}_{i<k}^{t+1}, \mathbf{\Psi}_{i\geq k}^t) + Q'_{\lambda_k}(\mathbf{\Psi}_k) \Big) \right)$.
      Update: $\mathbf{\Psi}_k^{t+1} \longleftarrow \text{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}} \left( \mathbf{\Psi}_k^t - \eta_k^t \nabla_k \Big( H(\mathbf{\Psi}_{i<k}^{t+1}, \mathbf{\Psi}_{i\geq k}^t) + Q'_{\lambda_k}(\mathbf{\Psi}_k) \Big) \right)$.
    end for
    Next initial step size: Compute Barzilai-Borwein step size $\eta_0^{t+1} = \min_k \eta_{k,0}^{t+1}$, where $\eta_{k,0}^{t+1}$ is computed via (9).
  end while
Output: Final iterates $\{\mathbf{\Psi}_k\}_{k=1}^K$.
```

D.1. Convergence Property

Consider a sequence of iterate $\{\mathbf{x}^t\}_{t\in\mathbb{N}}$ generated by a generic PALM algorithm for minimizing some objective function f. Specifically, assume

($\mathcal{H}_1$) $\inf f > -\infty$.

($\mathcal{H}_2$) The restriction of the function to its domain is a continuous function.

($\mathcal{H}_3$) The function satisfies the KL property.

Then, as in Theorem 2 of Attouch & Bolte (2009), if this objective function satisfying ($\mathcal{H}_1$), ($\mathcal{H}_2$), ($\mathcal{H}_3$) in addition satisfies the KL property with

$$\phi(s) = \alpha s^{1-\theta},$$

where $\alpha > 0$ and $\theta \in (0, 1]$. Then, for $\mathbf{x}^*$ some critical point of f, the following estimations hold

(i). If $\theta = 0$ then the sequence of iterates converges to $\mathbf{x}^*$ in a finite number of steps.

(ii). If $\theta \in (0, \frac{1}{2}]$ then there exist $\omega > 0$ and $\tau \in [0, 1)$ such that $\|\mathbf{x}^t - \mathbf{x}^*\| \leq \omega \tau^t$.

(iii). If $\theta \in (\frac{1}{2}, 1)$ then there exist $\omega > 0$ such that $\|\mathbf{x}^t - \mathbf{x}^*\| \leq \omega t^{-\frac{1-\theta}{1\theta-1}}$.

In the case of SG-PALM with non-convex regularizations, so long as the non-convex $\mathcal{L}$ satisfies the KL property with an exponent in $(0, \frac{1}{2}]$, the algorithm remains linearly convergent (to a critical point). We argue that this is true for SG-PALM with MCP or SCAD penalty. Li & Pong (2018) showed that penalized least square problems with such penalty functions satisfy the KL property with an exponent of $\frac{1}{2}$. The proof strategy for the convex case can be easily adopted, incorporating the KL results for MCP and SCAD in Li & Pong (2018), to show that the new $\mathcal{L}$ still has KL exponent of $\frac{1}{2}$. Therefore, SG-PALM with MCP or SCAD penalty converges linearly in the sense outlined above.

E. Additional Details of the Solar Flare Experiments

E.1. HMI and AIA Data

The Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) data characterize solar variability including the Sun's interior and the various components of magnetic activity; the SDO/Atmospheric Imaging Assembly (AIA) data contain a set of measurements of the solar atmosphere spectrum at various wavelengths. In general, HMI produces data that is particularly useful in determining the mechanisms of solar variability and how the physical processes inside the Sun that are related to surface magnetic field and activity. AIA contains structural information about solar flares, and the the high AIA pixel values are correlated with the flaring intensities. We are interested in examining if combination of multiple instruments enhances our understanding of the solar flares, comparing to the case of single instrument. Both HMI and AIA produce multi-band (or multi-channel) images, for this experiment we use all three channels of the HMI images and $9.4, 13.1, 17.1, 19.3$ nm wavelength channels of the AIA images. For a detailed descriptions of the instruments and all channels of the images, see `https://en.wikipedia.org/wiki/Solar_Dynamics_Observatory` and the references therein. Furthermore, for training and testing involved in this study, we used the data described in (Galvez et al., 2019), which are further pre-processed HMI and AIA imaging data for machine learning methods.

E.2. Classification of Solar Flares/Active Regions (AR)

The classification system for solar flares uses the letters A, B, C, M or X, according to the peak flux in watts per square metre (W/m^2) of X-rays with wavelengths 100 to 800 picometres (1 to 8 angstroms), as measured at the Earth by the GOES spacecraft (`https://en.wikipedia.org/wiki/Solar_flare#Classification`). Here, A usually refers to a "quite" region, which means that the peak flux of that region is not high enough to be classified as a real flare; B usually refers a "weak" region, where the flare is not strong enough to have impact on spacecrafts, earth, etc; and M or X refers to a "strong" region that is the most detrimental. Differentiating between a weak and a strong flare/region ahead of time is a fundamental task in space physics and has recently attracted attentions from the machine learning community (Chen et al., 2019; Jiao et al., 2019; Sun et al., 2019). In our study, we also focus on B and M/X flares and attempt to predict the videos that lead to either one of these two types of flares.

E.3. Run Time Comparison

We compare run times of the SG-PALM algorithm for estimating the precision matrix from the solar flare data with SyGlasso. Table 2 illustrates that the SG-PALM algorithm converges faster in wallclock time. Note that in this real dataset, which is potentially non-Gaussian, the convergence behavior of the algorithms is different compare to synthetic examples. Nonetheless, SG-PALM enjoys an order of magnitude speed-up over SyGlasso.

Table 2. Run time (in seconds) comparisons between SyGlasso and SG-PALM on solar flare data for different regularization parameters. Note that the SG-PALM is an order of magnitude faster that SyGlasso.

λ	**SyGlasso**		**SG-PALM**	
	iter	**sec**	**iter**	**sec**
0.28	47	5772.1	89	583.7
0.41	43	5589.0	86	583.4
0.54	45	5673.7	85	568.8
0.67	42	5433.0	77	522.6
0.79	39	4983.2	82	511.4
0.92	40	5031.9	72	498.0
1.05	39	4303.7	76	452.2
1.18	41	4234.7	64	437.6
1.30	40	4039.5	58	406.9
1.43	35	3830.7	64	364.9

E.4. Examples of Predicted Magnetogram Images

Figure 4 depicts examples of the predicted HMI channels by SG-PALM. We observe that the proposed method was able to reasonably capture various components of the magnetic field and activity. Note that the spatial behaviors of the HMI components are quite different from those of AIA channels, that is, the structures tend to be less smooth and continuous (e.g., separated holes and bright spots) in HMI.

E.5. Multi-instrument vs. Single Instrument Prediction

To illustrate the advantages of multi-instrument analysis, we compare the NRMSEs between an AIA-only (i.e., last four channels of the dataset) and an HMI&AIA (i.e., all seven channels of the dataset) study in predicting the last frames of 13-frame AIA videos, for each flare class, respectively, using the proposed SG-PALM. The results are depicted in Figure 5, where the average, standard deviation, and range of the NRMSEs across pixels are also shown for each error image. By leveraging the cross-instrument correlation structure, there is a $0.5\% - 1\%$ drop in the averaged error rates and a $2\% - 4\%$ drop in the range of the errors.

E.6. Illustration of the Difficulty of Predictions for Two Flares Classes

We demonstrate the difficulty of forward predictions of video frames. Figure 6 depicts two different channels of multiple frames from two videos leading to MX-class solar flares. Note that the current frame is the 13th frame in the sequence that we are trying to predict. We observe that the prediction task is particularly difficult if there is a sudden transition of either the brightness or spatial structure of the frames near the end of the video. These sudden transitions are more frequent for MX flares than for B flares. In addition, as MX flares are generally considered as rare events (i.e., less frequent than B flares), it is harder for SG-PALM or related methods to learn a common correlation structures from training data.

On the other hand, typical image sequences leading to B flares exhibit much smoother transitions from frame to frame. As shown in Figure 7, the SG-PALM was able to produce remarkably good predictions of the current frames.

E.7. Illustration of the Estimated Sylvester Generating Factors

Figure 8 illustrates the patterns of the estimated Sylvester generating factors ($\mathbf{\Psi}_k$'s) for each flare class. Here, the videos from both classes appear to form Markov Random Fields, that is, each pixel only depends on its close neighbors in space and time given all other pixels. This is demonstrated by observing that the temporal or each of the spatial generating factor, which can be interpreted as conditional dependence graph for the corresponding mode, has its energies concentrate around the diagonal and decay as the nodes move far apart (in space or time).

The spatial patterns are similar for different flares. Although the exact spatial patterns are different from one frame to another, they always have their energies being concentrated at certain region (i.e., the brightest spot) that is usually close to the center of the images. This is due to the way how these images were curated and pre-processed before analysis. On the other hand, the temporal structures are quite different. Specifically, B flares tend to have longer range dependencies, as the frames leading to these types flares are smooth, which is consistent with results from the previous section.

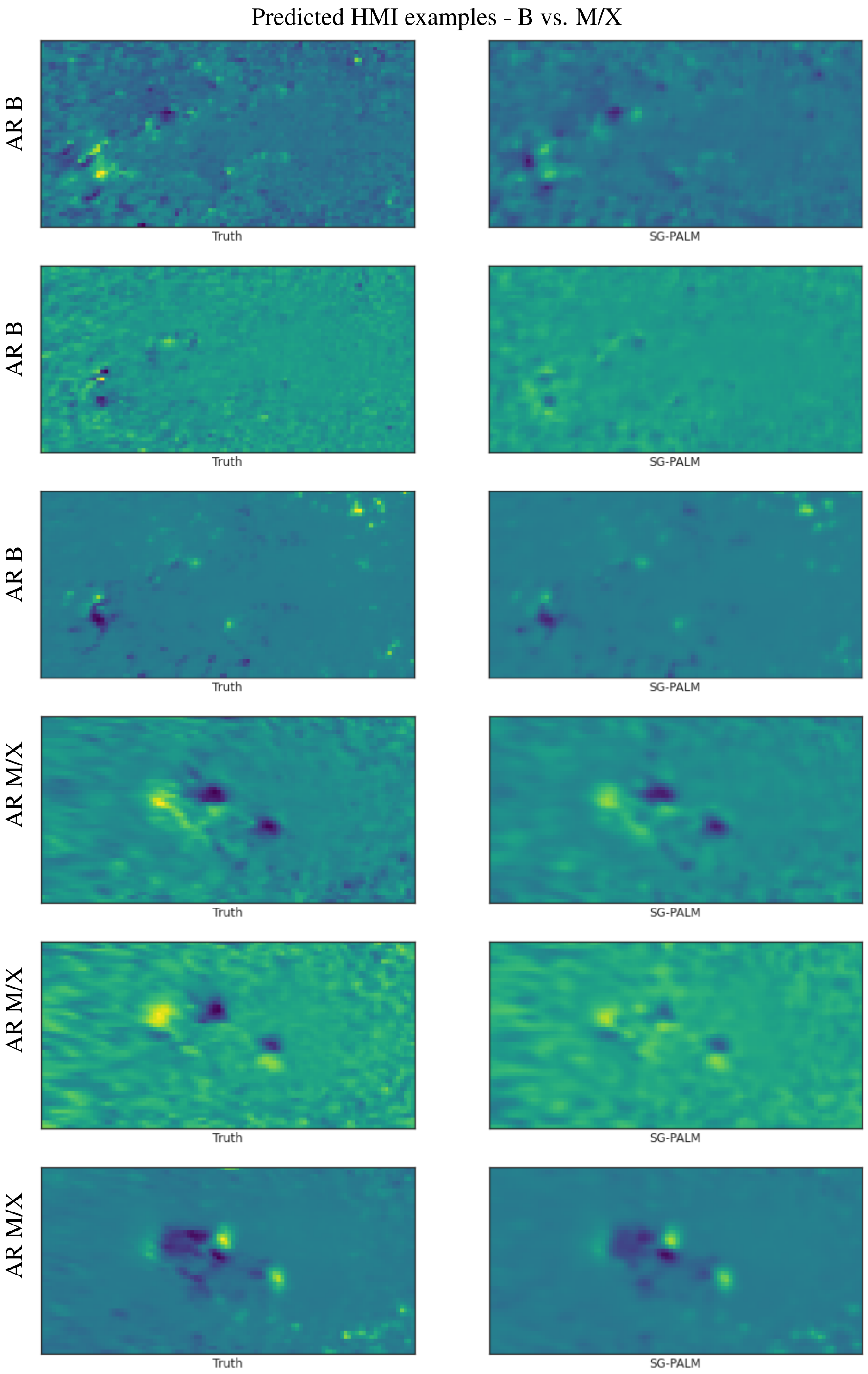


Figure 4. Examples of one-hour ahead prediction of the first three channels (HMI components) of ending frames of 13-frame videos, leading to B- (first three rows) and MX-class (last three rows) flares, produced by the SG-PALM, comparing to the real image (left column). Similarly to AIA predictions, linear forward predictors tend to underestimate the contrast ratio of the images. Nonetheless, the SG-PALM algorithm was able to both capture the spatial structures and intensities of the underlying magnetic fields. Note that the HMI images tend to be harder to predict, as indicated by the increased number and decreased degree of smoothness of features in the images, signifying the underlying magnetic activity on the solar surface.

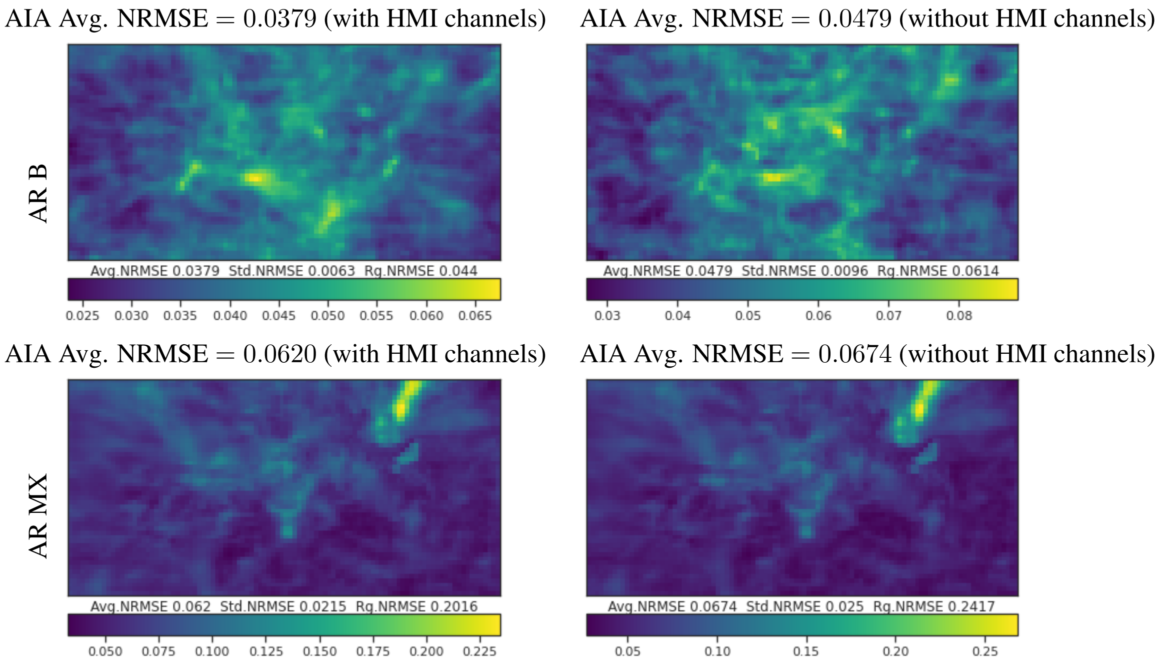


Figure 5. Comparison of the SG-PALM performance measured by NRMSE in predicting the AIA channels (i.e., last four channels) of the ending frame of 13-frame videos leading to B- and MX-class solar flares, by using all HMI&AIA channels (left column) and AIA-only channels (right column). The NRMSEs are computed by averaging across both testing samples and channels for each pixel. Note that there are improvements in both the averaged errors rates and the uncertainty in those errors (i.e., range of the errors) by including multi-instrument image channels.

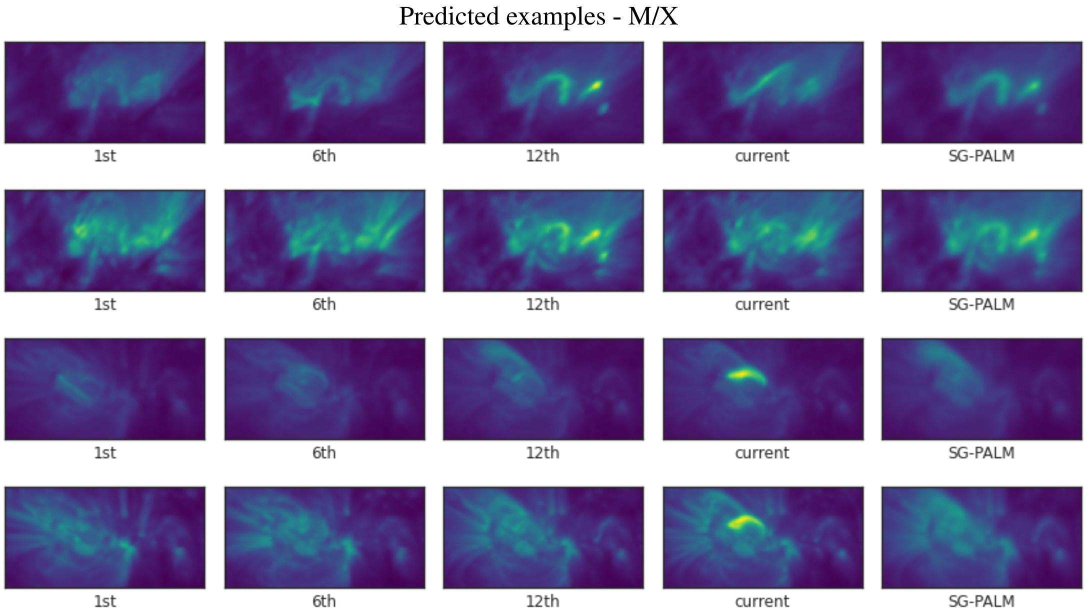


Figure 6. Examples of frames at various timestamps of videos preceding the predictions of the last frames (last column) that lead to MX flares. Here, the first two rows correspond to the same video as the last two rows in Figure 3. Note that the prediction tasks are difficult in these two extreme cases, where there are dramatic changes from the 12th to the current (13th) frames.

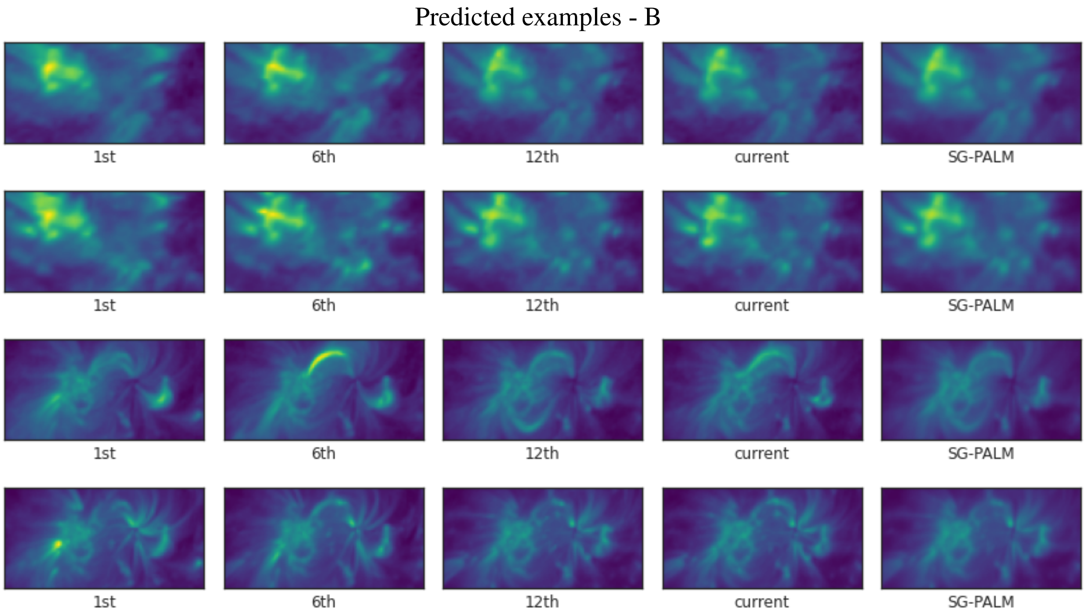
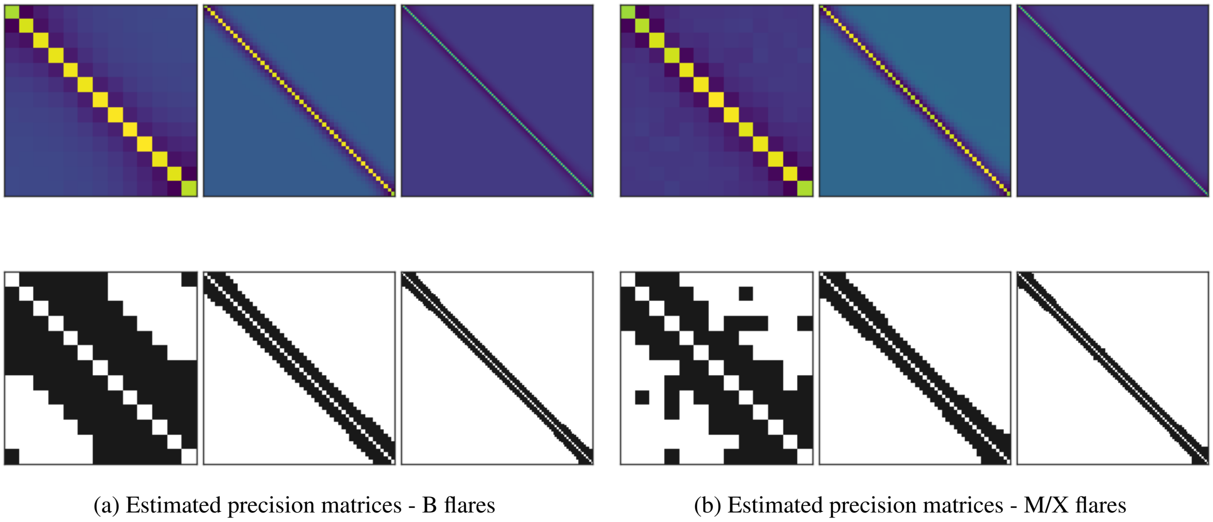


Figure 7. Examples of frames at various timestamps of videos preceding the predictions of the last frames (last column) that lead to B flares. Here, the first two rows correspond to the same video as the first two rows in Figure 3. Note that the prediction tasks are easier than those illustrated in Figure 6, since the transitions near the end of the videos are much smoother.



(a) Estimated precision matrices - B flares

(b) Estimated precision matrices - M/X flares

Figure 8. Estimated spatial and two (longitude and latitude) temporal Sylvester generating factors for B and MX solar flares, along with their off-diagonal sparsity patterns (second row in each subplot). Both classes exhibit autoregressive dependence structures (across time or space). Note the significant difference in the temporal components, where the B flares exhibit longer range dependency. This is consistent with the smooth transition property of the corresponding videos as illustrated previously.