Kronecker-structured covariance models for multiway data[*]

Yu Wang, Zeyu Sun, Dogyoon Song and Alfred Hero

University of Michigan,
Ann Arbor, MI 48109
e-mail: wayneyw@umich.edu*;* zeyusun@umich.edu*;* dogyoons@umich.edu*;* hero@umich.edu

Abstract: Many applications produce multiway data of exceedingly high dimension. Modeling such multi-way data is important in multichannel signal and video processing where sensors produce multi-indexed data, e.g. over spatial, frequency, and temporal dimensions. We will address the challenges of covariance representation of multiway data and review some of the progress in statistical modeling of multiway covariance over the past two decades, focusing on tensor-valued covariance models and their inference. We will illustrate through a space weather application: predicting the evolution of solar active regions over time.

Contents

1. Introduction

Probabilistic modeling of data is ubiquitous in statistics and machine learning, and arguably, estimating dependencies across covariates is one of the most fundamental modeling tasks. While first order "mean field" analysis can describe typical mean behavior of covariates, higher order analysis is required to capture interactions between covariates. Among the latter, covariance estimation is a second order analysis that has been extensively studied for empirically estimating the covariance or the inverse covariance (precision) matrix from multivariate data. Besides being of interest in their own right, covariance estimates are often a prerequisite in other applications, including: dimensionality reduction; multivariate clustering; supervised classification; linear prediction; and model selection [42, 63, 101, 5, 24, 79, 80, 81].

A striking feature of many modern datasets is their high-dimensionality and multiway nature, often involving a huge number of multi-indexed variables represented as tensor-valued data [46, 48]. Second order analysis of such tensor-valued data poses significant modeling and computational challenges due to the intrinsically high computational complexity of manipulating tensor representations [35]. To overcome these challenges, several methodological approaches have been developed over the past 50 years that reduce model complexity by imposing sparse or low-dimensional dependency structures in the covariance or inverse covariance. These include Kronecker generalizations of PCA (Kronecker PCA) [84, 30] and the matrix normal model [19], where the covariates are multivariate normal with multiway covariance structure modeled as a Kronecker product of matrices of much lower dimension. These Kronecker representations of covariance can have Kronecker factors that are either dense [22, 95] or sparse [2, 85].

The introduction of such parsimoniously structured tensor covariance models, and fast iterative computational algorithms for inferring these models, have made tensor covariance models more practical for many applications. In particular, such models have been applied to image classification [25], spatio-temporal image processing [30, 32, 65, 20], bioinformatics [83, 31, 82, 21], atmospheric science [54, 33], neuroscience [91, 78, 7], medical imaging [58, 57], and digital advertising [34]. Several other applications of multiway covariance models are discussed in the unpublished survey [76].

In this article, we survey some of the recent advances in second order analysis of multiway data. The focus is on Kronecker structured representations and approximations to multiway covariance. After introducing some notation in Section 2, we describe how multiway covariance structure arises in different applications, including in linear prediction of dynamic processes whose evolu-

tion is governed by differential equations. We then overview recent approaches to Kronecker modeling and estimation of multiway covariance and inverse covariance in Section 3. In Section 4, we summarize and compare these Kronecker models and estimators based on their theoretical performance properties and a numerical comparison study. The numerical study focuses on model comparison for an important application in the domain of space weather: prediction of solar active region evolution over time. In Section 5, we conclude the paper with a few open research directions in second order multiway data modeling. A software package [92] accompanies this survey with code implementing the methods described herein.

This survey is intended to complement previous surveys on aspects of high dimensional covariance modeling important to practitioners. These include surveys on robust structured covariance estimation [96, 44] and regularized covariance estimation [69, 49, 13]. Here, we focus on the class of Kronecker structured covariance models.

2. Background

2.1. Notation

Throughout the paper, scalars are denoted by lowercase letters, vectors by boldface lowercase letters, and matrices by boldface capital letters. All matrix and vector quantities are assumed to be real valued, although all the multiway covariance models discussed in this paper straightforwardly generalize to complex valued data. For a matrix $\mathbf{A} = (\mathbf{A}_{i,j}) \in \mathbb{R}^{d \times d}$, the spectral norm is denoted as $\|\mathbf{A}\|_2$, the Frobenius norm as $\|\mathbf{A}\|_F$, and the nuclear norm as $\|\mathbf{A}\|_*$. We define $\|\mathbf{A}\|_{1,\text{off}} := \sum_{i \neq j} |\mathbf{A}_{i,j}|$ as its off-diagonal ℓ_1 norm. For tensor algebra, we adopt the notations used by Kolda and Bader [46]. K-th order tensors are denoted by boldface Euler script letters, e.g., $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, with the $(i_1, \ldots, i_K)$-th element denoted as $\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_K}$. A fiber of a tensor is a higher order analogue of a row or a column of a matrix. It is obtained by fixing all but one of the indices of the tensor. The vectorization of $\boldsymbol{\mathcal{X}}$, denoted as $\text{vec}(\boldsymbol{\mathcal{X}})$, stacks the entries of the tensor into a d-dimensional column vector $(\boldsymbol{\mathcal{X}}_{1,1,\ldots,1}, \boldsymbol{\mathcal{X}}_{2,1,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{d_1,d_2,\ldots,d_k})^T$, where $d = \prod_{k=1}^K d_k$. Matricization, also known as unfolding, is the process of transforming a tensor into a matrix. The mode-k matricization of a tensor $\boldsymbol{\mathcal{X}}$, denoted by $\boldsymbol{\mathcal{X}}_{(k)}$, arranges the mode-$k$ fibers to be the columns of the resulting matrix. The k-mode product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ and a matrix $\mathbf{A} \in \mathbb{R}^{J \times d_k}$, denoted as $\boldsymbol{\mathcal{X}} \times_k \mathbf{A}$, is a tensor of size $d_1 \times \cdots \times d_{k-1} \times J \times d_{k+1} \times \ldots d_K$, with entries defined as $(\boldsymbol{\mathcal{X}} \times_k \mathbf{A})_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} := \sum_{i_k=1}^{d_k} \boldsymbol{\mathcal{X}}_{i_1,\ldots,i_K} A_{j,i_k}$. We define the K-way Kronecker product (KP) as $\bigotimes_{k=1}^K \mathbf{A}_k = \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_K$, and the K-way Kronecker sum (KS) as $\bigoplus_{k=1}^K \mathbf{A}_k = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_K = \sum_{k=1}^K \mathbf{I}_{[d_{1:k-1}]} \otimes \mathbf{A}_k \otimes \mathbf{I}_{[d_{k+1:K}]}$, where $\mathbf{I}_{[d_{k:\ell}]} = \mathbf{I}_{d_k} \otimes \cdots \otimes \mathbf{I}_{d_\ell}$. For the case of $K = 2$, $\mathbf{A}_1 \oplus \mathbf{A}_2 = \mathbf{A}_1 \otimes \mathbf{I}_{d_2} + \mathbf{I}_{d_1} \otimes \mathbf{A}_2$.

Statistical convergence rate will be denoted by the $O_P(\cdot)$ notation, which is defined as follows. Consider a sequence of real random variables $\{X_n\}_{n \in \mathbb{N}}$ defined

on a probability space $(\Omega, \mathcal{F}, P)$ and a deterministic (positive) sequence of reals $\{b_n\}_{n \in \mathbb{N}}$. By $X_n = O_P(1)$ is meant: $\sup_{n \in \mathbb{N}} P(|X_n| > K) \to 0$ as $K \to \infty$, where X_n is a sequence indexed by n. The notation $X_n = O_P(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_P(1)$. By $X_n = o_p(1)$ is meant $P(|X_n| > \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. By $\lambda_n \asymp b_n$ is meant $c_1 \leq \frac{\lambda_n}{b_n} \leq c_2$ for all n, where $c_1, c_2 > 0$ are absolute constants. The asymptotic notation $a_n = O(b_n)$ means $\lim \sum_{b \to \infty} |\frac{a_n}{b_n}| \leq C$ for some constant $C > 0$, while $c_n = \Omega(d_n)$ means $\liminf_{n \to \infty} |\frac{c_n}{d_n}| \geq C'$ for some constant $C' > 0$.

2.2. *Multiway data*

Multiway data frame   Multiway data are data that are indexed over multiple domains, called modes, and can be represented as a tensor-valued data frame. The number of modes is called the order of the tensor. Examples of multiway data include 3D images of the brain, where the modes are the 3 spatial dimensions, and spatio-temporal weather imaging data, a set of image sequences represented as 2 spatial modes and 1 temporal mode. One can also obtain a multiway data by stacking independent, identically distributed data samples.

Multiway patterned covariance matrix   Given a multiway data tensor of order K, the covariance matrix of the vectorized data has a special patterned structure. For instance, suppose that $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is a multiway data frame of order 2, (i.e., a matrix). Then the covariance matrix of $\text{vec}(\mathbf{X})$ is a $d_1 d_2$ by $d_1 d_2$ matrix $\mathbf{\Sigma} = (\Sigma_{ij})$ that corresponds to the covariance of the i-th and the j-th rows of $\mathbf{X}$. See Figure 1 for a visual illustration. As a special case, if the rows of $\mathbf{X}$ are independent and identically distributed with $d_2 \times d_2$ covariance $\mathbf{\Sigma}_{d_2}$, then $\text{cov}(\mathbf{X}) = \mathbf{I}_{d_1} \otimes \mathbf{\Sigma}_{d_2}$.

Utility of multiway data for prediction   Arguably, regression analysis is one of the most straightforward approaches for prediction as it attempts to estimate the relation between the covariates and the response variable. Despite its long and fruitful history, traditional regression techniques often fall short of handling multiway data encountered in modern applications, due to the ultra-high data dimensionality as well as complex data structures. Recently, several tensor regression models were proposed to exploit the special structure of tensor covariates [105, 34] or of tensor responses [71, 53, 106]. These models try to utilize the sparse and low-rank structures in the tensors – either in the regression coefficient tensor or the response tensor – to boost performance on the regression task by reducing the number of free parameters. Similarly, in Section 4.3, we demonstrate the utility of sparse precision/covariance matrices in linear prediction for tensor-valued data.

2.3. *Multiway covariance representations for diffusion processes*

While multiway representations have been successfully applied to many different types of data, they are particularly interpretable when modeling data generated
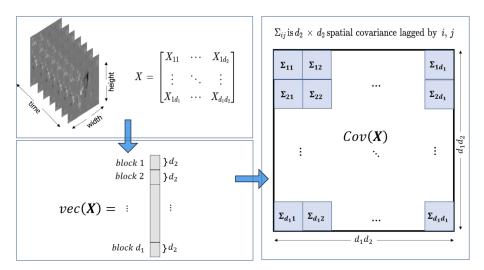
FIG 1. *Multiway data and patterned covariance.*

from physical diffusion processes. In particular, consider a random process u whose sample paths obey partial differential equations of the form

$$
\begin{aligned}
\mathcal{D}u &= f \quad \text{in } \Omega, \\
u &= g \quad \text{on } \partial\Omega,
\end{aligned}
\tag{1}
$$

where u is the unknown physical process, f is the driving process (e.g., white Gaussian noise), g is the function value of u on the boundary, $\mathcal{D}$ is some differential operator (e.g, a Laplacian or an Euler-Lagrange operator), and Ω is the domain. After finite difference discretization over the domain Ω, the model is equivalent to (ignoring discretization error) the matrix equation

$$
\mathbf{Du} = \mathbf{f}.
$$

Here, $\mathbf{D}$ is a sparse matrix since $\mathcal{D}$ is a differential operator. Additionally, as shown below, $\mathbf{D}$ admits the Kronecker structure as a mixture of Kronecker sums and Kronecker products.

The matrix $\mathbf{D}$ reduces to a Kronecker sum when $\mathcal{D}$ involves no mixed derivatives. As an example, we consider the Poisson equation, an elliptical PDE that governs many physical processes including electromagnetic induction, heat transfer, and convection. On a rectangular region $\Omega = (0, d_1) \times (0, d_2)$ in the 2D Cartesian plane, the Poisson equation with homogeneous Dirichlet boundary condition is expressed as

$$
\begin{aligned}
\mathcal{D}u = (\partial_x^2 + \partial_y^2)u &= f \quad \text{in } \Omega, \\
u &= 0 \quad \text{on } \partial\Omega
\end{aligned}
\tag{2}
$$

where $f : \Omega \to \mathbb{R}$ is the given source function and $u : \Omega \to \mathbb{R}$ is the unknown process of interest. Using the finite difference method with a square mesh grid

with unit spacing, the unknown and the source can be expressed as d_1-by-d_2 matrices, $\mathbf{U}$ and $\mathbf{F}$, respectively, that are related to each other via

$$U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{i,j} = F_{i,j} \tag{3}$$

for any interior grid point (i, j). Defining n-by-n square matrix

$$\mathbf{A}_n = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix},$$

the above relation can be expressed as the (vectorized) Sylvester equation with $K = 2$:

$$(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{u} = \mathbf{f}, \tag{4}$$

where $\mathbf{u} = \text{vec}(\mathbf{U})$, $\mathbf{f} = \text{vec}(\mathbf{F})$. Note that $\mathbf{A}_n$ is tridiagonal. In the case where $\mathbf{f}$ is white noise with variance σ^2, the inverse covariance matrix of $\mathbf{u}$ has the form $\text{cov}^{-1}(\mathbf{u}) = \sigma^{-2}(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})^T(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})$ and hence sparse.

More generally, any physical process generated from Equation (1) also has sparse inverse covariance matrices due to the sparsity of general discretized differential operators. Note that similar connections between continuous state physical processes and sparse "discretized" statistical models have been established by Lindgren, Rue and Lindström [56], who elucidated a link between Gaussian fields and Gauss-Markov random fields via stochastic partial differential equations.

In some applications, e.g. weather forecasting and multivariate time series, the data frame $\mathbf{X}$ can be modeled as having been generated by an underlying dynamical process obeying the aforementioned physical diffusion model. This model can be used for various tasks such as forecasting, classification, and two-sample comparison. For example, when the dynamical model is of known linear form, the Kalman filter can be used for these tasks [8]. In the case of non-linear dynamics, the ensemble Kalman filter (EnKF) [23] is applicable. Recently, a sparsity-penalized EnKF was introduced for modeling non-linear state dynamics [39]. This EnKF was implemented with a sparse inverse covariance estimator in order to stabilize the error covariance update of the Kalman filter. When the state is high dimensional and tensor-valued, [89] used Kronecker-structured models for the (inverse) covariance and demonstrated improvements in tracking certain classes of dynamical models. A similar application of Kronecker covariance modeling will be further illustrated in Section 4.2.

3. Multiway covariance models

In this section, we survey several covariance models for multiway data and review associated estimation algorithms that have been introduced in the literature. The common assumption of these models is that the multimodal covariance (or

TABLE 1

Overview of the models and the algorithms reviewed in Section 3. Here, Σ and Ω denote the covariance and the inverse covariance (precision), respectively. For the sake of brevity, the model in the second column is described for $K = 2$.

	Model	Algorithm	Note
KP-Covariance (Section 3.2)	$\boldsymbol{\Sigma} = \sum_{l=1}^{r} \mathbf{A}_l \otimes \mathbf{B}_l$	KPCA [84] Robust KPCA [31] AdaKron [52] TRCM [2] Gemini [104]	Algorithm requires $K = 2$ Algorithm requires $K = 2$ Analysis restricted to $K = 2$ Analysis restricted to $K = 2$ Analysis restricted to $K = 2$
KP-Precision (Section 3.3)	$\boldsymbol{\Omega} = \mathbf{A} \otimes \mathbf{B}$	KGlasso [2, 85] Tlasso [59]	Analysis restricted to $K = 2$ Applicable to $K \geq 2$
KS-Precision (Section 3.4)	$\boldsymbol{\Omega} = \mathbf{A} \oplus \mathbf{B}$	TeraLasso [33] BiGlasso [43]	Applicable to $K \geq 2$ Analysis restricted to $K = 2$
Sylvester GM (Section 3.5)	$\boldsymbol{\Omega} = (\mathbf{A} \oplus \mathbf{B})^2$	SyGlasso [91] SG-PALM [90]	Applicable to $K \geq 2$ Applicable to $K \geq 2$

inverse covariance) of multiway data has a factorized representation involving a Kronecker product or a Kronecker sum of smaller matrices, or as a linear combination thereof. Each of the smaller matrices corresponds to a single mode of data and these are often interpreted as mode-wise (inverse) covariances. The models and the algorithms reviewed in this section are summarized in Table 1.

3.1. The Kronecker product covariance model

For a K-mode tensor-valued data frame, the Kronecker product (KP) model of its covariance matrix $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 \otimes \cdots \otimes \boldsymbol{\Sigma}_K \tag{5}$$

where $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d_k \times d_k}, k = 1, \ldots, K$ are symmetric positive-semidefinite matrices, which are called the Kronecker factors of $\boldsymbol{\Sigma}$.

The KP model with $K = 2$ imposes a simple two-way covariance representation recast as $\boldsymbol{\Sigma} = \mathbf{A} \otimes \mathbf{B}$, arguably the simplest non-trivial one. Then, under this model, each $d_2 \times d_2$ block $[\boldsymbol{\Sigma}]_{ij}$ of the $d_1 d_2 \times d_1 d_2$ patterned covariance matrix $\boldsymbol{\Sigma}$ shown in Figure 1 is the same matrix, namely $[\boldsymbol{\Sigma}]_{ij} = \mathbf{B}$, up to a multiplicative factor. Despite its apparent simplicity, the KP model with $K = 2$ has been used effectively in many different applications, including spatio-temporal data frames arising in MIMO wireless communications [100, 95], geostatistics [16], genomics [98, 38], face recognition [102], and recommender systems [2].

The KP model was introduced by Dawid [19] in the context of the matrix normal distribution for data with $K = 2$ tensor modes. An alternating optimization algorithm for maximum likelihood estimation (MLE) of the $K = 2$ Kronecker factors was introduced by Mardia and Goodall [61] and analyzed by Dutilleul [22]. A different approach based on the covariance matching principle

was taken by Strobach [75], who cast the problem of estimating the $K = 2$ Kronecker factors as a minimum Frobenius norm covariance matching problem:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{S} - \mathbf{A} \otimes \mathbf{B}\|_F^2. \tag{6}$$

Using the fact that the matrix Frobenius norm is invariant to permutation of the entries, the above optimization (6) boils down to finding the rank-one approximation to the matrix $\mathcal{R}(\mathbf{S})$, where $\mathcal{R}$ is the rearrangement-permutation operator that maps the square matrix $\mathbf{S} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ to a rectangular matrix $\mathcal{R}(\mathbf{S}) \in \mathbb{R}^{d_1^2 \times d_2^2}$ [87]. The minimizers of (6) can thus be expressed in closed form using the left and right principal components $\mathbf{u}$ and $\mathbf{v}$ of $\mathcal{R}(\mathbf{S})$ as $\widehat{\mathbf{A}} = \mathcal{R}^{-1}(\mathbf{u})$ and $\widehat{\mathbf{B}} = \mathcal{R}^{-1}(\mathbf{v})$, and the resultant estimator can be represented as $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}}$. Compared to the MLE, this least square estimate does not require an iterative algorithm, but is generally not asymptotically efficient. A weighted Frobenius norm formulation based on (6) was considered by Werner, Jansson and Stoica [95], leading to an asymptotically efficient estimator under the matrix normal assumption.

KP models with $K > 2$ also arise in applications and are explored in the literature. Mardia and Goodall [61] considered a spatio-temporal multivariate environmental monitoring data set and adopted a KP covariance with $K = 3$. They were the first to propose a maximum likelihood estimation via triple iteration. As there were insufficient replicates, they simplified the model by assuming the temporal components were statistically independent. Galecki [26] discussed five 3-way KP models with different combinations of covariance factor structures, such as identity, compound symmetry, AR(1), and unstructured factors. Akdemir and Gupta [1] provided a heuristic method for $K \geq 2$, generalizing the flip-flop method originally proposed for matrices. The convergence of the algorithm was empirically verified using simulations with K up to 4, and using real world data with K up to 3. Hoff [36] discussed maximum likelihood estimation and Bayesian estimation for tensor normal models for general K, the latter approach verified on a 4-way international trade dataset. We note here that, in both Hoff [36] and Ohlson, Ahmad and Von Rosen [67], the KP model was referred to as the "separable covariance" model. An empirical study of the three-stage iterative algorithm for MLE of KP structured covariance with $K = 3$ was given by Manceur and Dutilleul [60], who provided numerical and simulation results showing the convergence of the algorithm. Pouryazdian, Beheshti and Krishnan [70]) proposed to select the order of canonical decomposition (CAN-DECOMP/PARAFAC) in modeling 3-way data under the assumption that the noise covariance is a KP model with $K = 3$ modes, justified by the structure of electroencephalogram (EEG) data spanning the time, frequency, and space dimensions.

3.2. Kronecker covariance decomposition and Kronecker PCA

There is a less restrictive model than (5) that approximates the covariance as a sum of several Kronecker products. For example, when $K = 2$, the covariance

under this model is expressed as

$$\boldsymbol{\Sigma} = \sum_{l=1}^{r} \mathbf{A}_l \otimes \mathbf{B}_l, \tag{7}$$

where $r \geq 1$ is a positive integer, $\mathbf{A}_l \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{B}_l \in \mathbb{R}^{d_2 \times d_2}$ for all $1 \leq l \leq r$. The sum of Kronecker product representation (7) is universal in the sense that any matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ can be expressed as a sum of r Kronecker products with $r \leq \min\{d_1^2, d_2^2\}$ in view of SVD, cf. Van Loan and Pitzianis [87, Section 2]. Indeed, even when $K \geq 3$, every tensor $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ can be represented by a sum of Kronecker products $\boldsymbol{\Sigma} = \sum_{l=1}^{r} \mathbf{A}_l^{(1)} \otimes \cdots \otimes \mathbf{A}_l^{(K)}$ where $d = \prod_{k=1}^{K}$ and $\mathbf{A}_l^{(k)} \in \mathbb{R}^{d_k \times d_k}$ for all $k \in [K]$ and $l \in [r]$. The minimum positive integer r such that $\boldsymbol{\Sigma}$ admits a representation of the form (7) is called the (tensor) rank of $\boldsymbol{\Sigma}$ [46]; note that the KP model from Section 3.1 corresponds to the special case with tensor rank 1 with additional constraints enforcing the factor matrices to be symmetric positive semidefinite.

A model of the form (7) (with small r) allows for a compact, yet quite flexible, representation of the matrix $\boldsymbol{\Sigma}$. Recently, Dantas, Cohen and Gribonval [18] investigated the utility of this model for hyperspectral image denoising through the lens of dictionary learning. We remark that they used the model (7) to approximate the first-order "mean-field" behavior of data rather than its covariance (cf. Section 1), and can be considered a higher-order analogue of low-rank matrix models.

In the context of second-order (covariance) analysis, Tsiligkaridis and Hero [84] proposed a penalized optimization approach for estimating a (tensor) rank r Kronecker product decomposition (7) of covariance, proposing an estimate that solves

$$\min_{\{\mathbf{A}_l, \mathbf{B}_l\}} \left\| \mathbf{S} - \sum_{l=1}^{r} \mathbf{A}_l \otimes \mathbf{B}_l \right\|_F^2 + \lambda \left\| \sum_{l=1}^{r} \mathbf{A}_l \otimes \mathbf{B}_l \right\|_*$$

for a user-supplied regularization parameter $\lambda > 0$. The solution to this penalized optimization is specified by the first r principal components of the singular value decomposition (SVD) of $\mathcal{R}(\mathbf{S})$ where r is determined by λ through a soft-thresholding of the SVD spectrum. In analogy to the ordinary PCA algorithm, the soft-thresholding SVD solution to this optimization problem was called Kronecker PCA (KPCA) in [30]. In [84], it was shown that, when the sample size $n > d_1 d_2$, KPCA provides a consistent estimator of $\boldsymbol{\Sigma}$. Bounds on statistical convergence rates were also established.

Subsequently, regularized variants of KPCA were proposed. Greenewald and Hero [30] proposed to add one pair of diagonal matrices in the representation (7), namely, $(\mathbf{A}_0, \mathbf{B}_0)$, to provide a better fit to homogeneous noise in spatiotemporal modeling. This was further extended to a robust Kronecker PCA [31], which is equipped with the additional capability to handle sparse unstructured "outlier correlations" that do not fit the vanilla KPCA model à la robust PCA [9, 11].

While the KPCA model is flexible in the sense that every covariance tensor admits a representation (7) for some r, there is no a priori guarantee that the

factors $\mathbf{A}_l, \mathbf{B}_l$ are symmetric, positive semidefinite (PSD) matrices. Indeed, one can see from a dimension counting argument[1] that if $d_1, d_2 \geq 2$, then there exists a PSD matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ that cannot be expressed as a sum of Kronecker products of symmetric matrices $\mathbf{A}_l \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{B}_l \in \mathbb{R}^{d_2 \times d_2}$. As a PSD matrix is symmetric, the aforementioned matrix $\boldsymbol{\Sigma}$ cannot be expressed as a sum of Kronecker products of PSD matrices. Unless the factors $\mathbf{A}_l, \mathbf{B}_l$ are PSD, we may not be able to interpret them as mode-wise 'principal covariance components.'

It remains an open question how closely an arbitrary covariance $\boldsymbol{\Sigma}$ can be approximated by a sum of Kronecker products of PSD matrices, or whether a typical covariance $\boldsymbol{\Sigma}$ can be represented as a sum of Kronecker products of PSD matrices under reasonable data-generative models.

3.3. Kronecker graphical Lasso and tensor Lasso

As discussed in Section 2.3, physical processes often give rise to sparse covariance or sparse inverse covariance. Furthermore, when the number of independent samples, n, is smaller than the data dimension, $d = \prod_{i=1}^{K} d_i$, imposition of a sparse structure can stabilize the multi-way covariance estimates. Leng and Pan [52] considered the problem of estimating a sparse covariance in the KP model (5) using a sparsity-penalized moment matching approach. Under the matrix normal model, this corresponds to a sparse model on the pairwise marginal dependencies in the covariates. It is sometimes more natural to impose sparsity on the conditional dependencies, equivalent to imposing sparsity on the inverse covariance instead of on the covariance itself, as demonstrated by Yuan and Lin [101] and Banerjee et al. [5]. A fast algorithm for sparse estimation of the inverse covariance in the matrix normal framework was proposed by Friedman et al. [24], who called the algorithm the graphical lasso (Glasso).

Allen and Tibshirani [3] extended the Glasso to the matrix normal model ($K = 2$ modes) and this algorithm, which we call the Kronecker graphical lasso (KGlasso), was shown to be consistent with statistical convergence bounds in [85]. The KGlasso uses an alternating flip-flop optimization on the sparsity penalized KP log-likelihood function, similarly to the algorithm proposed by Mardia and Goodall [61] for the original unpenalized KP maximum likelihood problem. Besides the diffusion processes mentioned in Section 2.3, the sparse KP inverse covariance structures also arise in classification problems where the predictors are matrix-valued [66].

The KGlasso solves the following non-convex optimization, equivalent to maximizing the penalized log likelihood,

$$\min_{\mathbf{A}, \mathbf{B}} \operatorname{tr}\{\mathbf{S}(\mathbf{A} \otimes \mathbf{B})\} - \log \det(\mathbf{A} \otimes \mathbf{B}) + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{B}\|_1 \qquad (8)$$

---

[1]The set of $d_1 d_2 \times d_1 d_2$ PSD matrices has nonempty interior in the vector space of $d_1 d_2 \times d_1 d_2$ symmetric matrices that has dimension $\binom{d_1 d_2 + 1}{2}$. Note that the tensor product of the vector space of $d_1 \times d_1$ symmetric matrices and that of $d_2 \times d_2$ symmetric matrices is the span of Kronecker products and has dimension $\binom{d_1 + 1}{2}\binom{d_2 + 1}{2}$, which is strictly smaller than $\binom{d_1 d_2 + 1}{2}$ if $d_1, d_2 \geq 2$.

where λ_1, λ_2 are sparsity regularization parameters applied to the ℓ_1 norms of the Kronecker product factors **A** and **B**. Here, the ℓ_1 norms are often restricted to range only over the off-diagonals of **A** and **B**. Like the Glasso, iterative optimization methods are required to solve the optimization problem (8), however, algorithmic implementation of such methods can be significantly more difficult because it is a non-convex problem.

The tensor Lasso (Tlasso) introduced by Lyu et al. [59] is a generalization of the KGlasso to $K \geq 2$ modes. As in the KGlasso, an alternating flip-flop maximization method is used to estimate the precision matrix factors. Consistency and a minimax property are established for the Tlasso. As pointed out in [59], normalization of the K inverse covariance factors is necessary to ensure model identifiability as the tensor normal model is invariant to reciprocal scaling and other transformations on the Kronecker factors.

Following KGlasso and Tlasso, recent works have focused on more efficient estimation algorithms for solving the non-convex problem posed by (8) with provable guarantees. For example, instead of alternatingly applying Glasso, Xu, Zhang and Gu [97] proposed an alternating block gradient descent algorithm that has been shown to converge linearly to the true unknown precision matrices with optimal statistical error rates. More recently, Min, Mai and Zhang [64] proposed a parallel estimation scheme as an alternative to previous alternating/cyclic approaches for estimating the K inverse covariance factors. They demonstrated that the parallelized version achieves similar convergence rates as alternating block descent.

3.4. Tensor graphical Lasso

The tensor graphical Lasso (TeraLasso) of Greenewald et al. [33] is a $K \geq 2$ generalization of the bigraphical lasso (BiGlasso) of Kalaitzis et al. [43]. For simplicity, we confine the discussion to the simpler $K = 2$ mode case. In contrast to the KGlasso/Tlasso, the TeraLasso replaces the Kronecker product representation of the inverse covariance, $\mathbf{\Omega} = \mathbf{A} \otimes \mathbf{B}$, with a Kronecker sum representation: $\mathbf{\Omega} = \mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B}$. The rationale of [43, 33] that motivates the Kronecker sum model is that, for the equal number of non-zero entries in factors **A** and **B**, it gives a sparser inverse covariance than the Kronecker product. This can be appreciated from the fact that the Kronecker sum representation of the inverse covariance corresponds to a Cartesian product graphical model instead of a direct product graphical model given by the Kronecker product representation. This KS structure has recently been studied in the context of matrix-valued time-series analysis – Jiang, Bigot and Maabout [40] introduced a novel multivariate autoregressive model to infer the graph topology encoded in the coefficient matrix, which captures the sparse Granger causality dependency structures describing both the sparse spatial relationship between sensors and the multiple measurement relationship. They decompose the graph by imposing a KS structure on the coefficient matrix.

Furthermore, the Kronecker sum representation converts non-convex optimization problem in the KGlasso (8) into the convex optimization problem:

$$\min_{\mathbf{A},\mathbf{B}} \operatorname{tr}\{\mathbf{S}(\mathbf{A} \oplus \mathbf{B})\} - \log \det(\mathbf{A} \oplus \mathbf{B}) + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{B}\|_1 \qquad (9)$$

where λ_1, λ_2 are sparsity regularization parameters applied to the ℓ_1 norms of the Kronecker sum factors $\mathbf{A}$ and $\mathbf{B}$. As in the KGlasso, these ℓ_1 norms range over the off-diagonals of $\mathbf{A}$ and $\mathbf{B}$. This objective can be solved via either an iterative application of Glasso [43] or by a projected proximal gradient descent algorithm [33]. More recent work has also extended these estimation schemes to be more scalable and to handle non-identifiability [99] and count data [55].

3.5. *Sylvester graphical Lasso and SG-PALM*

The Sylvester graphical lasso (SyGlasso) [91] models the precision matrix as the square of a sparse matrix that has Kronecker sum structure. Different from KGlasso, Tlasso and TeraLasso, the Sylvester graphical model uses the Sylvester tensor equation to define a generative process for the underlying multivariate tensor data. The Sylvester tensor equation has been studied in the context of finite-difference discretization of high-dimensional elliptical partial differential equations [29, 47]. Consider the simplest case of $K = 2$, any solution $\mathbf{X}$ to such a PDE must have the (discretized) form:

$$\mathbf{XA} + \mathbf{BX} = \mathbf{Z} \iff (\mathbf{A} \oplus \mathbf{B}) \operatorname{vec}(\mathbf{X}) = \operatorname{vec}(\mathbf{Z}), \qquad (10)$$

where $\mathbf{Z}$ is the driving source on the domain, and $\mathbf{A} \oplus \mathbf{B}$ represents the discretized differential operators for the PDE, e.g., Laplacian, Euler-Lagrange operators, and associated coefficients. These operators are often sparse and structured.

The Sylvester generative (SG) model (10) leads to a matrix-valued random variable $\mathbf{X}$ with a precision matrix $\boldsymbol{\Omega} = (\mathbf{A} \oplus \mathbf{B})^2$, given that $\mathbf{Z}$ is isotropic (white) Gaussian distributed. The Sylvester generating factors $\mathbf{A}$ and $\mathbf{B}$ can be obtained via minimization of the penalized negative log-pseudolikelihood

$$\min_{\mathbf{A},\mathbf{B}} \operatorname{tr}\{\mathbf{S}(\mathbf{A} \oplus \mathbf{B})^2\} - \log \det \left(\operatorname{diag}(\mathbf{A}) \oplus \operatorname{diag}(\mathbf{B})\right) + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{B}\|_1. \quad (11)$$

This differs from the true penalized Gaussian negative log-likelihood in the exclusion of off-diagonals of $\mathbf{A}$ and $\mathbf{B}$ in the log-determinant term. (11) is motivated and derived directly using the Sylvester equation defined in (10), from the perspective of solving a sparse linear system. This is a generalization and extension of the convex pseudolikelihood framework for high dimensional partial correlation estimation (CONCORD) [45] to tensor-variate variables. The pseudolikelihood function is convex and is maximized using an entrywise regression approach similar to CONCORD. It is known that inference using pseudolikelihood is consistent and enjoys the same $\sqrt{N}$ convergence rate as the MLE in general [6, 88].

Estimation of the generating parameters of the SG model is challenging since the sparsity penalty applies to the square root factors of the precision matrix, which leads to a highly coupled likelihood function. The nodewise regression

approach in SyGlasso recovers only the off-diagonal elements of each Sylvester factor. This is a deficiency in many applications where the factor-wise variances are desired. Moreover, the convergence rate of the cyclic coordinate-wise algorithm used in SyGlasso is unknown and the computational complexity of the algorithm is higher than other sparse Glasso-type procedures. To overcome these deficiencies, Wang et al. [90] proposed a proximal alternating linearized minimization method that is more flexible and versatile, called SG-PALM, for finding the minimizer of (11). SG-PALM is designed to exploit structures of the coupled objective function and yields simultaneous estimates for both off-diagonal and diagonal entries. It also achieves state-of-the-art iterative convergence results (i.e., linear convergence of the optimization error) and an order of magnitude improvement in practical runtime per iteration. Additionally, SG-PALM easily adapts to non-convex regularization functions (e.g., SCAD, MCP) and enjoys a similar rate of iterative convergence.

4. Illustrations and performance comparisons

In this section, we discuss theoretical guarantees on the multiway covariance estimation procedures described in Section 3 and provide numerical illustrations of them. In Section 4.1, we collect and compare theoretical guarantees on the statistical consistency and computational complexity for these covariance estimation algorithms. Then, in Section 4.2, we provide empirical performance comparisons when the covariance / precision matrix estimation methods are used for characterizing multiway physical processes. Finally, in Section 4.3, we demonstrate the use of tensor graphical models on the real-world application of solar active region evolution prediction. The numerical studies in Section 4.2 and 4.3 were performed using the publicly available software package [92] developed by the authors, and the readers can explore other properties beyond those presented here.

4.1. Theoretical guarantees

All of the methods described in Section 3 have non-asymptotic theoretical guarantees on the statistical consistency of the estimators and the rate of computational convergence, which can be converted to upper bounds on the sample- and computational complexity of the methods, respectively. The best known upper bounds on the estimation error (in Frobenius norm) and those on computational complexity (in the number of flops per iteration) are summarized in Table 2.

Statistical error   For each row of Table 2, the expression for statistical error applies only when the associated model is correct, i.e., there is no model mismatch (bias) between the model generating the observations and the model assumed to estimate covariance/precision. For structured covariance estimation, the more recent AdaKron algorithm achieves better statistical error (given that

TABLE 2

Comparison of known theoretical guarantees on sample complexity (statistical error) and computational complexity. Here, $M = \max\{d_1, d_2, N\}$, $m_k = \prod_{i \neq k} d_i$ is the co-dimension of the k-th mode, $d = \prod_{k=1}^{K} d_i$, and s_k characterizes the sparsity of each of the inverse covariance Kronecker factors $s_k = |\{(i,j) : i \neq j, [\mathbf{\Omega}_k]_{i,j} \neq 0\}|$, s is the sparsity of the full inverse covariance $s = |\{(i,j) : i \neq j, \mathbf{\Omega}_{i,j} \neq 0\}|$ and $s = \sum_{k=1}^{K} m_k s_k$ if $\mathbf{\Omega}$ satisfies the Kronecker sum model.

Model	Algorithm	Statistical Error	Computational Complexity
KP-Covariance (Section 3.2)	KPCA [84]	$O_P\left(\sqrt{\frac{r(d_1^2 + d_2^2 + \log M)}{N}}\right)$	$O(d_1^3 d_2^3)$
	Robust KPCA [31]	$O_P\left(\sqrt{\frac{r(d_1^2 + d_2^2 + \log M)}{N}}\right)$	$O(d_1^3 d_2^3)$
	AdaKron [52]	$O_P\left(\sum_{i=1}^{2} \sqrt{\frac{(s_i + d_i)\log d_i}{N}}\right)$	$O(d_1^3 + d_2^3)$
KP-Precision (Section 3.3)	KGlasso [2, 85]	$O_P\left(\sqrt{\frac{(d_1 + d_2)\log M}{N}}\right)$	$O(d_1^3 + d_2^3 + N d_1 d_2)$
	Tlasso [59]	$O_P\left(\sqrt{\frac{d_k(d_k + s_k)\log d_k}{Nd}}\right)$ mode-k	$O\left(\sum_{k=1}^{K}(d_k^3 + N d_k m_k^2)\right)$
KS-Precision (Section 3.4)	TeraLasso [33]	$O_P\left(\sqrt{K+1} \cdot \sqrt{\frac{(d+s)\log d}{N \min_k m_k}}\right)$	$O(dK + \sum_{k=1}^{K} d_k^3)$
Sylvester GM (Section 3.5)	SyGlasso [91]	$O_P\left(\sqrt{K} \cdot \max_k \sqrt{\frac{s_k d_k \log d}{N}}\right)$	$O\left(d \sum_{k=1}^{K}(N d_k + \sum_{j \neq k}(d_k + d_j))\right)$
	SG-PALM [90]	$O_P\left(\sqrt{K} \cdot \max_k \sqrt{\frac{s_k d_k \log d}{N}}\right)$	$O\left(\sum_{k=1}^{K}(d_k^3 + N \sum_{j \neq k} d_j m_j^2)\right)$

the true covariance matrix is sparse) than KPCA/Robust KPCA as the dependence on d_i's is linear. For structured precision matrix estimation, all the structured methods enjoy significantly reduced statistical error compared to the traditional Glasso. TeraLasso outperforms Tlasso and SyGlasso/SG-PALM approximately by a factor of d_k and s_k, respectively. Moreover, it is noteworthy that Tlasso and TeraLasso achieve single-sample convergence of the ℓ_2 norm (operator norm) errors that are not included in the table, meaning that the error vanishes with a single sample ($N = 1$) as the dimension grows to infinity [33, 59].

Computational complexity KPCA-based methods require expensive full SVD/eigenvalue factorization in general, which dominates their runtime complexity. Although speedup can be achieved via randomized algorithms or iterative algorithms for sparse matrices along with small r, the computational complexities of these improvements remain unknown for general problem instances.

All the structured precision estimation algorithms are variants of Glasso, implemented with techniques tailored to the model assumptions for speedup. Generally speaking, the resulting complexity consists of the mode-wise complexity (d_k^3) and the cost of updating the objective: dK for TeraLasso, $N \sum_k d_k m_k^2$ for Tlasso, and $N \sum_k \sum_{j \neq k} d_j m_j^2$ for SG-PALM. Note that AdaKron for structured covariance models uses a pre-processing step that avoids this operation involving d and/or N flops per-iteration, but this works only for $K = 2$. The mode-wise complexity of TeraLasso is dominated by matrix inversion, which is hard to scale for general problem instances. For Tlasso/KGlasso, the mode-wise complexity is the same as that of running a Glasso-type algorithm for each mode, which could be improved by applying state-of-the-art optimization techniques developed for

vector-variate Gaussian graphical models. For SG-PALM, the mode-wise operations involve only sparse-dense matrix multiplications, which could be improved to $O(d_k \cdot \text{nnz})$, where nnz counts the number of non-zero elements of the sparse matrix (i.e., the estimated $\mathbf{\Omega}_k$ at each iteration). This could greatly reduce the computational cost for extremely sparse $\mathbf{\Omega}_k$, e.g., with only $O(d_k)$ non-zero elements. Further, Tlasso and SG-PALM both incur a cost of $O(Nd_km_k^2)$ for each mode-wise update. This can also be reduced to be $\approx d$ for sparse estimated $\mathbf{\Omega}_k$'s at each iteration. Overall, for the sample-starved setting where we only have access to a handful of data samples, structured KP and KS models run similarly fast, while the Sylvester GM runs slower theoretically due to the extra and richer structures that it takes into account.

4.2. Numerical experiments with synthetic data

Here we compare the performance of the Kronecker covariance / precision estimation methods when used to estimate multiway physical processes, based on numerical experiments using synthetic data. To this end, we first describe two generative models that extend the spatial Poisson equation (cf. Section 2.3) to incorporate temporal dynamics, and the resulting multiway (inverse) covariance structure in Section 4.2.1. Thereafter, in Section 4.2.2, we illustrate the performance of the various (inverse) covariance estimation methods using synthetic data generated from the two models of spatio-temporal dynamics.

4.2.1. Two generative models for spatio-temporal dynamics

We describe two dynamic models that extend the spatial Poisson equation (cf. Section 2.3) to incorporate temporal dynamics, and the resulting multiway (inverse) covariance structure. These models will be used in our numerical experiments to generate data to compare the performance of estimation algorithms in Section 4.2.2.

Poisson-AR(1) process The first extension, which we call the Poisson-AR(1) process, imposes an autoregressive temporal model of order 1 on the source function f in the Poisson equation (2). Specifically, we say a sequence of discretized spatial observations $\{\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ indexed by time step $k = 1, \cdots, T$ is from a Poisson-AR(1) process if

$$(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}) \operatorname{vec}(\mathbf{U}^k) = \operatorname{vec}(\mathbf{Z}^k), \tag{12}$$

$$\operatorname{vec}(\mathbf{Z}^k) = a \operatorname{vec}(\mathbf{Z}^{k-1}) + \operatorname{vec}(\mathbf{W}^k), \quad |a| < 1, \tag{13}$$

where $\{\mathbf{W}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ is spatiotemporal white noise, i.e., $W_{i,j}^k \sim \mathcal{N}(0, \sigma_w^2)$, i.i.d. Assuming $\mathbf{Z}^0 = \mathbf{0}$ and defining the T-by-T matrix

$$\mathbf{B} = \begin{bmatrix} 1 & -a & & \\ & 1 & \ddots & \\ & & \ddots & -a \\ & & & 1 \end{bmatrix},$$

the above linear system of equations can be written as $(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{U}\mathbf{B} = \mathbf{W}$, or equivalently,

$$\left(\mathbf{B}^T \otimes (\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\right) \mathrm{vec}(\mathbf{U}) = \mathrm{vec}(\mathbf{W}), \tag{14}$$

where $\mathbf{U} = [\mathrm{vec}(\mathbf{U}^1)\,\mathrm{vec}(\mathbf{U}^2)\dots\mathrm{vec}(\mathbf{U}^T)] \in \mathbb{R}^{d_1 d_2 \times T}$ and $\mathbf{W}$ is defined likewise. The inverse covariance of $\mathbf{U}$, despite having a large size of $d_1 d_2 T \times d_1 d_2 T$, is sparse and has a mixed Kronecker sum and product structure.

Convection-diffusion process The second time-varying extension of the Poisson PDE model (2) is based on the convection-diffusion process [12]

$$\frac{\partial u}{\partial t} = \theta \sum_{i=1}^{2} \frac{\partial^2 u}{\partial x_i^2} - \epsilon \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}. \tag{15}$$

Here, $\theta > 0$ is the diffusivity; and $\epsilon \in \mathbb{R}$ is the convection velocity of the quantity along each coordinate. Note that for simplicity of discussion here, we assume these coefficients do not change with space and time (see, Stocker [74], for example, for a detailed discussion). These equations are closely related to the Navier-Stokes equation commonly used in stochastic modeling for weather and climate prediction [12, 74]. Coupled with Maxwell's equations, these equations can be used to model magneto-hydrodynamics [73], which characterize solar activities including flares.

A solution of Equation (15) can be approximated similarly as in the Poisson equation case, through a finite difference approach. Denote the discrete spatial samples of $u(\mathbf{x}, t)$ at time t_k as a matrix $\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}$. We obtain a discretized update propagating $u(\mathbf{x}, t)$ in space and time, which locally satisfies

$$\begin{aligned} \frac{U_{i,j}^k - U_{i,j}^{k-1}}{\Delta t} &= \theta \left(\frac{U_{i+1,j}^k + U_{i-1,j}^k + U_{i,j+1}^k + U_{i,j-1}^k - 4U_{i,j}^k}{h^2} \right) \\ &- \epsilon \left(\frac{U_{i+1,j}^k - U_{i-1,j}^k + U_{i,j+1}^k - U_{i,j-1}^k}{2h} \right), \end{aligned} \tag{16}$$

where $\Delta t = t_{k+1} - t_k$ is the time step and h is the mesh step (spatial grid spacing). Similarly to the Poisson-AR(1) process, in the following, we consider a "blocked" version of the convection-diffusion process.

We define the first-order and second-order discretized differential operators,

denote by **D** and **A**, respectively:

$$
\mathbf{D} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.
$$

Then, Equation (16) can be written as

$$
\begin{aligned}
\frac{1}{\Delta t}(\mathbf{D} \otimes \mathbf{I} \otimes \mathbf{I}) \operatorname{vec} \mathbf{U} = {} & \frac{\theta}{h^2}(\mathbf{I} \otimes \mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{A}) \operatorname{vec} \mathbf{U} \\
& - \frac{\epsilon}{2h}(\mathbf{I} \otimes \mathbf{D} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{D}) \operatorname{vec} \mathbf{U},
\end{aligned}
\tag{17}
$$

where $\mathbf{U} = [\operatorname{vec}(\mathbf{U}^1) \operatorname{vec}(\mathbf{U}^2) \dots \operatorname{vec}(\mathbf{U}^T)] \in \mathbb{R}^{d_1 d_2 \times T}$. Assuming the process is driven by some white noise **W**, similarly defined as in the Poisson-AR equation, the inverse covariance of **U** is again sparse and has a mixed Kronecker sum and product structure.

4.2.2. Experiments with synthetic data: Performance comparisons

We consider spatio-temporal processes (2D space + time) that evolve $T = 50$ time steps on a 8×8 spatial grid according to the Poisson-AR(1) and the Convection-Diffusion dynamics (cf. Section 4.2.1). We generated $N = 50$ independent realizations of each type of process. Figure 2 shows one realization for each type of process, with spatial resolution increased for better illustration. Assuming $K = 2$ with a spatial dimension 64 and a temporal dimension 50, we estimated the covariance / inverse covariance using several sparse multiway inverse covariance estimation methods including Glasso [24], KPCA (Section 3.2), Tlasso (Section 3.3), TeraLasso (Section 3.4), SG-PALM (Section 3.5). Note that none of the above-mentioned models operate under the true generative processes (i.e., there is a model mismatch with the data). Here, the sparsity-regularized methods are all implemented with an ℓ_1 penalty function, and the penalty parameters were selected similarly and guided by the theoretical results in Table 2. For example, for SG-PALM, we use a penalty parameter of $\lambda_k = C\sqrt{\frac{d_k \log d}{N}}$ where C is chosen by optimizing a normalized Frobenius norm error between the estimate and the truth, over a range of λ values parameterized by C. For the KPCA algorithm, both the nuclear norm penalty parameter and the separation rank are selected by optimizing a normalized Frobenius norm error via grid search.

A summary of the accuracy of estimation (normalized Frobenius norm error), as well as support recovery (Mathews Correlation Coefficient [62]), are reported in Table 3. In Figure 3 and Figure 4 we show the true and the estimated inverse covariance matrices obtained for all the methods except KPCA, under both the Poisson-AR (Figure 3) and the Convection-Diffusion processes (Figure 4).
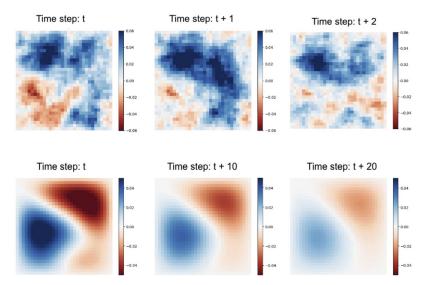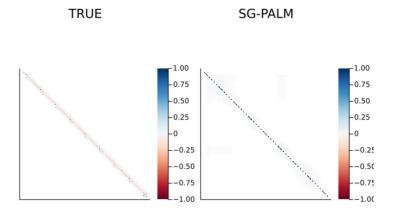
FIG 2. *Poisson-AR(1) (top) and 2D Convection-diffusion (bottom) state variables of size 64×64 at three different time steps, which are chosen such that t is sufficiently large for the processes to be in steady state and that the sampling times are adjusted to the temporal dynamics of each process. Here, model parameters used for simulations are: $a = -0.5$ for Poisson-AR, $\theta = 0.05$ and $\epsilon = 0.01$ for convection-diffusion process; the driving noise is generated from white Gaussian with $\sigma_w = 0.1$ in both cases.*
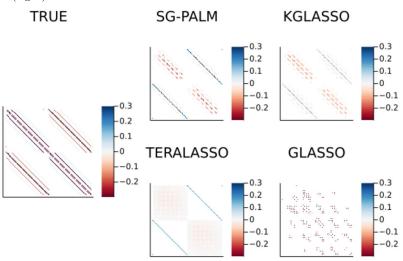
TABLE 3
Comparisons of performances measured by $\log\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F \backslash \|\boldsymbol{\Sigma}\|_F\right)$ for KPCA as well as

$\log\left(\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_F \backslash \|\boldsymbol{\Omega}\|_F\right)$ and the Mathews Correlation Coefficient (MCC) for SG-PALM, Tlasso, TeraLasso, Glasso. The MCC is a measure of the quality of sparsity recovery considered as a binary classification problem, where ± 1 indicates perfect agreement or disagreement between the truth and the estimation. Here the Frobenius norm errors are included in the first row under each generating type while the MCCs are in the second row. Note that the best performers under each type/criterion are highlighted.

Type	Metric	SG-PALM	KGlasso	TeraLasso	Glasso	KPCA
P-AR	Fnorm	**−0.2622**	1.1777	0.6312	0.9775	0.3289
	MCC	**0.4300**	0.3395	0.2061	0.0560	N/A
C-D	Fnorm	**−0.0420**	1.4919	−0.0208	2.2041	0.0642
	MCC	**0.2122**	0.1884	0.2018	0.0349	N/A

The inverse covariances under both generating processes admit structures with a mix of Kronecker sums and Kronecker products of sparse matrices. In both of the cases, the SG-PALM produces estimates having the lowest error. We believe this is due to the goodness of fit because the Sylvester graphical model imposes a squared KS structure on the precision matrix. Tlasso has comparable performances and achieves the best matrix approximation error under the convection-diffusion generating process. This might be due to the fact that the KP model corresponds to an underlying spatio-temporal autoregressive process.
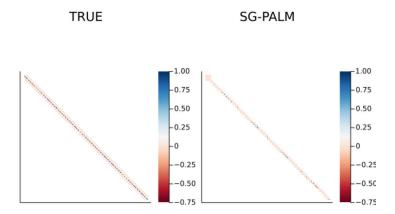
(a) Poisson-AR inverse covariance structure (left) and the estimate obtained by SG-PALM (right).
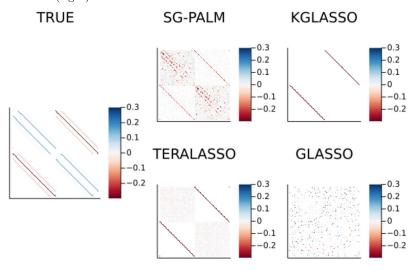

(b) Zoomed-in (middle 128 rows and columns) Poisson-AR inverse covariance structure (left) and the estimate obtained by SG-PALM, KGlasso, Glasso, TeraLasso (right, clockwise).

FIG 3. *Visualizations of the inverse covariance structures for Poisson-AR(1) dynamics and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.*
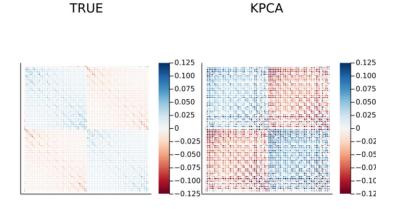
(a) Convection-diffusion inverse covariance structure (left) and the estimate obtained by SG-PALM (right).

(b) Zoomed-in (middle 128 rows and columns) convection-diffusion inverse covariance structure (left) and the estimate obtained by SG-PALM, KGlasso, Glasso, TeraLasso (right, clockwise).

FIG 4. *Visualizations of the inverse covariance structures for the Convection-Diffusion dynamics and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.*
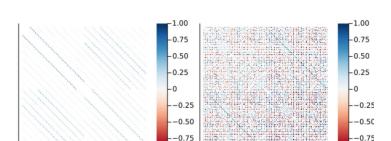
TeraLasso seems to produce the biggest model mismatch as indicated by the MCC scores. In Figure 5, we compare the true covariance matrices and the estimates obtained by KPCA. Here, although the KPCA model does not match

(a) Poisson-AR covariance structure (left) and the estimate obtained by KPCA (right).

(b) Convection-Diffusion covariance structure (left) and the estimate obtained by KPCA (right).

FIG 5. *Visualizations of the middle* 128 *rows and columns of the covariance structures for Poisson-AR(1) and Convection-Diffusion dynamics and their estimates, which show two temporal blocks of spatial correlation structures, each of size* 64 × 64, *with the diagonal elements removed for clearer visualization of the pattern. Here, white entries indicate zeros in the covariance matrices. Since the covariances are not sparse in general, all matrices are thresholded for clearer inspections of patterns.*

TABLE 4
*Average runtime and its standard error (in seconds) of estimating spatio-temporal (inverse)
covariance matrices of size $d \times 50$, where d is varying, using various algorithms.
Comparisons under various problem sizes (i.e., different d and N) are shown. Note the
sparse multiway precision models (SG-PALM, KGlasso, TeraLasso) are comparably fast and
are all faster than Glasso (for large problems) and KPCA.*

d	N	**Glasso**	**SG-PALM**	**TeraLasso**	**KGlasso**	**KronPCA**
		sec	**sec**	**sec**	**sec**	**sec**
8^2	25	0.40(0.20)	0.46(0.15)	0.15(0.35)	0.65(0.11)	37.22(0.20)
	50	0.48(0.21)	0.47(0.08)	0.22(0.50)	0.70(0.10)	38.22(0.55)
	100	0.76(0.05)	0.44(0.13)	0.26(0.28)	0.69(0.30)	39.09(1.05)
16^2	25	6.43(1.45)	3.37(1.09)	5.38(0.58)	5.14(1.99)	495.47(2.69)
	50	9.12(0.98)	3.27(1.81)	4.62(1.98)	3.39(2.00)	516.64(2.19)
	100	11.84(2.01)	4.85(1.10)	6.71(0.72)	5.67(0.57)	498.04(4.01)

the underlying generating processes, the estimates were able to capture certain blocking patterns that similarly exist in the true covariance.

Computational efficiencies of the various covariance/precision estimation algorithms are also vitally important in practice to facilitate real-time tracking of physical systems. Table 4 shows the runtime of different covariance and inverse covariance estimation algorithms for the synthetic experiments. It shows that by recognizing and exploiting multiway structures in the data, sparse multiway inverse covariance estimation methods, TeraLasso, Tlasso, and SG-PALM significantly reduce the runtime complexity of Glasso that ignores such special multiway structures. Remark that KPCA runs considerably slower than other methods as it involves expensive singular value decomposition of a large-dimensional re-arranged sample covariance matrix of the data.

4.3. Application to predicting evolution of solar active regions

Solar active regions are temporary centers of strong and complex magnetic fields on the sun, the principal source of violent eruptions such as solar flares [86]. While weak flares of, for example, B-class, have only limited terrestrial effect, strong flares of M- and X-class can produce tremendous amounts of electromagnetic radiation, causing disturbance or damage to satellites, power grids, and communication systems. Therefore, it would be of great value to be able to predict how active regions evolve before the onset of solar flares.

Although there are numerous studies that use active region images or physical parameters to predict flare activities [50, 14, 41, 93, 77], fewer studies have attempted to predict the complicated preflare evolution of active regions without physical modeling [4]. In this section, to demonstrate the application of the tensor graphical models, we use multiwavelength active region observations to predict the evolution of two types of active regions that lead to either a weak (B-class) flare or a strong (M- or X-class) flare [2].

---

[2]A similar and more comprehensive version of this numerical study was included in Wang and Hero [90].

We construct a multiwavelength active region video dataset from the curated dataset generated by Galvez et al. [27]. The video data are taken in four wavelengths (94Å, 131Å, 171Å, and 193Å) by the Atmospheric Imaging Assembly [AIA, 51] aboard the Solar Dynamics Observatory (SDO) satellite. Each video is a 24-hour image sequence of an active region at 1-hour cadence before a strong (M- or X-class) or a weak (B-class) flare occurs in the region. We spatially interpolate the videos so that each video is represented as a $d_1 \times d_2 \times d_3 \times d_4$ tensor, where $d_1 = 13$ denotes the number of frames in the video, $d_2 = 50$ denotes the height of the frames after interpolation, $d_3 = 100$ denotes the width of the frames after interpolation, and $d_4 = 4$ represents the number of different channels/wavelength at which the images are recorded. To prevent information leakage, we chronologically split the active region videos into a training set (year 2011 to 2014) and a test set (year 2011 to 2014). In the training set, there are 186 active region videos that lead to a B-class flare and 48 active region videos that lead to an M/X-class flare. In the test set, the sample sizes are 93 and 24 for the B-class and the M/X-class, respectively.

To perform active region prediction, we first fit the tensor graphical models on the training set to estimate the covariance or prediction matrices for each of the two types of active region videos, and then we use the best linear predictor to predict the last frame from all previous frames for videos in the test set. The forward linear predictor is constructed in a multi-output least squares regression setting as

$$\widehat{\mathbf{y}}_t = -\mathbf{\Omega}_{2,2}^{-1}\mathbf{\Omega}_{2,1}\mathbf{y}_{t-1:t-(p-1)} \tag{18}$$

when the precision estimate is available. Here, $t = d_1$ for predicting the last frame of a video. For notational convenience, let $p = d_1$ and $q = d_2d_3d_4$, then $\mathbf{y}_{t-1:t-(p-1)} = \mathbf{y}_{p-1:1} \in \mathbb{R}^{(p-1)q}$ is the stacked set of pixel values from the previous $p-1$ time instances and $\mathbf{\Omega}_{2,1} \in \mathbb{R}^{q \times (p-1)q}$ and $\mathbf{\Omega}_{2,2} \in \mathbb{R}^{q \times q}$ are submatrices of the $pq \times pq$ estimated precision matrix:

$$\widehat{\mathbf{\Omega}} = \begin{pmatrix} \mathbf{\Omega}_{1,1} & \mathbf{\Omega}_{1,2} \\ \mathbf{\Omega}_{2,1} & \mathbf{\Omega}_{2,2} \end{pmatrix}.$$

The tensor graphical models to be evaluated are SG-PALM (Section 3.5), Tlasso (Section 3.4), and TeraLasso (Section 3.3), with regularization parameters chosen by optimizing the prediction NRMSE (i.e., the root mean squared error normalized by the difference between maximum and minimum pixels) on the training set over a range of λ values guided by the convergence rates presented in Table 2. In particular, the SG-PALM estimator was implemented using a regularization parameter $\lambda_N = C_1\sqrt{\frac{\min(d_k)\log(d)}{N}}$ for all k with the constant C_1 chosen by optimizing the prediction NRMSE on the training set over a range of λ values parameterized by C_1. The TeraLasso estimator and the Tlasso estimator were implemented using $\lambda_{N,k} = C_2\sqrt{\frac{\log(d)}{N\prod_{i\neq k}d_i}}$ and $\lambda_{N,k} = C_3\sqrt{\frac{\log(d_k)}{Nd}}$ for $k = 1, 2, 3$, respectively, with C_2, C_3 optimized in a similar manner. Also considered is a baseline estimator, referred to as IndLasso, obtained by applying independent and separate ℓ_1-penalized regressions to each pixel in $\mathbf{y}_t$. Each

TABLE 5
Comparison of the SG-PALM, Tlasso, TeraLasso, IndLasso performances measured by NRMSE in predicting the last frame of 13-frame video sequences leading to B- and M/X-class solar flares. The NRMSEs are computed by averaging across testing samples and AIA channels for each pixel. SG-PALM achieves the best overall NRMSEs across pixels. B flares are generally easier to predict due to both a larger number of samples in the training set and typically smoother transitions from frame to frame within a video.

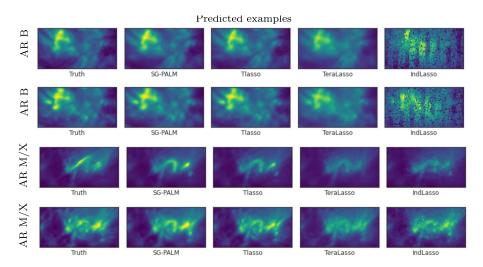| Flare Type	SG-PALM	Tlasso	TeraLasso	IndLasso
	NRMSE	NRMSE	NRMSE	NRMSE
B	0.0379	0.0386	0.0579	0.1628
M/X	0.0620	0.0790	0.0913	0.1172

FIG 6. *Examples of one-hour ahead prediction of the first two AIA channels of last frames of 13-frame videos, leading to B- (first two rows) and M/X-class (last two rows) flares, produced by the SG-PALM, Tlasso, TeraLasso, IndLasso algorithms, comparing to the ground truth image (far left column). The column ordering from left to right is in the order of increasing normalized mean squared error (NMSE) - see Table 5. Note that in general linear forward predictors tend to underestimate the contrast ratio of the images. The proposed SG-PALM produced the best-quality images in terms of both the spatial structures and contrast ratios. This figure is excerpted from Wang and Hero [90] (cf. Figure 3 in the reference).*

sparse regression in the IndLasso estimator was implemented and tuned independently with regularization parameters chosen in a similar manner. Note that we did not include Glasso and KPCA (Section 3.2) in the comparison due to their prohibitively expensive runtime complexity for large multiway data.

Table 5 shows the NRMSE over the testing samples for the forecasts based on SG-PALM, Tlasso, TeraLasso, and IndLasso. We observe that SG-PALM outperforms all three other methods, indicated by the lowest NRMSEs on the test set. Figure 6 depicts examples of predicted images, comparing with the ground truth. The SG-PALM estimates produced the most realistic image predictions that capture the spatially varying structures and closely approximate

the pixel values (i.e., maintaining contrast ratios). The latter is important as the flares are being classified into weak (B-class) and strong (M/X-class) categories based on the brightness of the images, and stronger flares are more likely to lead to catastrophic events, such as those that can lead to damaging terrestrial perturbations and health of astronauts.

5. Conclusion and discussion

Physical interpretability   While the Kronecker products expansion used in KPCA captures dense structures in the covariance matrix of data generated from more complex spatio-temporal physical processes, it lacks physical interpretability. In contrast to the case of Sylvester graphical model and Poisson-AR(1) processes, it is not obvious whether the sum of Kronecker products structure corresponds to any true physical models. Recent work in quantum informatics [15] has demonstrated a link between the estimation of the density matrix for entangled quantum states and the structured tensor approximation via $\sum_{i=1}^r \mathbf{A}_i \otimes \mathbf{B}_i$. Further characterizing and extending these connections to other classes of discretized PDEs is an interesting future direction. Furthermore, in both the Poisson-AR(1) and convection-diffusion examples, a mixed Kronecker sum and Kronecker product structure emerges that can be related to the state covariance of a dynamical system.

Heavy-tailed multiway covariance/precision models   Most existing work on multiway covariance and inverse covariance models focus on modeling Gaussian variables. It would be interesting to explore whether the pseudo-likelihood framework we adopted for SyGlasso and SG-PALM can be extended to non-Gaussian heavy-tailed models, e.g., using copula's or elliptically contoured distributions. This could have important practical applications, in particular to solar flare and active region prediction problems presented in Section 4.3. The images that characterize the active regions generally include a small number of pixels of extremely high intensity. These pixels might not be captured by a Gaussian-like distribution. Recently, there have been advances [94, 44] in covariance estimation for heavy-tailed, non-Gaussian vector-variate data. Multiway (inverse) covariance estimation is an open problem.

Kronecker-structured autoencoders   Low-rank covariance models have close connections with variational autoencoders (VAEs). Dai et al. [17] studied the relationship between (robust) PCA and VAEs. Since the Kronecker product for matrices is a generalization of the outer product for vectors, KPCA can be considered as a generalization of the low-rank approximation method of PCA. It is thus natural to exploit similar relationships between KPCA and VAEs. In this case VAE may be considered as a nonlinear/non-Gaussian extension to KPCA for low separation rank covariance models. Additionally, recent advances in efficient training of the VAE-type neural network architecture (e.g., using stochastic gradient descent) could improve the computational complexity of KPCA that is currently limited by an expensive singular value decomposition [84, 31].

Computational complexity Estimating the Kronecker product model is generally hard due to: (i) tensor factorization is NP-hard [35] when $K \geq 3$; and (ii) the optimization objective is nonconvex. The NP-hardness implies computational intractability for large-scale problems in the worst case. However, this does not necessarily preclude the existence of tractable solutions for most cases. Specifically, it is believed that the model can be estimated within polynomial time when the signal-to-noise ratio (SNR) is sufficiently high[3], but only when it is significantly higher than the information-theoretical threshold for model identifiability [72, 37]. While such studies have been completed for some important problems e,g, the statistical-computational gap [10, 103] and the optimization landscape [28], an analogous framework for Kronecker covariance remains to be developed.

While Kronecker sum models do not suffer from the nonconvexity of the Kronecker product model, scalable estimation methods that apply to very large covariance and precision matrices remain computationally challenging. Therefore, it would be worthwhile to develop improvements for the Kronecker sum that mirror the successes in related problems, e.g., introducing more amenable surrogate objectives [6, 45].

Software A software package called `TensorGraphicalModels` [92] has been developed to accompany this review article. Detailed code and instructions for reproducing the numerical results presented in this article (and beyond) are provided in the reference. The software provides a suite of tools in `Julia`, which was designed specifically for numerical and scientific computing and has recently become popular for large-scale scientific simulation and modeling, such as those conducted by climate scientists [68].

References

[1] AKDEMIR, D. and GUPTA, A. K. (2011). Array Variate Random Variables with Multiway Kronecker Delta Covariance Matrix Structure. *Journal of algebraic statistics* **2**.

[2] ALLEN, G. I. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics* **4** 764.

[3] ALLEN, G. I. and TIBSHIRANI, R. (2012). Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 721–743.

[4] BAI, L., BI, Y., YANG, B., HONG, J.-C., XU, Z., SHANG, Z.-H., LIU, H., JI, H.-S. and JI, K.-F. (2021). Predicting the evolution of photospheric magnetic field in solar active regions using deep learning. *Research in Astronomy and Astrophysics* **21** 113.

---

[3]For example, the convergence analysis of Tlasso [59] requires a 'good' initialization; a high SNR is required for a random initialization to satisfy the condition.

[5] Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research* **9** 485–516. MR2417243

[6] Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* 616–618.

[7] Bijma, F., De Munck, J. C. and Heethaar, R. M. (2005). The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. *NeuroImage* **27** 402–415.

[8] Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[9] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)* **58** 1–37. MR2811000

[10] Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences* **110** E1181–E1190.

[11] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization* **21** 572–596.

[12] Chandrasekhar, S. (1943). Stochastic problems in physics and astronomy. *Reviews of modern physics* **15** 1.

[13] Chen, Y., Wiesel, A. and Hero, A. O. (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing* **59** 4097–4107.

[14] Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., DuFumier, B., Zhou, T., Wang, X., Zhu, H., Sun, Z. and Gombosi, T. I. (2019). Identifying solar flare precursors using time series of SDO/HMI images and SHARP parameters. *Space Weather* **17** 1404–1426.

[15] Chu, M. T. and Lin, M. M. (2021). Nonlinear Power-Like and SVD-Like Iterative Schemes with Applications to Entangled Bipartite Rank-1 Approximation. *SIAM Journal on Scientific Computing* **0** S448–S474. MR4331954

[16] Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

[17] Dai, B., Wang, Y., Aston, J., Hua, G. and Wipf, D. (2018). Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research* **19** 1573–1614.

[18] Dantas, C. F., Cohen, J. E. and Gribonval, R. (2019). Learning tensor-structured dictionaries with application to hyperspectral image denoising. In *2019 27th European Signal Processing Conference (EUSIPCO)* 1–5. IEEE.

[19] Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68** 265–274.

[20] Deng, Y., Tang, X. and Qu, A. (2021). Correlation Tensor Decomposition and Its Application in Spatial Imaging Data. *Journal of the American Statistical Association* **just-accepted** 1–35.

[21] Du, J., Yan, X., Liu, Z., Cui, L., Ding, P., Tan, X., Li, X., Zhou, H., Gu, Q. and Xu, J. (2017). cBinderDB: a covalent binding agent database. *Bioinformatics* **33** 1258–1260.

[22] Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of statistical computation and simulation* **64** 105–123.

[23] Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99** 10143–10162.

[24] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[25] Fu, Y. and Huang, T. S. (2008). Image classification using correlation tensor analysis. *IEEE Transactions on Image Processing* **17** 226–234. MR2446011

[26] Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods* **23** 3105–3119.

[27] Galvez, R., Fouhey, D. F., Jin, M., Szenicer, A., Muñoz-Jaramillo, A., Cheung, M. C., Wright, P. J., Bobra, M. G., Liu, Y., Mason, J. et al. (2019). A machine-learning data set prepared from the NASA solar dynamics observatory mission. *The Astrophysical Journal Supplement Series* **242** 7.

[28] Ge, R. and Ma, T. (2020). On the optimization landscape of tensor decompositions. *Mathematical Programming* 1–47.

[29] Grasedyck, L. (2004). Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing* **72** 247–265.

[30] Greenewald, K. H. and Hero, A. O. (2014). Kronecker PCA based spatio-temporal modeling of video for dismount classification. In *Algorithms for Synthetic Aperture Radar Imagery XXI* **9093** 90930V. International Society for Optics and Photonics.

[31] Greenewald, K. and Hero, A. O. (2015). Robust kronecker product PCA for spatio-temporal covariance estimation. *IEEE Transactions on Signal Processing* **63** 6368–6378.

[32] Greenewald, K., Zelnio, E. and Hero, A. H. (2016). Robust SAR STAP via Kronecker decomposition. *IEEE Transactions on Aerospace and Electronic Systems* **52** 2612–2625.

[33] Greenewald, K., Zhou, S. and Hero, A. (2019). Tensor graphical lasso (TeraLasso). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81** 901–931.

[34] Hao, B., Wang, B., Wang, P., Zhang, J., Yang, J. and Sun, W. W. (2021). Sparse tensor additive regression. *Journal of Machine Learning Research* **22** 1–43.

[35] Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)* **60** 1–39. MR3144915

[36] Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis* **6**

179–196.

[37] Hopkins, S. B., Shi, J. and Steurer, D. (2015). Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory* 956–1006. PMLR.

[38] Hori, T., Montcho, D., Agbangla, C., Ebana, K., Futakuchi, K. and Iwata, H. (2016). Multi-task Gaussian process for imputing missing data in multi-trait and multi-environment trials. *Theoretical and Applied Genetics* **129** 2101–2115.

[39] Hou, E., Lawrence, E. and Hero, A. O. (2021). Penalized ensemble Kalman filters for high dimensional non-linear systems. *PloS one* **16** e0248046.

[40] Jiang, Y., Bigot, J. and Maabout, S. (2021). Online Graph Topology Learning from Matrix-valued Time Series. *arXiv preprint arXiv:2107.08020*.

[41] Jiao, Z., Sun, H., Wang, X., Manchester, W., Gombosi, T., Hero, A. and Chen, Y. (2020). Solar flare intensity prediction with machine learning models. *Space Weather* **18** e2020SW002440.

[42] Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis* 129–155. Springer.

[43] Kalaitzis, A., Lafferty, J., Lawrence, N. D. and Zhou, S. (2013). The bigraphical lasso. In *International Conference on Machine Learning* 1229–1237. PMLR.

[44] Ke, Y., Minsker, S., Ren, Z., Sun, Q. and Zhou, W.-X. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science* **34** 454–471.

[45] Khare, K., Oh, S.-Y. and Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 803–825. MR3382598

[46] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.

[47] Kressner, D. and Tobler, C. (2010). Krylov subspace methods for linear systems with tensor product structure. *SIAM journal on matrix analysis and applications* **31** 1688–1714.

[48] Landsberg, J. (2011). *Tensors: Geometry and Applications: Geometry and Applications* **128**. American Mathematical Soc.

[49] Ledoit, O. and Wolf, M. (2020). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*.

[50] Leka, K. and Barnes, G. (2003). Photospheric magnetic field properties of flaring versus flare-quiet active regions. II. Discriminant analysis. *The Astrophysical Journal* **595** 1296.

[51] Lemen, J. R., Akin, D. J., Boerner, P. F., Chou, C., Drake, J. F., Duncan, D. W., Edwards, C. G., Friedlaender, F. M., Heyman, G. F., Hurlburt, N. E. et al. (2011). The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO). In *The solar*

dynamics observatory 17–40. Springer.

[52] LENG, C. and PAN, G. (2018). Covariance estimation via sparse Kronecker structures. *Bernoulli* **24** 3833–3863.

[53] LI, L. and ZHANG, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112** 1131–1146.

[54] LI, Z., CHAO, Y., MCWILLIAMS, J. C. and IDE, K. (2008). A three-dimensional variational data assimilation scheme for the regional ocean modeling system. *Journal of Atmospheric and Oceanic Technology* **25** 2074–2090.

[55] LI, S., LÓPEZ-GARCÍA, M., LAWRENCE, N. D. and CUTILLO, L. (2022). Scalable Bigraphical Lasso: Two-way Sparse Network Inference for Count Data. *arXiv preprint arXiv:*2203.07912.

[56] LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 423–498. MR2853727

[57] LIU, G., TAN, X., DANG, C., TAN, S., XING, S., HUANG, N., PENG, K., XIE, C., TANG, X. and ZENG, J. (2019). Regional shape abnormalities in thalamus and verbal memory impairment after subcortical infarction. *Neurorehabilitation and Neural Repair* **33** 476–485.

[58] LLOSA-VITE, C. and MAITRA, R. (2022). Reduced-Rank Tensor-on-Tensor Regression and Tensor-variate Analysis of Variance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

[59] LYU, X., SUN, W. W., WANG, Z., LIU, H., YANG, J. and CHENG, G. (2019). Tensor graphical model: Non-convex optimization and statistical inference. *IEEE transactions on pattern analysis and machine intelligence* **42** 2024–2037.

[60] MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics* **239** 37–49.

[61] MARDIA, K. V. and GOODALL, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate environmental statistics* **6** 347–385. MR1268438

[62] MATTHEWS, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405** 442–451.

[63] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics* **34** 1436–1462. MR2278363

[64] MIN, K., MAI, Q. and ZHANG, X. (2022). Fast and separable estimation in high-dimensional tensor Gaussian graphical models. *Journal of Computational and Graphical Statistics* **31** 294–300. MR4387228

[65] MO, S., XIA, J., TAN, X. and RAJ, B. (2022). Point3D: tracking actions as moving points with 3D CNNs. *arXiv preprint arXiv:*2203.10584.

[66] MOLSTAD, A. J. and ROTHMAN, A. J. (2019). A penalized likelihood

method for classification with matrix-valued predictors. *Journal of Computational and Graphical Statistics* **28** 11–22.

[67] OHLSON, M., AHMAD, M. R. and VON ROSEN, D. (2013). The multilinear normal distribution: Introduction and some basic properties. *Journal of Multivariate Analysis* **113** 37–47.

[68] PERKEL, J. M. (2019). Julia: come for the syntax, stay for the speed. *Nature* **572** 141–143.

[69] POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science* **26** 369–387.

[70] POURYAZDIAN, S., BEHESHTI, S. and KRISHNAN, S. (2016). CANDE-COMP/PARAFAC model order selection based on reconstruction error in the presence of kronecker structured colored noise. *Digital Signal Processing* **48** 12–26.

[71] RABUSSEAU, G. and KADRI, H. (2016). Low-rank regression with tensor responses. *Advances in Neural Information Processing Systems* **29** 1867–1875.

[72] RICHARD, E. and MONTANARI, A. (2014). A statistical model for tensor PCA. *Advances in Neural Information Processing Systems* **27** 2897–2905.

[73] ROBERTS, B. (2006). Slow magnetohydrodynamic waves in the solar atmosphere. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **364** 447–460.

[74] STOCKER, T. (2011). *Introduction to climate modelling.* Springer Science & Business Media.

[75] STROBACH, P. (1995). Low-rank detection of multichannel Gaussian signals using block matrix approximation. *IEEE Transactions on Signal Processing* **43** 233–242.

[76] SUN, W. W., HAO, B. and LI, L. Tensors in Modern Statistical Learning.

[77] SUN, Z., BOBRA, M., WANG, X., WANG, Y., SUN, H., GOMBOSI, T., CHEN, Y. and HERO, A. (2021). Predicting Solar Flares using CNN and LSTM on Two Solar Cycles of Active Region Data. *Earth and Space Science Open Archive* 32.

[78] TAN, X., ROSS, C. A., MILLER, M. and TANG, X. (2018). Changepoint analysis of putamen and thalamus subregions in premanifest huntington's disease. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 531–535. IEEE.

[79] TAN, X., QI, Z., SEYMOUR, C. W. and TANG, L. (2022a). RISE: Robust Individualized Decision Learning with Sensitive Variables. *arXiv preprint arXiv:*2211.06569.

[80] TAN, X., ABBERBOCK, J., RASTOGI, P. and TANG, G. (2022b). Identifying Principal Stratum Causal Effects Conditional on a Post-treatment Intermediate Response. In *Conference on Causal Learning and Reasoning* 734–753. PMLR.

[81] TAN, X., YANG, S., YE, W., FARIES, D. E., LIPKOVICH, I. and KADZIOLA, Z. (2022c). When Doubly Robust Methods Meet Machine Learning for Estimating Treatment Effects from Real-World Data: A Comparative Study. *arXiv preprint arXiv:*2204.10969.

[82] TAN, X., CHANG, C.-C. H., ZHOU, L. and TANG, L. (2022d). A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources. In *International Conference on Machine Learning* 21013–21036. PMLR.

[83] TENG, S. L. and HUANG, H. (2009). A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments. *Journal of the American Statistical Association* **104** 465–473. MR2751431

[84] TSILIGKARIDIS, T. and HERO, A. O. (2013). Covariance estimation in high dimensions via kronecker product expansions. *IEEE Transactions on Signal Processing* **61** 5347–5360.

[85] TSILIGKARIDIS, T., HERO, A. O. and ZHOU, S. (2013). On convergence of kronecker graphical lasso algorithms. *IEEE transactions on signal processing* **61** 1743–1755.

[86] VAN DRIEL-GESZTELYI, L. and GREEN, L. M. (2015). Evolution of active regions. *Living Reviews in Solar Physics* **12** 1–98.

[87] VAN LOAN, C. F. and PITSIANIS, N. (1993). Approximation with Kronecker products. In *Linear algebra for large scale and real-time applications* 293–314. Springer.

[88] VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* 5–42.

[89] WANG, Y. and HERO, A. (2021a). Multiway Ensemble Kalman Filter. *arXiv preprint arXiv:*2112.04322.

[90] WANG, Y. and HERO, A. (2021b). SG-PALM: a Fast Physically Interpretable Tensor Graphical Model. *International Conference on Machine Learning (ICML), arXiv preprint arXiv:*2105.12271.

[91] WANG, Y., JANG, B. and HERO, A. (2020). The sylvester graphical lasso (syglasso). In *International Conference on Artificial Intelligence and Statistics* 1943–1953. PMLR.

[92] WANG, Y., SUN, Z. and HERO, A. (2022). TensorGraphicalModels: A Julia toolbox for multiway covariance models and ensemble Kalman filter. *Software Impacts* **13** 100308.

[93] WANG, X., CHEN, Y., TOTH, G., MANCHESTER, W. B., GOMBOSI, T. I., HERO, A. O., JIAO, Z., SUN, H., JIN, M. and LIU, Y. (2020). Predicting solar flares with machine learning: investigating solar cycle dependence. *The Astrophysical Journal* **895** 3.

[94] WEI, X. and MINSKER, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems* (I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN and R. GARNETT, eds.) **30**. Curran Associates, Inc.

[95] WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing* **56** 478–491.

[96] WIESEL, A., ZHANG, T. et al. (2015). Structured robust covariance estimation. *Foundations and Trends® in Signal Processing* **8** 127–216.

[97] Xu, P., Zhang, T. and Gu, Q. (2017). Efficient algorithm for sparse tensor-variate gaussian graphical models via gradient descent. In *Artificial Intelligence and Statistics* 923–932. PMLR.

[98] Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of multivariate analysis* **107** 119–140.

[99] Yoon, J. H. and Kim, S. (2020). EiGLasso: Scalable estimation of Cartesian product of sparse inverse covariance matrices. In *Conference on Uncertainty in Artificial Intelligence* 1248–1257. PMLR.

[100] Yu, K., Bengtsson, M., Ottersten, B., McNamara, D., Karlsson, P. and Beach, M. (2001). Second order statistics of NLOS indoor MIMO channels based on 5.2 GHz measurements. In *GLOBECOM'01. IEEE Global Telecommunications Conference (Cat. No. 01CH37270)* **1** 156–160. IEEE.

[101] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

[102] Zhang, Y. and Schneider, J. G. (2010). Learning multiple tasks with a sparse matrix-normal penalty. In *NIPS* **6** 2.

[103] Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory* **64** 7311–7338.

[104] Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics* **42** 532–562.

[105] Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108** 540–552. MR3174640

[106] Zhou, J., Sun, W. W., Zhang, J. and Li, L. (2021). Partially observed dynamic tensor response regression. *Journal of the American Statistical Association* 1–16.