ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 5, *pp.* 901–931

Tensor graphical lasso (TeraLasso)

Kristjan Greenewald,

IBM Research, Cambridge, USA

Shuheng Zhou

University of California, Riverside, USA

and Alfred Hero III

University of Michigan, Ann Arbor, USA

[Received May 2017. Final revision August 2019]

Summary. The paper introduces a multiway tensor generalization of the bigraphical lasso which uses a two-way sparse Kronecker sum multivariate normal model for the precision matrix to model parsimoniously conditional dependence relationships of matrix variate data based on the Cartesian product of graphs. We call this tensor graphical lasso generalization TeraLasso. We demonstrate by using theory and examples that the TeraLasso model can be accurately and scalably estimated from very limited data samples of high dimensional variables with multiway co-ordinates such as space, time and replicates. Statistical consistency and statistical rates of convergence are established for both the bigraphical lasso and TeraLasso estimators of the precision matrix and estimators of its support (non-sparsity) set respectively. We propose a scalable composite gradient descent algorithm and analyse the computational convergence rate, showing that the composite gradient descent algorithm is guaranteed to converge at a geometric rate to the global minimizer of the TeraLasso objective function. Finally, we illustrate TeraLasso by using both simulation and experimental data from a meteorological data set, showing that we can accurately estimate precision matrices and recover meaningful conditional dependence graphs from high dimensional complex data sets.

Keywords: Convergence guarantees; Covariance modelling for array-valued data; Kronecker sum; Non-separable factor models; Precision matrix estimation; Sparsity

1. Introduction

The increasing availability of matrix and tensor-valued data with complex dependences has fed the fields of statistics and machine learning. Examples of tensor-valued data include medical and radar imaging modalities, spatial and meteorological data collected from sensor networks and weather stations over time, and biological, neuroscience and spatial gene expression data aggregated over trials and time points. Learning useful structures from these large-scale, complex and high dimensional data in the low sample regime is an important task in statistical machine learning, biology and signal processing.

As the precision matrix (inverse covariance matrix) encodes interactions and, for tensor-valued Gaussian distributions, conditional independence relationships between and among variables, multivariate statistical models, such as the matrix normal model (Dawid, 1981), have been proposed for estimation of these matrices. However, the number of parameters of the

Address for correspondence: Kristjan Greenewald, MIT–IBM Watson AI Laboratory, 75 Binney Street, Cambridge, MA 02142, USA.
E-mail: kristjan.greenewald@gmail.com

1369–7412/19/81901

precision matrix of a K-way data tensor $X \in \mathbb{R}^{d_1 \times \ldots \times d_K}$ grows as $\Pi_{i=1}^K d_i^2$. Therefore in high dimensions unstructured precision matrix estimation is impractical, requiring very large sample sizes. Undirected graphs are often used to describe high dimensional distributions. Under sparsity conditions, the graph can be estimated by using l_1-penalization methods, such as the graphical lasso (Friedman *et al.*, 2008) and multiple (nodewise) regressions (Meinshausen and Bühlmann, 2006). Under suitable conditions, such approaches yield consistent (and sparse) estimation in terms of graphical structure and fast convergence rates with respect to the operator and Frobenius norm for the covariance matrix and its inverse. However, many of the statistical models that have been considered still tended to be overly simplistic and not fully reflective of reality. For example, in neuroscience one must take into account temporal correlations as well as spatial correlations, which reflect the connectivity that is formed by the neural pathways. Yet, the line of high dimensional statistical literature mentioned above has primarily focused on estimating linear or graphical models with independent and identically distributed samples. In the case of graphical models, the data matrix is usually assumed to have independent rows or columns that follow the same distribution. The independence assumptions substantially simplify mathematical derivations but they tend to be very restrictive. For instance, cortical circuits can change over time because of activities such as motor learning, attention or visual stimulation. These data typically have a complex structure that is organized by the experimental design, with one or more experimental factors varying according to a predefined pattern.

On the theoretical and methodological front, recent work demonstrated another regime where further reductions in the sample size are possible under additional structural assumptions on the conditional dependence graphs which arise naturally in the above-mentioned contexts when handling data with complex dependences. For example, the matrix normal model as studied in Tsiligkaridis *et al.* (2013) and Zhou (2014) restricts the topology of the graph to tensor product graphs where the precision matrix corresponds to a Kronecker product representation. Moreover, Zhou (2014) showed that one can estimate the covariance and inverse covariance matrices well by using only one instance from the matrix variate normal distribution. Along the same lines, the bigraphical lasso framework was proposed to model parsimoniously conditional dependence relationships of matrix variate data based on the Cartesian product of graphs (Kalaitzis *et al.*, 2013) as opposed to the direct product graphs of the matrix normal models above. These models naturally generalize to multilinear settings with more than two axes of structure as demonstrated in the present work. The present work addresses the problem of sparse modelling of a structured precision matrix for tensor-valued data; more precisely, we aim to estimate the structure and parameters for a class of Gaussian graphical models by restricting the topology to the class of Cartesian product graphs, with precision matrices represented by a Kronecker sum for data with complex dependences.

Towards these goals, we shall introduce the tensor graphical lasso procedure called 'Tera-Lasso' for estimating sparse K-way decomposable precision matrices. We shall show that our concentration-of-measure analysis enables a significant reduction in the sample size requirement to estimate parameters and the associated conditional dependence graphs along different co-ordinates such as space, time and experimental conditions. We establish consistency for both the bigraphical lasso and tensor graphical lasso estimators and obtain optimal rates of convergence in the operator and Frobenius norm for estimating the associated precision matrix, and for structure recovery. Finally, we demonstrate by using simulations and real data that the Kronecker sum precision model has excellent potential for improving computational scalability, structural interpretation and its applications to classification, prediction and visualization for complex data analysis.

A philosophical motivation of TeraLasso's Kronecker sum (Cartesian graph) model is that it achieves the maximum entropy among all models for which the tensor component projections of the covariance matrix are fixed; see Section 3. A compelling justification for the proposed Kronecker sum model for the precision matrix is that similar models have been successfully used in other fields, including regularization of multivariate splines, design of physical networks and decomposition of solutions of partial differential equations governing many physical processes. Additional discussion of these practical motivations for the model is in Section 1.3 below.

1.1. The multiway Kronecker sum precision matrix model

We follow the notation and terminology of Kolda and Bader (2009) for modelling tensor-valued data arrays. Define the vector of component dimensions $\mathbf{p}=(d_1,\ldots,d_K)$ and let p denote the product of these dimensions:

$$p=\prod_{k=1}^{K} d_k$$

and

$$m_k=\prod_{i\neq k} d_i=p/d_k.$$

To simplify the multiway Kronecker notation, we define

$$I_{[d_{k:l}]}=\underbrace{I_{d_k}\otimes\ldots\otimes I_{d_l}}_{l-k+1\text{ factors}}$$

where '$\otimes$' denotes the Kronecker (direct) product and $l\geqslant k$. Using this notation, the K-way Kronecker sum of matrix components $\{\Psi_k\}_{k=1}^{K}$ can be written as

$$\Psi_1\oplus\ldots\oplus\Psi_K=\sum_{k=1}^{K} I_{[d_{1:k-1}]}\otimes\Psi_k\otimes I_{[d_{k+1:K}]}. \qquad (1)$$

In the special case of $K=2$ this Kronecker sum representation reduces to the more familiar $\Psi_1\oplus\Psi_2=\Psi_1\otimes I_{d_1}+I_{d_2}\otimes\Psi_2$. The vectorization of a K-way tensor X is denoted as vec(X) and is defined as in Kolda and Bader (2009). Likewise, we define the transpose of a K-way tensor $X^{\mathrm{T}}\in\mathbb{R}^{d_K\times\ldots\times d_1}$ analogously to the matrix transpose, i.e. $(X^{\mathrm{T}})_{i_1,\ldots,i_K}=X_{i_K,\ldots,i_1}$.

When the precision matrix Ω has a decomposition of the form (1), the Kronecker sum components $\{\Psi_k\}_{k=1}^{K}$ are sparse and the K-way data X have a multivariate Gaussian distribution, the sparsity pattern of Ψ_k corresponds to a conditional independence graph across the kth dimension of the data.

Fig. 1 illustrates the Kronecker sum model proposed in equation (1) for $K=3$ and $d_k=4$. Specifically, Ψ_k, $k=1,2,3$, are identical 4×4 tridiagonal precision matrices corresponding to a one-dimensional auto-regressive (AR(1)) process. The precision matrix $\Omega=\Psi_1\oplus\Psi_2\oplus\Psi_3$ is shown in Fig. 1(a) and the covariance $\Sigma=\Omega^{-1}$ in Fig. 1(b). The entries of each Ψ_k are replicated $m_k=16$ times across Ω for each k. This regular structure permits the aggregation of corresponding entries in the sample covariance matrix, resulting in variance reduction in estimating Ω. This Kronecker sum gives Ω a non-separable and interlocking repeating block structure in the covariance matrix.

We propose the following sparse Kronecker sum estimator of the precision matrix Ω in equation (1), which we call the tensor graphical lasso, TeraLasso. TeraLasso minimizes the negative l_1-penalized Gaussian log-likelihood function over the domain $\mathcal{K}_{\mathbf{p}}^{\sharp}$ of precision matrices Ω having

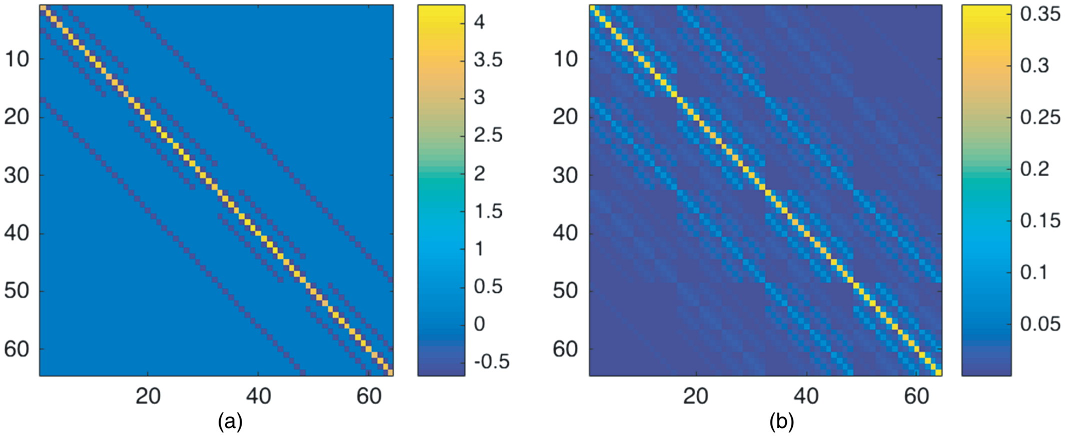


Fig. 1. Illustration of the Kronecker sum model for a tensor-valued AR(1) process (unlike the Kronecker product precision model, the nested block structure in Σ is not representable by a product of component factors): (a) sparse 4 × 4 × 4 precision matrix $\Omega = \Psi_1 \oplus \Psi_2 \oplus \Psi_3$, where Ψ_k are identical tridiagonal precision matrices corresponding to one-dimensional AR(1) models; (b) covariance matrix $\Sigma = \Omega^{-1}$

Kronecker sum form

$$\hat{\Omega} = \arg \min_{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}, \|\Omega\|_2 \leqslant \kappa} \left\{ -\log|\Omega| + \langle \hat{S}, \Omega \rangle + \sum_{k=1}^{K} m_k \sum_{i \neq j} g_{\rho_k}([\Psi_k]_{ij}) \right\} \tag{2}$$

where

$$\hat{S} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{vec}(X_i^{\mathrm{T}}) \mathrm{vec}(X_i^{\mathrm{T}})^{\mathrm{T}}, \tag{3}$$

$g_\rho(t)$ is a sparsity inducing regularization function parameterized by a regularization parameter ρ and

$$\mathcal{K}_{\mathbf{p}}^{\sharp} = \{A \succeq 0 : \exists\, B_k \in \mathbb{R}^{d_k \times d_k} \text{ subject to } A = B_1 \oplus \ldots \oplus B_K\} \tag{4}$$

is the set of positive semidefinite matrices that are decomposable into a Kronecker sum of fixed factor dimensions $\mathbf{p} = (d_1, \ldots, d_K)$. In this paper we consider (μ, γ) amenable regularizers g_ρ (Loh and Wainwright, 2017). The norm constraint $\|\Omega\|_2 \leqslant \kappa$ is required for the solution to be well defined when g_ρ is not a convex penalty. These penalties include non-convex regularizers such as smoothly clipped absolute deviation (SCAD) and the minimax convex penalty (MCP), as well as the traditional l_1-regularizer $g_\rho(t) = \rho|t|$.

Observe that sparsity in the off-diagonal elements of Ψ_k directly creates sparsity in Ω. As in the graphical lasso, incorporating an l_1-penalty over entries of Ω with the tensor-valued Gaussian or matrix normal (pseudo)log-likelihood promotes a sparse graphical structure in Ω; see for example Banerjee *et al.* (2008), Yuan and Lin (2007), Zhou (2014) and Zhou *et al.* (2011). In this work, we allow for the more general case of non-convex regularization functions g_ρ as considered in Loh and Wainwright (2017). Although sometimes difficult to tune in practice, non-convex regularization provides strong non-asymptotic guarantees on the elementwise estimation error of Ω, implying strong, single-sample support recovery guarantees when the smallest non-zero element of Ω is bounded from below.

The contributions of this paper are as follows. The sparse multivariate normal bigraphical lasso model called 'BiGLasso' is extended to the sparse tensor variate ($K > 2$) TeraLasso, allowing the modelling of data with arbitrary tensor degree K. A new sub-Gaussian concentration inequality (corollary 19 in the on-line supplement) is presented that gives rates of statistical convergence (theorems 1–3) of the TeraLasso estimator as well as the BiGLasso estimator, when the sample size is low (even equal to 1). TeraLasso's generalization of BiGlasso from two-way to K-way decompositions is important as it expands the domain of application, allowing a data scientist to group variables into their natural domains along multiple tensor axes. For example, with a health data set that is collected over space, time, people and replicates, TeraLasso's three-way tensor decomposition (time×space×people) can account for possible dependence structure between people, whereas a two-way BiGLasso or KLasso (Tsiligkaridis *et al.*, 2013) approach decomposing over (time×space) would unnecessarily enforce an assumption of independence between people. Alternatively, BiGLasso or KLasso could group two axes together (e.g. (time×space)×people); however, this would create a large unstructured factor whose estimation would require many more replicates than the three-way decomposition that TeraLasso uses to give each axis its own factor.

A highly scalable, first-order algorithm based on iterative soft thresholding is proposed to minimize the TeraLasso objective function. We prove (theorem 25 in the on-line supplement) that it converges to the global optimum with a geometric convergence rate, and we demonstrate its practical advantages on high dimensional problems. Compared with the alternating block co-ordinate descent algorithm that was proposed by Kalaitzis *et al.* (2013) for BiGLasso, the proposed iterative soft thresholding algorithm enjoys a per-iteration computational speed-up over BiGLasso of order $\Theta(p)$. Our numerical results show that the BiGLasso algorithm often requires many more iterations to converge than does our iterative soft thresholding method. Numerical comparisons are presented demonstrating that TeraLasso significantly improves performance in small sample regimes. To demonstrate the application of TeraLasso to real world data we use it to estimate the precision matrix of spatiotemporal meteorological data collected by the National Center for Environmental Prediction. Our results show that the TeraLasso precision matrix estimator degrades much more slowly than other estimators as we reduce the number of samples that are available to fit the model. The intuitive graphical structure, the robust eigenstructure and a maximum entropy interpretation make the TeraLasso model a compelling choice for modelling tensor data, much as the bigraphical lasso provides a meaningful alternative to the matrix normal model.

1.2. Relevant prior work

The use of tensor product models for multiway data has a long history. In the statistical context, directly fitting a Kronecker product to multiway data yields a first-order approximation corresponding to fitting the mean (Kolda and Bader, 2009) when the fitting criterion is the Frobenius norm of the residuals. Many such methods involve low rank factor decompositions including parallel factor analysis and CANDECOMP as in Harshman and Lundy (1994) and Faber *et al.* (2003), Tucker decomposition-based methods such as Tucker (1966) and Hoff (2016), and hybrid methods such as Johndrow *et al.* (2017). In contrast, second-order methods have been used to approximate multiway structure of the covariance (Werner *et al.*, 2008; Pouryazdian *et al.*, 2016). Series decomposition methods have been proposed for fitting the covariance matrix in Frobenius norm by using sums of Kronecker products (Tsiligkaridis and Hero, 2013; Greenewald and Hero, 2015; Rudelson and Zhou, 2017; Greenewald *et al.*, 2017).

Kronecker product approximations to the inverse covariance have fitted matrix normal models (Allen and Tibshirani, 2010) and sparse matrix normal models (Leng and Tang, 2012; Zhou, 2014; Tsiligkaridis *et al.*, 2013). In contrast with the Kronecker sum model (1) for the precision matrix Ω, the K-way Kronecker product model is $\Omega = \Psi_1 \otimes \ldots \otimes \Psi_K$. The Kronecker product decomposition implies a separable property of the precision matrix across the K data dimensions, which we might expect to become an increasingly restrictive condition as K increases. In this paper we show that the proposed Kronecker sum model (1) can be a worthwhile alternative representation.

A two-factor ($K=2$) sparse Kronecker sum model for the precision matrix Ω was introduced and studied in Kalaitzis *et al.* (2013). The model was fitted to the sample covariance matrix by using an iterative procedure called BiGlasso, which required the diagonal entries of Ω to be known. Conditions guaranteeing convergence were not provided. Here we extend the BiGlasso model to arbitrary $K \geqslant 2$ and unknown diagonal entries of Ω, provide a faster converging optimization algorithm and obtain strong convergence guarantees and bounds on the convergence rate for all K, including $K=2$. For completeness, we also obtain (appendix B of the on-line supplement) bounds on the convergence rate for the known diagonal setting of Kalaitzis *et al.* (2013).

The qualitative differences between the Kronecker product and Kronecker sum models for the precision matrix can be better appreciated by considering the product graphs that are induced by them. For given sparse Kronecker factors $\Psi_1, \ldots, \Psi_K$, the Kronecker product model corresponds to the direct (tensor) product of the component graphs whereas the Kronecker sum model corresponds to the Cartesian product of these components (Hammack *et al.*, 2011). (The Cartesian product of two graphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$ is a graph with vertices being the Cartesian product of V_1 and V_2, and with edges such that node (u,u') is adjacent to (v,v') if and only if either $u=v$ and u' is adjacent to v' in G_2, or $u'=v'$ and u is adjacent to v in G_1.) The direct product graph and Cartesian product graph differ greatly; the former has a number of edges equal to

$$\frac{1}{2}\prod_{k=1}^{K}(2|E_k|+|V_k|)-\prod_{k=1}^{K}|V_k|,$$

whereas the latter has a number of edges equal to

$$\sum_{k=1}^{K}|E_k|\prod_{i\neq k}|V_i|,$$

where V_i and E_i denote the node and edge sets of the ith component graph. (The notation $|V_i|=d_i$ denotes the row dimension of Ψ_i and $|E_i|$ denotes the number of non-zero upper triangular entries of Ψ_i.) To illustrate, if the number of non-zero entries of Ψ_k is cd_i for some c, the number of edges that are induced in the direct product graph by inserting a single new edge into the first component graph is equal to $\frac{1}{2}(2c+1)^K p/d_1 - p$, where we recall that $p=\Pi_{k=1}^K d_i$ is the number of covariates (rows of Ω). In contrast, for the Cartesian product graph it is only p/d_1 regardless of c. Hence, as c and K increase, using the Kronecker product model a single edge in Ψ_1 can create a proliferation of edges whereas the number of new edges in the Kronecker sum model is fixed, independent of K. A concrete example of these differences is illustrated in Fig. 2. The qualitative differences between the Kronecker product and Kronecker sum models for the precision matrix are summarized in Table 1.

1.3. Rationale for the proposed multiway Kronecker sum model

This paper develops a scalable, fast and accurate estimation procedure, TeraLasso, for multiway precision matrices Ω by using higher order Kronecker sum models. To justify the practical utility of TeraLasso we illustrate it on a spatiotemporal meteorological data set. We have also applied it to other applications which are not presented here. Although a comprehensive validation of the model on a larger corpus of real data is beyond the scope of this paper, there is ample evidence that the model will have many statistical applications. We base this assessment on the

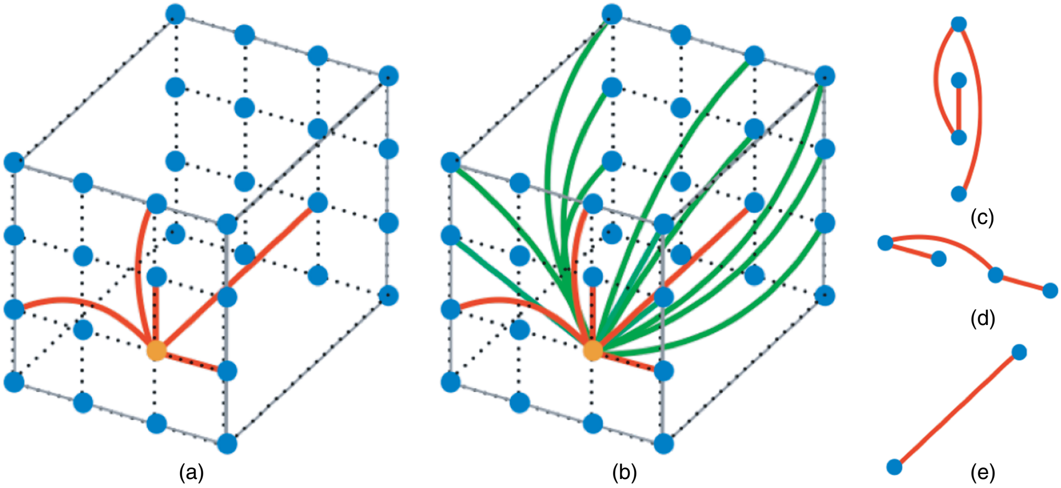


Fig. 2. Comparison of (a) the Kronecker sum (Cartesian product graph) and (b) Kronecker product (direct product graph): the products are formed from the component graphs (c), (d) and (e); the number of factors in the product graphs is $K=3$ and the dimensions are $d_1=d_2=4$ and $d_3=2$, leading to product graphs with 32 nodes, arranged in a regular three-dimensional grid in the figures at the bottom; only the edges emanating from the orange node are indicated (red and green edges); the Kronecker sum model has a total of 64 edges whereas the Kronecker product model is much less sparse, having a total of 184 edges

Table 1. Qualitative differences between the multiway Kronecker sum (TeraLasso) and multiway Kronecker product (BiGlasso) models for high dimensional precision matrix estimation

	Multiway Kronecker product	*Multiway Kronecker sum*
Covariance model	Precision matrix Ω is separable across K tensor components	Precision matrix is non-separable across tensor components, motivated by maximum entropy considerations
Graphical model	Graph is the direct product of the K graph components	Graph is the Cartesian product of the K graph components
Sparsity	Number of edges in Ω grows as the *product* of the number of edges in each component	Number of edges in Ω grows as the *sum* of the number of edges in each component
Graphical model interpretability	Edges in sparse factors contribute to large numbers of edges multiplicatively	Each edge in the sparse factors directly map to edges in the overall precision Ω; sparsity pattern follows Cartesian Markov-like network
Inference	Non-convex (multilinear) maximum likelihood estimator, alternative estimators usually favoured	Maximum likelihood estimator is convex

wide use of Kronecker sum models, equivalently Cartesian product graph models, in biology, physics, social sciences and network engineering, among other fields (Imrich *et al.*, 2008; Van Loan, 2000). In particular the Kronecker sum arises in solving the celebrated Sylvester equation for a matrix X which, for $K=2$, takes the form $XA+BX=N$. The Sylvester equation can be solved by expressing the equation in vectorized form as $A \oplus B\,\mathrm{vec}(X)=\mathrm{vec}(N)$ (for arbitrary K this becomes the tensor Sylvester equation $(A_1 \oplus \ldots \oplus A_K)\mathrm{vec}(X)=\mathrm{vec}(N)$), but this is often impractical in high dimension. Such equations result from the discretization of separable K-dimensional partial differential equations with tensorized finite elements (Grasedyck, 2004; Kressner and Tobler, 2010; Beckermann *et al.*, 2013; Shi *et al.*, 2013; Ellner, 1986). As a result

Kronecker sums come in many areas of applied mathematics, including beam propagation physics (Andrianov, 1997), control theory (Luenberger, 1966; Chapman *et al.*, 2014), fluid dynamics (Dorr, 1970) and spatiotemporal neural processes (Schmitt *et al.*, 2001).

Closer to home, the Kronecker sum model arises in multivariate spline data analysis, e.g. as applied to harmonic analysis on graphs (Kotzagiannidis and Dragotti (2019)). More recently, Fey *et al.* (2018) have proposed tensor B-splines defined over a Cartesian product basis for geometric convolutional neural networks. Kronecker sums have been proposed as precision matrices for weighting the quadratic regularizer in smoothed multivariate spline regression. In particular, Wood (2006) observed that, compared with the Kronecker product, the Kronecker sum reduces the coupling between the axes when used as a spline smoothing penalty for generalized additive mixed model regression. This observation motivated Wood (2006) and Eilers and Marx (2003) to use the inverse of a Kronecker sum matrix as a penalty, or prior, for smoothing K-dimensional regressions (see also work by Lee and Durbán (2011) and Wood *et al.* (2016)). This approach has been applied to spatiotemporal forest health modelling (for which $K=3$) (Augustin *et al.*, 2009), brain development modelling (Holland *et al.*, 2014) and analysis of the effect of climate and weather on spatiotemporal patterns of beetle populations (Preisler *et al.*, 2012), among other applications. In these spline regression problems the Kronecker sum appears as a precision matrix parameterizing a Gaussian prior on the spline coefficient vector β, where the prior is of the form $p(\beta) \propto \exp\{-\beta^{\mathrm{T}}(\lambda_1 S_1 \oplus \ldots \oplus \lambda_K S_K)\beta/2\}$. Here, λ_i are regularization coefficients and S_i are co-ordinatewise smoothing matrices, $i=1,\ldots,K$.

Instead of using the Kronecker sum to model the *a priori* precision matrix of a set of spline parameters, this paper proposes the Kronecker sum as a model for the precision matrix of the multiway data in the likelihood function, where the data matrix X takes the place of the spline coefficient vector β. The stated advantages of the Kronecker sum model for the spline regression setting (Wood, 2006) can be expected to carry over to the precision matrix estimation setting of TeraLasso. In particular, like the spline regression prior, TeraLasso smooths each axis separately, while summing over the others, thereby reducing coupling between the tensor axes compared with the Kronecker product. For data that have structure similar to that imposed by Wood (2006) on the spline regression coefficients this should result in a more accurate fit. Indeed, if a population of regression spline problems was available, in principle one could apply TeraLasso to estimating the best precision matrix of the spline coefficients that would minimize the population-averaged fitting error.

1.4. Outline

The remainder of the paper is organized as follows. We introduce notation and some preliminary results in Section 2, and our proposed TeraLasso model in Section 3. High dimensional consistency results are presented in Section 4, first with convex l_1-regularizers and then with non-convex sparsity regularizers. The first-order iterative soft thresholding optimization algorithm is described in Section 5, and conditions are specified for which the algorithm converges geometrically to the global optimum. Finally, Sections 6 and 7 illustrate the proposed TeraLasso estimator on simulated and real data, with Section 8 concluding the paper. We place all technical proofs in the on-line supplementary material, along with additional experiments and further exploration of the properties and implications of the Kronecker sum subspace $\mathcal{K}_{\mathbf{p}}$ and the associated identifiable parameterization.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`

2. Notation and preliminaries

We use upper case letters, e.g. A, for matrices and tensors, and bold lower case, e.g. $\mathbf{a}$, for vectors, and we denote the (i, j) element of a matrix A as A_{ij} and the $(i_1, i_2, \ldots, i_K)$ element of a tensor A as $A_{i_1,i_2,\ldots,i_K}$. Fibres are the higher order analogue of matrix rows and columns. A fibre of a tensor is obtained by fixing every index except one; the mode k fibre of tensor X is denoted as the column vector $X_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_K}$. Following the definition by Kolda and Bader (2009), tensor unfolding or matricization of X along the kth mode is denoted as $X_{(k)}$, formed by arranging the mode k fibres as columns of the resulting matrix of dimension $d_k \times m_k$. The column ordering is not important so long as it is consistent.

For a vector $\mathbf{y} = (y_1, \ldots, y_p)$ in $\mathbb{R}^p$, denote by $\|\mathbf{y}\|_2 = \sqrt{\Sigma_j y_j^2}$ the Euclidean norm of $\mathbf{y}$. The operator and Frobenius norms of a matrix A are denoted as $\|A\|_2$ and $\|A\|_{\mathrm{F}}$ respectively; the notation $\mathrm{vec}(A)$ denotes the vectorization of the matrix A; $\|A\|_\infty$ denotes the matrix ∞ norm and $\|A\|_{\max} = \max_{ij} |A_{ij}|$ denotes the max-norm. The determinant is denoted as $|A|$. We use the inner product $\langle A, B\rangle = \mathrm{tr}(A^{\mathrm{T}} B)$ throughout. Define the set of $p \times p$ matrices with Kronecker sum structure of fixed dimensions $d_1, \ldots, d_K$:

$$\mathcal{K}_{\mathbf{p}} = \{A \in \mathbb{R}^{p\times p} : \exists\, B_k \in \mathbb{R}^{d_k\times d_k} \text{ subject to } A = B_1 \oplus \ldots \oplus B_K\} \tag{5}$$

where the set of matrices that is defined in expression (4) is obtained by restricting $\mathcal{K}_{\mathbf{p}}$ to the positive cone, i.e.

$$\mathcal{K}_{\mathbf{p}}^{\sharp} = \{A \succeq 0 | A \in \mathcal{K}_{\mathbf{p}}\}.$$

The set $\mathcal{K}_{\mathbf{p}}$ (5) is linearly spanned by the K components, since there are no non-linear interactions between any of the parameters. Thus $\mathcal{K}_{\mathbf{p}}$ is a linear subspace of $\mathbb{R}^{p\times p}$, and we can define a unique projection operator onto $\mathcal{K}_{\mathbf{p}}$:

$$\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A) = \arg \min_{M\in\mathcal{K}_{\mathbf{p}}} \|A - M\|_{\mathrm{F}}^2.$$

A closed form expression for $\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A)$ is given in section A.3 of the on-line supplementary material. Note that the dimensionality of the $\mathcal{K}_{\mathbf{p}}$-subspace is $1 - K + \Sigma_{k=1}^K d_k^2$, which is often significantly smaller than the ambient dimension $p^2 = \Pi_{k=1}^K d_k^2$.

2.1. Parameterization of $\mathcal{K}_{\mathbf{p}}$ by Ψ_k

Note that $\Omega = \Psi_1 \oplus \ldots \oplus \Psi_K$ does not uniquely determine $\{\Psi_k\}_{k=1}^K$, i.e. without further constraints the Kronecker sum parameterization is not fully identifiable. It is easy to verify, however, that both $\mathrm{offd}(\Psi_k)$ and $\mathrm{diag}(\Omega)$ are identifiable, where we define the notation $\mathrm{offd}(M) = M - \mathrm{diag}(M)$. We can then write the identifiable decomposition

$$\hat{\Omega} = \mathrm{diag}(\hat{\Omega}) + \mathrm{offd}(\hat{\Psi}_1) \oplus \ldots \oplus \mathrm{offd}(\hat{\Psi}_K), \tag{6}$$

and correspondingly $\Omega_0 = \mathrm{diag}(\Omega_0) + \mathrm{offd}(\Psi_{0,1}) \oplus \ldots \oplus \mathrm{offd}(\Psi_{0,K})$. Note that, whereas the off-diagonal factors can take on any values, $\mathrm{diag}(\Omega_0)$ is not completely free (for a fully orthogonal parameterization see section 4 of the on-line supplement).

2.2. Interpretation of correlation coefficients

The quantities $[\Psi_k]_{ij}/\sqrt{([\Psi_k]_{ii}[\Psi_k]_{jj})}$ do not by themselves correspond to correlation coefficients. Because of the repeating structure of the Kronecker sum each element $[\Psi_k]_{ij}$ will appear in m_k distinct $d_k \times d_k$ symmetric subblocks of Ω, and in each (lth) subblock it will have a correlation coefficient that is uniquely defined for that subblock:

$$\rho_{k,ij,l} = \frac{[\Psi_k]_{ij}}{\surd\{([\Psi_k]_{ii} + c_l/d_k)([\Psi_k]_{jj} + c_l/d_k)\}}$$

where $c_l =$ tr(lth subblock of Ω) $-$ tr(Ψ_k). The overall correlation structure is preserved across the m_k blocks; simply the strength of the correlations are modulated by the contributions of the other $K-1$ additive factors in the block. (Recall that the Ψ_k need not be positive definite and c_l need not be greater than 0.)

3. Models and methods

Let $X_1, \ldots, X_n$ be n independent realizations of the K-way tensor X. Define $\mathbf{x}_i = \mathrm{vec}(X_i^{\mathrm{T}})$ for all $i = 1, \ldots, n$. Define $\hat{S} = (1/n)\Sigma_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ as the sample covariance. The mode k Gram matrix S_k and factorwise covariance $\Sigma^{(k)} = \mathbb{E}[S_k]$ are given by

$$S_k = \frac{1}{nm_k} \sum_{i=1}^n X_{i,(k)} X_{i,(k)}^{\mathrm{T}} \quad \text{and} \quad \Sigma^{(k)} = \frac{1}{m_k} \mathbb{E}[X_{(k)} X_{(k)}^{\mathrm{T}}], \qquad k = 1, \ldots, K,$$

noting that the elements of these matrices are effectively inner products between $(K-1)$-order tensors. S_k is the sample covariance of the data unfolded across the kth tensor axis, whereas $\Sigma^{(k)}$ denotes the population covariance matrix along the same axis. These Gram matrices S_k can be represented as elementwise aggregations over entries in the full sample covariance (3), with locations indexed by $\Psi_{k,i,j}$ as

$$[S_k]_{ij} = \frac{1}{m_k} \langle \hat{S}, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}} \otimes I_{[d_{k+1:K}]} \rangle. \tag{7}$$

In tensor covariance modelling when the dimension p is much larger than the number of samples n, the Gram matrices S_k are often used to model the rows and columns separately, notably in the matrix variate estimation methods of Zhou (2014) and Kalaitzis *et al.* (2013). Observe that the TeraLasso estimator (2) of the precision matrix can be expressed as

$$\hat{\Omega} = \arg \min_{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}, \|\Omega\|_2 \leqslant \kappa} \left[-\log|\Omega| + \sum_{k=1}^K m_k \left\{ \langle S_k, \Psi_k \rangle + \sum_{i \neq j} g_{\rho_k}([\Psi_k]_{ij}) \right\} \right] \tag{8}$$

where $\mathcal{K}_{\mathbf{p}}^{\sharp}$ is the set of positive semidefinite Kronecker sum matrices (4).

Ignoring regularization, the objective function in square brackets can be written as $-\log\{p(\hat{S}|\Omega)\}$ where $p(\hat{S}|\Omega) = \alpha_\Omega \Pi_{k=1}^K p(S_k|\Psi_k)$ and $p(S_k|\Psi_k) = \exp(-\langle m_k S_k, \Psi_k \rangle)$, with α_Ω a normalizing constant. The non-negativity of the Kullback–Liebler divergence

$$\int p(S|\Omega) \log\left\{ \frac{p(S|\Omega)}{\alpha_\Omega \prod_{k=1}^K p(S_k|\Psi_k)} \right\} \mathrm{d}S$$

implies that the Kronecker sum model is a *maximum entropy* model, as previously pointed out for the case of $K = 2$ by Kalaitzis *et al.* (2013). Alternatively, Kronecker sum models can be characterized as regularizing the precision matrix estimation problem with a minimally informative prior over the set $\mathcal{K}_{\mathbf{p}}^{\sharp}$.

The class of Kronecker sum matrices is a highly structured, lower dimensional subspace of $\mathbb{R}^{p \times p}$. By definition of the Kronecker sum (1), each entry of Ψ_k appears in $m_k = p/d_k$ entries of Ω. By imposing that the precision matrix has both Kronecker sum structure and sparse structure through the penalty g_ρ, TeraLasso can effectively regularize the precision estimation problem.

We assume that the penalty g_ρ is (μ, γ) amenable in the sense of Loh and Wainwright (2017).

Definition 1 ((μ, γ) amenable regularizer). A regularizer $g_\rho(t)$ is (μ, γ) amenable when $\mu \geqslant 0$ and $\gamma \in (0, \infty)$ if

(a) g_ρ is symmetric around zero and $g_\rho(0) = 0$,
(b) $g_\rho(t)$ and $g_\rho(t)/t$ are both non-decreasing on $\mathbb{R}^+$,
(c) $g_\rho(t)$ is differentiable for all $t \neq 0$,
(d) the function $g_\rho(t) + (\mu/2)t^2$ is convex,
(e) $\lim_{t \to 0^+} g'_\rho(t) = \rho$ and
(f) $g'_\rho(t) = 0$ for all $t \geqslant \gamma\rho$.

Note that the l_1-regularizer is $(0, \infty)$ amenable. Example non-convex penalties in this class include the SCAD penalty (Fan and Li, 2001) and the MCP penalty (Zhang, 2010), both defined in appendix C of the on-line supplement.

Observe that for non-zero μ (i.e. non-convex g_ρ) the constraint on the spectral norm of Ω ($\|\Omega\|_2 \leqslant \kappa$) in the TeraLasso objective function (8) is necessary since without it a global minimum may not exist (Loh and Wainwright, 2017). For a spectral norm constraint parameter set to $\kappa = \surd(2/\mu)$, we show (lemma 21 in the supplement) that objective function (8) with g_ρ (μ, γ) amenable is convex and has a unique global minimizer. For the l_1-penalty, the objective is always convex and κ can be set to ∞.

4. High dimensional consistency of TeraLasso

Let $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}}$ be an isotropic ψ_2-sub-Gaussian random vector with independent entries v_j satisfying $\mathbb{E}[v_j] = 0$, $1 = \mathbb{E}[v_j^2] \leqslant \|v_j\|_{\psi_2} \leqslant K$. The ψ_2-condition on a scalar random variable V is equivalent to sub-Gaussian decay of the tails of V, implying that $\mathbb{P}(|V| > t) \leqslant 2\exp(-t^2/c^2)$ for all $t > 0$. The extension to random vectors is straightforward. Specifically, $\mathbf{x}$ is a sub-Gaussian random vector with positive definite covariance $\Sigma \in \mathbb{R}^{p \times p}$ when

$$\mathbf{x} = \Sigma^{1/2}\mathbf{v}, \tag{9}$$

where $\Sigma^{1/2}$ denotes a positive definite square-root factor of Σ. We then call $X \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_K}$ an order K sub-Gaussian random tensor with covariance Σ when $\mathbf{x} = \mathrm{vec}(X^{\mathrm{T}})$ is a sub-Gaussian random vector in $\mathbb{R}^p$ defined as in equation (9).

We assume that the data $X_1, X_2, \ldots, X_n$ are independent and identically distributed sub-Gaussian random tensors whose inverse covariance follows the Kronecker sum model (1), namely, that $\mathrm{vec}(X_i^{\mathrm{T}}) \sim \mathbf{x}$, where $\mathbf{x}$ is a sub-Gaussian random vector in $\mathbb{R}^p$ as defined in equation (9). A special case of the sub-Gaussian model is the Gaussian model, for which the 0s in the precision matrix define the conditional independences among the variables X_i. This conditional independence relationship does not hold for the general sub-Gaussian case, but nonetheless strong convergence of the TeraLasso precision matrix estimator is preserved.

In addition to the sub-Gaussian generative model given above, we make the following technical assumptions on the true model, guaranteeing sparsity in Ω and its eigenvalues being bounded away from 0 and ∞.

Assumption 1. Define the support set of the kth Kronecker sum component Ψ_k of the precision matrix by $\mathcal{S}_k = \{(i, j) : i \neq j, [\Psi_k]_{ij} \neq 0\}$ for $k = 1, \ldots, K$. We assume that $\mathcal{S}_k$ is sparse, i.e. $\mathrm{card}(\mathcal{S}_k) \leqslant s_k$.

Assumption 2. The minimal eigenvalue satisfies $\phi_{\min}(\Omega) = \Sigma_{k=1}^K \phi_{\min}(\Psi_k) \geqslant \underline{k}_\Omega > 0$, and the maximum eigenvalue satisfies $\phi_{\max}(\Omega) = \Sigma_{k=1}^K \phi_{\max}(\Psi_k) \leqslant \bar{k}_\Omega < \infty$.

Defining the support set of Ω as $\mathcal{S}=\{(i,j):i\neq j,\}$, assumption 1 implies that $\text{card}(\mathcal{S})\leqslant s=\Sigma_{k=1}^{K}m_k s_k$.

4.1. Regularization with l_1-penalty

With $g_\rho(t)=\rho|t|$, the constraint on $\|\Omega\|_2$ is unnecessary, and objective function (8) becomes

$$\hat{\Omega}=\arg\min_{\Omega\in\mathcal{K}_{\mathbf{p}}^{\sharp}}\left\{-\log|\Omega|+\sum_{k=1}^{K}m_k(\langle S_k,\Psi_k\rangle+\rho_k|\Psi_k|_{1,\text{off}})\right\}\tag{10}$$

where $|\Psi_k|_{1,\text{off}}=\Sigma_{i\neq j}|[\Psi_k]_{ij}|$ is the off-diagonal l_1-norm. The objective (10) is jointly convex, and its minimization over $\Omega\in\mathcal{K}_{\mathbf{p}}^{\sharp}$ has a unique solution (see section 2.6 of the on-line supplement). We require an additional assumption.

Assumption 3. The sample size n and the component dimensions d_k satisfy the following condition:

$$n(\min_k m_k)^2\geqslant C^2\kappa(\Sigma_0)^4(s+p)(K+1)^2\log(p)\tag{11}$$

where $m_k=p/d_k$ and $\kappa(\Sigma_0)=\phi_{\max}(\Sigma_0)/\phi_{\min}(\Sigma_0)$ is the condition number of Σ_0.

This assumption holds for $n=1$ and sufficiently large $(\min_k m_k)^2>O(p)$, which can hold for any $K>2$. We obtain the following bounds on the Frobenius and operator norm error of the TeraLasso estimator (10). The constants (c,C_1,C_2,C_3) are given in the proof (see the on-line supplement) and do not depend on K, n, s or $\mathbf{p}$.

Theorem 1 (Frobenius error bound). Suppose that assumptions 1–3 hold, and that $\hat{\Omega}$ is the minimizer of expression (10) with $\rho_k\asymp(1/\underline{k}_{\Omega})\surd\{\log(p)/(nm_k)\}$. Then with probability at least $1-2(K+1)\exp\{-c\log(p)\}$

$$\|\hat{\Omega}-\Omega_0\|_{\text{F}}\leqslant\frac{2C_1\|\Sigma_0\|_2}{\phi_{\min}^2(\Sigma_0)}\sqrt{\left\{(K+1)(s+p)\frac{\log(p)}{n\min_k m_k}\right\}}.$$

Theorem 2 (factorwise and L2 error bounds). Suppose that the conditions of theorem 1 hold. Then with probability at least $1-2(K+1)\exp\{-c\log(p)\}$:

$$\frac{\|\text{diag}(\hat{\Omega})-\text{diag}(\Omega_0)\|_2^2}{(K+1)\max_k d_k}+\sum_{k=1}^{K}\frac{\|\text{offd}(\hat{\Psi}_k-\Psi_{0,k})\|_{\text{F}}^2}{d_k}\leqslant C_2(K+1)\left(1+\sum_{k=1}^{K}\frac{s_k}{d_k}\right)\frac{\log(p)}{n\min_k m_k}\tag{12}$$

and as a result

$$\|\hat{\Omega}-\Omega_0\|_2\leqslant C_3(K+1)\sqrt{\left\{\frac{p}{(\min_k m_k)^2}\left(1+\sum_{k=1}^{K}\frac{s_k}{d_k}\right)\frac{\log(p)}{n}\right\}}.$$

Theorems 1 and 2 are proved in section 5 of the on-line supplement. Observe that theorem 2 predicts result (12) that, for fixed n and $K>2$, the estimation error of the parameters of Ω converges to 0 as the dimensions $\{d_k\}$ go to ∞ (recall that $p=\Pi_{k=1}^{K}d_k$). This implies that for increasing dimensions TeraLasso will converge even for a single sample $n=1$. Because of the repeating structure and increasing dimension of Ω, the parameter estimates can converge without the overall Frobenius error $\|\hat{\Omega}-\Omega_0\|_{\text{F}}$ converging.

4.1.1. *Comparison with graphical lasso*

The Frobenius norm bound in theorem 1 improves on the sub-Gaussian graphical lasso rate of

Rothman *et al.* (2008) and Zhou *et al.* (2011) by a factor of $\min_k m_k$. If the dimensions are equal ($d_k = p^{1/K}$ and s_k are constant over k) and K is fixed, theorem 2 implies that

$$\|\Delta_{\mathrm{k}}\|_{\mathrm{F}} = O_p\left[\sqrt{\left\{\frac{(d_k + s_k)\log(p)}{m_k n}\right\}}\right],$$

indicating that TeraLasso with n replicates estimates the identifiable representation of Ψ_k with an error rate equivalent to that of the graphical lasso with $\Omega = \Psi_k$ and nm_k available replicates.

4.1.2. Independence along an axis

Suppose that the data tensor X is independent and identically distributed along the first axis, i.e. $\Psi_1 = I_{d_1}$. Then, instead of a K-way TeraLasso, a $(K-1)$-model with nd_1 replicates would suffice, yielding a factorwise error bound (theorem (2)) of

$$O\left[\sqrt{\left\{\left(1 + \sum_{k=2}^{K}\frac{s_k}{d_k}\right)\frac{\log(p/d_1)}{nd_1 \min_{k>1}(m_k/d_1)}\right\}}\right],$$

compared with the factorwise error bound of

$$O\left[\sqrt{\left\{\left(1 + \sum_{k=2}^{K}\frac{s_k}{d_k}\right)\frac{\log(p)}{n \min_k m_k}\right\}}\right]$$

associated with the full K-way model (since $s_1 = 0$). Hence having *a priori* knowledge of independence (allowing the use of the $(K-1)$-model) does not meaningfully improve the rate over the original K-way model so long as $\min_{k>1} m_k \approx \min_k m_k$. A similar satisfying result holds for the Frobenius error bound in theorem 1.

4.2. Non-convex regularizers and single-sample support recovery

Non-convex regularization will provide non-asymptotic guarantees on the elementwise estimation error, implying strong, single-sample support recovery guarantees when the smallest non-zero element of Ω_0 is bounded from below. However, these stronger results require more restrictive assumptions on sparsity of the precision matrix and its smallest non-zero element. Specifically, we shall require the following assumptions.

Assumption 4. The degree (maximum number of non-zero edges connected to a node) of the sparsity graph of each factor Ψ_k is bounded by a constant d.

Assumption 5. The sample size satisfies $n \min_k m_k \geqslant c_0 d^2 \log(p)$ for some c_0 sufficiently large.

Assumption 6. There are constants c_∞ and c_3 such that $\|(\Omega_0 \otimes \Omega_0)_{\mathcal{SS}}\|_\infty \leqslant c_\infty$ and

$$\min_{[i,j]\in\mathcal{S}} |[\Omega_0]_{ij}| \geqslant \rho(\gamma + 2c_\infty) + c_3\sqrt{\left\{\frac{\log(p)}{n \min_k m_k}\right\}}.$$

In assumption 6 the notation $A_{\mathcal{SS}}$ denotes the submatrix of A formed by extracting the rows and columns corresponding to the index set $\mathcal{S}$. Under these assumptions we have the following result.

Theorem 3 (non-convex regularizers). Suppose that the regularizer g_ρ in objective function (8) is (μ, γ) amenable, and $\kappa = \surd(2/\mu)$. Then with probability at least $1 - 2(K+1)\exp\{-c\log(p)\}$ as in theorem 1, expression (8) has a unique stationary point $\hat{\Omega}$ (given by the oracle estimator defined in the on-line supplement), with (for all k)

$$\|\text{offd}(\hat{\Psi}_k - \Psi_{0,k})\|_{\max} \leqslant \|\hat{\Omega} - \Omega_0\|_{\max} \leqslant c_3(K+1)\sqrt{\left\{\frac{\log(p)}{n\min_k m_k}\right\}},$$

$$\|\text{offd}(\hat{\Psi}_k - \Psi_{0,k})\|_{\text{F}} \leqslant c_3(K+1)\sqrt{\left\{\frac{s_k\log(p)}{n\min_k m_k}\right\}},$$

$$\|\hat{\Omega} - \Omega_0\|_{\text{F}} \leqslant c_3(K+1)\sqrt{\left\{\frac{(s+p)\log(p)}{n\min_k m_k}\right\}},$$

$$\|\hat{\Omega} - \Omega_0\|_2 \leqslant c_3 d(K+1)\sqrt{\left\{\frac{\log(p)}{n\min_k m_k}\right\}}.$$

The proof of theorem 3 is given in Section 7 in the on-line supplement and uses arguments that are analogous to those of Loh and Wainwright (2017) along with concentration inequalities arising from the structure of the TeraLasso model.

Theorem 3 implies that the elements (of both Ω and the offdiagonals of Ψ_k), and thus the support (of both Ω and the Ψ_k), can be estimated by using a single sample ($n=1$) provided that $\min_k m_k$ is sufficiently large. The Frobenius norm convergence rates (both factorwise and overall) for the convex and non-convex regularizers remain effectively the same (comparing theorem 3 with theorems 1 and 2); hence the primary benefit of the non-convex bound is the ability to guarantee support recovery in exchange for additional assumptions.

5. Tensor graphical iterative soft thresholding algorithm

In this section, we introduce an iterative soft thresholding method, restricted to the convex set $\mathcal{K}_{\mathbf{p}}^{\sharp}$ of possible positive semidefinite Kronecker sum precision matrices, to implement the TeraLasso optimization (8). We call this tensor graphical iterative soft thresholding implementation TG-ISTA.

5.1. *Composite gradient descent and proximal first-order methods*

Our goal is to solve the objective (8). This objective function can be decomposed into the sum of a differentiable function f and a lower semicontinuous but non-smooth function g: for $\Omega \in \mathcal{K}_{\mathbf{p}}$

$$Q(\Psi_1, \ldots, \Psi_K) = f(\Omega) + g(\Omega),$$

where, for $\langle \hat{S}, \Omega\rangle = \Sigma_{k=1}^K m_k \langle S_k, \Psi_k\rangle$,

$$\begin{aligned} f(\Omega) &= -\log|\Omega| + \langle \hat{S}, \Omega\rangle\Big|_{\Omega\in\mathcal{K}_{\mathbf{p}}}, \\ g(\Omega) &= \sum_{k=1}^K m_k \sum_{i\neq j} g_{\rho_k}([\Psi_k]_{ij}). \end{aligned} \tag{13}$$

For objectives of this form, Nesterov (2013) proposed a first-order method called composite gradient descent. Composite gradient descent has been specialized to the case of $g = |\cdot|_1$ and is widely known as iterative soft thresholding (see for example Tseng (2010), Combettes and Wajs (2005), Beck and Teboulle (2009) and Nesterov (1983, 2004)). An extension to non-convex regularizers g was given by Loh and Wainwright (2013).

The linearity of the constraint set $\mathcal{K}_{\mathbf{p}}$ suggests the use of gradient descent where the gradients are projected onto the associated $(1 - K + \Sigma_{k=1}^K d_k^2)$-dimensional *linear subspace*. The positive definite restriction can then be handled in a similar way to how Guillot *et al.* (2012) did for the original graphical lasso. We therefore derive composite gradient descent in the linear subspace $\mathcal{K}_{\mathbf{p}}$ of $\mathbb{R}^{p^2}$, creating a positive definite sequence of iterates $\{\Omega_t\}$ given by the recursion

$$\Omega_{t+1} \in \arg \min_{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}} (\tfrac{1}{2} \|\Omega - [\Omega_t - \zeta_t \mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}\{\nabla f(\Omega_t)\}]\|_{\mathrm{F}}^2 + \zeta_t g(\Omega)), \tag{14}$$

where the initial matrix $\Omega_0 \in \mathcal{K}_{\mathbf{p}}^{\sharp}$ can be chosen as the identity. We enforce the positive semidefinite constraint at each step by performing backtracking line search to find a suitable step size ζ_t (see algorithm 1 in Table 2 in Section 5.2) (Guillot *et al.*, 2012). We decompose and solve problem (14) for the case of the TeraLasso objective in Section 5.2.

5.2. TG-ISTA implementation of TeraLasso

To apply this form of composite gradient descent to the TeraLasso objective, the projected gradient of $f(\Omega)$ is required for expression (13). For simplicity, consider the l_1-regularized case. The general non-convex case is described in the next section and the on-line supplement. Since the gradient of $\langle \hat{S}, \Omega \rangle$ with respect to Ω is $\hat{S}$ (lemma 33 in the on-line supplementary material)

$$\nabla_{\Omega \in \mathcal{K}_{\mathbf{p}}}(\langle \hat{S}, \Psi_1 \oplus \ldots \oplus \Psi_k \rangle) = \mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\hat{S}) = \tilde{S}_1 \oplus \ldots \oplus \tilde{S}_K = \tilde{S} \qquad \tilde{S}_k = S_k - \frac{K-1}{K} \frac{\mathrm{tr}(S_k)}{d_k} I_{d_k}. \tag{15}$$

Although many different conventions for parameterizing the projection by using the $\tilde{S}_k$ are possible, the projection remains unique. Alternative parameterizations will not affect the convergence or output of the algorithm. Since the gradient of $-\log|\Omega|$ with respect to Ω is Ω^{-1} (Boyd and Vandenberghe, 2009), the projected gradient takes the form

$$\nabla_{\Omega \in \mathcal{K}_{\mathbf{p}}}(-\log|\Omega|) = \mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Omega^{-1}) = G_1^t \oplus \ldots \oplus G_K^t. \tag{16}$$

The matrices $G_k^t \in \mathbb{R}^{d_k \times d_k}$ are computed via the expressions that are given in lemma 33 in the on-line supplement. Combining expressions (15) and (16), the projected gradient of the objective $f(\Omega_t)$ is

$$\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}\{\nabla f(\Omega_t)\} = \tilde{S} - (G_1^t \oplus \ldots \oplus G_K^t). \tag{17}$$

Lemma 1 (decomposition of objective). For $\Omega_t, \Omega \in \mathcal{K}_{\mathbf{p}}$ of the form

$$\Omega_t = \Psi_1^t \oplus \ldots \oplus \Psi_K^t$$

and

$$\Omega = \Psi_1 \oplus \ldots \oplus \Psi_K,$$

the unique solution to problem (14) with $g_\rho = |\cdot|_1$ is given by $\Omega_{t+1} = \Psi_1^{t+1} \oplus \ldots \oplus \Psi_K^{t+1}$ where

$$\Psi_k^{t+1} = \arg \min_{\Psi_k \in \mathbb{R}^{d_k \times d_k}} \tfrac{1}{2} \|\Psi_k - \{\Psi_k^t - \zeta_t(\tilde{S}_k - G_k^t)\}\|_{\mathrm{F}}^2 + \zeta_t \rho_k |\Psi_k|_{1,\mathrm{off}}. \tag{18}$$

The proof is in the on-line supplement section 2.5. The right-hand side of equation (18) is the proximal operator of the l_1-penalty on the off-diagonal entries. The solution has closed form, as given in Beck and Teboulle (2009),

$$\Psi_k^{t+1} = \mathrm{shrink}_{\zeta_t \rho_k}^{-}\{\Psi_k^t - \zeta_t(\tilde{S}_k - G_k^t)\}, \tag{19}$$

where we define the off-diagonal shrinkage operator $\mathrm{shrink}_\rho^{-}(\cdot)$ as

$$[\mathrm{shrink}_\rho^{-}(M)]_{ij} = \begin{cases} \mathrm{sgn}(M_{ij})(|M_{ij}| - \rho)_+ & i \neq j, \\ M_{ij} & \text{otherwise}. \end{cases} \tag{20}$$

Table 2. Algorithm 1: TG-ISTA implementation of TeraLasso (high level)

```
input: sample covariance matrix factors $S_k$, regularization parameters $\rho_i$, backtracking constant
  $c \in (0,1)$, initial step size $\zeta_{1,0}$, initial iterate $\Omega_{\mathrm{init}} = I \in \mathcal{K}_{\mathbf{p}}^{\sharp}$
while not converged do
    compute the subspace gradient $\mathrm{proj}_{\mathcal{K}_{\mathbf{p}}}(\Omega_t^{-1}) = G_1^t \oplus \ldots \oplus G_K^t$
    line search, let step size $\zeta_t$ be the largest element of $\{c^j \zeta_{t,0}\}_{j=1,\ldots}$ such that the following are
      satisfied for $\Psi_k^{t+1} = \mathrm{shrink}_{\zeta_t \rho_k}^{-}\{\Psi_k^t - \zeta_t(\tilde{S}_k - G_k^t)\}$

          $\Psi_1^{t+1} \oplus \ldots \oplus \Psi_K^{t+1} \succ 0$

      and

          $f(\{\Psi_k^{t+1}\}) \leqslant \mathcal{Q}_{\zeta_t}(\{\Psi_k^{t+1}\}, \{\Psi_k^{t+1}\})$

    for $k = 1, \ldots, K$ do
        composite objective gradient update,

          $\Psi_k^{t+1} \leftarrow \mathrm{shrink}_{\zeta_t \rho_k}^{-}\{\Psi_k^t - \zeta_t(\tilde{S}_k - G_k^t)\}$

    end for
    compute Barzilai–Borwein step size $\zeta_{t+1,0}$ via expression (27) in on-line supplement section 2.2
end while
return $\{\Psi_k^{t+1}\}_{k=1}^K$
```

The composite gradient descent algorithm is given in algorithm 1 in Table 2. In section 8 of the on-line supplement, a scalable geometric rate of convergence of TG-ISTA to the global minimum is derived (theorem 25). In section 3.2 of the supplement we show that each iteration can be computed in $O(pK + \Sigma_{k=1}^K d_k^3)$ floating point operations.

5.3. TG-ISTA for a non-convex regularizer

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

$$\Omega^{t+1} = \mathrm{shrink}_{\zeta\rho}^{-}\{\Omega^t - \zeta \nabla \bar{\mathcal{L}}_n(\Omega^t)\} \tag{21}$$

where ζ is the step size and

$$\bar{\mathcal{L}}_n(\Omega) = -\log|\Omega| + \langle \hat{S}, \Omega \rangle + \sum_{k=1}^K m_k \sum_{i \neq j} \{g_\rho([\Psi_k]_{ij}) - \rho|[\Psi_k]_{ij}|\}.$$

The update (21) can be decomposed into the factorwise updates

$$\Psi_k^{t+1} = \mathrm{shrink}_{\zeta\rho}^{-}[\Psi_k^t - \zeta\{\tilde{S}_k - G_k^t + q_\rho'(\Psi_k)\}]$$

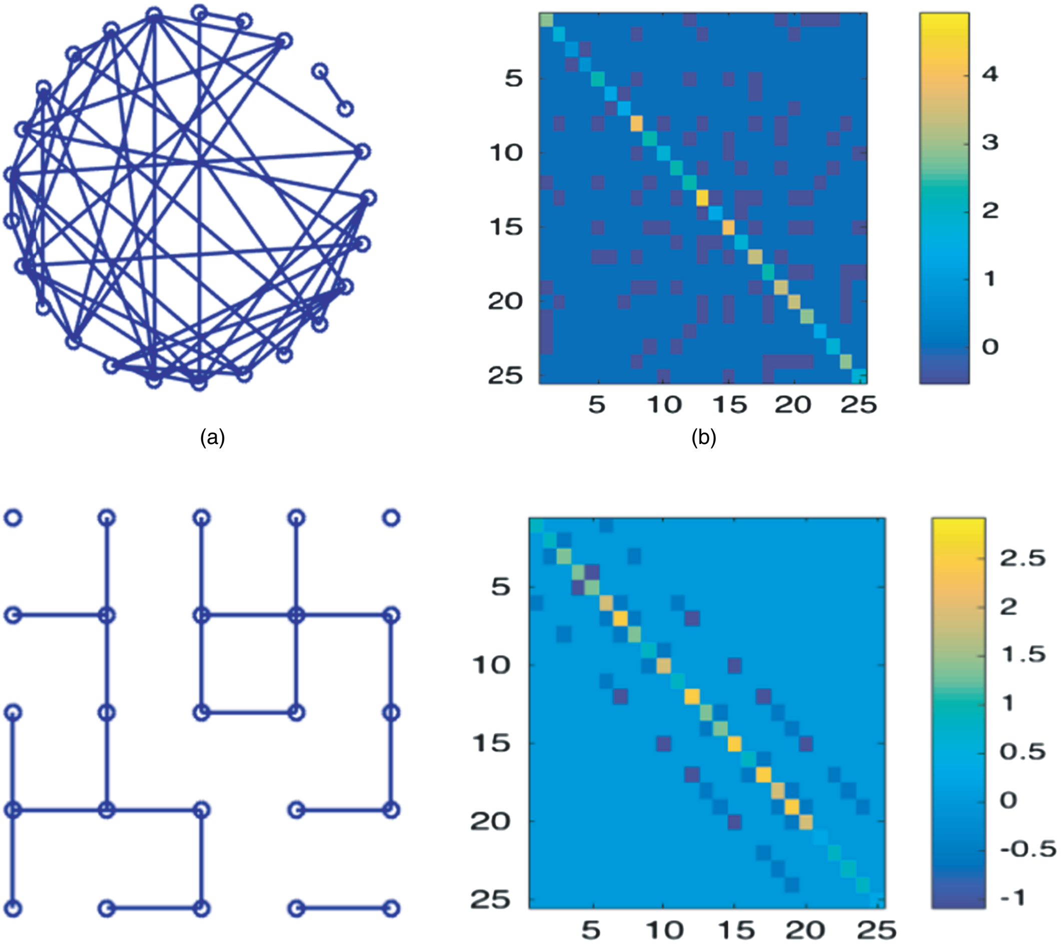
where $q_\rho'(t) = \mathrm{d}\{g_\rho(t) - \rho|t|\}/\mathrm{d}t$ for $t \neq 0$ and $q_\rho'(0) = 0$. These updates can be inserted into the framework of algorithm 1, with an added step of enforcing the $\|\Omega\|_2 \leqslant \kappa$ constraint, e.g. via step size line search. The algorithm is summarized in algorithm 2 in the on-line supplement section 2.1.

Theorem 4 (convergence of algorithm 2). Algorithm 2 will converge to the global optimum when the norm constraint parameter κ is chosen to be less than or equal to $\sqrt{(2/\mu)}$.

Proof. The proof follows since for $\kappa \leqslant \sqrt{(2/\mu)}$ the objective (8) is convex on the convex constraint set $\{\Omega \in \mathcal{K}_{\mathbf{p}} | \Omega \succ 0, \|\Omega\|_2 \leqslant \kappa\}$ (lemma 21 in the on-line supplement).

6. Validation on synthetic data

Random graphs were created for each factor Ψ_k by using both an Erdős–Renyi (ER) topology and a random grid graph topology. (Code for the experiments is included in the supplementary material and can be found at `https://github.com/kgreenewald/teralasso`.) These ER-type graphs were generated according to the method of Zhou *et al.* (2010). Initially we set $\Psi_k = 0.25 I_{n\times n}$, where $n = 100$, and randomly select q edges and update Ψ_k as follows: for each new edge (i, j), a weight $a > 0$ is chosen uniformly at random from $[0.2, 0.4]$; we subtract a from $[\Psi_k]_{ij}$ and $[\Psi_k]_{ji}$, and increase $[\Psi_k]_{ii}$ and $[\Psi_k]_{jj}$ by a. This keeps Ψ_k positive definite. We repeat this process until all edges are added. Finally, we form $\Omega = \Psi_1 \oplus \ldots \oplus \Psi_K$. An example 25-node, $q = 25$ ER graph and precision matrix are shown in Fig. 3. The random grid graph is produced in a similar way, with the exception that edges are allowed between adjacent nodes only, where the nodes are arranged on a square grid (Fig. 3(c)). Algorithm 1 in section 2.3 of the on-line supplement describes how the random vector $\mathbf{x} = \mathrm{vec}(X^{\mathrm{T}})$ is generated under the Kronecker sum model.



(a) (b)

(c) (d)

Fig. 3. (a), (b) Example ER graph with 25 nodes and 50 edges and (c), (d) random grid graph (square) with 25 nodes and 26 edges: (a), (c) graphical representation; (b),(d) corresponding precision matrix Ψ

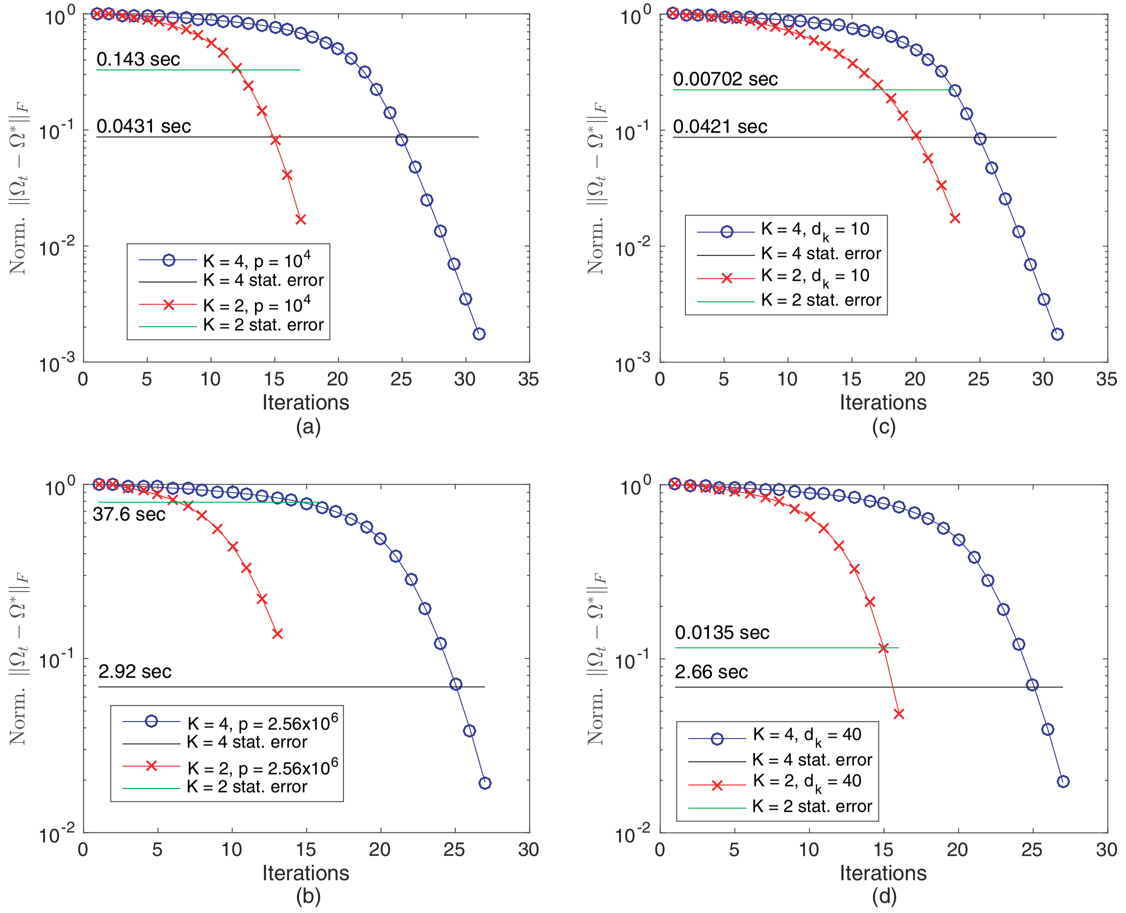
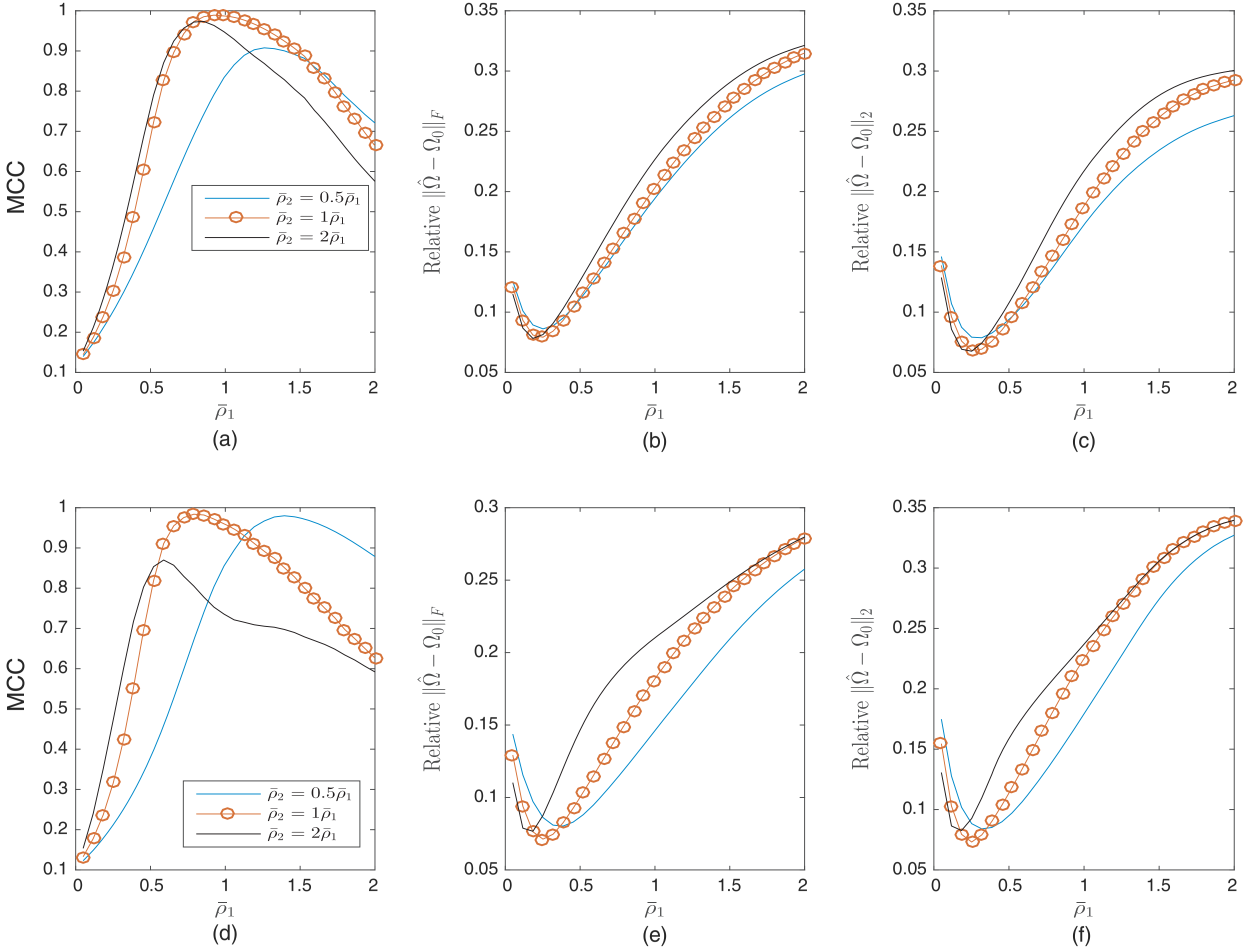


Fig. 4. Linear geometric convergence of the convex (l_1-penalized) TG-ISTA implementation of TeraLasso (shown is the normalized Frobenius norm $\|\Omega_t - \Omega^*\|_F$ of the difference between the estimate at the tth iteration and the optimal Ω^*); (a),(b) results comparing $K = 2$ and $K = 4$ on the same data with the same value of p (different d_k); (c), (d) results comparing the same value of d_k (different p); for (a), (c) n = 100 sample size and (b), (d) n = 1 sample size; also included are the statistical error levels, and the computation times required to reach them; observe the consistent and rapid linear convergence rate, with logarithmic dependence on K and dimension d_k



(a) (b) (c)

(d) (e) (f)

Fig. 5. Setting tuning parameters with $K = 3$, $n = 1$ and $d_1 = d_3 = 64$ (shown are the MCC, relative Frobenius error and relative L2-error of the TeraLasso estimate as the scaled tuning parameters ρ_i are varied; also shown are deviations of $\bar{\rho}_2$ from the theoretically dictated $\bar{\rho}_2 = \bar{\rho}_1 = \bar{\rho}_3$): (a)–(c) equal dimensions, $d_1 = d_2 = d_3$ (the first and third factors are random ER graphs with d_k edges, and the second factor is a random grid graph with $d_k/2$ edges); (d)–(f) dimensions $d_2 = 2d_1$ (each factor is a random ER graph with d_k edges); note in these scenarios that using $\bar{\rho}_1 = \bar{\rho}_2$ is near optimal, as theoretically predicted

6.1. Validation of theoretical algorithmic convergence rates

To verify the geometric convergence of the TG-ISTA implementation (theorem 25 in the on-line supplement), we generated Kronecker sum inverse covariance graphs and plotted the Frobenius norm between the inverse covariance iterates Ω_t and the optimal point Ω^*. We set the Ψ_k to be random ER graphs with d_k edges where $d_1 = \ldots = d_K$, and determined the value for $\rho_k = \rho$ by using cross-validation. Fig. 4 shows the results as a function of iteration, for a variety of d_k- and K-configurations and the l_1 convex regularization. Fig. 13 in the on-line supplement section 2.1 repeats these experiments with the non-convex SCAD and MCP penalties, using the same random seed. For comparison, the statistical error of the optimal point is also shown, as optimizing beyond this level provides reduced benefit. As predicted, linear or better convergence to the global optimum is observed. The small number of iterations combined with the low computational cost per iteration confirm the algorithmic efficiency of the TG-ISTA implementation of TeraLasso. Additional numerical experiments demonstrating fast convergence on larger-scale problems are given in section 3.2 of the supplement.

6.2. Regularization with l_1-penalty

In the TeraLasso objective (10), the sparsity of the estimate is controlled by K distinct tuning parameters ρ_k for $k = 1, \ldots, K$. The convergence condition on ρ_k in theorem 1 suggests that the ρ_k can be set as $\rho_k = \bar{\rho}\surd\{\log(p)/nm_k\}$ with $\bar{\rho}$ being a single scalar tuning parameter, depending on absolute constants and $\|\Sigma\|_2$. Below, we experimentally validate the reliability of this tuning strategy.

The performance is empirically evaluated by using several metrics including the Frobenius norm ($\|\hat{\Omega} - \Omega_0\|_F$) and spectral norm ($\|\hat{\Omega} - \Omega_0\|_2$) error of the precision matrix estimate $\hat{\Omega}$ and the Matthews correlation coefficient to quantify the edge misclassification error. Let the number of true positive edge detections be TP, true negative detections TN, false positive detections FP and false negative detections FN. The Matthews correlation coefficient is defined as (Matthews, 1975)

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\surd\{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})\}},$$

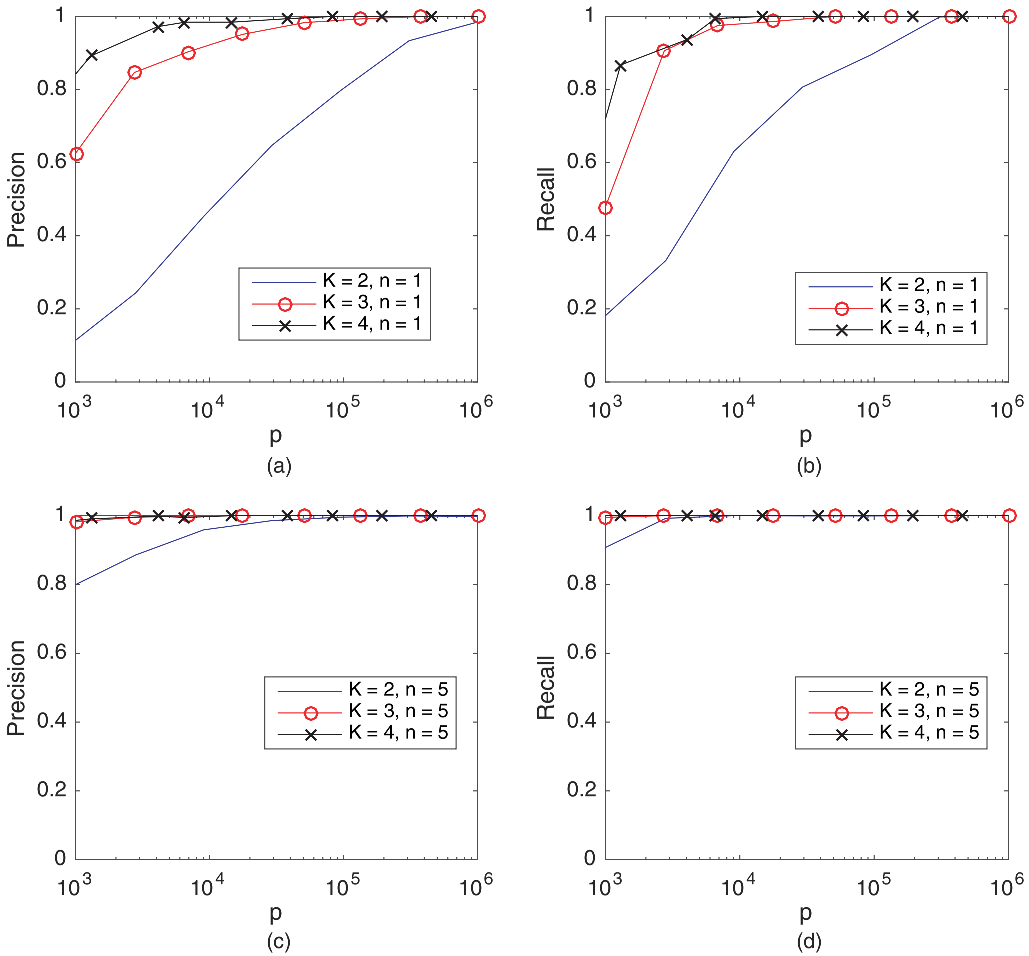
where each non-zero off-diagonal element of Ψ_k is considered as a single edge. Larger values of MCC imply better edge estimation performance, with $\mathrm{MCC} = 0$ implying complete failure and $\mathrm{MCC} = 1$ perfect edge set estimation.

Shown in Fig. 5 are the MCC, normalized Frobenius error and spectral norm error as functions of $\bar{\rho}_1$ and $\bar{\rho}_2$ where the $\bar{\rho}_k$ constants give $\rho_k = \bar{\rho}_k/\surd\{\log(p)/(nm_k)\}$. Note that $\bar{\rho}_1 = \bar{\rho}_2 = \bar{\rho}_3$ achieves near optimal results.

Having verified the single-tuning-parameter approach, hereafter we shall cross-validate only $\bar{\rho}$. In the on-line supplement section 3.3, we provide experimental verification in a wide variety of experimental settings (including varying the relative size of the tensor dimensions d_k) that our bounds on the rate of convergence for the l_1 regularized model are tight. Fig. 6 illustrates how increasing dimension p and K improves single-sample performance. Shown are the average TeraLasso edge detection precision and recall values for various values of K in the single and five-sample regimes, all increasing to 1 (perfect structure estimation) as p, K and n increase.

6.3. Non-convex regularization

Here the l_1-penalized TeraLasso is compared with TeraLasso with non-convex regularization (8). Shown in Fig. 7 are the MCC, normalized Frobenius error and spectral norm error for



(a) (b) (c) (d)

Fig. 6. Edge support estimation on random ER graphs, with the ρ_k set according to theorem 1 (graphical model edge detection precision and recall curves are shown as a function of data dimension $p = \Pi_{k=1}^{K} d_k$; for each value of the tensor order K, we set $d_k = p^{1/K}$; observe single-sample convergence as the dimension p increases and as increasing K creates additional structure): (a), (b) $n = 1$; (c), (d) $n = 5$

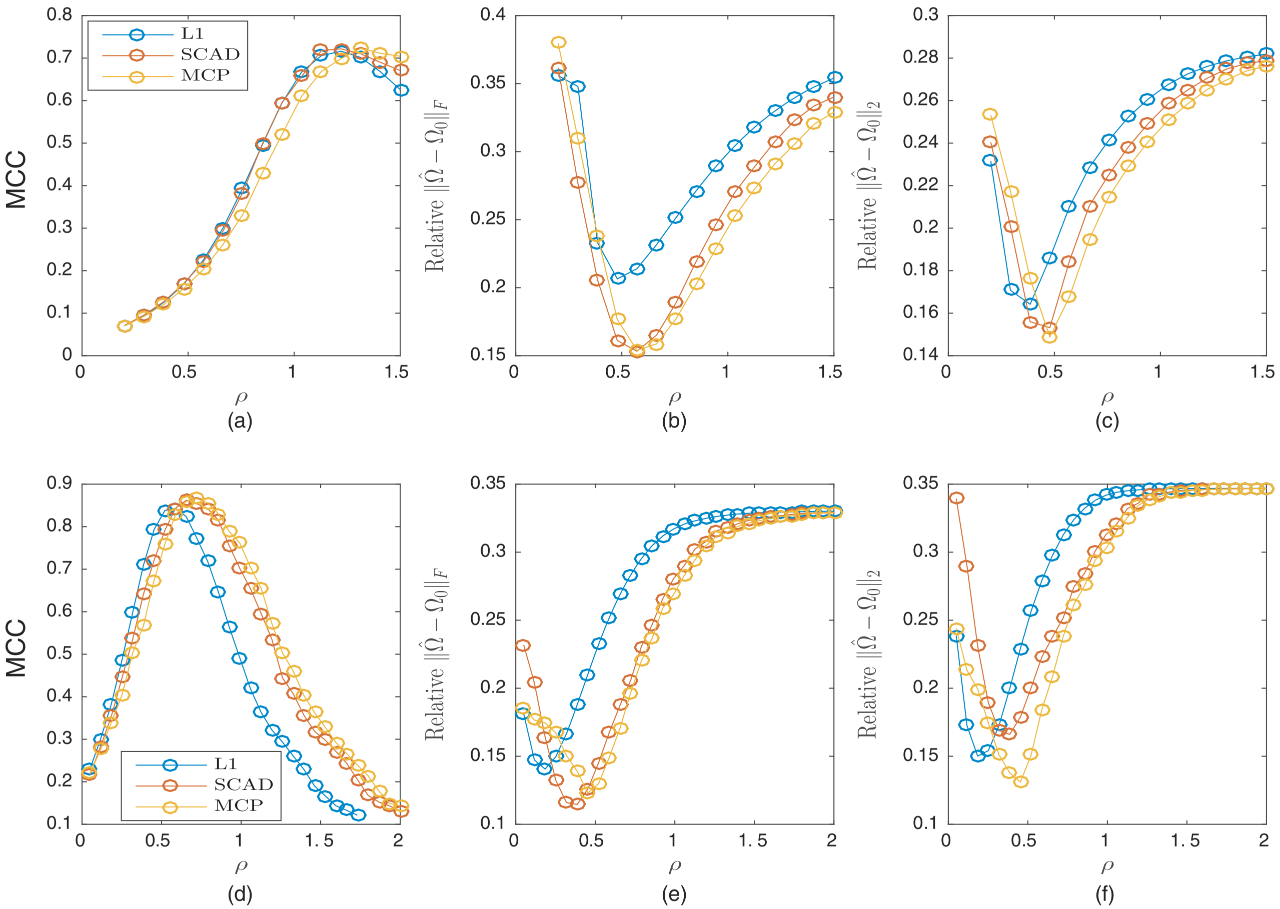
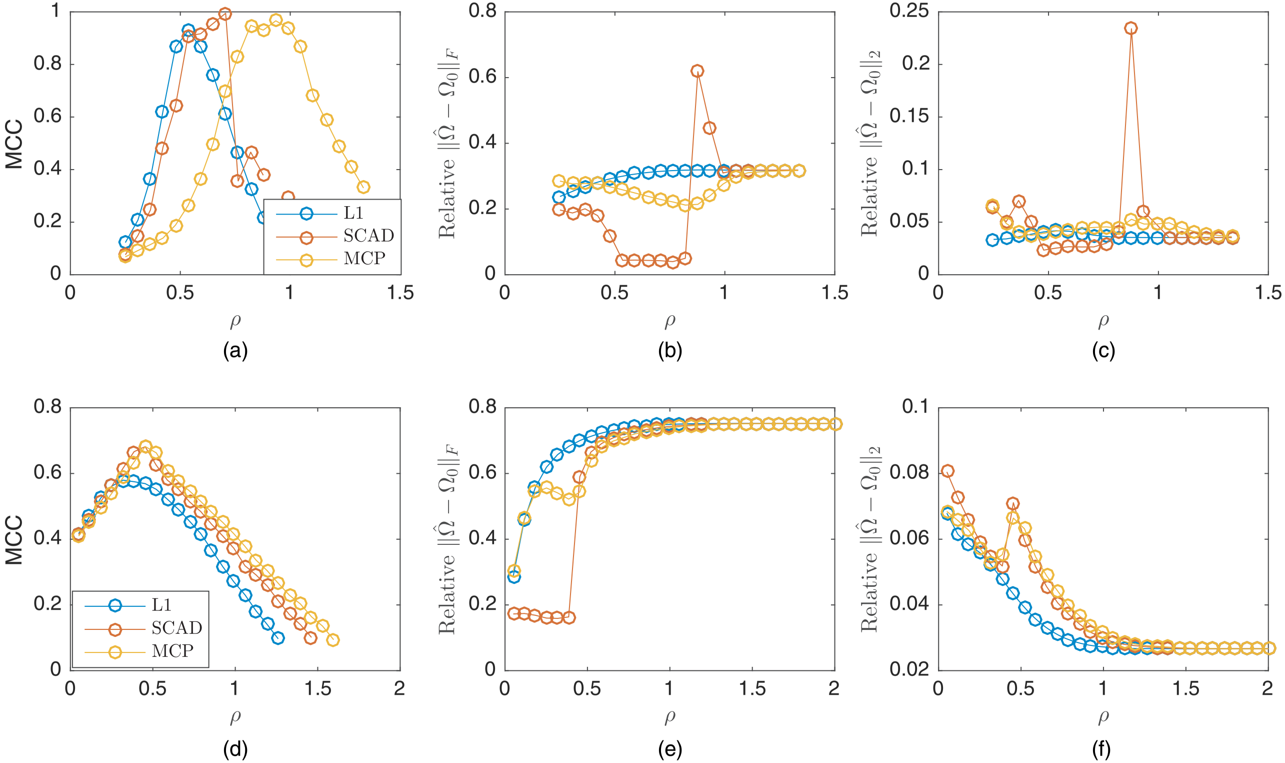


Fig. 7. Non-convex regularizers in the single-sample regime ($n = 1$, Ψ_k ER with d_k edges) (shown are the MCC, relative Frobenius error and relative L2-error as a function of ρ; note that non-convex regularization improves performance): (a)–(c) $K = 2$, $d_1 = d_2 = 1024$; (d)–(f) $K = 3$, $d_1 = d_2 = d_3 = 32$



(a) (b) (c)

(d) (e) (f)

Fig. 8. Non-convex regularizers with spiked identity factors Ψ_k (shown are the MCC and relative Frobenius error as a function of ρ; note that non-convex regularization improves performance when ρ is chosen correctly): (a)–(c) $K = 2$, $d_1 = d_2 = 256$, $n = 10$, $\Psi_k = 0.5I_{d_k} + 0.5(\mathbf{1}_8; \mathbf{0}_{248})(\mathbf{1}_8; \mathbf{0}_{248})^{\mathrm{T}}$; (d)–(f) $K = 3$, $d_1 = d_2 = d_3 = 32$, $n = 1$, $\Psi_k = 0.5I_{d_k} + 0.5(\mathbf{1}_{16}; \mathbf{0}_{16})(\mathbf{1}_{16}; \mathbf{0}_{16})^{\mathrm{T}}$

estimating $K=2$ and $K=3$ ER graphs as functions of regularization parameter ρ for each of l_1, SCAD (96) and MCP (97) regularizers in a variety of configurations. Fig. 8 shows similar results for Ψ_k a variant of the spiked identity model of Loh and Wainwright (2017). Observe that non-convex regularization improves performance slightly, not only for structure estimation (MCC) but for the Frobenius norm error (due to the reduction in bias) as well. This improvement is increased in the spiked identity case.

7. National Center for Environmental Prediction wind speed data

The TeraLasso model is illustrated on a meteorological data set. The US National Center for Environmental Prediction maintains records of average daily wind velocities in the lower troposphere, with daily readings beginning in 1948. The data are available on line from `ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.dailyavgs/surface`. Velocities are recorded globally, in a 144 × 73 latitude–longitude grid with spacings of 2.5° in each co-ordinate. Over bounded areas, the spacing is approximately a rectangular grid, suggesting a $K=2$ model (latitude *versus* longitude) for the spatial covariance, and a $K=3$ model (latitude *versus* longitude *versus* time) for the full spatiotemporal covariance.

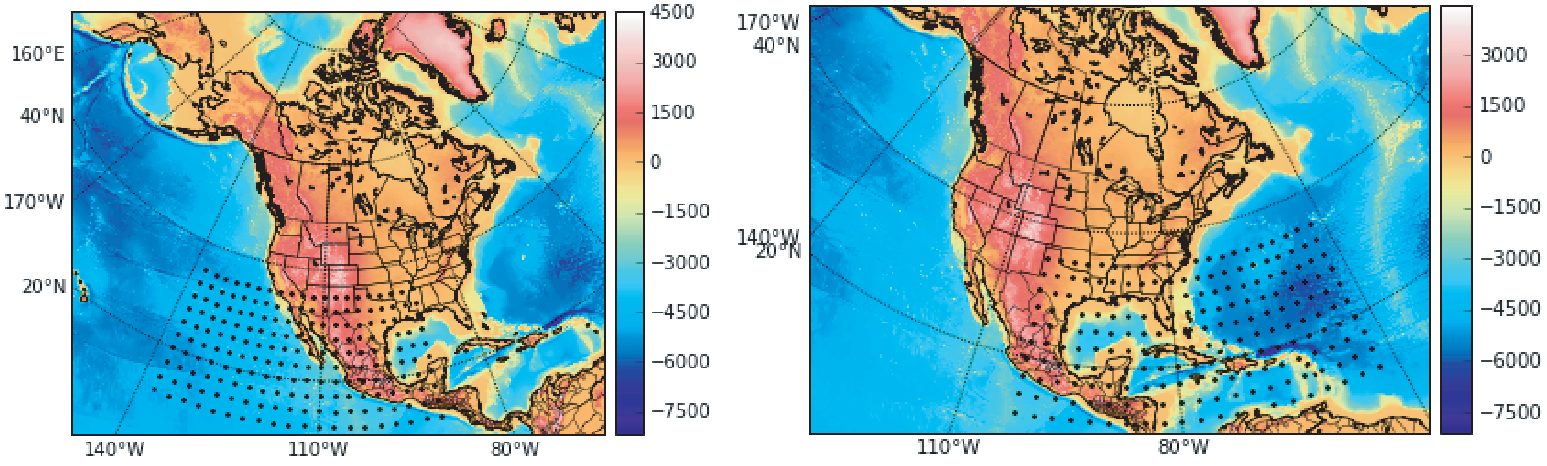
Consider the time series of daily average wind speeds. Following Tsiligkaridis and Hero (2013), we regress out the mean for each day in the year via a 14th-order polynomial regression on the entire history from 1948 to 2015. We extract two 20 × 10 spatial grids, one from eastern North America, and one from western North America (Fig. 9). Figs 10 and 11 show the TeraLasso estimates for latitude and longitude factors by using time samples from January in n years following 1948, for both the eastern and the western grids. Observe the approximate AR(1) structure, and the break in correlation (Fig. 11(b)) in the western longitude factor. The location of this break corresponds to the high elevation line of the Rocky Mountains. In the on-line supplement, we compare the TeraLasso estimator with the unstructured shrinkage estimator, the non-sparse Kronecker sum estimator (TeraLasso estimator with sparsity parameter $\rho=0$) and the Gemini sparse Kronecker product estimator of Zhou (2014). It is shown that TeraLasso provides a significantly better fit to the data.

To illustrate the utility of the estimated precision matrices, we use them to construct a season classifier. National Center for Environmental Prediction wind speed records are taken from the 51-year span from 1948 to 2009. We estimate spatial precision matrices on n consecutive days in January and June of a training year, and running anomaly detection on $m=30$-day sequences of observations in the remaining 50 testing years. We report average classifier performance by averaging over all 51 possible partitions of the 51-year data into one training and 50 testing years. The sequences are labelled as summer (June) and winter (January), and we compute the classification error rate for the winter *versus* summer classifier obtained by choosing the season that is associated with the larger of the likelihood functions

$$\log|\hat{\Omega}_{\text{summer}}| - \sum_{i=1}^{m} (\mathbf{x}_i - \mu_i)^{\mathrm{T}} \hat{\Omega}_{\text{summer}} (\mathbf{x}_i - \mu_i),$$

$$\log|\hat{\Omega}_{\text{winter}}| - \sum_{i=1}^{m} (\mathbf{x}_i - \mu_i)^{\mathrm{T}} \hat{\Omega}_{\text{winter}} (\mathbf{x}_i - \mu_i).$$

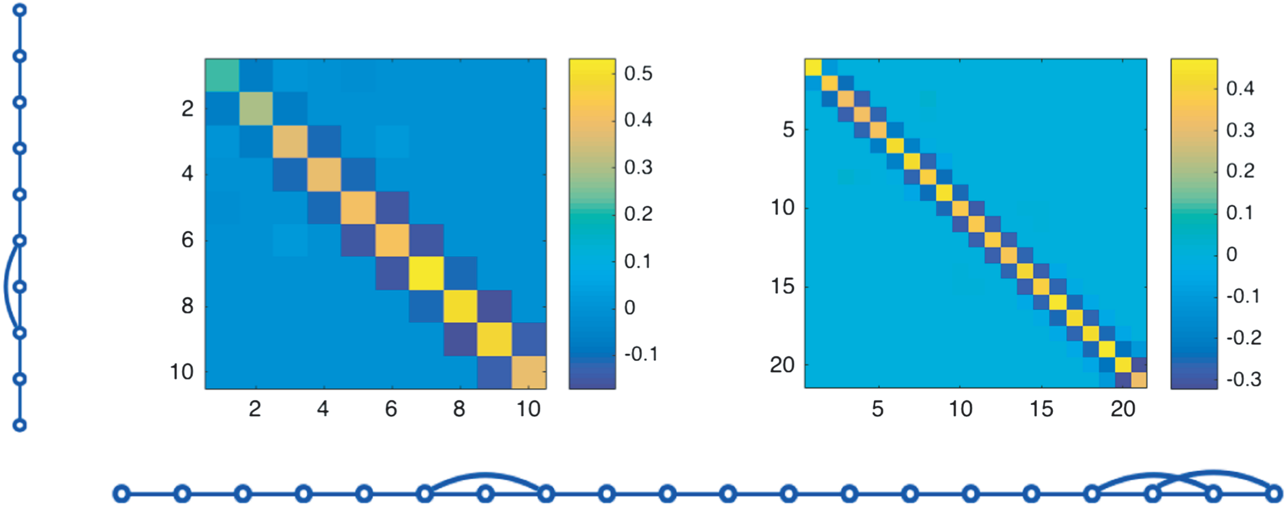
We consider the $K=3$ spatial–temporal precision matrix for a spatial–temporal array of size $10 \times 20 \times T$, with the first (10×10)-factor corresponding to the latitude axis of the spatial array, the second a (20×20)-factor corresponding to the longitude axis and the third factor a $(T \times T)$-factor corresponding to a temporal axis of length T. The spatial–temporal array is created by concatenating T temporally consecutive 10×20 spatial samples. We use l_1-regularization.



(a)

(b)

Fig. 9. Rectangular 10 × 20 latitude–longitude grids of wind speed locations (●) (the elevation colour map is in metres): (a) 'western grid'; (b) 'eastern grid'



(a) (b)

Fig. 10. TeraLasso estimate factors, $K = 2$—eastern grid: (a) graphical representation of latitude (left, 10 nodes) and (b) longitude factors (bottom, 20 nodes) with the corresponding precision estimates (note the simple AR(1)-type structure in (b))

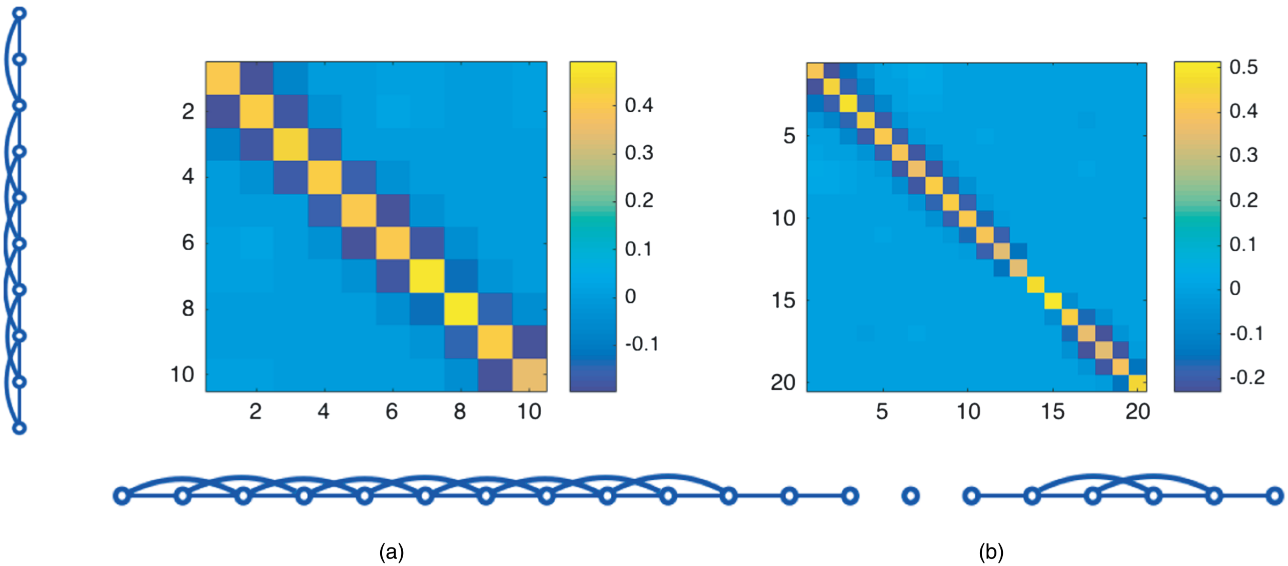


Fig. 11. TeraLasso estimate factors, $K = 2$—western grid: graphical representation of (a) latitude (left) and (b) longitude factors (bottom) with the corresponding precision estimates (observe that the decorrelation (longitude factor entries connecting nodes 1–13 to nodes 14–20 are essentially 0) in the western longitudinal factor, corresponding to the high elevation line of the Rocky Mountains)

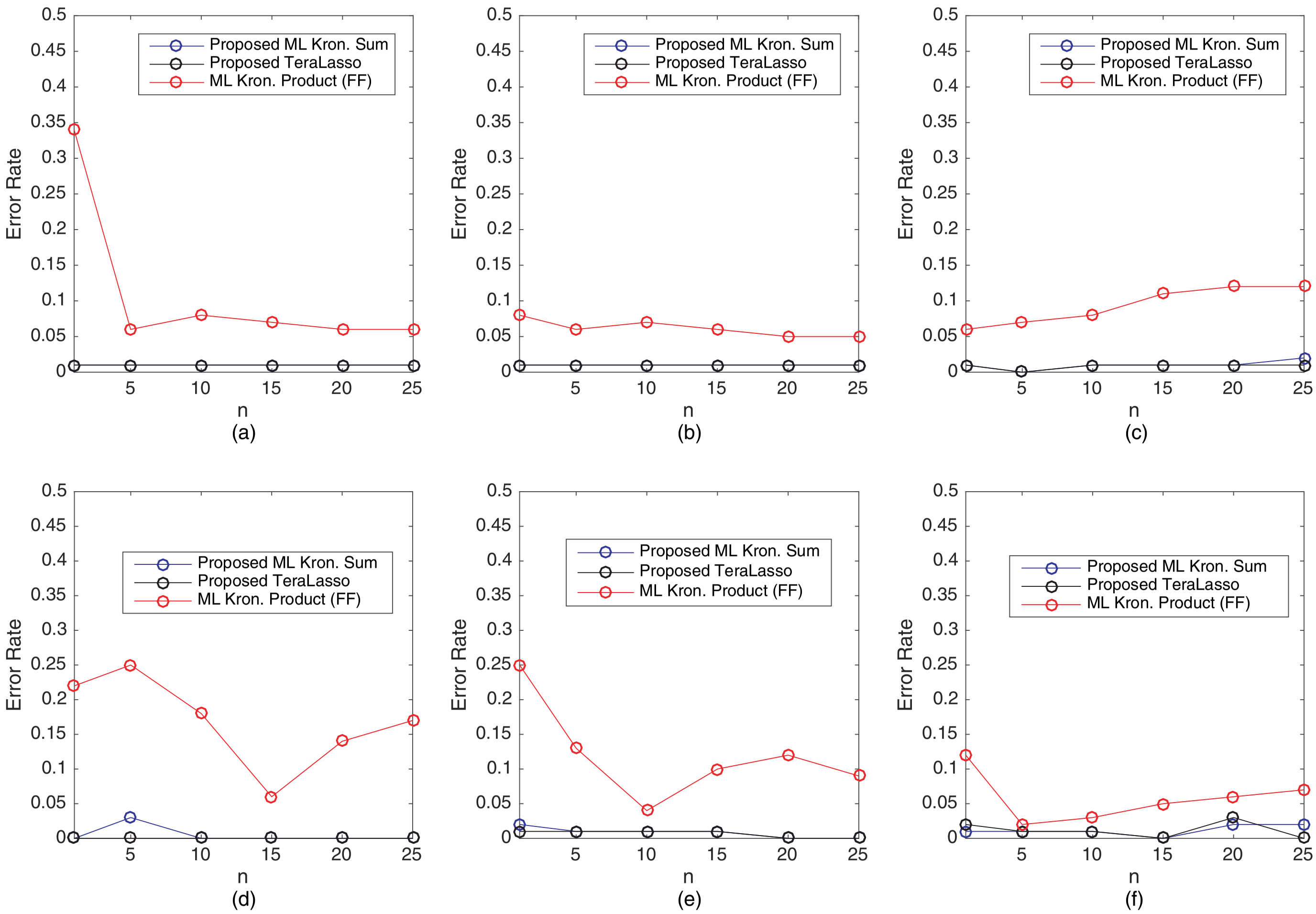


Fig. 12. Classification using Gaussian log-likelihood and estimated spatiotemporal ($K = 3$) precision matrices for each season, where T is the temporal dimension in days (shown is wind speed summer *versus* winter classification error rate as a function of sample size n and length of temporal window T; note the stability of the Kronecker sum estimate in the $n = 1$ case with low error rate): (a)–(c) eastern grid; (d)–(f) western grid; (a), (d) $T = 5$; (b), (e) $T = 10$; (c), (f) $T = 15$

Results for various sized temporal covariance extents ($T = d_3$) are shown in Fig. 12 for TeraLasso, with the unregularized TeraLasso (maximum likelihood Kronecker sum) and maximum likelihood Kronecker product estimator (Werner *et al.*, 2008; Tsiligkaridis *et al.*, 2013) results shown for comparison. In this experiment, we use the maximum likelihood Kronecker product estimator instead of Gemini, as for this maximum likelihood classification task the maximum-likelihood-based approach performs significantly better than the factorwise objective approach of the Gemini estimators, which is not surprising as the Kronecker product is not a good fit for these data (section 3.4 of the on-line supplement). Note the superior performance and increased single-sample robustness of the proposed maximum likelihood Kronecker sum and TeraLasso estimates compared with the Kronecker product estimate, confirming the better fit of TeraLasso. In each case, the non-monotonic behaviour of the Kronecker product curves is due partly to randomness that is associated with the small test sample size, and partly because the Kronecker product in $K = 3$ has overly strong coupling across tensor directions, giving large bias.

8. Conclusion

A factorized model, called TeraLasso, is proposed for the precision matrix of tensor-valued data that uses Kronecker sum structure and sparsity to regularize the precision matrix estimate. An optimization algorithm like iterative soft thresholding is presented that scales to high dimensions. Statistical and algorithmic convergences are established for TeraLasso that quantify performance gains relative to other structured and unstructured approaches. Numerical results demonstrate single-sample convergence as well as tightness of the bounds. Finally, an application to real tensor-valued ($K = 3$) meteorological data is considered, where the TeraLasso model is shown to fit the data well and enables improved single-sample performance for estimation and anomaly detection. Future work includes combining first-moment tensor representation methods for mean estimation such as parallel factor analysis (Harshman and Lundy, 1994) with the second-order TeraLasso method that is introduced in this paper for estimating the covariance.

Acknowledgements

The research reported in this paper was partially supported by US Army Research Office grant W911NF-15-1-0479, US Department of Energy grant DE-NA0002534, National Science Foundation grant DMS-1316731 and the Elizabeth Caroline Crosby Research Award from the Advance Program at the University of Michigan.

References

Allen, G. I. and Tibshirani, R. (2010) Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Statist.*, **4**, 764–790.

Andrianov, S. N. (1997) A matrix representation of lie algebraic methods for design of nonlinear beam lines. *AIP Conf. Proc.*, **391**, 355–360.

Augustin, N. H., Musio, M., von Wilpert, K., Kublin, E., Wood, S. N. and Schumacher, M. (2009) Modeling spatiotemporal forest health monitoring data. *J. Am. Statist. Ass.*, **104**, 899–911.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

Beck, A. and Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imgng Sci.*, **2**, 183–202.

Beckermann, B., Kressner, D. and Tobler, C. (2013) An error analysis of Galerkin projection methods for linear systems with tensor product structure. *SIAM J. Numer. Anal.*, **51**, 3307–3326.

Boyd, S. and Vandenberghe, L. (2009) *Convex Optimization*. New York: Cambridge University Press.

Chapman, A., Nabi-Abdolyousefi, M. and Mesbahi, M. (2014) Controllability and observability of network-of-networks via Cartesian products. *IEEE Trans. Autom. Control*, **59**, 2668–2679.

Combettes, P. L. and Wajs, V. R. (2005) Signal recovery by proximal forward-backward splitting. *Multsc. Modlng Simuln*, **4**, 1168–1200.

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

Dorr, F. W. (1970) The direct solution of the discrete Poisson equation on a rectangle. *SIAM Rev.*, **12**, 248–263.

Eilers, P. H. and Marx, B. D. (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometr. Intell. Lab. Syst.*, **66**, 159–174.

Ellner, N. S. (1986) New ADI model problem applications. In *Proc. Association for Computing Machinery Fall Jt Computer Conf.*, pp. 528–534. Institute of Electrical and Electronics Engineers Computer Society Press.

Faber, N. K. M., Bro, R. and Hopke, P. K. (2003) Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometr. Intell. Lab. Syst.*, **65**, 119–137.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fey, M., Lenssen, J. E., Weichert, F. and Müller, H. (2018) Splinecnn: fast geometric deep learning with continuous b-spline kernels. In *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 869–877. Institute of Electrical and Electronics Engineers.

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

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

Greenewald, K. and Hero, A. (2015) Robust Kronecker product PCA for spatio-temporal covariance estimation. *IEEE Trans. Signl Process.*, **63**, 6368–6378.

Greenewald, K., Park, S., Zhou, S. and Giessing, A. (2017) Time-dependent spatially varying graphical models, with application to brain fMRI data analysis. In *Advances in Neural Information Processing Systems 30* (eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett), pp. 5832–5840. Red Hook: Curran Associates.

Guillot, D., Rajaratnam, B., Rolfs, B., Maleki, A. and Wong, I. (2012) Iterative thresholding algorithm for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pp. 1574–1582.

Hammack, R., Imrich, W. and Klavžar, S. (2011) *Handbook of Product Graphs*. Boca Raton: CRC Press.

Harshman, R. A. and Lundy, M. E. (1994) PARAFAC: parallel factor analysis. *Computnl Statist. Data Anal.*, **18**, 39–72.

Hoff, P. D. (2016) Equivariant and scale-free Tucker decomposition models. *Baysn Anal.*, **11**, 627–648.

Holland, D., Chang, L., Ernst, T. M., Curran, M., Buchthal, S. D., Alicata, D., Skranes, J., Johansen, H., Hernandez, A. and Yamakawa, R. (2014) Structural growth trajectories and rates of change in the first 3 months of infant brain development. *J. Am. Med. Ass. Neurol.*, **71**, 1266–1274.

Imrich, W., Klavžar, S. and Rall, D. F. (2008) *Topics in Graph Theory: Graphs and Their Cartesian Product*. Boca Raton: Peters–CRC Press.

Johndrow, J. E., Bhattacharya, A. and Dunson, D. B. (2017) Tensor decompositions and sparse log-linear models. *Ann. Statist.*, **45**, 1–38.

Kalaitzis, A., Lafferty, J., Lawrence, N. and Zhou, S. (2013) The bigraphical lasso. In *Proc. Int. Conf. Machine Learning*, pp. 1229–1237. Norristown: Omnipress.

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

Kotzagiannidis, M. S. and Dragotti, P. L. (2019) Splines and wavelets on circulant graphs. *Appl. Computnl Harm. Anal.*, **47**, 481–515.

Kressner, D. and Tobler, C. (2010) Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matr. Anal. Appl.*, **31**, 1688–1714.

Lee, D.-J. and Durbán, M. (2011) P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statist. Modllng*, **11**, 49–69.

Leng, C. and Tang, C. Y. (2012) Sparse matrix graphical models. *J. Am. Statist. Ass.*, **107**, 1187–1200.

Loh, P.-L. and Wainwright, M. J. (2013) Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484. Red Hook: Curran Associates.

Loh, P.-L. and Wainwright, M. J. (2017) Support recovery without incoherence: a case for nonconvex regularization. *Ann. Statist.*, **45**, 2455–2482.

Luenberger, D. (1966) Observers for multivariable systems. *IEEE Trans. Autom. Control*, **11**, 190–197.

Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.

Nesterov, Y. (1983) A method of solving a convex programming problem with convergence rate o (1/k2). *Sov. Math. Dokl.*, **27**, 372–376.

Nesterov, Y. (2004) *Introductory Lectures on Convex Optimization: Applied Optimization*. Boston: Kluwer.

Nesterov, Yu. (2013) Gradient methods for minimizing composite objective function. *Math. Programmng*, **140**, 125–161.

Pouryazdian, S., Beheshti, S. and Krishnan, S. (2016) CANDECOMP/PARAFAC model order selection based on reconstruction error in the presence of Kronecker structured colored noise. *Digtl Signl Process.*, **48**, 12–26.

Preisler, H. K., Hicke, J. A., Ager, A. A. and Hayes, J. L. (2012) Climate and weather influences on spatial temporal patterns of mountain pine beetle populations in Washington and Oregon. *Ecology*, **93**, 2421–2434.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008) Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, **2**, 494–515.

Rudelson, M. and Zhou, S. (2017) High dimensional errors-in-variables models with dependent measurements. *Electron. J. Statist.*, **11**, 1699–1797.

Schmitt, U., Louis, A. K., Darvas, F., Buchner, H. and Fuchs, M. (2001) Numerical aspects of spatio-temporal current density reconstruction from EEG-/MEG-data. *IEEE Trans. Med. Imgng*, **20**, 314–324.

Shi, X., Wei, Y. and Ling, S. (2013) Backward error and perturbation bounds for high order Sylvester tensor equation. *Lin. Multlin. Alg.*, **61**, 1436–1446.

Tseng, P. (2010) Approximation accuracy, gradient methods, and error bound for structured convex optim. *Math. Progrmmng*, **125**, 263–295.

Tsiligkaridis, T. and Hero, A. (2013) Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Trans. Signl. Process.*, **61**, 5347–5360.

Tsiligkaridis, T., Hero, A. and Zhou, S. (2013) On convergence of Kronecker graphical lasso algorithms. *IEEE Trans. Signl Process.*, **61**, 1743–1755.

Tucker, L. R. (1966) Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311.

Van Loan, C. F. (2000) The ubiquitous Kronecker product. *J. Computnl Appl. Math.*, **123**, 85–100.

Werner, K., Jansson, M. and Stoica, P. (2008) On estimation of cov. matrices with Kronecker product structure. *IEEE Trans. Signl. Process.*, **56**, 478–491.

Wood, S. N. (2006) Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, **62**, 1025–1036.

Wood, S. N., Pya, N. and Saefken, B. (2016) Smoothing parameter and model selection for general smooth models. *J. Am. Statist. Ass.*, **111**, 1548–1563.

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

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zhou, S. (2014) Gemini: graph estimation with matrix variate normal instances. *Ann. Statist.*, **42**, 532–562.

Zhou, S., Lafferty, J. and Wasserman, L. (2010) Time varying undirected graphs. *Mach. Learn.*, **80**, 295–319.

Zhou, S., Rütimann, P., Xu, M. and Bühlmann, P. (2011) High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.*, **12**, 2975–3026.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material for Tensor graphical lasso (TeraLasso)’.