JRSS Ser B
B19012 -
Greenewald
et al.

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

$$\Psi_1 \oplus \cdots \oplus \Psi_K = \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \Psi_k \otimes I_{[d_{k+1:K}]}. \tag{1}$$

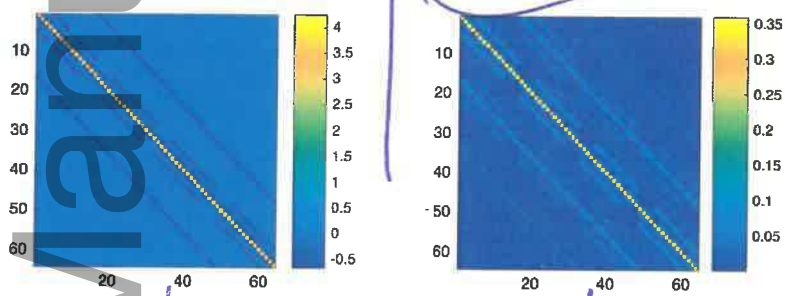
In the special case of $K = 2$ this Kronecker sum representation reduces to the more familiar $\Psi_1 \oplus \Psi_2 = \Psi_1 \otimes I_{d_1} + I_{d_2} \otimes \Psi_2$. The vectorization of a K-way tensor X is denoted as $\mathrm{vec}(X)$ and is defined as in Kolda and Bader (2009). Likewise, we define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

enlarge to
width
33 picas

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 has a multivariate Gaussian distribution, the sparsity pattern of Ψ_k corresponds to a conditional independence graph across the k-th dimension of the data.

Figure 1 illustrates the Kronecker sum model proposed in (1) for $K = 3$ and $d_k = 4$. Specifically, $\Psi_k, k = 1, 2, 3$ are identical 4×4 tridiagonal precision matrices corresponding to a one dimensional autoregressive-1 (AR-1) process. In the Figure the precision matrix $\Omega = \Psi_1 \oplus \Psi_2 \oplus \Psi_3$ is shown on the left and covariance $\Sigma = \Omega^{-1}$ on the right. The entries of each Ψ_k are replicated $m_k =$

less #



(a) (b)

Insert (a), (b)
in 8pt Helvetica
final size

Fig. 1: Illustration of the Kronecker sum model for a tensor valued AR(1) process. Left: Sparse $4 \times 4 \times 4$ 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. Right: Covariance matrix $\Sigma = \Omega^{-1}$. Unlike the Kronecker product precision model, the nested block structure in Σ is not reperesentable by a product of component factors.

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 nonseparable and interlocking repeating block structure in the covariance matrix.

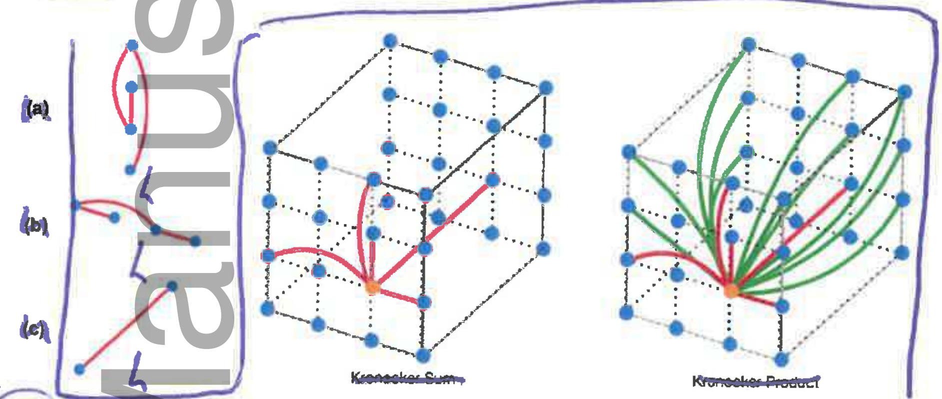
We propose the following sparse Kronecker sum estimator of the precision matrix Ω in (1), which we call the Tensor Graphical Lasso (TeraLasso). The

JRSS Ser B
B19012 -
Greenewald
et al.

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 one 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 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 $K \geq 2$ 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 supplement) bounds on the convergence rate for the known-diagonal setting of Kalaitzis et al. (2013).

Width
33 picas



(c)
(d)
(a)/(e)

Insert (a)-(e)
in 8pt
Helvetica
final size

(b)/

Fig. 2: Comparison of Kronecker sum (Cartesian product graph) at center and Kronecker product (direct product graph) at right. The products are formed from the component graphs shown in (a)-(b)-(c); the number of factors in the product graphs is $K = 3$ and the dimensions are $d_1 = d_2 = 4$, $d_3 = 2$, leading to product graphs with 32 nodes, arranged in a regular 3 dimensional grid in the figures at 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 while the Kronecker product model is much less sparse, having a total of 184 edges.

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 while the Kronecker sum model corresponds

JRSS Ser B
B19012 -
Greenewald
et al.

where $q'_\rho(t) = \frac{d}{dt}(g_\rho(t) - \rho|t|)$ 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 $\|\Omega\|_2 \leq \kappa$ constraint, e.g. via step size line search. The algorithm is summarized in Algorithm 2 in Supplement 2.1.

THEOREM 5 (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. Follows since for $\kappa \leq \sqrt{2/\mu}$ the objective (8) is convex on the convex constraint set $\{\Omega \in \mathcal{K}_\mathbf{p} | \Omega \succ 0, \|\Omega\|_2 \leq \kappa\}$ (Lemma 21, supplement).

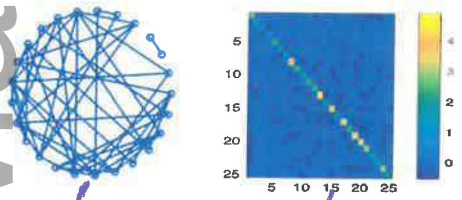
enlarge
to width
33 picas

6. Validation on synthetic data

Random graphs were created for each factor Ψ_k using both an Erdos-Renyi (ER) topology and a random grid graph topology¶. 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 $[\Psi_k]_{ii}$, $[\Psi_k]_{jj}$ by a. This keeps Ψ_k positive definite. We repeat this process until all edges are added. Finally, we form $\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K$. An example 25-node, $q = 25$ ER graph and precision matrix are shown in Figure 3. The random grid graph

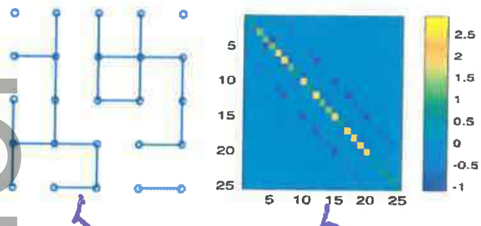
Insert (a)-(d)
in 8pt
Helvetica
final size

(a) (b)



(a) Random Erdos-Renyi (ER) graph with 25 nodes and 50 edges

(c) (d)



(b) Random grid graph (square) with 25 nodes and 26 edges

Fig. 3: Example Erdos-Renyi and random grid graphs. Left: Graphical representation. Right: Corresponding precision matrix Ψ.

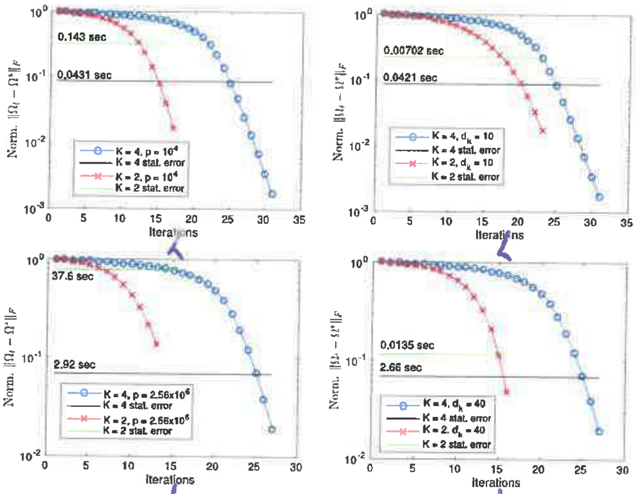
¶Code for experiments can be found at https://github.com/kgreenewald/teralasso.

enlarge to landscape width 40 picas

JRSS Ser-B B19012 — Greenewald et al.

~~(a) $\ell 1$ penalty, $n = 100$ sample size~~

~~(b) $\ell 1$ penalty, $n = 1$ sample size~~



(a)

(b)

(c)

(d)

Insert (a)-(d) in 8pt Helvetica final size

Fig. 4: ~~Linear geometric convergence of the convex ($\ell 1$ 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^*. On the left are results comparing $K = 2$ and $K = 4$ on the same data with the same value of p (different d_k), on the right they are compared for the same value of d_k (different p). 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.

Shown in Figure 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 giving $\rho_k = \frac{\bar{\rho}_k}{\sqrt{(\log p)/(nm_k)}}$. Note $\bar{\rho}_1 = \bar{\rho}_2 = \bar{\rho}_3$ achieves near optimal results.

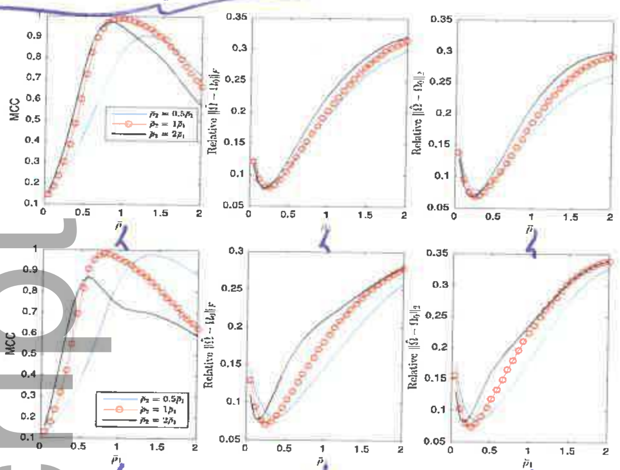
Having verified the single tuning parameter approach, hereafter we will cross-validate only $\bar{\rho}$. In supplement Section 3.3, we provide experimental verification in a wide variety of experimental settings (including varying the relative size of the tensor dimensions d_k) that our bounds on the rate of convergence for the $\ell 1$ regularized model are tight. Figure 6 illustrates how increasing dimension p and K improves single sample performance. Shown are the average TeraLasso edge detection precision and recall values for different values of K in the single and 5-sample regimes, all increasing to 1 (perfect structure estimation) as p, K, and n increase.

6.3. Nonconvex Regularization

Here the $\ell 1$ penalized TeraLasso is compared to TeraLasso with nonconvex regularization (8). Shown in Figure 7 are the MCC, normalized Frobenius error, and spectral norm error for estimating $K = 2$ and $K = 3$ Erdos-Renyi graphs

landscape – enlarge to width 40 picas

JRSS Ser B B19012 – Greenewald et al.



(a) (b) (c) (d) (e) (f)

Insert (a)-(f) in 8pt Helvetica final size

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. Shown are deviations of $\bar{\rho}_2$ from the theoretically dictated $\bar{\rho}_2 = \bar{\rho}_1 = \bar{\rho}_3$. Top: Equal dimensions, $d_1 = d_2 = d_3$. First and third factors are random ER graphs with d_k edges, and the second factor is random grid graph with $d_k/2$ edges. Bottom: Dimensions $d_2 = 2d_1$, each factor is a random ER graph with d_k edges. Notice in these scenarios that using $\bar{\rho}_1 = \bar{\rho}_2$ is near optimal, as theoretically predicted.

as functions of regularization parameter ρ for each of $\ell 1$, SCAD (96), and MCP (97) regularizers in a variety of configurations. Figure 8 shows similar results for Ψ_k a variant of the spiked identity model of Loh et al. (2017). Observe that nonconvex 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. NCEP Windspeed Data

The TeraLasso model is illustrated on a meteorological dataset. The US National Center for Environmental Prediction (NCEP) maintains records of average daily wind velocities in the lower troposphere, with daily readings beginning in 1948. The data is available online at 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 degrees in each coordinate. Over bounded areas, the spacing is approximately a rectangular grid, suggesting a $K = 2$ model (latitude vs. longitude) for the spatial covariance, and a $K = 3$ model (latitude vs. longitude vs. time) for the full spatio-temporal covariance.

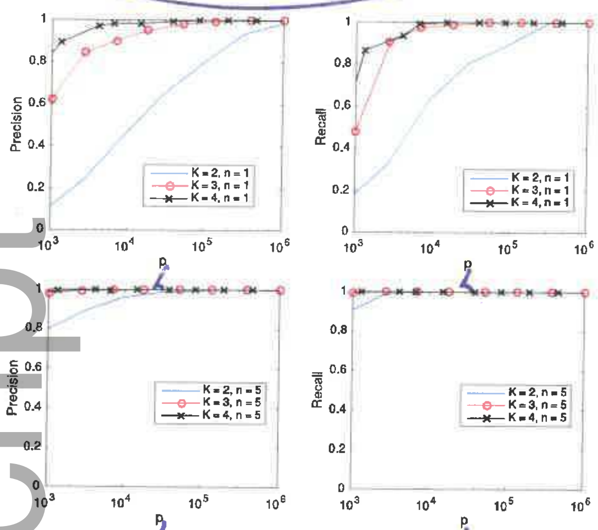
Consider the time series of daily-average wind speeds. Following Tsiligkaridis

Landscape - enlarge to width 40 picas

JRSS Ser B
B19012 - Greenewald et al.

~~(a) $n = 1$~~ (a)

~~(b) $n = 5$~~ (b)



(c) (d)

Insert (a)-(d) in 8pt Helvetica final size

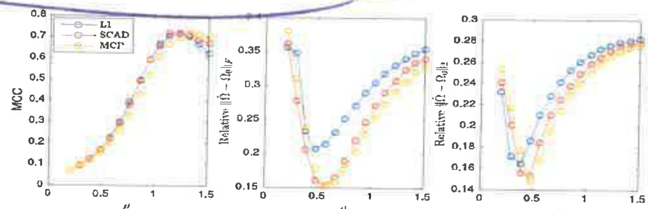
Fig. 6: Edge support estimation on random ER graphs, with the ρ_k set according to Theorem 1. Graphical model edge detection precision and recall curves are shown as a function of data dimension $p = \prod_{k=1}^{K} d_k$. For each value of the tensor order K, we set $d_k = p^{1/K}$. Observe single sample convergence as the dimension p increases and as increasing K creates additional structure.

and Hero (2013), we regress out the mean for each day in the year via a 14-th order polynomial regression on the entire history from 1948-2015. We extract two 20×10 spatial grids, one from eastern North America, and one from western North America (Figure 9). Figure 10 shows the TeraLasso estimates for latitude and longitude factors using time samples from January in n years following 1948, for both the eastern and western grids. Observe the approximate AR structure, and the break in correlation (Figure 10 (b), longitude factor) in the Western Longitude factor. The location of this break corresponds to the high elevation line of the Rocky Mountains. In the supplement, we compare the TeraLasso estimator to 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 the 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. NCEP windspeed records are taken from the 51-year span from 1948-2009. We estimate spatial precision matrices on n consecutive days in January and June of a training year respectively, 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 1 training and 50 testing years.

landscape – enlarge to width 40 picas

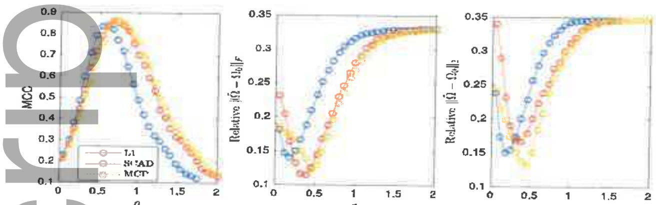
JRSS Ser B
B19012 –
Greenewald et al



(a)L (b)L (c)L

(a) $K = 2, d_1 = d_2 = 1024$

Insert (a)–(f) in 8pt Helvetica final size



(d)L (e)L (f)L

(b) $K = 3, d_1 = d_2 = d_3 = 32$

Fig. 7: Nonconvex 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 nonconvex regularization improves performance.

The sequences are labeled as summer (June), and winter (January), and we compute the classification error rate for the winter vs. summer classifier obtained by choosing the season associated with the larger of the likelihood functions

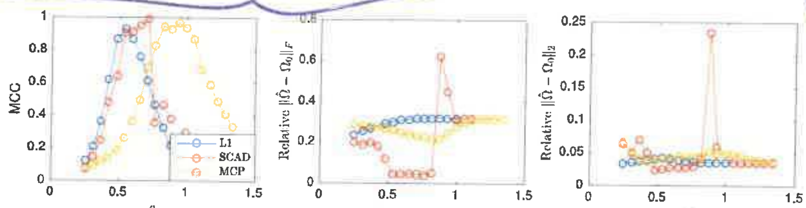
$$\log|\widehat{\Omega}_{\text{summer}}| - \sum_{i=1}^{m}(\mathbf{x}_i - \mu_i)^T\widehat{\Omega}_{\text{summer}}(\mathbf{x}_i - \mu_i)$$
$$\log|\widehat{\Omega}_{\text{winter}}| - \sum_{i=1}^{m}(\mathbf{x}_i - \mu_i)^T\widehat{\Omega}_{\text{winter}}(\mathbf{x}_i - \mu_i).$$

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 $\ell 1$ regularization.

Results for different sized temporal covariance extents ($T = d_3$) are shown in Figure 11 for TeraLasso, with unregularized TeraLasso (ML 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 ML Kronecker product estimator instead of the 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 this data (Section 3.4 of the supplement). Note the superior perfor-

landscape – enlarge to width 40 picas

JRSS Ser B B19012 – Greenewald et al.

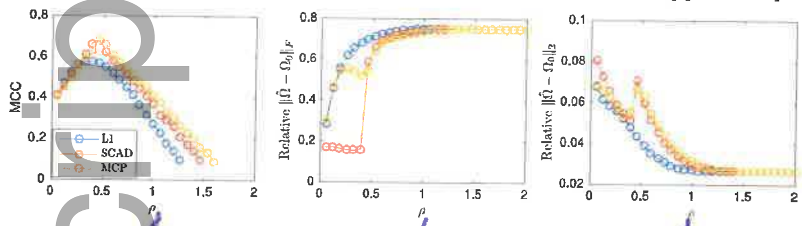


(a)/

~~(a) $K = 2$, $d_1 = d_2 = 256$, $n = 10$, $\Psi_k = 0.5 I_{d_k} + 0.5 [1_8; 0_{248}] [1_8; 0_{248}]^T$~~

(b)/ (c)/

Insert (a)–(f) in 8pt Helvetica final size



(d)/

~~(b) $K = 3$, $d_1 = d_2 = d_3 = 32$, $n = 1$, $\Psi_k = 0.5 I_{d_k} + 0.5 [1_{16}; 0_{16}] [1_{16}; 0_{16}]^T$~~

(e)/ (f)/

Fig. 8: Nonconvex regularizers with spiked identity factors Ψ_k. Shown are the MCC and relative Frobenius error as a function of ρ. Note nonconvex regularization improves performance when ρ is chosen correctly.

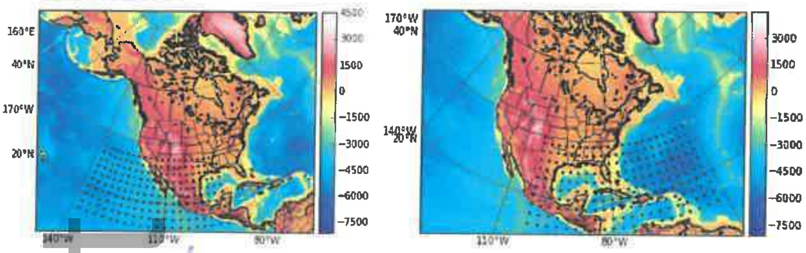
mance and increased single sample robustness of the proposed ML Kronecker Sum and TeraLasso estimates as compared to the Kronecker product estimate, confirming the better fit of TeraLasso. In each case, the nonmonotonic behavior of the Kronecker product curves is due partly to randomness associated with the small test sample size, and partly due to the fact that the Kronecker product in $K = 3$ has overly strong coupling across tensor directions, giving large bias.

8. Conclusion

A factorized model, called the 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 ISTA-like optimization algorithm is presented that scales to high dimensions. Statistical and algorithmic convergence are established for the 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 enable improved single-sample performance for estimation and anomaly detection. Future work includes combining first moment tensor representation methods for mean estimation such as PARAFAC (Harshman and Lundy, 1994) with the second order TeraLasso method introduced in this paper for estimating the covariance.

landscape - enlarge to width 48 picas

JRSS Ser B B19012 - Greenewald et al.



(a)

(b)

Fig. 9: Rectangular 10 × 20 latitude-longitude grids of windspeed locations shown as black dots. Elevation colormap shown in meters. Left: "Western grid", Right: "Eastern grid".

Insert (a), (b) in 8pt Helvetica final size

9. Acknowledgement

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, NSF 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. *The Annals of Applied Statistics*, **4**, 764–790.

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

Augustin, N. H., Musio, M., von Wilpert, K., Kublin, E., Wood, S. N. and Schumacher, M. (2009) Modeling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association*, **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. *The Journal of Machine Learning Research*, **9**, 485–516.

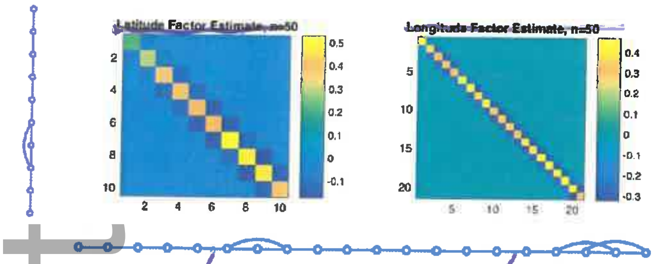
Beck, A. and Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**, 183–202.

Beckermann, B., Kressner, D. and Tobler, C. (2013) An error analysis of galerkin projection methods for linear systems with tensor product structure. *SIAM Journal on Numerical Analysis*, **51**, 3307–3326.

JRSS Ser B
B14012 – Greenwald et al.

Both figs landscape
– enlarge to width 40 picas

Fig. 10



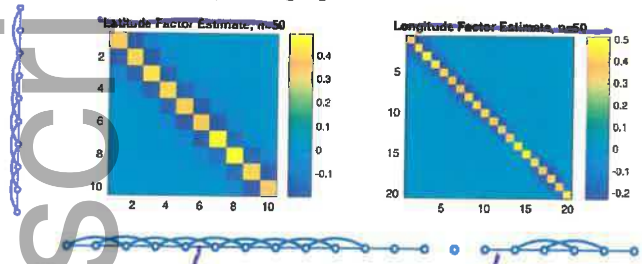
(a)

(a) Eastern grid. Graphical representation of latitude (left, 10 nodes) and longitude factors (bottom, 20 nodes) with the corresponding precision estimates. Note the simple AR-1 type structure of the longitude graph.

(b)

Insert (a), (b)
in 8pt Helvetica
final size

Fig. 11



(a)

(b)

(b) Western grid. Graphical representation of latitude (left) and longitude factors (bottom) with the corresponding precision estimates. Observe the decorrelation (longitude factor entries connecting nodes 1-13 to nodes 14-20 are essentially zero) in the Western longitudinal factor, corresponding to the high-elevation line of the Rocky Mountains.

~~Fig. 10: TeraLasso estimate factors, $K = 2$.~~

Boyd, S. and Vandenberghe, L. (2009) *Convex optimization*. Cambridge university press.

Chapman, A., Nabi-Abdolyousefi, M. and Mesbahi, M. (2014) Controllability and observability of network-of-networks via cartesian products. *IEEE Transactions on Automatic Control*, **59**, 2668–2679.

Combettes, P. L. and Wajs, V. R. (2005) Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, **4**, 1168–1200.

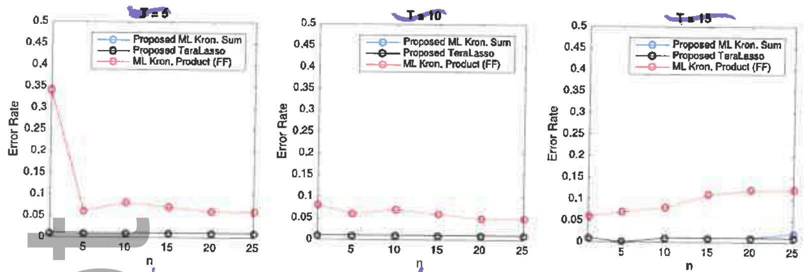
Dawid, A. P. (1981) Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, **68**, 265–274.

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

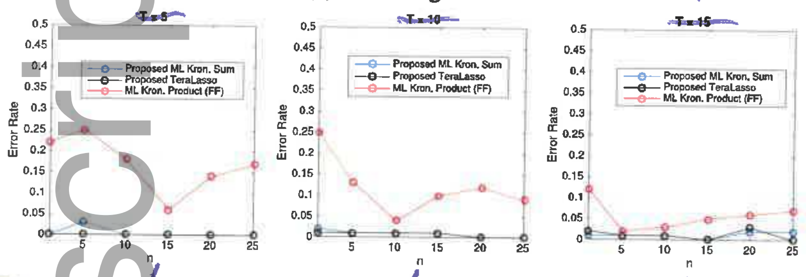
Eilers, P. H. and Marx, B. D. (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, **66**, 159–174.

landscape - enlarge to width 48 picas

JRSS Ser B B19012 - Greenewald et al.



(a) Eastern grid



(b) Western grid

Insert (a)-(f) in 8pt Helvetica final size

Fig. ~~11~~ 12: Classification using Gaussian loglikelihood and estimated spatio-temporal ($K = 3$) precision matrices for each season, where T is the temporal dimension in days. Shown is windspeed summer vs. 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.

Ellner, N. S. et al. (1986) New ADI model problem applications. In *Proceedings of 1986 ACM Fall joint computer conference*, 528–534.

Faber, N. K. M., Bro, R. and Hopke, P. K. (2003) Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, **65**, 119–137.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348–1360.

Fey, M., Eric Lenssen, J., Weichert, F. and Müller, H. (2018) Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 869–877.

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 ap-

J. R. Statist. Soc. B (2019)
81, Part 5, pp. 000–000

SI 32 COL
JRSS Ser B
B19012 – Greenewald et al.

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. This paper introduces a multi-way tensor generalization of the Bigraphical Lasso (BiGLasso), which uses a two-way sparse Kronecker-sum multivariate-normal model for the precision matrix to parsimoniously model conditional dependence relationships of matrix-variate data based on the Cartesian product of graphs. We call this generalization the Tensor graphical Lasso (TeraLasso). We demonstrate 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 coordinates such as space, time and replicates. Statistical consistency and statistical rates of convergence are established for both the BiGLasso 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 analyze 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 the TeraLasso using both simulation and experimental data from a meteorological dataset, showing that we can accurately estimate precision matrices and recover meaningful conditional dependency graphs from high dimensional complex datasets.

Keywords:

1. Introduction

The increasing availability of matrix and tensor-valued data with complex dependencies 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.

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

fl ns ? Au: confirm

1

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 precision matrix of a K-way data tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ grows as $\prod_{i=1}^{K} d_i^2$. 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 using ℓ_1-penalization methods, such as the graphical Lasso ~~(GLasso)~~ (Friedman et al., 2008) and multiple (nodewise) regressions (Meinshausen et al., 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 formed by the neural pathways. Yet, this line of high dimensional statistical literature mentioned above has primarily focused on estimating linear or graphical models with ~~i.i.d.~~ 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, the cortical circuits can change over time ~~due to~~ activities such as motor learning, attention or visual stimulation. ~~This~~ data typically ~~has~~ a complex structure that is organized by the experiments design, with one or more experimental factors varying according to a predefined pattern.

Gk cap pi, upright Π

by

that is

independent and identically distributed

because of

have

cal

These

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 dependency graphs which arise naturally in the above mentioned contexts when handling data with complex dependencies. 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 a Kronecker product representation. Moreover, (Zhou, 2014) showed that one can estimate the covariance and inverse covariance matrices well using only one instance from the matrix-variate normal distribution. Along the same lines, the Bigraphical Lasso framework was proposed to parsimoniously model 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 modeling of a structured precision matrix

to

by

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 dependencies.

procedure called

Towards these goals, we ~~will~~ shall introduce the tensor graphical Lasso (TeraLasso) ~~procedure~~ for estimating sparse K-way decomposable precision matrices. We ~~will~~ shall show that our concentration of measure analysis enables a significant reduction in the sample size requirement ~~in order~~ to estimate parameters and the associated conditional dependence graphs along different coordinates 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 multi-way Kronecker sum precision matrix model

We follow the notation and terminology of Kolda and Bader (2009) for modeling 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, so:

$$p = \prod_{k=1}^{K} d_k \quad \text{and} \quad m_k = \prod_{i \neq k} d_i = \frac{p}{d_k}.$$

display

To simplify the multiway Kronecker notation, we define

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

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

$$\Psi_1 \oplus \cdots \oplus \Psi_K = \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \Psi_k \otimes I_{[d_{k+1:K}]}. \tag{1}$$

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 $\text{vec}(X)$ and is defined as in Kolda and Bader (2009). Likewise, we define the transpose of a K-way tensor $X^T \in \mathbb{R}^{d_K \times \cdots \times d_1}$ analogously to the matrix transpose, i.e. $[X^T]_{i_1,\ldots,i_K} = X_{i_K,\ldots,i_1}$.

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 ~~has~~ have a multivariate Gaussian distribution, the sparsity pattern of Ψ_k corresponds to a conditional independence graph across the k-th dimension of the data.

Fig. 1 equation

Figure 1 illustrates the Kronecker sum model proposed in (1) for $K = 3$ and $d_k = 4$. Specifically, $\Psi_k, k = 1, 2, 3$ are identical 4×4 tridiagonal precision matrices corresponding to a one-dimensional autoregressive ~~-1 (AR-1)~~ process. (AR(1)) ~~In the Figure~~ the precision matrix $\Omega = \Psi_1 \oplus \Psi_2 \oplus \Psi_3$ is shown ~~on the left~~ in Fig. 1(a) and the covariance $\Sigma = \Omega^{-1}$ ~~on the right~~ in Fig. 1(b). The entries of each Ψ_k are replicated $m_k =$

also sized separately

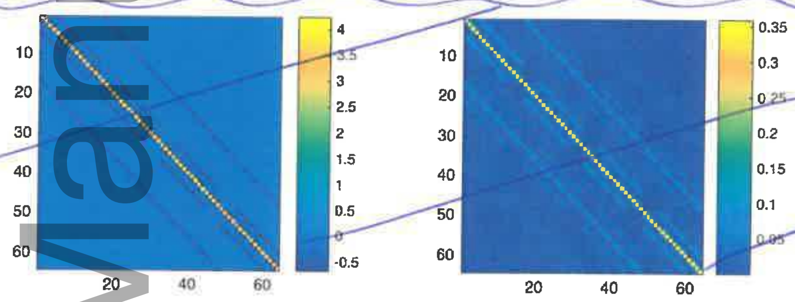


Fig. 1: Illustration of the Kronecker sum model for a tensor valued AR(1) process. ~~Left:~~ Sparse $4 \times 4 \times 4$ precision matrix $\Omega = \Psi_1 \oplus \Psi_2 \oplus \Psi_3$, where Ψ_k are identical tridiagonal precision matrices corresponding to one-dimensional AR(1) models : (a) ~~Right:~~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

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.

equation

We propose the following sparse Kronecker sum estimator of the precision matrix Ω in (1), which we call the Tensor Graphical Lasso (TeraLasso). ~~The~~

TeraLasso minimizes the negative ℓ_1-penalized Gaussian log-likelihood function over the domain $\mathcal{K}_{\mathbf{p}}^\sharp$ of precision matrices Ω having Kronecker sum form

smaller circumflex throughout

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

rom

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

#

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

normal #s

subject to

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

is the set of positive semidefinite matrices that are decomposable into a Kronecker sum of fixed factor dimensions $d_1, \ldots, d_K$. In this paper we consider (μ, γ)-amenable regularizers g_ρ (Loh et al., 2017). The norm constraint $\|\Omega\|_2 \leq \kappa$ is required for the solution to be well defined when g_ρ is not a convex penalty. These penalties includes nonconvex regularizers such as SCAD and MCP, as well as the traditional $\ell 1$ regularizer $g_\rho(t) = \rho|t|$.

smoothly clipped absolute deviation

the minimax convex penalty

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)-loglikelihood promotes a sparse graphical structure in Ω; see for example (Banerjee et al., 2008; Yuan and Lin, 2007; Zhou, 2014; Zhou et al., 2011). In this work, we allow for the more general case of nonconvex regularization functions g_ρ as considered in Loh et al. (2017). ~~While~~ sometimes difficult to tune in practice, nonconvex regularization provides strong nonasymptotic guarantees on the elementwise estimation error of Ω, implying strong, single sample support recovery guarantees when the smallest nonzero element of Ω is bounded from below.

, Although

and

The contributions of this paper are as follows. The sparse multivariate normal Bigraphical Lasso (BiGLasso) model is extended to the sparse tensor-variate ($K > 2$) TeraLasso ~~model~~, allowing the modeling of data with arbitrary tensor degree K. A new subgaussian concentration inequality (Corollary 19 in the 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 ~~one~~). TeraLasso's generalization of BiGlasso from 2-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 3-way tensor decomposition (time×space×people) can account for possible dependency structure between people, ~~while~~ a 2-way BiGLasso or KLasso approach decom-

called

on-line

1

two-

three

whereas

?Au: KLasso

two

posing over (time×space) would unnecessarily enforce an assumption of independence between people. Alternately, 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 three the ~~3~~-way decomposition that TeraLasso uses to give each axis its own factor. on iterative soft thresholding

A highly scalable, first-order ~~ISTA~~-based algorithm is proposed to minimize the TeraLasso objective function. We prove (Theorem 25 in the supplement) on-line that it converges to the global optimum with a geometric convergence we rate, and demonstrate its practical advantages on high dimensional problems. with As compared to the alternating block coordinate descent algorithm proposed by that was Kalaitzis et al. (2013) for ~~the~~ BiGLasso, the proposed ~~ISTA~~ algorithm enjoys a iterative soft thresholding (x2) per-iteration computational speedup over BiGLasso of order $\Theta(p)$. Our numerical results show that the BiGLasso algorithm often requires many more iterations to converge than our ~~ISTA~~ method. does 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 spatio-temporal meteorological data collected by the National Center for Environmental Prediction ~~(NCEP)~~. Our results show that the TeraLasso precision matrix estimator degrades much that are more slowly than other estimators as ~~one~~ we reduces the number of samples 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 modeling 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 criteria is the Frobenius norm of the residuals. Many such methods involve low-rank factor decompositions including: ~~PARAFAC~~ parallel factor analysis and CANDECOMP as in Harshman and Lundy (1994), Faber et al. (2003); ?Au: CANDECOMP 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 using by 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 to 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 ~~one~~ 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 $K \geq 2$ 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).

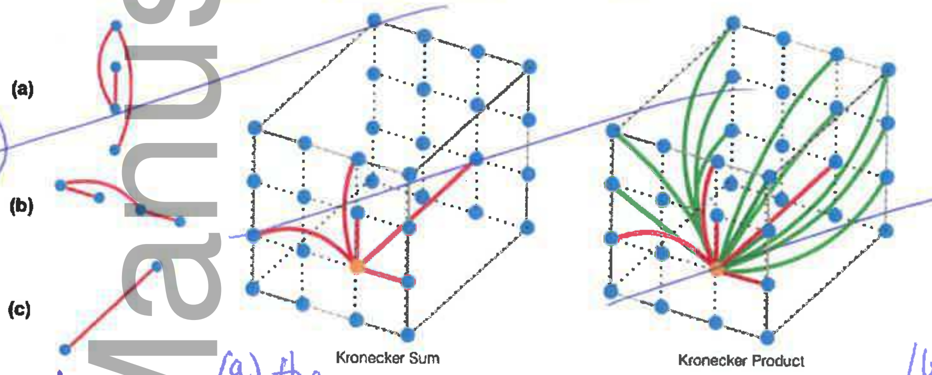


Fig. 2 Comparison of (a) the Kronecker sum (Cartesian product graph) ~~at center~~ and (b) Kronecker product (direct product graph) ~~at right~~. The products are formed from the component graphs (b), (c) and (d); ~~shown in (a)-(b)-(c);~~ the number of factors in the product graphs is $K = 3$ and the dimensions are $d_1 = d_2 = 4$ and $d_3 = 2$, leading to product graphs with 32 nodes, arranged in a regular three-~~3~~ 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 ~~while~~ the Kronecker product model is much less sparse, having a total of 184 edges.

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 ~~while~~ the Kronecker sum model corresponds light face

set non-justified

	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 favored.	Maximum likelihood estimator is convex.

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

to the Cartesian product* of these components (Hammack et al., 2011). The direct product graph and Cartesian product graph differ greatly; the former has a number of edges equal to $\frac{1}{2}\prod_{k=1}^{K}(2|E_k|+|V_k|)-\prod_{k=1}^{K}|V_k|$, ~~while~~ the latter has a number of edges equal to $\sum_{k=1}^{K}|E_k|\prod_{i\neq k}|V_i|$, where V_i, E_i denote the node and edge sets of the i-th component graph†. To illustrate, if the number of non-zero entries of Ψ_k is cd_i for some c, the number of edges 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 $p=\prod_{k=1}^{K}d_i$ is the number of covariates (rows of Ω). ~~On the other hand,~~ 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 ~~while~~ the number of new edges in the Kronecker sum model is fixed, independent of K. A concrete example of these differences is illustrated in Figure 2. The qualitative differences between the Kronecker product and Kronecker sum models for the precision matrix are summarized in Table 1.

display — whereas — and — that we — Gk cap pi — In contrast, — whereas — Fig.2 — Table 1

1.3. Rationale for the proposed multiway Kronecker sum model

This paper develops a scalable, fast and accurate estimation procedure, ~~the~~ TeraLasso, for multiway precision matrices Ω using higher order Kronecker sum models. To justify the practical utility of ~~the~~ TeraLasso we illustrate it on a spatio-temporal meteorological dataset. We have also applied it to other applications not presented here. ~~While~~ comprehensive validation of the model

by — which are — Although a

*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 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.

partial differential equations

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 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 $(A_1 \oplus \cdots \oplus A_K)\mathrm{vec}(X) = \mathrm{vec}(N)$), but this is often impractical in high dimension. Such equations result from the discretization of separable K-dimensional ~~PDEs~~ with tensorized finite elements (Grasedyck, 2004; Kressner and Tobler, 2010; Beckermann et al., 2013; Shi et al., 2013; Ellner et al., 1986). As a result Kronecker sums come in many areas of applied math, including beam propagation physics (Andrianov (1997)); control theory (Luenberger, 1966; Chapman et al., 2014); fluid dynamics (Dorr, 1970); and spatio-temporal neural processes (Schmitt et al., 2001). cematics

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 (2017)). More recently, Fey et al. (2018) ~~has~~ have proposed tensor B-splines defined over a Cartesian product basis for geometric Convolutional Neural Networks ~~(CNN)~~. Kronecker sums have been proposed as precision matrices for weighting the quadratic regularizer in smoothed multivariate spline regression. In particular, Wood (2006) observed that, ~~as~~ compared ~~to~~ 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 spatio-temporal forest health modeling (for which $K = 3$) (Augustin et al., 2009), brain development modeling (Holland et al., 2014), and analysis of the ~~impact~~ effect of climate and weather on spatio-temporal 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 $p(\beta) \propto \exp\{-\beta^T(\lambda_1 S_1 \oplus \cdots \oplus \lambda_K S_K)\beta/2\}$. Here, λ_i are regularization coefficients and S_i are coordinate-wise 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, the TeraLasso smooths each axis separately, while summing over the others, thereby reducing coupling between the tensor axes as compared to the Kronecker product. For data that has 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 the TeraLasso to estimating the best precision matrix of the spline coefficients that would minimize the population-averaged fitting error.

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 $\ell 1$ regularizers and then with non-convex sparsity regularizers. The first order ISTA 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 supplementary material, along with additional experiments and further exploration of the properties and implications of the Kronecker sum subspace $\mathcal{K}_{\mathrm{p}}$ and the associated identifiable parameterization.

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

2. Notation and Preliminaries

We use upper case letters, e.g. A for matrices and tensors, bold lower case $\mathbf{a}$ for vectors, and 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 $A_{i_1, i_2, \ldots, i_K}$. Fibers are the higher-order analogue of matrix rows and columns. A fiber of a tensor is obtained by fixing every index but one, the mode-k fiber of tensor X is denoted as the column vector $X_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_K}$. Following 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 fibers 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 $y = (y_1, \ldots, y_p)$ in $\mathbb{R}^p$, denote by $\|y\|_2 = \sqrt{\sum_j y_j^2}$ the Euclidean norm of y. The operator and Frobenius norms of a matrix A are denoted as $\|A\|_2$ and $\|A\|_F$ respectively; the notation vec(A) denotes the vectorization of the matrix A; $\|A\|_\infty$ denotes the matrix infinity norm and $\|A\|_{\max} = \max_{ij} |A_{ij}|$ denotes the max norm. The determinant is denoted as $|A|$. We use the inner product $\langle A, B \rangle = \mathrm{tr}(A^T B)$ throughout. Define the set of $p \times p$ matrices with Kronecker sum structure of fixed dimensions $d_1, \ldots, d_K$:

$$\mathcal{K}_{\mathrm{p}} = \{A \in \mathbb{R}^{p \times p} : \exists\, B_k \in \mathbb{R}^{d_k \times d_k} \text{ s.t. } A = B_1 \oplus \cdots \oplus B_K\} \tag{5}$$

where the set of matrices defined in (4) is obtained by restricting $\mathcal{K}_{\mathrm{p}}$ to the

https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets positive cone, i.e.,

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

~~Note that~~ the set $\mathcal{K}_{\mathbf{p}}$ (5) is linearly spanned by the K components, since there are no nonlinear 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 $\mathcal{K}_{\mathbf{p}}$:

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

rom

Gk cap sigma pi

on-line

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

(B) 2.1. **Parameterization of $\mathcal{K}_{\mathbf{p}}$ by Ψ_k.** Note that $\Omega = \Psi_1 \oplus \dots \oplus \Psi_K$ 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

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

whereas

and correspondingly $\Omega_0 = \mathrm{diag}(\Omega_0) + \mathrm{offd}(\Psi_{0,1}) \oplus \dots \oplus \mathrm{offd}(\Psi_{0,K})$. Note that ~~while~~ the offdiagonal factors can take on any values, $\mathrm{diag}(\Omega_0)$ is not completely free (for a fully orthogonal parameterization see Section 4 of the supplement).

on-line

(B) 2.2. **Interpretation of correlation coefficients.** The quantities $\frac{[\Psi_k]_{ij}}{\sqrt{[\Psi_k]_{ii}[\Psi_k]_{jj}}}$ do not by themselves correspond to correlation coefficients. ~~Due to~~ 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 uniquely defined for that subblock:

Because of

that is

$$\rho_{k,ij,\ell} = \frac{[\Psi_k]_{ij}}{\sqrt{([\Psi_k]_{ii} + c_\ell/d_k)([\Psi_k]_{jj} + c_\ell/d_k)}}$$

where $c_\ell = \mathrm{tr}(\ell\text{th subblock of } \Omega) - \mathrm{tr}(\Psi_k)$. 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.

(A)

3. Models and Methods

Let $X_1, \dots, X_n$ be n independent realizations of the K-way tensor X. Define $\mathbf{x}_i = \mathrm{vec}\left(X_i^T\right)$ for all $i = 1, \dots, n$. Define $\widehat{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T$ as the sample

greater than

[Recall that the Ψ_k need not be positive definite and c_ℓ need not be > 0.]

covariance. The mode-k Gram matrix S_k and factor-wise covariance $\Sigma^{(k)} = \mathbb{E}[S_k]$ are given by

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

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, while $\Sigma^{(k)}$ 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

whereas

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

In tensor covariance modeling 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

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

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

Gk cap pi

Ignoring regularization, the objective function in ~~curly~~ brackets can be written as $-\log p(\widehat{S}|\Omega)$ where $p(\widehat{S}|\Omega) = \alpha_\Omega \prod_{k=1}^{K} p(S_k|\Psi_k)$ and $p(S_k|\Psi_k) = \exp\left(-\langle m_k S_k, \Psi_k\rangle\right)$, with α_Ω a normalizing constant. The non-negativity of the Kullback-Liebler divergence display $\int p(S|\Omega)\log\left\{\frac{p(S|\Omega)}{\alpha_\Omega\prod_{k=1}^{K}p(S_k|\Psi_k)}\right\}dS$ 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 ha~~ve~~s both Kronecker sum structure and sparse structure through the penalty g_ρ, TeraLasso ~~is able to~~ can effectively regularize the precision estimation problem. that

We assume the penalty g_ρ is (μ,γ)-amenable in the sense of Loh et al. (2017).

1/2 line #

DEFINITION 1 ((μ,γ) AMENABLE REGULARIZER). *A regularizer $g_\rho(t)$ is (μ,γ)-amenable when $\mu \geq 0$ 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 nondecreasing on* $\mathbb{R}^+$,

(c) $g_\rho(t)$ *is differentiable for all* $t \neq 0$,

(d) *The function* $g_\rho(t) + \frac{\mu}{2}t^2$ *is convex*,

(e) $\lim_{t\to 0^+} g'_\rho(t) = \rho$, and

(f) $g'_\rho(t) = 0$ *for all* $t \geq \gamma\rho$.

rom unless marked otherwise

Note that the $\ell 1$ regularizer is $(0, \infty)$-amenable. Example nonconvex penalties in this class include the SCAD penalty (Fan and Li, 2001) and the MCP penalty (Zhang et al., 2010), both defined in Appendix C of the supplement. on-line

Observe that for nonzero μ (i.e. nonconvex g_ρ) the constraint on the spectral norm of Ω ($\|\Omega\|_2 \leq \kappa$) in the TeraLasso objective function (8) is necessary since without it a global minimum may not exist (Loh et al., 2017). For spectral norm a constraint parameter set to $\kappa = \sqrt{2/\mu}$, 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 $\ell 1$ penalty, the objective is always convex and κ can be set to infinity. ∞.

4. High Dimensional Consistency of the TeraLasso

Let $\mathbf{v} = [v_1, \ldots, v_p]^T$ be an isotropic ψ_2-subgaussian random vector with independent entries v_j satisfying $\mathbb{E}v_j = 0$, $1 = \mathbb{E}v_j^2 \leq \|v_j\|_{\psi_2} \leq K$. The ψ_2 condition on a scalar random variable V is equivalent to subgaussian decay of the tails of V, implying that $\mathbb{P}(|V| > t) \leq 2\exp(-t^2/c^2)$ for all $t > 0$. The extension to random vectors is straightforward. Specifically, $\mathbf{x}$ is a subgaussian 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 $X \in \mathbb{R}^{d_1\times d_2\times\ldots\times d_K}$ to be an order-K subgaussian random tensor with covariance Σ when that $\mathbf{x} = \mathrm{vec}(X^T)$ is a subgaussian random vector in $\mathbb{R}^p$ defined as in equation (9).

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

In addition to the subgaussian 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 zero and infinity. ∞.

½ line #

Assumption 1.

(A1) 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. $\text{card}(\mathcal{S}_k) \leq s_k$.

½ line #   Gk cap sigma

Assumption 2.

(A2) The minimal eigenvalue satisfies $\phi_{\min}(\Omega) = \sum_{k=1}^K \phi_{\min}(\Psi_k) \geq \underline{k}_\Omega > 0$, and the maximum eigenvalue satisfies $\phi_{\max}(\Omega) = \sum_{k=1}^K \phi_{\max}(\Psi_k) \leq \bar{k}_\Omega < \infty$.

Defining the support set of Ω as $\mathcal{S} = \{(i,j) : i \neq j, \}$, assumption (A1) implies that $\text{card}(\mathcal{S}) \leq s = \sum_{k=1}^K m_k s_k$.

4.1. *Regularization with ℓ_1 penalty*

B

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

normal #s

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

Gk cap sigma

where $|\Psi_k|_{1,\text{off}} = \sum_{i \neq j} |[\Psi_k]_{ij}|$ 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 supplement). We require an additional assumption on-line

Assumption 3.   ½ line #

(A3) The sample size n and the component dimensions d_k satisfy the following condition:

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

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.

½ line #

Note 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 supplement) on-line and do not depend on K, n, s, or $\mathbf{p}$.

display

THEOREM 1 (FROBENIUS ERROR BOUND). *Suppose that the assumptions (A1)-(A3) hold, and that $\widehat{\Omega}$ is the minimizer of expression (10) with $\rho_k \asymp \frac{1}{\underline{k}_\Omega}\sqrt{\frac{\log p}{n m_k}}$. Then with probability at least* $1 - 2(K+1)\exp\{-c \log p\}$ rom

$$\|\widehat{\Omega} - \Omega_0\|_F \leq \frac{2C_1 \|\Sigma_0\|_2}{\phi^2_{\min}(\Sigma_0)} \sqrt{\{(K+1)(s+p) \frac{\log p}{n \min_k m_k}\}}$$

THEOREM 2 (FACTORWISE AND L2 ERROR BOUNDS). *Suppose* [that] *the conditions of Theorem 1 hold. Then with probability at least* $1-2(K+1)\exp(-c\log p)$:

display

$$\frac{\|\mathrm{diag}(\widehat{\Omega}) - \mathrm{diag}(\Omega_0)\|_2^2}{(K+1)\max_k d_k} + \sum_{k=1}^{K} \frac{\|\mathrm{offd}(\widehat{\Psi}_k - \Psi_{0,k})\|_F^2}{d_k} \leq C_2(K+1)\left(1+\sum_{k=1}^{K}\frac{s_k}{d_k}\right)\frac{\log p}{n\min_k m_k} \qquad (12)$$

and as a result

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

? Au: which result

Ck cap pi

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

4.1.1. **Comparison ~~to~~ [with] [graphical] GLasso.** The Frobenius norm bound in Theorem 1 improves on the subgaussian [graphical] GLasso 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] $\|\Delta_k\|_F = O_p\left(\sqrt{\frac{(d_k+s_k)\log p}{m_k n}}\right)$, indicating that TeraLasso with n replicates estimates the identifiable representation of Ψ_k with an error rate equivalent to that of [the graphical] GLasso with $\Omega = \Psi_k$ and nm_k available replicates.

4.1.2. **Independence along an axis.** Suppose that the data tensor X is ~~i.i.d.~~ [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

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

~~as compared to~~ [with] the factorwise error bound of

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

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 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. Nonconvex Regularizers and Single Sample Support Recovery

Nonconvex regularization will provide nonasymptotic guarantees on the elementwise estimation error, implying strong, single sample support recovery guarantees when the smallest nonzero element of Ω_0 is bounded from below. ~~On the other hand,~~ these stronger results require more restrictive assumptions on sparsity of the precision matrix and its smallest nonzero element. Specifically, we ~~will~~ require the following:

(A4) The degree (maximum number of nonzero edges connected to a node) of the sparsity graph of each factor Ψ_k is bounded by a constant d.

(A5) The sample size satisfies $n \min_k m_k \geq c_0 d^2 \log p$ for some c_0 large ~~enough~~.

(A6) There exist constants c_∞, c_3 such that $\|(\Omega_0 \otimes \Omega_0)_{SS}\|_\infty \leq c_\infty$ and

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

In (A6) the notation A_{SS} denotes the submatrix of A formed by extracting the rows and columns corresponding to the index set S. Under these assumptions we have the following result.

THEOREM 3 (NONCONVEX REGULARIZERS). *Suppose the regularizer g_ρ in* (8) *is (μ, γ)-amenable, and $\kappa = \sqrt{2/\mu}$. Then with probability at least $1 - 2(K+1)\exp(-c\log p)$ as in Theorem 1,* (8) *has a unique stationary point $\widehat{\Omega}$ (given by the oracle estimator defined in the supplement), with (for all k)*

$$\|\mathrm{offd}(\widehat{\Psi}_k - \Psi_{0,k})\|_{\max} \leq \|\widehat{\Omega} - \Omega_0\|_{\max} \leq c_3(K+1)\sqrt{\frac{\log p}{n \min_k m_k}},$$

$$\|\mathrm{offd}(\widehat{\Psi}_k - \Psi_{0,k})\|_F \leq c_3(K+1)\sqrt{\frac{s_k \log p}{n \min_k m_k}},$$

$$\|\widehat{\Omega} - \Omega_0\|_F \leq c_3(K+1)\sqrt{\frac{(s+p)\log p}{n \min_k m_k}},$$

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

The proof is given in Section 7 in the supplement, and uses arguments analogous to those of Loh et al. (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 using a that sufficiently

single sample ($n = 1$) provided $\min_k m_k$ is large ~~enough~~. The Frobenius norm convergence rates (both factorwise and overall) for the convex and nonconvex regularizers remain effectively the same (comparing Theorem 3 to Theorems 1 and 2), hence the primary benefit of the nonconvex bound is the ability to guarantee support recovery in exchange for additional assumptions. with

;

Tensor graphical iterative soft thresholding

A

5. ~~TG-ISTA~~ Algorithm

In this section, we introduce an iterative soft thresholding (ISTA) 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 implementation Tensor Graphical Iterative Soft Thresholding (TG-ISTA).

B

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 semi-continuous but nonsmooth function g: for $\Omega \in \mathcal{K}_{\mathbf{p}}$,

Gk cap sigma

$$Q(\Psi_1, \dots, \Psi_K) = f(\Omega) + g(\Omega), \text{ where for } \langle \widehat{S}, \Omega \rangle = \sum_{k=1}^{K} m_k \langle S_k, \Psi_k \rangle,$$

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

centre

For objectives of this form, Nesterov (2007) 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 (ISTA) (see for example Tseng (2010); Combettes and Wajs (2005); Beck and Teboulle (2009); Nesterov (1983, 2004)). An extension to nonconvex regularizers g is given in Loh and Wainwright (2013). and by was

The linearity of the constraint set $\mathcal{K}_{\mathbf{p}}$ suggests the use of gradient descent where the gradients are projected onto the associated $(1 - K + \sum_{k=1}^{K} d_k^2)$ dimensional *linear subspace*. The positive definite restriction can then be handled in a similar way as Guillot et al. (2012) did for the vanilla GLasso. 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

Gk cap sigma to how ?Au: vanilla graphical

small fraction

$$\Omega_{t+1} \in \arg \min_{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta_t \text{Proj}_{\mathcal{K}_{\mathbf{p}}} \{ \nabla f(\Omega_t) \} \right) \right\|_F^2 + \zeta_t g(\Omega) \right\}, \quad (14)$$

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 stepsize ζ_t (see Algorithm 1) (Guillot et al., 2012). We decompose and solve the problem (14) for the case of the TeraLasso objective in Section 5.2 below.

in Table 2
in Section 5.2

B

5.2. TG-ISTA implementation of TeraLasso

expression

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

on-line

smaller tilde ~ throughout

6# 3pi car

$$\nabla_{\Omega \in \mathcal{K}_{\mathbf{p}}}(\langle \widehat{S}, \Psi_1 \oplus \cdots \oplus \Psi_k \rangle) = \mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\widehat{S}) = \widetilde{S}_1 \oplus \cdots \oplus \widetilde{S}_K = \widetilde{S} \quad \text{~~where~~}$$

$$\widetilde{S}_k = S_k - \frac{K-1}{K}\frac{\mathrm{tr}(S_k)}{d_k} I_{d_k}. \tag{15}$$

Although

by

~~While~~ many different conventions for parameterizing the projection using the $\widetilde{S}_k$ 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

smaller

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

that are

on-line

expressions

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 (15) and (16), the projected gradient of the objective $f(\Omega_t)$ is

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

rom

LEMMA 4 (DECOMPOSITION OF OBJECTIVE). *For $\Omega_t, \Omega \in \mathcal{K}_{\mathbf{p}}$ of the form*

display

$$\Omega_t = \Psi_1^t \oplus \cdots \oplus \Psi_K^t \quad \text{and} \quad \Omega = \Psi_1 \oplus \cdots \oplus \Psi_K,$$

display

problem rom

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

small fraction

rom

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

the on-line

equation

The proof is in supplement Section 2.5. The right hand side of (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),

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

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

6#

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

Table 2

on-line

in Table 2.

The composite gradient descent algorithm is given in Algorithm 1. In Section 8 of the 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 $O\left(pK + \sum_{k=1}^{K} d_k^3\right)$ floating point operations.

Gk cap sigma

Table 2.

light face

?AU: SCM

Algorithm 1 TG-ISTA implementation of TeraLasso (high level)

1: Input: SCM factors S_k, regularization parameters ρ_i, backtracking constant $c \in (0, 1)$, initial step size $\zeta_{1,0}$, initial iterate $\Omega_{\text{init}} = I \in \mathcal{K}_{\mathbf{p}}^{\sharp}$.

2: **while** not converged **do**

3: Compute the subspace gradient $\text{proj}_{\mathcal{K}_{\mathbf{p}}}(\Omega_t^{-1}) = G_1^t \oplus \cdots \oplus G_K^t$.

4: *Line search:* Let stepsize ζ_t be the largest element of $\{c^j \zeta_{t,0}\}_{j=1,\ldots}$ such that the following are satisfied for $\Psi_k^{t+1} = \text{shrink}^-_{\zeta_t \rho_k}\{\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t)\}$

set non-justified

vert

display

$$\Psi_1^{t+1} \oplus \cdots \oplus \Psi_K^{t+1} \succ 0 \quad \text{and} \quad f(\{\Psi_k^{t+1}\}) \leq Q_{\zeta_t}(\{\Psi_k^{t+1}\}, \{\Psi_k^t\})$$

vert

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

6: *Composite objective gradient update:*

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

7: **end for**

expression on-line section

8: Compute Barzilai-Borwein stepsize $\zeta_{t+1,0}$ via (27) in supplement 2.2.

9: **end while**

10: Return $\{\Psi_k^{t+1}\}_{k=1}^K$

B

5.3. TG-ISTA for a nonconvex 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

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

where ζ is the step size and

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

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

$$\Psi_k^{t+1} = \text{shrink}^-_{\zeta\rho}\left[\Psi_k^t - \zeta\{\widetilde{S}_k - G_k^t + q'_\rho(\Psi_k)\}\right]$$

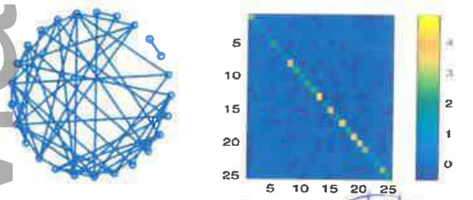
where $q'_\rho(t) = \frac{d}{dt}\{q_\rho(t) - \rho|t|\}$ for $t \neq 0$ and $q'_\rho(0) = 0$. These updates can be inserted into the framework of Algorithm 1, with an added step of enforcing the $\|\Omega\|_2 \leq \kappa$ constraint, e.g. via step size line search. The algorithm is summarized in Algorithm 2 in Supplement 2.1.

THEOREM 5 (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 $\sqrt{2/\mu}$.*

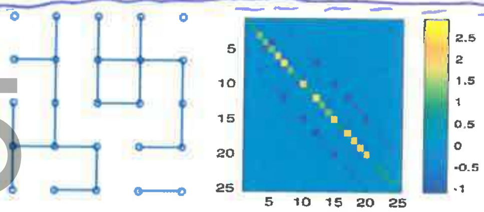
PROOF. Follows since for $\kappa \leq \sqrt{2/\mu}$ the objective (8) is convex on the convex constraint set $\{\Omega \in \mathcal{K}_{\mathbf{p}} | \Omega \succ 0, \|\Omega\|_2 \leq \kappa\}$ (Lemma 21, supplement).

6. Validation on synthetic data

Random graphs were created for each factor Ψ_k using both an Erdos-Renyi (ER) topology and a random grid graph topology¶. 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 $[\Psi_k]_{ii}$ $[\Psi_k]_{jj}$ by a. This keeps Ψ_k positive definite. We repeat this process until all edges are added. Finally, we form $\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K$. An example 25-node, $q = 25$ ER graph and precision matrix are shown in Figure 3. The random grid graph



(a) Random Erdos-Renyi (ER) graph with 25 nodes and 50 edges



(b) Random grid graph (square) with 25 nodes and 26 edges

Fig. 3. Example Erdos-Renyi and random grid graphs. Left: Graphical representation. Right: Corresponding precision matrix Ψ.

¶Code for experiments can be found at https://github.com/kgreenewald/teralasso.

only c on-line

is produced in a similar way, with the exception that edges are ~~only~~ allowed between adjacent nodes, where the nodes are arranged on a square grid (Figure 3(b)). Algorithm 1 in Section 2.3 of the supplement describes how the random vector $\mathbf{x} = \text{vec}(X^T)$ is generated under the Kronecker sum model.

B

6.1. Validation of theoretical algorithmic convergence rates

on-line Fig. 4 section 7 by the on-line

To verify the geometric convergence of the TG-ISTA implementation (Theorem 25 in the 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 $d_1 = \cdots = d_K$, and determined the value for $\rho_k = \rho$ using cross validation. Figure 4 shows the results as a function of iteration, for a variety of d_k and K configurations and the ℓ_1 convex regularization. Figure 13 in Supplement 2.1 repeats these experiments with the nonconvex 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.

B

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 $\rho_k = \bar{\rho}\sqrt{\frac{\log(p)}{nm_k}}$ 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.

by detections

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

fixed 2pt # fine #

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

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

(a) $\ell 1$ penalty,
$n = 100$ sample size

also sized separately

(b) $\ell 1$ penalty,
$n = 1$ sample size

landscape

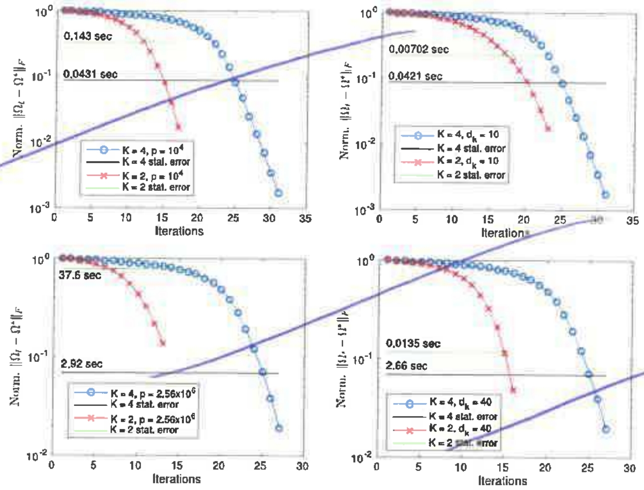


Fig. 4. Linear geometric convergence of the convex ($\ell 1$ penalized) TG-ISTA implementation of TeraLasso. Shown is the normalized Frobenius norm $\|\Omega_t - \Omega^*\|_F$ of the difference between the estimate at the tth iteration and the optimal Ω^*. On the left are results comparing $K = 2$ and $K = 4$ on the same data with the same value of p (different d_k); on the right they are compared for the same value of d_k (different p). 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) from

(c), (d) results comparing

; for (a), (c) n = 100 sample size and (b), (d) n = 1 sample sizes

Fig.5

Shown in Figure 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 giving $\rho_k = \frac{\bar{\rho}_k}{\sqrt{(\log p)/(nm_k)}}$. Note $\bar{\rho}_1 = \bar{\rho}_2 = \bar{\rho}_3$ achieves near optimal results. that

Having verified the single tuning parameter approach, hereafter we will cross-validate only $\bar{\rho}$. In 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 $\ell 1$ regularized model are tight. Figure 6 illustrates how increasing dimension p and K improves single sample performance. Shown are the average TeraLasso edge detection precision and recall values for different values of K in the single and 5-sample regimes, all increasing to 1 (perfect structure estimation) as p, K, and n increase.

shall

the on-line

Fig.6

five

various

(B)

6.3. *Nonconvex Regularization*

Fig.7

Here the $\ell 1$ penalized TeraLasso is compared to TeraLasso with nonconvex regularization (8). Shown in Figure 7 are the MCC, normalized Frobenius error and spectral norm error for estimating $K = 2$ and $K = 3$ Erdos-Renyi graphs

with

ER

also sized separately

landscape

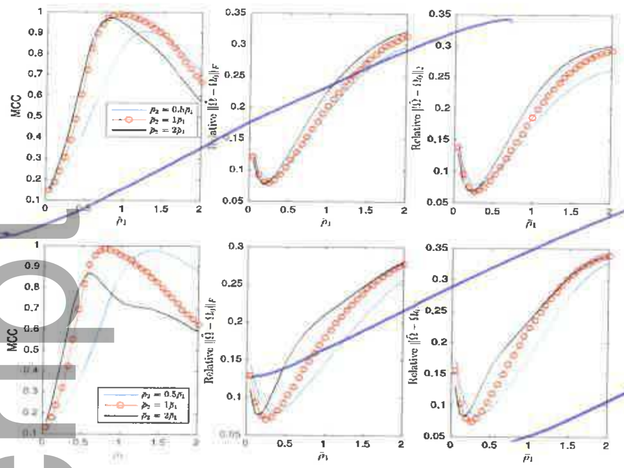


Fig. 5. Setting tuning parameters with $K = 3$, $n = 1$, and $d_1 = d_3 = 64$. Shown are the MCC, relative Frobenius error, and relative L2 error of the TeraLasso estimate as the scaled tuning parameters ρ_i are varied. Shown are deviations of $\bar{\rho}_2$ from the theoretically dictated $\bar{\rho}_2 = \bar{\rho}_1 = \bar{\rho}_3$. Top: Equal dimensions, $d_1 = d_2 = d_3$. First and third factors are random ER graphs with d_k edges, and the second factor is random grid graph with $d_k/2$ edges. Bottom: Dimensions $d_2 = 2d_1$, each factor is a random ER graph with d_k edges. Notice in these scenarios that using $\bar{\rho}_1 = \bar{\rho}_2$ is near optimal, as theoretically predicted.

(the (a)–(c) (d)–(f)

Fig. 8

as functions of regularization parameter ρ for each of $\ell 1$, SCAD (96), and MCP (97) regularizers in a variety of configurations. Figure 8 shows similar results for Ψ_k a variant of the spiked identity model of Loh et al. (2017). Observe that nonconvex 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.

National Center for Environmental Prediction

7. NCEP Windspeed Data

The TeraLasso model is illustrated on a meteorological dataset. The US National Center for Environmental Prediction (NCEP) maintains records of average daily wind velocities in the lower troposphere, with daily readings beginning in 1948. The data is available online at 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 degrees in each coordinate. Over bounded areas, the spacing is approximately a rectangular grid, suggesting a $K = 2$ model (latitude vs. longitude) for the spatial covariance, and a $K = 3$ model (latitude vs. longitude vs. time) for the full spatio-temporal covariance.

versus versus

Consider the time series of daily-average wind speeds. Following Tsiligkaridis

(a) $n = 1$

(b) $n = 5$

also sized separately

landscape

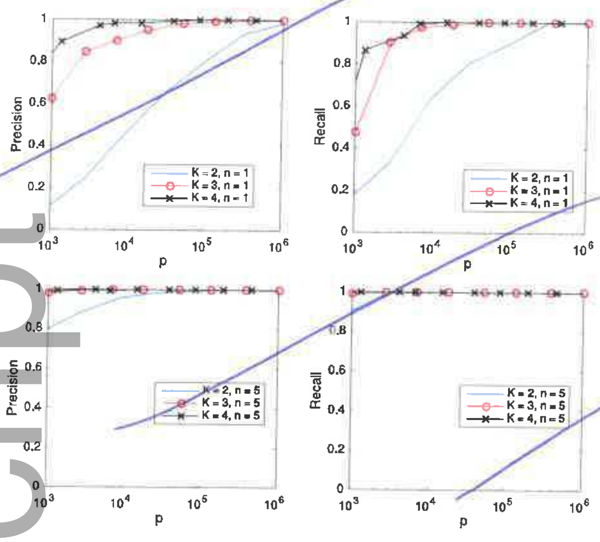


Fig. 6: Edge support estimation on random ER graphs, with the ρ_k set according to Theorem 1. Graphical model edge detection precision and recall curves are shown as a function of data dimension $p = \prod_{k=1}^{K} d_k$. For each value of the tensor order K, we set $d_k = p^{1/K}$. Observe single sample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

and Hero (2013), we regress out the mean for each day in the year via a 14-th order polynomial regression on the entire history from 1948-2015. We extract two 20×10 spatial grids, one from eastern North America, and one from western North America (Figure 9). Figure 10 shows the TeraLasso estimates for latitude and longitude factors using time samples from January in n years following 1948, for both the eastern and the western grids. Observe the approximate AR structure, and the break in correlation (Figure 10 (b), longitude factor) in the Western Longitude factor. The location of this break corresponds to the high elevation line of the Rocky Mountains. In the supplement, we compare the TeraLasso estimator to 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 the TeraLasso provides a significantly better fit to the data.

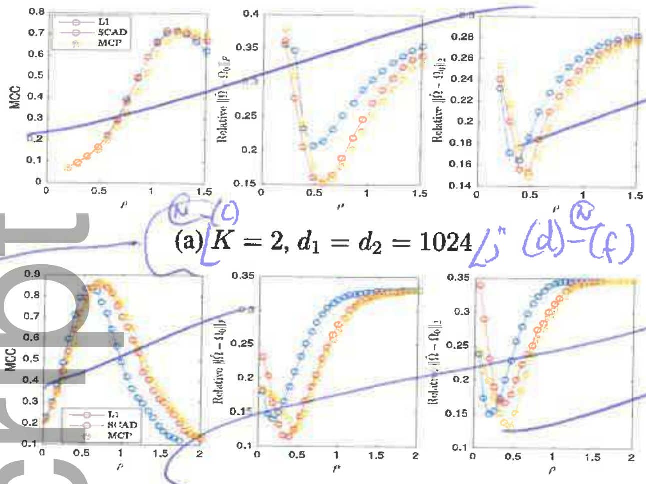
to | and 11 | figs 9, 10, 11 | by | 2(1) | on-line | with | to

To illustrate the utility of the estimated precision matrices, we use them to construct a season classifier. NCEP windspeed records are taken from the 51-year span from 1948-2009. We estimate spatial precision matrices on n consecutive days in January and June of a training year respectively, 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 1 training and 50 testing years.

to | one | run on

National Center for Environmental Prediction

also sized separately



(a) $K = 2$, $d_1 = d_2 = 1024$ (d)–(f)

landscape

(b) $K = 3$, $d_1 = d_2 = d_3 = 32$

Fig. 7: Nonconvex 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 nonconvex regularization improves performance.

that

The sequences are labeled as summer (June) and winter (January), and we compute the classification error rate for the winter vs. summer classifier obtained by choosing the season associated with the larger of the likelihood functions

that is — versus

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

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 $\ell 1$ regularization.

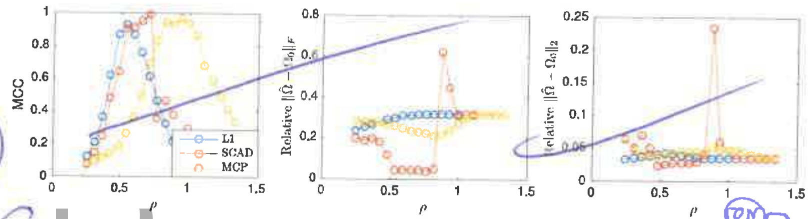
various

Results for different sized temporal covariance extents ($T = d_3$) are shown in Figure 11 for TeraLasso, with unregularized TeraLasso (ML 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 ML Kronecker product estimator instead of the 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 this data (Section 3.4 of the supplement). Note the superior perfor-

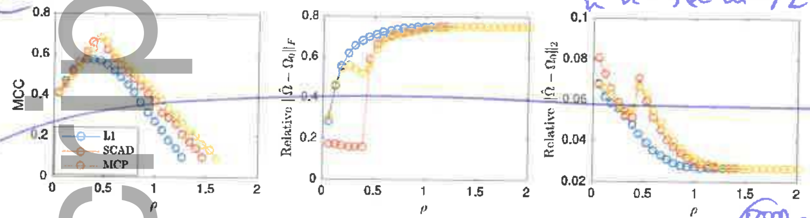
Fig. 12 — (maximum likelihood — the — these — on-line

maximum likelihood

also sized separately



(a) $K = 2$, $d_1 = d_2 = 256$, $n = 10$, $\Psi_k = 0.5 I_{d_k} + 0.5 [1_8; 0_{248}][1_8; 0_{248}]^T$; (d)–(f)



(b) $K = 3$, $d_1 = d_2 = d_3 = 32$, $n = 1$, $\Psi_k = 0.5 I_{d_k} + 0.5 [1_{16}; 0_{16}][1_{16}; 0_{16}]^T$

landscape

Fig. 8. Nonconvex regularizers with spiked identity factors Ψ_k. Shown are the MCC and relative Frobenius error as a function of ρ. Note nonconvex regularization improves performance when ρ is chosen correctly; (a)–(c)

mance and increased single sample robustness of the proposed ~~ML~~ Kronecker Sum and TeraLasso estimates ~~as~~ compared ~~to~~ the Kronecker product estimate, confirming the better fit of TeraLasso. In each case, the nonmonotonic behavior of the Kronecker product curves is due partly to randomness associated with the small test sample size, and partly ~~due to the fact that~~ the Kronecker product in $K = 3$ has overly strong coupling across tensor directions, giving large bias.

with maximum likelihood

that is

because

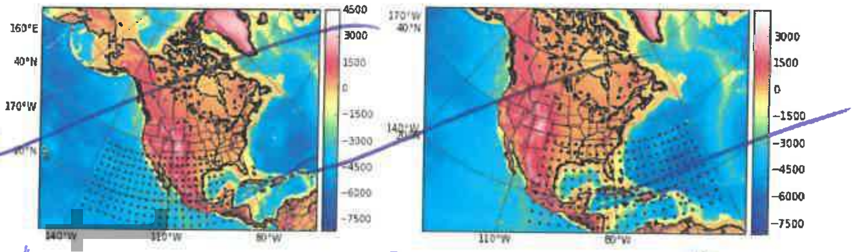
8. Conclusion

A factorized model, called ~~the~~ 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 ISTA-like optimization algorithm is presented that scales to high dimensions. Statistical and algorithmic convergence are established for ~~the~~ 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 enable improved single-sample performance for estimation and anomaly detection. Future work includes combining first moment tensor representation methods for mean estimation such as PARAFAC (Harshman and Lundy, 1994) with the second order TeraLasso method introduced in this paper for estimating the covariance.

iterative soft thresholding

parallel factor analysis that is

also sized separately



landscape

Fig. 9: Rectangular 10 × 20 latitude-longitude grids of wind speed locations ~~shown as black dots.~~ Elevation colormap shown in meters. Left: Western grid, Right: Eastern grid.

9. Acknowledgement

National Science Foundation

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, NSF 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. *The Annals of Applied Statistics*, **4**, 764–790.

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

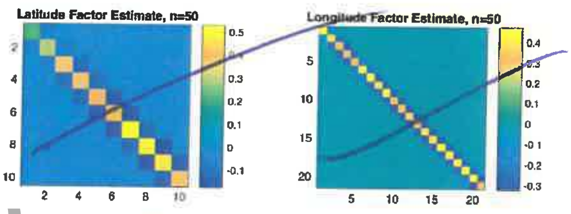
Augustin, N. H., Musio, M., von Wilpert, K., Kublin, E., Wood, S. N. and Schumacher, M. (2009) Modeling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association*, **104**, 899–911.

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

Beck, A. and Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **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 Journal on Numerical Analysis*, **51**, 3307–3326.

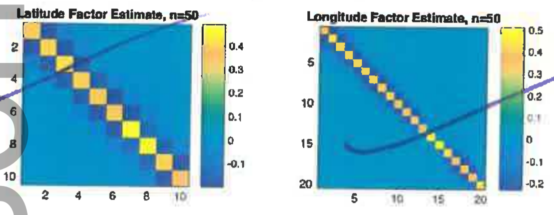
Fig. 10. TeraLasso estimate factors, K=2



landscape

also sized separately

(a) Eastern grid. Graphical representation of latitude (left, 10 nodes) and longitude factors (bottom, 20 nodes) with the corresponding precision estimates. Note the simple AR-1 type structure of the longitude graph.



landscape

(b) Western grid. Graphical representation of latitude (left) and longitude factors (bottom) with the corresponding precision estimates. Observe the decorrelation (longitude factor entries connecting nodes 1-13 to nodes 14-20 are essentially zero) in the Western longitudinal factor, corresponding to the high-elevation line of the Rocky Mountains.

zero

Fig. 10. TeraLasso estimate factors, $K = 2$

Boyd, S. and Vandenberghe, L. (2009) *Convex optimization*. Cambridge university press.

Chapman, A., Nabi-Abdolyousefi, M. and Mesbahi, M. (2014) Controllability and observability of network-of-networks via cartesian products. *IEEE Transactions on Automatic Control*, **59**, 2668–2679.

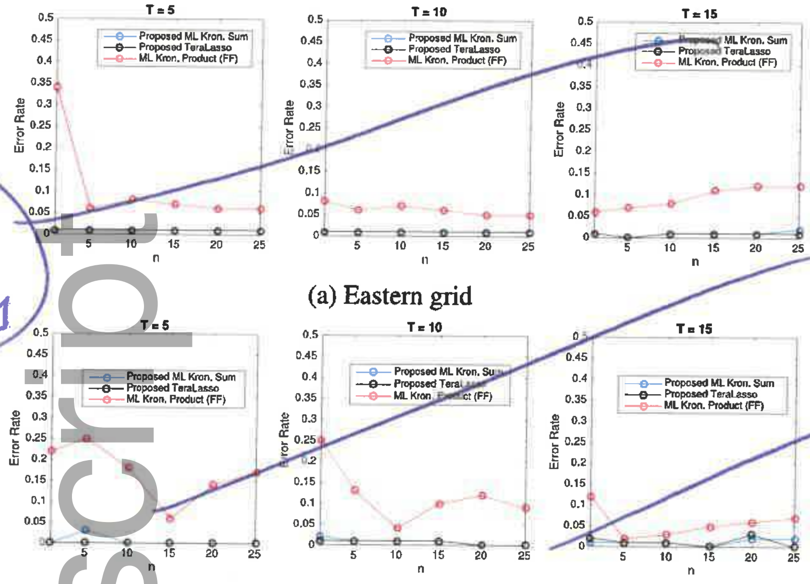
Combettes, P. L. and Wajs, V. R. (2005) Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, **4**, 1168–1200.

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 review*, **12**, 248–263.

Eilers, P. H. and Marx, B. D. (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, **66**, 159–174.

New York:



(a) Eastern grid

(b) Western grid

also sized separately

landscape 12.

(a)–(c) eastern grid; (d)–(f) western grid; (a), (d) $T=5$; (b), (e) $T=10$; (c), (f) $T=15$

Fig. ~~11~~: Classification using Gaussian loglikelihood and estimated spatio-temporal ($K = 3$) precision matrices for each season, where T is the temporal dimension in days. Shown is windspeed summer ~~vs.~~ 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.

?Au: all auths eds

Ellner, N. S. et al. (1986) New ADI model problem applications. In *Proceedings ~~of 1986 ACM~~ Fall joint computer conference*, 528–534. Association for Computing Machinery

Faber, N. K. M., Bro, R. and Hopke, P. K. (2003) Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, **65**, 119–137.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348–1360.

?Au: towns

Fey, M., Eric Lenssen, J., Weichert, F. and Müller, H. (2018) Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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. on Sig. Proc.*, **63**, 6368–6378.

?Au: town

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. Curran Associates, Inc.

?Au: NIPS eds

Guillot, D., Rajaratnam, B., Rolfs, B., Maleki, A. and Wong, I. (2012) Iterative thresholding algorithm for sparse inverse covariance estimation. In *NIPS*, 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. *Computational Statistics & Data Analysis*, **18**, 39–72.

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

?Au: all aus

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

?Au: town

Imrich, W., Klavžar, S. and Rall, D. F. (2008) *Topics in graph theory: Graphs and their Cartesian product*. AK Peters/CRC Press.

?Au: all aus

Johndrow, J. E., Bhattacharya, A., Dunson, D. B. et al. (2017) Tensor decompositions and sparse log-linear models. *The Annals of Statistics*, **45**, 1–38.

?Au: eds publish

Kalaitzis, A., Lafferty, J., Lawrence, N. and Zhou, S. (2013) The bigraphical lasso. In *Proceedings of the International Conference on Machine Learning*, pp. 1229–1237.

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

?Au: book f.d.

Kotzagiannidis, M. S. and Dragotti, P. L. (2017) Splines and wavelets on circulant graphs. *Applied and Computational Harmonic Analysis*.

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.

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

Leng, C. and Tang, C. Y. (2012) Sparse matrix graphical models. *~~Journal of the American Statistical Association~~*, **107**, 1187–1200.

?Au: eds, publish

Loh, P.-L. and Wainwright, M. J. (2013) Regularized m-estimators with non-convexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484.

Loh, P.-L., Wainwright, M. J. et al. (2017) Support recovery without incoherence: A case for nonconvex regularization. *~~The~~ Annals of Statistics*, **45**, 2455–2482.

Luenberger, D. (1966) Observers for multivariable systems. *IEEE Trans~~actions on~~ Automatic Control*, **11**, 190–197.

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.

?Au: all aus

Meinshausen, N., Bühlmann, P. et al. (2006) High-dimensional graphs and variable selection with the lasso. *~~The~~ annals of statistics*, **34**, 1436–1462.

Nesterov, Y. (1983) A method of solving a convex programming problem with convergence rate o (1/k2). ~~In~~ *Soviet Mathematics Doklady*, ~~vol.~~ 27, 372–376.

Nesterov, Y. — (2004) *Introductory Lectures on Convex Optimization*. Applied Optimization, ~~vol.~~ 87, Kluwer ~~Academic Publishers~~, Boston.

Nesterov, Y. — (2007) Gradient methods for minimizing composite objective function. *CORE report*.

?Au: CORE town

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. *Digital Signal Processing*, **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.

?Au: all aus

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J. et al. (2008) Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, **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 Transactions on Medical Imaging*, **20**, 314–324.

Shi, X., Wei, Y. and Ling, S. (2013) Backward error and perturbation bounds for high order sylvester tensor equation. *Linear and Multilinear Algebra*, **61**, 1436–1446.

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

Tsiligkaridis, T. and Hero, A. (2013) Covariance estimation in high dimensions via kronecker product expansions. *IEEE Trans. on Sig. Proc.*, **61**, 5347–5360.

Tsiligkaridis, T., Hero, A. and Zhou, S. (2013) On convergence of kronecker graphical lasso algorithms. *IEEE Trans. Signal Proc.*, **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. *Journal of computational and applied mathematics*, **123**, 85–100.

Werner, K., Jansson, M. and Stoica, P. (2008) On estimation of cov. matrices with kronecker product structure. *IEEE Trans. on Sig. Proc.*, **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 Sfken, B. (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, **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. et al. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, **38**, 894–942.

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

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

Zhou, S., Rütimann, P., Xu, M. and Bühlmann, P. (2011) High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research*, **12**, 2975–3026.

Running heads verso K. Greenewald, S. Zhou and A. Hero

Tensor Graphical Lasso recto

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

Supplementary Material for Tensor Graphical Lasso (TeraLasso)