Tensor Graphical Lasso (TeraLasso)

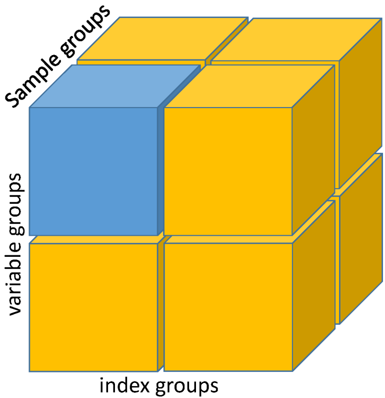
Alfred Hero

Michigan Institute for Data Science (MIDAS)
Dept of Electrical Engineering and Computer Science (EECS), Dept of Biomedical Engineering (BME), Dept of Statistics
Program in Computational Medicine and Bioinformatics
Program in Applied and Interdisciplinary Mathematics
University of Michigan - Ann Arbor

May 28, 2018

Multi-indexed data

- Sample groups
  $\Rightarrow$ Data from mixed populations[1]
- Variable groups
  $\Rightarrow$ Data from different modalities[2]
- Index groups
  $\Rightarrow$ Data indexed over diverse domains[3]



[1] G. McLachlan and D. Peel. Finite mixture models. John Wiley and Sons, 2004.

[2] Y. Yilmaz and A. Hero, Multimodal Event Detection in Twitter Hashtag Networks, Journ. of Signal Processing Systems, pp.1-16, 2016.

[3] K. Greenewald, S. Zhou, A.O. Hero, The Tensor Graphical Lasso (TeraLasso) arxiv1705.03983, 2017.

Data cube of heterogeneous data

Multi-indexed data array:

$$\mathbf{Z} = \{Z_{i_1,\ldots,i_K}\}_{i_1,\ldots,i_K}^{d_1,\ldots,d_K}, \qquad d = \prod_{i=1}^{K} d_k$$

Applications:

- predicting genome expression over time
- predicting weather from spatio-temporal climatological data
- human gait detection from low resolution video
- anomaly detection from videos of crowds
- classifying fMRI brain activation image sequences

Data cube of heterogeneous data

Multi-indexed data array:

$$\mathbf{Z} = \{Z_{i_1,\ldots,i_K}\}_{i_1,\ldots,i_K}^{d_1,\ldots,d_K}, \qquad d = \prod_{i=1}^{K} d_k$$

Applications:

- predicting genome expression over time
- predicting weather from spatio-temporal climatological data
- human gait detection from low resolution video
- anomaly detection from videos of crowds
- classifying fMRI brain activation image sequences

Common traits of these applications:

- the indices index over different types of domains
- the content of the arrays is complex and high dimensional
- the content is noisy and of high variability

Data cube of heterogeneous data

Multi-indexed data array:

$$\mathbf{Z} = \{Z_{i_1,\ldots,i_K}\}_{i_1,\ldots,i_K}^{d_1,\ldots,d_K}, \qquad d = \prod_{i=1}^{K} d_k$$

Applications:

- predicting genome expression over time
- predicting weather from spatio-temporal climatological data
- human gait detection from low resolution video
- anomaly detection from videos of crowds
- classifying fMRI brain activation image sequences

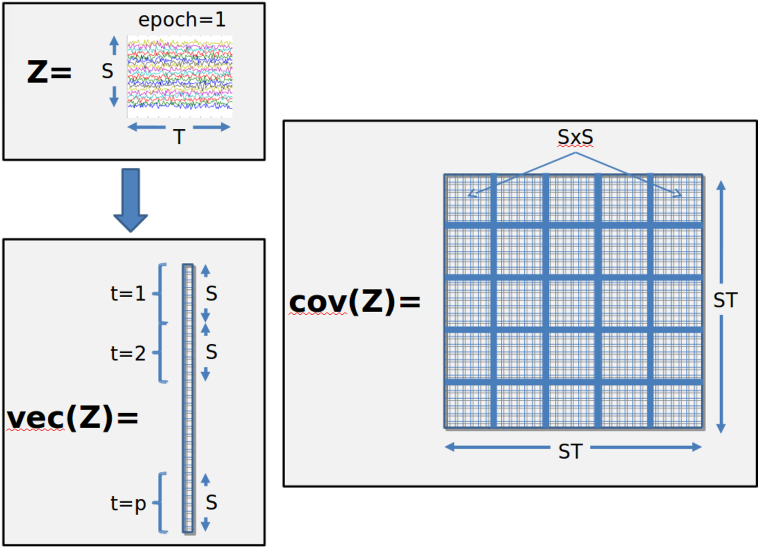
Common traits of these applications:

- the indices index over different types of domains
- the content of the arrays is complex and high dimensional
- the content is noisy and of high variability

First order models not sufficient: much of the information lies in covariance

Covariance representation $d_1 = S, d_2 = T$

$\mathbf{\Sigma}$ =cov($\mathbf{Z}$) takes partitioned form



Measurement matrix and sample covariance

	Sample 1	Sample 2	...	Sample n
Variable 1	Z_{11}	Z_{12}	...	Z_{1n}
Variable 2	Z_{21}	Z_{22}	...	Z_{2n}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
Variable d	Z_{d1}	Z_{n2}	...	Z_{dn}

$d \times n$ measurement matrix $\mathbb{Z}$ has i.i.d. columns $\mathbf{Z}_i$ with $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{Z}_i)$

$$\mathbb{Z} = \begin{bmatrix} Z_{11} & \cdots & \cdots & Z_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ Z_{d1} & \cdots & \cdots & Z_{dn} \end{bmatrix} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_n]$$

Measurement matrix and sample covariance

	Sample 1	Sample 2	...	Sample n
Variable 1	Z_{11}	Z_{12}	...	Z_{1n}
Variable 2	Z_{21}	Z_{22}	...	Z_{2n}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
Variable d	Z_{d1}	Z_{n2}	...	Z_{dn}

$d \times n$ measurement matrix $\mathbb{Z}$ has i.i.d. columns $\mathbf{Z}_i$ with $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{Z}_i)$

$$\mathbb{Z} = \begin{bmatrix} Z_{11} & \cdots & \cdots & Z_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ Z_{d1} & \cdots & \cdots & Z_{dn} \end{bmatrix} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_n]$$

$d \times d$ *sample covariance matrix*:

$$\begin{aligned} \mathbf{S}_n &= (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{Z}_i - \overline{\mathbf{Z}})(\mathbf{Z}_i - \overline{\mathbf{Z}})^T \\ &= (n-1)^{-1} \mathbb{Z}(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbb{Z}^T \end{aligned}$$

Empirical estimation of covariance under multivariate normal model

Multivariate normal model for $\mathbb{Z}$

$$\log f(\mathbb{Z}|\mathbf{\Sigma}) = \frac{n}{2}\log\det(\mathbf{\Sigma}^{-1}) - \frac{n-1}{2}\mathrm{tr}\left(\mathbf{S}_n\mathbf{\Sigma}^{-1}\right)$$

Maximum likelihood covariance estimator $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Omega}}^{-1}$ =soln to convex program[1]

$$\hat{\mathbf{\Omega}} = \mathrm{argmin}_{\mathbf{\Omega}}\mathrm{tr}(\mathbf{S}_n\mathbf{\Omega}) - c_n\log\det(\mathbf{\Omega})$$

Note that there are $p = d(d+1)/2 = O(d^2)$ unknown parameters in $\mathbf{\Sigma}$

Requirements:

1. $n \geq O(d)$ for MLE to be unique and have finite variance
2. $n \geq O(d^2)$ for estimator to be accurate (Bühlmann&VanDeGeer, 2011)

$$E[\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2] = O\left(d^2/n\right)$$

[1] $c_n = \frac{n}{n-1}$

Empirical estimation of covariance under multivariate normal model

Multivariate normal model for $\mathbb{Z}$

$$\log f(\mathbb{Z}|\boldsymbol{\Sigma}) = \frac{n}{2}\log\det(\boldsymbol{\Sigma}^{-1}) - \frac{n-1}{2}\mathrm{tr}\left(\mathbf{S}_n\boldsymbol{\Sigma}^{-1}\right)$$

Maximum likelihood covariance estimator $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Omega}}^{-1}$ =soln to convex program[1]

$$\hat{\boldsymbol{\Omega}} = \mathrm{argmin}_{\boldsymbol{\Omega}}\mathrm{tr}(\mathbf{S}_n\boldsymbol{\Omega}) - c_n\log\det(\boldsymbol{\Omega})$$

Note that there are $p = d(d+1)/2 = O(d^2)$ unknown parameters in $\boldsymbol{\Sigma}$

Requirements:

1. $n \geq O(d)$ for MLE to be unique and have finite variance
2. $n \geq O(d^2)$ for estimator to be accurate (Bühlmann&VanDeGeer, 2011)

$$E[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2] = O\left(d^2/n\right)$$

If $\boldsymbol{\Omega}$ has structure can obtain $E[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2] = O\left(d/n\right)$, e.g. for sparse $\boldsymbol{\Omega}$

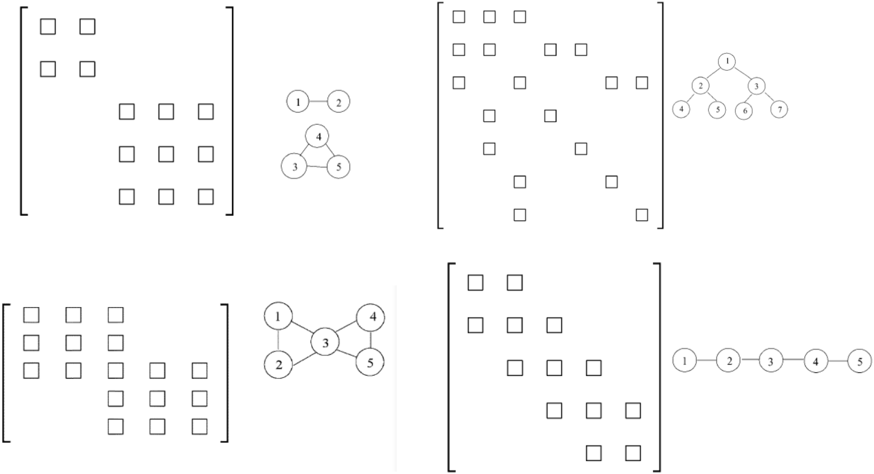
$$\hat{\boldsymbol{\Omega}} = \mathrm{argmin}_{\boldsymbol{\Omega}}\mathrm{tr}(\mathbf{S}_n\boldsymbol{\Omega}) - c_n\log\det(\boldsymbol{\Omega}) + \lambda\|\boldsymbol{\Omega}\|_1 \qquad (GLasso)$$

[1] $c_n = \frac{n}{n-1}$

Many types of covariance structure have been studied

- Sparse structured covariance
  - Banded covariance model: Bickel-Levina (2008)
  - Sparse eigendecomposition model: Johnstone-Lu (2007)
  - Stein shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
- Sparse inverse covariance: Gaussian graphical models
  - Glasso: Freidman-Hastie,-Tibshirani(2007), Wiesel-Eldar-H (2010).
  - l_1-regularized logistic: Wainright, Ravikumar, Lafferty (2006)
  - Bayesian sparse cov estimation: Rajaratnam-Massam-Carvalho (2008)
  - Sparse Kronecker GGM (Matrix Normal):Allen-Tibshirani (2010), Tsiligkaridis-Zhou-H (2012), Greenewald-H (2014)
- Covariance sparsity testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980), Eagleson (1983), Jiang (2004), Zhou (2007), Cai and Jiang (2011)
  - Independence screening (Fan 2007, Rajaratnam-H, 2011)
- Inverse covariance sparsity testing
  - Partial correlation hub screening (Rajaratnam-H 2012)
  - Sparse prediction and regression by correlation screening (SPARCS) (Firouzi-Rajaratnam-H 2017)

Graphical representations of $\mathbf{\Sigma}$ or $\mathbf{\Sigma}^{-1}$



Wiesel, Eldar and H IEEE TSP 2010

Sparsity in covariance vs sparsity in inverse covariance

Sparsity models

- Sparse correlation ($\mathbf{\Sigma}$) graphical models:
  - Most correlation are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation ($\mathbf{K} = \mathbf{\Sigma}^{-1}$) graphical models
  - Most inverse covariance entries are zero, few conditional dependencies
  - Examples: Markov random fields, autoregressive (AR) processes, global latent variables

- Sometimes correlation matrix and its inverse are both sparse
- Often only one of them is sparse

Refs: Meinshausen-Bühlmann (2006), Friedman (2007), Bannerjee (2008), Wiesel-Eldar-H (2010) .

Most physical processes are sparse in inverse covariance

Let $p^t(x, y)$ be a space-time process satisfying Poisson equation

$$\frac{\nabla^2 p^t}{\nabla x^2} + \frac{\nabla^2 p^t}{\nabla y^2} = W^t$$

where $W^t = W^t(x, y)$ is driving process.
For small Δ_x, Δ_y, p satisfies the difference equation:

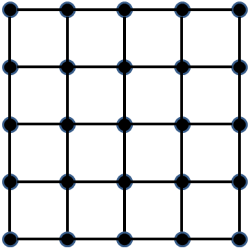
$$X^t_{i,j} = \frac{(X^t_{i+1,j} + X^t_{i-1,j})\Delta^2 y + (X^t_{i,j+1} + X^t_{i,j-1})\Delta^2 x - W^t_{i,j}\Delta^2 x \Delta^2 y}{2(\Delta^2 x + \Delta^2 y)}$$

In matrix form, as before: $[\mathbf{I} - \mathbf{A}]\mathbf{X}^t = \mathbf{W}^t$ and

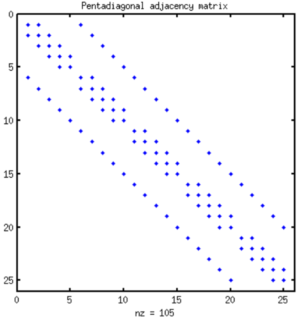
$$\mathbf{\Omega} = \mathrm{cov}^{-1}(\mathbf{X}^t) = \sigma^2_W [\mathbf{I} - \mathbf{A}][\mathbf{I} - \mathbf{A}]^T$$

$\mathbf{A}$ is sparse "pentadiagonal" matrix.

Example: 5×5 Poisson random field graphical model



Graph G_K on $\mathbb{R}^2$



corresp. **K** adjacency matrix

Gauss-Markov random field from Poisson equation on 90×90 grid

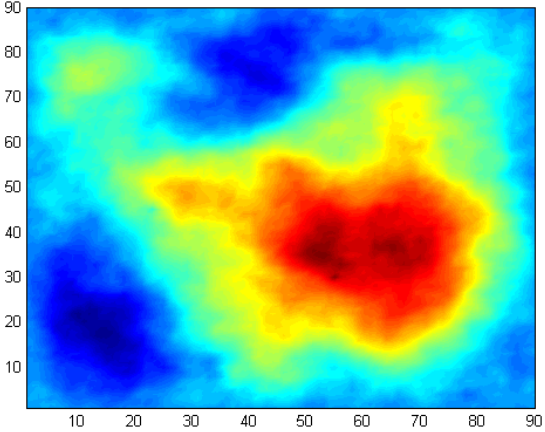
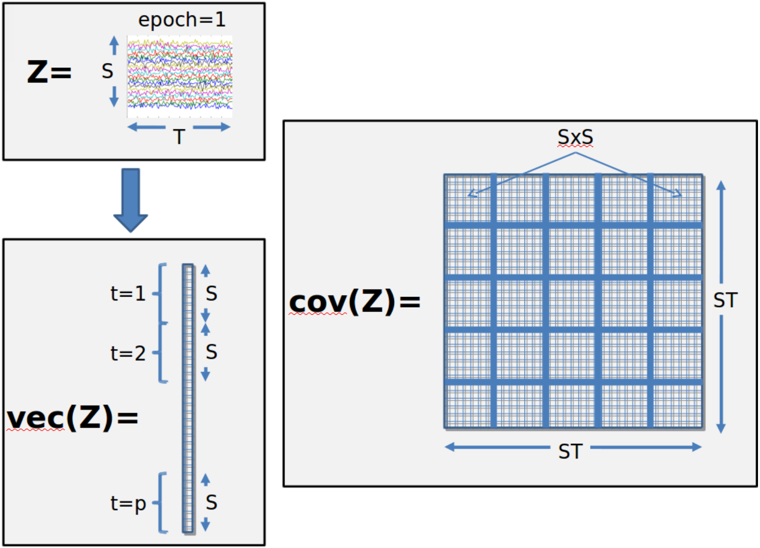


Figure: Poisson random field. $\mathbf{W}^t = \mathbf{N}_{iso} + sin(\omega_1 t)\mathbf{e}_1 + sin(\omega_2 t)\mathbf{e}_2$ ($\omega_1 = 0.025$, $\omega_2 = 0.02599$, SNR=0dB).

Kronecker product models: covariance representation

$\mathbf{\Sigma} = \text{cov}(\mathbf{Z})$ takes partitioned form



Kronecker product representation of covariance matrices

Let **A** be a $p \times p$ matrix and **B** be a $q \times q$ matrix. For $d = pq$ the Kronecker product of **A** and **B** is the $d \times d$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \cdots & a_{pp}\mathbf{B} \end{bmatrix}$$

Kronecker product representation of covariance matrices

Let $\mathbf{A}$ be a $p \times p$ matrix and $\mathbf{B}$ be a $q \times q$ matrix. For $d = pq$ the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is the $d \times d$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \cdots & a_{pp}\mathbf{B} \end{bmatrix}$$

Kronecker product properties (VanLoan-Pitsianis 1992)[a]:

- $\mathbf{A} \otimes \mathbf{B}$ is p.d. if $\mathbf{A}$ and $\mathbf{B}$ are p.d.
- $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ if $\mathbf{A}$ and $\mathbf{B}$ are invertible.
- $\det(\mathbf{A} \otimes \mathbf{B}) = \det\mathbf{A} \det\mathbf{B}$
- For any $pq \times pq$ matrix $\mathbf{D}$

$$\|\mathbf{D} - \mathbf{A} \otimes \mathbf{B}\|_F^2 = \|\mathcal{R}(\mathbf{D}) - \mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{B})^T\|_F^2$$

Van Loan and Pitsianis (1993). "Approximation with Kronecker products." *Linear algebra for large scale and real-time applications*

Kronecker product model for inverse covariance

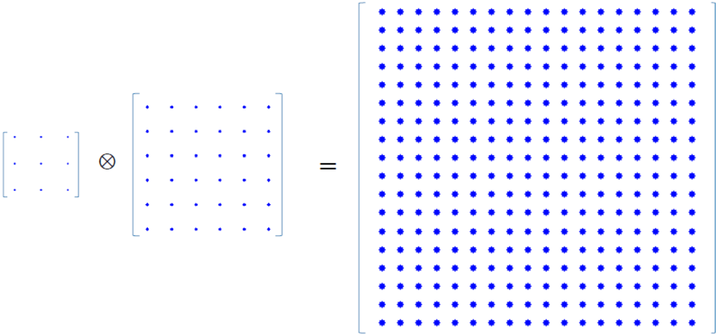


Figure: 18×18 covariance matrix has 18*17/2=153 unknown cross-correlation parameters. Kronecker product covariance model reduces this to $3 + 15 = 18$ parameters.

Leads to Kronecker MLE (matrix normal): Dawid (1981), Werner-Jansson-Stoica (2008), Tsiligkaridis-H-Zhou (2013)

Aka: matrix normal model, transposable covariance model, flip-flop covariance

Sparse Kronecker product model for inverse covariance

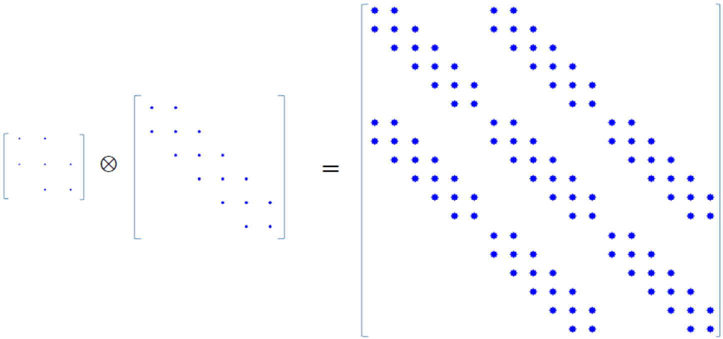


Figure: A sparse Kronecker product covariance model reduces number of parameters from 153 to 7 unknown correlation parameters.

Leads to KGlasso (sparse matrix normal): Allen-Tibshirani (2010), Yin-Li (2012), Tsiligkaridis-H-Zhou (2013)

Sparse matrix variate normal model: the KGlasso

Let $\mathbf{Z}$ ($p \times q$) follow the matrix normal normal distribution with inverse covariance matrix $\mathbf{\Omega} = \mathbf{X} \otimes \mathbf{Y}$. If

$$\|\mathbf{\Omega}\|_0 \leq O(pq)$$

then the matrix variate normal model is said to be sparse.

Maximize sparsity penalized likelihood function to estimate $\mathbf{\Omega}$

$$(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = \mathrm{argmin}_{\mathbf{X},\mathbf{Y}} J_\lambda(\mathbf{X}, \mathbf{Y})$$

where

$$J_\lambda(\mathbf{X}, \mathbf{Y}) = \mathrm{tr}\{\mathbf{S}_n(\mathbf{X} \otimes \mathbf{Y})\} - \log(|\mathbf{X}|\,|\mathbf{Y}|) + \lambda_X \|\mathbf{X}\|_1 + \lambda_Y \|\mathbf{Y}\|_1$$

Alternating minimization (flip-flop) algorithms

- Transposable regularized covariance algorithm (Allen-Tibshirani 2010)
- KGlasso algorithm (Tsiligkaridis-Zhou-H 2012)

Application: windspeed prediction (10×10 patch - Arctic to Norwegian sea)

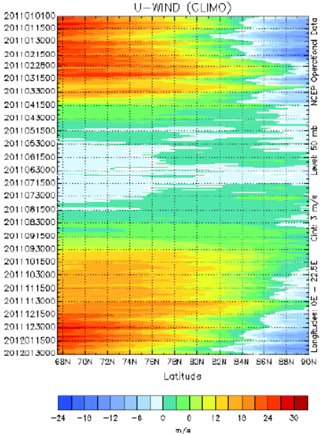
Region extending over latitudes 90-67.5 degrees N and longitudes 0-22.5 degrees E

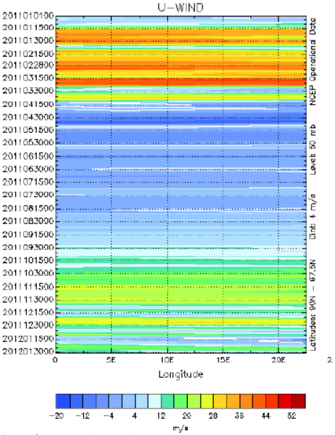
- $S = 100$ (10×10 spatial grid)
- $T = 8$ (2 day time window)
- $n = 224$ (over period 2003-2007)

Tsiligkaridis, H (2013). Covariance estimation in high D via kronecker product expansions. *IEEE Trans on Signal Processing*

Application: NCEP 10×10 patch over Arctic to Norwegian sea

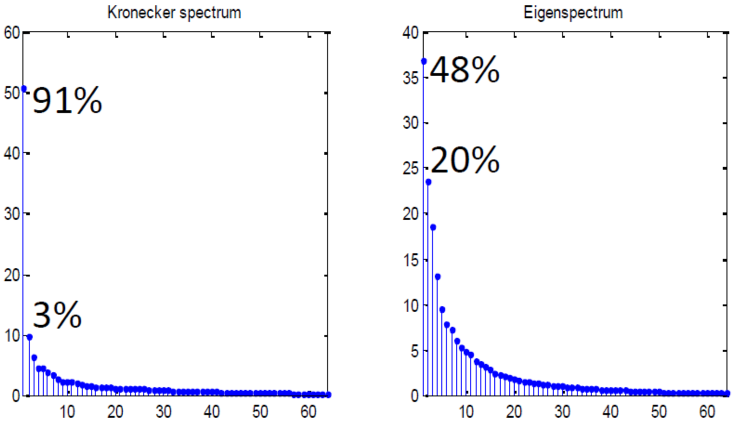
U component of windspeed





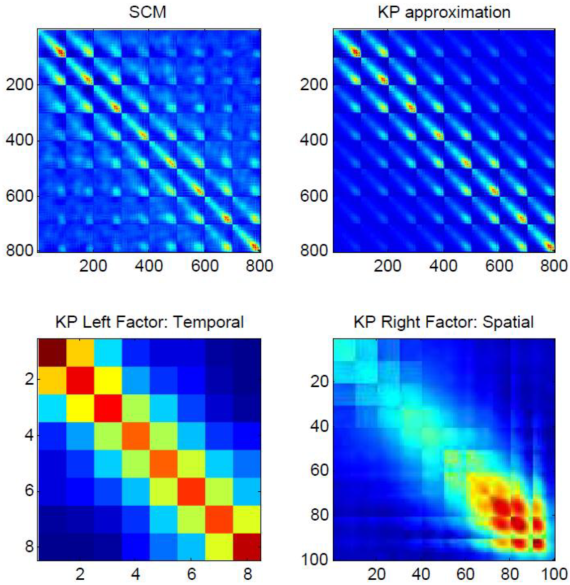
Lat (90-67.5) and long (0-22.5) is over 2.5 degree increments

Application: NCEP 10×10 over Arctic to Norwegian sea



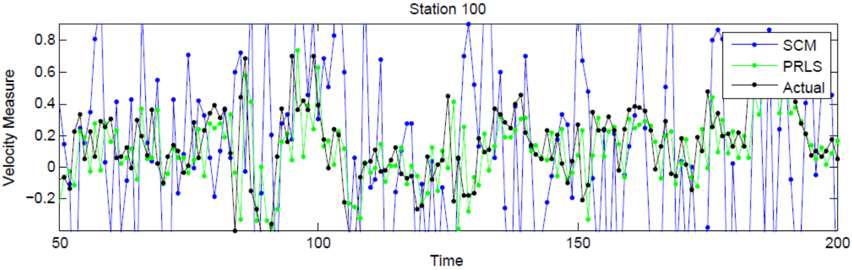
- Kronecker spectrum (left) significantly more concentrated than eigenspectrum (right)

Application: NCEP 10×10 patch over Arctic to Norwegian sea



Application: NCEP 10×10 patch over Arctic to Norwegian sea

When use PRLS for prediction get higher prediction accuracy



- Prediction by regularized sample covariance matrix (SCM):
  - Sample covariance $\hat{\mathbf{\Sigma}} = \hat{\mathbf{S}}_n$ is rank deficient
  - Min-norm (Moore-Penrose inverse) used on OLS predictor
- Prediction by Kronecker approximation to covariance
  - $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_X \otimes \hat{\mathbf{\Sigma}}_Y$ is full rank
  - Use $\hat{\mathbf{\Omega}}$ directly in OLSE predictor

Tsiligkaridis and Hero (2013). Covariance estimation in high D via kronecker product expansions. *IEEE Trans on Signal Processing*

TeraLasso: Kronecker sum covariance and inverse covariance model

Define Kronecker sum for two matrices $\mathbf{A} \in \mathbb{R}^{m \times m}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$

$$\mathbf{A} \oplus \mathbf{B} \stackrel{\text{def}}{=} \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}$$

- Kronecker sum covariance model for $p = 2$ (Rudelson 2017)[a]

$$\mathbf{\Sigma} = \mathbf{A} \oplus \mathbf{B}$$

- Sparse Kronecker sum precision model for $p = 2$ (BiGlasso) (Kalaitzis 2013)[b]

$$\mathbf{\Sigma}^{-1} = \mathbf{A} \oplus \mathbf{B}$$

- Tensor-graphical Lasso (TeraLasso): (Greenewald 2017)[b]

$$\mathbf{\Sigma}^{-1} = \mathbf{\Psi}_1 \oplus \ldots \oplus \mathbf{\Psi}_p$$

[a] Rudelson and Zhou (2017). Errors-in-variables models with dependent measurements. *Elec Journ of Statistics*

[b] Kalaitzis, Lafferty, Lawrence, Zhou (2013). The bigraphical lasso. *Int. Conf. on Machine Learning, ICML*

[c] Greenewald, Zhou, and Hero (2017). Tensor Graphical Lasso (TeraLasso). arXiv:1705.03983.

Properties of Kronecker sum

Kronecker sum $\mathbf{A} \oplus \mathbf{B} \stackrel{\text{def}}{=} \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}$ satisfies[2]

- Associativity: $\mathbf{A}_1 \oplus (\mathbf{A}_2 \oplus \mathbf{A}_3) = (\mathbf{A}_1 \oplus \mathbf{A}_2) \oplus \mathbf{A}_3$

[2] A.J. Laub (2005), *Matrix analysis for scientists and engineers,"* *SIAM Press*

Properties of Kronecker sum

Kronecker sum $\mathbf{A} \oplus \mathbf{B} \stackrel{\text{def}}{=} \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}$ satisfies[2]

- Associativity: $\mathbf{A}_1 \oplus (\mathbf{A}_2 \oplus \mathbf{A}_3) = (\mathbf{A}_1 \oplus \mathbf{A}_2) \oplus \mathbf{A}_3$
- Linearity in $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$

[2] A.J. Laub (2005), *Matrix analysis for scientists and engineers," SIAM Press*

Properties of Kronecker sum

Kronecker sum $\mathbf{A} \oplus \mathbf{B} \stackrel{\text{def}}{=} \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}$ satisfies[2]

- Associativity: $\mathbf{A}_1 \oplus (\mathbf{A}_2 \oplus \mathbf{A}_3) = (\mathbf{A}_1 \oplus \mathbf{A}_2) \oplus \mathbf{A}_3$
- Linearity in $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$
- Distributivity:

$$\begin{aligned} & c_A(\mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \mathbf{A}_3) + c_B(\mathbf{B}_1 \oplus \mathbf{B}_2 \oplus \mathbf{B}_3) \\ & \quad = (c_A\mathbf{A}_1 + \mathbf{c}_B\mathbf{B}_1) \oplus (c_A\mathbf{A}_2 + \mathbf{c}_B\mathbf{B}_2) \oplus (c_A\mathbf{A}_3 + \mathbf{c}_B\mathbf{B}_3) \end{aligned}$$

[2] A.J. Laub (2005), *Matrix analysis for scientists and engineers," SIAM Press*

Properties of Kronecker sum

Kronecker sum $\mathbf{A} \oplus \mathbf{B} \stackrel{\text{def}}{=} \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}$ satisfies[2]

- Associativity: $\mathbf{A}_1 \oplus (\mathbf{A}_2 \oplus \mathbf{A}_3) = (\mathbf{A}_1 \oplus \mathbf{A}_2) \oplus \mathbf{A}_3$
- Linearity in $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$
- Distributivity:

$$c_A(\mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \mathbf{A}_3) + c_B(\mathbf{B}_1 \oplus \mathbf{B}_2 \oplus \mathbf{B}_3) \\ = (c_A\mathbf{A}_1 + \mathbf{c}_B\mathbf{B}_1) \oplus (c_A\mathbf{A}_2 + \mathbf{c}_B\mathbf{B}_2) \oplus (c_A\mathbf{A}_3 + \mathbf{c}_B\mathbf{B}_3)$$

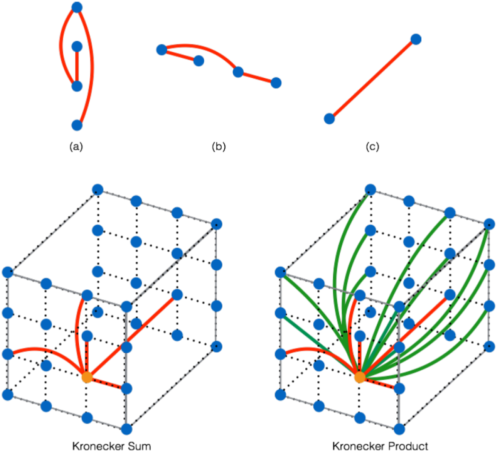
- Eigendecomposition: if $\mathbf{A}_k = \mathbf{U}_k \boldsymbol{\Lambda}_k \mathbf{U}_k^T$ is SVD

$$\mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_K = (\mathbf{U}_1 \oplus \cdots \oplus \mathbf{U}_K)(\boldsymbol{\Lambda}_1 \oplus \cdots \oplus \boldsymbol{\Lambda}_K)(\mathbf{U}_1 \oplus \cdots \oplus \mathbf{U}_K)^T$$

[2] A.J. Laub (2005), *Matrix analysis for scientists and engineers," SIAM Press*

Kronecker sum vs Kronecker product

The Kronecker sum gives more flexible and sparser covariance representations



$\mathbf{A} \oplus \mathbf{B} \oplus \mathbf{C}$ $\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}$

64 total edges 184 total edges

TeraLasso: a sparse Kronecker sum precision matrix estimator

TeraLasso estimator of $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} \in \kappa_p^\sharp$

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega} \in \kappa_p^\sharp} \left\{ -\log|\mathbf{\Omega}| + \mathrm{tr}(\hat{\mathbf{S}}\mathbf{\Omega}) + \sum_{k=1}^{K} \rho_k m_k |\mathbf{\Psi}_k|_{1,\mathrm{off}} \right\}$$

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

$$\hat{S} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{vec}(\mathbf{Z}_i^T)\mathrm{vec}(\mathbf{Z}_i^T)^T$$

and

$$\kappa_{\mathrm{p}}^\sharp = \{\mathbf{A} \succ 0 : \exists B_k \in \mathbb{R}^{d_k \times d_k} \text{ s.}t.\ \mathbf{A} = \mathbf{B}_1 \oplus \ldots \oplus \mathbf{B}_K\}$$

Convergent convex optimization via gradient-based descent, e.g., first order FISTA[a]

[a] Greenewald, Zhou, and Hero (2017). Tensor Graphical Lasso (TeraLasso). arXiv:1705.03983

TeraLasso: statistical convergence rates

(**A1**) Sparsity assumption card$(\mathcal{S}_k) \leq s_k$, where $\mathcal{S}_k$ is support set of k-th Kronecker factor Ψ_k.

(**A2**) Eigenvalues of $\mathbf{\Omega}$ satisfy $0 < \lambda_i < \infty$

Theorem

*(**Frobenius Error Bound**)[a]. Suppose the assumptions (A1)-(A2) hold, and that $\hat{\mathbf{\Omega}}$ is a minimizer with $\rho_k = a\frac{1}{k}\sqrt{\frac{\log d}{nm_k}}$. Then with probability at least $1 - 2(K+1)exp(-c\log d)$*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\| \leq \frac{C\|\mathbf{\Sigma}_0\|_2}{\lambda_{\min}(\mathbf{\Sigma}_0)}\sqrt{(K+1)(s+d)\frac{\log d}{n \min_k m_k}}$$

Compare to Glasso (Rothman 2008)[b], (Zhou)[c]. If $\Omega \in \kappa_p^\sharp$ then

$\Rightarrow$ TeraLasso statistical convergence rate is faster by a factor of $\min_k m_k$.

[a] Greenewald, Zhou, and Hero (2017). Tensor Graphical Lasso (TeraLasso). arXiv:1705.03983.

[b] Rothman, Bickel, Levina, Zhu (2008). Sparse permutation invariance covariance estimation, *Elec. Journ. Statistics*.

[c] Zhou, Rütimann, Xi, Bülmann (2011). High dimensional covariance estiamtion based on GGMs. *NIPS*.

TeraLasso: runtime comparisons for random Erdös-Renyi

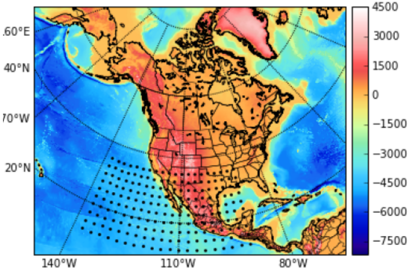
K	p	d_k	n	TeraLasso Runtime (s)	BiGLasso Runtime (s)
2	100	10	10	.0131	.84
2	625	25	10	.0147	6.81
2	2500	50	10	.0272	161
2	5625	75	10	.0401	1690
2	10^4	100	10	.0664	
2	2.5×10^5	500	10	1.62	
2	10^6	1000	10	23.2	
2	4×10^6	2000	10	427	
3	10^6	100	10	3.52	NA
3	8×10^6	200	10	11.2	NA
3	1.25×10^8	500	10	32.6	NA
3	1×10^9	1000	10	70.0	NA
4	10^4	10	10	.281	NA
4	1.6×10^5	20	10	.649	NA
4	6.25×10^6	50	10	10.8	NA
4	1.00×10^9	178	10	88.4	NA
5	1.16×10^9	65	10	124	NA

- TeraLasso speedup wrt BiGlasso by 2 to 4 orders of magnitude ($10^2 - 10^4$)
- Teralasso is scalable to many many variables

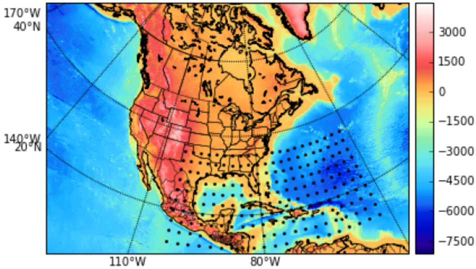
$$\text{Terralasso/iter} \Rightarrow O(\sum_{k=1}^{K} d_k^3) \qquad \text{Glasso/iter}^{a} \Rightarrow O(p^3) = O\left(\prod_{k=1}^{K} d_k^3\right)$$

[a] Guillot, Rajaratnam, Rolfs, Maleki, Wong (2012). Iterative thresholding algorithm for sparse inverse covariance estimation. *NIPS*.

Application to NCEP wind speed data

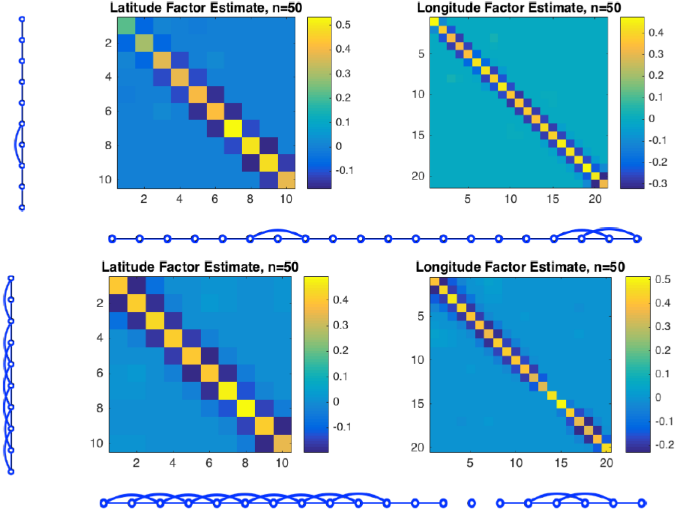


10 × 20 grid for western US



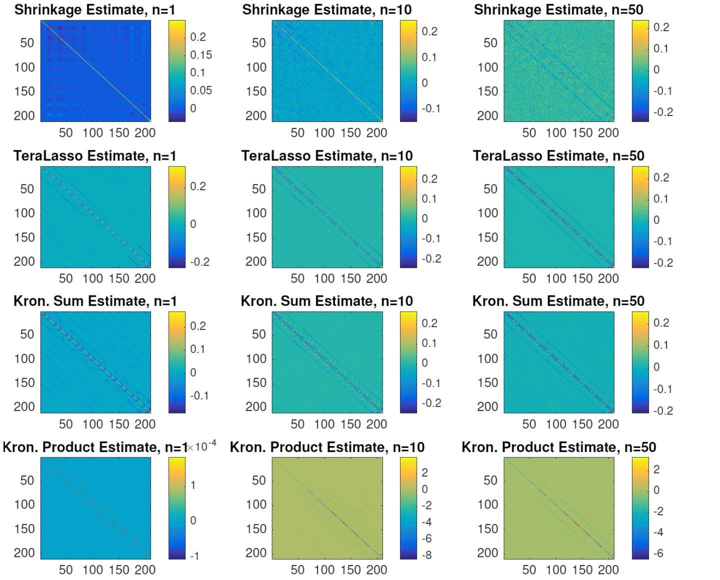
10 × 20 grid for eastern US

TeraLasso factors for spatial-only model ($K = 2$) (NCEP data)



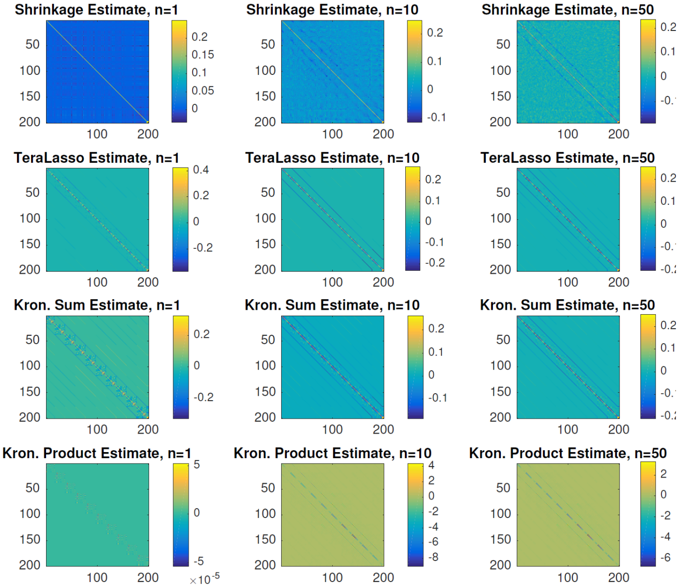
Top: Eastern grid. Bottom: Western grod. Kronecker sum factors for $d_1 = 10, d_2 = 20$

Comparisons for spatial-only model ($K = 2$) (NCEP data)



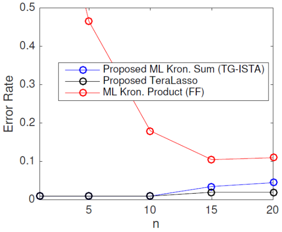
Eastern grid: inverse spatial covariance estimates for $d_1 = 10, d_2 = 20$

Comparisons for spatial-only model ($K = 2$) (NCEP data)

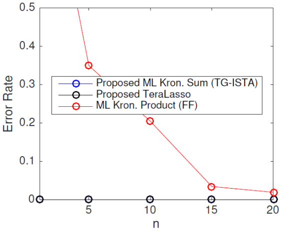


Western grid: inverse spatial covariance estimates for $d_1 = 10, d_2 = 20$

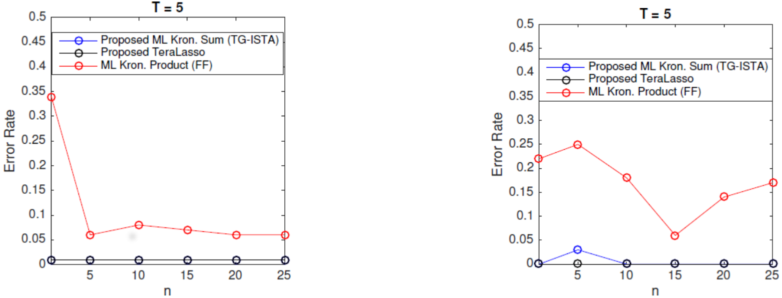
$K = 2$ vs $K = 3$ TeraLasso for season classification (NCEP data)



(a) Eastern grid



(b) Western grid

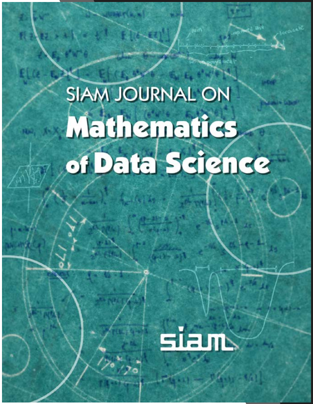


Top: $K = 2$. Bottom: $K = 3$ w/ $d_1 = 10, d_2 = 20, d_3 = 5$

Conclusions

- Value of Kronecker covariance representations for multi-indexed arrays
- Sparse Kronecker product model for matrix normal distributions (KLasso)
  - Reduces complexity of model from $O(p^2q^2)$ to $O(p+q)$
  - Bilinear non-convex objective function
- Sparse Kronecker sum model for matrix normal distributions (TeraLasso)
  - Reduces complexity of model even further than KLasso
  - Linear convex objective function

New journal



- Publishes work that advances mathematical, statistical, and computational methods in the context of data and information sciences. We invite papers that present significant advances in this context, including applications to science, engineering, business, and medicine
- Editor-in-chief: Tamara G. Kolda (Sandia)
- Section Editors:
  - Alfred Hero (Michigan)
  - Michael I. Jordan (Berkeley)
  - Robert Nowak (Wisconsin)
  - Joel A. Tropp (Caltech)
- SIAM's newest journal will begin to take author submissions in Spring 2018
- http://www.siam.org/journals/simods.php