Kronecker PCA

Alfred O. Hero

University of Michigan - Ann Arbor

Oct 9, 2013

Acknowledgement

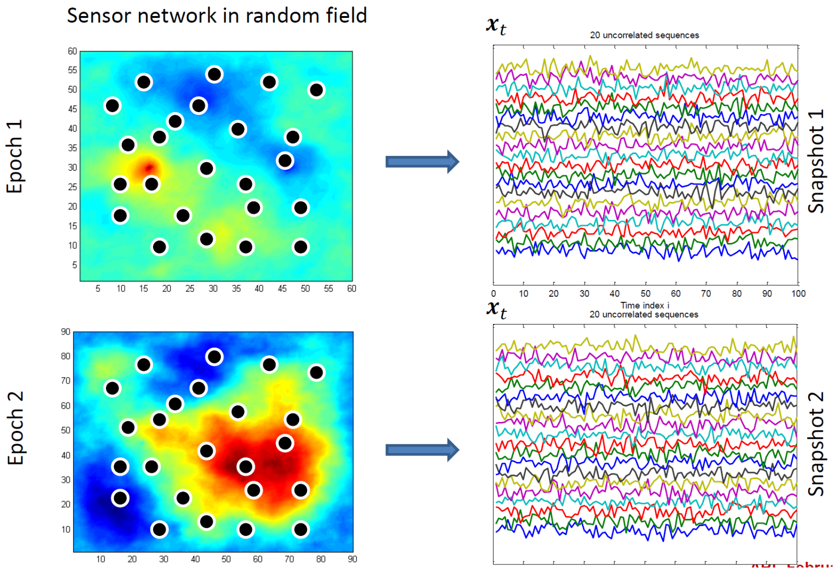
The work presented in this talk is joint with

- Theodoros Tsiligkarides, UM EECS
- Kristjan Greenwald, UM EECS
- Shuheng Zhou, UM Dept of Statistics

Papers:

- Tsiligkardis, H, Zhou, "Convergence properties of Kronecker graphical lasso algorithms," IEEE Trans on SP, Vol. 61, No 7, pp. 1743-1755, 2013. Also see arXiv:1204.0585, Mar 2012.
- Tsiligkardis and H, "Covariance estimation in high dimensions using Kronecker product expansions," to appear IEEE Trans on SP in 2013, Also see arXiv:1302.2686, Feb. 2013.
- Greenwald, Tsiligkardis and H, "Kronecker Sum Decompositions of Space-Time Data," to appear Proc. of IEEE CAMSAP, Dec 2013.

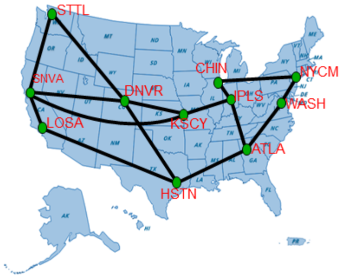
Sensor network measurements: $p = 24, q = 100, n = 3000$



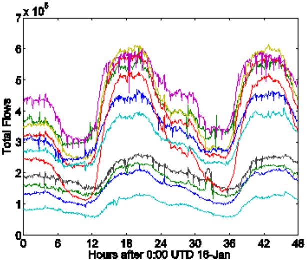
Internetwork traffic: $p = 11, q = 12, n = 365$

Correlation models and networks

11 node Abilene network



11 x 576 NetFlow measurements



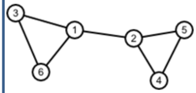
11x11 spatial correlation matrix

$$\begin{bmatrix} 1 & 0.5 & -0.5 & 0.1 & 0.2 & 0.7 & 0.1 & \dots & 0.2 \\ 0.5 & 1 & 0.1 & 2 & -0.7 & 0.1 & -0.3 & \dots & -0.1 \\ -0.5 & 0.1 & 1 & 0.4 & 0.1 & 0.6 & 0.2 & \dots & 0.1 \\ 0.1 & 0.5 & 0.4 & 1 & 0.8 & 0.2 & 0.1 & \dots & 0.2 \\ 0.2 & -0.7 & 0.1 & 0.8 & 1 & 0.1 & 0.3 & \dots & 0.1 \\ 0.7 & 0.1 & 0.6 & 0.2 & 0.1 & 1 & 0.1 & \dots & 0.1 \\ 0.1 & -0.3 & 0.2 & 0.1 & 0.3 & 0.1 & 1 & \dots & 0.2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & & 0.1 \\ 0.1 & -0.1 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 & 0.1 & 1 \end{bmatrix}$$

11x11 adjacency matrix

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Correlation graph

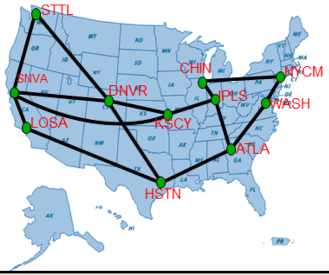


Patwari, H and Pacholski, "Manifold learning visualization of network traffic data," SIGCOMM 2005.

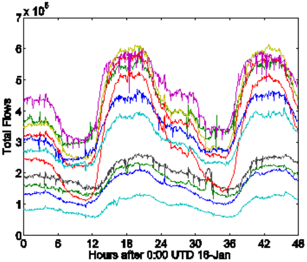
Internetwork traffic: $p = 11, q = 12, n = 365$

Correlation models and networks

11 node Abilene network



11 x 576 NetFlow measurements



1
0.5
−0.5
0.1
0.2
0.7
0.1
...
0.1

Spatio-temporal extension of spatial correlation model

1. Analysis window (block) of 1 hour durationn: q=12 time samples/block
2. Number of variables per sample: pq=132 = corr matrix dimension
3. Number of unknowns per sample: pq(pq-1)/2=8646
4. Number of samples (non-overlapping blocks) over 1 year: n=365

Patwari, Hero and Pacholski, "Manifold learning visualization of network traffic data," SIGCOMM 2005.

Statistical estimation of precision: Wishart cov model

Sample covariance matrix constructed from i.i.d. $\mathbf{z}_i \sim N(0, \mathbf{\Sigma})$

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T$$

- Then: $\mathbf{S}_n$ Wishart distributed with mean $\mathbf{\Sigma}$

$$\mathbf{S}_n \ \sim \ f(S_n; \mathbf{\Sigma}) \ \propto \ |\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\mathrm{tr}\{S_n \mathbf{\Sigma}^{-1}\}\right)$$

- Penalized MLE of $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$

$$\hat{\mathbf{\Theta}} = \mathrm{amin}_{\mathbf{\Theta}} \left(\mathrm{tr}\left\{\mathbf{S}_n \mathbf{\Theta}\right\} - \log|\mathbf{\Theta}| + R_1(\mathbf{\Theta})\right)$$

- Gaussian graphical models (GGM): Lauritzen (1996),
- Sparsity regularization: Meinshausen-Buhlmann (2006), Yuan-Lin (2007), Banerjee-ElGhaoui-d'Aspremont (2008), Friedman-Hastie-Tibshirani (2008), Chiquet (2010).

Statistical estimation of covariance: Gaussian cov model

- Assumption: $\mathbf{S}_n$ i.i.d. Gaussian distributed with mean $\boldsymbol{\Sigma}$

$$\mathbf{S}_n \sim f(S_n; \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2\sigma^2}\|S_n - \boldsymbol{\Sigma}\|_F^2\right)$$

Frobenius norm of square matrix $\mathbf{A}$:

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{d}|\mathbf{A}_{i,j}|$$

- Penalized MLE of $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\Sigma}} = \mathrm{amin}_{\boldsymbol{\Sigma}}\|\mathbf{S}_n - \boldsymbol{\Sigma}\|_F^2 + R_2(\boldsymbol{\Sigma})$$

- Cov estimation: Furrer-Bengtsson (2007), Gini-Greco (2002), Lounici (2012), Vershynin (2011)
- Shrinkage regularization: Ledoit-Wolf (2000), Schafer-Strimmer (2007), Chen-Wiesel-Eldar-H (2010)
- Sparsity reg: Dempster (1972), Rothman-Bickel-Levina (2008)

Sparse covariance and precision models

Two types of sparse covariance models:

- Sparse correlation graphical models:
    - Most correlations are zero, few marginal dependencies
    - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation graphical models
    - Most inverse covariance entries are zero, few conditional dependencies
    - Examples: Markov random fields, autoregressive (AR) processes, global latent variables

Sparse covariance and precision models

Two types of sparse covariance models:

- Sparse correlation graphical models:
  - Most correlations are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation 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.

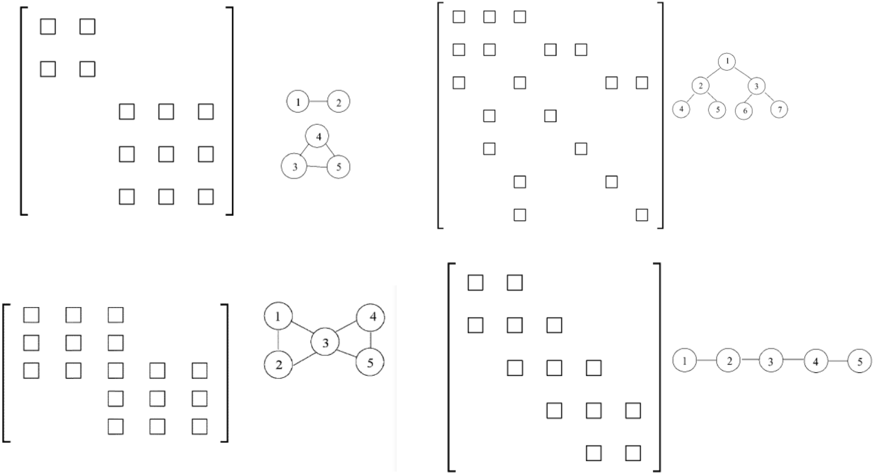
Sparse covariance and precision models

Two types of sparse covariance models:

- Sparse correlation graphical models:
  - Most correlations are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation 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.

Gallery of sparsity patterns and associated graphs



Wiesel, Eldar and H IEEE TSP 2010

Sparsity and multivariate dependency

Two types of sparse covariance models:

- Sparse correlation graphical models:
  - Most correlations are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation 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.

• Another way to reduce parameter dimension: Kronecker product constraint.

Kronecker product model for covariance matrix

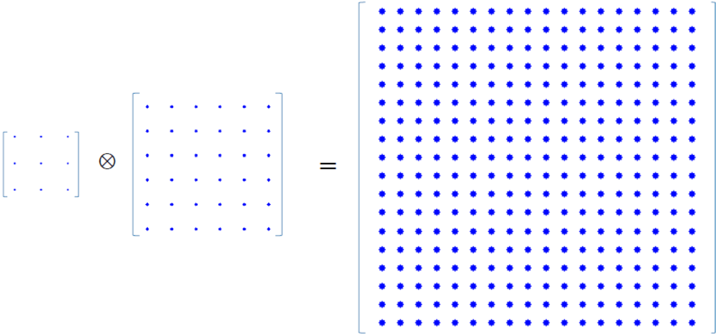


Figure: A saturated model with 18×18 covariance matrixl has 18*17/2=153 unknown correlation parameters. A Kronecker product covariance model reduces number of parameters to $3 + 15 = 18$ unknown correlation parameters.

Sparse Kronecker product model for covariance matrix

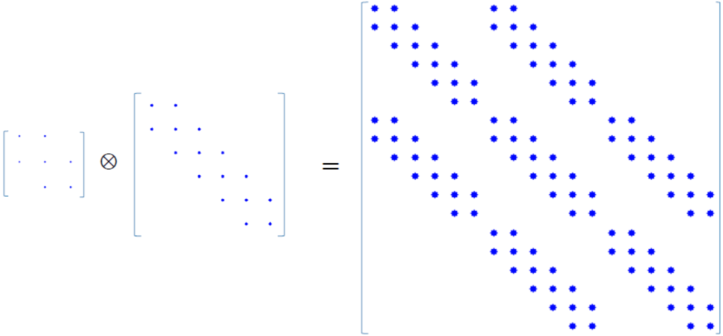


Figure: A sparse Kronecker product covariance model reduces number of parameters from 153 to 7 unknown correlation parameters.

Kronecker products of square matrices

Let $\mathbf{A}$ be a $p \times p$ matrix and $\mathbf{B}$ be a $q \times q$ matrix. For $d = pq$ define the $d \times d$ matrix $\mathbf{C}$ by the Kronecker product factorization $\mathbf{C} = \mathbf{A} \bigotimes \mathbf{B}$ where

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

Kronecker product properties (VanLoan-Pitsianis 1992):

- $\mathbf{C}$ is p.d. if $\mathbf{A}$ and $\mathbf{B}$ are p.d.
- $\mathbf{C}^{-1} = \mathbf{A}^{-1} \bigotimes \mathbf{B}^{-1}$ if $\mathbf{A}$ and $\mathbf{B}$ are invertible.
- $|\mathbf{C}| = |\mathbf{A}| \; |\mathbf{B}|$
- Linear Kronecker equations benefit from reduced computation

$$(\mathbf{A} \bigotimes \mathbf{B})\mathbf{z} = \mathbf{r} \;\; \Leftrightarrow \;\; \mathbf{B}\mathbf{Z}\mathbf{A}^T = \mathbf{R}$$

Matrix variate normal likelihood model

Let $\mathbf{z} \in \mathbb{R}^d$, where $d = pq$, be zero mean with covariance matrix $\mathbf{\Sigma} = \mathbf{A} \bigotimes \mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathbf{B} \in \mathbb{R}^{q \times q}$ are p.d.

When $\mathbf{z}$ is Gaussian distributed it is said to follow the matrix variate normal distribution (Dawid 1981, Gupta-Nagar 1999)

$$\begin{aligned} f(\mathbf{z}; \mathbf{A}, \mathbf{B}) &\propto (|\mathbf{A}|\,|\mathbf{B}|)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{z}^T(\mathbf{A}^{-1} \bigotimes \mathbf{B}^{-1})\mathbf{z}\right) \\ &= (|\mathbf{X}|\,|\mathbf{Y}|)^{1/2} \exp\left(-\frac{1}{2}\mathbf{z}^T(\mathbf{X} \bigotimes \mathbf{Y})\mathbf{z}\right) \end{aligned}$$

$\mathbf{X} = \mathbf{A}^{-1}$, $\mathbf{Y} = \mathbf{B}^{-1}$.

Matrix variate normal likelihood model

Let $\mathbf{z} \in \mathbb{R}^d$, where $d = pq$, be zero mean with covariance matrix $\mathbf{\Sigma} = \mathbf{A} \bigotimes \mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathbf{B} \in \mathbb{R}^{q \times q}$ are p.d.

When $\mathbf{z}$ is Gaussian distributed it is said to follow the matrix variate normal distribution (Dawid 1981, Gupta-Nagar 1999)

$$\begin{aligned} f(\mathbf{z}; \mathbf{A}, \mathbf{B}) &\propto (|\mathbf{A}|\,|\mathbf{B}|)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{z}^T(\mathbf{A}^{-1} \bigotimes \mathbf{B}^{-1})\mathbf{z}\right) \\ &= (|\mathbf{X}|\,|\mathbf{Y}|)^{1/2} \exp\left(-\frac{1}{2}\mathbf{z}^T(\mathbf{X} \bigotimes \mathbf{Y})\mathbf{z}\right) \end{aligned}$$

$\mathbf{X} = \mathbf{A}^{-1}$, $\mathbf{Y} = \mathbf{B}^{-1}$.

MLE of $\mathbf{\Sigma} = \mathbf{A} \bigotimes \mathbf{B}$ and $\mathbf{\Theta} = \mathbf{X} \bigotimes \mathbf{Y}$

- Likelihood function is biconvex in $\mathbf{X}$ and $\mathbf{Y}$
- Can solve for MLE using alternating maximization methods
  - "Flip-flop" algorithms proposed by Dutilleul (1999) and Werner-Jansen-Stoica (2008)

Sparse matrix variate normal model: the KGlasso

Let $\mathbf{z}$ follow the matrix normal normal distribution with precision matrix $\mathbf{\Theta} = \mathbf{X} \bigotimes \mathbf{Y}$. If

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

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

Algorithms for estimating sparse $\mathbf{\Theta}$ try to solve

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

where

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

Alternating minimization algorithms

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

KGlasso Algorithm

Algorithm 1 KGlasso (Tsiligkaridis et al. [2012])

Input: $\hat{\mathbf{S}}_n$, p, f, n, $\lambda_X > 0$, $\lambda_Y > 0$
Output: $\hat{\boldsymbol{\Theta}}_{KGlasso}$
Initialize $\mathbf{A}_{init}$ to be positive definite.
$\hat{\mathbf{A}} \leftarrow \mathbf{A}_{init}$
repeat
  $\hat{\mathbf{B}} \leftarrow \frac{1}{p}\sum_{i,j=1}^{p} [\hat{\mathbf{A}}^{-1}]_{i,j}\hat{\mathbf{S}}_n(j,i)$
  $\check{\mathbf{Y}} \leftarrow \arg\min_{\mathbf{Y}\in S_{++}^{q}} \mathrm{tr}(\mathbf{Y}\hat{\mathbf{B}}) - \log|\mathbf{Y}| + \lambda_Y|\mathbf{Y}|_1$
  $\hat{\mathbf{A}} \leftarrow \frac{1}{q}\sum_{k,l=1}^{q} [\hat{\mathbf{B}}^{-1}]_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)$
  $\check{\mathbf{X}} \leftarrow \arg\min_{\mathbf{X}\in S_{++}^{p}} \mathrm{tr}(\mathbf{X}\hat{\mathbf{A}}) - \log|\mathbf{X}| + \lambda_X|\mathbf{X}|_1$
until convergence
$\hat{\boldsymbol{\Theta}}_{KGlasso} \leftarrow \check{\mathbf{X}} \otimes \check{\mathbf{Y}}$

Computational complexity: $\mathcal{O}(p^3 + q^3)$ (KGlasso) vs $\mathcal{O}(p^3 q^3)$ (Glasso).

Limit Points of KGlasso (Tsiligkaridis et al. [2012])

Define $J_\lambda^{(k)} = J_\lambda(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})$ for $k = 0, 1, 2, \ldots$.

Theorem

1. *If $\hat{\mathbf{S}}_n$ is positive definite, KGlasso converges to a fixed point. Also, we have $J_\lambda^{(k)} \searrow J_\lambda^{(\infty)}$.*
2. *Assume that $n > pq$ and that the KGlasso is not initialized at a local maximum. Then the algorithm converges to a local minimum.*

High dimensional convergence rates for p.d. $\hat{\mathbf{S}}_n \in \mathbb{R}^{d\times d}$

Convergence rates in n, p, q [Yuan-Lin 2008, Tsiligkaridis-H 2012]

1. $\hat{\mathbf{\Theta}}_{SCM} = \mathrm{amin}_{\mathbf{\Theta}}\{\mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n - \log|\mathbf{\Theta}|)\}$. Then:
   $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F = O\left(\sqrt{p^2q^2/n}\right)$

High dimensional convergence rates for p.d. $\hat{\mathbf{S}}_n \in \mathbb{R}^{d \times d}$

Convergence rates in n, p, q [Yuan-Lin 2008, Tsiligkaridis-H 2012]

1. $\hat{\mathbf{\Theta}}_{SCM} = \mathrm{amin}_{\mathbf{\Theta}}\{\mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n - \log|\mathbf{\Theta}|)\}$. Then:
   $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F = O\left(\sqrt{p^2 q^2/n}\right)$
2. $\hat{\mathbf{\Theta}}_{GLasso} = \mathrm{amin}_{\mathbf{\Theta}}\{\mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log|\mathbf{\Theta}|) + \lambda|\mathbf{\Theta}|_1\}$. If $\lambda \asymp \sqrt{pq\log(pq)/n}$. Then:
   $\|\hat{\mathbf{\Theta}}_{GLasso} - \mathbf{\Theta}\|_F = O\left(\sqrt{pq\log(pq)/n}\right)$

High dimensional convergence rates for p.d. $\hat{\mathbf{S}}_n \in \mathbb{R}^{d\times d}$

Convergence rates in n, p, q [Yuan-Lin 2008, Tsiligkaridis-H 2012]

1. $\hat{\mathbf{\Theta}}_{SCM} = \mathrm{amin}_{\mathbf{\Theta}}\{\mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n - \log|\mathbf{\Theta}|)\}$. Then: $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F = O\left(\sqrt{p^2q^2/n}\right)$
2. $\hat{\mathbf{\Theta}}_{GLasso} = \mathrm{amin}_{\mathbf{\Theta}}\{\mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log|\mathbf{\Theta}|) + \lambda|\mathbf{\Theta}|_1\}$. If $\lambda \asymp \sqrt{pq\log(pq)/n}$. Then: $\|\hat{\mathbf{\Theta}}_{GLasso} - \mathbf{\Theta}\|_F = O\left(\sqrt{pq\log(pq)/n}\right)$
3. $\hat{\mathbf{\Theta}}_{FF} = \mathrm{amin}_{\mathbf{X},\mathbf{Y}}\{\mathrm{tr}((\mathbf{X}\bigotimes\mathbf{Y})\hat{\mathbf{S}}_n) - \log(|\mathbf{X}|\,|\mathbf{Y}|)\}$. Then: $\|\mathbf{\Theta}_{FF} - \mathbf{\Theta}\|_F = O\left(\sqrt{(p^2+q^2)\log(p+q)/n}\right)$

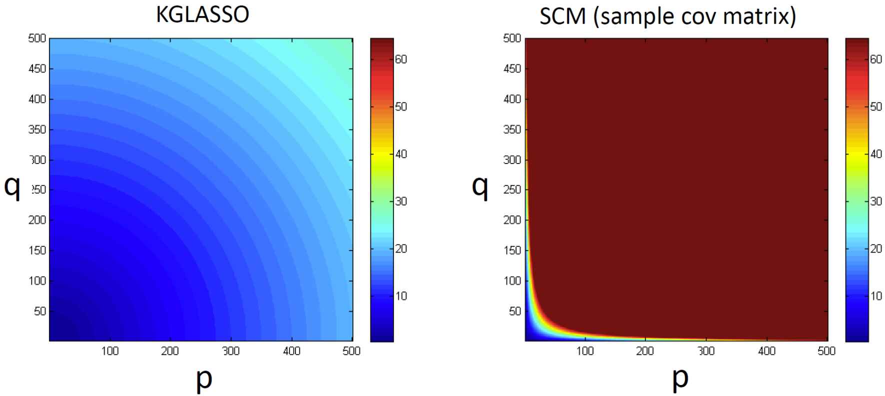
High dimensional convergence rates for p.d. $\hat{\mathbf{S}}_n \in \mathbb{R}^{d\times d}$

Convergence rates in n, p, q [Yuan-Lin 2008, Tsiligkaridis-H 2012]

1. $\hat{\mathbf{\Theta}}_{SCM} = \text{amin}_{\mathbf{\Theta}}\{\text{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n - \log|\mathbf{\Theta}|)\}$. Then:
   $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F = O\left(\sqrt{p^2q^2/n}\right)$
2. $\hat{\mathbf{\Theta}}_{GLasso} = \text{amin}_{\mathbf{\Theta}}\{\text{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log|\mathbf{\Theta}|) + \lambda|\mathbf{\Theta}|_1\}$. If $\lambda \asymp \sqrt{pq\log(pq)/n}$. Then:
   $\|\hat{\mathbf{\Theta}}_{GLasso} - \mathbf{\Theta}\|_F = O\left(\sqrt{pq\log(pq)/n}\right)$
3. $\hat{\mathbf{\Theta}}_{FF} = \text{amin}_{\mathbf{X},\mathbf{Y}}\{\text{tr}((\mathbf{X}\bigotimes\mathbf{Y})\hat{\mathbf{S}}_n) - \log(|\mathbf{X}|\,|\mathbf{Y}|)\}$. Then:
   $\|\mathbf{\Theta}_{FF} - \mathbf{\Theta}\|_F = O\left(\sqrt{(p^2+q^2)\log(p+q)/n}\right)$
4. $\hat{\mathbf{\Theta}}_{KGlasso} = \text{amin}_{\mathbf{X},\mathbf{Y}}\{\text{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log|\mathbf{\Theta}| + \lambda_X|\mathbf{X}|_1 + \lambda_Y|\mathbf{Y}|_1\}$. If $\lambda_X, \lambda_Y \asymp \sqrt{(p+q)\log(p+q)/n}$. Then:
   $\|\mathbf{\Theta}_{KGlasso} - \mathbf{\Theta}\|_F = O\left(\sqrt{(p+q)\log(p+q)/n}\right)$

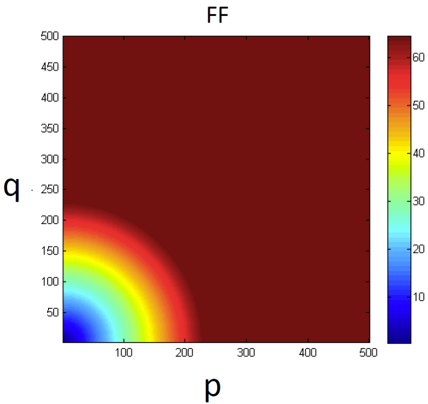
Phase transition maps

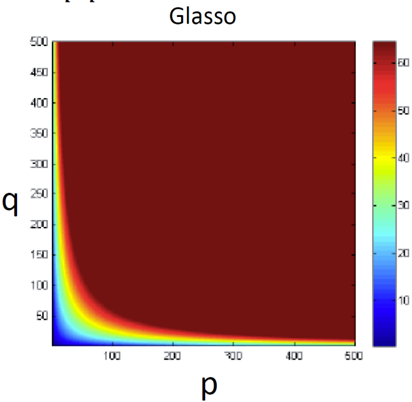
- KGlasso phase transition occurs when $p + q > n$
- SCM phase transition occurs when $p^2 q^2 > n$



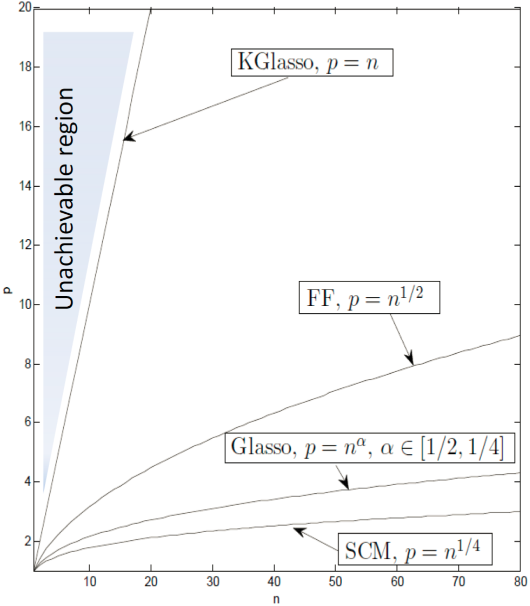
Phase transition maps

- FF has phase transition that occurs when $p^2+q^2 > n$
- Glasso phase transition occurs when $pq > n$





Phase transition boundaries in p vs. n plane ($p = q$)



Kronecker sum decomposition

Theorem (VanLoan-Pitsianis 1993)

Any $pq \times pq$ matrix $\mathbf{C}$ can be represented as

$$\mathbf{C} = \sum_{i=1}^{r} \alpha_i \mathbf{A}_i \bigotimes \mathbf{B}_i$$

for some r, some sequence of $\mathbf{A}_i \in \mathbb{R}^{p \times p}$ and $\mathbf{B}_i \in \mathbb{R}^{q \times q}$ and some coefficients α_i. For given p, q the minimal feasible value of $r = r(p, q)$ is called the (p, q)-separation rank of $\mathbf{C}$.

Towards a Kronecker sum approximation to $\mathbf{S}_n$

Theorem (Eckart-Young (1936), Schmidt (1907))

For matrix $\mathbf{D} \in \mathbb{R}^{m,n}$ and $r > 0$ the solution to

$$\min_{\mathbf{C}:\mathrm{rank}(\mathbf{C})\leq r} \|\mathbf{D} - \mathbf{C}\|_F$$

is the truncated SVD of $\mathbf{D}$

$$\mathbf{C} = \sum_{i=1}^{r} \sigma_i \, \boldsymbol{\xi}_i \boldsymbol{\nu}_i^T$$

where $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$ are singular values and $\{\boldsymbol{\xi}_i, \boldsymbol{\nu}_i\}_{i=1}^{\min(m,n)}$ are associated singular vectors of $\mathbf{D}$.

Towards a Kronecker sum approximation to $\mathbf{S}_n$

Nuclear norm convex relaxation of EY (Fazel 2002, Recht-Fazel-Parillo 2007, Hiriart-Urruty and Le 2011):

$$\min_{\mathbf{C}\in\mathbb{R}^{m,n}} \{\|\mathbf{D}-\mathbf{C}\|_F^2 + \beta\|\mathbf{C}\|_*\}$$

where $\|\mathbf{C}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{C})$.

Towards a Kronecker sum approximation to $\mathbf{S}_n$

Nuclear norm convex relaxation of EY (Fazel 2002, Recht-Fazel-Parillo 2007, Hiriart-Urruty and Le 2011):

$$\min_{\mathbf{C}\in\mathbb{R}^{m,n}} \{\|\mathbf{D}-\mathbf{C}\|_F^2 + \beta\|\mathbf{C}\|_*\}$$

where $\|\mathbf{C}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{C})$.

$\Rightarrow$ Applied to covariance matrices $\mathbf{S}_n \in \mathbb{R}^{d\times d}$ (Lounici, 2012):

$$\min_{\mathbf{C}\in\mathcal{S}_+^d} \{\|\mathbf{S}_n-\mathbf{C}\|_F^2 + \beta\|\mathbf{C}\|_*\}$$

where $S_+^d = \{\mathbf{C} : \mathbf{C} \in \mathbb{R}^{d\times d} \text{ and } \mathbf{C} \text{ p.s.d.}\}$.

- Lounici gave oracle bounds on accuracy of the solution $\mathbf{C}_\beta$.

Towards a Kronecker sum approximation to $\mathbf{S}_n$

Nuclear norm convex relaxation of EY (Fazel 2002, Recht-Fazel-Parillo 2007, Hiriart-Urruty and Le 2011):

$$\min_{\mathbf{C}\in\mathbb{R}^{m,n}} \{\|\mathbf{D}-\mathbf{C}\|_F^2 + \beta\|\mathbf{C}\|_*\}$$

where $\|\mathbf{C}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{C})$.

$\Rightarrow$ Applied to covariance matrices $\mathbf{S}_n \in \mathbb{R}^{d\times d}$ (Lounici, 2012):

$$\min_{\mathbf{C}\in\mathcal{S}_+^d} \{\|\mathbf{S}_n-\mathbf{C}\|_F^2 + \beta\|\mathbf{C}\|_*\}$$

where $S_+^d = \{\mathbf{C} : \mathbf{C} \in \mathbb{R}^{d\times d} \text{ and } \mathbf{C} \text{ p.s.d.}\}$.

- Lounici gave oracle bounds on accuracy of the solution $\mathbf{C}_\beta$.

$\Rightarrow$ convex relaxation for Kronecker sum approximation?

Reduced rank Kronecker sum approximation

Reduced rank Kronecker sum approximation

$$\min_{\mathbf{S}(\mathbf{A},\mathbf{B}):\mathrm{rank}(\mathbf{S}(\mathbf{A},\mathbf{B}))\leq r} \|\mathbf{S}_n - \mathbf{S}(\mathbf{A},\mathbf{B})\|_F^2$$

where $\mathbf{S}_n$ is the sample covariance matrix and

$$\mathbf{S}(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{r} \alpha_i \mathbf{A}_i \bigotimes \mathbf{B}_i$$

Here α_i, $\mathbf{A}_i$, $\mathbf{B}_i$ are to be determined while p, q are fixed.

Reduced rank Kronecker sum approximation

Reduced rank Kronecker sum approximation

$$\min_{\mathbf{S}(\mathbf{A},\mathbf{B}):\mathrm{rank}(\mathbf{S}(\mathbf{A},\mathbf{B}))\leq r} \|\mathbf{S}_n - \mathbf{S}(\mathbf{A},\mathbf{B})\|_F^2$$

where $\mathbf{S}_n$ is the sample covariance matrix and

$$\mathbf{S}(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{r} \alpha_i \mathbf{A}_i \bigotimes \mathbf{B}_i$$

Here α_i, $\mathbf{A}_i$, $\mathbf{B}_i$ are to be determined while p, q are fixed.

Solution of this problem appears difficult (Hiriarty-Urruty and Le 2013).

Alternative: use permuted approximation error

Consider the following rank one permuted representation of the Frobenius norm of Kronecker products (VanLoan-Pitsianis 1992).

For any matrices $\mathbf{D} \in \mathbb{R}^{pq \times pq}$, $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathbf{B} \in \mathbb{R}^{q \times q}$:

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

where

- $\mathrm{vec}(\bullet)$ is the vectorization operator, e.g. when applied to $\mathbf{A}$ it maps $\mathbb{R}^{p \times p}$ to $\mathbb{R}^{p^2}$
- $\mathcal{R}(\bullet)$ is a permutation rearrangement operator mapping $\mathbb{R}^{pq \times pq}$ to $\mathbb{R}^{p^2 \times q^2}$

Permutation-rearrangement operator pair $\mathcal{R}$, $\mathcal{R}^{-1}$

Let $\mathbf{D} \in \mathbb{R}^{pq \times pq}$ have the $q \times q$ block partition

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{1:q,1:q} & \cdots & \mathbf{D}_{1:q,(p-1)q:pq} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{(p-1)q:pq,1:q} & \cdots & \mathbf{D}_{(p-1)q:pq,(p-1)q:pq} \end{bmatrix}$$

Define the rearrangement operator and its inverse

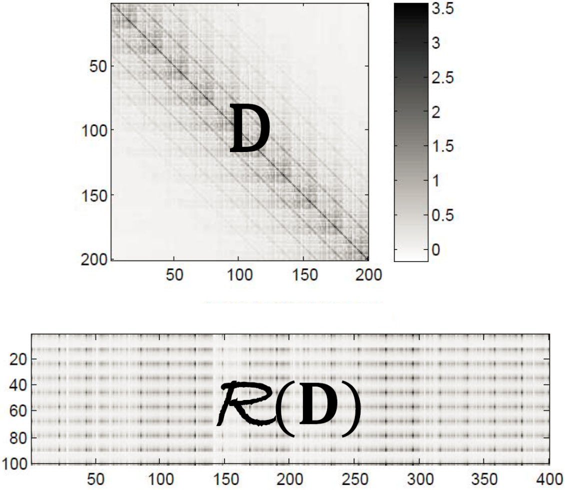
- $\mathcal{R} : \mathbb{R}^{pq \times pq} \to \mathbb{R}^{p^2 \times q^2}$ is the rearrangement operator

$$\mathcal{R}(\mathbf{D}) = \begin{bmatrix} \left(\mathrm{vec}\mathbf{D}_{1:q,1:q}\right)^T \\ \vdots \\ \left(\mathrm{vec}\mathbf{D}_{(p-1)q:pq,(p-1)q:pq}\right)^T \end{bmatrix}$$

- $\mathcal{R}^{-1} : \mathbb{R}^{p^2 \times q^2} \to \mathbb{R}^{pq \times pq}$ is the operator inverse of $\mathcal{R}$

$$\mathcal{R}^{-1}(\mathcal{R}(\mathbf{D})) = \mathbf{D}$$

Illustration of permutation operator



Permuted Rank Least Squares (PRLS)

Consider the nuclear norm PLS optimization

$$\hat{\mathbf{R}} = \mathrm{amin}_{\mathbf{R}}\{\|\mathcal{R}(\mathbf{S}_n) - \mathbf{R}\|_F^2 + \beta\|\mathbf{R}\|_*\}$$

The minimization is over $\mathbf{R} \in \mathbb{R}^{p^2 \times q^2}$ and we define the PRLS Kronecker sum approximation to $\mathbf{S}_n$ as $\hat{\mathbf{\Sigma}} = \mathcal{R}^{-1}(\hat{\mathbf{R}})$.

Theorem (Tsiligkaridis-H 2013)

The PRLS Kronecker sum approximation is

$$\hat{\mathbf{\Sigma}} = \mathcal{R}^{-1}\left(\hat{\mathbf{R}}\right), \quad \hat{\mathbf{R}} = \sum_{i=1}^{\min(p^2,q^2)} \left(\sigma_i - \frac{\beta}{2}\right)_+ \mathbf{u}_i \mathbf{v}_i^T$$

- $(\sigma_k, \mathbf{u}_k, \mathbf{v}_k)$ *is the* k*-th component of the SVD of* $\mathcal{R}(\mathbf{S}_n)$

PRLS Kronecker sum decomposition properties

Permuted-rank least squares (PRLS) algorithm produces a solution

$$\hat{\mathbf{\Sigma}} = \mathcal{R}^{-1}(\hat{\mathbf{R}})$$

that satisfies the following properties (Tsiligkaridis-H 2013).

PRLS solution of separation rank r is

- a positive definite matrix (w.p.1) if $n \geq pq$
- a symmetric matrix: $\hat{\mathbf{\Sigma}}^T = \hat{\mathbf{\Sigma}}$
- a Kronecker sum: $\hat{\mathbf{\Sigma}} = \sum_{\gamma=1}^{r} \hat{\alpha}_i \, \hat{\mathbf{A}}_i \bigotimes \hat{\mathbf{B}}_i$

PRLS Kronecker sum decomposition properties

Permuted-rank least squares (PRLS) algorithm produces a solution

$$\hat{\mathbf{\Sigma}} = \mathcal{R}^{-1}(\hat{\mathbf{R}})$$

that satisfies the following properties (Tsiligkaridis-H 2013).

PRLS solution of separation rank r is

- a positive definite matrix (w.p.1) if $n \geq pq$
- a symmetric matrix: $\hat{\mathbf{\Sigma}}^T = \hat{\mathbf{\Sigma}}$
- a Kronecker sum: $\hat{\mathbf{\Sigma}} = \sum_{\gamma=1}^{r} \hat{\alpha}_i \, \hat{\mathbf{A}}_i \bigotimes \hat{\mathbf{B}}_i$

PRLS solution of separation rank r is not

- a solution of algebraic rank r
- an orthonormal basis decomposition: $\mathrm{tr}\{(\hat{\mathbf{A}}_i \bigotimes \hat{\mathbf{B}}_i)(\hat{\mathbf{A}}_j \bigotimes \hat{\mathbf{B}}_j)\} \neq \delta_{i-j}$

MSE convergence rates

Theorem (Tsiligkaridis-H 2013)

Assume $\mathbf{S}_n \in \mathbb{R}^{pq \times pq}$ *is p.d and let* $M = \max(p, q, n)$. *Let* λ *satisfy*

$$\lambda = C(p^2 + q^2 + \log(M))/n$$

Then, with probability at least $1 - 2M^{-1/4C}$ *the matched PRLS estimator* $\hat{\mathbf{\Sigma}}_{p.q.r}$ *of* $\mathbf{\Sigma}$ *satisfies:*

$$\begin{aligned} \|\hat{\mathbf{\Sigma}}_{p.q.r} - \mathbf{\Sigma}\|_F^2 \quad \leq \quad & \min_{\mathbf{R}:\mathrm{rank}(\mathbf{R}) \leq r} \|\mathbf{R} - \mathcal{R}(\mathbf{\Sigma})\|_F^2 \\ & + C^{'} \left(r(p^2 + q^2 + \log(M))/n \right) \end{aligned}$$

where $C^{'} = (1.5(1 + \sqrt{2})C)^2$.

Proof: largely inspired by Lounici 2012.

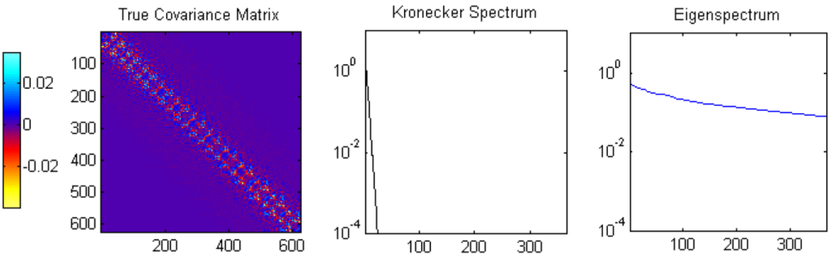
Simulation: Block Toeplitz covariance

Step 1 : Generate vector AR(1) process $\mathbf{z}_t \in \mathbb{R}^p$

$$\mathbf{z}_t = \mathbf{\Phi}\mathbf{z}_{t-1} + \mathcal{E}_t, \quad t = 1, 2, \ldots$$

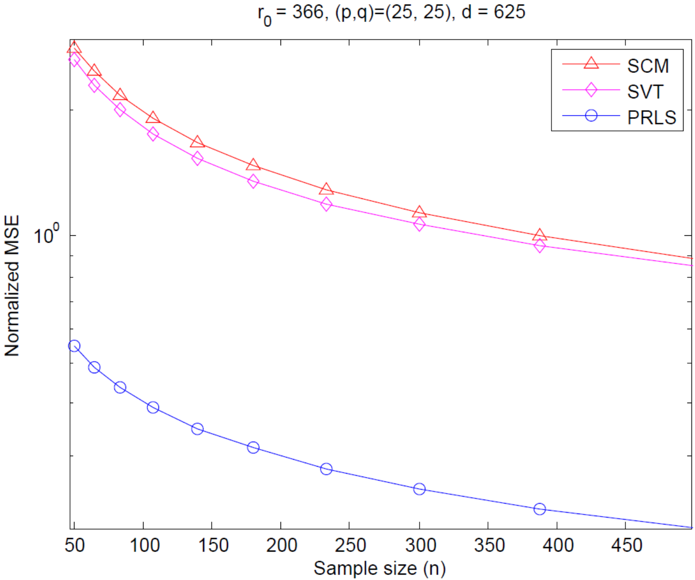
Step 2 : Concatenate AR(1) vectors into $\mathbf{Z}_t \in \mathbb{R}^{pq}$

$$\mathbf{Z}_t = [\mathbf{z}_{p+m}^T, \mathbf{z}_{p+2m}^T, \ldots, \mathbf{z}_{p+qm}^T]^T$$



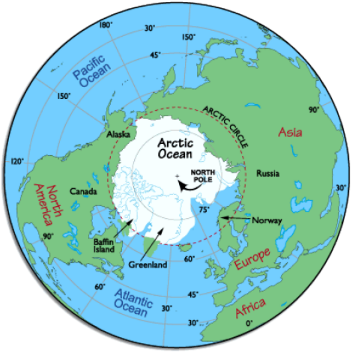
$p = q = 25$, $\|\mathbf{\Phi}\|_2 = 0.95$.

Simulation: Block Toeplitz covariance



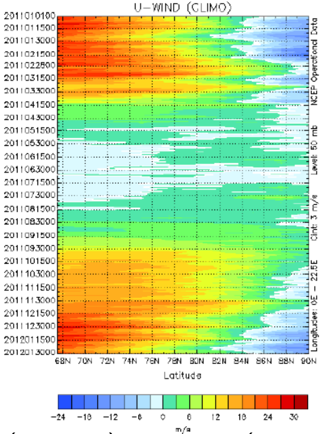
Application: National Center for Environmental Prediction

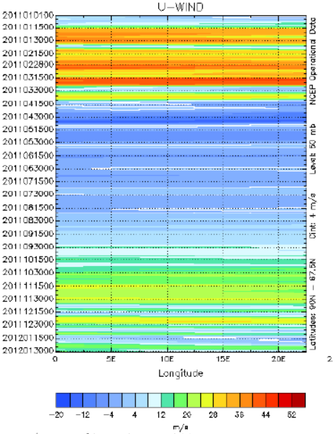
- Wind speed data (1948-2012)
- 100 stations in 10x10 grid
- 2-day time windows (8 sample snapshots)
- Period: 2001 to 2007
- p=100, q=8, n=224



Application: National Center for Environmental Prediction

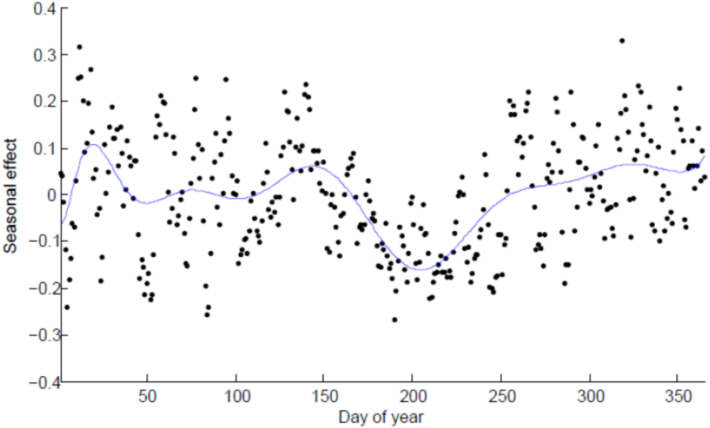
U component of windspeed



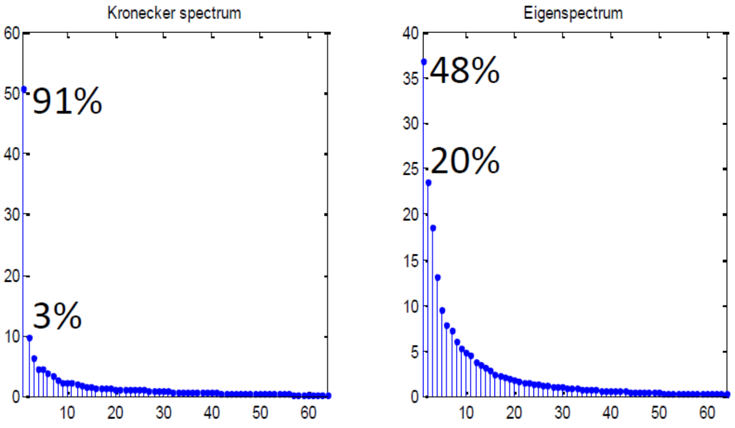


Lat (90-67.5) and long (0-22.5) is over 2.5 degree increments

Application: National Center for Environmental Prediction

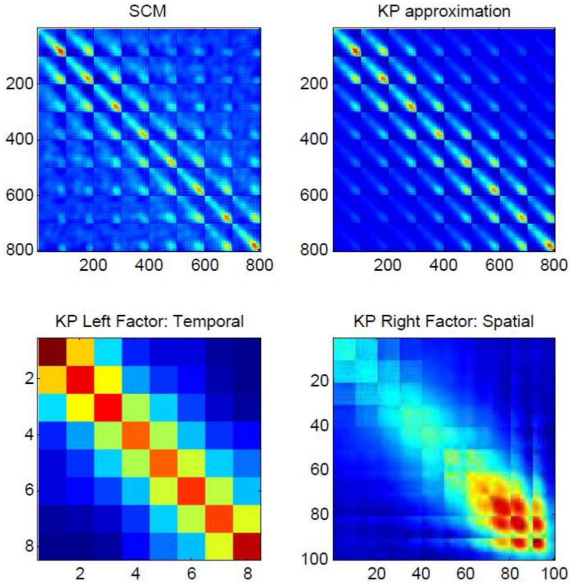


Application: National Center for Environmental Prediction



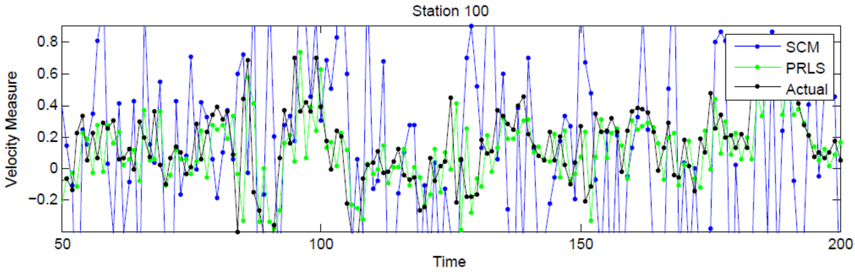
- Kronecker spectrum (left) significantly more concentrated than eigenspectrum (right)

Application: National Center for Environmental Prediction



Application: National Center for Environmental Prediction

When use PRLS for prediction get higher prediction accuracy



- SCM:
  - $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$ is rank deficient
  - Prediction by min-norm (Moore-Penrose inverse) linear regression
- PRLS
  - $\hat{\mathbf{\Sigma}}$ for PRLS is full rank
  - Prediction by standard linear regression

Conclusion

- Kronecker product covariance models are scalable.

Conclusion

- Kronecker product covariance models are scalable.
- Significant advantages in MSE convergence rate and phase transition behavior

Conclusion

- Kronecker product covariance models are scalable.
- Significant advantages in MSE convergence rate and phase transition behavior
- These rates guide scaling of regularization parameters λ

Conclusion

- Kronecker product covariance models are scalable.
- Significant advantages in MSE convergence rate and phase transition behavior
- These rates guide scaling of regularization parameters λ
- Good match to certain experimental data streams
    - wind speed network prediction
    - activity recognition
    - video classification

Conclusion

- Kronecker product covariance models are scalable.
- Significant advantages in MSE convergence rate and phase transition behavior
- These rates guide scaling of regularization parameters λ
- Good match to certain experimental data streams
  - wind speed network prediction
  - activity recognition
  - video classification
- Open problems
  - Missing data
  - MLE for Kronecker sum models
  - Relation between separation rank and algebraic rank

References

Neal Patwari, III Alfred O. Hero, and Adam Pacholski. Manifold learning visualization of network traffic data. In *MineNet '05: Proceeding of the 2005 ACM SIGCOMM workshop on Mining network data*, pages 191–196, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-026-4. doi: http://doi.acm.org/10.1145/1080173.1080182.

T. Tsiligkaridis and A.O. Hero. Covariance estimation in high dimensions via kronecker product expansions. *arXiv preprint arXiv:1302.2686*, 2013.

T. Tsiligkaridis, Alfred Hero, and Shuheng Zhou. Convergence properties of Kronecker Graphical Lasso algorithms. *arXiv:1204.0585*, April 2012.