Computational and Statistical Methodology for Highly-Structured Data

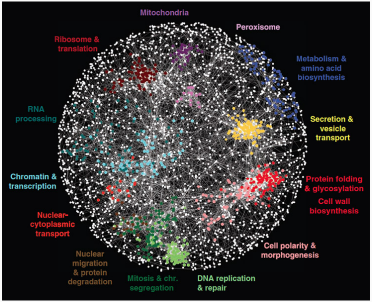
Ph.D.Thesis Defense: 2020-09-10

Michael Weylandt

Slides Available Online at `https://tinyurl.com/WeylandtDefense`

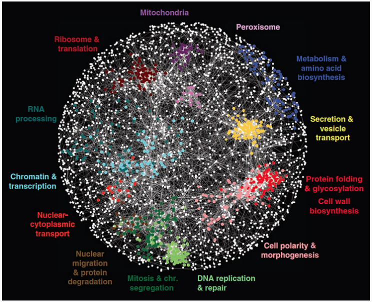
Rice University, Department of Statistics

Big-Data



"Big Data" enables "Big Models"

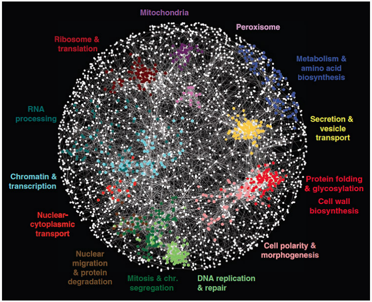
Big-Data



"Big Data" enables "Big Models"

"Big Data" is never IID

Big-Data

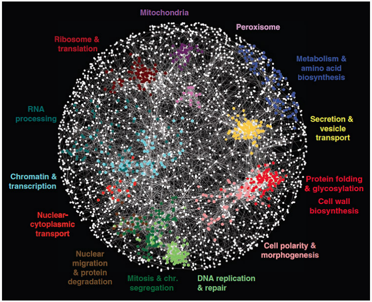


"Big Data" enables "Big Models"

"Big Data" is never IID

"Big Data" **requires** "Big Models"

Big-Data



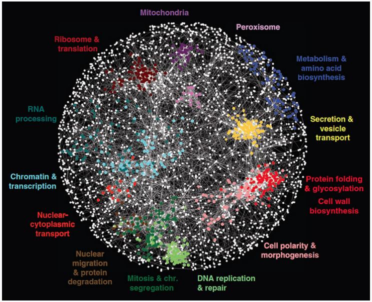
"Big Data" enables "Big Models"

"Big Data" is never IID

"Big Data" **requires** "Big Models"

Highly-structured data requires flexible but powerful models to reflect and capture dependencies in data

Big-Data



"Big Data" enables "Big Models"

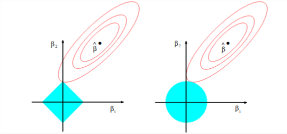
"Big Data" is never IID

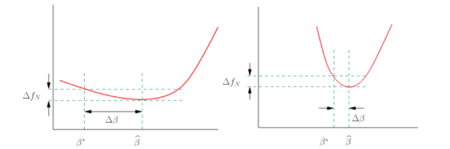
"Big Data" **requires** "Big Models"

Highly-structured data requires flexible but powerful models to reflect and capture dependencies in data

Big data allows us to fit such models

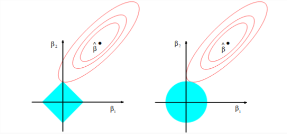
Convex Revolution in Statistical Machine Learning

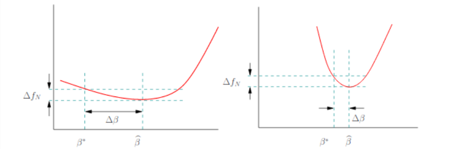




Biggest advance in 21st c. statistics – convex analysis and optimization:

Convex Revolution in Statistical Machine Learning

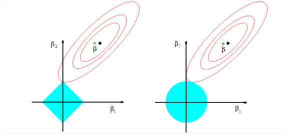


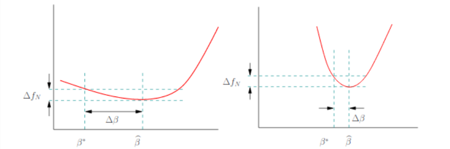


Biggest advance in 21st c. statistics – convex analysis and optimization:

- Development of novel regularized estimation schemes

Convex Revolution in Statistical Machine Learning

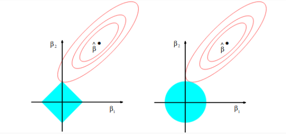


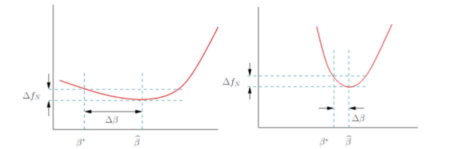


Biggest advance in 21st c. statistics – convex analysis and optimization:

- Development of novel regularized estimation schemes
- Algorithms that efficiently scale to enormous data sets

Convex Revolution in Statistical Machine Learning

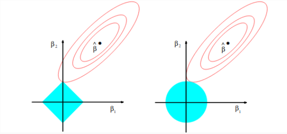


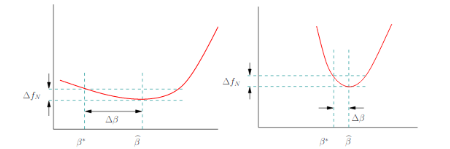


Biggest advance in 21st c. statistics – convex analysis and optimization:

- Development of novel regularized estimation schemes
- Algorithms that efficiently scale to enormous data sets
- Theoretical advances based on powerful convex analysis

Convex Revolution in Statistical Machine Learning





Biggest advance in 21st c. statistics – convex analysis and optimization:

- Development of novel regularized estimation schemes
- Algorithms that efficiently scale to enormous data sets
- Theoretical advances based on powerful convex analysis

Develop Methodology for Big Highly-Structured Data Built on Powerful Convex Analysis and Optimization

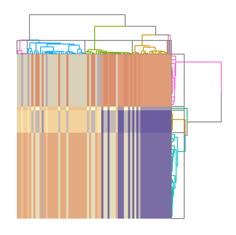
Agenda

Splitting Methods for Clustering

Tensor Co-Clustering

Co-Clustering:

- Simultaneous clustering along all faces of a tensor
- Discover "checkerboard" patterns in data
- "Cluster Heatmap" for 2-tensors
- Manifold learning for K-tensors (Mishne *et al.*, 2019)



Convex Bi-Clustering

Convex formulation of co-clustering: Chi *et al.* (2017) and Chi *et al.* (2018)

- Frobenius norm loss $\implies$ approximate observed data
- Convex fusion penalty $\implies$ encourages clustering

Convex Bi-Clustering

Convex formulation of co-clustering: Chi *et al.* (2017) and Chi *et al.* (2018)

- Frobenius norm loss $\implies$ approximate observed data
- Convex fusion penalty $\implies$ encourages clustering

Matrix (2-tensor) case:

$$\hat{\boldsymbol{U}} = \operatorname*{arg\,min}_{\boldsymbol{U} \in \mathbb{R}^{n \times p}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{n} w_{ij} \|\boldsymbol{U}_{i\cdot} - \boldsymbol{U}_{j\cdot}\|_q + \sum_{\substack{k,l=1 \\ k \neq l}}^{p} \tilde{w}_{kl} \|\boldsymbol{U}_{\cdot k} - \boldsymbol{U}_{\cdot l}\|_q \right)$$

Convex Bi-Clustering

Convex formulation of co-clustering: Chi *et al.* (2017) and Chi *et al.* (2018)

- Frobenius norm loss $\implies$ approximate observed data
- Convex fusion penalty $\implies$ encourages clustering

Matrix (2-tensor) case:

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times p}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{n} w_{ij} \|\boldsymbol{U}_{i\cdot} - \boldsymbol{U}_{j\cdot}\|_q + \sum_{\substack{k,l=1 \\ k \neq l}}^{p} \tilde{w}_{kl} \|\boldsymbol{U}_{\cdot k} - \boldsymbol{U}_{\cdot l}\|_q \right)$$

Simultaneous clustering of rows and columns:

- Rows are clustered together if $\hat{\boldsymbol{U}}_{i\cdot} = \hat{\boldsymbol{U}}_{j\cdot}$
- Columns are clustered together if $\hat{\boldsymbol{U}}_{\cdot k} = \hat{\boldsymbol{U}}_{\cdot l}$
- Each element of $\boldsymbol{X}$ is assigned to a single bi-cluster

Convex Bi-Clustering

Convex formulation of co-clustering: Chi *et al.* (2017) and Chi *et al.* (2018)

- Frobenius norm loss $\implies$ approximate observed data
- Convex fusion penalty $\implies$ encourages clustering

Matrix (2-tensor) case:

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times p}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{n} w_{ij} \|\boldsymbol{U}_{i\cdot} - \boldsymbol{U}_{j\cdot}\|_q + \sum_{\substack{k,l=1 \\ k \neq l}}^{p} \tilde{w}_{kl} \|\boldsymbol{U}_{\cdot k} - \boldsymbol{U}_{\cdot l}\|_q \right)$$

Simultaneous clustering of rows and columns:

- Rows are clustered together if $\hat{\boldsymbol{U}}_{i\cdot} = \hat{\boldsymbol{U}}_{j\cdot}$
- Columns are clustered together if $\hat{\boldsymbol{U}}_{\cdot k} = \hat{\boldsymbol{U}}_{\cdot l}$
- Each element of $\boldsymbol{X}$ is assigned to a single bi-cluster

λ controls the number of co-clusters smoothly

Splitting Methods for Convex Bi-Clustering

Simplified form:

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times p}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\underbrace{\|\boldsymbol{D}_{\text{row}}\boldsymbol{U}\|_{\text{row},q}}_{P_{\text{row}}(\boldsymbol{D}_{\text{row}}\boldsymbol{U})} + \underbrace{\|\boldsymbol{U}\boldsymbol{D}_{\text{col}}\|_{\text{col},q}}_{P_{\text{col}}(\boldsymbol{U}\boldsymbol{D}_{\text{col}})} \right)$$

Splitting Methods for Convex Bi-Clustering

Simplified form:

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times p}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\underbrace{\|\boldsymbol{D}_{\text{row}}\boldsymbol{U}\|_{\text{row},q}}_{P_{\text{row}}(\boldsymbol{D}_{\text{row}}\boldsymbol{U})} + \underbrace{\|\boldsymbol{U}\boldsymbol{D}_{\text{col}}\|_{\text{col},q}}_{P_{\text{col}}(\boldsymbol{U}\boldsymbol{D}_{\text{col}})} \right)$$

Current state of the art:

COBRA - Dykstra-Like Proximal Algorithm

(Bauschke and Combettes, 2008; Chi and Lange, 2015)

- Alternating row- and column-wise convex clustering

Convex clustering subproblems are still slow, so COBRA doesn't scale

Splitting Methods for Convex Bi-Clustering

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U}\in\mathbb{R}^{n\times p}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\underbrace{\|\boldsymbol{D}_{\text{row}}\boldsymbol{U}\|_{\text{row},q}}_{P_{\text{row}}(\boldsymbol{D}_{\text{row}}\boldsymbol{U})} + \underbrace{\|\boldsymbol{U}\boldsymbol{D}_{\text{col}}\|_{\text{col},q}}_{P_{\text{col}}(\boldsymbol{U}\boldsymbol{D}_{\text{col}})} \right)$$

Can we develop a fast splitting approach?

Splitting Methods for Convex Bi-Clustering

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times p}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(\underbrace{\|\boldsymbol{D}_{\text{row}}\boldsymbol{U}\|_{\text{row},q}}_{P_{\text{row}}(\boldsymbol{D}_{\text{row}}\boldsymbol{U})} + \underbrace{\|\boldsymbol{U}\boldsymbol{D}_{\text{col}}\|_{\text{col},q}}_{P_{\text{col}}(\boldsymbol{U}\boldsymbol{D}_{\text{col}})} \right)$$

Can we develop a fast splitting approach?

Davis and Yin (2017) three-block ADMM:

1. $\boldsymbol{U}^{(k+1)} = \boldsymbol{X} - \boldsymbol{D}_{\text{row}}^T \boldsymbol{Z}_{\text{row}}^{(k)} - \boldsymbol{Z}_{\text{col}}^{(k)} \boldsymbol{D}_{\text{col}}^T$

2(a). $\boldsymbol{V}_{\text{row}}^{(k+1)} = \text{prox}_{\lambda/\rho\, P_{\text{row}}(\cdot)}(\boldsymbol{D}_{\text{row}}\boldsymbol{U}^{(k+1)} + \boldsymbol{Z}_{\text{row}}^{(k)})$

2(b). $\boldsymbol{V}_{\text{col}}^{(k+1)} = \text{prox}_{\lambda/\rho\, P_{\text{col}}(\cdot)}(\boldsymbol{U}^{(k+1)}\boldsymbol{D}_{\text{col}} + \boldsymbol{Z}_{\text{col}}^{(k)})$

3(a). $\boldsymbol{Z}_{\text{row}}^{(k+1)} = \boldsymbol{Z}_{\text{row}}^{(k)} + \rho(\boldsymbol{D}_{\text{row}}\boldsymbol{U}^{(k+1)} - \boldsymbol{V}_{\text{row}}^{(k+1)})$

3(b). $\boldsymbol{Z}_{\text{col}}^{(k+1)} = \boldsymbol{Z}_{\text{col}}^{(k)} + \rho(\boldsymbol{U}^{(k+1)}\boldsymbol{D}_{\text{col}} - \boldsymbol{V}_{\text{col}}^{(k+1)})$

Equivalent to AMA and to prox-gradient on the dual - very slow!

(Tseng, 1991)

Splitting Methods for Convex Bi-Clustering

Why not apply ADMM directly?

Why not apply ADMM directly? Lifted problem:

$$\underset{\substack{\boldsymbol{U} \in \mathbb{R}^{n \times p} \\ (\boldsymbol{V}_{\text{row}}, \boldsymbol{V}_{\text{col}}) \in \mathbb{R}^{\#\text{row} \times p} \times \mathbb{R}^{n \times \#\text{col}}}}{\arg\min} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\|_F^2 + \lambda \left(P_{\text{row}}(\boldsymbol{V}_{\text{row}}) + P_{\text{col}}(\boldsymbol{V}_{\text{col}}) \right)$$

subject to

$$\mathfrak{L}_1 \boldsymbol{U} - (\boldsymbol{V}_{\text{row}}, \boldsymbol{V}_{\text{col}}) = 0 \text{ where } \mathfrak{L}_1 \boldsymbol{U} = (\boldsymbol{D}_{\text{row}} \boldsymbol{U}, \boldsymbol{U} \boldsymbol{D}_{\text{col}})$$

Isomorphic, but much better computationally!

$\boldsymbol{V}, \boldsymbol{Z}$ updates as before (separable penalties + Cartesian structure)
$\boldsymbol{U}$ more complicated

Splitting Methods for Convex Bi-Clustering

$\boldsymbol{U}$-subproblem:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times p}}{\arg\min}\ \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{U}\|_F^2+\frac{\rho}{2}\|\boldsymbol{D}_{\text{row}}\boldsymbol{U}-\boldsymbol{V}_{\text{row}}^{(k)}+\rho^{-1}\boldsymbol{Z}_{\text{row}}^{(k)}\|_F^2+\frac{\rho}{2}\|\boldsymbol{U}\boldsymbol{D}_{\text{col}}-\boldsymbol{V}_{\text{col}}^{(k)}+\rho^{-1}\boldsymbol{Z}_{\text{col}}^{(k)}\|_F^2$$

Stationary condition - Sylvester equation:

$$\boldsymbol{X}+\boldsymbol{D}_{\text{row}}^T(\boldsymbol{V}_{\text{row}}^{(k)}-\rho^{-1}\boldsymbol{Z}_{\text{row}}^{(k)})+(\boldsymbol{V}_{\text{col}}^{(k)}-\rho^{-1}\boldsymbol{Z}_{\text{col}}^{(k)})\boldsymbol{D}_{\text{col}}^T=\boldsymbol{U}+\rho\boldsymbol{D}_{\text{row}}^T\boldsymbol{D}_{\text{row}}\boldsymbol{U}+\rho\boldsymbol{U}\boldsymbol{D}_{\text{col}}\boldsymbol{D}_{\text{col}}^T$$

Splitting Methods for Convex Bi-Clustering

$\boldsymbol{U}$-subproblem:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times p}}{\arg\min}\ \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{U}\|_F^2+\frac{\rho}{2}\|\boldsymbol{D}_{\text{row}}\boldsymbol{U}-\boldsymbol{V}_{\text{row}}^{(k)}+\rho^{-1}\boldsymbol{Z}_{\text{row}}^{(k)}\|_F^2+\frac{\rho}{2}\|\boldsymbol{U}\boldsymbol{D}_{\text{col}}-\boldsymbol{V}_{\text{col}}^{(k)}+\rho^{-1}\boldsymbol{Z}_{\text{col}}^{(k)}\|_F^2$$

Stationary condition - Sylvester equation:

$$\boldsymbol{X}+\boldsymbol{D}_{\text{row}}^T(\boldsymbol{V}_{\text{row}}^{(k)}-\rho^{-1}\boldsymbol{Z}_{\text{row}}^{(k)})+(\boldsymbol{V}_{\text{col}}^{(k)}-\rho^{-1}\boldsymbol{Z}_{\text{col}}^{(k)})\boldsymbol{D}_{\text{col}}^T=\boldsymbol{U}+\rho\boldsymbol{D}_{\text{row}}^T\boldsymbol{D}_{\text{row}}\boldsymbol{U}+\rho\boldsymbol{U}\boldsymbol{D}_{\text{col}}\boldsymbol{D}_{\text{col}}^T$$

Alternative: add quadratic term to make $\boldsymbol{U}$-subproblem easier to solve (Deng and Yin, 2016)

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times p}}{\arg\min}\cdots+\alpha\|\boldsymbol{U}\|_F^2-\rho\|\mathfrak{L}_1\boldsymbol{U}\|^2 \text{ where } \mathfrak{L}_1\boldsymbol{U}=(\boldsymbol{D}_{\text{row}}\boldsymbol{U},\boldsymbol{U}\boldsymbol{D}_{\text{col}})$$

$$\begin{aligned}\boldsymbol{U}^{(k+1)}=\Big(&\alpha\boldsymbol{U}^{(k)}+\boldsymbol{X}+\rho\boldsymbol{D}_{\text{row}}^T(\boldsymbol{V}^{(k)}-\rho^{-1}\boldsymbol{Z}_{\text{row}}^{(k)}-\boldsymbol{D}_{\text{row}}\boldsymbol{U}^{(k)})\\&+\rho(\boldsymbol{V}_{\text{col}}^{(k)}-\rho^{-1}\boldsymbol{Z}_{\text{col}}^{(k)}-\boldsymbol{U}^{(k)}\boldsymbol{D}_{\text{col}})\boldsymbol{D}_{\text{col}}^T\Big)/(1+\alpha)\end{aligned}$$

Results

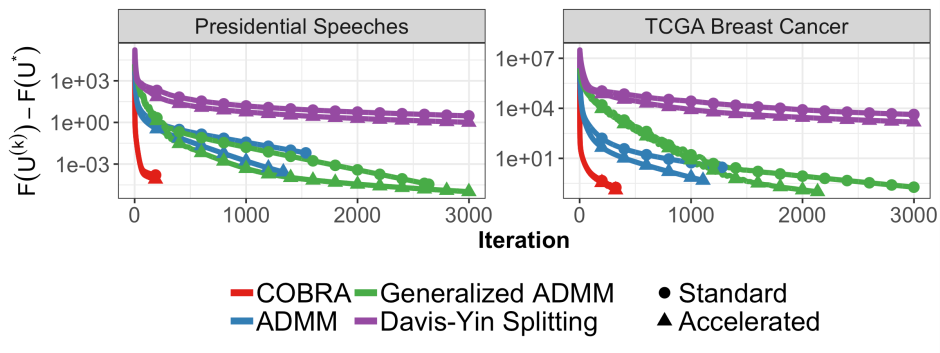
Compare:

- ADMM
- Generalized ADMM
- Davis-Yin Three-Block ADMM
- COBRA $\implies$ alternating row- and column-clustering sub-problems

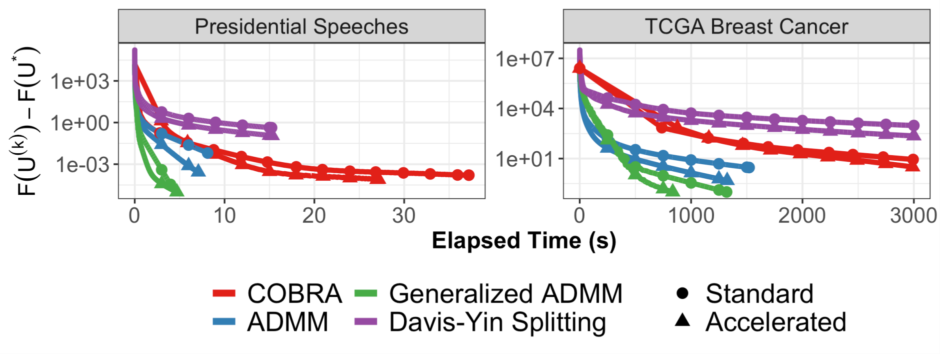
Data:

- Presidents $\in \mathbb{R}^{44 \times 75}$
- TCGA Breast Cancer $\in \mathbb{R}^{438 \times 353}$

Results: Iteration Count



Results: Elapsed Time



Higher-Order Extensions

$$\hat{\mathcal{U}} = \arg\min_{\mathcal{U}} \frac{1}{2}\|\mathcal{X} - \mathcal{U}\|_F^2 + \lambda \sum_{j=1}^{J} \|\mathcal{U} \times_j \mathcal{D}_j\|_{j,q}$$

Higher-Order Extensions

$$\hat{\mathcal{U}} = \arg\min_{\mathcal{U}} \frac{1}{2}\|\mathcal{X} - \mathcal{U}\|_F^2 + \lambda \sum_{j=1}^{J} \|\mathcal{U} \times_j \mathcal{D}_j\|_{j,q}$$

Same “lifting” approach works for Generalized ADMM and Davis-Yin:

$$\mathcal{U}^{(k+1)}_{\text{Gen-ADMM}} = \frac{\alpha}{1+\alpha}\mathcal{U}^{(k)} + \frac{\mathcal{X}}{1+\alpha} + \frac{\rho}{1+\alpha}\sum_{j=1}^{J}(\mathcal{V}_j^{(k)} - \rho^{-1}\mathcal{Z}_j^{(k)} - \mathcal{U}^{(k)} \times_j \mathcal{D}_j) \times_j \mathcal{D}_j^{\mathcal{T}}$$

$$\mathcal{U}^{(k+1)}_{\text{DY / AMA}} = \mathcal{X} - \sum_{j=1}^{J} \mathcal{Z}_j^{(k)} \times_j (\mathcal{D}_j)^{\mathcal{T}}$$

$$\mathcal{V}_j^{(k+1)} = \operatorname*{prox}_{\lambda/\rho\|\cdot\|_{j,q}}\left(\mathcal{U}^{(k+1)} \times_j \mathcal{D}_j + \rho^{-1}\mathcal{Z}_j^{(k)}\right) \quad \forall j \in \{1, \dots, J\}$$

$$\mathcal{Z}_j^{(k+1)} = \mathcal{Z}_j^{(k)} + \rho(\mathcal{U}^{(k+1)} \times_j \mathcal{D}_j - \mathcal{V}_j^{(k+1)}) \quad \forall j \in \{1, \dots, J\}$$

Higher-Order Extensions

$$\hat{\mathcal{U}} = \arg\min_{\mathcal{U}} \frac{1}{2}\|\mathcal{X} - \mathcal{U}\|_F^2 + \lambda \sum_{j=1}^{J} \|\mathcal{U} \times_j \mathcal{D}_j\|_{j,q}$$

Same "lifting" approach works for Generalized ADMM and Davis-Yin:

$$\mathcal{U}^{(k+1)}_{\text{Gen-ADMM}} = \frac{\alpha}{1+\alpha}\mathcal{U}^{(k)} + \frac{\mathcal{X}}{1+\alpha} + \frac{\rho}{1+\alpha}\sum_{j=1}^{J}(\mathcal{V}_j^{(k)} - \rho^{-1}\mathcal{Z}_j^{(k)} - \mathcal{U}^{(k)} \times_j \mathcal{D}_j) \times_j \mathcal{D}_j^{\mathcal{T}}$$

$$\mathcal{U}^{(k+1)}_{\text{DY / AMA}} = \mathcal{X} - \sum_{j=1}^{J} \mathcal{Z}_j^{(k)} \times_j (\mathcal{D}_j)^{\mathcal{T}}$$

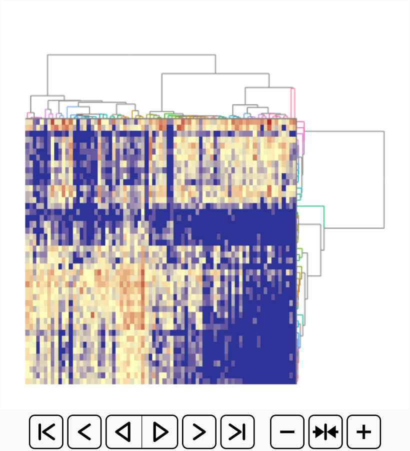
$$\mathcal{V}_j^{(k+1)} = \operatorname*{prox}_{\lambda/\rho\|\cdot\|_{j,q}}\left(\mathcal{U}^{(k+1)} \times_j \mathcal{D}_j + \rho^{-1}\mathcal{Z}_j^{(k)}\right) \quad \forall j \in \{1, \dots, J\}$$

$$\mathcal{Z}_j^{(k+1)} = \mathcal{Z}_j^{(k)} + \rho(\mathcal{U}^{(k+1)} \times_j \mathcal{D}_j - \mathcal{V}_j^{(k+1)}) \quad \forall j \in \{1, \dots, J\}$$

Standard ADMM $\Longrightarrow$ *tensor Sylvester equation*:

$$\mathcal{X} + \rho\sum_{j=1}^{J}(\mathcal{V}_j^{(k)} - \rho^{-1}\mathcal{Z}_j^{(k)}) \times_j \mathcal{D}_j^{\mathcal{T}} = \mathcal{U}_{\text{ADMM}} + \rho\sum_{j=1}^{J}\mathcal{U}_{\text{ADMM}} \times_j \mathcal{D}_j \times_j \mathcal{D}_j^{\mathcal{T}}.$$

Implications



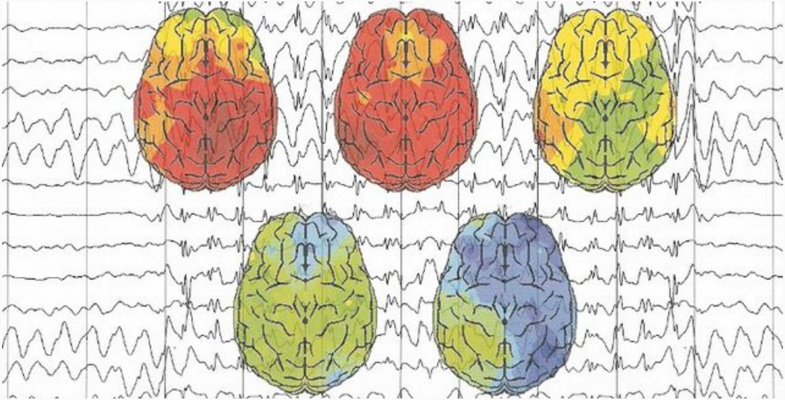
Efficient Convex Clustering Algorithm

Embed in More Complex Schemes (*e.g.*, 2D trend filtering)

"Lifting" Trick Useful for Multiply-Regularized Problems

Multi-Rank Regularized PCA

Motivation



Principal Components Analysis:

- Exploratory Data Analysis
- Pattern Recognition
- Dimension Reduction
- Data Visualization

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Advantages:

- Identify patterns in rows and columns of $\boldsymbol{X}$
- $\boldsymbol{u}$, $\boldsymbol{v}$, d calculated using SVD of $\boldsymbol{X}$

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Advantages:

- Identify patterns in rows and columns of $\boldsymbol{X}$
- $\boldsymbol{u}$, $\boldsymbol{v}$, d calculated using SVD of $\boldsymbol{X}$

PCA is consistent under standard ($n \to \infty$) asymptotics (Anderson, 1963)

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Advantages:

- Identify patterns in rows and columns of $\boldsymbol{X}$
- $\boldsymbol{u}, \boldsymbol{v}, d$ calculated using SVD of $\boldsymbol{X}$

PCA is consistent under standard ($n \to \infty$) asymptotics (Anderson, 1963)

Convergence is slow: RMT asymptotics ($p/n \to c$) more relevant

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Advantages:

- Identify patterns in rows and columns of $\boldsymbol{X}$
- $\boldsymbol{u}, \boldsymbol{v}, d$ calculated using SVD of $\boldsymbol{X}$

PCA is consistent under standard ($n \to \infty$) asymptotics (Anderson, 1963)

Convergence is slow: RMT asymptotics ($p/n \to c$) more relevant

PCA in high-dimensions is inconsistent (Johnstone and Lu, 2009)

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Advantages:

- Identify patterns in rows and columns of $\boldsymbol{X}$
- $\boldsymbol{u}, \boldsymbol{v}, d$ calculated using SVD of $\boldsymbol{X}$

PCA is consistent under standard ($n \to \infty$) asymptotics (Anderson, 1963)

Convergence is slow: RMT asymptotics ($p/n \to c$) more relevant

PCA in high-dimensions is inconsistent (Johnstone and Lu, 2009)

Low-rank model is *always* high-dimensional along one axis

Regularization in PCA

Low-rank model for PCA - estimate low-rank mean of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{E} \text{ where } \boldsymbol{E} \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\arg\min_{\boldsymbol{u},\boldsymbol{v},d} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2 \Leftrightarrow \arg\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$$

Advantages:

- Identify patterns in rows and columns of $\boldsymbol{X}$
- $\boldsymbol{u}, \boldsymbol{v}, d$ calculated using SVD of $\boldsymbol{X}$

PCA is consistent under standard ($n \to \infty$) asymptotics (Anderson, 1963)

Convergence is slow: RMT asymptotics ($p/n \to c$) more relevant

PCA in high-dimensions is inconsistent (Johnstone and Lu, 2009)

Low-rank model is *always* high-dimensional along one axis

Regularization Needed

Sparse and Functional PCA

Sparse and Functional PCA: Allen and W., (DSW 2019)

$$\underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S_u}}, \boldsymbol{v} \in \overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$$

where

$$\overline{\mathbb{B}}^n_{\boldsymbol{S_u}} = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}^T \boldsymbol{S_u} \boldsymbol{x} = \boldsymbol{x}^T (\boldsymbol{I} + \alpha_{\boldsymbol{u}} \boldsymbol{\Omega_u}) \boldsymbol{x} \leq 1 \right\}$$

Sparse and Functional PCA

Sparse and Functional PCA: Allen and W., (DSW 2019)

$$\underset{\boldsymbol{u}\in\overline{\mathbb{B}}^n_{\boldsymbol{S_u}},\boldsymbol{v}\in\overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max}\ \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} - \lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v})$$

where

$$\overline{\mathbb{B}}^n_{\boldsymbol{S_u}} = \left\{\boldsymbol{x}\in\mathbb{R}^n : \boldsymbol{x}^T\boldsymbol{S_u}\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{I}+\alpha_{\boldsymbol{u}}\boldsymbol{\Omega_u})\boldsymbol{x}\leq 1\right\}$$

Sparse and Functional PCA:

- Smoothness in $\boldsymbol{u}$ - structure $\boldsymbol{S_u}$ + strength $\alpha_{\boldsymbol{u}}$
- Sparsity in $\boldsymbol{u}$ - structure $P_{\boldsymbol{u}}$ + strength $\lambda_{\boldsymbol{u}}$
- Smoothness in $\boldsymbol{v}$ - structure $\boldsymbol{S_v}$ + strength $\alpha_{\boldsymbol{v}}$
- Sparsity in $\boldsymbol{v}$ - structure $P_{\boldsymbol{v}}$ + strength $\lambda_{\boldsymbol{v}}$

Sparse and Functional PCA

Sparse and Functional PCA: Allen and W., (DSW 2019)

$$\underset{\boldsymbol{u}\in\overline{\mathbb{B}}^n_{\boldsymbol{S_u}},\boldsymbol{v}\in\overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max}\ \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} - \lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v})$$

where

$$\overline{\mathbb{B}}^n_{\boldsymbol{S_u}} = \left\{\boldsymbol{x}\in\mathbb{R}^n : \boldsymbol{x}^T\boldsymbol{S_u}\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{I}+\alpha_{\boldsymbol{u}}\boldsymbol{\Omega_u})\boldsymbol{x}\leq 1\right\}$$

Sparse and Functional PCA:

- Smoothness in $\boldsymbol{u}$ - structure $\boldsymbol{S_u}$ + strength $\alpha_{\boldsymbol{u}}$
- Sparsity in $\boldsymbol{u}$ - structure $P_{\boldsymbol{u}}$ + strength $\lambda_{\boldsymbol{u}}$
- Smoothness in $\boldsymbol{v}$ - structure $\boldsymbol{S_v}$ + strength $\alpha_{\boldsymbol{v}}$
- Sparsity in $\boldsymbol{v}$ - structure $P_{\boldsymbol{v}}$ + strength $\lambda_{\boldsymbol{v}}$

Well-posed and non-degenerate

Computing SFPCA

$$\text{SFPCA:} \quad \underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S_u}}, \boldsymbol{v} \in \overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$$

Bi-concave in $\boldsymbol{u}$ and $\boldsymbol{v}$ $\implies$ alternating maximization strategy

Projection + (accelerated) proximal gradient to solve sub-problems

Computing SFPCA

$$\text{SFPCA:} \quad \underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S_u}}, \boldsymbol{v} \in \overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$$

Bi-concave in $\boldsymbol{u}$ and $\boldsymbol{v}$ $\implies$ alternating maximization strategy

Projection + (accelerated) proximal gradient to solve sub-problems

Theorem

1. The $\boldsymbol{u}$-update converges to a solution of

$$\underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S_u}}}{\arg\min} \; \frac{1}{2} \|\boldsymbol{X}\boldsymbol{v} - \boldsymbol{u}\|_2^2 + \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) + \frac{\alpha_{\boldsymbol{u}}}{2} \boldsymbol{u}^T \boldsymbol{\Omega}_{\boldsymbol{u}} \boldsymbol{u}$$

2. $\boldsymbol{u}$-update finds global optimum for fixed $\boldsymbol{v}$
3. Converges to block-coordinate-wise global optima (Nash points)

Fast!

Multi-Rank SFPCA

Orthogonality of PCs: interpretation and statistical independence

Can we do the same for SFPCA?

Multi-Rank SFPCA

Orthogonality of PCs: interpretation and statistical independence

Can we do the same for SFPCA?

Multi-rank extension of SFPCA:

$$\text{MR-SFPCA:} \quad \underset{\boldsymbol{U} \in \mathcal{V}_{\boldsymbol{S_u}}^{n \times k}, \boldsymbol{V} \in \mathcal{V}_{\boldsymbol{S_v}}^{p \times k}}{\arg\max} \operatorname{Tr}(\boldsymbol{U}^T \boldsymbol{X} \boldsymbol{V}) - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{U}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{V})$$

where $\mathcal{V}_{\boldsymbol{S_u}}^{n \times k}$ is the k$^{\text{th}}$ order generalized Stiefel manifold in $\mathbb{R}^n$:

$$\boldsymbol{U} \in \mathcal{V}_{\boldsymbol{S_u}}^{n \times k} \Leftrightarrow \boldsymbol{U}^T \boldsymbol{S_u} \boldsymbol{U} = \boldsymbol{I}_k$$

Multi-Rank SFPCA

Orthogonality of PCs: interpretation and statistical independence

Can we do the same for SFPCA?

Multi-rank extension of SFPCA:

$$\text{MR-SFPCA:} \quad \underset{\boldsymbol{U} \in \mathcal{V}_{\boldsymbol{S_u}}^{n \times k}, \boldsymbol{V} \in \mathcal{V}_{\boldsymbol{S_v}}^{p \times k}}{\arg\max} \operatorname{Tr}(\boldsymbol{U}^T \boldsymbol{X} \boldsymbol{V}) - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{U}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{V})$$

where $\mathcal{V}_{\boldsymbol{S_u}}^{n \times k}$ is the k$^{\text{th}}$ order generalized Stiefel manifold in $\mathbb{R}^n$:

$$\boldsymbol{U} \in \mathcal{V}_{\boldsymbol{S_u}}^{n \times k} \Leftrightarrow \boldsymbol{U}^T \boldsymbol{S_u} \boldsymbol{U} = \boldsymbol{I}_k$$

Generalized Stiefel manifold constraint $\implies$ manifold optimization (Absil *et al.*, 2007)

As with R1-SFPCA, alternating (partial) maximization

Manifold Proximal Gradient

Standard SFPCA $\boldsymbol{u}$-subproblem updates:

$$\boldsymbol{u} := \underset{\frac{\lambda_{\boldsymbol{u}}}{L_{\boldsymbol{u}}} P_{\boldsymbol{u}}(\cdot)}{\text{prox}} \left(\boldsymbol{u} + L_{\boldsymbol{u}}^{-1} \left(\boldsymbol{X}\hat{\boldsymbol{v}} - \boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}\right)\right) \qquad \hat{\boldsymbol{u}} := \begin{cases} \boldsymbol{u} & \|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} \leq 1 \\ \boldsymbol{u}/\|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} & \text{otherwise} \end{cases}$$

Proximal + projected gradient descent

Manifold Proximal Gradient

Standard SFPCA $\boldsymbol{u}$-subproblem updates:

$$\boldsymbol{u} := \underset{\frac{\lambda_{\boldsymbol{u}}}{L_{\boldsymbol{u}}} P_{\boldsymbol{u}}(\cdot)}{\text{prox}} \left(\boldsymbol{u} + L_{\boldsymbol{u}}^{-1} \left(\boldsymbol{X}\hat{\boldsymbol{v}} - \boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}\right)\right) \qquad \hat{\boldsymbol{u}} := \begin{cases} \boldsymbol{u} & \|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} \leq 1 \\ \boldsymbol{u}/\|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} & \text{otherwise} \end{cases}$$

Proximal + projected gradient descent

Multi-Rank SFPCA $\boldsymbol{U}$-subproblem updates - Manifold Prox Gradient: (Chen *et al.*, 2020a)

$$\hat{\boldsymbol{D}} = \underset{\boldsymbol{D} \in \mathbb{R}^{n \times k}}{\arg\min} -\langle \boldsymbol{X}\hat{\boldsymbol{V}}, \boldsymbol{D} \rangle_F + \lambda_{\boldsymbol{U}} P_{\boldsymbol{U}}(\boldsymbol{U}^{(k)} + \boldsymbol{D})$$
$$\text{s.t. } \boldsymbol{D}^T \boldsymbol{S}_{\boldsymbol{u}} \boldsymbol{U}^{(k)} + (\boldsymbol{U}^{(k)})^T \boldsymbol{S}_{\boldsymbol{u}} \boldsymbol{D} = \boldsymbol{0}$$
$$\boldsymbol{U}^{(k+1)} = \text{Retr}_{\boldsymbol{U}^{(k)}}(\eta \hat{\boldsymbol{D}})$$

Manifold Proximal Gradient

Standard SFPCA $\boldsymbol{u}$-subproblem updates:

$$\boldsymbol{u} := \underset{\frac{\lambda_{\boldsymbol{u}}}{L_{\boldsymbol{u}}} P_{\boldsymbol{u}}(\cdot)}{\text{prox}} \left(\boldsymbol{u} + L_{\boldsymbol{u}}^{-1}\left(\boldsymbol{X}\hat{\boldsymbol{v}} - \boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}\right)\right) \qquad \hat{\boldsymbol{u}} := \begin{cases} \boldsymbol{u} & \|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} \leq 1 \\ \boldsymbol{u}/\|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} & \text{otherwise} \end{cases}$$

Proximal + projected gradient descent

Multi-Rank SFPCA $\boldsymbol{U}$-subproblem updates - Manifold Prox Gradient: (Chen *et al.*, 2020a)

$$\hat{\boldsymbol{D}} = \underset{\boldsymbol{D} \in \mathbb{R}^{n \times k}}{\arg\min} -\langle \boldsymbol{X}\hat{\boldsymbol{V}}, \boldsymbol{D}\rangle_F + \lambda_{\boldsymbol{U}} P_{\boldsymbol{U}}(\boldsymbol{U}^{(k)} + \boldsymbol{D})$$

$$\text{s.t. } \boldsymbol{D}^T \boldsymbol{S}_{\boldsymbol{u}} \boldsymbol{U}^{(k)} + (\boldsymbol{U}^{(k)})^T \boldsymbol{S}_{\boldsymbol{u}} \boldsymbol{D} = \boldsymbol{0}$$

$$\boldsymbol{U}^{(k+1)} = \text{Retr}_{\boldsymbol{U}^{(k)}}(\eta \hat{\boldsymbol{D}})$$

One step of each subproblem $\implies$ convergence to stationary point:

- Constraint set smooth $\implies$ Stationary points isolated
- Guaranteed descent at each iteration (Chen *et al.*, 2020b)

Manifold ADMM

Easier $\boldsymbol{U}$-updates from Manifold ADMM: (Kovnatsky *et al.*, 2016)

$$\begin{aligned}
\boldsymbol{U}^{(k+1)} &= \underset{\boldsymbol{U} \in \mathcal{V}_{n \times k}^{\boldsymbol{S_u}}}{\arg\min} - \text{Tr}(\boldsymbol{U}^T \boldsymbol{X} \boldsymbol{V}) + \frac{\rho}{2} \| \boldsymbol{U} - \boldsymbol{W}^{(k)} + \boldsymbol{Z}^{(k)} \|_F^2 \\
\boldsymbol{W}^{(k+1)} &= \underset{\boldsymbol{W} \in \mathbb{R}^{n \times k}}{\arg\min} \lambda_{\boldsymbol{U}} P_{\boldsymbol{U}}(\boldsymbol{W}) + \frac{\rho}{2} \| \boldsymbol{U}^{(k+1)} - \boldsymbol{W} + \boldsymbol{Z}^{(k)} \|_F^2 \\
&= \underset{\lambda_{\boldsymbol{U}}/\rho P_{\boldsymbol{U}}(\cdot)}{\text{prox}} \left(\boldsymbol{U}^{(k+1)} + \boldsymbol{Z}^{(k)} \right) \\
\boldsymbol{Z}^{(k+1)} &= \boldsymbol{Z}^{(k)} + \boldsymbol{U}^{(k+1)} - \boldsymbol{W}^{(k+1)}
\end{aligned}$$

Manifold ADMM

Easier $\boldsymbol{U}$-updates from Manifold ADMM: (Kovnatsky *et al.*, 2016)

$$\begin{aligned}
\boldsymbol{U}^{(k+1)} &= \underset{\boldsymbol{U} \in \mathcal{V}_{n \times k}^{\boldsymbol{S_u}}}{\arg\min} - \text{Tr}(\boldsymbol{U}^T \boldsymbol{X} \boldsymbol{V}) + \frac{\rho}{2} \|\boldsymbol{U} - \boldsymbol{W}^{(k)} + \boldsymbol{Z}^{(k)}\|_F^2 \\
\boldsymbol{W}^{(k+1)} &= \underset{\boldsymbol{W} \in \mathbb{R}^{n \times k}}{\arg\min} \lambda_{\boldsymbol{U}} P_{\boldsymbol{U}}(\boldsymbol{W}) + \frac{\rho}{2} \|\boldsymbol{U}^{(k+1)} - \boldsymbol{W} + \boldsymbol{Z}^{(k)}\|_F^2 \\
&= \underset{\lambda_{\boldsymbol{U}}/\rho P_{\boldsymbol{U}}(\cdot)}{\text{prox}} \left(\boldsymbol{U}^{(k+1)} + \boldsymbol{Z}^{(k)} \right) \\
\boldsymbol{Z}^{(k+1)} &= \boldsymbol{Z}^{(k)} + \boldsymbol{U}^{(k+1)} - \boldsymbol{W}^{(k+1)}
\end{aligned}$$

First step is a *generalized unbalanced Procrustes problem* - analytical solution via SVD of $\boldsymbol{S_u}^{-1/2} \boldsymbol{X} \hat{\boldsymbol{V}} + \rho \boldsymbol{S_u}^{1/2} (\boldsymbol{W}^{(k)} - \boldsymbol{Z}^{(k)})$

Typically converges quickly and to a good solution (no theory)

Deflation Techniques

$$\text{Rank-1 SFPCA:} \quad \underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S_u}}, \boldsymbol{v} \in \overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$$

How to get additional *nested* SFPCA components?

Deflation Techniques

$$\text{Rank-1 SFPCA:} \quad \underset{\boldsymbol{u}\in\overline{\mathbb{B}}^{n}_{\boldsymbol{S_u}},\boldsymbol{v}\in\overline{\mathbb{B}}^{p}_{\boldsymbol{S_v}}}{\arg\max}\ \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} - \lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v})$$

How to get additional *nested* SFPCA components?

Deflation:

- Hotelling: $\boldsymbol{X}_t^{\text{HD}} := \boldsymbol{X}_{t-1} - \boldsymbol{U}_t(\boldsymbol{U}_t^T\boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^T\boldsymbol{X}_{t-1}\boldsymbol{V}_t(\boldsymbol{V}_t^T\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^T$
- Projection: $\boldsymbol{X}_t^{\text{PD}} := (\boldsymbol{I}_n - \boldsymbol{U}_t(\boldsymbol{U}_t^T\boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^T)\boldsymbol{X}_{t-1}(\boldsymbol{I}_p - \boldsymbol{V}_t(\boldsymbol{V}_t^T\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^T)$
- Schur Complement: $\boldsymbol{X}_t^{\text{SD}} := \boldsymbol{X}_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t(\boldsymbol{U}_t^T\boldsymbol{X}_{t-1}\boldsymbol{V}_t)^{-1}\boldsymbol{U}_t^T\boldsymbol{X}_{t-1}$

HD doesn't fully remove signal and may re-introduce if estimated PCs non-orthogonal

Deflation Techniques

$$\text{Rank-1 SFPCA:} \quad \underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S_u}}, \boldsymbol{v} \in \overline{\mathbb{B}}^p_{\boldsymbol{S_v}}}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$$

How to get additional *nested* SFPCA components?

Deflation:

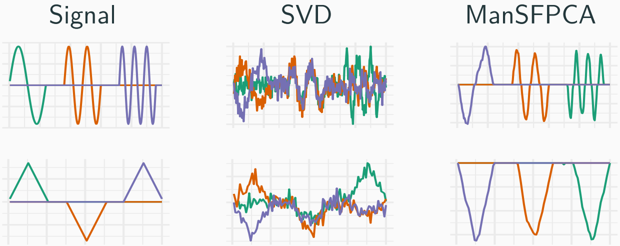
- Hotelling: $\boldsymbol{X}_t^{\text{HD}} := \boldsymbol{X}_{t-1} - \boldsymbol{U}_t(\boldsymbol{U}_t^T\boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^T\boldsymbol{X}_{t-1}\boldsymbol{V}_t(\boldsymbol{V}_t^T\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^T$
- Projection: $\boldsymbol{X}_t^{\text{PD}} := (\boldsymbol{I}_n - \boldsymbol{U}_t(\boldsymbol{U}_t^T\boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^T)\boldsymbol{X}_{t-1}(\boldsymbol{I}_p - \boldsymbol{V}_t(\boldsymbol{V}_t^T\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^T)$
- Schur Complement: $\boldsymbol{X}_t^{\text{SD}} := \boldsymbol{X}_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t(\boldsymbol{U}_t^T\boldsymbol{X}_{t-1}\boldsymbol{V}_t)^{-1}\boldsymbol{U}_t^T\boldsymbol{X}_{t-1}$

HD doesn't fully remove signal and may re-introduce if estimated PCs non-orthogonal

Method	Two-Way 0-ing $\boldsymbol{u}_t^T\boldsymbol{X}_t\boldsymbol{v}_t = 0$	One-Way 0-ing $\boldsymbol{u}_t^T\boldsymbol{X}_t, \boldsymbol{X}_t\boldsymbol{v}_t = \boldsymbol{0}$	Subsequent 0-ing ($\forall s \geq 0$) $\boldsymbol{u}_t^T\boldsymbol{X}_{t+s}, \boldsymbol{X}_{t+s}\boldsymbol{v}_t = \boldsymbol{0}$	Robust to Scale of $\boldsymbol{u}_t, \boldsymbol{v}_t$
Hotelling	✓	✗	✗	✗
Projection	✓	✓	✗	✗
Schur	✓	✓	✓	✓

Simulation: "On Model" Signal Recovery

Scenario 1: $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ Orthogonal - SNR $\approx$ 1.2

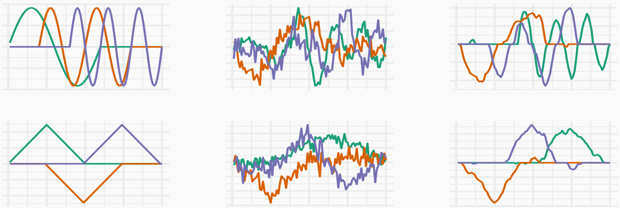


		HD	PD	SD	ManSFPCA
	PC1	15.92%	21.05%	**21.87%**	
CPVE	PC2	22.21%	29.42%	**30.59%**	**37.12%**
	PC3	26.80%	35.57%	**37.09%**	
rSS-Error	$\boldsymbol{U}$	129.54%	129.55%	128.35%	**68.66%**
	$\boldsymbol{V}$	143.01%	143.72%	141.15%	**36.98%**

Simulation: ‘Off-Model” Signal Recovery

Scenario 2: $\boldsymbol{U}^*$ *and* $\boldsymbol{V}^*$ ***Not*** *Orthogonal - SNR* $\approx$ *1.7*

Signal SVD ManSFPCA



		HD	PD	SD	ManSFPCA
	PC1	8.85%	19.74%	**29.80%**	
CPVE	PC2	13.03%	28.30%	**39.87%**	**50.85%**
	PC3	16.16%	34.22%	**46.48%**	
rSS-Error	$\boldsymbol{U}$	215.73%	206.30%	205.74%	**97.77%**
	$\boldsymbol{V}$	211.15%	207.77%	204.38%	**78.26%**

Implications

Principaled Approach to Multi-Rank PCA:

- Unifies **many** regularized PCA variants
- Extension to multiple PCs without loosing orthogonality
- Deflation techniques applicable to all PCA techniques

Implications

Principaled Approach to Multi-Rank PCA:

- Unifies **many** regularized PCA variants
- Extension to multiple PCs without loosing orthogonality
- Deflation techniques applicable to all PCA techniques

Tensor extensions:

- Rank-1 SFPCA yields regularized CP
- MR-SFPCA yields regularized Tucker

Implications

Principaled Approach to Multi-Rank PCA:

- Unifies **many** regularized PCA variants
- Extension to multiple PCs without loosing orthogonality
- Deflation techniques applicable to all PCA techniques

Tensor extensions:

- Rank-1 SFPCA yields regularized CP
- MR-SFPCA yields regularized Tucker

Additional Multivariate Methods:

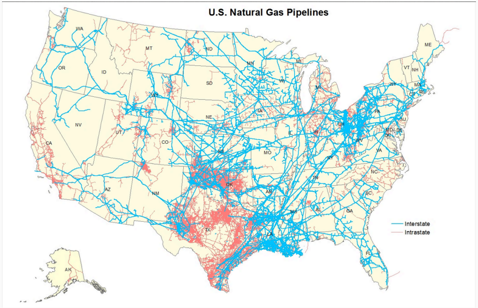
- CCA, LDA, PLS, *etc.* all SVD - can all be similarly treated

Multivariate Models for Gas Markets

Natural Gas Markets

LNG Markets:

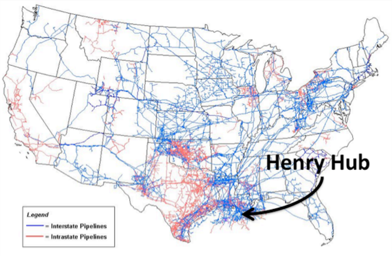
- 32% of all US Electricity (1.273 PWh in 2017)
- 3M Miles of NG Pipelines
- 150+ Trading Spots



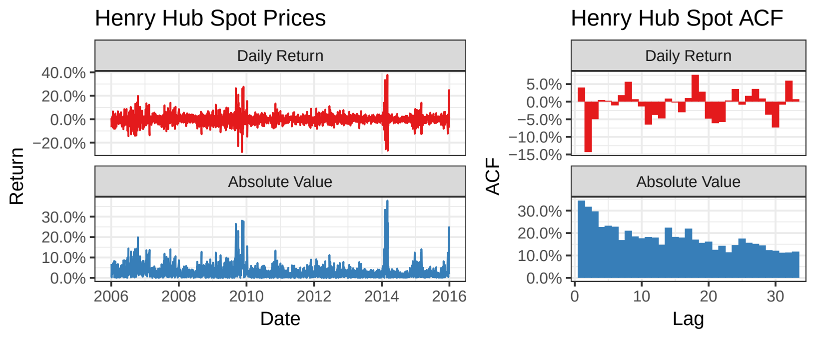
Natural Gas Markets

Henry Hub:

- $14B+ Futures Volume Daily
- Common Proxy for Domestic NG Markets Broadly

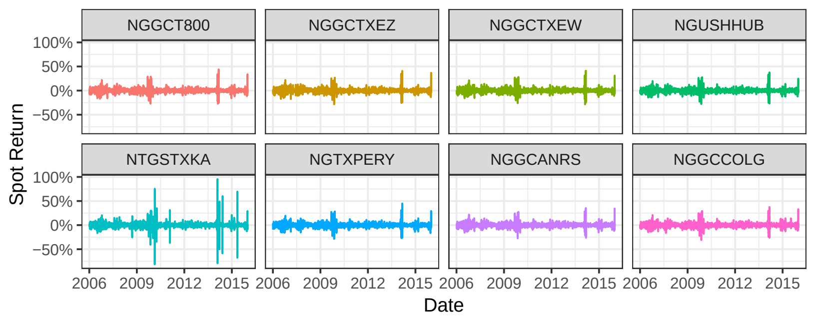


Natural Gas Markets



Equity-like dynamics: vol clustering, heavy-tails, 2nd moment autocorrelation

Natural Gas Markets



High inter-spot correlation: PC1 (74%) suggests single-factor model

Natural Gas Markets

We want to capture:

- High-Dimensional Multivariate Time Series
- Irregular Data Availability
    - NG Futures Priced Near-Continuously on Lit Markets
    - NG Spots Traded Over-the-Counter
- GARCH Type Behavior + Single-Factor Structure

Natural Gas Markets

We want to capture:

- High-Dimensional Multivariate Time Series
- Irregular Data Availability
    - NG Futures Priced Near-Continuously on Lit Markets
    - NG Spots Traded Over-the-Counter
- GARCH Type Behavior + Single-Factor Structure

Realized Beta GARCH Model (Hansen *et al.*, 2014) combining:

- Intra-Day Futures Realized Volatility
- End-of-Day Spot Volatility
- 2^{nd} Moment Single-Factor Dynamics

Bayesian Realized Beta GARCH

- Bi-Variate GARCH Model (Multivariate Skew Normal Specification):
  - "Realized" (High-Frequency) Volatility: improved estimate of σ_t^2
  - "Beta" Volatility linkage: Volatility at Henry $\implies$ volatility in spots

Bayesian Realized Beta GARCH

- Bi-Variate GARCH Model (Multivariate Skew Normal Specification):
  - "Realized" (High-Frequency) Volatility: improved estimate of σ_t^2
  - "Beta" Volatility linkage: Volatility at Henry $\Longrightarrow$ volatility in spots
- Bayesian Estimation:
  - Priors calibrated to S&P 500 (equity) markets: improve estimation
  - Coherent uncertainty propagation
  - Improved out of sample forecasts

Does it work?

Fitting strategy:

- Fit to 250 window: refit every 50 days
- One-day rolling predictions for out-of-sample

Does it work?

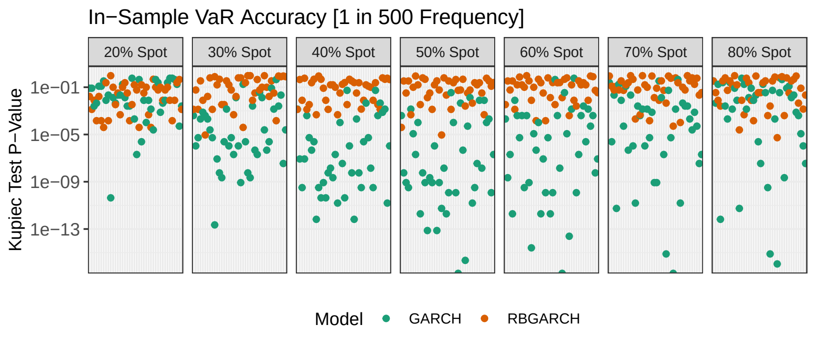
Fitting strategy:

- Fit to 250 window: refit every 50 days
- One-day rolling predictions for out-of-sample

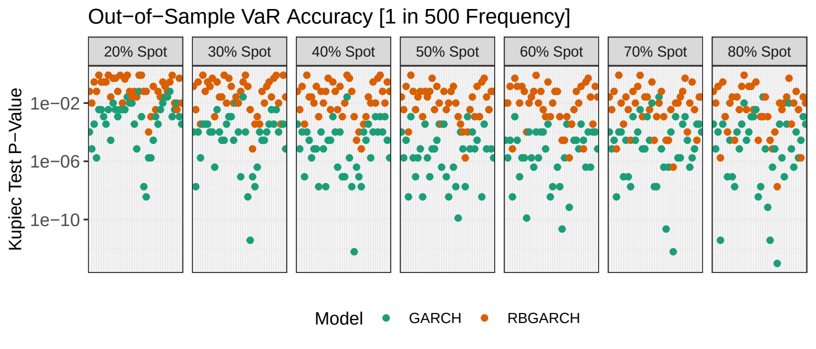
Measures of Fit:

- In-sample VaR Test (Kupiec, 1995)
  - Binomial test for number of VaR exceedances
- Out-of-sample VaR Test (Kupiec, 1995)
  - Estimated out-of-sample log-likelihood

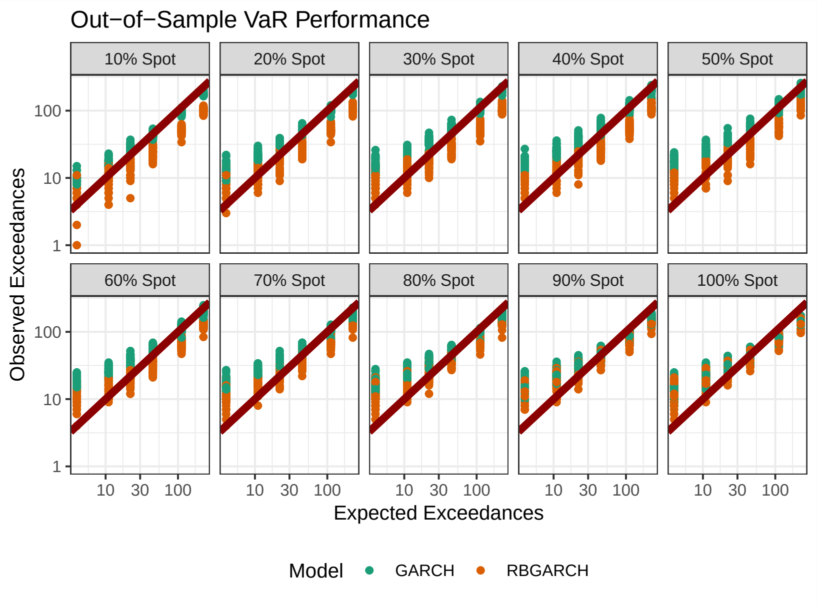
Application to Tail Forecasting



Application to Tail Forecasting

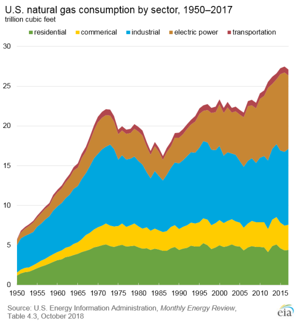


Application to Tail Forecasting



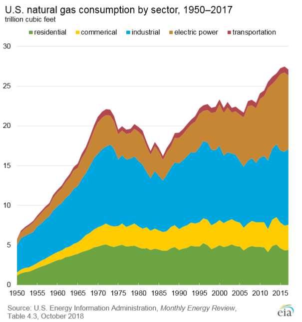
Implications

- Multivariate and Multi-Resolution Model for NG Volatility
    - Daily and intra-day volatility measures
    - Multivariate Treatment of 50+ NG Trading Hubs



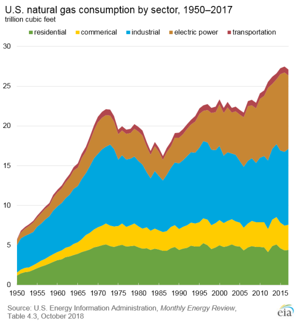
Implications

- Multivariate and Multi-Resolution Model for NG Volatility
  - Daily and intra-day volatility measures
  - Multivariate Treatment of 50+ NG Trading Hubs
- Bayesian Approach
  - Market Calibrated Priors



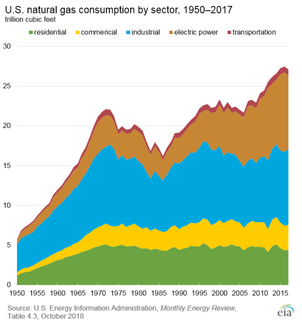
Implications

- Multivariate and Multi-Resolution Model for NG Volatility
    - Daily and intra-day volatility measures
    - Multivariate Treatment of 50+ NG Trading Hubs
- Bayesian Approach
    - Market Calibrated Priors
- Improved Out-of-Sample Prediction
    - VaR Estimates: more accurate + more conservative



Implications

- Multivariate and Multi-Resolution Model for NG Volatility
    - Daily and intra-day volatility measures
    - Multivariate Treatment of 50+ NG Trading Hubs
- Bayesian Approach
    - Market Calibrated Priors
- Improved Out-of-Sample Prediction
    - VaR Estimates: more accurate + more conservative
- Amenable to all commodities markets with irregular data availability



Complex Convex Analysis

Motivation

Complex-data arise in many domains:

- Signal and radar processing (Schreier and Scharf, 2010; Candès *et al.*, 2015; Mechlenbrauker *et al.*, 2017)
- Neuroscience (Yu *et al.*, 2018; Adrian *et al.*, 2018)
- Geostatistics (de Iaco *et al.*, 2003; Mandic *et al.*, 2009)
- Astronomy (Zechmeister and Kürster, 2009)
- Econometrics (Granger and Engle, 1983)

Motivation

Complex-data arise in many domains:

- Signal and radar processing (Schreier and Scharf, 2010; Candès *et al.*, 2015; Mechlenbrauker *et al.*, 2017)
- Neuroscience (Yu *et al.*, 2018; Adrian *et al.*, 2018)
- Geostatistics (de Iaco *et al.*, 2003; Mandic *et al.*, 2009)
- Astronomy (Zechmeister and Kürster, 2009)
- Econometrics (Granger and Engle, 1983)

Major sources:

- Spectral analysis (Fourier transforms)
- 2D directional data

Why Complex?

Why not treat complex data as real?

Why Complex?

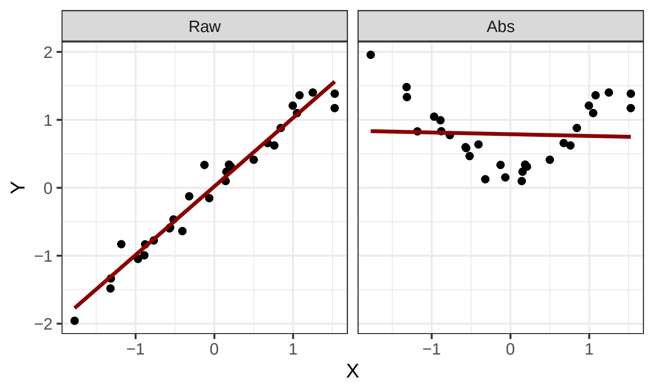
Why not treat complex data as real?

- Discards phase data (Canolty *et al.*, 2006)

Why Complex?

Why not treat complex data as real?

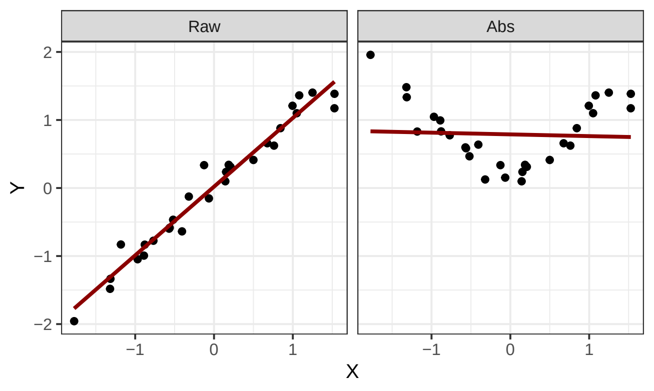
- Discards phase data (Canolty *et al.*, 2006)
- Corrupts statistical relationships:



Why Complex?

Why not treat complex data as real?

- Discards phase data (Canolty *et al.*, 2006)
- Corrupts statistical relationships:



- Natural domain for "spectral" phenomena

Proper and Improper RVs

Let Z be a *univariate* complex random variable:

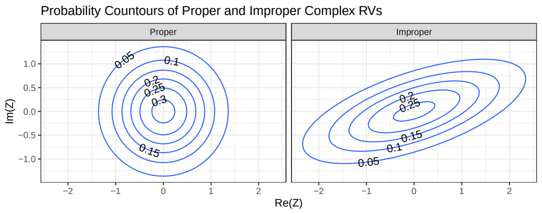
- Secretly "multivariate": correlation between $\Re(Z)$ and $\Im(Z)$

Proper and Improper RVs

Let Z be a *univariate* complex random variable:

- Secretly "multivariate": correlation between $\Re(Z)$ and $\Im(Z)$

Important case: *proper* (circular) RVs are those where $\Re(Z) \perp\!\!\!\perp \Im(Z)$

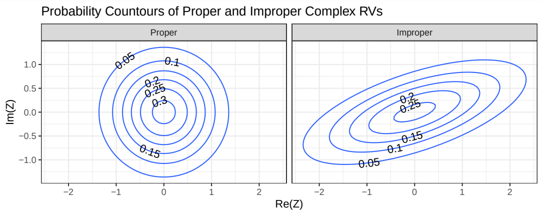


Proper and Improper RVs

Let Z be a *univariate* complex random variable:

- Secretly "multivariate": correlation between $\Re(Z)$ and $\Im(Z)$

Important case: *proper* (circular) RVs are those where $\Re(Z) \perp\!\!\!\perp \Im(Z)$



Complex variables often arise as Fourier transform of *stationary* processes

- Only *relative* phase of multivariate signal matters
- Absolute phase is meaningless

$$\text{Law}[Z] = \text{Law}[e^{i\theta} Z] \implies Z \text{ proper}$$

Complex Gaussian Distribution

The *complex* Gaussian is a **three** parameter distribution:

- Mean: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Z}]$
- Covariance: $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^H] = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})\overline{(\boldsymbol{Z} - \boldsymbol{\mu})}^T]$
- Pseudo-Covariance: $\boldsymbol{\Gamma} = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^T]$

Complex Gaussian Distribution

The *complex* Gaussian is a **three** parameter distribution:

- Mean: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Z}]$
- Covariance: $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^H] = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})\overline{(\boldsymbol{Z} - \boldsymbol{\mu})}^T]$
- Pseudo-Covariance: $\boldsymbol{\Gamma} = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^T]$

Requirements:

- $\boldsymbol{\Sigma}$ is positive-definite: $\Sigma_{ij} = \mathbb{E}[z_i \overline{z_j}]$ - non-negative when $i = j$
- $\boldsymbol{\Gamma}$ is indefinite: $\Gamma_{ij} = \mathbb{E}[z_i z_j]$ - possibly negative when $i = j$
- Additionally $\overline{\boldsymbol{\Sigma}} - \boldsymbol{\Gamma}^H \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma} \succeq 0$

Complex Gaussian Distribution

The *complex* Gaussian is a **three** parameter distribution:

- Mean: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Z}]$
- Covariance: $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^H] = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})\overline{(\boldsymbol{Z} - \boldsymbol{\mu})}^T]$
- Pseudo-Covariance: $\boldsymbol{\Gamma} = \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^T]$

Requirements:

- $\boldsymbol{\Sigma}$ is positive-definite: $\Sigma_{ij} = \mathbb{E}[z_i \overline{z_j}]$ - non-negative when $i = j$
- $\boldsymbol{\Gamma}$ is indefinite: $\Gamma_{ij} = \mathbb{E}[z_i z_j]$ - possibly negative when $i = j$
- Additionally $\overline{\boldsymbol{\Sigma}} - \boldsymbol{\Gamma}^H \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma} \succeq 0$

Proper complex normal $(\boldsymbol{\Gamma} = \mathbf{0})$ almost universally assumed in statistics (Wooding, 1956; Goodman, 1963; Graczyk *et al.*, 2003)

Non-stationary DSP sometimes uses general case (van den Bos, 1995; Schreier and Scharf, 2010; Adali *et al.*, 2011)

Complex Convex Analysis: Subgradient Analysis

Penalized *M*-Estimation Paradigm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathscr{L}(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda \mathscr{P}(\boldsymbol{\beta})$$

What if $\boldsymbol{X}, \boldsymbol{y}, \beta$ complex?

Complex Convex Analysis: Subgradient Analysis

Penalized *M*-Estimation Paradigm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathscr{L}(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda \mathscr{P}(\boldsymbol{\beta})$$

What if $\boldsymbol{X}, \boldsymbol{y}, \beta$ complex? Well defined if $\mathscr{L}, \mathscr{P} \to \mathbb{R}$ (*e.g.* norms!)

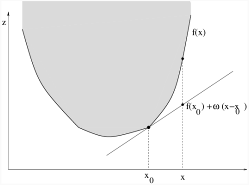
Complex Convex Analysis: Subgradient Analysis

Penalized *M*-Estimation Paradigm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathscr{L}(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda \mathscr{P}(\boldsymbol{\beta})$$

What if $\boldsymbol{X}, \boldsymbol{y}, \beta$ complex? Well defined if $\mathscr{L}, \mathscr{P} \to \mathbb{R}$ (*e.g.* norms!)

Analysis: first order sub-gradient conditions



γ is a sub-gradient of f at x if

$$f(y) \geq f(x) + \gamma(y - x) \quad \text{for all } y$$

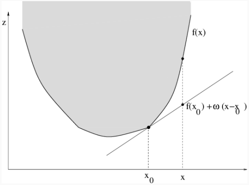
Complex Convex Analysis: Subgradient Analysis

Penalized *M*-Estimation Paradigm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathscr{L}(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda \mathscr{P}(\boldsymbol{\beta})$$

What if $\boldsymbol{X}, \boldsymbol{y}, \beta$ complex? Well defined if $\mathscr{L}, \mathscr{P} \to \mathbb{R}$ (*e.g.* norms!)

Analysis: first order sub-gradient conditions



γ is a sub-gradient of f at x if

$$f(y) \geq f(x) + \gamma(y - x) \quad \text{for all } y$$

If x, y are *complex*, this is ill-defined!

The Wirtinger Fix

Wirtinger (1927) studied differentiability of functions $\mathbb{C} \to \mathbb{R}$

Never Cauchy-Riemann differentiable (holomorphic) $\implies$ traditional complex analysis does not apply

The Wirtinger Fix

Wirtinger (1927) studied differentiability of functions $\mathbb{C} \to \mathbb{R}$

Never Cauchy-Riemann differentiable (holomorphic) $\implies$ traditional complex analysis does not apply

Wirtinger's idea: write $f(z) = f(z, \bar{z})$ and differentiate with respect to z while holding $\bar{z}$ fixed

The Wirtinger Fix

Wirtinger (1927) studied differentiability of functions $\mathbb{C} \to \mathbb{R}$

Never Cauchy-Riemann differentiable (holomorphic) $\implies$ traditional complex analysis does not apply

Wirtinger's idea: write $f(z) = f(z, \bar{z})$ and differentiate with respect to z while holding $\bar{z}$ fixed

Occasionally used as $\mathbb{CR}$-calculus (Kreutz-Delgado, 2009)

The Wirtinger Fix

Wirtinger (1927) studied differentiability of functions $\mathbb{C} \to \mathbb{R}$

Never Cauchy-Riemann differentiable (holomorphic) $\implies$ traditional complex analysis does not apply

Wirtinger's idea: write $f(z) = f(z, \bar{z})$ and differentiate with respect to z while holding $\bar{z}$ fixed

Occasionally used as $\mathbb{CR}$-calculus (Kreutz-Delgado, 2009)

GOAL: rigorously define this derivative and connect it to optimization

The Wirtinger Fix

$\mathbb{C}^p$ is an *inner product space* over a field $\mathbb{F}$:

- Can add vectors (elements of $\mathbb{C}^p$) and multiply by $\mathbb{F}$
- Inner product $\langle \cdot, \cdot \rangle : \mathbb{C}^p \times \mathbb{C}^p \to \mathbb{F}$

Typically $\mathbb{F} = \mathbb{C}$

The Wirtinger Fix

$\mathbb{C}^p$ is an *inner product space* over a field $\mathbb{F}$:

- Can add vectors (elements of $\mathbb{C}^p$) and multiply by $\mathbb{F}$
- Inner product $\langle \cdot, \cdot \rangle : \mathbb{C}^p \times \mathbb{C}^p \to \mathbb{F}$

Typically $\mathbb{F} = \mathbb{C}$

In this work, $\mathbb{F} = \mathbb{R}$!

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \frac{\boldsymbol{a}^H \boldsymbol{b} + \boldsymbol{a}^T \overline{\boldsymbol{b}}}{2}$$

Sub-gradient inequality becomes

$$f(\boldsymbol{w}) \geq f(\boldsymbol{z}) + \langle \gamma, \boldsymbol{w} - \boldsymbol{z} \rangle \quad \text{for all } \boldsymbol{w} \in \mathbb{C}^p$$

All terms real $\implies$ well-defined!

The Wirtinger Fix

Is this valid?

The Wirtinger Fix

Is this valid?

Identity and existence of minimizer are *topological* properties

$$\langle \boldsymbol{a}, \boldsymbol{a} \rangle = \sum_{i=1}^{p} |a_i|^2$$

- Real inner product $\implies$ the same norm
- Same norm $\implies$ same topology
- Same topology $\implies$ minimizer unchanged

The Wirtinger Fix

Is this valid?

Identity and existence of minimizer are *topological* properties

$$\langle \boldsymbol{a}, \boldsymbol{a} \rangle = \sum_{i=1}^{p} |a_i|^2$$

- Real inner product $\implies$ the same norm
- Same norm $\implies$ same topology
- Same topology $\implies$ minimizer unchanged

We have freedom to change *algebraic* structure used to analyze problem

Likelihood still defined with "regular" complex multiplication

The Wirtinger Fix

Convex Analysis for Wirtinger ($\mathbb{C}^p \to \mathbb{R}$) Functions:

The Wirtinger Fix

Convex Analysis for Wirtinger ($\mathbb{C}^p \to \mathbb{R}$) Functions:

- **W. Theorem 2.2**: If Wirtinger f is convex, it has sub-gradients.

The Wirtinger Fix

Convex Analysis for Wirtinger ($\mathbb{C}^p \to \mathbb{R}$) Functions:

- **W. Theorem 2.2**: If Wirtinger f is convex, it has sub-gradients.
- **W. Theorem 2.6**: Sub-gradients given by the Wirtinger (formal) derivative.

The Wirtinger Fix

Convex Analysis for Wirtinger ($\mathbb{C}^p \to \mathbb{R}$) Functions:

- **W. Theorem 2.2**: If Wirtinger f is convex, it has sub-gradients.
- **W. Theorem 2.6**: Sub-gradients given by the Wirtinger (formal) derivative.
- **W. Theorem 2.1**: If the Wirtinger derivative is 0, global minimum.

The Wirtinger Fix

Convex Analysis for Wirtinger ($\mathbb{C}^p \to \mathbb{R}$) Functions:

- **W. Theorem 2.2**: If Wirtinger f is convex, it has sub-gradients.
- **W. Theorem 2.6**: Sub-gradients given by the Wirtinger (formal) derivative.
- **W. Theorem 2.1**: If the Wirtinger derivative is 0, global minimum.
- **W. Theorem 2.8**: Key result for sparse models:

$$\partial |z| = \begin{cases} \angle z = z/|z| & z \neq 0 \\ \{w \in \mathbb{C} : |w| \leq 1\} & z = 0 \end{cases}$$

The Wirtinger Fix

Convex Analysis for Wirtinger ($\mathbb{C}^p \to \mathbb{R}$) Functions:

- **W. Theorem 2.2**: If Wirtinger f is convex, it has sub-gradients.
- **W. Theorem 2.6**: Sub-gradients given by the Wirtinger (formal) derivative.
- **W. Theorem 2.1**: If the Wirtinger derivative is 0, global minimum.
- **W. Theorem 2.8**: Key result for sparse models:

$$\partial|z| = \begin{cases} \angle z = z/|z| & z \neq 0 \\ \{w \in \mathbb{C} : |w| \leq 1\} & z = 0 \end{cases}$$

Change definition of "multiplication"
convex analysis still "works!"

Example: Complex OLS

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{C}^p} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$$

Set Wirtinger derivative to $\mathbf{0}$:

$$\begin{aligned} f &= \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\beta)^H(\boldsymbol{y} - \boldsymbol{X}\beta) \\ &= (\boldsymbol{y}^H - \overline{\beta}\boldsymbol{X}^H)(\boldsymbol{y} - \boldsymbol{X}\beta) \\ \mathbf{0} = \frac{\partial f}{\partial \boldsymbol{\beta}} &= -\boldsymbol{y}^H\boldsymbol{X} + \overline{\boldsymbol{\beta}}^T\boldsymbol{X}^H\boldsymbol{X} \\ \implies \overline{\boldsymbol{\beta}} &= (\boldsymbol{X}^H\boldsymbol{X})^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}} \\ \boldsymbol{\beta} &= (\boldsymbol{X}^H\boldsymbol{X})^{-1}\boldsymbol{X}^H\boldsymbol{y} \end{aligned}$$

Intuition from real-domain translates to complex-domain!

Concentration Inequalities: Regression Noise (W., Lemma 3.9)

Suppose Z is mean-zero sub-Gaussian with variance proxy σ^2:

(a) If Z is real:

$$P[|Z| \geq t] \leq 2\exp\{-t^2/2\sigma^2\}$$

Concentration Inequalities: Regression Noise (W., Lemma 3.9)

Suppose Z is mean-zero sub-Gaussian with variance proxy σ^2:

(a) If Z is real:

$$P[|Z| \geq t] \leq 2 \exp\{-t^2/2\sigma^2\}$$

(b) If Z is proper and complex:

$$P[|Z| \geq t] \leq 4 \exp\{-t^2/2\sigma^2\}$$

Proper and complex Z concentrates like real 2-vector.

Concentration Inequalities: Regression Noise (W., Lemma 3.9)

Suppose Z is mean-zero sub-Gaussian with variance proxy σ^2:

(a) If Z is real:

$$P[|Z| \geq t] \leq 2 \exp\{-t^2/2\sigma^2\}$$

(b) If Z is proper and complex:

$$P[|Z| \geq t] \leq 4 \exp\{-t^2/2\sigma^2\}$$

(c) If Z is complex:

$$P[|Z| \geq t] \leq 2 \exp\{-t^2/16\sigma^2\}$$

Proper and complex Z concentrates like real 2-vector.

Penalty for unknown dependence in $\Re(Z)$ and $\Im(Z)$

Concentration Inequalities: Regression Noise (W., Lemma 3.9)

Suppose Z is mean-zero sub-Gaussian with variance proxy σ^2:

(a) If Z is real:

$$P[|Z| \geq t] \leq 2\exp\{-t^2/2\sigma^2\}$$

(b) If Z is proper and complex:

$$P[|Z| \geq t] \leq 4\exp\{-t^2/2\sigma^2\}$$

(c) If Z is complex:

$$P[|Z| \geq t] \leq 2\exp\{-t^2/16\sigma^2\}$$

Proper and complex Z concentrates like real 2-vector.

Penalty for unknown dependence in $\Re(Z)$ and $\Im(Z)$

Similar rates for effective noise $\|\boldsymbol{X}^H\boldsymbol{\epsilon}\|_\infty$ depending on $\boldsymbol{\epsilon}$

Concentration Inequalities: Sample Covariance (W., Lemma 4.3)

Suppose $\boldsymbol{Z}$ is mean-zero Gaussian vector with variance $\boldsymbol{\Sigma}^*$:
$[\sigma^2 = \max(\Sigma^*_{ii})]$

(a) If Z is real:

$$P[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max} \geq t\sigma^2] \leq 3p^2 e^{-nt^2/8}$$

Concentration Inequalities: Sample Covariance (W., Lemma 4.3)

Suppose $\boldsymbol{Z}$ is mean-zero Gaussian vector with variance $\boldsymbol{\Sigma}^*$:
$[\sigma^2 = \max(\Sigma^*_{ii})]$

(a) If Z is real:

$$P[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max} \geq t\sigma^2] \leq 3p^2 e^{-nt^2/8}$$

(b) If Z is proper and complex :

$$P[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max} \geq t\sigma^2] \leq 3p^2 e^{-nt^2/4}$$

Better concentration for proper complex $\boldsymbol{Z}$ than real $\boldsymbol{Z}$!

Intuition: $\Re(\boldsymbol{Z}) \perp\!\!\!\perp \Im(\boldsymbol{Z}) \implies$ double sample size

Concentration Inequalities: Sample Covariance (W., Lemma 4.3)

Suppose $\boldsymbol{Z}$ is mean-zero Gaussian vector with variance $\boldsymbol{\Sigma}^*$:
$[\sigma^2 = \max(\Sigma^*_{ii})]$

(a) If Z is real:

$$P[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max} \geq t\sigma^2] \leq 3p^2 e^{-nt^2/8}$$

(b) If Z is proper and complex :

$$P[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max} \geq t\sigma^2] \leq 3p^2 e^{-nt^2/4}$$

(c) If Z is complex:

$$P[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\| \geq t\sigma^2] \leq 3p^2 e^{-nt^2/64}$$

Better concentration for proper complex $\boldsymbol{Z}$ than real $\boldsymbol{Z}$!

Intuition: $\Re(\boldsymbol{Z}) \perp\!\!\!\perp \Im(\boldsymbol{Z}) \implies$ double sample size

The Complex Lasso

Real Lasso: (Tibshirani, 1996; Chen *et al.*, 1998)

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

ℓ_1-norm penalty $\implies$ $\hat{\boldsymbol{\beta}}$ will be *sparse* (have exact zeros):

- Compressed Sensing: can estimate $\boldsymbol{\beta}^*$ well even with $p \ll n$ elements
- Automatic Variable Selection: can guess the exact zeros in $\boldsymbol{\beta}^*$ so long as $\boldsymbol{X}$ is not "too correlated"

The Complex Lasso

Real $\textsc{Lasso}$: (Tibshirani, 1996; Chen *et al.*, 1998)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

ℓ_1-norm penalty $\implies$ $\hat{\boldsymbol{\beta}}$ will be *sparse* (have exact zeros):

- Compressed Sensing: can estimate $\boldsymbol{\beta}^*$ well even with $p \ll n$ elements
- Automatic Variable Selection: can guess the exact zeros in $\boldsymbol{\beta}^*$ so long as $\boldsymbol{X}$ is not "too correlated"

Rich theoretical literature Fu and Knight (2000), Greenshtein and Ritov (2004), Zhao and Yu (2006), Bickel *et al.* (2009), Zhang and Huang (2008), Bunea *et al.* (2007), Meinshausen and Yu (2009), and van de Geer and Bühlmann (2009) *etc.*

Results all translate to the complex lasso ($\textsc{CLasso}$)!

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{C}^p} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

The Complex Lasso

W., Theorem 3.3: Under standard assumptions (Wainwright, 2009), the Complex-Lasso ($\textsc{CLasso}$) is model selection consistent with probability

(a) $\geq 1 - 2\exp\{-(\tau - 2)/2\log(p - s)\}$ for real ϵ (real $\boldsymbol{X}, \boldsymbol{y}$)

$$\epsilon \overset{\text{IID}}{\sim} \text{subG}(0, \sigma^2) \quad \lambda_{\min}(\boldsymbol{X}_S^H \boldsymbol{X}_S / n) \geq c > 0 \quad \max_{j \in S^c} \|(\boldsymbol{X}_S^H \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^H \boldsymbol{X}_j\|_1 \leq 1 - \gamma$$

The Complex Lasso

W., Theorem 3.3: Under standard assumptions (Wainwright, 2009), the Complex-Lasso ($\textsc{CLasso}$) is model selection consistent with probability

(a) $\geq 1 - 2\exp\{-(\tau-2)/2\log(p-s)\}$ for real ϵ (real $\boldsymbol{X}, \boldsymbol{y}$)

(b) $\geq 1 - 4\exp\{-(\tau-2)/2\log(p-s)\}$ for real ϵ (complex $\boldsymbol{X}, \boldsymbol{y}$)

$$\epsilon \overset{\text{IID}}{\sim} \text{subG}(0, \sigma^2) \quad \lambda_{\text{min}}(\boldsymbol{X}_S^H \boldsymbol{X}_S/n) \geq c > 0 \quad \max_{j \in S^c} \|(\boldsymbol{X}_S^H \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^H \boldsymbol{X}_j\|_1 \leq 1 - \gamma$$

The Complex Lasso

W., Theorem 3.3: Under standard assumptions (Wainwright, 2009), the Complex-Lasso ($\textsc{CLasso}$) is model selection consistent with probability

(a) $\geq 1 - 2\exp\{-(\tau - 2)/2 \log(p - s)\}$ for real $\boldsymbol{\epsilon}$ (real $\boldsymbol{X}, \boldsymbol{y}$)
(b) $\geq 1 - 4\exp\{-(\tau - 2)/2 \log(p - s)\}$ for real $\boldsymbol{\epsilon}$ (complex $\boldsymbol{X}, \boldsymbol{y}$)
(c) $\geq 1 - 4\exp\{-(\tau - 16)/16 \log(p - s)\}$ for complex $\boldsymbol{\epsilon}$

$$\boldsymbol{\epsilon} \overset{\text{IID}}{\sim} \text{subG}(0, \sigma^2) \quad \lambda_{\min}(\boldsymbol{X}_S^H \boldsymbol{X}_S / n) \geq c > 0 \quad \max_{j \in S^c} \|(\boldsymbol{X}_S^H \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^H \boldsymbol{X}_j\|_1 \leq 1 - \gamma$$

W., Theorem 3.3: Under standard assumptions (Wainwright, 2009), the Complex-Lasso ($\textsc{CLasso}$) is model selection consistent with probability

(a) $\geq 1 - 2\exp\{-(\tau-2)/2\log(p-s)\}$ for real $\boldsymbol{\epsilon}$ (real $\boldsymbol{X}, \boldsymbol{y}$)
(b) $\geq 1 - 4\exp\{-(\tau-2)/2\log(p-s)\}$ for real $\boldsymbol{\epsilon}$ (complex $\boldsymbol{X}, \boldsymbol{y}$)
(c) $\geq 1 - 4\exp\{-(\tau-16)/16\log(p-s)\}$ for complex $\boldsymbol{\epsilon}$
(d) $\geq 1 - 8\exp\{-(\tau-2)/2\log(p-s)\}$ for proper complex $\boldsymbol{\epsilon}$

$$\boldsymbol{\epsilon} \overset{\text{IID}}{\sim} \text{subG}(0, \sigma^2) \quad \lambda_{\min}(\boldsymbol{X}_S^H \boldsymbol{X}_S/n) \geq c > 0 \quad \max_{j \in S^c} \|(\boldsymbol{X}_S^H \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^H \boldsymbol{X}_j\|_1 \leq 1-\gamma$$

The Complex Lasso

W., Theorem 3.3: Under standard assumptions (Wainwright, 2009), the Complex-Lasso ($\textsc{CLasso}$) is model selection consistent with probability

(a) $\geq 1 - 2\exp\{-(\tau - 2)/2\log(p - s)\}$ for real ϵ (real $\boldsymbol{X}, \boldsymbol{y}$)

(b) $\geq 1 - 4\exp\{-(\tau - 2)/2\log(p - s)\}$ for real ϵ (complex $\boldsymbol{X}, \boldsymbol{y}$)

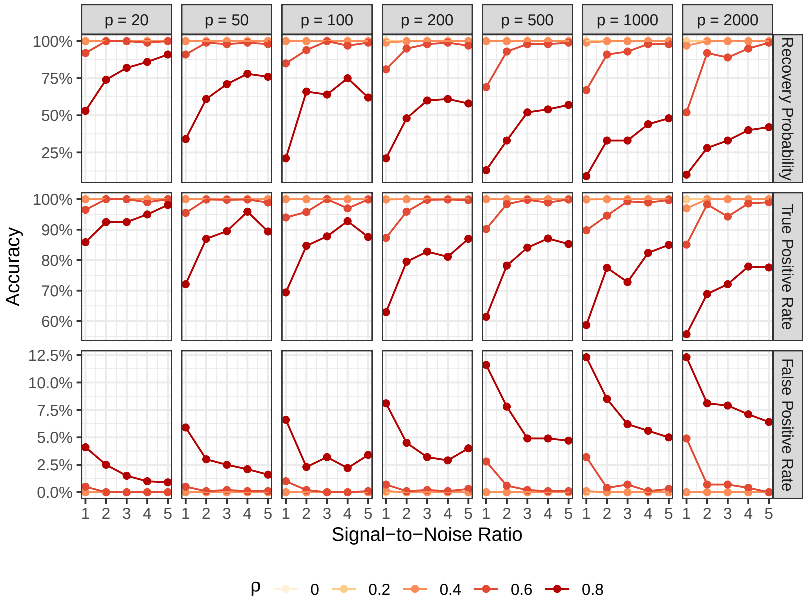
(c) $\geq 1 - 4\exp\{-(\tau - 16)/16\log(p - s)\}$ for complex ϵ

(d) $\geq 1 - 8\exp\{-(\tau - 2)/2\log(p - s)\}$ for proper complex ϵ

$$\epsilon \overset{\text{IID}}{\sim} \text{subG}(0, \sigma^2) \quad \lambda_{\min}(\boldsymbol{X}_S^H\boldsymbol{X}_S/n) \geq c > 0 \quad \max_{j \in S^c} \|(\boldsymbol{X}_S^H\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^H\boldsymbol{X}_j\|_1 \leq 1-\gamma$$

First precise finite sample results for $\textsc{CLasso}$: previously studied by Yang and Zhang (2011), Maleki *et al.* (2013), and Mechlenbrauker *et al.* (2017)

Model Selection Consistency of CLasso



The Complex Graphical Lasso

Suppose $\boldsymbol{Z}$ is drawn from a p-variate complex Gaussian with precision matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\Sigma}^*)^{-1}$. $\textsc{CGLasso}$ gives a sparse estimate of $\boldsymbol{\Theta}^*$:

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta} \in \mathbb{C}^{p \times p}_{\succeq 0}} - \log\det \boldsymbol{\Theta} + \text{Tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) + \lambda \|\boldsymbol{\Theta}\|_{1,\text{off-diag}}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance (Yuan and Lin, 2007; Friedman *et al.*, 2008; Ravikumar *et al.*, 2011).

The Complex Graphical Lasso

Suppose $\boldsymbol{Z}$ is drawn from a p-variate complex Gaussian with precision matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\Sigma}^*)^{-1}$. $\text{CGL}\textsc{asso}$ gives a sparse estimate of $\boldsymbol{\Theta}^*$:

$$\hat{\boldsymbol{\Theta}} = \operatorname*{arg\,min}_{\boldsymbol{\Theta} \in \mathbb{C}^{p \times p}_{\succeq 0}} -\log\det\boldsymbol{\Theta} + \text{Tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) + \lambda \|\boldsymbol{\Theta}\|_{1,\text{off-diag}}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance (Yuan and Lin, 2007; Friedman *et al.*, 2008; Ravikumar *et al.*, 2011).

W. Theorem 4.1: If $\boldsymbol{Z}$ is a proper complex Gaussian satisfying standard assumptions, then the $\text{CGL}\textsc{asso}$ recovers the true sparsity pattern of $\boldsymbol{\Theta}^*$ with probability

$$\geq 1 - 3p^2 \exp\left\{-\frac{\log p}{64\sigma^2} - \frac{\delta n}{256\sigma^2}\right\}$$

The Complex Graphical Lasso

Suppose $\boldsymbol{Z}$ is drawn from a p-variate complex Gaussian with precision matrix $\mathbf{\Theta}^* = (\mathbf{\Sigma}^*)^{-1}$. $\text{CGL}\textsc{asso}$ gives a sparse estimate of $\mathbf{\Theta}^*$:

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta} \in \mathbb{C}^{p \times p}_{\succeq 0}}{\arg\min} - \log\det\mathbf{\Theta} + \mathrm{Tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) + \lambda\|\mathbf{\Theta}\|_{1,\text{off-diag}}$$

where $\hat{\mathbf{\Sigma}}$ is the sample covariance (Yuan and Lin, 2007; Friedman *et al.*, 2008; Ravikumar *et al.*, 2011).

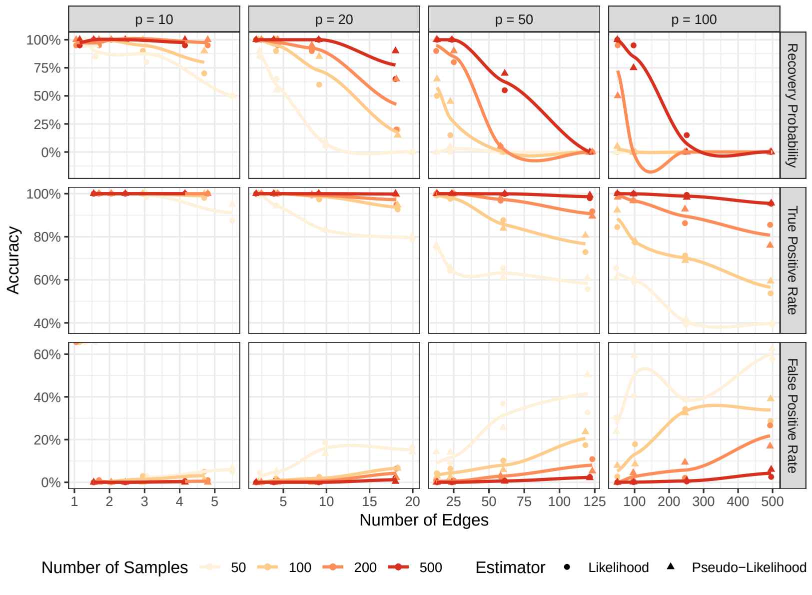
W. Theorem 4.1: If $\boldsymbol{Z}$ is a proper complex Gaussian satisfying standard assumptions, then the $\text{CGL}\textsc{asso}$ recovers the true sparsity pattern of $\mathbf{\Theta}^*$ with probability

$$\geq 1 - 3p^2 \exp\left\{-\frac{\log p}{64\sigma^2} - \frac{\delta n}{256\sigma^2}\right\}$$

Similar results for neighborhood selection (regularized pseudo-likelihood) (Meinshausen and Bühlmann, 2006)

First theoretical results for $\text{CGL}\textsc{asso}$: Tugnait (2018, 2019a, 2019b) gave experimental results and algorithms

Model Selection Consistency of CGLasso



Improper and Dependent Observations

In addition to incoherence of $\mathbf{\Theta}^*$, CGLasso requires only that

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\max} < C\sigma\sqrt{\frac{\log p}{n}} \quad \text{for fixed } C$$

Improper and Dependent Observations

In addition to incoherence of $\mathbf{\Theta}^*$, CGLasso requires only that

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\max} < C\sigma\sqrt{\frac{\log p}{n}} \quad \text{for fixed } C$$

This condition can be satisfied under weaker assumptions,

- Improper $\boldsymbol{Z}$: consistent with same rate, worse constants
- Time series spectrum (Fiecas *et al.*, 2019): $\mathcal{O}(p^2/n)$

Improper and Dependent Observations

In addition to incoherence of $\mathbf{\Theta}^*$, $\textsc{CGLasso}$ requires only that

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\text{max}} < C\sigma\sqrt{\frac{\log p}{n}} \quad \text{for fixed } C$$

This condition can be satisfied under weaker assumptions,

- Improper $\boldsymbol{Z}$: consistent with same rate, worse constants
- Time series spectrum (Fiecas *et al.*, 2019): $\mathcal{O}(p^2/n)$

W. Theorem 4.1 + Fiecas *et al.* (2019) + Dahlhaus (2000)

Suppose $\mathcal{Z} = \{\boldsymbol{Z}_t\}_{t=1}^T$ is a stationary Gaussian linear p-dimensional time series with spectrum $\mathbf{\Gamma}(k) \in \mathbb{C}_{\succeq 0}^{p \times p}$, such that $\mathbf{\Gamma}^{-1}(k)$ satisfies the incoherence conditions at all k. Let $\hat{\mathbf{\Gamma}}(k)$ be the sample avereged periodogram based on T observations. Then the graphical model $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}})$ with

$$(i,j) \notin \hat{\mathcal{E}} \Leftrightarrow \hat{\Theta}_{ij}^{(k)} = 0 \quad \text{for } \hat{\mathbf{\Theta}}^{(k)} = \textsc{CGLasso}(\hat{\mathbf{\Gamma}}(k)) \quad \text{for all } k < T/2$$

correctly estimates the conditional independence structures in $\mathcal{Z}$ at all lags with probability $\geq 1 - Cp^2/T$ for T sufficiently large.

Implications

Developed Wirtinger convex analysis and applied it to CLasso and CGLasso estimators:

Implications

Developed Wirtinger convex analysis and applied it to CLasso and CGLasso estimators:

- Foundational Optimization and Statistical Theory for *Complex Machine Learning*

Implications

Developed Wirtinger convex analysis and applied it to CLasso and CGLasso estimators:

- Foundational Optimization and Statistical Theory for *Complex Machine Learning*
- First Finite-Sample Results for Complex *M*-Estimation

Implications

Developed Wirtinger convex analysis and applied it to CLasso and CGLasso estimators:

- Foundational Optimization and Statistical Theory for *Complex Machine Learning*
- First Finite-Sample Results for Complex M-Estimation
- First Statistical Study of Improper Gaussian Distributions

Implications

Developed Wirtinger convex analysis and applied it to CLasso and CGLasso estimators:

- Foundational Optimization and Statistical Theory for *Complex Machine Learning*
- First Finite-Sample Results for Complex *M*-Estimation
- First Statistical Study of Improper Gaussian Distributions
- Exciting Implications for Multivariate Time Series

Conclusion & Discussion

Conclusions

Novel Methodologies with Sophisticated Structure: tensors, low-rank estimation, misaligned & high-frequency time series, complex-variates

Conclusions

Novel Methodologies with Sophisticated Structure: tensors, low-rank estimation, misaligned & high-frequency time series, complex-variates

Efficient algorithms + strong convergence guarantees built on robust convex analysis

Conclusions

Novel Methodologies with Sophisticated Structure: tensors, low-rank estimation, misaligned & high-frequency time series, complex-variates

Efficient algorithms + strong convergence guarantees built on robust convex analysis

Ability to flexibly incorporate problem structure into estimation methodology directly

Conclusions

Novel Methodologies with Sophisticated Structure: tensors, low-rank estimation, misaligned & high-frequency time series, complex-variates

Efficient algorithms + strong convergence guarantees built on robust convex analysis

Ability to flexibly incorporate problem structure into estimation methodology directly

Extensions of classical M-estimation and convex analysis to Wirtinger functions

Acknowledgements

M	G	Z	L	U	K	E	M	I	N	J	I	E	F	M
M	A	J	F	G	G	L	I	B	B	Y	U	X	R	I
A	A	R	O	A	N	D	E	R	S	E	N	N	E	T
K	U	T	G	H	I	S	G	E	O	R	G	E	D	C
E	I	G	T	A	A	K	R	I	S	T	E	N	Y	H
A	I	M	U	T	R	N	J	U	L	I	A	H	G	T
N	U	G	A	S	O	E	M	E	W	V	T	G	U	E
D	U	T	A	T	T	M	T	A	T	A	K	B	R	R
R	G	I	U	U	T	I	R	Z	K	N	L	E	D	E
E	R	A	K	Y	T	E	N	I	L	N	U	N	C	N
A	F	N	K	O	V	A	O	E	O	U	I	H	O	C
B	P	Y	M	E	W	N	M	N	H	C	S	A	F	E
N	F	I	N	B	T	A	N	E	J	C	M	A	E	G
I	S	E	Y	O	O	A	L	A	J	I	I	M	S	P
E	G	F	O	B	C	D	R	J	O	N	E	S	S	E

W. “Splitting Methods for Convex Bi-Clustering and Co-Clustering.” *DSW 2019*

W. “Multi-Rank Sparse and Functional PCA: Manifold Optimization and Iterative Deflation Techniques.” *CAMSAP 2019*

W, Y. Han, and K. B. Ensor. “Multivariate Modeling of Natural Gas Spot Trading Hubs Incorporating Futures Market Realized Volatility.” *Under revision at JASA: ACS*

2020 ASA Business & Economic Statistics Student Paper Award

W. “High-Dimensional Spectral Graph Estimation via Wirtinger Graphical Models for Complex-Valued Data.” *In preparation for AoS*

Thank you!