Sparse & Functional Principal Components Analysis

Genevera I. Allen

Department of Statistics and Electrical and Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital.

April 4, 2014

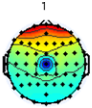
Structured Big-Data

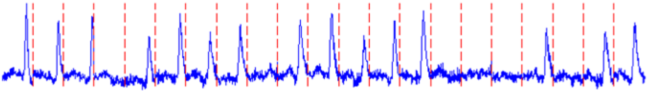
Structured Data = Data associated with locations.

- Time Series, Longitudinal & Spatial Data.
- Image data & Network Data.
- Tree Data.
- *Object-Oriented Data*

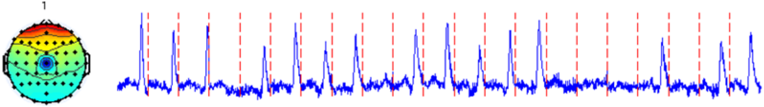
Examples of Massive Structured Data:

- Neuroimaging and Neural Recordings: MRI, fMRI, EEG, MEG, PET, DTI, direct neuronal recordings (spike trains), optigenetics.





Structured Big-Data



Data matrix, $\mathbf{X}_{L \times T}$, of L brain locations by T time points.

Goal: Unsupervised analysis of brain activation patterns and temporal neural activation patterns.

Principal Components Analysis:

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

Review: PCA Models

1. Covariance: Leading eigenspace of Gaussian covariance.
   - Model: $\mathbf{X} \sim N(0, \mathbf{I} \otimes \mathbf{\Sigma}_{p \times p})$ and estimate leading eigenspace of $\mathbf{\Sigma}$.
   - Empirical Optimization Problem:

$$\underset{\mathbf{v}_k}{\text{maximize}} \quad \mathbf{v}_k^T \mathbf{X}^T \mathbf{X} \mathbf{v}_k \quad \text{subject to} \quad \mathbf{v}_k^T \mathbf{v}_k = 1 \ \& \ \mathbf{v}_k^T \mathbf{v}_j = 0 \ \forall \ k > j.$$

2. Matrix Factorization: Low-rank mean structure.
   - Model: $\mathbf{X} = \mathbf{M} + \mathbf{E}$ for mean matrix $\mathbf{M}$ that is low-rank and $\mathbf{E}$ iid additive noise.
   - Empirical Optimization Problem:

$$\underset{\mathbf{U},\mathbf{D},\mathbf{V}}{\text{minimize}} \ || \mathbf{X} - \mathbf{U} \mathbf{D} \mathbf{V}^T ||_F^2 \ \text{subject to} \ \mathbf{U}^T \mathbf{U} = \mathbf{I} \ \& \ \mathbf{V}^T \mathbf{V} = \mathbf{I}.$$

Solution: Singular Value Decomposition (SVD).

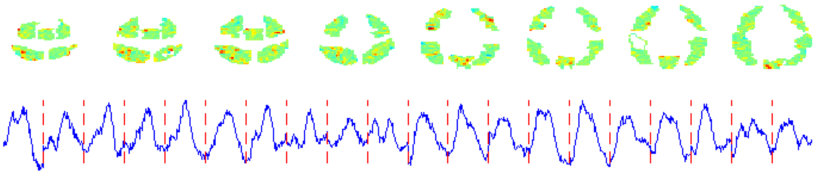
Review: Regularized PCA

Big-Data settings, regularize the leading eigenvalues leading to . . .

1. Functional PCA.
   - Encourage PC factors to be smooth with respect to known data structure. (Rice & Silverman, 1991; Silverman, 1996; Ramsay, 2006; Huang et al., 2008).
   - Leads to consistent PC estimates. (Silverman, 1996)
2. Sparse PCA.
   - Automatic feature / variable selection. (Jollieffe et al., 2003; Zou et al., 2006; d'Aspermont et al., 2007; Shen & Huang, 2008)
   - Leads to consistent PC estimates. (Johnstone & Lu, 2009; Amini & Wainwright, 2009; Shen et al., 2012, Vu & Lei, 2013)

Why Sparse & Functional PCA?

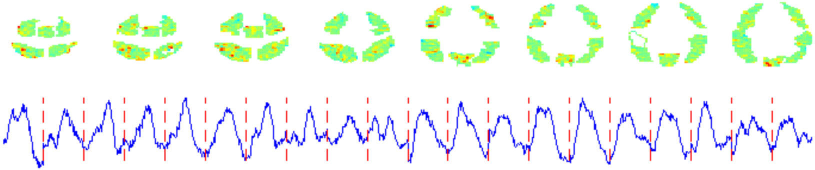
1. Applied Motivation (Neuroimaging):



2. Improved signal recovery, feature selection, interpretation, data visualization.
3. Question: Is there a general mathematical framework for regularization in the context of PCA?

Why Sparse & Functional PCA?

1. Applied Motivation (Neuroimaging):



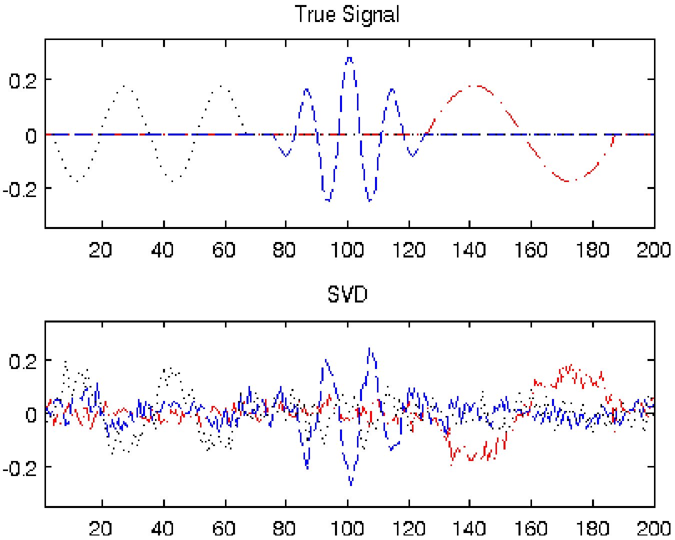
Objectives

(i) Formulate a (good) optimization framework to achieve SFPCA.

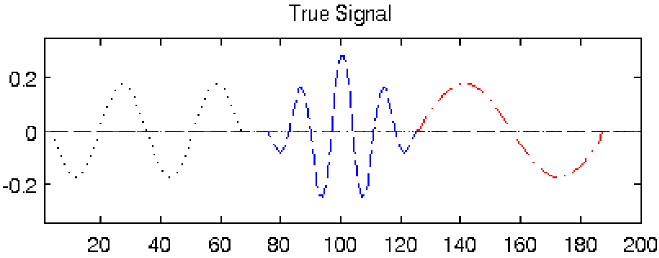
(ii) Develop a scalable algorithm to fit SFPCA.

(iii) Carefully study the properties of the model and algorithm from an optimization perspective.

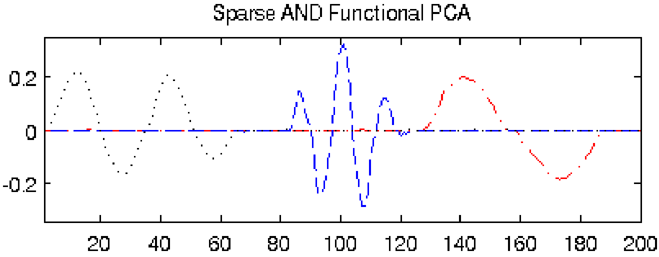
Preview


True Signal
0.2
0
-0.2
20
40
60
80
100
120
140
160
180
200
SVD
0.2
0
-0.2
20
40
60
80
100
120
140
160
180
200

Preview



True Signal
0.2
0
-0.2
20
40
60
80
100
120
140
160
180
200



Sparse AND Functional PCA
0.2
0
-0.2
20
40
60
80
100
120
140
160
180
200

Setup

Matrix Factorization: Low-Rank Mean Model.

$$\mathbf{X}_{n \times p} = \sum_{k=1}^{K} d_k \, \mathbf{u}_k \, \mathbf{v}_k^T + \epsilon$$

- Assume data, $\mathbf{X}$, previously centered.
- PC Factors $\mathbf{v}$ and/or $\mathbf{u}$ are sparse and/or smooth.
- ϵ is iid noise; d_k is fixed, but unknown.
- Rows and / or columns of $\mathbf{X}$ arise from discretized curves or other features associated with locations.

Functional PCA & Two-Way FPCA

Encouraging smoothness:

- Continuous functions: 2^{nd} derivatives quantify curvature.
- Penalty: $\int f''(t) f''(t) dt$. Penalizes average squared curvature.
- Discrete extension: Squared second differences between adjacent variables.
- Discrete extension in matrix form:

$$\beta^T \mathbf{\Omega} \beta = \sum (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$$

  $\mathbf{\Omega}_{p \times p} \succ 0$ is the second differences matrix.

- Other possible choices $\mathbf{\Omega}$ that encourages smoothness. (Ramsay 2006).

Functional PCA & Two-Way FPCA

Functional PCA

$$\underset{\mathbf{v}}{\text{maximize}} \quad \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} \quad \text{subject to} \quad \mathbf{v}^T (\mathbf{I} + \alpha_{\mathbf{v}} \boldsymbol{\Omega}_{\mathbf{v}}) \mathbf{v} = 1.$$

Silverman (1996)

Equivalent to:

- Regression Approach: (Huang et al., 2008)

$$\hat{\mathbf{u}} = \text{argmin}_{\mathbf{u}} \left\{ || \mathbf{X} \hat{\mathbf{v}} - \mathbf{u} ||_2^2 \right\}.$$
$$\hat{\mathbf{v}} = \text{argmin}_{\mathbf{v}} \left\{ || \mathbf{X}^T \hat{\mathbf{u}} - \mathbf{v} ||_2^2 + \alpha \, \mathbf{v}^T \boldsymbol{\Omega} \, \mathbf{v} \right\}.$$

- Half-Smoothing.

Functional PCA & Two-Way FPCA

Two-Way Functional PCA

$$\underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \quad \mathbf{u}^T \mathbf{X} \mathbf{v} - \frac{1}{2} \mathbf{u}^T (\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}}) \mathbf{u} \, \mathbf{v}^T (\mathbf{I} + \alpha_{\mathbf{v}} \boldsymbol{\Omega}_{\mathbf{v}}) \mathbf{v} .$$

Huang et al. (2009)

- Equivalent to two-way half-smoothing.
- Related to two-way ℓ_2 penalized regression.

Sparse PCA & Two-Way SPCA

Sparse PCA via Penalized Regression

$$\hat{\mathbf{u}} = \operatorname{argmin}_{\mathbf{u}} \left\{ || \mathbf{X} \hat{\mathbf{v}} - \mathbf{u} ||_2^2 \right\}.$$

$$\hat{\mathbf{v}} = \operatorname{argmin}_{\mathbf{v}} \left\{ || \mathbf{X}^T \hat{\mathbf{u}} - \mathbf{v} ||_2^2 + \lambda || \mathbf{v} ||_1 \right\}.$$

Shen & Huang (2008).

Other SPCA approaches:

- Semi-definite programming.
- Covariance thresholding.

Sparse PCA & Two-Way SPCA

Two-Way Sparse PCA

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \quad & \mathbf{u}^T \mathbf{X} \mathbf{v} - \lambda_{\mathbf{u}} || \mathbf{u} ||_1 - \lambda_{\mathbf{v}} || \mathbf{v} ||_1 \\ \text{subject to} \quad & \mathbf{u}^T \mathbf{u} \leq 1 \;\&\; \mathbf{v}^T \mathbf{v} \leq 1. \end{aligned}$$

Allen et al. (2011); Lagrangian of Witten et al. (2009); Related to Lee et al. (2010).

- SPCA of Shen & Huang (2008) a special case.
- Related to two-way penalized regression.

Two-Way Regularized PCA

Alternating Penalized Regressions

$$\hat{\mathbf{u}} = \operatorname{argmin}_{\mathbf{u}} \left\{ || \mathbf{X} \hat{\mathbf{v}} - \mathbf{u} ||_2^2 + \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) \right\}.$$

$$\hat{\mathbf{v}} = \operatorname{argmin}_{\mathbf{v}} \left\{ || \mathbf{X}^T \hat{\mathbf{u}} - \mathbf{v} ||_2^2 + \lambda_{\mathbf{v}} P_{\mathbf{v}}(\mathbf{v}) \right\}.$$

Questions:

1. Are sparse AND smoothing ℓ_2 penalties permitted?
2. What penalty types lead to convergent solutions?
3. What optimization problem is this class of algorithms solving?

We consider ...

Objective

Flexible combinations of smoothness and / or sparsity on the row and / or column PC factors.

Four Penalties:

- Row Sparsity: $\lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u})$, for example $\lambda_{\mathbf{u}} || \, \mathbf{u} \, ||_1$.
- Row Smoothness: $\alpha_{\mathbf{u}} \, \mathbf{u}^T \, \boldsymbol{\Omega}_{\mathbf{u}} \, \mathbf{u}$; $\quad \boldsymbol{\Omega}_{\mathbf{u}}^{(n \times n)} \succ 0$.
- Column Sparsity: $\lambda_{\mathbf{v}} P_{\mathbf{v}}(\mathbf{v})$ for example $\lambda_{\mathbf{v}} || \, \mathbf{v} \, ||_1$.
- Column Smoothness: $\alpha_{\mathbf{v}} \, \mathbf{v}^T \, \boldsymbol{\Omega}_{\mathbf{v}} \, \mathbf{v}$; $\quad \boldsymbol{\Omega}_{\mathbf{v}}^{(p \times p)} \succ 0$.

Approach: Iteratively solve for the best rank-one solution in a greedy manner.

Formulating an Optimization Problem

We want . . .

1. To generalize existing methods for PCA, SPCA, FPCA, and two-way SPCA and FPCA.
   - These should all be special cases when regularization parameters are active.
2. Desirable numerical properties:
   - Identifiable PC factors.
   - Non-degenerate and Well-scaled solution and PC factors
   - Balanced regularization (NO regularization masking).
     - Regularization Masking: $\exists$ a $\lambda, \alpha > 0$ where the solution does not depend on both λ and α.
3. A computationally tractable algorithm in big-data settings.

Optimization Approaches: Natural Extensions?

Question:

1. Why not add ℓ_1 and smoothing ℓ_2 penalties to the Frobenius-norm loss of the SVD problem?

$$\underset{\mathbf{u}(:\mathbf{u}^T\mathbf{u}\leq 1),\mathbf{v}(:\mathbf{v}^T\mathbf{v}\leq 1)}{\text{minimize}} ||\mathbf{X} - (d)\,\mathbf{u}\,\mathbf{v}^T\,||_F^T + \lambda_{\mathbf{u}}||\,\mathbf{u}\,||_1 + \lambda_{\mathbf{v}}||\,\mathbf{v}\,||_1 + \alpha_{\mathbf{u}}\,\mathbf{u}^T\,\boldsymbol{\Omega}_{\mathbf{u}}\,\mathbf{u} + \alpha_{\mathbf{v}}\,\mathbf{v}^T\,\boldsymbol{\Omega}_{\mathbf{v}}\,\mathbf{v}\,.$$

   Unidentifiable, degenerate, does not generalize.

2. Why not add ℓ_1 penalties to the two-way FPCA optimization problem of Huang et al. (2009)?

$$\underset{\mathbf{u}(:\mathbf{u}^T\mathbf{u}\leq 1),\mathbf{v}(:\mathbf{v}^T\mathbf{v}\leq 1)}{\text{maximize}} \mathbf{u}^T\,\mathbf{X}\,\mathbf{v} - \frac{1}{2}\,\mathbf{u}^T(\mathbf{I} + \alpha_{\mathbf{u}}\,\boldsymbol{\Omega}_{\mathbf{u}})\,\mathbf{u}\,\mathbf{v}^T(\mathbf{I} + \alpha_{\mathbf{v}}\,\boldsymbol{\Omega}_{\mathbf{v}})\,\mathbf{v} - \lambda_{\mathbf{u}}||\,\mathbf{u}\,||_1 - \lambda_{\mathbf{v}}||\,\mathbf{v}\,||_1$$

3. Why not add smoothing ℓ_2 penalties to the two-way SPCA problem of Witten et al. (2009)?

$$\underset{\mathbf{u}(:\mathbf{u}^T\mathbf{u}\leq 1),\mathbf{v}(:\mathbf{v}^T\mathbf{v}\leq 1)}{\text{maximize}} \quad \mathbf{u}^T\,\mathbf{X}\,\mathbf{v} - \lambda_{\mathbf{u}}||\,\mathbf{u}\,||_1 - \lambda_{\mathbf{v}}||\,\mathbf{v}\,||_1 - \alpha_{\mathbf{u}}\,\mathbf{u}^T\,\boldsymbol{\Omega}_{\mathbf{u}}\,\mathbf{u} - \alpha_{\mathbf{v}}\,\mathbf{v}^T\,\boldsymbol{\Omega}_{\mathbf{v}}\,\mathbf{v}\,.$$

   Regularization masking, computationally challenging.

Assumptions on Penalties

A1 $\mathbf{\Omega_u} \succeq 0$ & $\mathbf{\Omega_v} \succeq 0$.

A2 $P()$'s are positive, homogeneous of order one, i.e. $P(cx) = cP(x) \; \forall \; c > 0$.

- Sparse penalties: ℓ_1-norm, SCAD, MC+, etc.
- Structured sparse: group lasso, fused lasso, generalized lasso, etc.

A3 If $P()$'s non-convex, then $P()$ can be decomposed into the difference of two convex functions, $P(x) = P_1(x) - P_2(x)$ for $P_1()$ and $P_2()$ convex.

- Includes common non-convex penalties such as SCAD, MC+, log-concave, etc.

SFPCA Optimization Problem

Rank-one Sparse & Functional PCA the solution to:

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \quad & \mathbf{u}^T \mathbf{X} \mathbf{v} - \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) - \lambda_{\mathbf{v}} P_{\mathbf{v}}(\mathbf{v}) \\ \text{subject to} \quad & \mathbf{u}^T (\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}}) \mathbf{u} \leq 1 \ \& \ \mathbf{v}^T (\mathbf{I} + \alpha_{\mathbf{v}} \boldsymbol{\Omega}_{\mathbf{v}}) \mathbf{v} \leq 1. \end{aligned}$$

- Penalty Parameters: $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} \geq 0$.

Relationship to Other PCA Approaches

Theorem

(i) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to PCA / the SVD of $\mathbf{X}$.

Relationship to Other PCA Approaches

Theorem

(i) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to PCA / the SVD of $\mathbf{X}$.

(ii) If $\lambda_{\mathbf{u}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to Sparse PCA (Shen & Huang, 2008).

Relationship to Other PCA Approaches

Theorem

(i) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to PCA / the SVD of $\mathbf{X}$.

(ii) If $\lambda_{\mathbf{u}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to Sparse PCA (Shen & Huang, 2008).

(iii) If $\alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to a special case of the two-way Sparse PCA (Allen et al., 2011), which is the Lagrangian of (Witten et al., 2009) and closely related to that (Lee et al., 2010).

Relationship to Other PCA Approaches

Theorem

(i) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to PCA / the SVD of $\mathbf{X}$.

(ii) If $\lambda_{\mathbf{u}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to Sparse PCA (Shen & Huang, 2008).

(iii) If $\alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to a special case of the two-way Sparse PCA (Allen et al., 2011), which is the Lagrangian of (Witten et al., 2009) and closely related to that (Lee et al., 2010).

(iv) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}} = 0$, then equivalent to the Functional PCA solution of (Silverman, 1996) and (Huang et al., 2008).

Relationship to Other PCA Approaches

Theorem

(i) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to PCA / the SVD of $\mathbf{X}$.

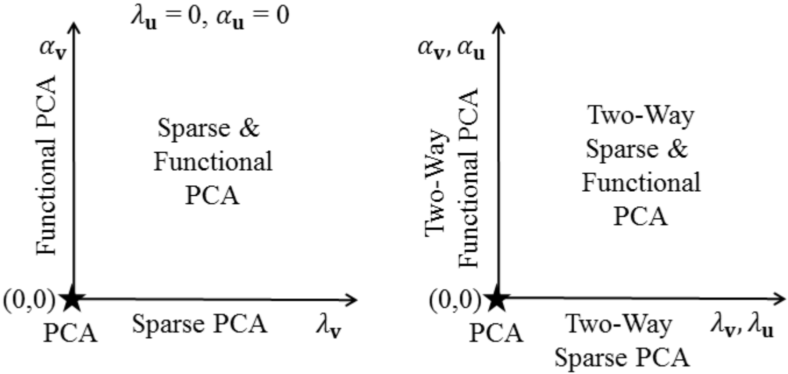
(ii) If $\lambda_{\mathbf{u}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to Sparse PCA (Shen & Huang, 2008).

(iii) If $\alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then equivalent to a special case of the two-way Sparse PCA (Allen et al., 2011), which is the Lagrangian of (Witten et al., 2009) and closely related to that (Lee et al., 2010).

(iv) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}} = 0$, then equivalent to the Functional PCA solution of (Silverman, 1996) and (Huang et al., 2008).

(v) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}} = 0$, then equivalent to the two-way FPCA solution of (Huang et al., 2009).

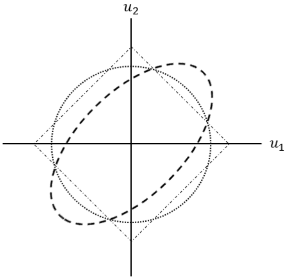
Relationship to Other PCA Approaches



Desirable Numerical Properties

Theorem

1. Identifiable ($\mathbf{u}$ and $\mathbf{v}$) up to a sign change.
2. Balanced regularization (no *regularization masking*):
   - $\exists$ a $\lambda_{\mathbf{u}}^{max}$ s.t. $\mathbf{u}^* = 0$.
   - $(\mathbf{u}^*, \mathbf{v}^*)$ depend on all non-zero regularization parameters.
3. Well-scaled and non-degenerate.
   - Either $\mathbf{u}^{*,T}(\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}})\, \mathbf{u}^* = 1$ and $\mathbf{v}^{*,T}(\mathbf{I} + \alpha_{\mathbf{v}} \boldsymbol{\Omega}_{\mathbf{v}})\, \mathbf{v}^* = 1$,
   - Or $\mathbf{u}^* = 0$ and $\mathbf{v}^* = 0$.



SFPCA Optimization Problem

Rank-one Sparse & Functional PCA the solution to:

$$\underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \quad \mathbf{u}^T \mathbf{X} \mathbf{v} - \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) - \lambda_{\mathbf{v}} P_{\mathbf{v}}(\mathbf{v})$$
$$\text{subject to} \quad \mathbf{u}^T (\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}}) \mathbf{u} \leq 1 \ \& \ \mathbf{v}^T (\mathbf{I} + \alpha_{\mathbf{v}} \boldsymbol{\Omega}_{\mathbf{v}}) \mathbf{v} \leq 1.$$

- Non-convex, non-differentiable, QCQP.
- $P()$ convex $\Longrightarrow$ bi-convex.
  - Convex in $\mathbf{v}$ with $\mathbf{u}$ fixed as well as the converse.

Idea: Alternating optimization.

Relationship to Penalized Regression

Theorem

The solution to the SFPCA problem with respect to $\mathbf{u}$ is given by:

$$\hat{\mathbf{u}} = \text{argmin}_{\mathbf{u}} \left\{ \frac{1}{2} || \mathbf{X} \mathbf{v} - \mathbf{u} ||_2^2 + \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) + \frac{\alpha_{\mathbf{u}}}{2} \mathbf{u}^T \boldsymbol{\Omega}_u \mathbf{u} \right\}$$

$$\mathbf{u}^* = \begin{cases} \hat{\mathbf{u}} / ||\hat{\mathbf{u}}||_{\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}}} & ||\hat{\mathbf{u}}||_{\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}}} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Analogous to re-scaled Elastic Net problem! (Zou & Hastie, 2005).
- Result holds because of A2, order-one penalties.

Relationship to Penalized Regression

$$\underset{\mathbf{u}}{\text{minimize}} \quad \frac{1}{2} || \mathbf{X} \, \mathbf{v} - \mathbf{u} \, ||_2^2 + \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) + \frac{\alpha_{\mathbf{u}}}{2} \, \mathbf{u}^T \, \mathbf{\Omega}_u \, \mathbf{u}$$

- Can be solved by (accelerated) proximal gradient descent.
- Geometric convergence, $O(k \log k)$, for convex penalties.
- A3, difference of convex, ensures convergence for non-convex penalties.
- Closed form proximal operator for many penalties: $\text{prox}_P(\mathbf{y}, \lambda) = \text{argmin}_{\mathbf{x}} \{ \frac{1}{2} ||\mathbf{x} - \mathbf{y}||_2^2 + \lambda P(\mathbf{x}) \}$.
  - Some form of thresholding (i.e. soft-thresholding for ℓ_1 penalty).

SFPCA Algorithm

Rank-One Algorithm

1. Initialize $\mathbf{u}$ and $\mathbf{v}$ to that of the rank-1 SVD of $\mathbf{X}$. Set $\mathbf{S_u} = \mathbf{I} + \alpha_\mathbf{u} \mathbf{\Omega_u}$ and $L_\mathbf{u} = \lambda_{max}(\mathbf{S_u})$; set $\mathbf{S_v} = \mathbf{I} + \alpha_\mathbf{v} \mathbf{\Omega_v}$ and $L_\mathbf{v} = \lambda_{max}(\mathbf{S_v})$.
2. Repeat until convergence:
    1. Estimate $\hat{\mathbf{u}}$. Repeat until convergence:
    $\mathbf{u}^{(t+1)} = \text{prox}_{P_\mathbf{u}} \left(\mathbf{u}^{(t)} + \frac{1}{L}(\mathbf{X}\,\mathbf{v}^* - \mathbf{S_u}\,\mathbf{u}^{(t)}), \frac{\lambda_\mathbf{u}}{L_\mathbf{u}} \right)$.
    2. Set $\mathbf{u}^* = \begin{cases} \hat{\mathbf{u}}/\|\hat{\mathbf{u}}\|_{\mathbf{S_u}} & \|\hat{\mathbf{u}}\|_{\mathbf{S_u}} > 0 \\ 0 & \text{otherwise.} \end{cases}$
    3. Estimate $\hat{\mathbf{v}}$. Repeat until convergence:
    $\mathbf{v}^{(t+1)} = \text{prox}_{P_\mathbf{v}} \left(\mathbf{v}^{(t)} + \frac{1}{L}(\mathbf{X}^T\,\mathbf{u}^* - \mathbf{S_v}\,\mathbf{v}^{(t)}), \frac{\lambda_\mathbf{v}}{L_\mathbf{v}} \right)$.
    4. Set $\mathbf{v}^* = \begin{cases} \hat{\mathbf{v}}/\|\hat{\mathbf{v}}\|_{\mathbf{S_v}} & \|\hat{\mathbf{v}}\|_{\mathbf{S_v}} > 0 \\ 0 & \text{otherwise.} \end{cases}$
3. Return $\mathbf{u} = \mathbf{u}^* / \|\,\mathbf{u}^*\,\|_2$, $\mathbf{v} = \mathbf{v}^* / \|\,\mathbf{v}^*\,\|_2$, and $d = \mathbf{u}^T\,\mathbf{X}\,\mathbf{v}$.

Convergence

Theorem: Convergence of rank-one SFPCA.

For $P_{\mathbf{u}}$ and $P_{\mathbf{v}}$ convex,

- The SFPCA Algorithm convergences to a stationary point of the SFPCA problem.
- The solution is unique given an initial starting point.

Convergence

Theorem: Convergence of rank-one SFPCA.

For $P_{\mathbf{u}}$ and $P_{\mathbf{v}}$ convex,

- The SFPCA Algorithm convergences to a stationary point of the SFPCA problem.
- The solution is unique given an initial starting point.

- Multi-rank solutions can be computed in a greedy manner (power method).
- Several deflation schemes available (Mackey, 2009).
- Subtraction deflation most common:
  - Fit rank-one SFPCA to $\mathbf{X}$ to estimate 1st PC factors.
  - Subtract rank one fit, $\mathbf{X} - d_1 \, \mathbf{u}_1 \, \mathbf{v}_1^T$, and apply SFPCA to estimate 2nd PC factors.

Extensions: Non-Negativity Constraints

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \quad & \mathbf{u}^T \mathbf{X} \mathbf{v} - \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) - \lambda_{\mathbf{v}} P_{\mathbf{v}}(\mathbf{v}) \\ \text{subject to} \quad & \mathbf{u}^T (\mathbf{I} + \alpha_{\mathbf{u}} \boldsymbol{\Omega}_{\mathbf{u}}) \mathbf{u} \leq 1 \ \& \ \mathbf{v}^T (\mathbf{I} + \alpha_{\mathbf{v}} \boldsymbol{\Omega}_{\mathbf{v}}) \mathbf{v} \leq 1, \\ & \mathbf{u} \geq 0 \ \& \ \mathbf{v} \geq 0. \end{aligned}$$

Corollary

Replace $\text{prox}_P()$ with the *positive proximal operator*, $\text{prox}_P^+(\mathbf{y}, \lambda) = \text{argmin}_{\mathbf{x}:\mathbf{x}\geq 0}\{\frac{1}{2}||\mathbf{x} - \mathbf{y}||_2^2 + \lambda P(\mathbf{x})\}$.

- Same convergence guarantees hold.

Selecting Regularization Parameters

- Cross-validation (CV), Generalized CV, BIC, etc.
- Nested vs. Grid search.

Ideas:

- Select $\alpha_{\mathbf{u}}$ and $\lambda_{\mathbf{u}}$ together.
- Exploit connection to Elastic Net to compute degrees of freedom for BIC.

Example for ℓ_1 penalty:

- Let $\mathcal{A}(\mathbf{u})$ be the active set of $\mathbf{u}$.

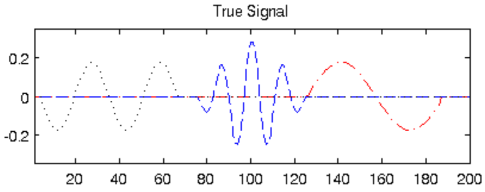
$$\hat{df}(\alpha_{\mathbf{u}}, \lambda_{\mathbf{u}}) = \mathrm{trace}\left[\left(\mathbf{I}_{\mathcal{A}(\mathbf{u})} - \frac{\alpha_{\mathbf{u}}}{2} \boldsymbol{\Omega}_{\mathbf{u}}(\mathcal{A}(\mathbf{u}), \mathcal{A}(\mathbf{u})) \right)^{-1} \right].$$

Simulation I

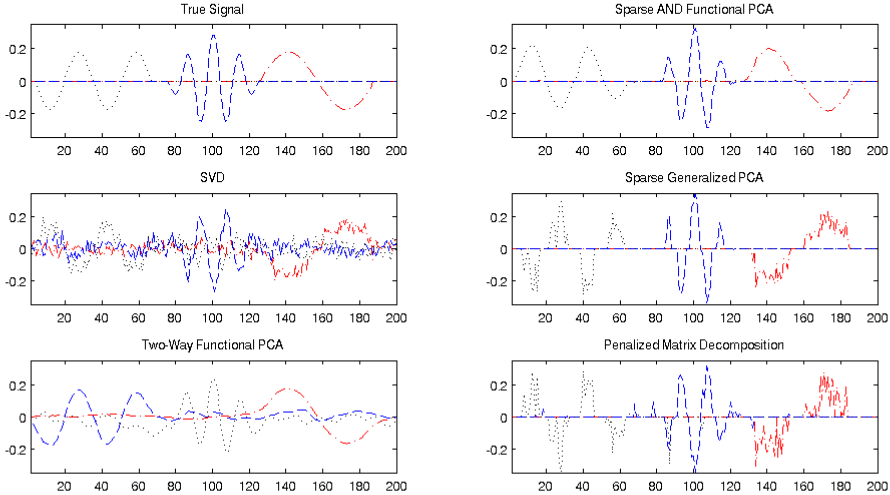
Setup: Rank-3 Model with sparse & smooth right factors.

$$\mathbf{X}_{n \times p} = \sum_{k=1}^{K} d_k \, \mathbf{u}_k \, \mathbf{v}_k^T + \epsilon$$

- $\epsilon_{ij} \overset{iid}{\sim} N(0, 1)$.
- $\mathbf{u}_k$ random orthonormal vectors of length n;
  $D = \mathrm{diag}([n/4, n/5, n/6]^T)$.
- $\mathbf{v}_k$ fixed signal vectors of length $p = 200$:



Simulation I



Simulation I

Table: $n = 100$ Results.

		TWFPCA	SSVD	PMD	SGPCA	SFPCA
$\mathbf{v}_1$	TP	-	0.897	0.568	0.768	**0.935**
	FP	-	0.323	0.001	0.006	**0.052**
	r$\angle$	**0.153**	0.625	2.220	0.726	0.189
$\mathbf{v}_2$	TP	-	0.783	0.657	0.445	**0.713**
	FP	-	0.320	0.106	0.002	**0.047**
	r$\angle$	5.980	0.549	0.597	0.829	**0.438**
$\mathbf{v}_3$	TP	-	0.771	0.514	0.499	**0.883**
	FP	-	0.316	0.066	0.004	**0.054**
	r$\angle$	3.660	0.855	1.270	1.010	**0.468**
	rSE	0.668	0.760	1.000	0.737	**0.450**

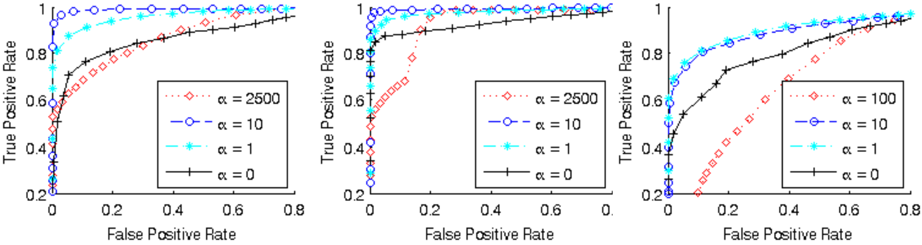
Simulation I

Table: $n = 300$ Results.

		TWFPCA	SSVD	PMD	SGPCA	SFPCA
$\mathbf{v}_1$	TP	-	0.973	0.509	0.921	**0.987**
	FP	-	0.322	0.000	0.005	**0.068**
	r$\angle$	0.768	0.487	15.700	0.553	**0.152**
$\mathbf{v}_2$	TP	-	0.919	0.773	0.839	**0.967**
	FP	-	0.319	0.000	0.038	**0.048**
	r$\angle$	52.300	0.428	1.310	0.488	**0.320**
$\mathbf{v}_3$	TP	-	0.943	0.530	0.849	**0.972**
	FP	-	0.314	0.000	0.015	**0.060**
	r$\angle$	33.100	0.545	5.940	0.631	**0.131**
	rSE	1.170	0.790	3.380	0.809	**0.655**

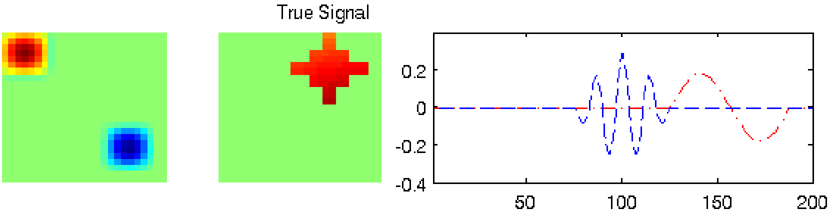
Simulation I

SFPCA also improves feature selection . . .

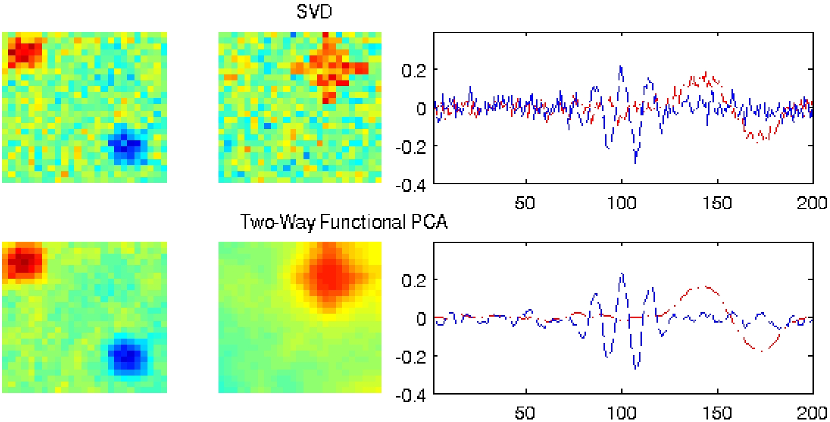


Simulation II

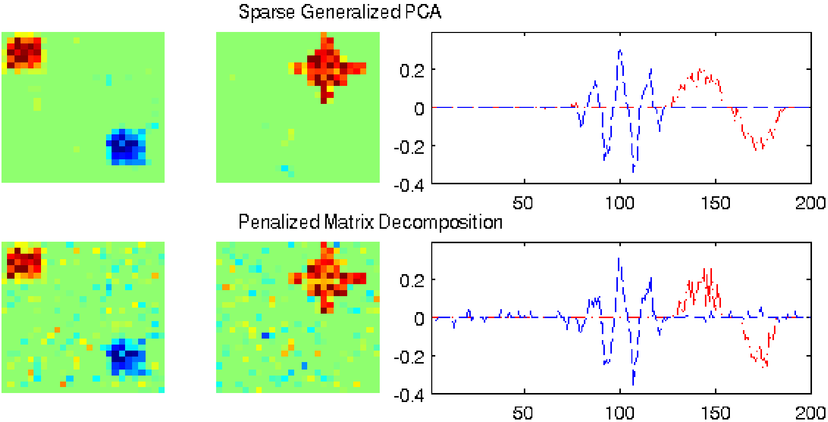
Setup: Rank-2 Model with sparse & smooth spatial (25×25 grid) and temporal (200-length vector) factors.



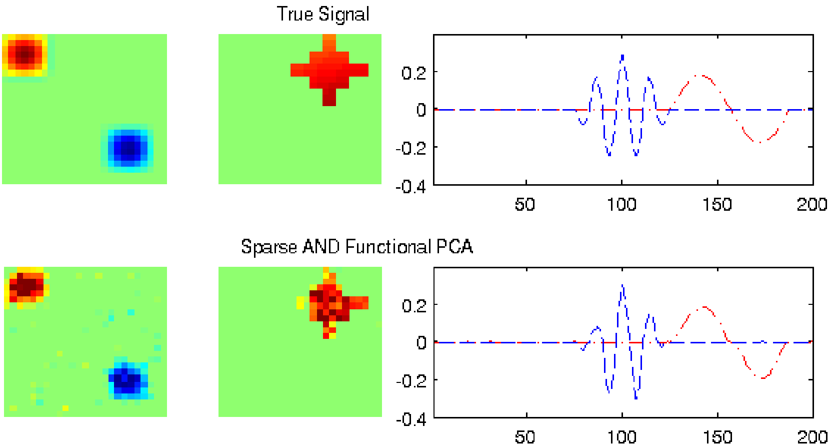
Simulation II



Simulation II



Simulation II



1. Motivation
2. Background & Challenges: Regularized PCA
3. Sparse & Functional PCA Model
4. Sparse & Functional PCA Algorithm
5. Simulation Studies
6. Case Study: EGG Data

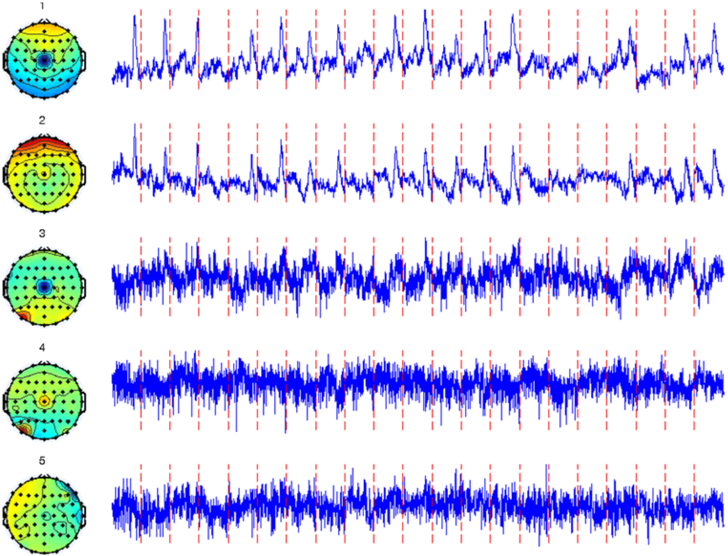
EEG Predisposition to Alcoholism

Data:

- EEG measures electrical signals in the active brain over time.
- Sampled from 64 channels at 256Hz.
- Consider 1^{st} alcoholic subject over epochs relating to non-matching image stimuli.
- Data matrix: 57×5376, channel location by epoch time points (21 epochs of 256 time points each).
- $\mathbf{\Omega_u}$ weighted squared second differences matrix using spherical distances between channel locations.
- $\mathbf{\Omega_v}$ squared second differences matrix between time points.

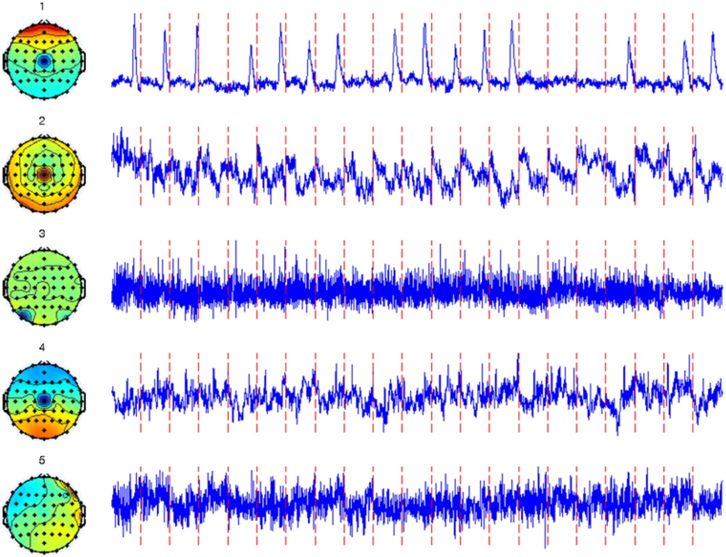
EEG Results

PCA Results:



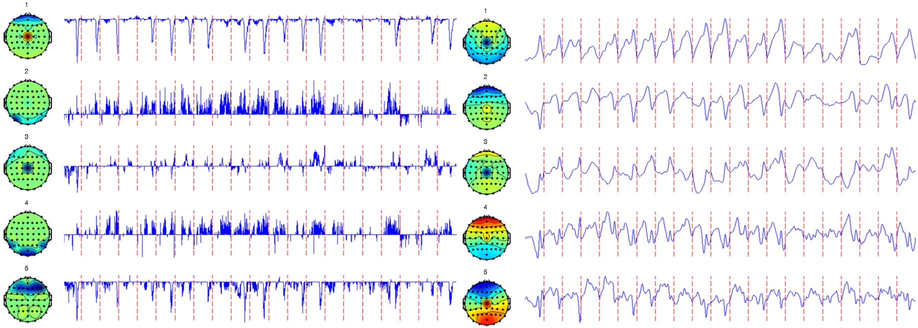
EEG Results

Independent Component Analysis (ICA) Results:



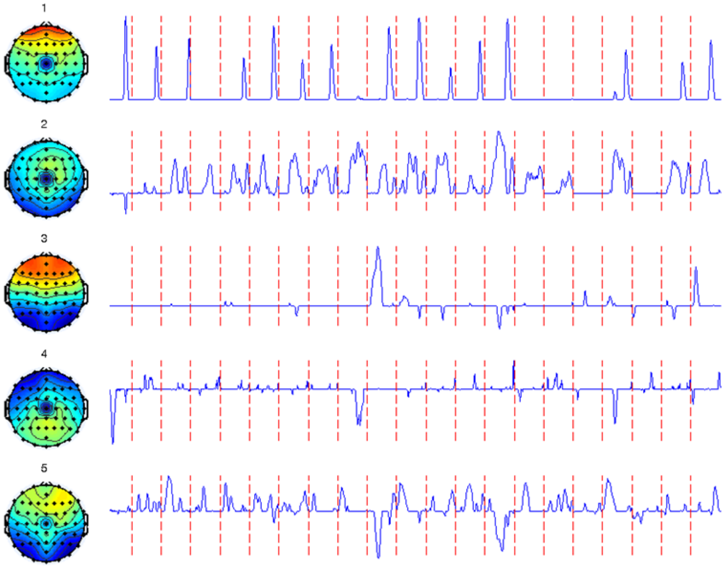
EEG Results

Penalized Matrix Decomposition & Two-Way FPCA Results:



EEG Results

Sparse & Functional PCA Results:



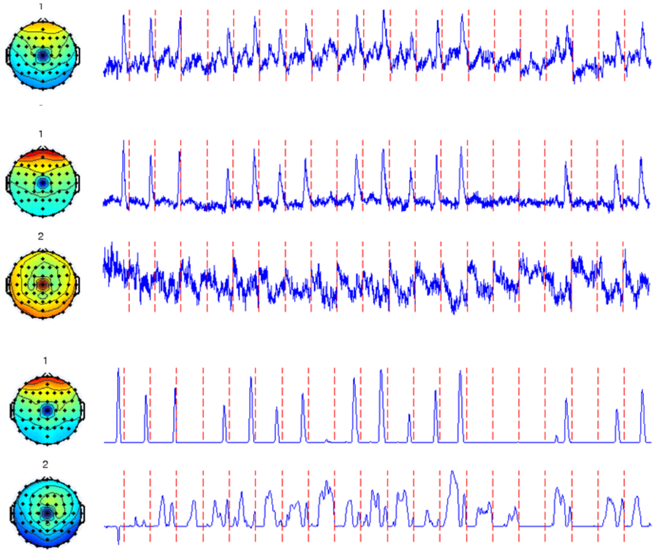
EEG Results

Comparison:

PCA:

ICA:

SFPCA:



EEG Results

SFPCA Notes:

- 3.28 seconds to converge. (Software entirely in Matlab).
- BIC selected $\lambda_{\mathbf{u}} = 0$ (spatial sparsity) for first 5 components.
- BIC selected $\alpha_{\mathbf{u}} = 10 - 12$, $\alpha_{\mathbf{v}} = 0.5 - 10$, and $\lambda_{\mathbf{v}} = 1 - 2.5$ for first 5 components.

Flexible, data-driven selection of appropriate amount of regularization.

Summary & Future Work

Summary

- SFPCA generalizes much of the existing literature on regularized PCA via alternating regressions.
- SFPCA has the flexibility to permit many types of regularization in a data-driven manner.
- SFPCA results in better signal recovery and more interpretable factors as well as improved feature selection.

Future Statistical Work:

- Statistical consistency, especially in high-dimensional settings.
- Extensions to other multivariate methods: CCA, PLS, LDA, Clustering, and etc.

The Bigger Picture: Modern Multivariate Analysis

Goal: Flexible, data-driven approaches for analyzing complex big-data.

Approach: Alternating penalized regressions framework and deflation for any eigenvalue or singular value problems.

Can mix and match any of the following:

- Generalizations that permit non-iid noise: Generalized PCA (Allen et al., 2013).
- Non-negativity constraints. (Today's talk; Zaas, ; Allen and Maletic-Savatic, 2011).
- Higher-order data and multi-way arrays. (Allen, 2012; Allen, 2013).
- Structured Signal: Sparsity and/or Smoothness. (Today's talk).

Acknowledgments

Funding:

National Science Foundation, Division of Mathematical Sciences 1209017 & 1264058.

Software available at:

`http://www.stat.rice.edu/~gallen/software.html`

References

G. I. Allen, “Sparse and Functional Principal Components Analysis”, arXiv:1309.2895, 2013.

G. I. Allen, L. Grosenick, and J. Taylor, ”A Generalized Least Squares Matrix Decomposition”, (To Appear) Journal of the American Statistical Association, Theory & Methods, 2014.

G. I. Allen and M. Maletic-Savatic, ”Sparse Non-negative Generalized PCA with Applications to Metabolomics”, Bioinformatics, 27:21, 3029-3035, 2011.

G. I. Allen, ”Sparse Higher-Order Principal Components Analysis”, In Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, 2012.

G. I. Allen, C. Peterson, M. Vannucci, and M. Maletic-Savatic, ”Regularized Partial Least Squares with an Application to NMR Spectroscopy”, Statistical Analysis and Data Mining, 6:4, 302-314, 2013.

G. I. Allen, ”Multi-way Functional Principal Components Analysis”, In IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2013.

J. Huang, H. Shen and A. Buja, “The analysis of two-way functional data using two-way regularized singular value decompositions”, Journal of the American Statistical Association, Theory & Methods, 104:488, 2009.

D. M. Witten, R. Tibshirani, and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis”, Biostatistics, 10:3, 515-534, 2009.