# Sparse Principal Component Analysis for Frequency Data

Tobias Bork

Institute for Numerical Simulation

December 10, 2019





#### Sparse Principal Component Analysis

Tobias Bork

ntroduction

1 0

lde:

racu

iviatnematical Formulatio

imite of Heability

Application

ipaise i CA

Mathematical Formulatio

Sparsity and Norms

Numerical Solution
Application

Further Analysis

References

# **Dimensionality Reduction**

Sparse Principal Component Analysis

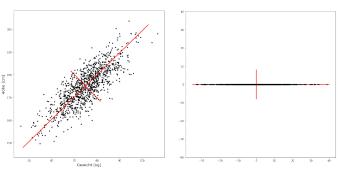
Tobias Bork

#### Introduction

Problems in high dimensions:

- Time and storage space
- Visualizing data set
- Curse of dimensionality
- ▶ Idea: Reduce the number of variables while preserving structure in the data
- Approach: Feature selection methods
- **Approach:** Feature extraction methods

### Idea of PCA



data set

(a) Finding principal axis on a (b) Linear projection of data to first principal axis

#### Sparse Principal Component Analysis

Tobias Bork

Idea

### Mathematical Formulation

Let  $\mathbf{X} \in \mathbb{R}^{n \times p}$  be a centered data matrix with n samples and p variables. We find the first principal axis by

$$v_1 = \underset{\|v\|_2=1}{\operatorname{arg max}} \operatorname{Var}[\mathbf{X}v] = \underset{\|v\|_2=1}{\operatorname{arg max}} v^T \mathbf{\Sigma} v$$

where  $\Sigma = \frac{\mathbf{X}^T \mathbf{X}}{n}$  is the sample covariance matrix.

Sparse Principal Component Analysis

Tobias Bork

ntroduction

CA

Mathematical Formulations

heorems

Carres DCA

rse PCA

Mathematical Formulation Sparsity and Norms

Numerical Solution
Application

urther Analysis

References

### Mathematical Formulation

Let  $\mathbf{X} \in \mathbb{R}^{n \times p}$  be a centered data matrix with n samples and p variables. We find the first principal axis by

$$v_1 = \underset{\|v\|_2=1}{\operatorname{arg max}} \operatorname{Var}[\mathbf{X}v] = \underset{\|v\|_2=1}{\operatorname{arg max}} v^T \mathbf{\Sigma} v$$

where  $\Sigma = \frac{\mathbf{X}^T \mathbf{X}}{n}$  is the sample covariance matrix. We compute the following principal axis successively

$$v_{k+1} = rg \max_{\|v\|=1} v^T \mathbf{\Sigma} v$$

subject to 
$$v_{k+1}^T v_l = 0 \quad \forall 1 \le l \le k$$

#### Sparse Principal Component Analysis

Tobias Bork

ntroduction

PCA

Idea

Mathematical Formulations

Theorems Limits of Usa

arse PCA

hematical Formulat

parsity and Norms umerical Solution

Further Analysis

.

Mathematical Formulations

Limits of Usab

parse PCA

Mathematical Formulation Sparsity and Norms

Application
Further Analysis

rurther Analysis

. ..

Appendix

Let  $\mathbf{X} \in \mathbb{R}^{n \times p}$  be a centered data matrix with n samples and p variables. We find the first principal axis by

$$v_1 = \underset{\|v\|_2=1}{\operatorname{arg\,max}} \operatorname{Var}[\mathbf{X}v] = \underset{\|v\|_2=1}{\operatorname{arg\,max}} v^T \mathbf{\Sigma} v$$

where  $\Sigma = \frac{\mathbf{X}^T \mathbf{X}}{n}$  is the sample covariance matrix. We compute the following principal axis successively

$$v_{k+1} = arg \max_{\|v\|=1} v^T \mathbf{\Sigma} v$$

subject to 
$$v_{k+1}^T v_l = 0 \quad \forall 1 \le l \le k$$

The new principal components are defined by  $Z_i = \mathbf{X}v_i$ 

# Mathematical Formulation using SVD

The principal axis can also be computed via the eigendecomposition of  $\Sigma$ .

$$\pmb{\Sigma} = \pmb{\mathsf{VLV}}^{T}$$

where  ${\bf L}$  is a diagonal matrix with eigenvalues  $\lambda_i$  and  ${\bf V}$  is the matrix of eigenvectors.

#### Sparse Principal Component Analysis

#### Tobias Bork

ntroduction

PCA

Idea

Mathematical Formulations

Theorems Limits of Usab

orse PCA

arse PCA

Sparsity and Norms

Numerical Soluti Application

urther Analysis

eferences

# Mathematical Formulation using SVD

The principal axis can also be computed via the eigendecomposition of  $\Sigma$ .

$$\Sigma = VLV^T$$

where **L** is a diagonal matrix with eigenvalues  $\lambda_i$  and **V** is the matrix of eigenvectors.

Closely related is the Singular Value Decomposition (SVD)

$$X = UDV^T$$

where **D** is a diagonal matrix with singular values  $d_1, \ldots, d_p$ , **U** a  $n \times p$  and **V** a  $p \times p$  orthogonal matrix.

Sparse Principal Component Analysis

Tobias Bork

ntroduction

PCA

Idea

Mathematical Formulations

Limits of Usabili

arse PCA

Sparsity and Norms

Application

rurtner Analysis

Reference

## PCA as a regression problem

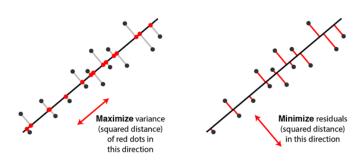


Figure: Two equivalent ways of finding principal axis

Sparse Principal Component Analysis

Tobias Bork

ntroduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability Application

Anthomatical Formulati

Sparsity and Norms

Application

Ť

References

# PCA as a regression problem

### Theorem

Let  $x_i$  be the *i*th row of **X**.

$$\hat{\mathbf{A}}_k = \underset{\mathbf{A}_k}{\operatorname{arg\,min}} \sum_{i=1}^n \left\| x_i - \mathbf{A}_k \mathbf{A}_k^T x_i \right\|^2 + \lambda \sum_{j=1}^k \|\beta_j\|^2$$

subject to 
$$\mathbf{A}_k^T \mathbf{A}_k = \mathbf{I}_{k \times k}$$

Then, if we normalize each column  $\tilde{\mathbf{A}}_k = \begin{bmatrix} \hat{\alpha}_1 \\ \|\hat{\alpha}_1\| \end{bmatrix} \cdots \begin{bmatrix} \hat{\alpha}_k \\ \|\hat{\alpha}_1\| \end{bmatrix}$  we recover the first k principal axis.

#### Sparse Principal Component Analysis

Tobias Bork

itroduction

PCA

Idea

Mathematical Formulations

Theorems Limits of Usability

Application

athematical Formulation

Sparsity and Norms Numerical Solution

Further Analysis

recrements

### Theorems

The Success of PCA is due to the following optimal properties:

- Principal Components sequentially capture the maximum variability
- Principal Components are uncorrelated
- Eckart-Young-Mirsky-Theorem

#### Sparse Principal Component Analysis

Tobias Bork

Theorems

# Limits of Usability

### Sparse Principal Component Analysis

Tobias Bork

Limits of Usability

#### Drawbacks:

- Linear Relationship between variables
- Completeness of data set
- Outliers in data set
- ▶ PCA is inconsistent when  $p \gg n$
- Interpreation of principal axis

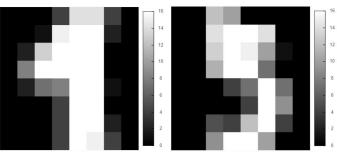
# Application to handwritten digits

### **Data Set Characteristics:**

► Number of Instances: 1797

► Number of Attributes: 64

▶ Attribute Information: 8 × 8 image of integer pixels



(a) Handwritten digit 1

(b) Handwritten digit 5

#### Sparse Principal Component Analysis

#### Tobias Bork

ntroduction

PCA

Idea

Mathematical Form

Theorems

Application

#### DCA

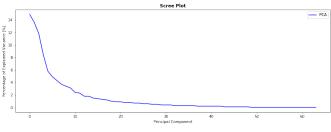
Mathematical Formulation Sparsity and Norms

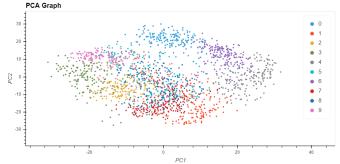
Application

Further Analysis

References

# Application to handwritten digits





#### Sparse Principal Component Analysis

#### Tobias Bork

ntroduction

#### PC.

. . .

idea

Theorems

nits of Usability

#### Application

#### arse PCA

Mathematical Formulation

Numerical Solution

Further Analysis

#### References

# Sparse PCA

**Problem:** Principal Components are hard to interpret **Approach:** Require sparse loadings when performing PCA

$$\max v^T \Sigma v$$

subject to 
$$\|v\|_2 = 1$$
,  $\|v\|_0 \le t$ 

#### Relaxation:

- a regression framework
- a convex semidefinite programming framework
- a generalized power method framework
- ▶ an alternating maximization framework
- forward-backward greedy search and exact methods using branch-and-bound techniques
- ► Bayesian formulation framework

#### Sparse Principal Component Analysis

Tobias Bork

ntroduction

Idea

Markey

Theorems

Limits of Usabilit

Application

Sparse PCA

#### Sparse F CA

Sparsity and Norms

Numerical Solution

Further Analysis

References

Theorems

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms Numerical Solution Application

■ Further Analysis

Appendix

We will use a regression framework to derive sparse PCA.

### **Problem Formulation:**

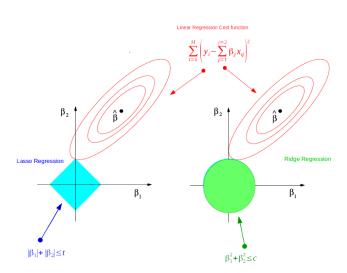
Let  $\mathbf{B} = \left[\beta_1 \middle| \cdots \middle| \beta_k \right]$ . The Sparse PCA Criterion is defined by

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}, \mathbf{B}}{\operatorname{arg \, min}} \sum_{i=1}^{n} \left\| x_{i} - \mathbf{A} \mathbf{B}^{T} x_{i} \right\|^{2} + \lambda \sum_{j=1}^{k} \|\beta_{j}\|^{2} + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_{j}\|_{1}$$

subject to 
$$\mathbf{A}^T \mathbf{A} = I_{k \times k}$$

Then,  $\beta_i$  represent the newly found sparse principal axis and  $Z_i = \mathbf{X}\beta_i$  the sparse principal components.

# Sparsity inducing norms



#### Sparse Principal Component Analysis

#### Tobias Bork

#### ntroduction

#### PC

1.1.

Mathematical Ec

Theorems

imits of Usabilit

#### Sparse PCA

Mathematical Formulation

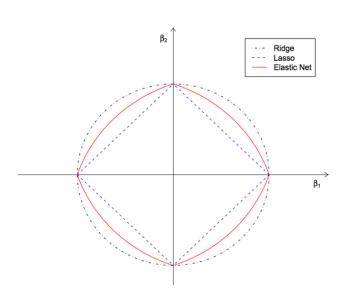
#### Sparsity and Norms

Numerical Soluti

Further Analysis

#### References

# Sparsity inducing norms



#### Sparse Principal Component Analysis

#### Tobias Bork

Introduction

#### PC

Ido

Mathematical Formulatio

Theorems

Application

#### parse PCA

Asthematical Formulation

Sparsity and Norms

Numerical Solut

Further Analysi

#### References

### **Numerical Solution**

#### Sparse Principal Component Analysis

#### Tobias Bork

atroduction

CA

PCA

Idea

Theorems

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution Application

Turtiler Allalysis

References

Appendix

**Problem:** How do we minimize the Sparse PCA criterion?

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A}, \mathbf{B}} \sum_{i=1}^{n} \left\| x_{i} - \mathbf{A} \mathbf{B}^{T} x_{i} \right\|^{2} + \lambda \sum_{j=1}^{k} \|\beta_{j}\|^{2} + \sum_{j=1}^{k} \lambda_{1, j} \|\beta_{j}\|_{1}$$

subject to 
$$\mathbf{A}^T \mathbf{A} = I_{k \times k}$$

▶ **B given A:** For each j, let  $Y^* = \mathbf{X}\alpha_j$ . We minimize over  $\hat{\mathbf{B}} = \begin{bmatrix} \hat{\beta}_1, \dots, \hat{\beta}_k \end{bmatrix}$  by solving k elastic net problems

$$\hat{\beta}_{j} = \operatorname*{min}_{\beta_{j}} \left\| \boldsymbol{Y}^{*} - \boldsymbol{X} \beta_{j} \right\|^{2} + \lambda \left\| \beta_{j} \right\|^{2} + \lambda_{1,j} \left\| \beta_{j} \right\|_{1}$$

Sparse Principal Component Analysis

Tobias Bork

ntroduction

PCA

Idea

Mathematical Formulation

imits of Usability

arse PCA

atnematical Formulation

Numerical Solution

Further Analysis

References

$$\hat{\beta}_{j} = \operatorname*{arg\,min}_{\beta_{j}} \left\| \boldsymbol{Y}^{*} - \boldsymbol{\mathsf{X}} \beta_{j} \right\|^{2} + \lambda \left\| \beta_{j} \right\|^{2} + \lambda_{1,j} \left\| \beta_{j} \right\|_{1}$$

▶ A given B: We can ignore the penalties and minimize

$$\sum_{i=1}^{n} \left\| x_i - \mathbf{A} \mathbf{B}^T x_i \right\|^2 = \left\| \mathbf{X} - \mathbf{X} \mathbf{B} \mathbf{A}^T \right\|_F^2$$

subject to 
$$\mathbf{A}^T \mathbf{A} = I_{k \times k}$$

This problem has an explicit solution which is obtained by computing the SVD of

$$(\mathbf{X}^T\mathbf{X})\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Sparse Principal Component Analysis

Tobias Bork

ntroduction

. .

Idea

Mathematical Formula

Theorems Limits of Usability

Application

#### parse PCA

Mathematical Formulation

#### Numerical Solution

Application

Further Analysis

Reference

#### Sparse Principal Component Analysis

#### Tobias Bork

Numerical Solution

Theorem (Reduced Rank Procrustes Rotation)

Let  $\mathbf{M} \in \mathbb{R}^{n \times p}$  and  $\mathbf{N} \in \mathbb{R}^{n \times k}$  be two matrices. Consider the constrained minimization problem

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{arg min}} \| \mathbf{M} - \mathbf{N} \mathbf{A}^T \|_F^2$$
 subject to  $\mathbf{A}^T \mathbf{A} = I_{k \times k}$ 

Suppose the SVD of  $\mathbf{M}^T \mathbf{N}$  is  $\mathbf{UDV}^T$ , then

$$\hat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$$
.

 $\mathbf{A} \leftarrow \mathbf{V}[1: k]$ , the loadings of the first k ordinary principal components

3: while not converged do

4: Given a fixed  $\mathbf{A} = [\alpha_1, \dots, \alpha_k]$ , solve k elastic net problems

$$\beta_{j} = \mathop{\arg\min}_{\beta} \left\| \left\| \mathbf{X} \alpha_{j} - \mathbf{X} \beta \right\|^{2} + \lambda \left\| \beta \right\|^{2} + \lambda_{1,j} \left\| \beta \right\|_{1}$$

For a fixed  $\mathbf{B} = [\beta_1, \dots, \beta_k]$ , compute the SVD of 5:

$$\mathbf{X}^T \mathbf{X} \mathbf{B} = \mathbf{U} \mathbf{D} \mathbf{V}^T$$

$$A \leftarrow UV^T$$

6: end while

7: 
$$\hat{V}_j \leftarrow \frac{\beta_j^2}{\|\beta_i\|}$$
 for  $j = 1, \dots, k$ 

8: end procedure

#### Sparse Principal Component Analysis

Tobias Bork

Numerical Solution

#### Sparse Principal Component Analysis

#### Tobias Bork

troduction

#### PCA

lde

Mathematical Formulatio

I heorems

Application

#### Sparse PCA

Mathematical Formulation

Non-sized Colonia

Application

Further Analysis

References

# Further Analysis

- ► Consistency theorem for Sparse PCA when  $p \gg n$
- ▶ Efficient implementation when  $p \gg n$
- Computation of ajusted variances
- ▶ Identify differences in Sparse PCA implementations across different platforms (R, Python)
- ► Application to frequency data set

#### Sparse Principal Component Analysis

Tobias Bork

Introduction

PCA

Idea

Mathematical Formulation

Theorems

Application

arse PCA

hamatial Familia

Mathematical Formulation

lumerical Solution

Further Analysis

Furtner Analys

Reterences

### References

#### Sparse Principal Component Analysis

#### Tobias Bork

#### ntroduction

#### PCA

lde

Mathematical Formulation

Limits of Usabil

Application

#### Sparse PCA

Mathematical Formulation

Numerical Solution

Further Analys

### References

# Theorem (Eckart-Young-Mirsky-Theorem)

Let  $\widehat{\mathbf{A}}^* = \mathbf{U}_1 \mathbf{D}_1 \mathbf{V}_1^{\top}$  be the truncated singular value decomposition. Then  $\widehat{\mathbf{A}}^*$  solves the matrix rank approximation problem

$$\min_{\mathrm{rank}(\widehat{\mathbf{A}}) \leq r} \|\mathbf{A} - \widehat{\mathbf{A}}\|_F = \|\mathbf{A} - \widehat{\mathbf{A}}^*\|_F = \sqrt{\sigma_{r+1}^2 + \dots + \sigma_m^2}$$

where  $\sigma_i$  are the singular values of **A**.

#### Sparse Principal Component Analysis

#### Tobias Bork

Introduction

1 0,

Idea

Mathematical Formulation

Limits of Usability

Sparce DCA

parse PCA

lathematical Formulation parsity and Norms

Application

Further Analysis

References

Appendix

Consider a linear regression model with n observations and ppredictors. Let  $Y = (y_1, \dots, y_n)^T$  be the response vector and  $\mathbf{X} = [X_1 | \cdots | X_n]$ .

The linear regression model has the form

$$f(\mathbf{X}) = \beta_0 + \sum_{i=1}^p X_i \beta_i$$

where the  $\beta_i$ 's are unknown coefficients. We define the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2$$

# Ridge Regression

### **Problem Formulation**

$$\hat{\beta}^{\textit{ridge}} = \mathop{\arg\min}_{\beta} \| Y - \mathbf{X}\beta \|_2^2 + \lambda \quad \text{subject to } \| \beta \|_2^2 \leq t$$

or equivalently in Lagrangian Form

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \|\beta\|_2^2 \right\}$$

#### Sparse Principal Component Analysis

#### Tobias Bork

#### ntroduction

#### PCA

Idea

Mathematical Formu

Theorems

Limits of Usability

#### Sparse PCA

Sparsity and Norms

Numerical Solutio

Further Analysis

References

# LASSO Regression

### Problem Formulation

$$\hat{\beta}^{\textit{lasso}} = \mathop{\arg\min}_{\beta} \left\| Y - \mathbf{X} \beta \right\|_2^2 + \lambda \quad \text{subject to} \ \left\| \beta \right\|_1 \leq t$$

or equivalently in Lagrangian Form

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \|\beta\|_1 \right\}$$

#### Sparse Principal Component Analysis

#### Tobias Bork

Appendix

The elastic net penalty is a convex combination of the ridge and lasso penalties.

### Problem Formulation:

$$\hat{\beta}^{\textit{en}} = \operatorname*{arg\,min}_{\beta}(1+\lambda_{2})\left\{ \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|^{2} + \lambda_{2} \left\| \boldsymbol{\beta} \right\|_{2}^{2} + \lambda_{1} \left\| \boldsymbol{\beta} \right\|_{1} \right\}$$

Given a fixed  $\lambda_2$ , the LARS-EN algorithm (Zou and Hastie 2005) efficiently solves the elastic net problem for all  $\lambda_1$  with the computational cost of a single least squares fit.

### RUNTIME SPCA ALGORITHM

#### Sparse Principal Component Analysis

#### Tobias Bork

#### ntroduction

#### PC

Idea

Mathematical Formulation

Limits of Usabili

Application

#### oparse r CA

Mathematical Formulation

Numerical Solution

Application

Further Analys

#### References

Mathematical Formulation Theorems

Application

Sparse PCA

Mathematical Formulation

Numerical Solution

Further Analysis

Reference

Appendix

Let  $\mathbf{A}_{p \times k} = [\alpha_1, \dots, \alpha_k]$  and  $\mathbf{B}_{p \times k} = [\beta_1, \dots, \beta_k]$ . Since  $\mathbf{A}$  is orthonomal, let  $\mathbf{A}_\perp$  be any orthonormal matrix such that  $[\mathbf{A}; \mathbf{A}_\perp]$  is  $p \times p$  orthonormal. Then we ran reformulate the problem

$$\begin{split} \sum_{i=1}^{n} \left\| \mathbf{x}_{i} - \mathbf{A} \mathbf{B}^{T} \mathbf{x}_{i} \right\|^{2} &= \left\| \mathbf{X} - \mathbf{X} \mathbf{B} \mathbf{A}^{T} \right\|_{F}^{2} \\ &= \left\| \mathbf{X} \mathbf{A}_{\perp} \right\|_{F}^{2} + \left\| \mathbf{X} \mathbf{A} - \mathbf{X} \mathbf{B} \right\|_{F}^{2} \\ &= \left\| \mathbf{X} \mathbf{A}_{\perp} \right\|_{F}^{2} + \sum_{j=1}^{k} \left\| \mathbf{X} \alpha_{j} - \mathbf{X} \beta_{j} \right\|^{2} \end{split}$$