

Sparse Principal Component Analysis for Frequency Data

Tobias Bork

Institute for Numerical Simulation

December 10, 2019



Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Introduction

PCA

Idea
Mathematical Formulations
Theorems
Limits of Usability
Application

Sparse PCA

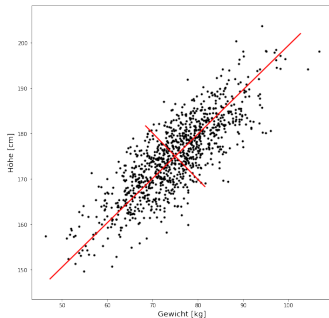
Mathematical Formulation
Sparsity and Norms
Numerical Solution
Application
Further Analysis

References

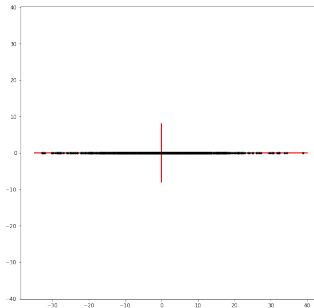
Appendix

- ▶ **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



(a) Finding principal axis on a data set



(b) Linear projection of data to first principal axis

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

$$v_1 = \arg \max_{\|v\|_2=1} \text{Var}[\mathbf{X}v] = \arg \max_{\|v\|_2=1} v^T \mathbf{\Sigma} v$$

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

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

$$v_1 = \arg \max_{\|v\|_2=1} \text{Var}[\mathbf{X}v] = \arg \max_{\|v\|_2=1} v^T \mathbf{\Sigma} v$$

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

We compute the following principal axis successively

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

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

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

$$v_1 = \arg \max_{\|v\|_2=1} \text{Var}[\mathbf{X}v] = \arg \max_{\|v\|_2=1} v^T \mathbf{\Sigma} v$$

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

We compute the following principal axis successively

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

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

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Mathematical Formulation using SVD

The principal axis can also be computed via the eigendecomposition of Σ .

$$\Sigma = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

where \mathbf{L} is a diagonal matrix with eigenvalues λ_i and \mathbf{V} is the matrix of eigenvectors.

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Mathematical Formulation using SVD

The principal axis can also be computed via the eigendecomposition of Σ .

$$\Sigma = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

where \mathbf{L} is a diagonal matrix with eigenvalues λ_i and \mathbf{V} is the matrix of eigenvectors.

Closely related is the Singular Value Decomposition (SVD)

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

where \mathbf{D} is a diagonal matrix with singular values d_1, \dots, d_p , \mathbf{U} a $n \times p$ and \mathbf{V} a $p \times p$ orthogonal matrix.

PCA as a regression problem

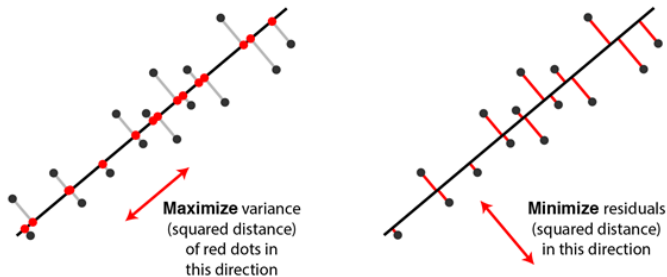


Figure: Two equivalent ways of finding principal axis

PCA as a regression problem

Theorem

Let x_i be the i th row of \mathbf{X} .

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

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

Then, if we normalize each column $\tilde{\mathbf{A}}_k = \left[\frac{\hat{\alpha}_1}{\|\hat{\alpha}_1\|} \mid \dots \mid \frac{\hat{\alpha}_k}{\|\hat{\alpha}_1\|} \right]$
we recover the first k principal axis.

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

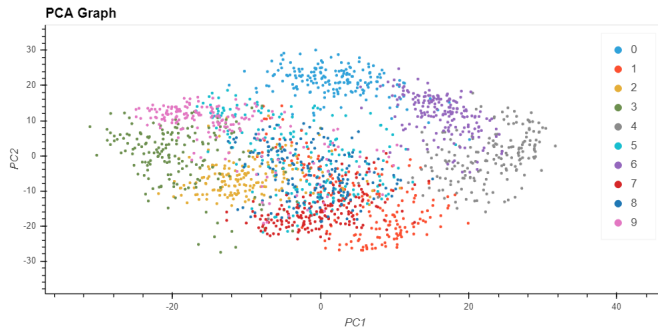
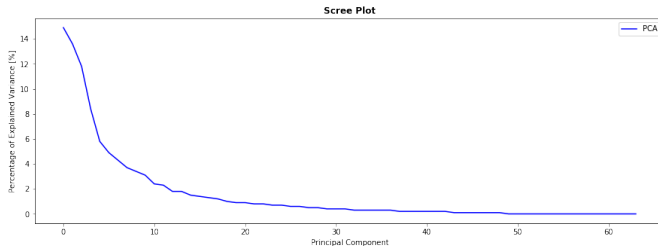
The Success of PCA is due to the following optimal properties:

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

Drawbacks:

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

Application to handwritten digits



Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Problem: Principal Components are hard to interpret

Approach: Require sparse loadings when performing PCA

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

$$\text{subject to } \|v\|_2 = 1, \quad \|v\|_0 \leq t$$

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

We will use a regression framework to derive sparse PCA.

Problem Formulation:

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

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg \min_{\mathbf{A}, \mathbf{B}} \sum_{i=1}^n \left\| x_i - \mathbf{A} \mathbf{B}^T x_i \right\|^2 + \lambda \sum_{j=1}^k \|\beta_j\|^2 + \sum_{j=1}^k \lambda_{1,j} \|\beta_j\|_1$$

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

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Sparsity inducing norms

Introduction

PCA

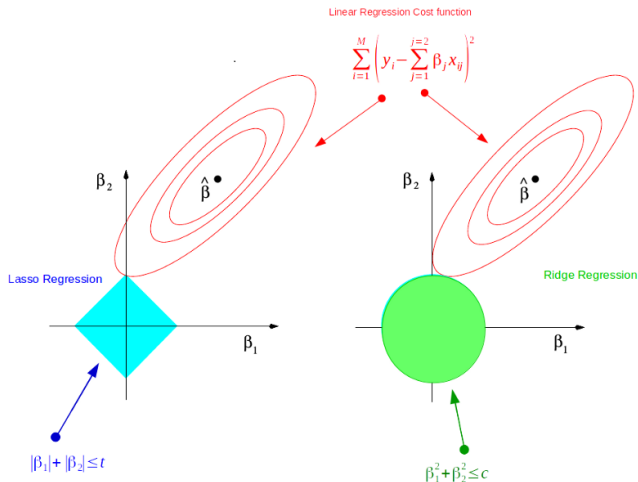
- Idea
- Mathematical Formulations
- Theorems
- Limits of Usability
- Application

Sparse PCA

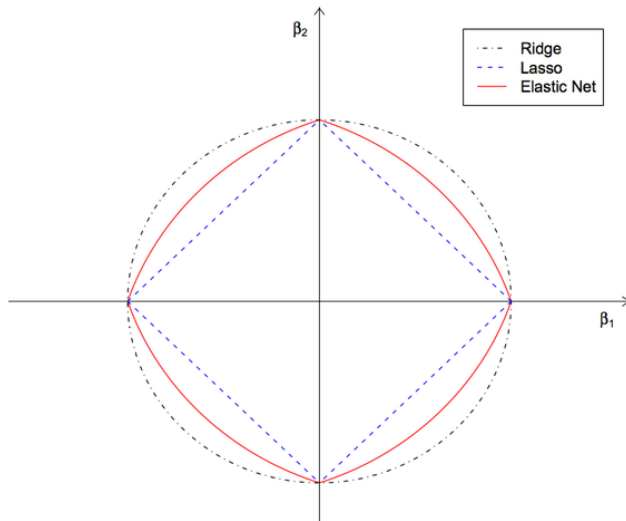
- Mathematical Formulation
- Sparsity and Norms**
- Numerical Solution
- Application
- Further Analysis

References

Appendix



Sparsity inducing norms



Introduction

PCA

- Idea
- Mathematical Formulations
- Theorems
- Limits of Usability
- Application

Sparse PCA

- Mathematical Formulation
- Sparsity and Norms**
- Numerical Solution
- Application
- Further Analysis

References

Appendix

Problem: How do we minimize the Sparse PCA criterion?

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg \min_{\mathbf{A}, \mathbf{B}} \sum_{i=1}^n \left\| \mathbf{x}_i - \mathbf{A} \mathbf{B}^T \mathbf{x}_i \right\|^2 + \lambda \sum_{j=1}^k \|\beta_j\|^2 + \sum_{j=1}^k \lambda_{1,j} \|\beta_j\|_1$$

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

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

$$\hat{\beta}_j = \arg \min_{\beta_j} \|Y^* - \mathbf{X}\beta_j\|^2 + \lambda \|\beta_j\|^2 + \lambda_{1j} \|\beta_j\|_1$$

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

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

$$\hat{\beta}_j = \arg \min_{\beta_j} \|\mathbf{Y}^* - \mathbf{X}\beta_j\|^2 + \lambda \|\beta_j\|^2 + \lambda_{1j} \|\beta_j\|_1$$

- **A given B:** We can ignore the penalties and minimize

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

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

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

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

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

$$\hat{\mathbf{A}} = \arg \min_{\mathbf{A}} \left\| \mathbf{M} - \mathbf{N} \mathbf{A}^T \right\|_F^2 \quad \text{subject to } \mathbf{A}^T \mathbf{A} = \mathbf{I}_{k \times k}$$

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

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

Algorithm 1 General SPCA Algorithm

- 1: **procedure** SPCA(A, B)
2: $\mathbf{A} \leftarrow \mathbf{V}[1:k]$, 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

$$\beta_j = \arg \min_{\beta} \|\mathbf{X}\alpha_j - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 + \lambda_{1,j} \|\beta\|_1$$

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

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

$$\mathbf{A} \leftarrow \mathbf{U} \mathbf{V}^T$$

- 6: **end while**
7: $\hat{\mathbf{V}}_j \leftarrow \frac{\beta_j}{\|\beta_j\|}$ for $j = 1, \dots, k$
8: **end procedure**
-

[Introduction](#)[PCA](#)[Idea](#)[Mathematical Formulations](#)[Theorems](#)[Limits of Usability](#)[Application](#)[Sparse PCA](#)[Mathematical Formulation](#)[Sparsity and Norms](#)[Numerical Solution](#)[Application](#)[Further Analysis](#)[References](#)[Appendix](#)

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix



Hastie et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition. 2009,. Springer-Verlag New York.



Zou et al. *Sparse Principal Component Analysis*. Journal of Computational and Graphical Statistics. Volume 15. Number 2. 2006.
<https://doi.org/10.1198/106186006X113430>.



Foucart, Rauhut. *A Mathematical Introduction to Compressive Sensing*. 1st edition. 2013. Birkhäuser Basel.

Sparsity and Norms

Further Analysis

References

Theorem (Eckart-Young-Mirsky-Theorem)

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

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

where σ_i are the singular values of \mathbf{A} .

Linear Regression

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

The linear regression model has the form

$$f(\mathbf{X}) = \beta_0 + \sum_{j=1}^p X_j \beta_j$$

where the β_j 's are unknown coefficients.

We define the residual sum of squares

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

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Problem Formulation

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

or equivalently in Lagrangian Form

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

Problem Formulation

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \quad \text{subject to } \|\beta\|_1 \leq t$$

or equivalently in Lagrangian Form

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

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

Problem Formulation:

$$\hat{\beta}^{en} = \arg \min_{\beta} (1 + \lambda_2) \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}$$

Given a fixed λ_2 , the LARS-EN algorithm (Zou and Hastie 2005) efficiently solves the elastic net problem for all λ_1 with the computational cost of a single least squares fit.

Introduction

PCA

Idea

Mathematical Formulations

Theorems

Limits of Usability

Application

Sparse PCA

Mathematical Formulation

Sparsity and Norms

Numerical Solution

Application

Further Analysis

References

Appendix

Sparse PCA Algorithm Complexity

► Case 1: $n > p$

- Compute the $p \times p$ matrix $\hat{\Sigma} = \mathbf{X}^T \mathbf{X}$ which requires np^2 operations (Same $\hat{\Sigma}$ is used in each step)
- Compute $\mathbf{X}^T \mathbf{X} \mathbf{B}$ which costs $p^2 k$ operations
- Compute the SVD of $\mathbf{X}^T \mathbf{X} \mathbf{B}$ which costs $\mathcal{O}(pk^2)$
- Compute each elastic net solution which requires at most $\mathcal{O}(p^3)$

Since $k \leq p$, the total computation cost is at most $np^2 + m\mathcal{O}(p^3)$ where m is the number of iterations before convergence

Sparse PCA Algorithm Complexity

► Case 2: $p \gg n$

The trick of using $\hat{\Sigma}$ is no longer applicable, because $\hat{\Sigma}$ is a huge $p \times p$ matrix in this case. The most consuming step is solving each elastic net, whose cost is of order $\mathcal{O}(pnJ + J^3)$ for a positive λ , where J is the number of nonzero coefficients.

Generally speaking the total cost is of order $mk\mathcal{O}(pnJ + J^3)$, which can be expensive for large J and p . Fortunately, there exists a special SPCA algorithm for efficiently dealing with $p \gg n$ data.

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 orthonormal, let \mathbf{A}_\perp be any orthonormal matrix such that $[\mathbf{A}; \mathbf{A}_\perp]$ is $p \times p$ orthonormal. Then we can reformulate the problem

$$\begin{aligned} \sum_{i=1}^n \left\| \mathbf{x}_i - \mathbf{A} \mathbf{B}^T \mathbf{x}_i \right\|^2 &= \left\| \mathbf{X} - \mathbf{X} \mathbf{B} \mathbf{A}^T \right\|_F^2 \\ &= \left\| \mathbf{X} \mathbf{A}_\perp \right\|_F^2 + \left\| \mathbf{X} \mathbf{A} - \mathbf{X} \mathbf{B} \right\|_F^2 \\ &= \left\| \mathbf{X} \mathbf{A}_\perp \right\|_F^2 + \sum_{j=1}^k \left\| \mathbf{X} \alpha_j - \mathbf{X} \beta_j \right\|^2 \end{aligned}$$

Introduction

PCA

- Idea
- Mathematical Formulations
- Theorems
- Limits of Usability
- Application

Sparse PCA

- Mathematical Formulation
- Sparsity and Norms
- Numerical Solution
- Application
- Further Analysis

References

Appendix