Sparse Principal Component Analysis
for Frequency Data

Tobias Bork

Institute for Numerical Simulation

December 10, 2019

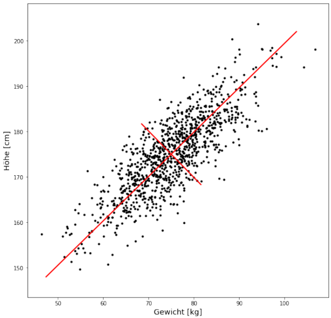
UNIVERSITÄT BONN

Fraunhofer SCAI

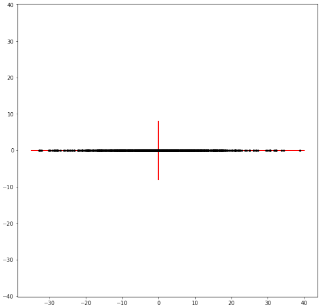
Dimensionality Reduction

- **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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\mathbf{\Sigma}$.

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

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

Mathematical Formulation using SVD

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

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

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

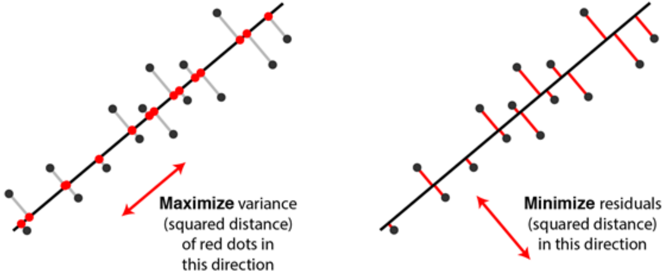


Figure: Two equivalent ways of finding principal axis

PCA as a regression problem

Theorem

Let x_i be the ith row of $\mathbf{X}$.

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

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

Then, if we normalize each column $\tilde{\mathbf{A}}_k = \left[\frac{\hat{\alpha}_1}{\|\hat{\alpha}_1\|} \middle| \cdots \middle| \frac{\hat{\alpha}_k}{\|\hat{\alpha}_1\|} \right]$

we recover the first k principal axis.

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

Limits of Usability

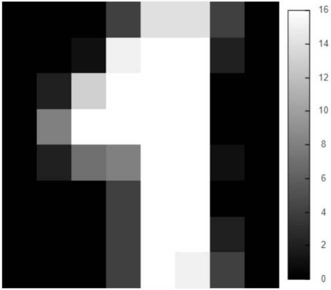
Drawbacks:

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

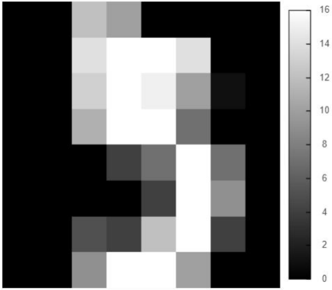
Application to handwritten digits

Data Set Characteristics:

- **Number of Instances:** 1797
- **Number of Attributes:** 64
- **Attribute Information:** 8×8 image of integer pixels

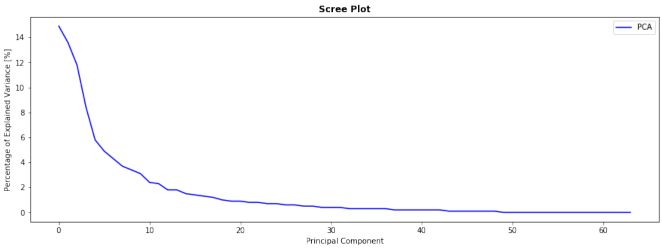


(a) Handwritten digit 1



(b) Handwritten digit 5

Application to handwritten digits



Sparse PCA

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

Mathematical Formulation

We will use a regression framework to derive sparse PCA.

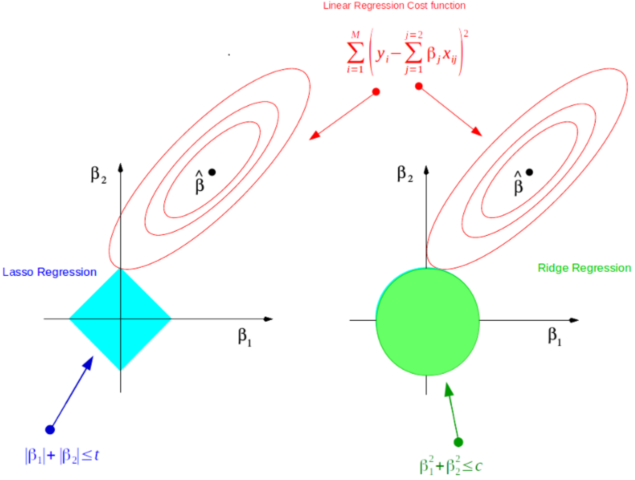
Problem Formulation:
Let $\mathbf{B} = \left[\beta_1 \middle| \cdots \middle| \beta_k\right]$. The Sparse PCA Criterion is defined by

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

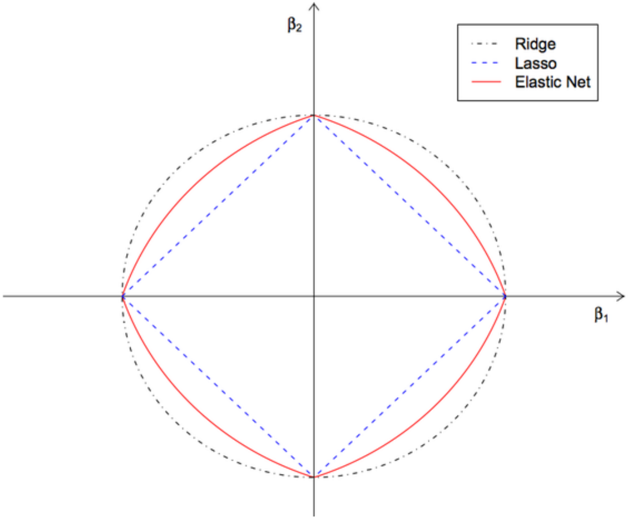
$$\text{subject to } \mathbf{A}^T\mathbf{A} = I_{k\times k}$$

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

Sparsity inducing norms



Sparsity inducing norms



Numerical Solution

Problem: How do we minimize the Sparse PCA criterion?

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

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

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

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

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

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

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

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

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

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

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

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*

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

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

```
procedure SPCA(A, B)
    A ← V[, 1 : k], the loadings of the first k ordinary principal components
    while not converged do
        Given a fixed A = [α_1, ..., α_k], solve k elastic net problems
```

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

```
        For a fixed B = [β_1, ..., β_k], compute the SVD of
```

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

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

```
    end while
    V̂_j ← β_j / ||β_j||  for j = 1, ..., k
end procedure
```

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

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

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

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

Linear Regression

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

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

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

Ridge Regression

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

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

LASSO Regression

Problem Formulation

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

or equivalently in Lagrangian Form

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

Elastic Net

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.

RUNTIME SPCA ALGORITHM

Numerical Solution

Let $\mathbf{A}_{p\times k} = [\alpha_1, \ldots, \alpha_k]$ and $\mathbf{B}_{p\times k} = [\beta_1, \ldots, \beta_k]$. 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

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