Sparse Principal Component Analysis

for Frequency Data

Tobias Bork

Institute for Numerical Simulation

December 6, 2019

Dimensionality Reduction

- What are DR methods?
- Why are DR methods important / what is the goal of DR methods?
- When to use DR methods?

Central Idea

- Reduce dimensionality of a data set while retaining as much information as possible
- extract the most important information from the data table;
- compress the size of the data set by keeping only this important information;
- simplify the description of the data set; and
- analyze the structure of the observations and the variables
- first component has most variance

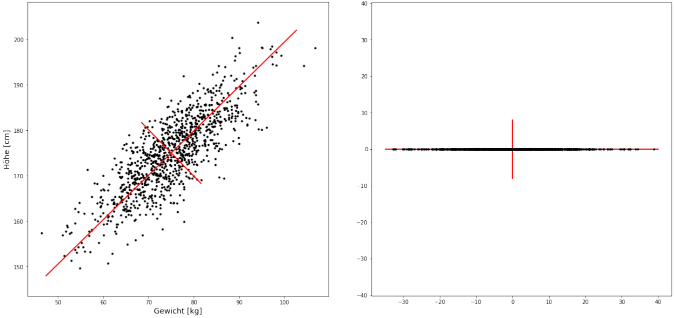


Figure: Some description of the plots

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 $\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X}$ is the sample covariance matrix. Then we can find the following principal axis successively

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

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

The new, transformed variables are defined by $Z_i = \mathbf{X} v_i$

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

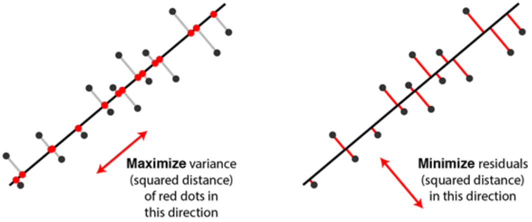
$$\mathbf{\Sigma} = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

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 $d_1, \ldots, d_p$, $\mathbf{U}$ a $n \times p$ and $\mathbf{V}$ a $p \times p$ orthogonal matrix.

PCA as a regression problem



Suppose we want to extract the first k principal axis.

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

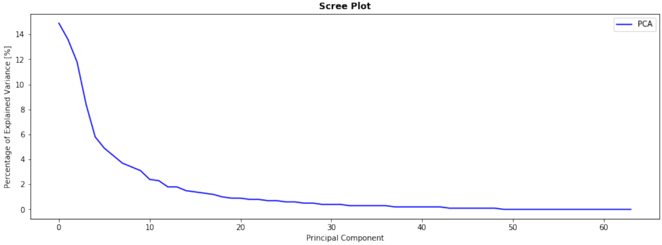
$$\text{subject to } \mathbf{V}_k^T \mathbf{V}_k = I_{k \times k}$$

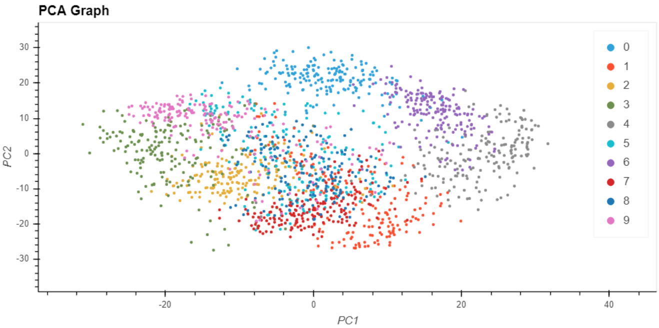
Succes of PCA is due to the following two important optimal properties

1. Principal Components sequentially capture the maximum variability (among the columns of X, thus guaranteeing minimal information loss)
2. Principal Components are uncorrelated, (so we can talk about one principal component without referring to others)

- Linear Relationship between variables
- Correlation of variables
- Completeness of data set
- Outliers
- Number of variables p ¿¿ Number of Samples n (Inconsistency Theorem)
- Interpreation of principal axis

Application to handwritten digits





Algorithm 1 General SPCA Algorithm

procedure $\mathrm{SPCA}(A, B)$
  $\mathbf{A} \leftarrow \mathbf{V}[, 1\colon k]$, the loadings of the first k ordinary principal components
  while not converged **do** ▷ Definiere Abbruchkriterium
    Given a fixed $\mathbf{A} = [\alpha_1, \ldots, \alpha_k]$, solve the elastic net problem

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

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

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

    $\mathbf{A} \leftarrow \mathbf{U}\mathbf{V}^T$
  end while
  $\hat{V}_j = \frac{\beta_j}{\|\beta_j\|}$ for $j = 1, \ldots, k$
end procedure

References

Beamer Paket
http://latex-beamer.sourceforge.net/

User's Guide to the Beamer

DANTE e.V. http://www.dante.de