Sparse Principal Component Analysis

for Frequency Data

Tobias Bork

Institute for Numerical Simulation

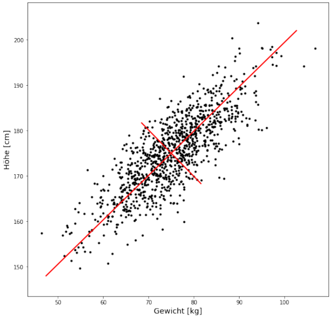
December 9, 2019

Dimensionality Reduction

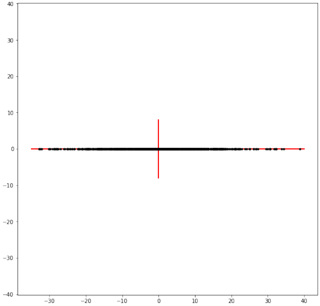
- **Problems** in high dimensions:
  - Time and storage space
  - Multi-collinearity
  - Visualizing data set
  - Curse of dimensionality
- **Idea:** Reduce the number of variables while preserving structure in the data
- **Approach:** Feature selection methods
- **Approach:** Feature extraction methods

Central Idea

- Reduce dimensionality while retaining as much information as possible
- Sequentially identify principal axis of greatest variability
- Represent data set regarding identified principal axis
- Linearly project the data to a space of fewer dimensions
- Yields natural order on principal components



(a) Finding principal axis on a data set



(b) Linear projection of data to a space of fewer dimensions

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.
We compute the following principal axis successively

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

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

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

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

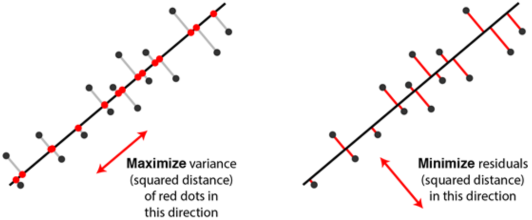
$$\mathbf{\Sigma} = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

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.

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

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

where x_i is the ith row of $\mathbf{X}$

Theorems

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)

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

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

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

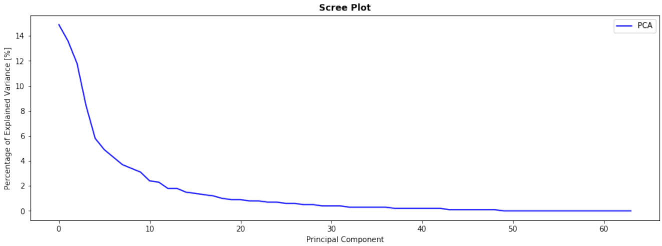
Theorem

PCA is inconsistent for $p \gg n$.

Limits of Usability

- Linear Relationship between variables
- Correlation of variables
- Completeness of data set
- Outliers
- Inconsistency theorem in $p \gg n$ case
- Interpreation of principal axis

Application to handwritten digits



Linear Regression

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

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = \| Y - \mathbf{X}\beta \|_2^2$$

$$\hat{\beta} = \arg\min_{\beta} RSS(\beta)$$

Ridge Regression

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \| Y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{subject to } \|\beta\|_2^2 \leq t$$

or equivalently in Lagrangian Form

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

LASSO Regression

Elastic Net

Sparse PCA

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

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

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

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. We add

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

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

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