Big Data Management & Analytics

EXERCISE 8 – TEXT PROCESSING, PCA

21st of December, 2015

Sabrina Friedl
LMU Munich

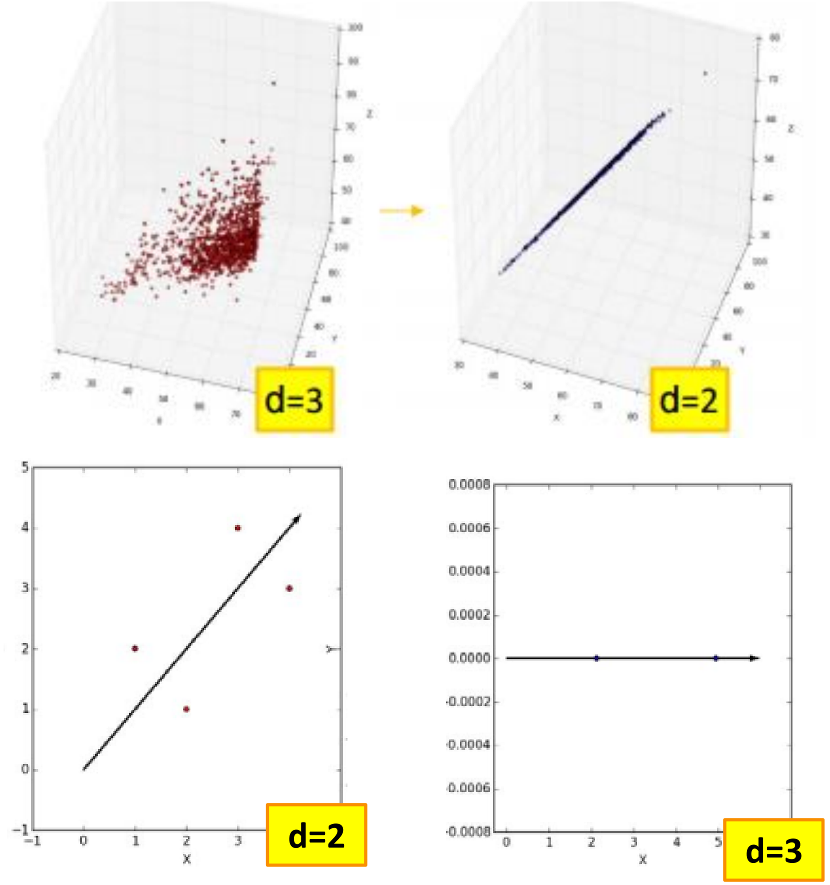
Product Component Analysis (PCA)

REVISION AND EXAMPLE

Goals of PCA

Find a lower-dimensional representation of data to:

- Detect hidden correlations
- Remove (summarize redundant, irrelevant or noisy features
- Fascilitate interpretation and visualization (actually visualization is possible only for few dimensions)
- Make storage and processing of data easier



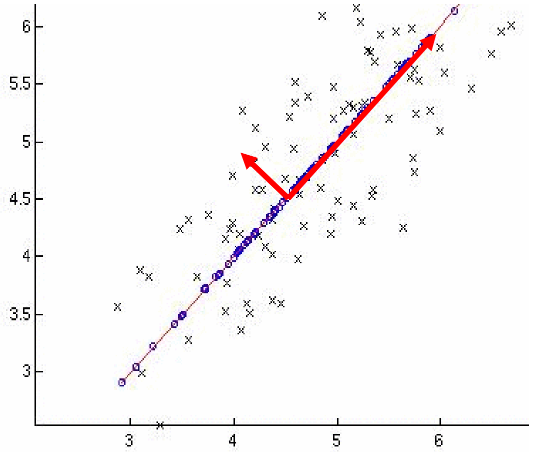
Idea of PCA

A good data representation retains the main differences between data points but eliminates irrelevant variances

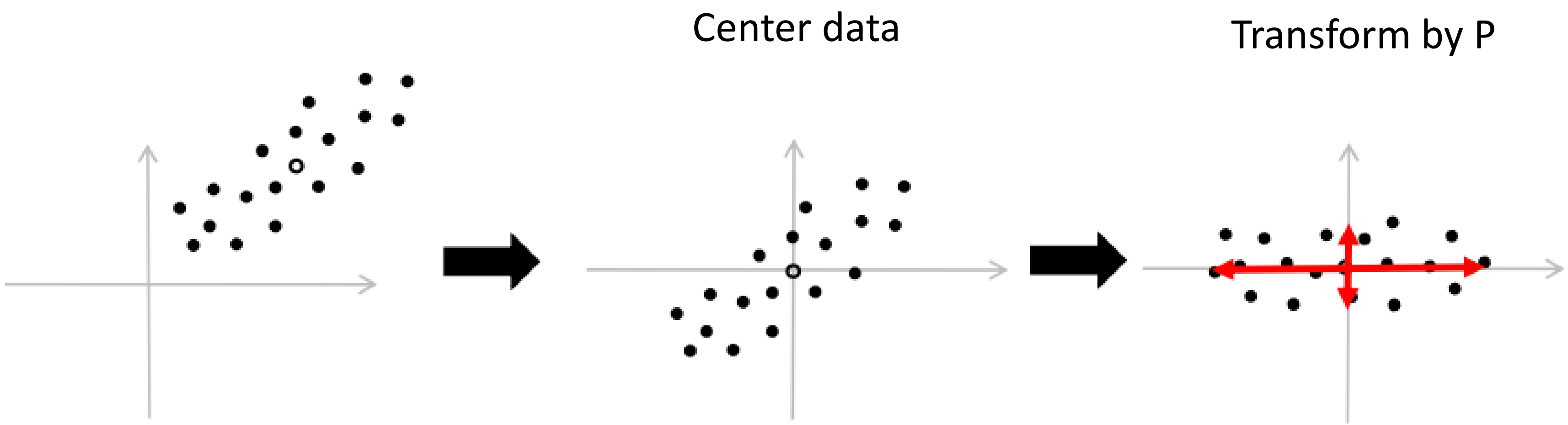
- Given matrix X: n data points with d dimensions (features)
- Find k directions (linear combinations of dimensions) with highest variance = principal components: $v_1, v_2, \dots v_k$
- Project data points onto these directions
- General Form: $XP = Y$

(n x d) * (d x k) = (n x k)

X = raw data matrix
P = (v_1, v_2,... v_k) transformation matrix
Y = k-dimensional representation of X



PCA – Graphical Intuition



Rotate the data space in a way that the direction with the largest variance is placed on an axis of the data space

How to get Principal Components?

Calculate the eigenvalues and eigenvectors of the covariance matrix

Sigma here is the name of the matrix, not the sum symbol!

$$\Sigma_D = \begin{pmatrix} VAR(X_1) & \cdots & COV(X_1, X_d) \\ \vdots & \ddots & \vdots \\ COV(X_d, X_1) & \cdots & VAR(X_d) \end{pmatrix}$$

$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$VAR(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = COV(X,X)$$

Describes the pairwise correlation between all features

For a **centralized data** matrix X with $\mu = 0$ we can calculate the covariance matrix as:

$$\frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = \Sigma_D = \begin{pmatrix} VAR(X_1) & \cdots & COV(X_1, X_d) \\ \vdots & \ddots & \vdots \\ COV(X_d, X_1) & \cdots & VAR(X_d) \end{pmatrix}$$

Eigenvalues and Eigenvectors

Let A be a square $d\ x\ d$ matrix. If there exists a real scalar λ and a $d\ x\ 1$ vector $v\ \neq 0$, such that:

$$Av = \lambda v,$$

then λ is called an **eigenvalue** of A and v is the associated **eigenvector**.

How to find eigenvalues / eigenvectors of A?

- Solving the equation: $\det(A - \lambda\, \mathrm{I}_{dxd}) = 0$ yields the eigenvalues
- For each eigenvalue λ_i, we find its eigenvector by solving the system of equations $(A - \lambda_i\, I_{dxd})\, v_i = 0$

Dimension Reduction

For n dimensions of X we get n eigevalues and eigenvectors. The transformation matrix is then constructed by putting the eigenvectors as columns into a matrix: $\mathrm{T} = (v_1, v_2, \dots v_n)$

Eigendecomposition: $\Sigma = T\Lambda T^T$

Σ = covariance matrix
$T = (v_1, v_2, \dots v_n)$ transformation matrix
Λ = diagonalised matrix with eigenvalues on diagonal

To get a k-dimensional representation Y of (centered) data X we take only the first k eigenvectors (principal components) of T and call this matrix P.

We calculate: $\boldsymbol{XP} = \boldsymbol{Y}$

To transform back: $\mathrm{Z} = YP^T$

PCA – Summary of Steps

1. Center the data X : $x_i - \mu_i$
2. Calculate the covariance-matrix: $\Sigma = \frac{1}{n} X^T X$
3. Calculate the eigenvalues and eigenvectors of Σ
   - Calculate eigenvalues λ by finding the zeros of the characteristic polynomial: $\det(\Sigma - \lambda I)$
   - Calculate the eigenvectors by solving $(\Sigma - \lambda I)v = 0$
4. Select the k eigenvectors with the biggest eigenvalues and create $\mathrm{P} = (v_1, v_2, \dots v_k)$
5. Transform the original (n x d) matrix X to a (n x k) representation: $XP = Y$

Useful links

- KDD II script: http://www.dbs.ifi.lmu.de/Lehre/KDD_II/WS1516/skript/KDD2-2-HDData.DimensionalityReduction.pdf

- A tutorial about PCA: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf