

Dimensionality Reduction

- Many dimensions are often interdependent (correlated);

We can:

- Reduce the dimensionality of problems;
- Transform interdependent coordinates into significant and independent ones;

Principal Component Analysis

Principal Component Analysis -- PCA (also called Karhunen-Loeve transformation)

- **PCA** transforms the original input space into a lower dimensional space, by constructing dimensions that are linear combinations of the given features;
- The objective is to consider **independent** dimensions along which data have **largest variance** (i.e., greatest variability);

Principal Component Analysis -- PCA

- **PCA** involves a linear algebra procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called **principal components**;
- The first principal component accounts for as much of the variability in the data as possible;
- Each succeeding component (orthogonal to the previous ones) accounts for as much of the remaining variability as possible.

Principal Component Analysis -- PCA

- So: PCA finds n linearly transformed components s_1, s_2, \dots, s_n so that they explain the maximum amount of variance;
- We can define PCA in an intuitive way using a recursive formulation:

Principal Component Analysis -- PCA

- Suppose data are first centered at the origin (i.e., their mean is $\mathbf{0}$);
- We define the direction of the first principal component, say \mathbf{w}_1 , as follows

$$\mathbf{w}_1 = \arg \max_{\|\mathbf{w}\|=1} E[(\mathbf{w}^T \mathbf{x})^2]$$

where \mathbf{w}_1 is of the same dimensionality q as the data vector \mathbf{x}

- Thus: the first principal component is the projection on the direction along which the variance of the projection is maximized.

Principal Component Analysis -- PCA

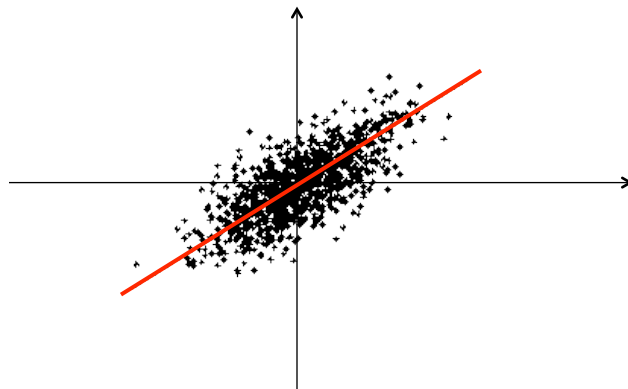
- Having determined the first $k-1$ principal components, the k -th principal component is determined as the principal component of the data residual:

$$\mathbf{w}_k = \arg \max_{\|\mathbf{w}\|=1} E\{[\mathbf{w}^T (\mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x})]^2\}$$

- The principal components are then given by:

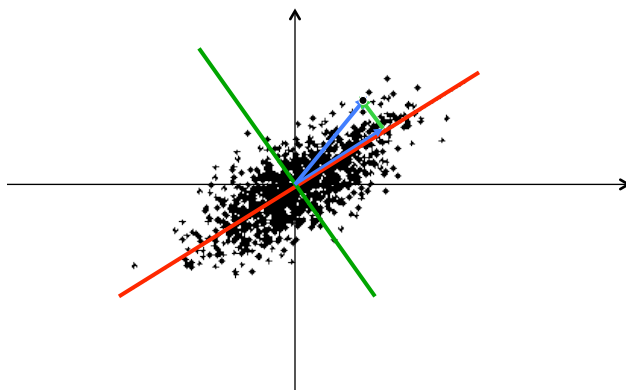
$$s_i = \mathbf{w}_i^T \mathbf{x}$$

Simple illustration of PCA



First principal component of a two-dimensional data set.

Simple illustration of PCA



Second principal component of a two-dimensional data set.

PCA – Geometric interpretation

Basically:

PCA rotates the data (centered at the origin) in such a way that the maximum variability is visible (i.e., aligned with the axes.)

PCA – How to compute the principal components

Let \mathbf{w} be the direction of the first principal component, with $\|\mathbf{w}\| = 1$

$s_i = \mathbf{w}^T \mathbf{x}_i$ is the projection of \mathbf{x}_i along \mathbf{w}

$$\bar{s} = \frac{1}{N} \sum_{i=1}^N s_i = \frac{1}{N} \sum_{i=1}^N \mathbf{w}^T \mathbf{x}_i$$

Variance of data along \mathbf{w} :

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N (s_i - \bar{s})^2 &= \\ \frac{1}{N} \sum_{i=1}^N \left(\mathbf{w}^T \mathbf{x}_i - \frac{1}{N} \sum_{j=1}^N \mathbf{w}^T \mathbf{x}_j \right)^2 & \end{aligned}$$

PCA – How to compute the principal components

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N (s_i - \bar{s})^2 &= \\ \frac{1}{N} \sum_{i=1}^N \left(\mathbf{w}^T \mathbf{x}_i - \frac{1}{N} \sum_{j=1}^N \mathbf{w}^T \mathbf{x}_j \right)^2 &= \\ \frac{1}{N} \sum_{i=1}^N \left[\mathbf{w}^T \left(\mathbf{x}_i - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j \right) \right]^2 &= \\ \frac{1}{N} \sum_{i=1}^N \left[\mathbf{w}^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right]^2 &= \\ \frac{1}{N} \sum_{i=1}^N \left[\mathbf{w}^T (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w} \right] &= \\ \mathbf{w}^T \left[\frac{1}{N} \sum_{i=1}^N [(\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T] \right] \mathbf{w} &= \mathbf{w}^T \Sigma \mathbf{w} \end{aligned}$$

Sample covariance matrix

PCA – How to compute the principal components

Thus : the variance of data along direction \mathbf{w} can be written as

$$\mathbf{w}^T \Sigma \mathbf{w}$$

Our objective is to find \mathbf{w} such that

$$\mathbf{w} = \arg \max_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w}$$

with the constraint $\mathbf{w}^T \mathbf{w} = 1$

By introducing one Lagrange multiplier λ , we obtain the following unconstrained optimization problem

$$\mathbf{w} = \arg \max_{\mathbf{w}} \left[\mathbf{w}^T \Sigma \mathbf{w} - \lambda (\mathbf{w}^T \mathbf{w} - 1) \right]$$

Setting $\frac{\partial}{\partial \mathbf{w}} = 0$ gives: $2\Sigma \mathbf{w} - 2\lambda \mathbf{w} = 0$

That is : $\Sigma \mathbf{w} = \lambda \mathbf{w}$

Our problem is reduced to an eigenvalue problem

PCA – How to compute the principal components

Thus : the variance of data along direction \mathbf{w} can be written as

$$\mathbf{w}^T \Sigma \mathbf{w}$$

Our objective is to find \mathbf{w} such that

$$\mathbf{w} = \arg \max_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w}$$

with the constraint $\mathbf{w}^T \mathbf{w} = 1$

By introducing one Lagrange multiplier λ , we obtain the following unconstrained optimization problem

$$\mathbf{w} = \arg \max_{\mathbf{w}} \left[\mathbf{w}^T \Sigma \mathbf{w} - \lambda (\mathbf{w}^T \mathbf{w} - 1) \right]$$

Setting $\frac{\partial}{\partial \mathbf{w}} = 0$ gives: $2\Sigma \mathbf{w} - 2\lambda \mathbf{w} = 0$

That is : $\Sigma \mathbf{w} = \lambda \mathbf{w}$

The solution \mathbf{w} is the eigenvector of Σ corresponding to the largest eigenvalue λ .

PCA -- Summary

- The computation of the w_i is accomplished by solving an eigenvalue problem for the sample **covariance** matrix (assuming data have 0 mean):

$$\Sigma = E[\mathbf{x} \mathbf{x}^T]$$

- The eigenvector associated with the **largest eigenvalue** corresponds to the **first** principal component; the eigenvector associated with the **second largest** eigenvalue corresponds to the **second** principal component; and so on...
- Thus: The w_i are the eigenvectors of Σ that correspond to the n largest eigenvalues of Σ

PCA -- In practice

- The basic goal of PCA is to reduce the dimensionality of the data. Thus, one usually chooses:

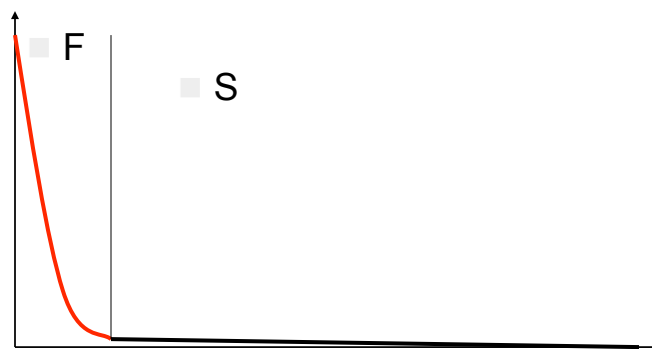
$$n \ll q$$

- But how do we select the number of components n ?

Determining the number of components

- Plot the eigenvalues – each eigenvalue is related to the amount of variation explained by the corresponding axis (eigenvector);
- If the points on the graph tend to level out (show an “elbow” shape), these eigenvalues are usually close enough to zero that they can be ignored.
- In general: Limit the variance accounted for.

Critical information lies in low dimensional subspaces



- A typical eigenvalue spectrum and its division into two orthogonal subspaces

