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 $s_1, s_2, \cdots, s_n$ 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 ***0***);
- We define the direction of the first principal component, say w_1, as follows

$$w_1 = \arg \max_{\|w\|=1} E[(w^T x)^2]$$

where w_1 is of the same dimensionality q as the data vector 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

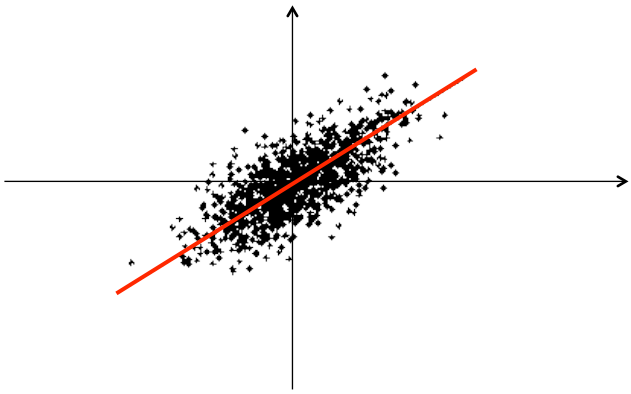
- Having determined the first *k-1* principal components, the *k*-th principal component is determined as the principal component of the data residual:

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

- The principal components are then given by:

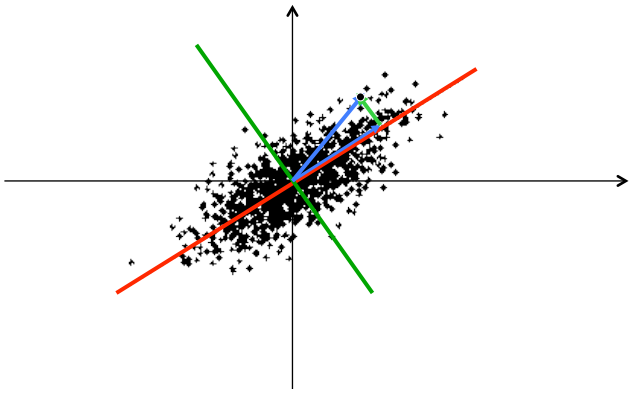
$$s_i = \boldsymbol{w}_i^T \boldsymbol{x}$$

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

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

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

Variance of data along $\boldsymbol{w}$:

$$\frac{1}{N}\sum_{i=1}^{N}\left(s_i - \bar{s}\right)^2 =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^T \boldsymbol{x}_i - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{w}^T \boldsymbol{x}_j\right)^2$$

PCA – How to compute the principal components

$$\frac{1}{N}\sum_{i=1}^{N}\left(s_i - \bar{s}\right)^2 =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^T \boldsymbol{x}_i - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{w}^T \boldsymbol{x}_j\right)^2 =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{w}^T\left(\boldsymbol{x}_i - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j\right)\right]^2 =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{w}^T\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\right]^2 =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{w}^T\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^T \boldsymbol{w}\right] =$$

$$\boldsymbol{w}^T\left\{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^T\right]\right\}\boldsymbol{w} = \boldsymbol{w}^T \Sigma \boldsymbol{w}$$

Sample covariance matrix

PCA – How to compute the principal components

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

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

Our objective is to find $\boldsymbol{w}$ such that

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

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

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

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

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

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

Our problem is reduced to an eigenvalue problem

PCA – How to compute the principal components

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

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

Our objective is to find $\boldsymbol{w}$ such that

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

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

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

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

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

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

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

PCA -- Summary

- The computation of the $\boldsymbol{w}_i$ is accomplished by solving an eigenvalue problem for the sample covariance matrix (assuming data have $\boldsymbol{0}$ mean):

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

- 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 $\boldsymbol{w}_i$ 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:

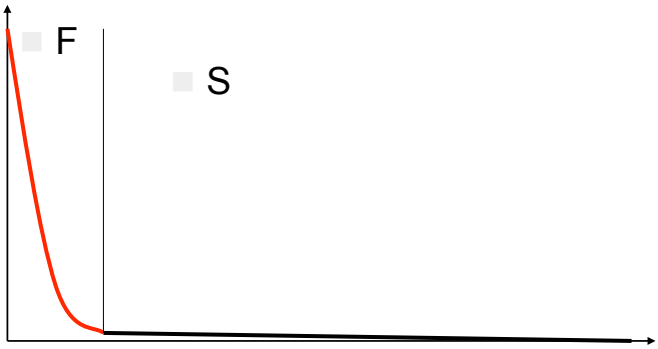
$$n << q$$

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

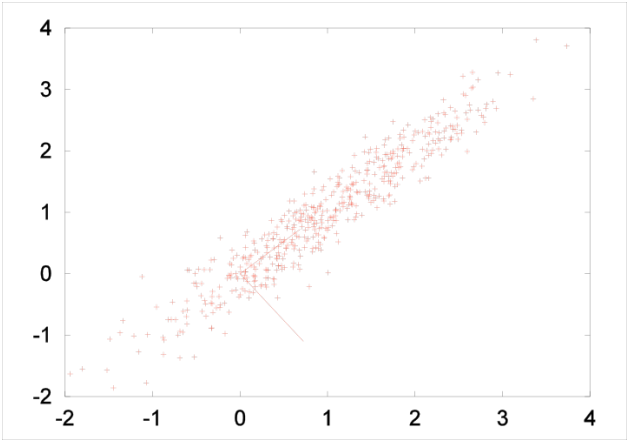
Determining the number of components

- 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



$\lambda_1 = 1.98, \quad \lambda_2 = 0.05$

Determining the number of components

$\boldsymbol{x}_i \in \Re^q, \quad i = 1, \cdots, N$

$\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_q$: q eigenvectors (principal component directions)

$\|\boldsymbol{w}_i\| = 1$ (the $\boldsymbol{w}_i$s are orthonormal vectors)

Representation of $\boldsymbol{x}_i$ in eigenvector space :

$$\boldsymbol{y}_i = \left(\boldsymbol{w}_1^T \boldsymbol{x}_i\right)\boldsymbol{w}_1 + \left(\boldsymbol{w}_2^T \boldsymbol{x}_i\right)\boldsymbol{w}_2 + \cdots + \left(\boldsymbol{w}_q^T \boldsymbol{x}_i\right)\boldsymbol{w}_q$$

Suppose we retain the first k principal components :

$$\boldsymbol{y}_i^k = \left(\boldsymbol{w}_1^T \boldsymbol{x}_i\right)\boldsymbol{w}_1 + \left(\boldsymbol{w}_2^T \boldsymbol{x}_i\right)\boldsymbol{w}_2 + \cdots + \left(\boldsymbol{w}_k^T \boldsymbol{x}_i\right)\boldsymbol{w}_k$$

Then :

$$\boldsymbol{y}_i - \boldsymbol{y}_i^k = \left(\boldsymbol{w}_{k+1}^T \boldsymbol{x}_i\right)\boldsymbol{w}_{k+1} + \cdots + \left(\boldsymbol{w}_q^T \boldsymbol{x}_i\right)\boldsymbol{w}_q$$

Determining the number of components

$$\left(y_i - y_i^k\right)^T \left(y_i - y_i^k\right) =$$

$$\left[\left(w_{k+1}^T x_i\right) w_{k+1} + \cdots + \left(w_q^T x_i\right) w_q\right]^T \left[\left(w_{k+1}^T x_i\right) w_{k+1} + \cdots + \left(w_q^T x_i\right) w_q\right] =$$

$$w_{k+1}^T \left(w_{k+1}^T x_i\right)^2 w_{k+1} + \cdots + w_q^T \left(w_q^T x_i\right)^2 w_q =$$

(note $w_i^T w_j = 0 \ \forall i \neq j$ since w_i and w_j are orthogonal vectors)

$$\left(w_{k+1}^T x_i\right)^2 w_{k+1}^T w_{k+1} + \cdots + \left(w_q^T x_i\right)^2 w_q^T w_q =$$

$$\left(w_{k+1}^T x_i\right)^2 + \cdots + \left(w_q^T x_i\right)^2 =$$

$$\left(w_{k+1}^T x_i\right)\left(x_i^T w_{k+1}\right) + \cdots + \left(w_q^T x_i\right)\left(x_i^T w_q\right) =$$

$$w_{k+1}^T \left(x_i x_i^T\right) w_{k+1} + \cdots + w_q^T \left(x_i x_i^T\right) w_q$$

Determining the number of components

$$\frac{1}{N} \sum_{i=1}^{N} \left(y_i - y_i^k\right)^T \left(y_i - y_i^k\right) =$$

Mean square error

$$\frac{1}{N} \sum_{i=1}^{N} \left[w_{k+1}^T \left(x_i x_i^T\right) w_{k+1} + \cdots + w_q^T \left(x_i x_i^T\right) w_q\right] =$$

$$w_{k+1}^T \left[\frac{1}{N} \sum_{i=1}^{N} \left(x_i x_i^T\right)\right] w_{k+1} + \cdots + w_q^T \left[\frac{1}{N} \sum_{i=1}^{N} \left(x_i x_i^T\right)\right] w_q =$$

$$w_{k+1}^T \Sigma w_{k+1} + \cdots + w_q^T \Sigma w_q$$

We have: $\Sigma w_{k+1} = \lambda_{k+1} w_{k+1}, \cdots, \Sigma w_q = \lambda_q w_q$

Thus:

$$w_{k+1}^T \Sigma w_{k+1} + \cdots + w_q^T \Sigma w_q =$$

$$w_{k+1}^T \lambda_{k+1} w_{k+1} + \cdots + w_q^T \lambda_q w_q =$$

$$\lambda_{k+1} + \cdots + \lambda_q$$

Determining the number of components

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \boldsymbol{y}_i^k\right)^T\left(\boldsymbol{y}_i - \boldsymbol{y}_i^k\right) = \lambda_{k+1} + \cdots \lambda_q$$

⟹ The mean square error of the truncated representation is equal to the sum of the remaining eigenvalues.

In general: choose *k* so that 90-95% of the variance of the data is captured.