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 $\boldsymbol{w}_1$, as follows

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

  where $\boldsymbol{w}_1$ is of the same dimensionality $\boldsymbol{q}$ as the data vector $\boldsymbol{x}$
- Thus: the first principal component is the projection on the direction along which the variance of the projection is maximized.

Principal Component Analysis -- PCA

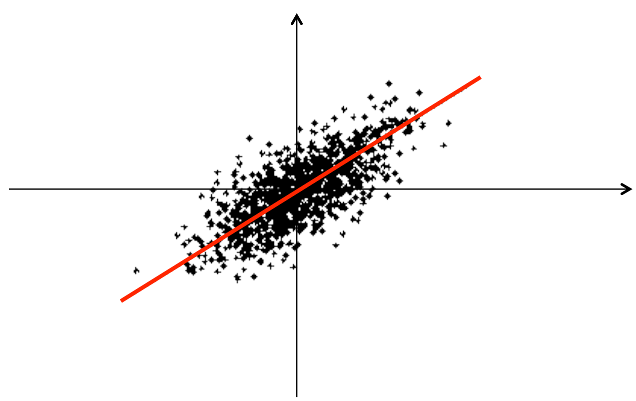
- Having determined the first *k-1* principal components, the *k*-th principal component is determined as the principal component of the data residual:

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

- The principal components are then given by:

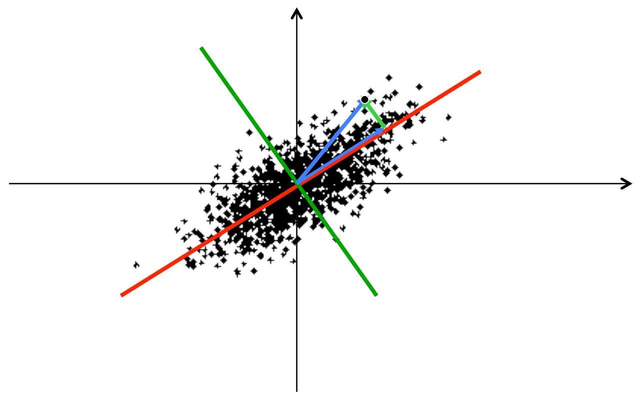
$$s_i = \boldsymbol{w}_i^T \boldsymbol{x}$$

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 $\boldsymbol{w}$:

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

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

PCA – How to compute the principal components

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

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

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

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

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

$$\boldsymbol{w}^T \underbrace{\boxed{\left\{ \frac{1}{N}\sum_{i=1}^{N} \left[(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \right] \right\}}}_{\text{Sample covariance matrix}} \boldsymbol{w} = \boldsymbol{w}^T \Sigma \boldsymbol{w}$$

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

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

The solution 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:

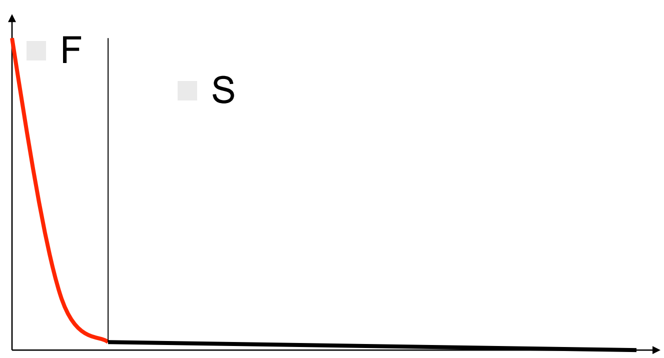
$$n << q$$

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

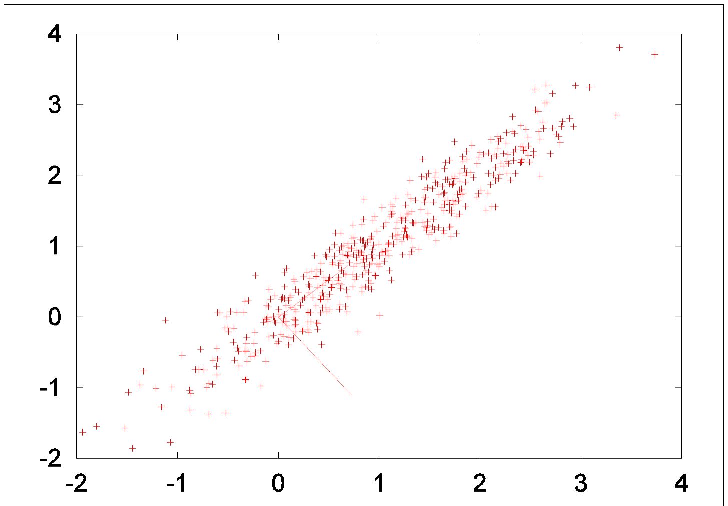
Determining the number of components

- 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

$x_i \in \Re^q, \quad i = 1, \cdots, N$

$w_1, w_2, \cdots, w_q$: q eigenvectors (principal component directions)

$\|w_i\| = 1$ (the w_is are orthonormal vectors)

Representation of x_i in eigenvector space :

$$y_i = (w_1^T x_i) w_1 + (w_2^T x_i) w_2 + \cdots + (w_q^T x_i) w_q$$

Suppose we retain the first k principal components :

$$y_i^k = (w_1^T x_i) w_1 + (w_2^T x_i) w_2 + \cdots + (w_k^T x_i) w_k$$

Then :

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

Determining the number of components

$$(y_i - y_i^k)^T (y_i - y_i^k) =$$

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

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

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

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

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

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

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

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

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{w}_{k+1}^T\left(\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{w}_{k+1} + \cdots + \mathbf{w}_q^T\left(\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{w}_q\right] =$$

$$\mathbf{w}_{k+1}^T\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i\mathbf{x}_i^T\right)\right]\mathbf{w}_{k+1} + \cdots + \mathbf{w}_q^T\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i\mathbf{x}_i^T\right)\right]\mathbf{w}_q =$$

$$\mathbf{w}_{k+1}^T\Sigma\mathbf{w}_{k+1} + \cdots + \mathbf{w}_q^T\Sigma\mathbf{w}_q$$

We have: $\Sigma\mathbf{w}_{k+1} = \lambda_{k+1}\mathbf{w}_{k+1}, \cdots, \Sigma\mathbf{w}_q = \lambda_q\mathbf{w}_q$

Thus:

$$\mathbf{w}_{k+1}^T\Sigma\mathbf{w}_{k+1} + \cdots + \mathbf{w}_q^T\Sigma\mathbf{w}_q =$$

$$\mathbf{w}_{k+1}^T\lambda_{k+1}\mathbf{w}_{k+1} + \cdots + \mathbf{w}_q^T\lambda_q\mathbf{w}_q =$$

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

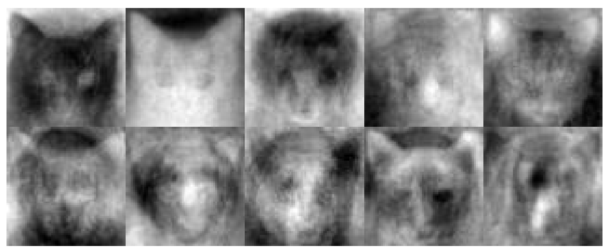
Determining the number of components

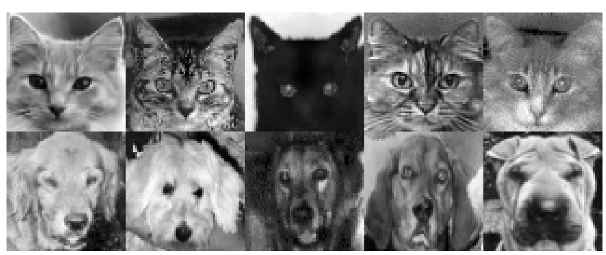
$$\boxed{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i^k\right)^T\left(y_i - y_i^k\right) = \lambda_{k+1} + \cdots \lambda_q}$$

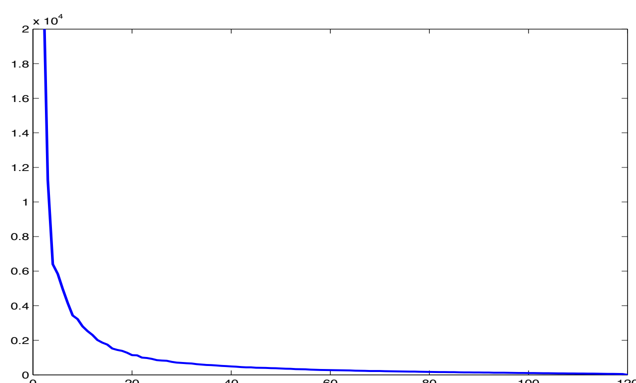
⇒ The mean square error of the truncated representation is equal to the sum of the remaining eigenvalues.

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

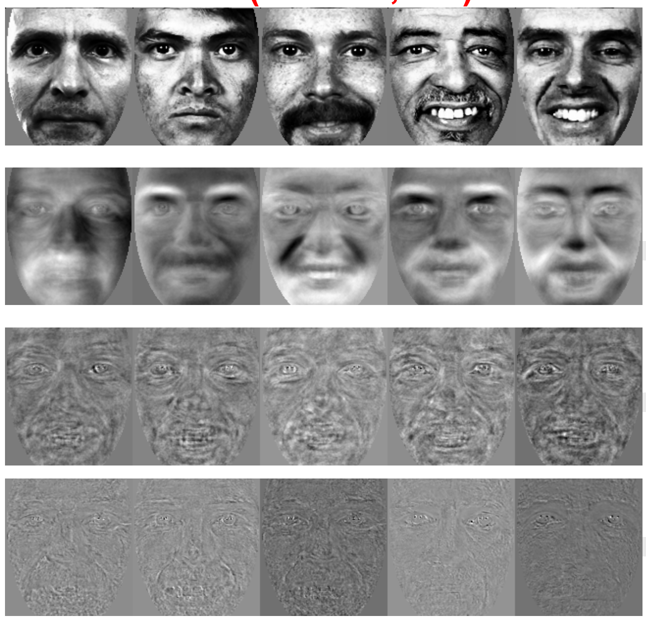
Cat and Dog faces: black & white 64 by 64 images (dim=4096)



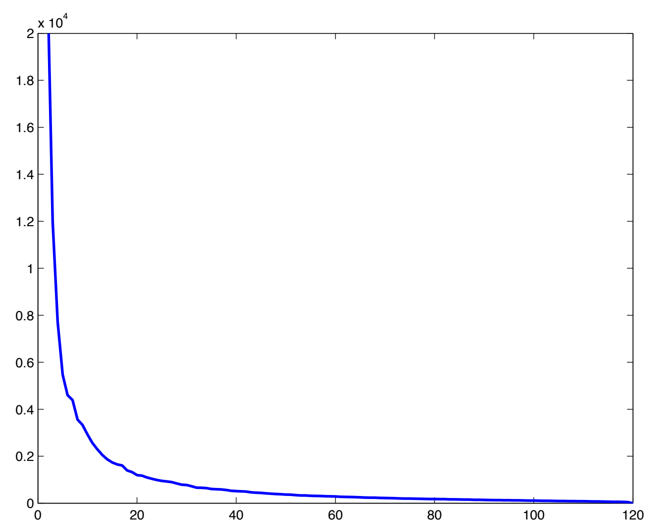




FERET face images; image size is 150 by 130 (dim=19,500)



- original
- 1 ~ 5
- 101~105
- 191~195



Advantages of PCA

- Optimal linear dimensionality reduction technique in the mean-square sense;
- Reduce the curse-of-dimensionality;
- Computational overhead of subsequent processing stages is reduced;
- Noise may be reduced;
- A projection into a subspace of a very low dimensionality, e.g. two dimensions, is useful for visualizing the data.