PCA -- In practice

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

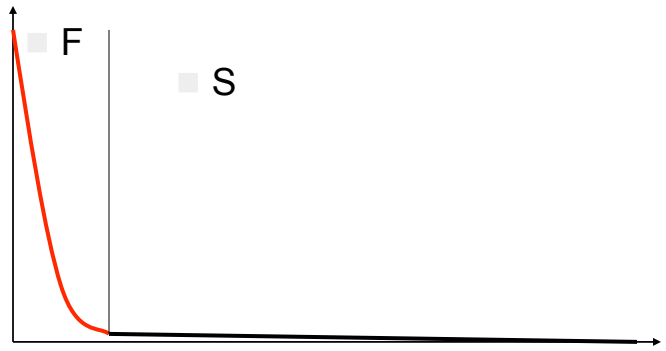
$$n << q$$

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

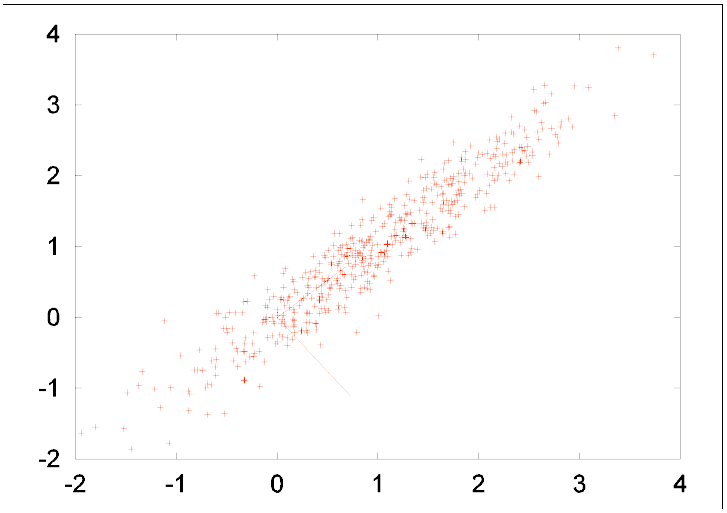
Determining the number of components

- 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

$$\left(\boldsymbol{y}_i - \boldsymbol{y}_i^k\right)^T \left(\boldsymbol{y}_i - \boldsymbol{y}_i^k\right) =$$

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

$$\boldsymbol{w}_{k+1}^T \left(\boldsymbol{w}_{k+1}^T \boldsymbol{x}_i\right)^2 \boldsymbol{w}_{k+1} + \cdots + \boldsymbol{w}_q^T \left(\boldsymbol{w}_q^T \boldsymbol{x}_i\right)^2 \boldsymbol{w}_q =$$

(note $\boldsymbol{w}_i^T \boldsymbol{w}_j = 0 \ \forall i \neq j$ since $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are orthogonal vectors)

$$\left(\boldsymbol{w}_{k+1}^T \boldsymbol{x}_i\right)^2 \boldsymbol{w}_{k+1}^T \boldsymbol{w}_{k+1} + \cdots + \left(\boldsymbol{w}_q^T \boldsymbol{x}_i\right)^2 \boldsymbol{w}_q^T \boldsymbol{w}_q =$$

$$\left(\boldsymbol{w}_{k+1}^T \boldsymbol{x}_i\right)^2 + \cdots + \left(\boldsymbol{w}_q^T \boldsymbol{x}_i\right)^2 =$$

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

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

Determining the number of components

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

Mean square error

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

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

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

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

Thus:

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

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

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

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.