CS 559 Machine Learning

Principal Component Analysis

Yue Ning
Department of Computer Science
Stevens Institute of Technology

Sharing examples or materials

- Course design is a collaborative effort that benefits from collective wisdom.
- **Extra credits:** Share examples or materials that you find helpful on Canvas

Eigenvalues and Eigenvectors

- For an $n \times n$ square matrix A, $\mathbf{e}$ is an eigenvector with eigenvalue λ if:

$$A\mathbf{e} = \lambda\mathbf{e}$$

or

$$(A - \lambda I)\mathbf{e} = 0$$

- If $(A - \lambda I)$ is invertible, the only solution is $\mathbf{e} = \mathbf{0}$ (trivial).

Eigenvalues and Eigenvectors

- For non-trivial solutions, the determinant of this matrix follows:

$$det(A - \lambda I) = 0$$

- This is called the "characteristic polynomial"
- Solutions are not unique because if $\mathbf{e}$ is an eigenvector $a\mathbf{e}$ is also an eigenvector.

Eigenvalues and Eigenvectors - simple example

- For a 2×2 matrix A:

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- Given:

$$det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

- We get: $a_{11}a_{22} - a_{12}a_{21} - (a_{11} + a_{22})\lambda + \lambda^2 = 0 \rightarrow$
  $1 \cdot 4 - 2 \cdot 2 - (1 + 4)\lambda + \lambda^2 = 0$
- Solutions are $\lambda = 0$ and $\lambda = 5$.

Eigenvalues and Eigenvectors - simple example

- The eigenvector for the first eigenvalue, $\lambda = 0$ is:

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1x + 2y \\ 2x + 4y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is $x = 2, y = -1$
- The eigenvector for the second eigenvalue, $\lambda = 5$ is:

$$\begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -4x + 2y \\ 2x - 1y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is $x = 1, y = 2$

Video on Eigenvalues and Eigenvectors

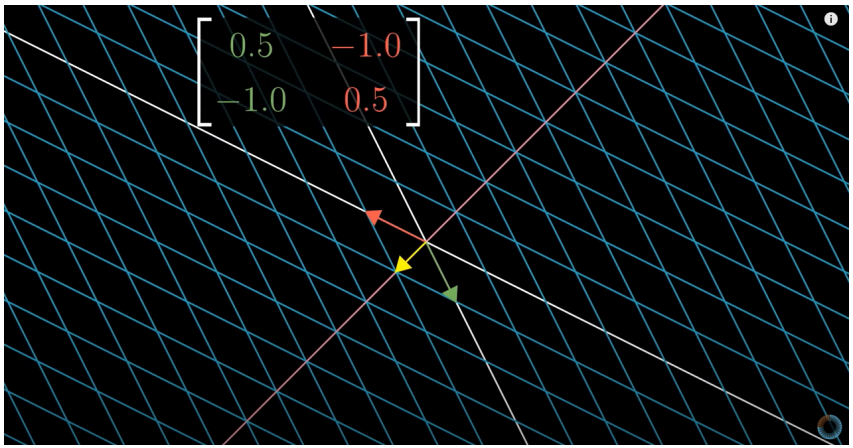


Figure: Eigenvectors and eigenvalues from YouTube

Eigenvalues and Eigenvectors - properties

- The product of the eigenvalues is the determinant of A: $det(A)$
- The sum of the eigenvalues $=$ trace(A)
- The eigenvectors are pairwise orthogonal

Dimensionality Reduction

- Many dimensions are often interdependent (correlated);
- Solution 1: reduce the dimensionality of problems;
- Solution 2: transform interdependent coordinates into significant and independent ones;

Principal Component Analysis

It is 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 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

- So: PCA finds n linearly transformed components:

$$s_1, s_2, \ldots, s_n$$

  so that they explain the maximum amount of variance;
- We can define PCA in an intuitive way using a recursive formulation:

Principal Component Analysis

- 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 d 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

- 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 Principal Component Analysis

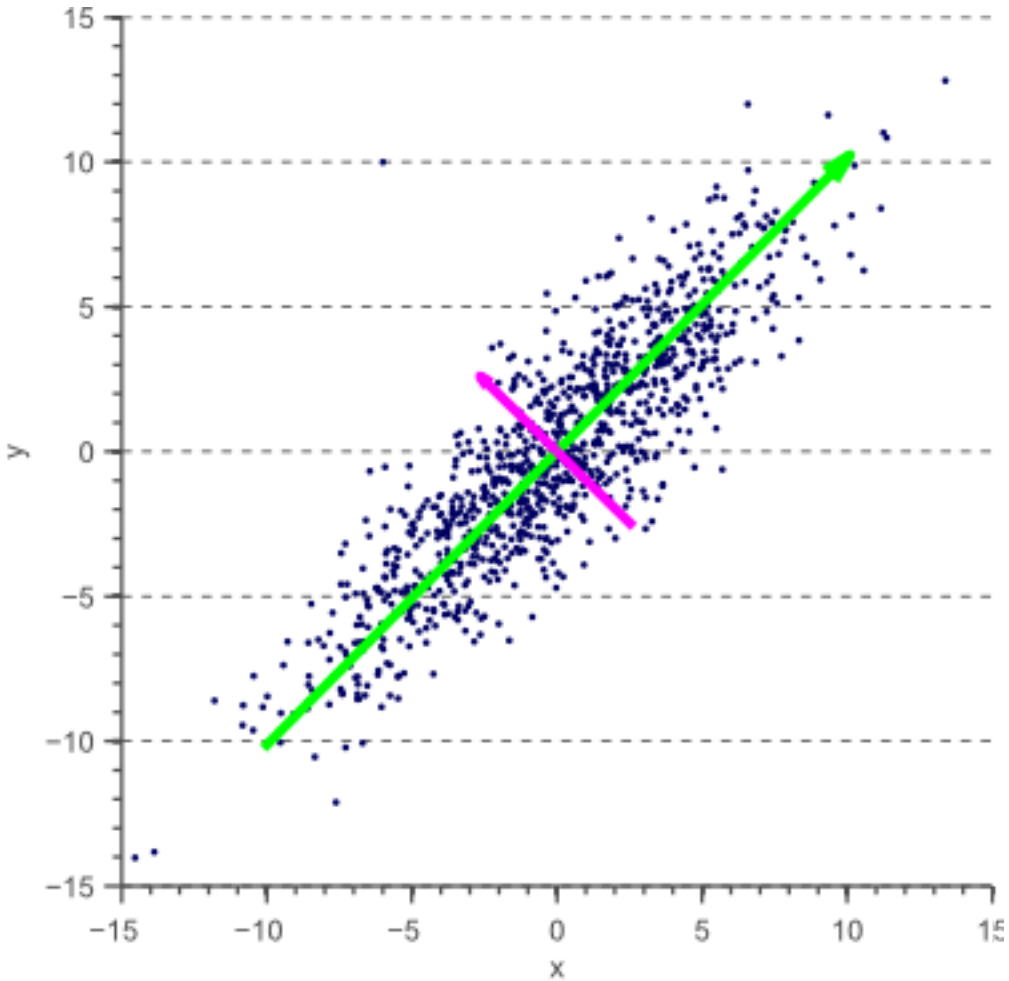


Figure: Green: first principal component of a two-dimensional dataset; Magenta: second principal component

PCA - Geometric interpretation

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 $\mathbf{w}$:

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

PCA - How to compute the principal components

$$\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2 = \ldots = \mathbf{w}^T \Sigma \mathbf{w}$$

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:

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

- Setting $\frac{\partial}{\partial \mathbf{w}} = 0$, gives $2\Sigma\mathbf{w} - 2\lambda\mathbf{w} = 0$
- That is $\Sigma\mathbf{w} = \lambda\mathbf{w}$ (reduced to an eigenvalue problem)

PCA - How to compute the principal components

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

$$\Sigma \mathbf{w} = \lambda \mathbf{w}$$

PCA - Summary

- The computation of the $\mathbf{w}_i$ 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 $\mathbf{w}_i$ are the eigenvectors of Σ that correspond to the i largest eigenvalues of Σ.

PCA - In practice

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

$$n \ll d$$

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

Determining the number of component

- 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.

Determining the number of component

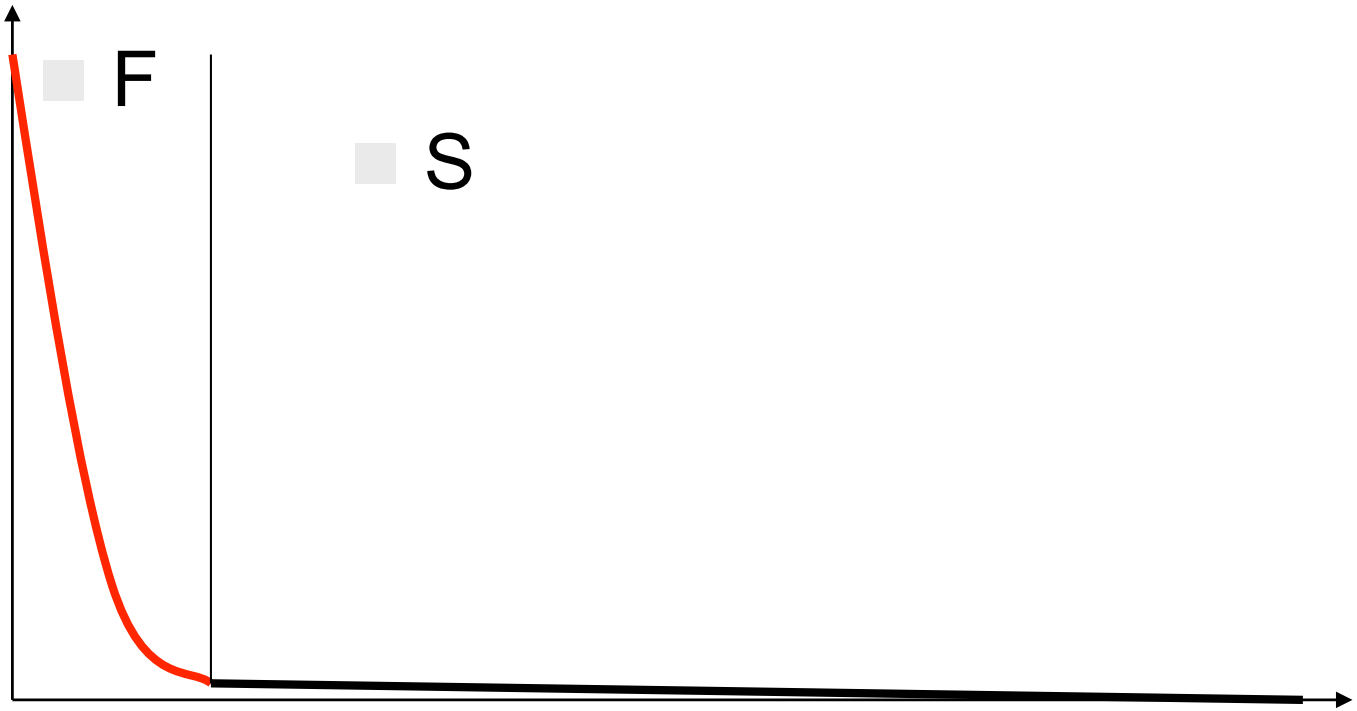
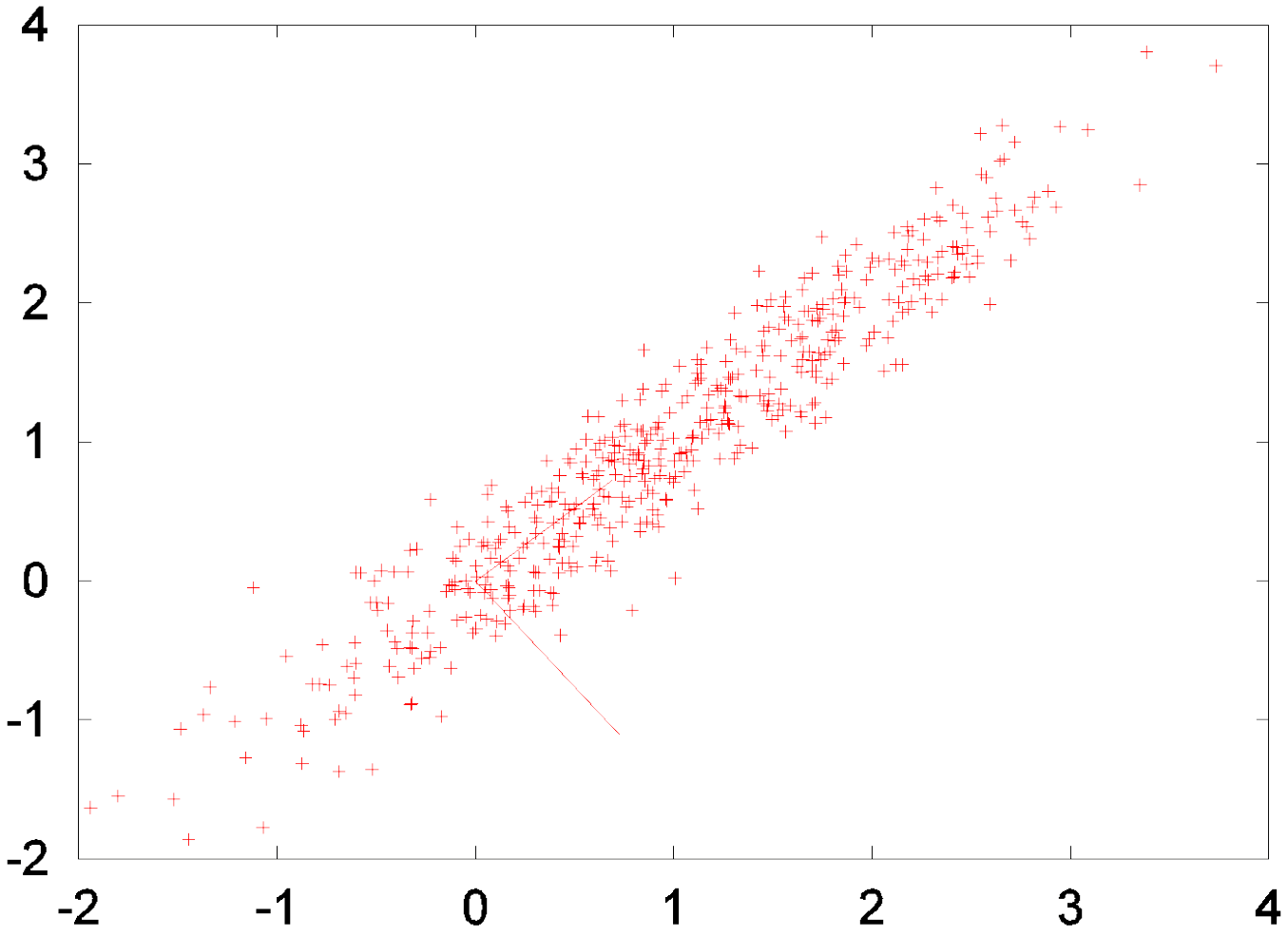


Figure: A typical eigenvalue spectrum and its division into two orthogonal subspaces

Determining the number of component



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

Determining the number of component

- $\mathbf{x}_i \in \mathbb{R}^d, i = 1, ..., N$
- $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d$: d eigenvectors (principal component directions)
- $||\mathbf{w}_i|| = 1$ (the $\mathbf{w}_i$s are orthonormal vectors)
- Representation of $\mathbf{x}_i$ in eigenvector space:

$$\mathbf{z}_i = (\mathbf{w}_1^T \mathbf{x}_i)\mathbf{w}_1 + (\mathbf{w}_2^T \mathbf{x}_i)\mathbf{w}_2 + ... + (\mathbf{w}_d^T \mathbf{x}_i)\mathbf{w}_d$$

- Suppose we retain the first k principal component:

$$\mathbf{z}_i^k = (\mathbf{w}_1^T \mathbf{x}_i)\mathbf{w}_1 + (\mathbf{w}_2^T \mathbf{x}_i)\mathbf{w}_2 + ... + (\mathbf{w}_k^T \mathbf{x}_i)\mathbf{w}_k$$

- Then:

$$\mathbf{z}_i - \mathbf{z}_i^k = (\mathbf{w}_{k+1}^T \mathbf{x}_i)\mathbf{w}_{k+1} + ... + (\mathbf{w}_d^T \mathbf{x}_i)\mathbf{w}_d$$

Determining the number of component

$$(\mathbf{z}_i - \mathbf{z}_i^k)^T(\mathbf{z}_i - \mathbf{z}_i^k) =$$

$$\left[(\mathbf{w}_{k+1}^T\mathbf{x}_i)\mathbf{w}_{k+1} + ... + (\mathbf{w}_d^T\mathbf{x}_i)\mathbf{w}_d\right]^T\left[(\mathbf{w}_{k+1}^T\mathbf{x}_i)\mathbf{w}_{k+1} + ... + (\mathbf{w}_d^T\mathbf{x}_i)\mathbf{w}_d\right]$$

$$= \mathbf{w}_{k+1}^T(\mathbf{w}_{k+1}^T\mathbf{x}_i)^2\mathbf{w}_{k+1} + ... + \mathbf{w}_d^T(\mathbf{w}_d^T\mathbf{x}_i)^2\mathbf{w}_d$$

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

$$= (\mathbf{w}_{k+1}^T\mathbf{x}_i)^2\mathbf{w}_{k+1}^T\mathbf{w}_{k+1} + ... + (\mathbf{w}_d^T\mathbf{x}_i)^2\mathbf{w}_d^T\mathbf{w}_d$$

$$= (\mathbf{w}_{k+1}^T\mathbf{x}_i)^2 + ... + (\mathbf{w}_d^T\mathbf{x}_i)^2$$

$$= (\mathbf{w}_{k+1}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{w}_{k+1}) + ... + (\mathbf{w}_d^T\mathbf{x}_i)(\mathbf{x}_i)^T\mathbf{w}_d$$

$$= \mathbf{w}_{k+1}^T(\mathbf{x}_i\mathbf{x}_i^T)\mathbf{w}_{k+1} + ... + \mathbf{w}_d^T(\mathbf{x}_i\mathbf{x}_i^T)\mathbf{w}_d$$

Determining the number of component

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \mathbf{z}_i^k)^T(\mathbf{z}_i - \mathbf{z}_i^k) =$$

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

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

$$= \mathbf{w}_{k+1}^T\Sigma\mathbf{w}_{k+1} + \ldots + \mathbf{w}_d^T\Sigma\mathbf{w}_d$$

(Note: $\Sigma\mathbf{w}_{k+1} = \lambda_{k+1}\mathbf{w}_{k+1}, \ldots, \Sigma\mathbf{w}_d = \lambda_d\mathbf{w}_d$)

$$= \mathbf{w}_{k+1}^T\lambda_{k+1}\mathbf{w}_{k+1} + \mathbf{w}_d^T\lambda_d\mathbf{w}_d$$

$$= \lambda_{k+1} + \ldots + \lambda_d$$

Determining the number of component

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \mathbf{z}_i^k)^T(\mathbf{z}_i - \mathbf{z}_i^k) = \lambda_{k+1} + ... + \lambda_d$$

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.

PCA - Advantages

- 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.

Acknowledgements

Slides adapted from Dr. Carlotta Domeniconi's *Pattern Recognition* at George Mason University.