Data Transformation

Irene Siragusa, PhD

CHILAB
COMPUTER HUMAN INTERACTION LABORATORY
Deep Learning and NLP in Healthcare
Università degli Studi di Palermo

Outline

- Dimensionality reduction with axis rotation
  - Principal Component Analysis
  - Singular Value Decomposition
  - Latent Semantic Analysis
- Dimensionality reduction with type transformation
  - Discrete Wavelet Transform
  - Muti Dimensional Scaling
  - Spectral Transformation

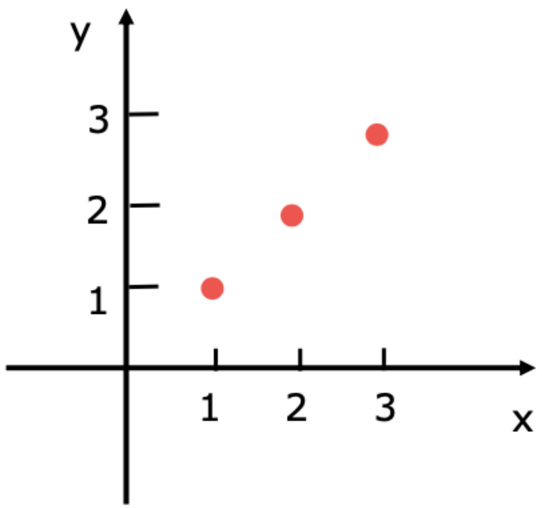
Dimensionality reduction with axis rotation

Consider the following 3 points in a 2-dimensional space

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\mathbf{x}_3 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$



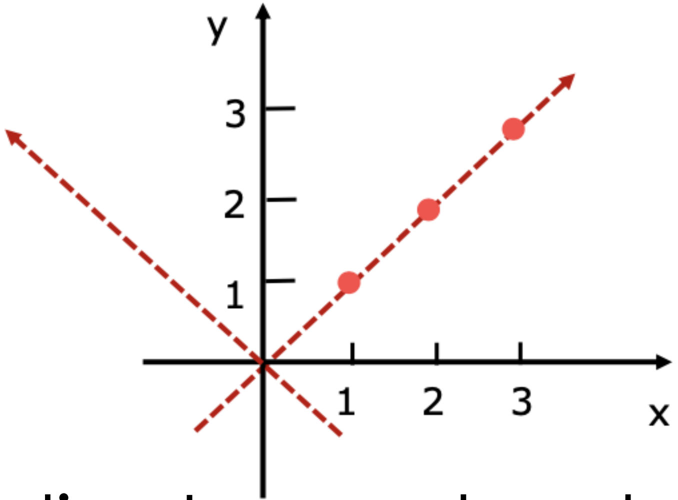
Dimensionality reduction with axis rotation

What is the new coordinates if we rotate the axis

$$\mathbf{x}_1 = \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix}$$

$$\mathbf{x}_3 = \begin{bmatrix} 3\sqrt{2} \\ 0 \end{bmatrix}$$



The second coordinate can be dropped

without information loss

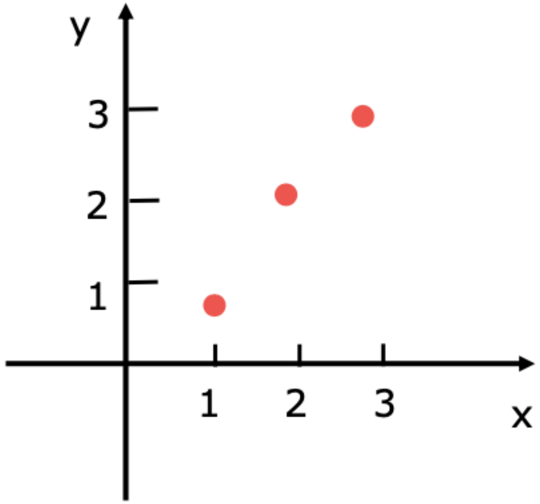
Dimensionality reduction with axis rotation

Consider the following 3 points in a 2-dimensional space

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0.9 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} 2.1 \\ 2 \end{bmatrix}$$

$$\mathbf{x}_3 = \begin{bmatrix} 2.9 \\ 3.1 \end{bmatrix}$$



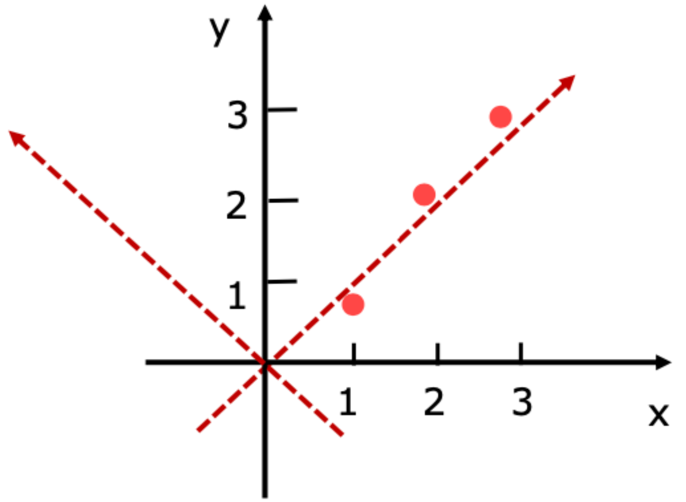
Dimensionality reduction with axis rotation

What is the new coordinates if we rotate the axis

$$\mathbf{x}_1 = \begin{bmatrix} 1.34 \\ 0.07 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} 2.89 \\ 0.07 \end{bmatrix}$$

$$\mathbf{x}_3 = \begin{bmatrix} 4.24 \\ -0.14 \end{bmatrix}$$



The second coordinate can be dropped
with little information loss

Dimensionality reduction with axis rotation

- Dimentionality reduction can be done when correlations exist among features

- Data highly correlated are concentrated along few preferred dimensions that can be used as new axis obtained via rotation

Dimensionality reduction with axis rotation

- Axis rotation
  - ☑ Remove correlations
  - ☑ Reduce dimensionality

- How to determine new axis system?
  - Principal componend analysis (PCA)
  - Singular value decomposition (SVD)
  - Latent Semantic Analysis

Axis rotation

- By default, the original coordinates are defined with respect to standard basis $\{\mathbf{e}_1,\ \mathbf{e}_2, \dots, \mathbf{e}_d\}$

$$\begin{bmatrix} x^1 \\ x^2 \\ \dots \\ x^d \end{bmatrix} \in \ \mathbb{R}^d \ \leftrightarrow \mathbf{x} = x^1\mathbf{e}_1 + x^2\mathbf{e}_2 + \cdots + x^d\mathbf{e}_d$$

- We can build an orthonormal basis $\{\mathbf{w}_1,\ \mathbf{w}_2, \dots, \mathbf{w}_d\}$ from which, calculate the associated orthonormal matrix $W = [\mathbf{w}_1,\ \mathbf{w}_2, \dots, \mathbf{w}_d]$

Axis rotation

- Then we can calculate a new representation of **x** with the new orthonormal basis

$$\mathbf{x} = WW^{\mathrm{T}}\mathbf{x} = \left(\sum_{i=1}^{d} \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}}\right)\mathbf{x} = \sum_{i=1}^{d} \mathbf{w}_i (\mathbf{w}_i^{\mathrm{T}}\mathbf{x})$$

$$= (\mathbf{w}_1^{\mathrm{T}}\mathbf{x})\mathbf{w}_1 + (\mathbf{w}_2^{\mathrm{T}}\mathbf{x})\mathbf{w}_2 + \cdots + (\mathbf{w}_d^{\mathrm{T}}\mathbf{x})\mathbf{w}_d$$

- Thus building new coordinates

$$\mathbf{y} = \begin{bmatrix} \mathbf{w}_1^{\mathrm{T}}\mathbf{x} \\ \mathbf{w}_2^{\mathrm{T}}\mathbf{x} \\ \cdots \\ \mathbf{w}_d^{\mathrm{T}}\mathbf{x} \end{bmatrix} \in \mathbb{R}^d$$

and by dropping some of the new coordinates, we reduce dimensionality

Principal Component Analysis

- Generally applied after mean centering, where data are centered in the origin by subtracting the mean to each data point.

- The goal of PCA is to rotate the data into an axis-system where the greatest amount of variance is captured in a small number of dimensions.

Principal Component Analysis

- Given a data set $\mathcal{D}$ with n data points and d dimensions, be C the covariance matrix

$$C = \frac{D^T D}{n} - \bar{\mu}^T \bar{\mu}$$

$$c_{ij} = \frac{\sum_{k=1}^{n} x_k^i x_k^j}{n} - \mu_i \mu_j \quad \forall\ i, j\ \in \{1, \dots, d\}$$

Principal Component Analysis

- C is a semidefinite matrix, meaning that $\bar{v}^T C \bar{v} > 0$ and doing this with a d-vector $\bar{v}$ is equal to calculate the variance of the 1-dimentional projection $D\bar{v}^T$ of the dataset $\mathcal{D}$ on $\bar{v}$

$$\bar{v}^T C \bar{v} = \frac{(D\bar{v})^T D\bar{v}}{n} - (\bar{\mu}\bar{v})^2$$

- Goal of PCA is to determine orthonormal vectors $\bar{v}$ maximizing $\bar{v}^T C \bar{v}$, that is the variance along the new directions

Principal Component Analysis

- But the covariance matrix is symmetric and positive semidefinite, the following diagonalization is possible

$$C = P\Lambda P^T$$

- Λ is a diagonal matrix with eigenvalues of C, $C\mathbf{v} = \lambda\mathbf{v}$
- Columns of the matrix P contain the orthonormal eigenvectors of C, representing successive orthogonal solutions to the optimization model of maximizing the variance $\bar{v}^T C\bar{v}$ along the unit direction $\bar{v}$.

Principal Component Analysis

- Eigenvalues represent the variances of the data along the corresponding eigenvectors
- Diagonal matrix Λ is the new covariance matrix after axis rotation
- We can re-arrange rows of P in decreasing order as for the eigenvalues and consider only the first k principal components for which the total variance is preserved since it is higher than a given threshold.

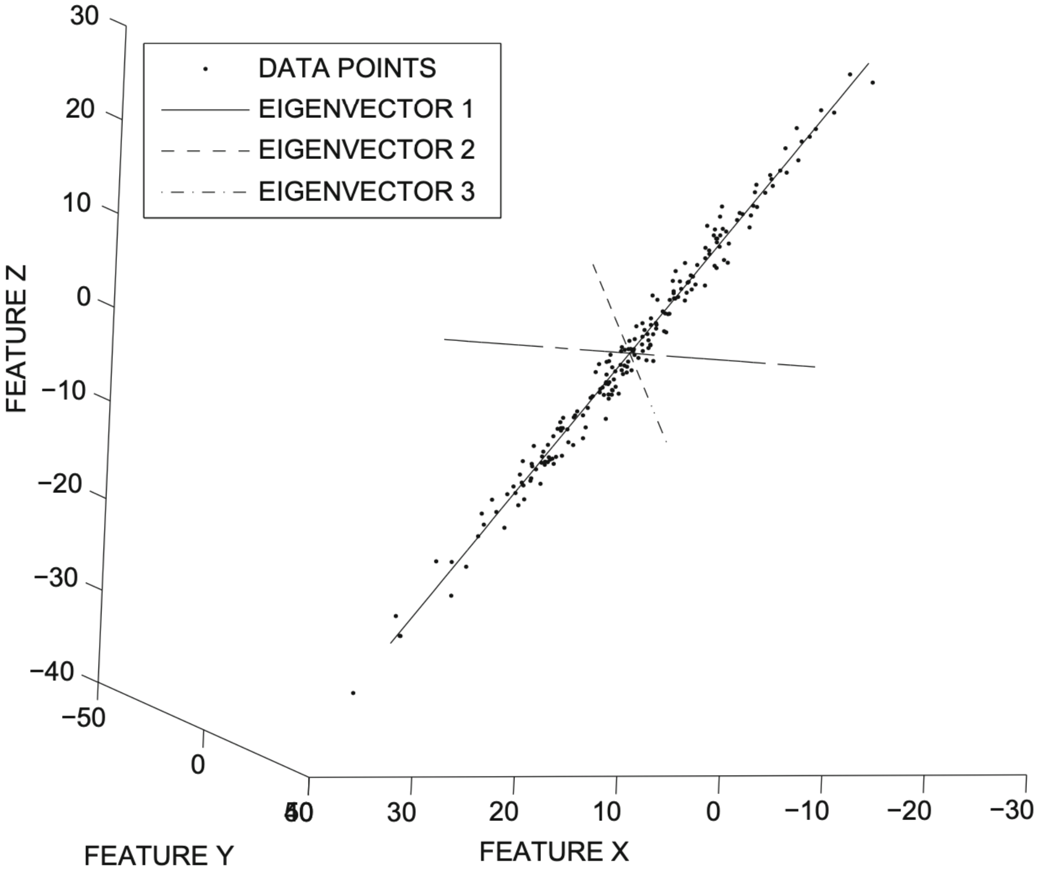
Principal Component Analysis

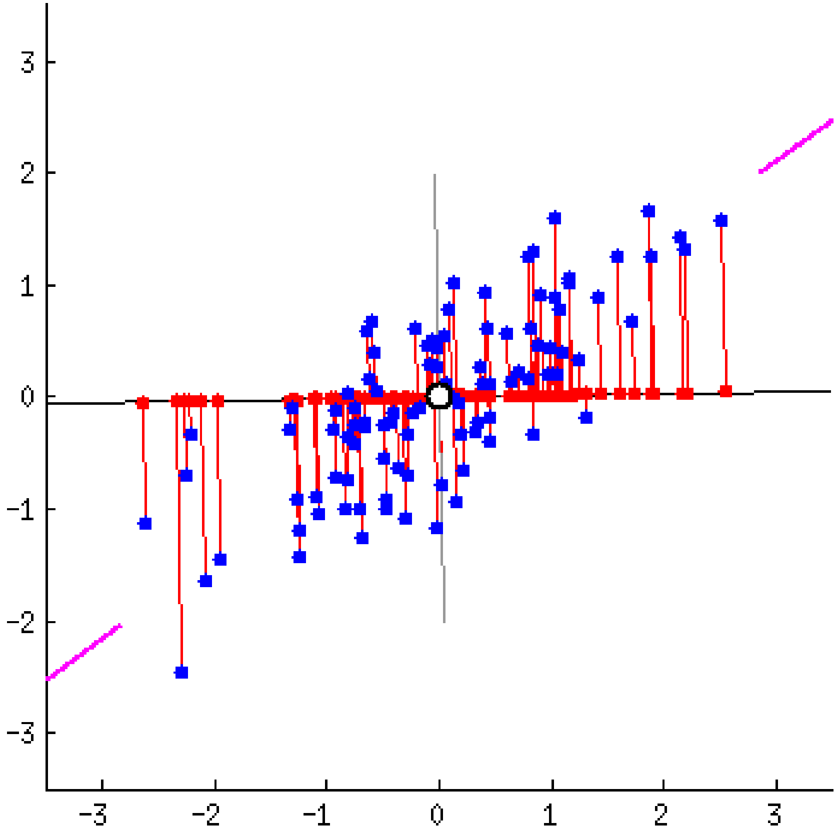
- From the re-arranged matrix P, the new data points can be calculated

$$D' = DP$$

- From matrix D' of size $n \times d$, only its first leftmost k columns will have the most variance.
- Remaining $d - k$ columns will be approximately equal to the mean of the data in the rotated axis system.
- Covariance matrix of D' is the diagonal matrix Λ where correlations have been removed.
- The variance of the D' is equal to the sum of the k corresponding eigenvalues.

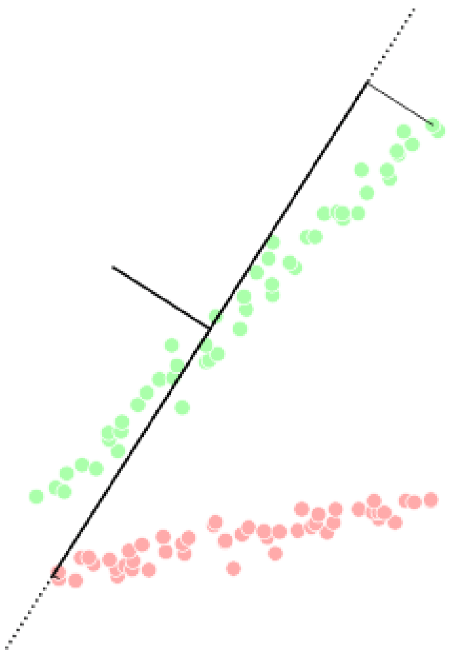
Principal Component Analysis



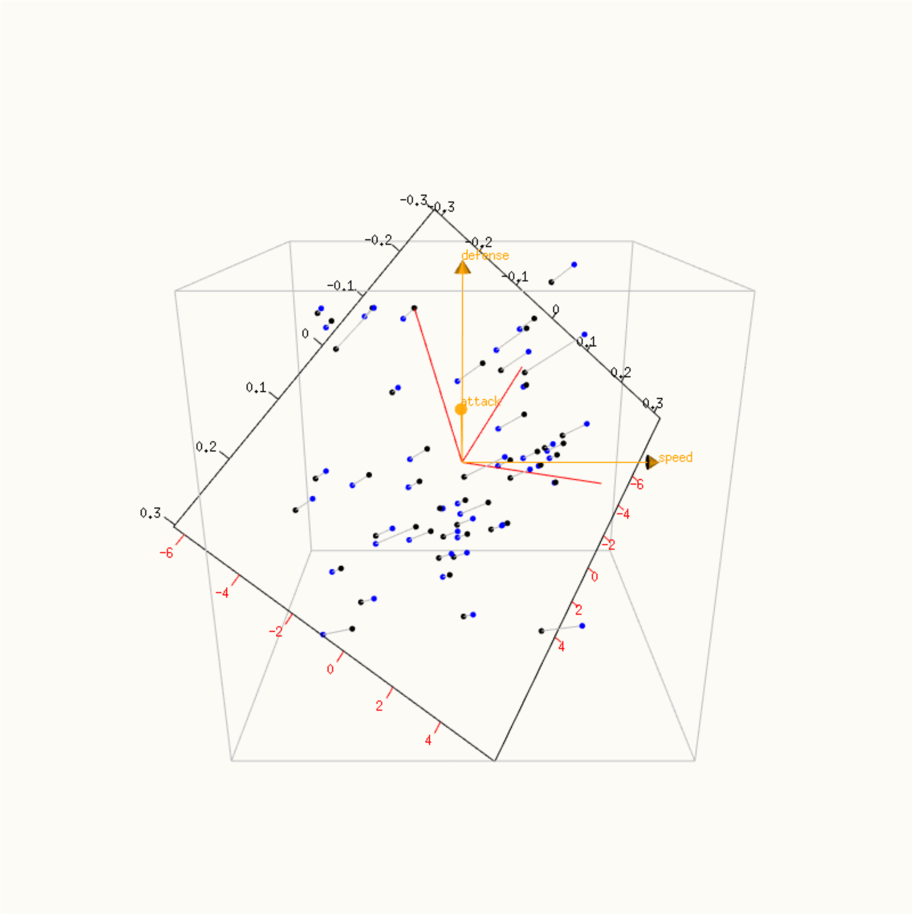


https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Component Analysis



Principal Component Analysis



Singular Value Decomposition

- Singular Value Decomposition (SVD) is more general than PCA because it provides two sets of basis vectors instead of one.

  ☑ Basis vectors of both the rows and columns of the datamatrix

  ✗ PCA only provides basis vectors of the rows of the data matrix.

Singular Value Decomposition

- SVD is equal to PCA for data sets in which the mean of each attribute is 0.

- The basis vectors of PCA are invariant to mean-translation, whereas those of SVD are not.
  - With mean centered data SVD coincides with PCA

- SVD is often applied without mean centering to sparse nonnegative data such as document-term matrices.

Singular Value Decomposition

- It is a factorization of data set $\mathcal{D}$ with n data points and d dimensions

$$D = Q\Sigma P^T$$

- Σ is an $n \times d$ diagonal matrix containing the nonnegative singular values, arranged in nonincreasing order and they are the latent representation of data points.
- Σ is rectangular and it is referred to as diagonal because only entries form Σ_{ii} are nonzero.

Singular Value Decomposition

- It is a factorization of data set $\mathcal{D}$ with n data points and d dimensions

$$D = Q\Sigma P^T$$

- Q is an $n \times n$ matrix with orthonormal columns, called left singular vectors and they are the eigenvectors of DD^T while $\Sigma\Sigma^T$ are their eigenvalues

$$DD^T = Q\Sigma(P^T P)\Sigma^T D^T = Q(\Sigma\Sigma^T)Q^T$$

Singular Value Decomposition

- It is a factorization of data set $\mathcal{D}$ with n data points and d dimensions

$$D = Q\Sigma P^T$$

- P is an $d \times d$ matrix with orthonormal columns, called right singular vectors and they are the eigenvectors of $D^T D$ while $\Sigma^T \Sigma$ are their eigenvalues

$$D^T D = P^T \Sigma^T (Q^T Q) \Sigma P = P^T (\Sigma^T \Sigma) P$$

- P provides the basis vectors as for the eigenvectors of the covariance matrix in PCA, and they are equal if SVD is applied to mean–centered data

Singular Value Decomposition

- Diagonal entries of Σ can be arranged in decreasing order, and columns of matrix P and Q ordered accordingly.

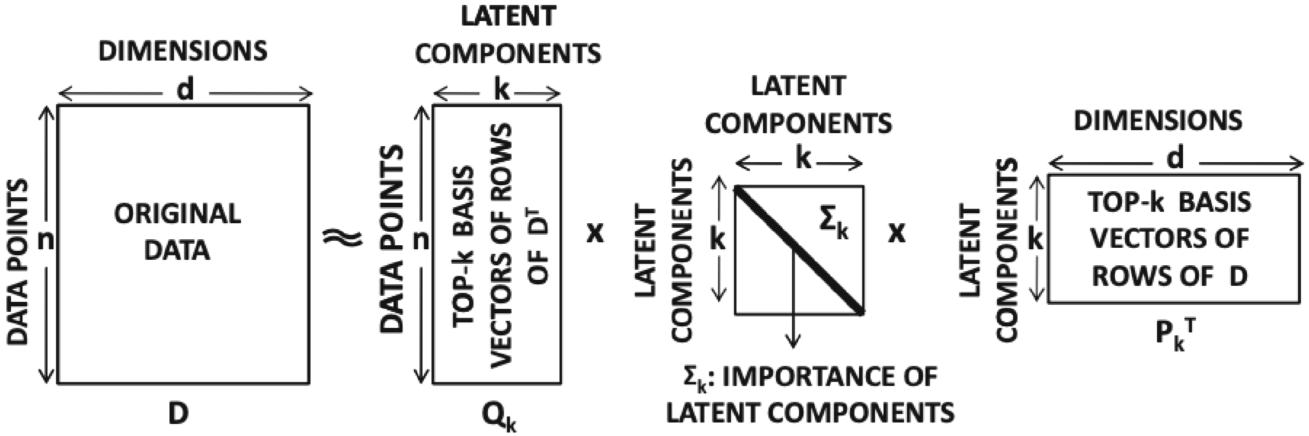
- P_k and Q_k are the truncated version of P and Q by selecting their first k columns.

- Σ_k is the $k \times k$ square matrix containing the top k singular values.

Singular Value Decomposition

- SVD factorization yields an approximate k-dimensional data representation of the original data set:

$$D \approx Q_k \Sigma_k P_k^T$$



SVD truncation maximizes the aggregate squared Euclidean distances (energy) of the transformed data points about the origin, thus making it more general than PCA

PCA vs SVD

- PCA
  - projects the data on a low-dimensional hyperplane passing through the data mean
  - captures as much of the variance (or, squared Euclidean distance about the mean) of the data as possible
- SVD
  - projects the data on a low-dimensional hyperplane passing through the origin.
  - captures as much of the aggregate squared Euclidean distance about the origin as possible.

PCA and SVD applications

- Noise reduction
  - Improvement on the quality of data
- Matrix inversion
  - SVD can be used for the inversion of a square $d \times d$ matrix D.
- Matrix algebra
  - Efficency in the application of algebraic operations.
- SVD and PCA are extraordinarily useful because matrix and linear algebra operations are ubiquitous in data mining.
- SVD and PCA facilitate such matrix operations by providing convenient decompositions and basis representations.

Latent Semantic Analysis

- Data matrix D is an $n \times d$ document-term matrix containing normalized word frequencies in the n documents, where d is the size of the lexicon:

$$
\begin{array}{r c}
 & \begin{array}{ccc} the & an & \end{array} \\
\begin{array}{r} The\ cat\ on\ the\ table \\ An\ apple \\ \dots \end{array} & \begin{bmatrix} 2 & 0 & \dots \\ 0 & 1 & \dots \\ \dots & \cdots & \dots \end{bmatrix}
\end{array}
$$

- D is a a large and sparse matrix, with low column mean
- Covariance matrix is proportional to $D^T D$

Latent Semantic Analysis

- It is possible to apply SVD's to the document-term matrix and obtaining a high decreasing in dimensionality and higher-quality data since noise effects of synonymy and polysemy are removed
  - High-energy singular vectors represent the directions of correlation in the data, and the appropriate context of the word is implicitly represented along these directions.
  - Low-energy singular vectors represent the variations at individual level.

Dimensionality reduction with type transformation

- Data are transformed from a more complex type to a less complex type, thus
  - ☑ data type portability
  - ☑ dimensionality reduction
  - ☑ less dependency-oriented data

- Methods
  - Time series to multidimensional via Discrete Wavelet Transform (DWT)
  - Weighted graphs to multidimensional via Multi-Dimensional Scaling (MDS) and Spectral Transformation

Discrete Wavelet Transform

- Discrete Wavelet Transform (DWT) or also referred to Haar Wavelet Transform

- Allow multigranularity decomposition and summarization of time-series data and their transformation as multidimensional data.

- Highlights the variation in time-series instead of the actual values that may be redundant (eg. sensors)

Discrete Wavelet Transform

- Wavelet technique creates a set of coefficient-weighted wavelet basis vectors. Each coefficient represents the rough variation of the time series between the two halves of a particular time range.

- Different order-level of coefficient can be determined
  - Higher-order coefficients represent broad trends in the series since they correspond to larger ranges.
  - Lower-order coefficients represent more localized trends in shorter series.

Discrete Wavelet Transform

- Given a temporal sequence $(t_i; x_i)$ with length q that is a power of 2, it can be recursively divided into subseries

$$S_k^i = \left[(i-1) * \frac{q}{2^{k-1}} + 1, i * \frac{q}{2^{k-1}}\right], \qquad k = 1, \dots, \log_2(q) + 1$$

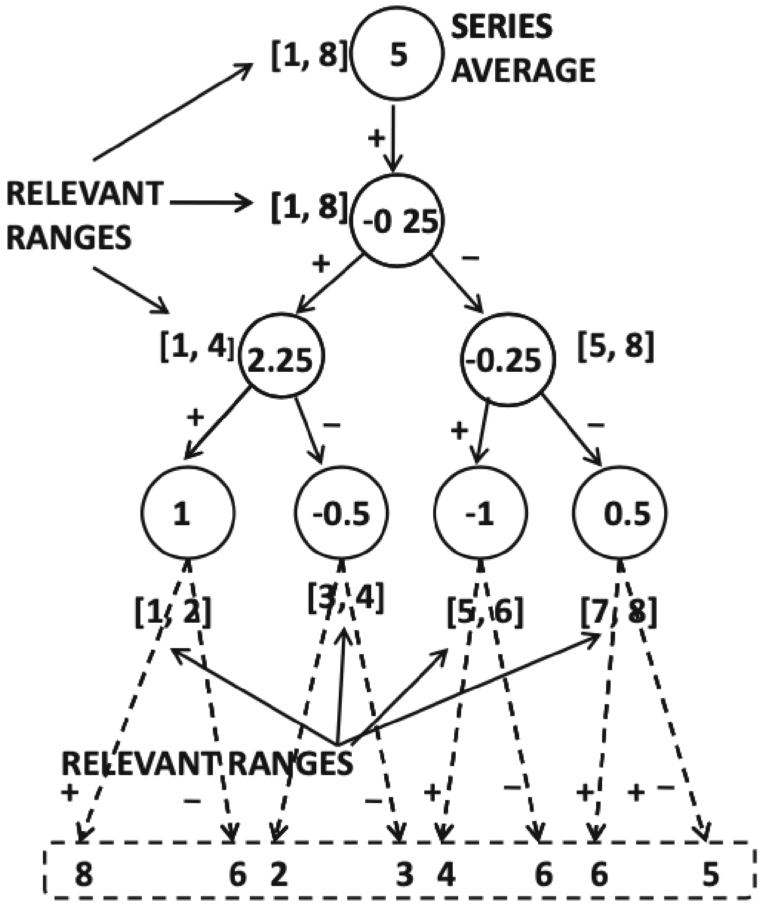
- i-th coefficient, for each k level of decomposition referring to the S_k^i subseries is

$$\psi_k^i = \frac{\Phi_{k+1}^{2*i-1} - \Phi_{k+1}^{2*i}}{2}$$

- where Φ_k^i are the average values of S_k^i sequence

Discrete Wavelet Transform

- DWT coefficients are defined by all the coefficients of order 1 to $\log_2(q)$.
- Global average Φ_1^1 is required for perfect reconstruction.
- The total number of coefficients is exactly equal to the length of the original series, thus the dimensionality reduction is obtained by discarding the smaller (normalized) coefficients.



Discrete Wavelet Transform

- Wavelet representation is a decomposition of the original time series of length q into the weighted sum of a set of q wavelets, simpler time series that are orthogonal to one another.

- Wavelets are the basis vectors, and the wavelet coefficients represent the weights of the different basis vectors in the decomposition.

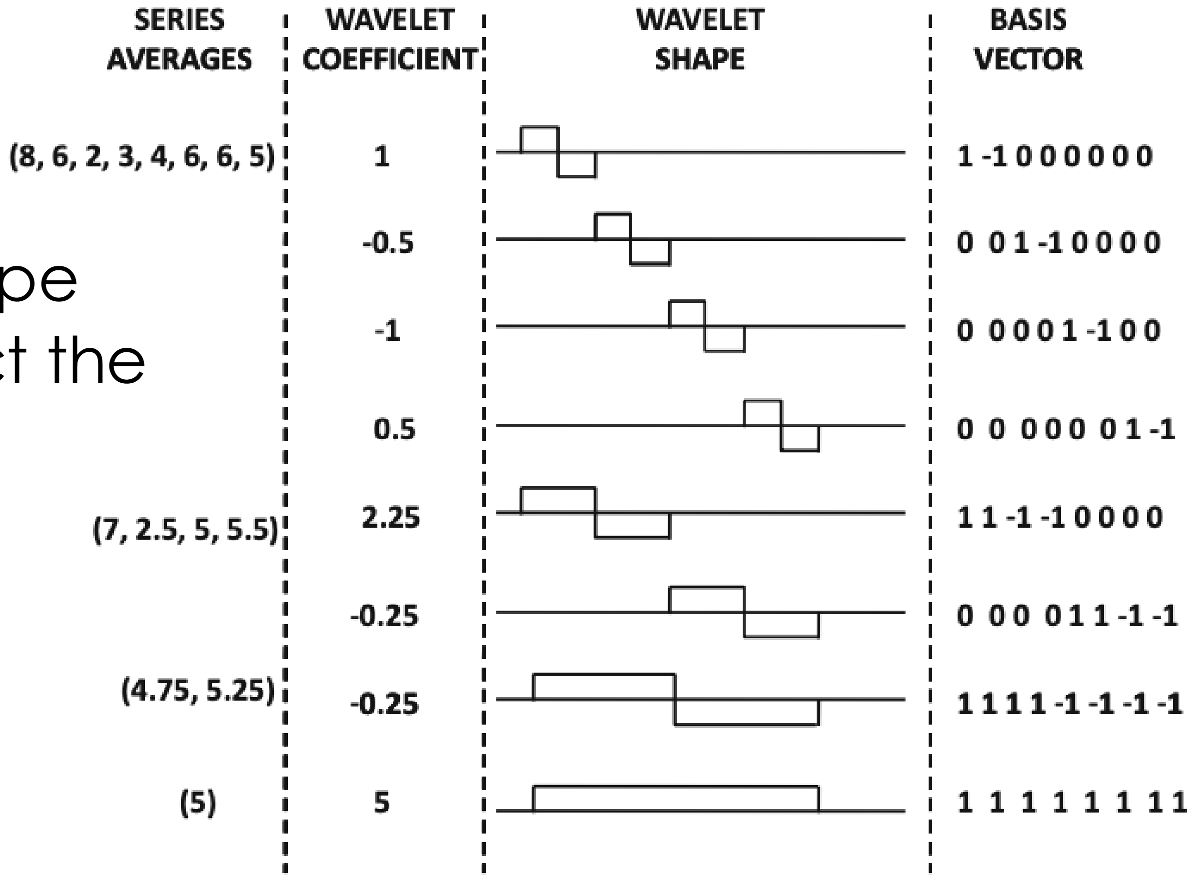
- Number of wavelet coefficients, basis vectors and their length is equal to q.

Discrete Wavelet Transform

- Basis vectors
  - values 1, -1 and 0
  - are orthonormal
  - have a simple wavelet shape
  - can be used for reconstruct the whole sequence

$$T = \sum_{i=1}^{q} a_i \overline{W_i} = \sum_{i=1}^{q} \boxed{a_i \|\overline{W_i}\|} \frac{\overline{W_i}}{\|\overline{W_i}\|}$$

normalized coefficients



Discrete Wavelet Transform

- DWT can be seen as an axis rotation.
- Dimensionality reduction can be achieved retaining the coefficients with the largest normalized values.
  - the error loss from the wavelet representation is minimized.
- DWT can be extended to multi-dimensional series case.
  - 2-d DWT as for spatial series

Multi Dimensional Scaling

- Given a graph with n nodes, and edge weight δ_{ij} denotes distance or similarity between two nodes and all pairwise weights are known.
- The objective is to map the n nodes to n *k*-dimensional vectors $\overline{X_1}, \dots, \overline{X_n}$, whose reciprocal distances (eg. Euclidian distance) corresponds to δ_{ij}.
- The *k* coordinates are treated as variables that need to be optimized

$$O = \sum_{i,j:i<j} \left(\left\|\overline{X_i} - \overline{X_j}\right\| - \delta_{ij}\right)^2$$

Multi Dimensional Scaling

- Distance matrix $\Delta = \left[\delta_{ij}^2\right]_{n \times n}$ can be converted into a symmetric doc-product matrix $S_{n \times n}$

$$\bar{X}_i \cdot \bar{X}_j = -\frac{1}{2}\left[\left\|\bar{X}_i - \bar{X}_j\right\|^2 - \left(\|\bar{X}_i\|^2 + \left\|\bar{X}_j\right\|^2\right)\right]$$

- Considering I as the identity matrix and U as a $n \times n$ matrix of 1s

$$S = -\frac{1}{2}\left(I - \frac{U}{n}\right)\Delta\left(I - \frac{U}{n}\right) \equiv DD^T$$

via SVD truncation

$$S \approx Q_k \Sigma_k^2 Q_k^T = (Q_k \Sigma_k)(Q_k \Sigma_k)^T \text{ where } D = Q_k \Sigma_k$$

Spectral Transformation

- Spectral methods are designed for preserving local distances.
- Spectral methods work with similarity graphs in which the weights on the edges represent similarity between adjacent nodes.

Spectral Transformation

- $G = (N, A)$ is an undirected graph with n nodes, node set N and edge set A.
- Similarities between nodes are stored in the symmetric and sparse square matrix W.
- Objective is to embed the nodes in a k-dimensional space where data similarity structure is preserved.

$$O = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(y_i - y_j \right)^2$$

Spectral Transformation

- Objective function can be rewritten in terms of the Laplacian matrix L of weight matrix W
$$L = \Lambda - W$$
- Λ is a diagonal matrix of $\Lambda_{ii} = \sum_{j=1}^{n} w_{ij}$
- Thus considering $\bar{y} = (y_1, \dots, y_n)^T$, O can be re-written as
$$O = 2\bar{y}^T L \bar{y}$$
- Considering also the following constraint $\bar{y}^T \Lambda \bar{y} = 1$, O is optimized by selecting the k smallest eigenvectors in
$$\Lambda^{-1} L \bar{y} = \lambda \bar{y}$$