

Lecture „Intelligent Systems“

Chapter 5: Similarities of Time-Series

Prof. Dr.-Ing. habil. Sven Tomforde / Intelligent Systems
Winter term 2020/2021

Contents

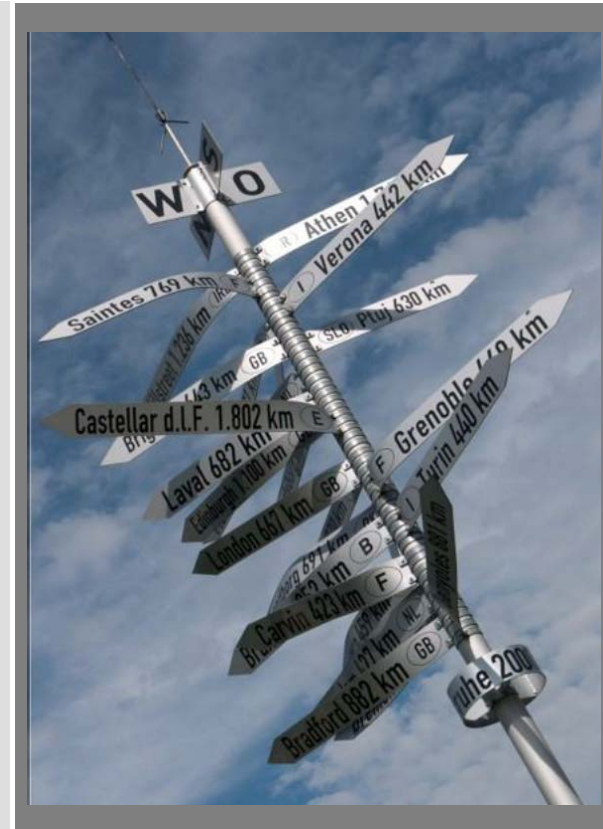
- Fundamentals of similarity measurement
- Dynamic similarity measures for time series
- Similarity measures for time series models
- Conclusion and further readings

Goals

Students should be able to:

- determine the distance of time series element by element.
- define and apply dynamic similarity measures for time series (LCSS, DTW, ED).
- explain the principle of similarity determination on time series models using examples.

- Fundamentals of similarity measurement
- Dynamic similarity measures for time series
- Similarity measures for time series models
- Conclusion and further readings



Similarity

- Comparison of samples (or segments), i.e. distance or similarity measurement, is necessary in all further machine learning steps considering sensor-based information in intelligent systems.
- Basically, large distances are associated with low similarity, small distances with high similarity and vice versa.
- Distance measurement of different elements can be performed:
 - Directly on **raw data** (features),
 - on a corresponding **representation** or
 - about a **model** created from the data.

Element-by-element distance

- Simplest distance between two patterns: Minkowsky norms
- For time series, e.g., application to feature vectors or evaluation by element.
- Assumption: Time-series have the same length (otherwise: apply e.g. interpolation, see missing values)
- Distance between two-time series X and Y (with same length N):

$$D_P(X, Y) = \left(\sum_{i=1}^N |x_i - y_i|^p \right)^{\frac{1}{p}}$$

– with x_i and y_i being the i -th elements of the two time-series

- $p = 1$: Manhattan Distance
- $p = 2$: Euclidian Distance
- ...

Multivariate time-series

- For multi-variate (i.e. m-dimensional) time-series, the distance can also be defined:

$$D(X, Y) = \frac{1}{N} \sum_{i=1}^N ||x_i - y_i||$$

where x_i and y_i represent the i-th elements (here: m-dimensional vectors), N is the number of samples

- Instead of the Euclidian distance between two vectors x and y , other dimensions can be used, e.g. the matrix norm:

$$|||x - y||_M := \sqrt{(x - y)^T M (x - y)}$$

for $x, y \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times n}$

- For:

$$M := \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{pmatrix}$$

with any real-valued diagonal elements, we get a so-called diagonal norm.

- For $M=I$ (thus all diagonal elements are 1) the result is again the Euclidean Norm (i.e. as a special case).

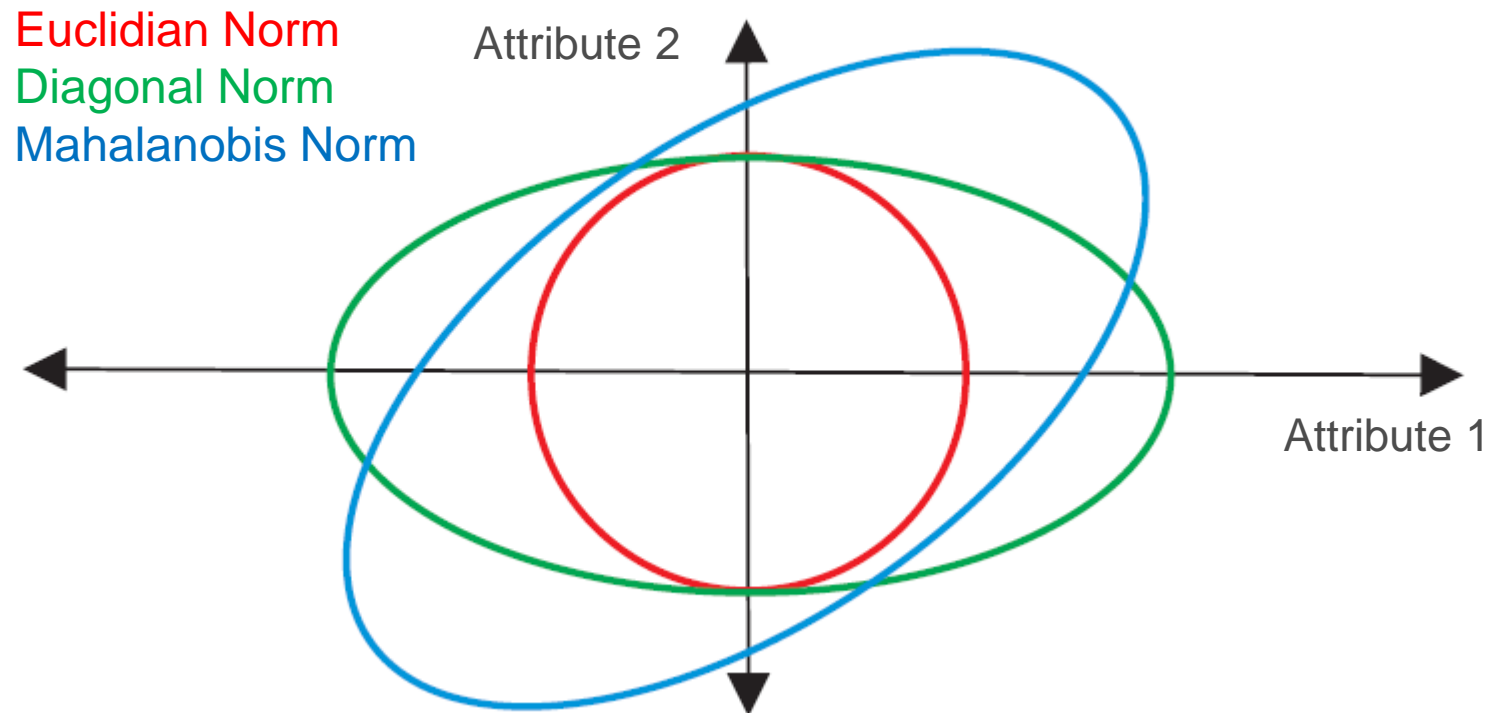
Mahalanobis norm

- Defined by the inverse of the covariance matrix of the data values:

$$\mathbf{M} := \left(\frac{1}{N-1} \sum_{k=1}^N (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T \right)^{-1}$$

- with the mean value of:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^N \mathbf{x}_k$$



- The points on the circle or ellipse have the same distance to the origin with respect to the selected norm (data sets are not shown here).

Further examples of distance dimensions:

- **Cosine distance**

- Normalised standard scalar product of two vectors (cosine of the angle)
- Defined as:

$$d(x, y) := \frac{\langle x | y \rangle}{\|x\| \cdot \|y\|}$$

- Remark: An alternative notation of $\langle x | y \rangle$ is $x^T y$.

Further examples of distance measures (continued):

- Hamming distance:
 - Measure for the difference of character strings
 - Named after Richard W. Hamming (1915-1998)
 - Often used for error detection/correction: Data elements received via a transmission path are compared with valid characters (correction then via probability if necessary).
 - Examples:

0	0	1	1	0
0	0	1	0	0

Hamming distance: 1

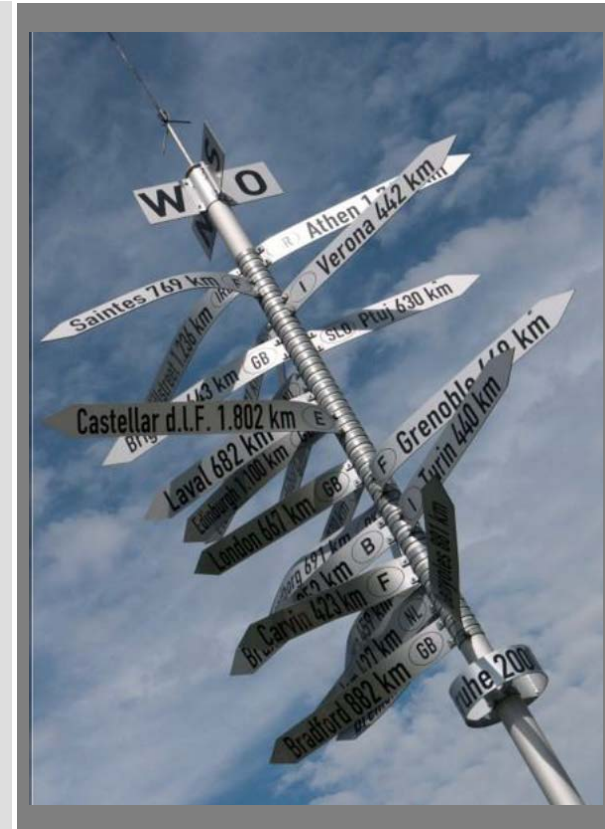
1	2	3	4	5
1	3	3	4	4

Hamming distance: 2

A	P	P	L	E
B	E	R	R	Y

Hamming distance: 5

- Fundamentals of similarity measurement
- **Dynamic similarity measures for time series**
- Similarity measures for time series models
- Conclusion and further readings



Similarity measures for time series

- Special similarity measures on time series for the consideration of dynamic temporal relationships
- Frequently additional processing of time series of different lengths possible
- Dynamic **hiding of different scales and translations** in the value and time range

Longest Common Subsequences (LCSS)

- Goal: Find the longest common partial sequence of several sequences.
- Important: Partial sequence does not necessarily mean that only one (coherent) "section" of the original sequence is possible.
- Example:
 - Sequence $X = \langle B, G, M, M, T, E, Y, R, F, F, B \rangle$
 - Sequence $Y = \langle G, D, F, F, T, E, R, R, A, S, U, B, B, W \rangle$
- The longest joint partial sequence of X and Y is: $\langle G, T, E, R, B \rangle$
- Application especially in bioinformatics (e.g. gene sequences)

Longest Common Subsequences

- Given are two time-series of the lengths n and m
 - $X = (x_1, \dots, x_n)$
 - $Y = (y_1, \dots, y_m)$
- Goal: Search for the longest common partial sequence of both time-series
 - Taking into account local scaling and translation in the value range
- Solution: LCCS on time-series

LCCS – Step 1: Atomic Matchings

- Two time-series $S = (s_1, \dots, s_w)$ and $T = (t_1, \dots, t_w)$ of length w
- S and T are called similar if the following holds:
$$|s_i - t_i| \leq \varepsilon$$
for a given $\varepsilon \in \mathbb{R}$ and $i = 1, \dots, w$.
- Initially, all connected and related sub-sequences of length w are extracted from the time-series X and Y :
 - $\tilde{S}_i = (x_i, \dots, x_{i+w-1})$ with $i = 1, \dots, n - w + 1$
 - $\tilde{T}_i = (y_i, \dots, y_{i+w-1})$ with $i = 1, \dots, n - w + 1$

LCCS – Step 1: Atomic Matchings

- All partial sequences are normalised (i.e., to $[0;1]$) or standardised (e.g., to mean 0 and standard deviation 1) using appropriate scaling
- The standardised sub-sequences S_i of the time-series X are now checked for similarity in pairs with all partial sequences T_j of the time-series Y .
- Any match between a sub-sequence of X and a subsequence of Y is called “atomic matching” (of a length w).

Note:

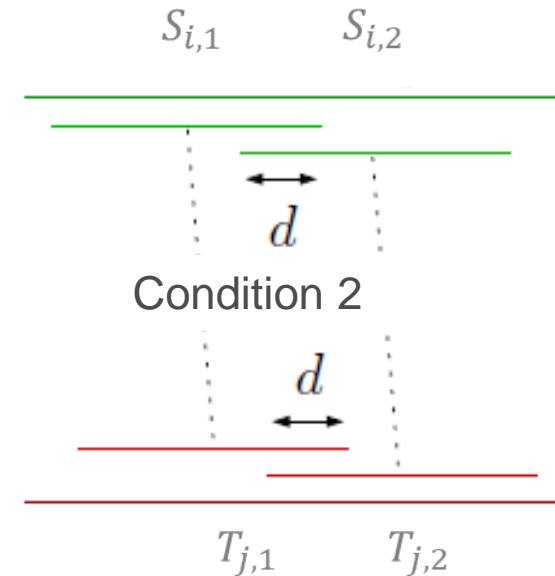
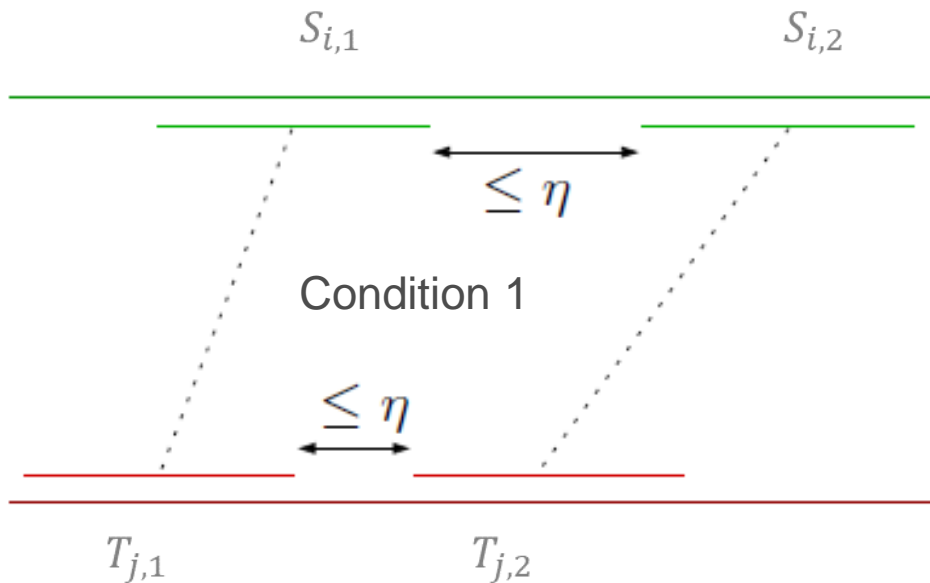
- Challenge: Comparison of $n - w + 1$ partial sequences of X with $m - w + 1$ partial sequences of Y .
- Alternative: An efficient search for similar partial sequences using a suitable index structure.

LCCS – Step 2: Formation of longer partial sequences

- Goal:
 - Atomic similarities are now combined to longer sequences
 - Combination is done with other atomic similarities or already combined sections
- Given:
 - Two matches $(S_{i,1}; T_{j,1})$ and $(S_{i,2}; T_{j,2})$ with $i_1 < i_2$ and $j_1 < j_2$
- Furthermore:
 - $\text{length}(S_{i,1})$ and $\text{length}(T_{j,1})$ are functions determining the lengths of the sequences in terms of number of samples / data point.

- Matchings can be combined into longer sequences if one of the following conditions is met:
 1. Not overlapping
 - Sequences $S_{i,1}$ and $S_{i,2}$ do no overlap on X (i.e., $i_1 + \text{length}(S_{i,1}) < i_2$)
 - Their distance is not greater than a fixed value $\eta \in \mathbb{N}$ (i.e., $i_1 + \text{length}(S_{i,1}) + \eta \geq i_2$).
 - The same condition applies for sequences $T_{j,1}$ and $T_{j,2}$ on Y .
 2. Overlapping
 - The two matches $(S_{i,1}; T_{j,1})$ and $(S_{i,2}; T_{j,2})$ overlap on both time-series by the same length $d = i_1 + \text{length}(S_{i,1}) - i_2 = j_1 + \text{length}(T_{j,1}) - j_2$
- Also, a certain similarity of the scaling factors used to scale the partial sequences involved may be required for a combination of matches.

Possibilities for constructing longer sub-sequences



LCSS – Step 3: Finding the longest match

- Now, all matches are combined as much as possible and k pairs of matches $(S'_1, T'_1), \dots, (S'_k, T'_k)$ are given.
- We are now looking for the subset $(S'_{l_1}, T'_{l_1}), \dots, (S'_{l_h}, T'_{l_h})$, for which the following requirements hold:
 1. The endpoint of S'_{l_i} is before the start point of S'_{l_j} on X and the endpoint of T'_{l_i} is before the start point of T'_{l_j} on Y for $1 \leq i < j \leq h$ (i.e., the subsequences S'_{l_i} and S'_{l_j} as well as T'_{l_i} and T'_{l_j} do not overlap on X and Y , correspondingly)
 2. The total length of all sequences is maximal:

$$\sum_{i=1}^h \text{length}(S'_{l_i}) + \sum_{i=1}^h \text{length}(T'_{l_i})$$

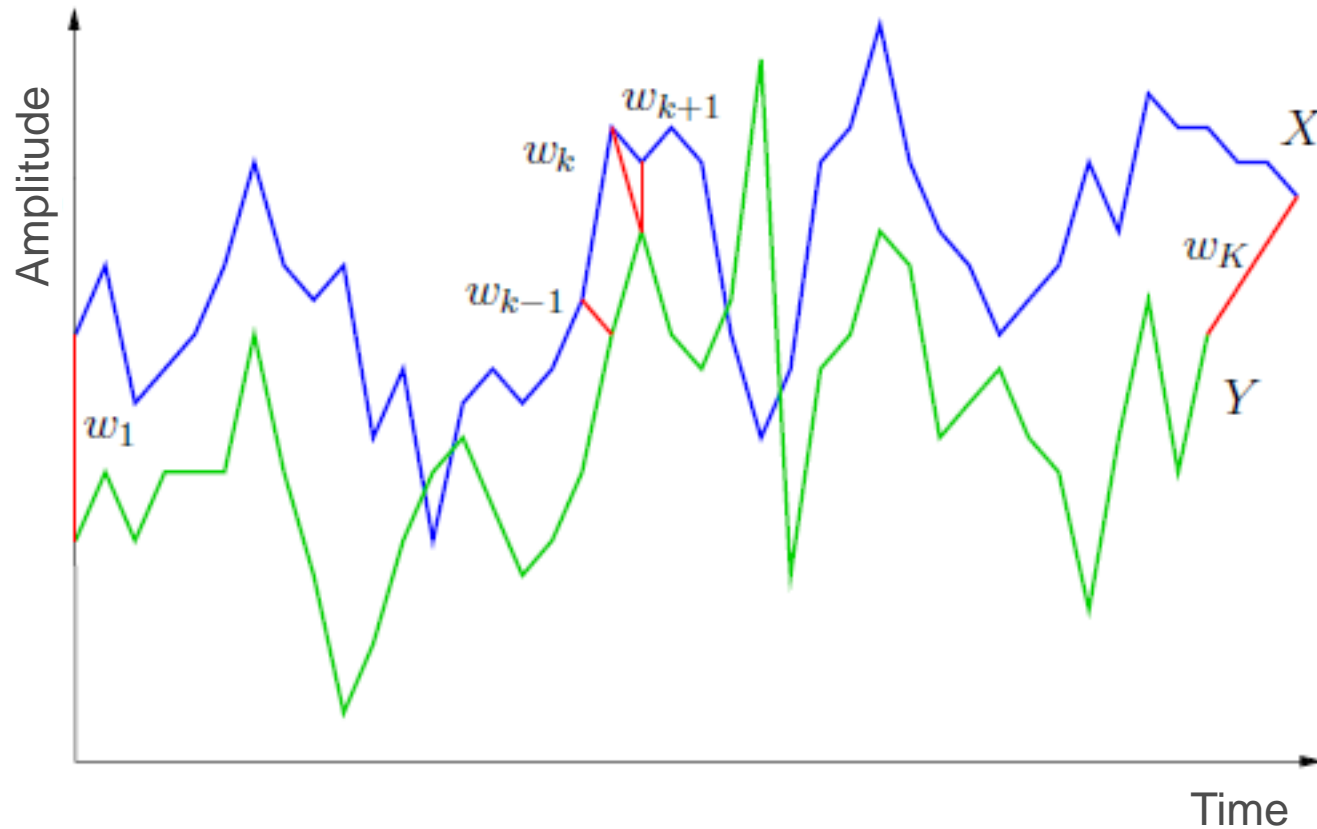
- Let \circ be the sequential composition of two sub-sequences.
- Then $X' = S'_{l_i} \circ \dots \circ S'_{l_h}$ and $Y' = T'_{l_i} \circ \dots \circ T'_{l_h}$ are the longest common sub-sequences of the time-series X and Y .
- In contrast to LCSS on symbolic sequences, X' and Y' can have different lengths (see condition 2 in step 2).
- Advantage of the procedure:
 - Flexible assignment of partial sequences of two time-series to each other.
 - Thus, robust comparison despite scaling, translation and longer sections that do not match
- Disadvantage:
 - High computing effort

Dynamic Time Warping

- Given:
 - Two time-series $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_m)$ of length n and m
- Goal:
 - Search for the so-called warping path $W = w_1, \dots, w_k$ of the length k
 - This path consists of assignments of both time-series of the form
$$w_k = (i_k, j_k)$$
 - The sum of all distances $d_k = d(x_{i_k}, y_{j_k})$ is minimal:

$$D(X, Y) = \min_w \sum_{k=1}^K d(x_{i_k}, y_{j_k})$$

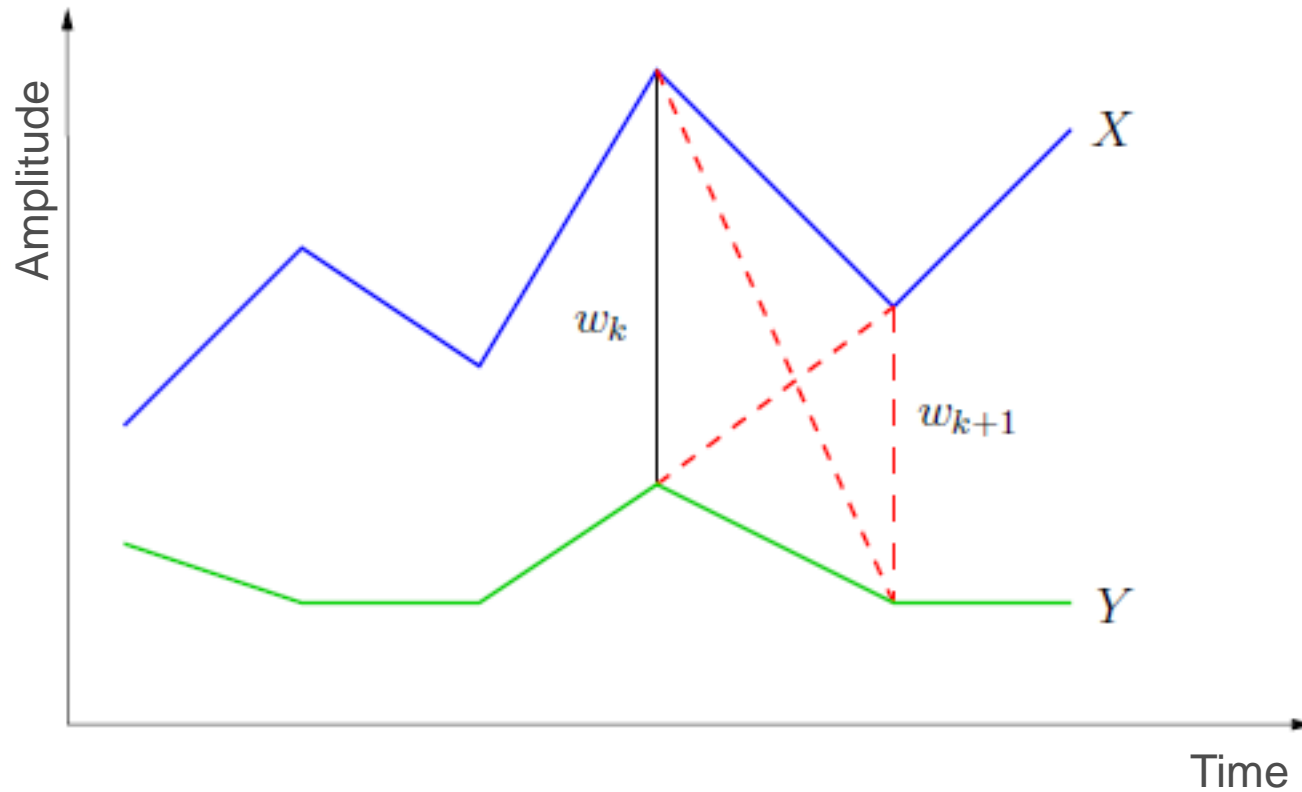
Example:



The following additional restrictions apply to the warping path:

- **Boundary condition:**
 - $w_1 = (1,1)$ and $w_k = (n, m)$
 - I.e.: the path begins with both first elements and ends with both last elements on the two time-series
- **Continuity:**
 - Let $w_k = (i_k, j_k)$ and $w_{k+1} = (i_{k+1}, j_{k+1})$ be two consecutive assignments
 - Then $i_{k+1} - i_k \leq 1$ and $j_{k+1} - j_k \leq 1$ must apply
 - I.e.: each warping path is continuous, so each element from both time-series occurs in at least one assignment
- **Monotony:**
 - Let $w_k = (i_k, j_k)$ and $w_{k+1} = (i_{k+1}, j_{k+1})$ be two consecutive assignments
 - Then $i_{k+1} - i_k \geq 1$ and $j_{k+1} - j_k \geq 1$ must apply
 - I.e.: The warping path assignments maintain the chronological order of the data points of both time-series.

Example



Determine the optimal warping path:

$$DTW(i, j) = d(x_i, y_j) + \begin{cases} 0 & \text{for } i = 1, j = 1 \\ DTW(i, j - 1) & \text{for } i = 1, j > 1 \\ DTW(i - 1, j) & \text{for } i > 1, j = 1 \\ \min \begin{pmatrix} DTW(i - 1, j), \\ DTW(i, j - 1), \\ DTW(i - 1, j - 1) \end{pmatrix} & \text{otherwise} \end{cases}$$

Remarks

- The minimum sum of the distances of all allocations $D(X, Y)$ from X and Y is given by:

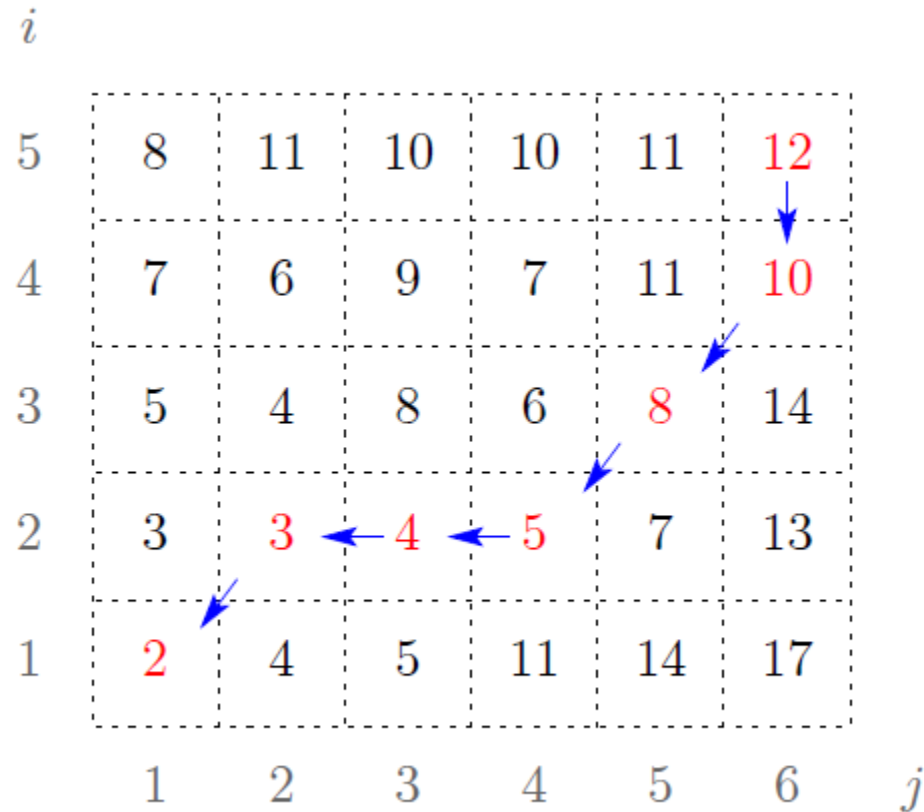
$$D(X, Y) = \text{DTW}(n, m)$$

- Distance measurement
 - For the distance calculation $d(x, y)$, different dimensions can be used
 - E.g. the Euclidian distance or (usually) the squared distance

Remarks on the process

- To obtain a formulation of the DTW distance independently of the total length of the time-series, this must be divided by the length of the warping path K .
- The warping path itself can be determined from the DTW matrix using backtracking.
 - The starting point is the position (n, m) , i.e. the end of the path
 - The smallest previous entry is determined step by step until the position $(1,1)$ is reached
 - I.e. from the last point to the starting point following the smallest values
- The number of steps results in K .

Backtracking within the DTW matrix:



Problem:

- DTW path may degenerate
- I.e.: optimal path is along the diagonal, unfavourable path is at the edge
- DTW path is restricted accordingly by a boundary condition.

Solution:

- Limitation of the warping path with regard to deviation from the diagonal

Constraints

- For all $w_k = (i_k, j_k)$ of the path with $1 \leq k \leq K$ and given angle α as well as $\beta = \arctan\left(\frac{n}{m}\right)$ must hold:

- Maximum absolute deviation (maximum “temporal” distance between two assignments:

$$|i_k - \tan(\beta) \cdot j_k| \leq w$$

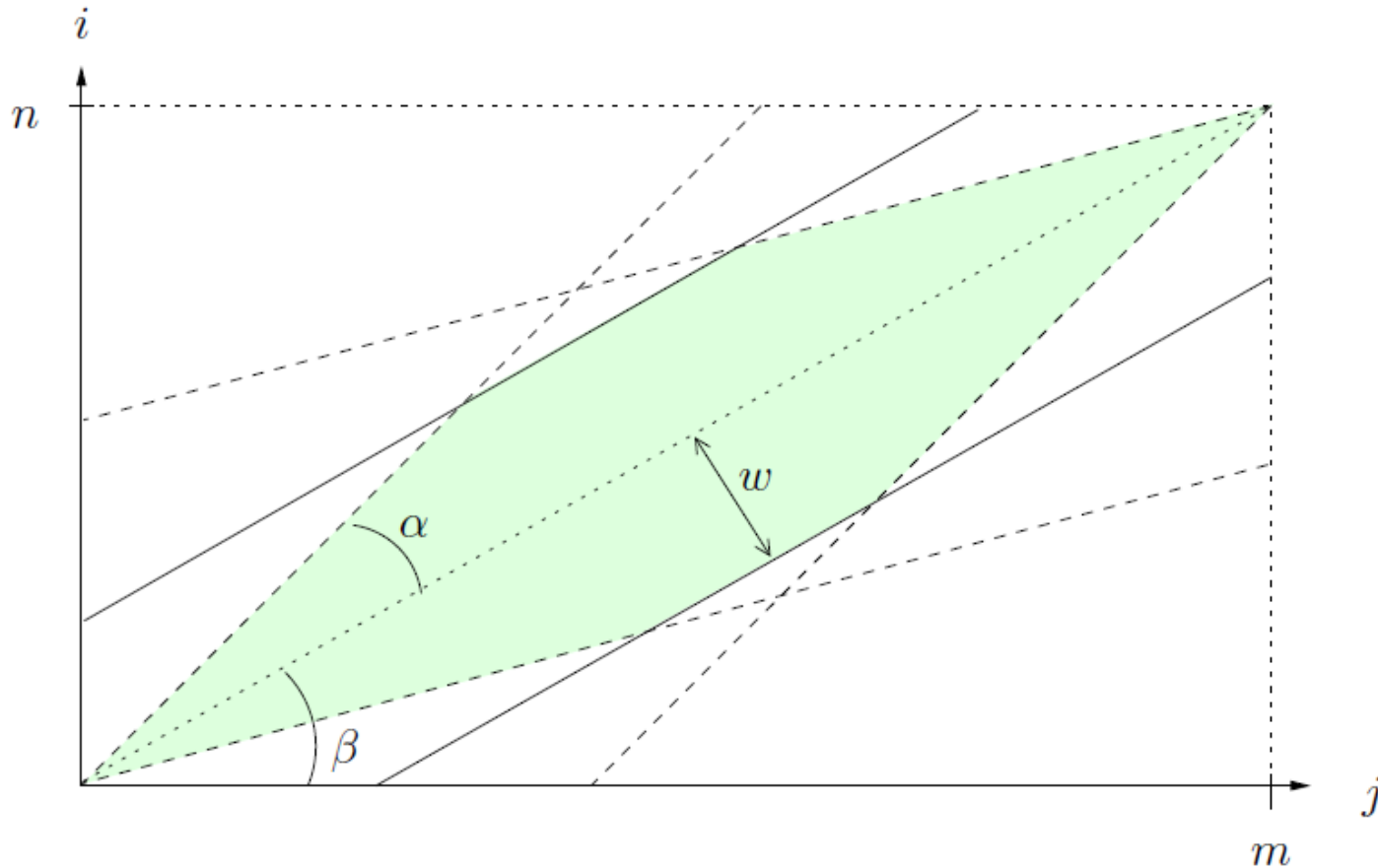
- Maximum relative deviation

$$i_k \leq \min(\tan(\beta + \alpha) \cdot j_k; n - \tan(\beta - \alpha) \cdot (m - j_k))$$

- And:

$$i_k \geq \min(\tan(\beta - \alpha) \cdot j_k; n - \tan(\beta + \alpha) \cdot (m - j_k))$$

Result:



- Another possibility to avoid a too large deviation of the path from the diagonal is the so-called slope factor $\phi \in \mathbb{R}^+$:

$$DTW(i, j) = d(x_i, y_j) + \begin{cases} 0 & \text{for } i = 1, j = 1 \\ \phi \cdot DTW(i, j - 1) & \text{for } i = 1, j > 1 \\ \phi \cdot DTW(i - 1, j) & \text{for } i > 1, j = 1 \\ \min \begin{pmatrix} \phi \cdot DTW(i - 1, j), \\ \phi \cdot DTW(i, j - 1), \\ DTW(i - 1, j - 1) \end{pmatrix} & \text{otherwise} \end{cases}$$

Edit Distance (ED)

- Edit distance, also called Levenshtein distance
- ED specifies the minimum number of insert, delete and replace operations necessary to convert one string to another.
- Calculation for two symbol sequences X and Y of the lengths n and m :

$$ED(i, j) = \min \begin{cases} ED(i-1, j-1) & \text{if } x_i = y_i \\ ED(i-1, j-1) + 1 & \text{(Replacement)} \\ ED(i, j-1) + 1 & \text{(Insertion)} \\ ED(i-1, j) + 1 & \text{(Deletion)} \end{cases}$$

with $ED(0,0) = 0$, $ED(i, 0) = i$ and $ED(0, j) = j$.

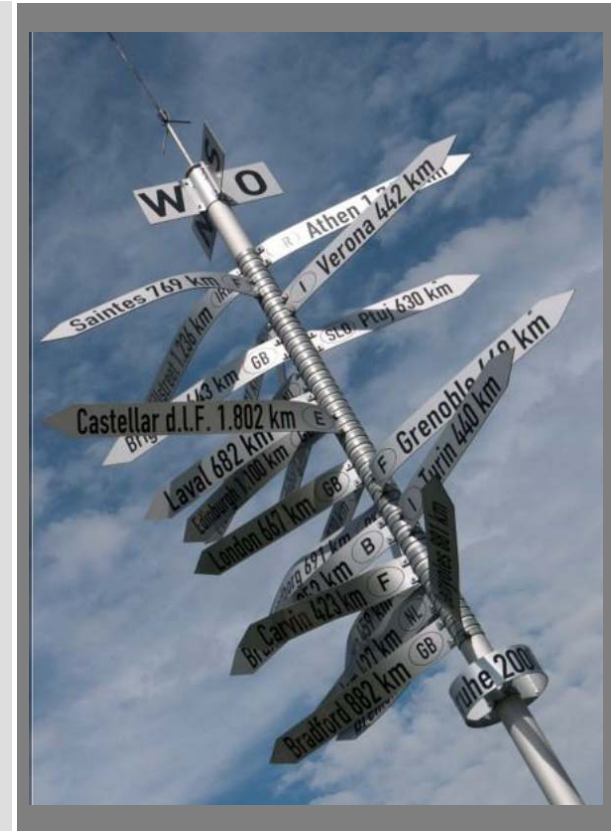
Edit Distance

- The total distance is given via $ED(n, m)$.
- If, in addition to the distance, the sequence of the operations is also of interest, a backtrace must be performed in the same way as for DTW.
- Special variants for time-series:
 - DER: Use of threshold values to map continuous distances between data values to “equal” or “unequal”
(Source: Chen, Öszu, Oria, Robust and Fast Similarity Search for Moving Object Trajectories, 2005)
 - TWED: temporal and spatial distance of data values during insertion, deletion, and adjustment
(Source: Marteau, Time Warp Edit Distance with Stiffness Adjustment for Time-Series Matching, 2009)

Further distance dimensions

- Many other distance measures and variations possible taking into account different aspects.
- Both on raw data and different representations or features, such as PCA, SVD, etc..
- A good overview of further measures can be found for example in Section 2.2.4 of [Mitsa 2010].

- Fundamentals of similarity measurement
- Dynamic similarity measures for time series
- **Similarity measures for time series models**
- Conclusion and further readings

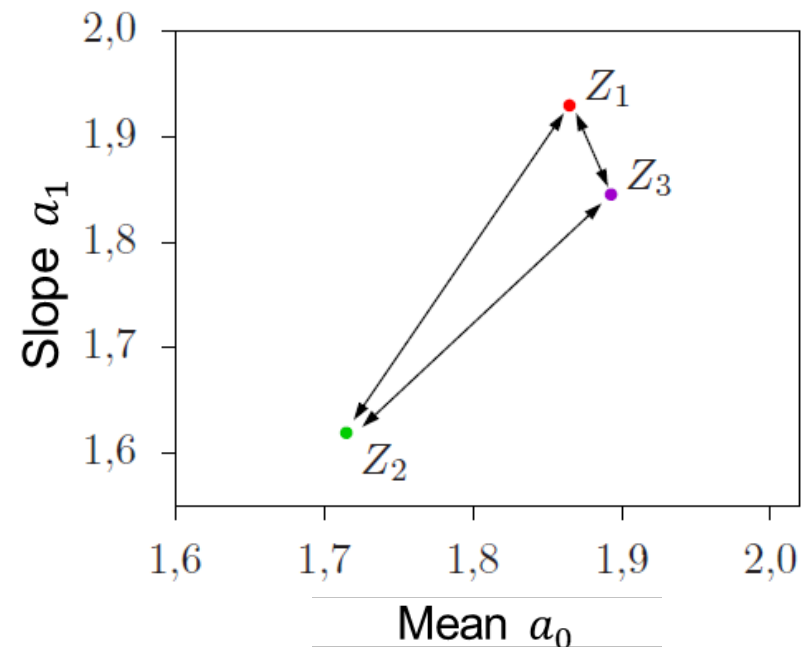


Distance measurements on models

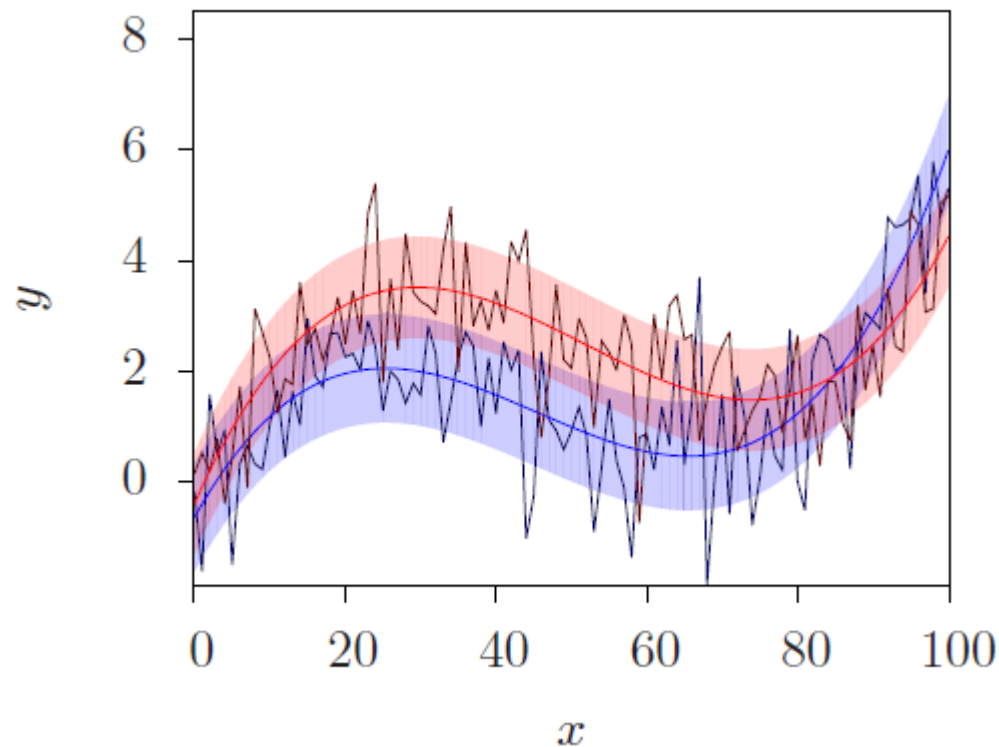
- Instead of a similarity measurement on raw data or a representation, time series models can also be used.
- Possibilities are available:
 - Comparison of **model parameters** (which can be treated similarly to feature vectors)
 - **Own distance measures** which compare models based on their properties (e.g. probabilistic models using divergence measures)
 - Comparison of **models with time series** (e.g. "How well does an unknown time series fit into a trained model?")

Shape Space Distance

- Comparison of the trend shares of each time-series
- Possible variations:
 - Additional consideration of the approximation error
 - Non-consideration of the average a_0
 - Etc.



Probabilistic model



- Interpretation of each model as a time-varying normal distribution
- Comparison of two models using divergence measures

Probabilistic model: Divergence measures

- Kullback-Leibler divergence

$$KL(u||v) = \int_{-\infty}^{+\infty} u(x) \ln \frac{u(x)}{v(x)} dx$$

- For normal distributions with means μ_u and μ_v as well as variances σ_u^2 and σ_v^2 :

$$KL(u||v) = \frac{(\mu_u - \mu_v)^2}{2\sigma_v^2} + \frac{1}{2} \left(\frac{\sigma_u^2}{\sigma_v^2} - 1 - \ln \frac{\sigma_u^2}{\sigma_v^2} \right)$$

- Symmetric variant:

$$KL_2(u||v) = KL(u||v) + KL(v||u)$$

Probabilistic model: Divergence measures

- Comparison of two time-series: Evaluation of two time-series models at temporal positions x_1, \dots, x_N and calculation of the average distance:

$$d_{KL_2}(X, Y) = \frac{\sigma_1^2}{2\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 1 + \frac{\sigma_1^{-2} + \sigma_2^{-2}}{2(N+1)} \sum_{n=0}^N (p_1(x_n) - p_2(x_n))^2$$

- Other divergence and distance measures defined for probability distributions such as Bhattacharyya or Hellinger distance are possible as well.

Fisher score

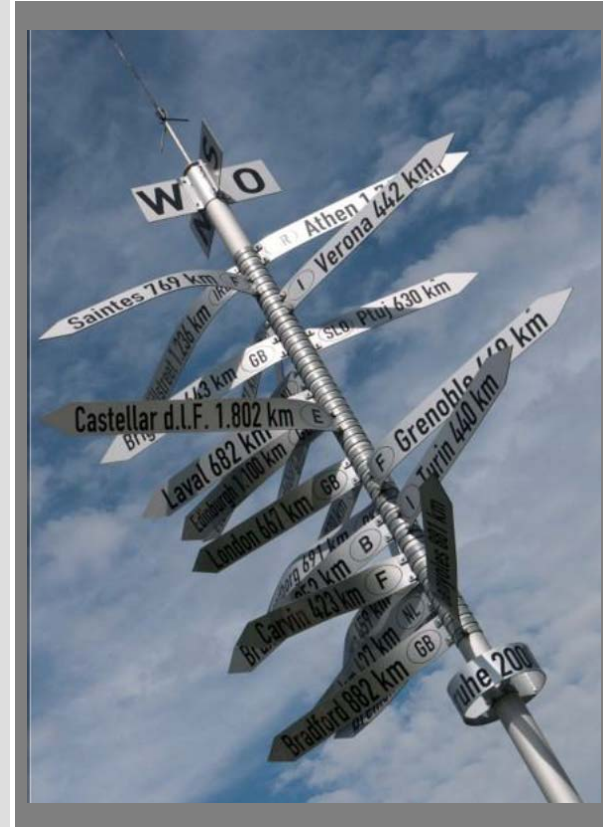
- By calculating the so-called Fisher score vector, it can be determined how well a given realisation O “fits” a trained model.
- For this purpose, the logarithm of the likelihood (so-called log-likelihood) of the realisation is derived according to the individual parameters.
- The Fisher score vector of a model $\theta = \{\theta_1, \dots, \theta_k\}$ with k model parameters and a given realisation O :

$$\nabla_{\Theta}(O) = \left(\frac{\partial \log p(O|\Theta)}{\partial \theta_1} \dots \frac{\partial \log p(O|\Theta)}{\partial \theta_k} \right)^T$$

Fisher score (cont.)

- For the comparison of two time-series X and Y the Fisher score vectors $\nabla_{\Theta}(X)$ and $\nabla_{\Theta}(Y)$ are calculated now.
- These can then be compared on vectors using any distance and similarity measures.
- Further literature on this subject:
[Taylor, Christianini, Kernel Methods for Pattern Analysis, 2004]

- Fundamentals of similarity measurement
- Dynamic similarity measures for time series
- Similarity measures for time series models
- Conclusion and further readings



Summary of the chapter

- Basics: distance directly based on raw data (features), based on a corresponding representation, or based on a model created from the data.
- Euclidian, Mahalanobis and Diagonal Norm, Hamming Distance
- Similarity measures for time series: special similarity measures on time series to consider dynamic temporal relationships, often additional processing of time series of different lengths possible.
- Techniques: Longest Common Subsequences vs. Dynamic Time Warping vs. Edit Distance
- Similarity measures on time series models: Comparison of model parameters, own distance measures (comparison of models based on their properties), comparison of models with time series

- [TC04] Taylor, Christianini: Kernel Methods for Pattern Analysis, 2004
- [CÖO05] Chen, Özsu, Oria: Robust and Fast Similarity Search for Moving Object Trajectories, 2005
- [Mitsa 2010] Mitsa, Theophano. Temporal data mining. CRC Press, 2010.
- [Marteau 2009] Marteau: Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching, 2009
- [AS95] Agrawal, Lin, Sawhney, Shim: Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases 1995

- Any questions...?