Christian-Albrechts-Universität zu Kiel

Intelligent Systems

Chapter 6: Similarities

Winter Term 2019 / 2020

Prof. Dr.-Ing. habil. Sven Tomforde
Institute of Computer Science / Intelligent Systems group

About this Chapter

Christian-Albrechts-Universität zu Kiel

Content

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

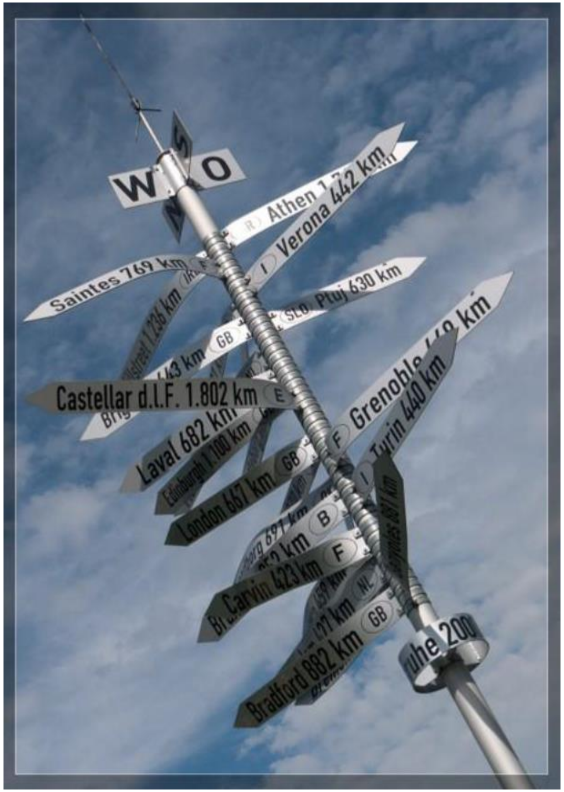
Goals

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.

Agenda

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



Basics

Similarity

- Comparison of samples, i.e. distance or similarity measurement, is necessary in all further 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
  - with regard to a model created from the data.

Basics (2)

Element-by-element distance

- Simplest distance between two patterns: Minkowsky norms
- For time series, application, for example, to feature vectors, or evaluation by element, provided that time series have the same length (otherwise, for example, interpolation).
- Distance between two time series X and Y (with same length N):

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

where x_i and y_i are the i-th elements of the two time series.

  - p = 1 - Manhattan Distance
  - p = 2 - Euclidean distance
  - …

Basics (3)

Multivariate time series

- For multivariate (n-dimensional) time series, the distance can also be defined:

$$D(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{y}_i||$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ represent the i-th elements (here: n-dimensional vectors).

- Instead of the Euclidean distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, 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 $M \in \mathbb{R}^{n \times n}$.

Basics (4)

- For:

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

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

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

Basics (5)

Christian-Albrechts-Universität zu Kiel

Mahalanobis norm

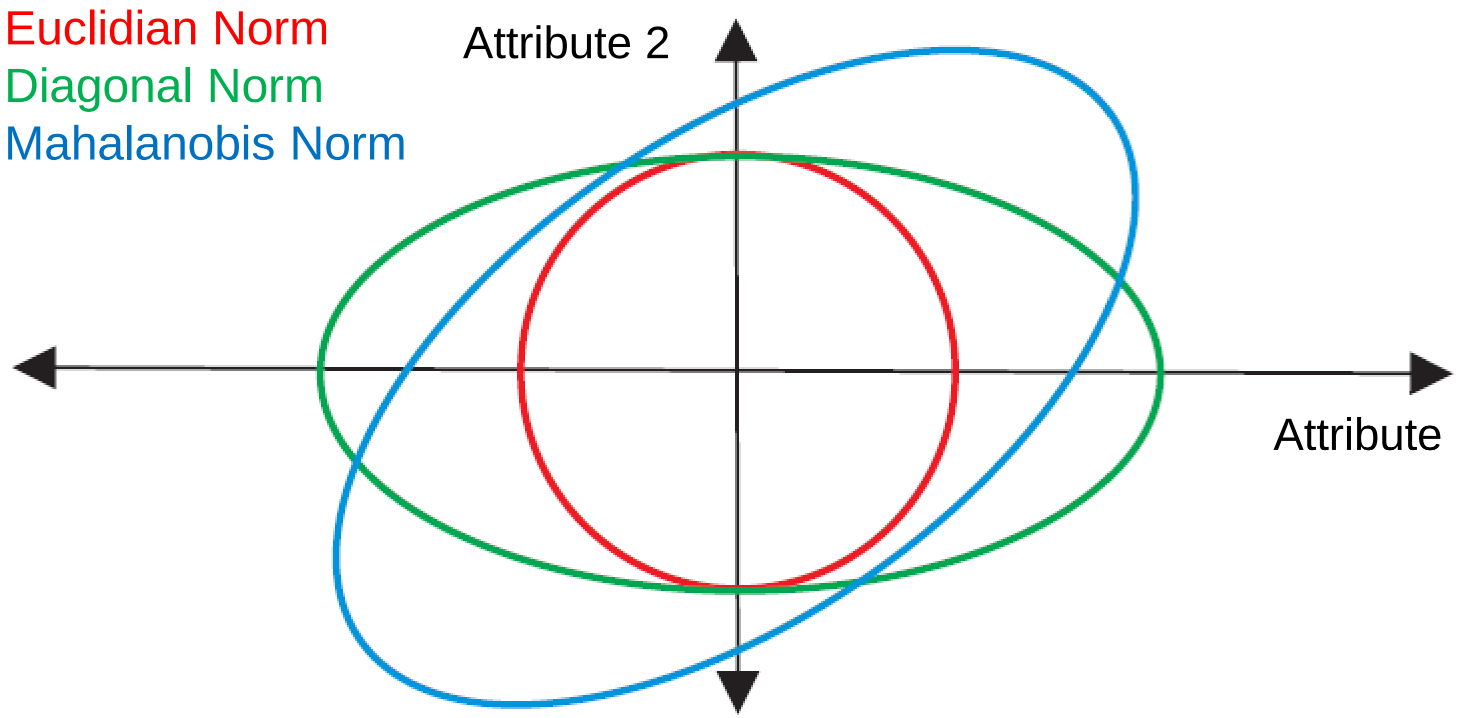
- Defined by the inverse of the covariance matrix of the data values:

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

- with the mean value of:

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

Basics (6)



- 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).

Basics (7)

Christian-Albrechts-Universität zu Kiel

Further examples of distance dimensions:

- Cosine distance: normalised standard scalar product of two vectors (cosine of the angle):

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

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

Basics (8)

Christian-Albrechts-Universität zu Kiel

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:

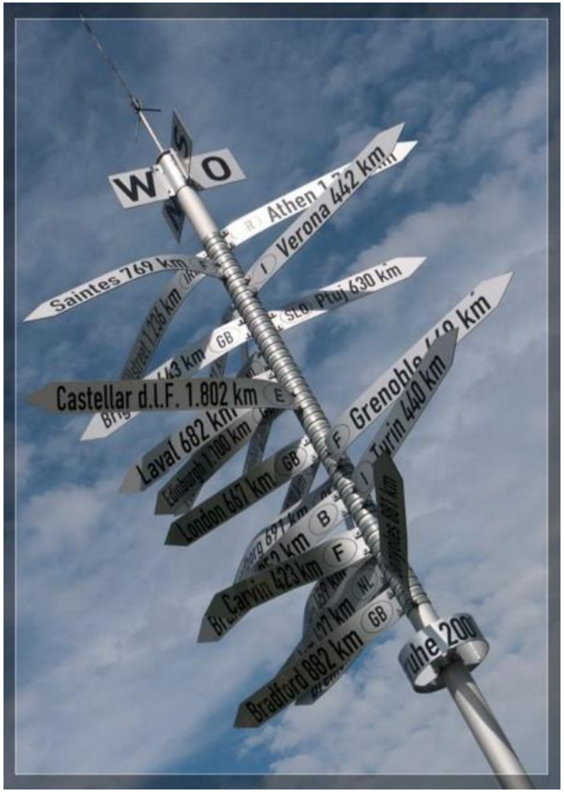
0 0 1 1 0
0 0 1 0 0
Hamming distance: 1

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

Agenda

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



Dynamic similarity measures

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

Dynamic similarity measures (2)

Longest Common Subsequences (LCSS)

- Goal: To 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)

Dynamic similarity measures (3)

Christian-Albrechts-Universität zu Kiel

Longest Common Subsequences (cont.)

- Given are two time series $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$ of the lengths n and m
- The search is for the longest common partial sequence of both time series, taking into account local scaling and translation in the value range
- Solution: Longest Common Subsequences on time series

Source: [Agrawal, Lin, Sawhney, Shim, Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases 1995]

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:

$$|s_i - t_i| \leq \epsilon$$

for a given $\epsilon \in \mathbb{R}$ and for $i = 1, \dots, w$.

- Initially, all connected and related sub-sequences of length w are extracted from the time series X and Y :

$$\tilde{S}_i = (x_i, \dots, x_{i+w-1}) \text{ mit } i = 1, \dots, n - w + 1$$
$$\tilde{T}_j = (y_j, \dots, y_{i+w-1}) \text{ mit } j = 1, \dots, n - w + 1$$

Dynamic similarity measures (5)

Christian-Albrechts-Universität zu Kiel

LCSS - Step 1: Atomic Matching (continued)

- All partial sequences are normalised (e.g. 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 subsequence of X and a subsequence of Y is called “Atomic Matching” (of the length w).

Note:

- Instead of a comparison of n - w + 1 partial sequences of X with m - w + 1 partial sequences of Y an efficient search for similar partial sequences using a suitable index structure is also possible!

Dynamic similarity measures (6)

LCSS - Step 2: Formation of longer partial sequences

- Goal: Atomic similarities are now combined (with other atomic similarities or already combined sections) to longer sequences.
- 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: $length(S_{i,1})$ and $length(T_{j,1})$ are functions determining the lengths of $S_{i,1}$ and $T_{j,1}$ (number of samples / data points)

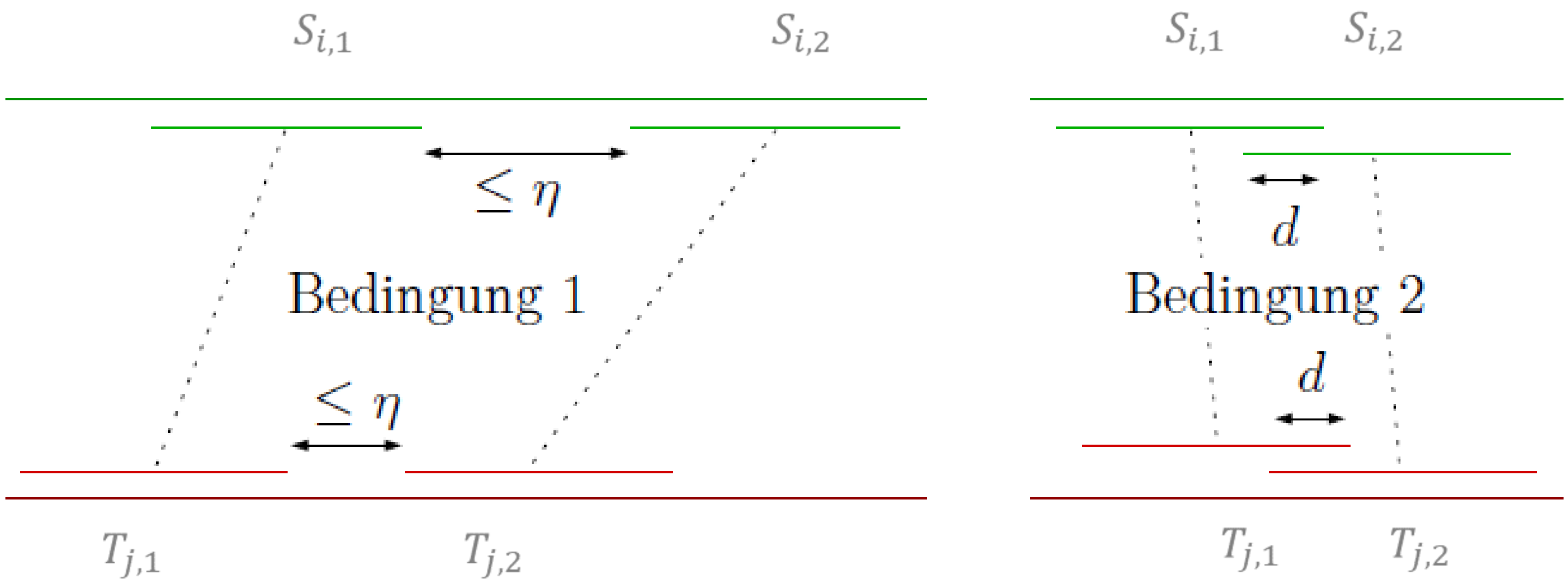
Dynamic similarity measures (7)

- Matchings can be combined into longer sequences if one of the following conditions is met:
  1. Sequences $S_{i,1}$ and $S_{i,2}$ do not overlap on X (i.e.: $i_1 + lenght(S_{i,1}) < i_2$) and their distance is not greater than a fixed value $\eta \in \mathbb{N}$ (i.e. $i_1 + length\left(S_{i,1}\right) + \eta \geq i_2$). The same conditions must also apply to sequences $T_{j,1}$ and $T_{j,2}$ on Y.
  2. 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 $d = i_1 + length\left(S_{i,1}\right) - i_2 = j_1 + length\left(T_{j,1}\right) - j_2$.
- In addition, a certain similarity of the scaling factors used to scale the partial sequences involved may be required for a combination of matches.

Dynamic similarity measures (8)

Christian-Albrechts-Universität zu Kiel

Possibilities for constructing longer sub-sequences



Dynamic similarity measures (9)

LCSS - Step 3: Finding the longest match

- Now, all matches are combined as much as possible and k pairs of matches $(S'_1; T'_1), \ldots, (S'_k; T'_k)$ are given.
- We are now looking for the subset $(S'_{l_1}; T'_{l_1}), \ldots, (S'_{l_h}; T'_{l_h})$, for which the following requirements hold:
  - The end point of S'_{l_i} is before the start point of S'_{l_j} on X and the end point of T'_{l_i} is before the start point of T'_{l_j} on Y for $\leq i < j \leq h$ (i.e. the sub-sequences 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)
  - The total length of all sequences $\sum_{i=1}^{h} length(S'_{l_i}) + \sum_{i=1}^{h} length(T'_{l_i})$ is maximal.

Dynamic similarity measures (10)

Christian-Albrechts-Universität zu Kiel

- Let $\circ$ be the sequential composition of two sub-sequences, then $X' = S'_{l_1} \circ \ldots \circ S'_{l_h}$ and $Y' = T'_{l_1} \circ \ldots \circ T'_{l_h}$ are the longest common sub-sequences of the time series X and Y.
- In contrast to LCSS on symbol sequences, X' and Y' can have different lengths (see condition 1 in step 2).
- Advantage of the procedure: flexible assignment of partial sequences of two time series to each other and thus robust comparison despite scaling, translation and longer sections that do not match.
- Disadvantage: high computing effort

Dynamic similarity measures (11)

Christian-Albrechts-Universität zu Kiel

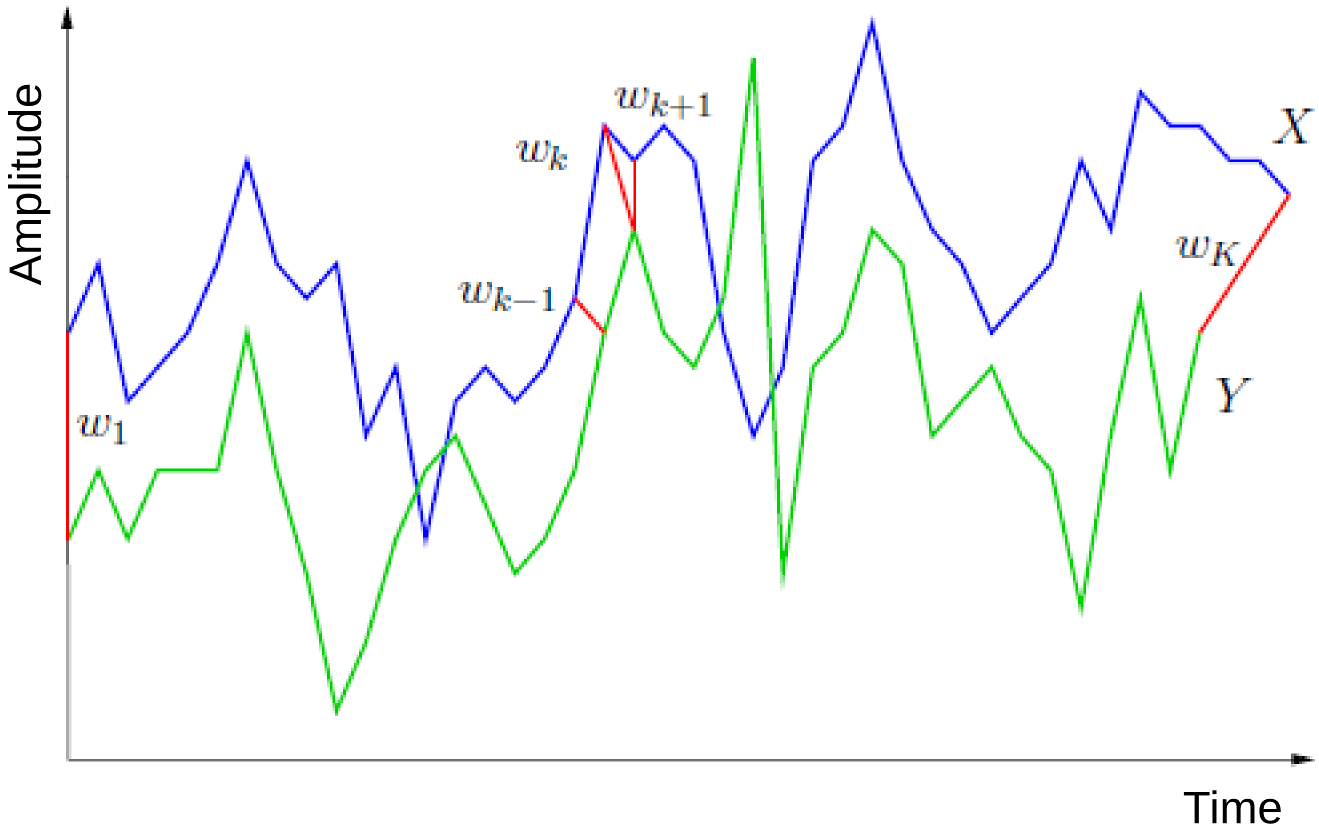
Dynamic Time Warping (DTW)

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

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

Dynamic similarity measures (12)

Example



Dynamic similarity measures (13)

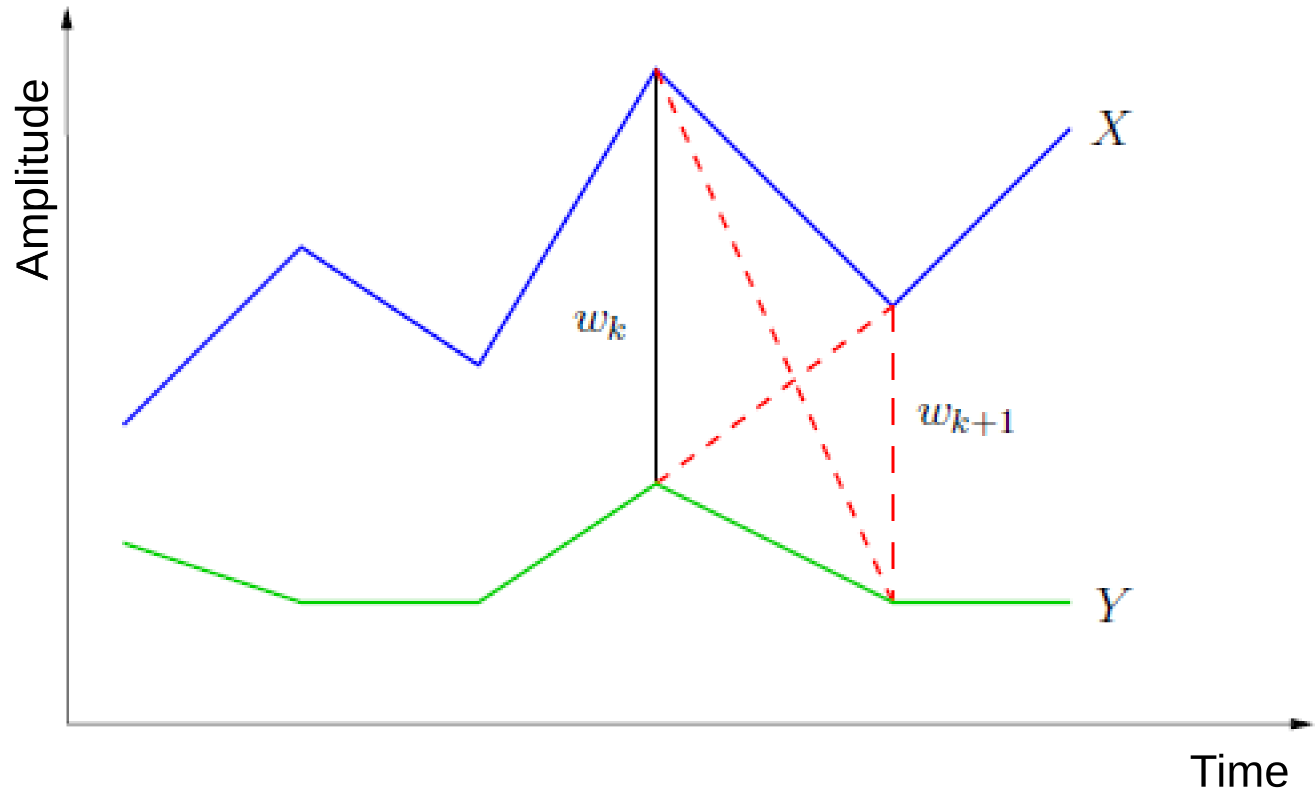
The following additional restrictions apply to the warping path:

- Boundary condition: $w_1 = (1,1)$ and $w_K = (n, m)$, i.e. the path begins with the first element and ends on both with the last element on both 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 contiguous, 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 $i_{k+1} - i_k \geq 0$ and $j_{k+1} - j_k \geq 0$ must apply (i.e., the warping path assignments maintain the chronological order of the data points of both time series).

Dynamic similarity measures (14)

Christian-Albrechts-Universität zu Kiel

Example



Dynamic similarity measures (15)

Determine the optimal warping path:

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

Dynamic similarity measures (16)

Christian-Albrechts-Universität zu Kiel

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

$$D(X,Y) = DTW(n,m)$$

- For the distance calculation $d(x,y)$ different dimensions can be used, e.g. the Euclidean distance or (usually) the squared distance.

Dynamic similarity measures (17)

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

Backtracking within the DTW matrix:

i							
5	8	11	10	10	11	12	
4	7	6	9	7	11	10	
3	5	4	8	6	8	14	
2	3	3	4	5	7	13	
1	2	4	5	11	14	17	
	1	2	3	4	5	6	j

Dynamic similarity measures (19)

Problem:

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

Solution:

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

Dynamic similarity measures (20)

Christian-Albrechts-Universität zu Kiel

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

- Maximum absolute deviation (maximum "temporal" distance between two assignments)

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

- Maximum relative deviation

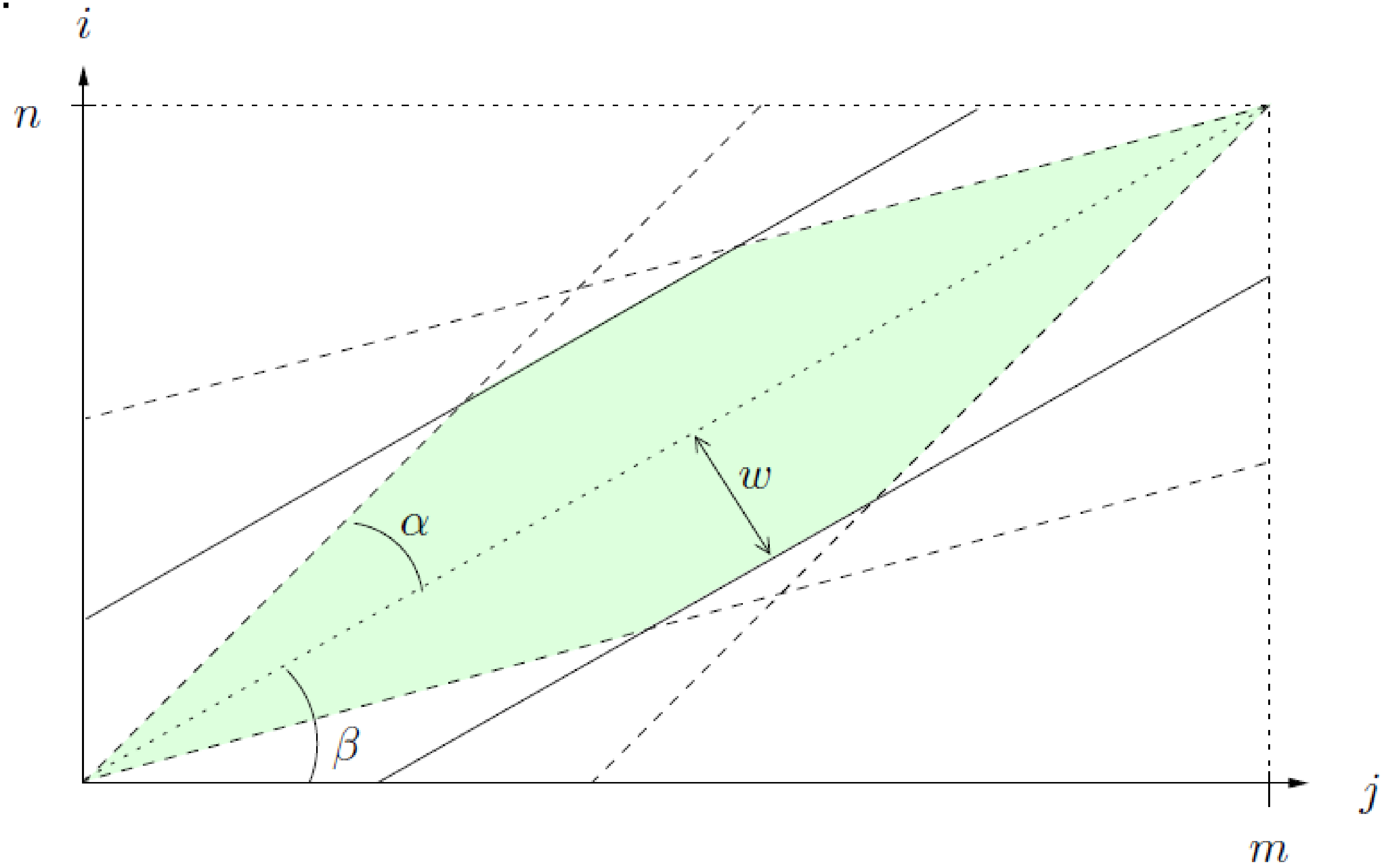
$$i_k \leq \min(\tan(\beta + \alpha) \cdot j_k \,, n - \tan(\beta - \alpha) \cdot (m - j_k))$$

- And:

$$i_k \geq \max(\tan(\beta - \alpha) \cdot j_k \,, n - \tan(\alpha + \beta) \cdot (m - j_k))$$

Dynamic similarity measures (21)

Result:



Dynamic similarity measures (22)

Christian-Albrechts-Universität zu Kiel

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

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

Christian-Albrechts-Universität zu Kiel

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 $ED(0,j) = j$

Dynamic similarity measures (24)

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 on time series:
  - EDR: Use of threshold values to map continuous distances between data values to "equal" or "unequal"

    (Source: Chen, Özsu, 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)

Dynamic similarity measures (25)

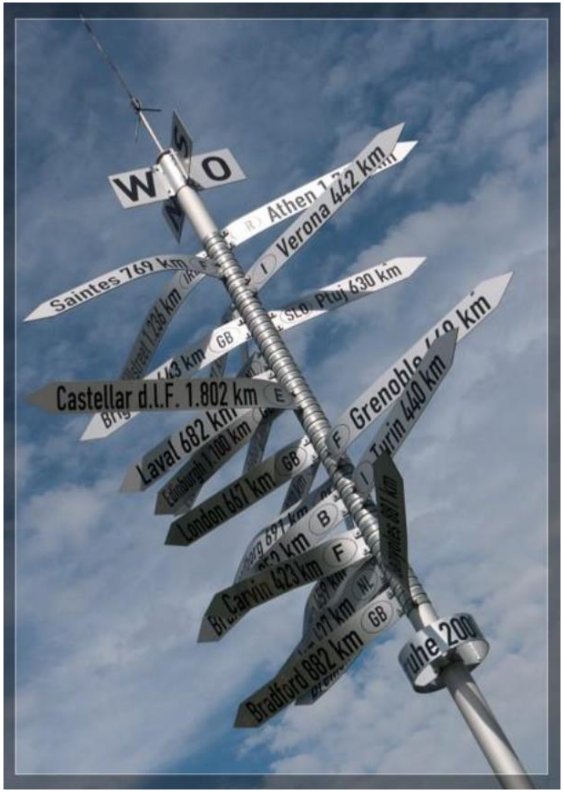
Christian-Albrechts-Universität zu Kiel

Further distance dimensions

- Many other distance measures and variations possible taking into account different aspects.
- Both on raw data and on 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].

Agenda

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



Similarity measures for models

Christian-Albrechts-Universität zu Kiel

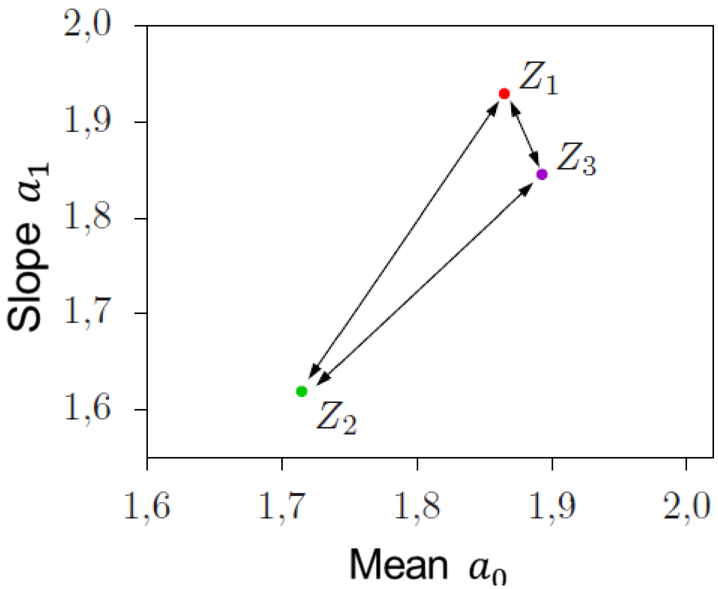
Distance measurements on models

- 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?")

Similarity measures for models (2)

Shape Space Distance

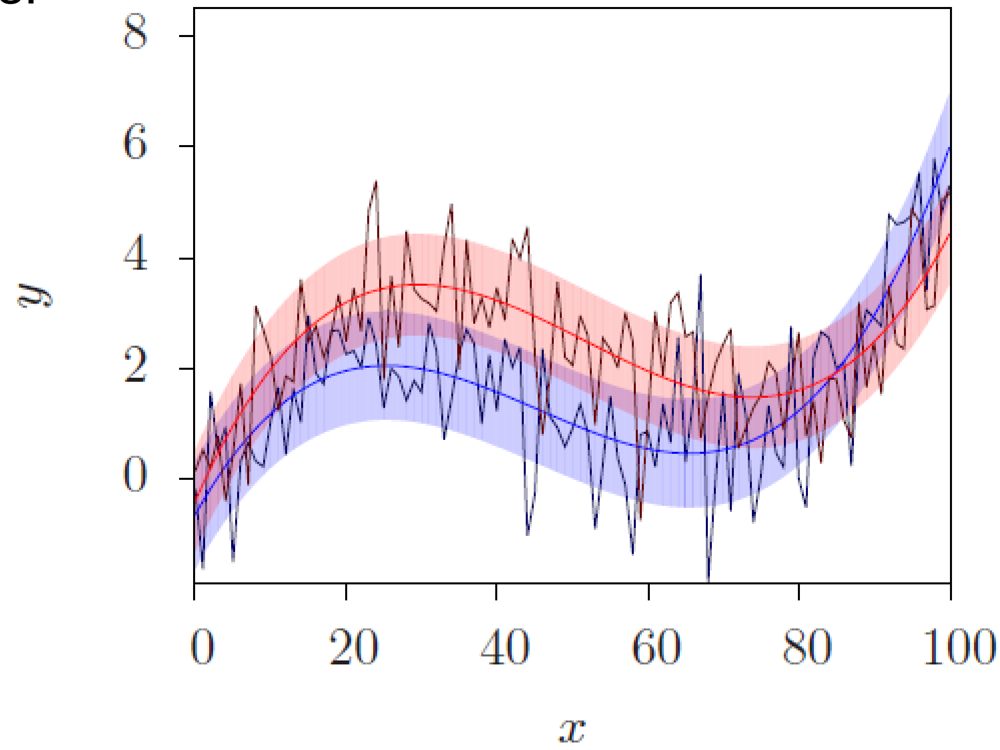
- Comparison of the trend shares of each time series
- Possible variations: additional consideration of the approximation error, non-consideration of the average a_0, etc.



Similarity measures for models (3)

Christian-Albrechts-Universität zu Kiel

Probabilistic model



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

Similarity measures for models (4)

Christian-Albrechts-Universität zu Kiel

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:

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

Similarity measures for models (5)

Christian-Albrechts-Universität zu Kiel

Probabilistic model: divergence measures (continued)

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

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

- Other divergence and distance measures on probability distributions such as Bhattacharyya or Hellinger distance possible

Similarity measures for models (6)

Christian-Albrechts-Universität zu Kiel

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.
- Fisher score vector of a model $\Theta = \{\theta_1, \ldots, \theta_k\}$ with k model parameters and given realisation O :

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

Similarity measures for models (7)

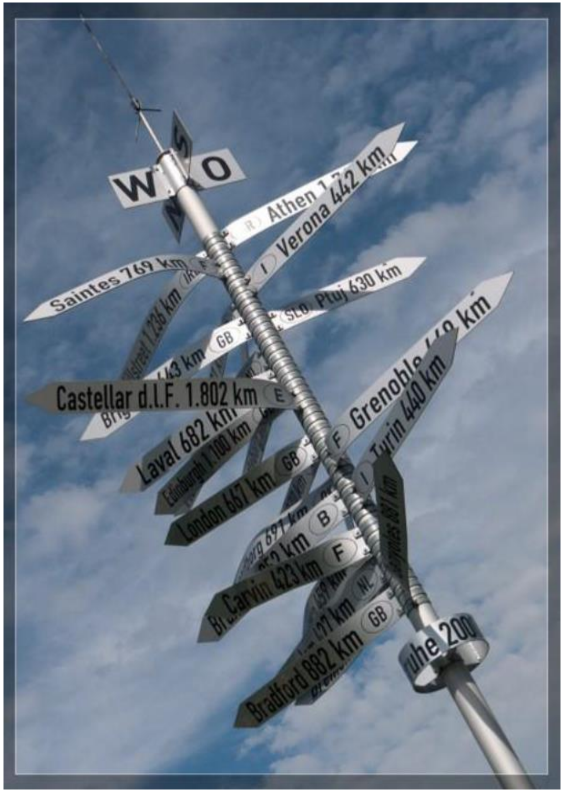
Christian-Albrechts-Universität zu Kiel

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]

Agenda

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



Conclusion

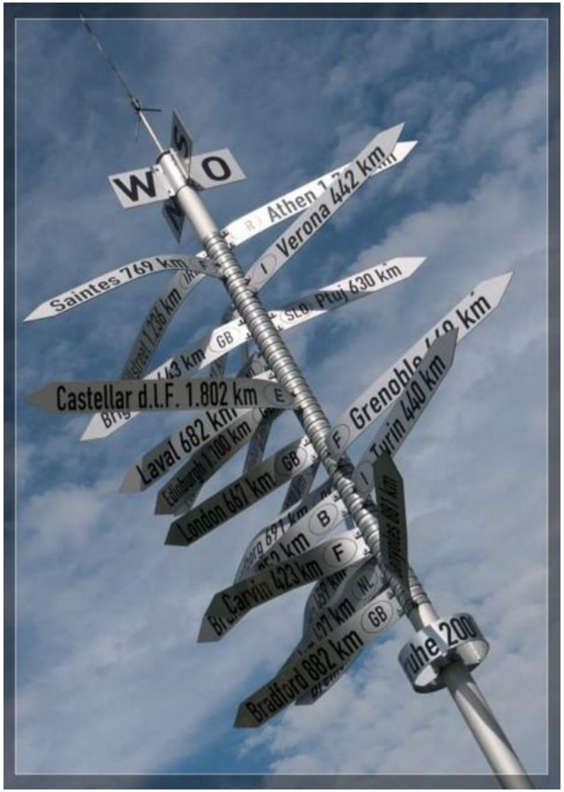
Christian-Albrechts-Universität zu Kiel

- Basics: distance directly based on raw data (features), based on a corresponding representation, or based on a model created from the data.
- Euclid, 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

Agenda

Christian-Albrechts-Universität zu Kiel

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



Further readings

- [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

End

Christian-Albrechts-Universität zu Kiel

- Questions….?