Christian-Albrechts-Universität zu Kiel

Intelligent Systems

Chapter 10: Motif Search and Anomalies

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

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings

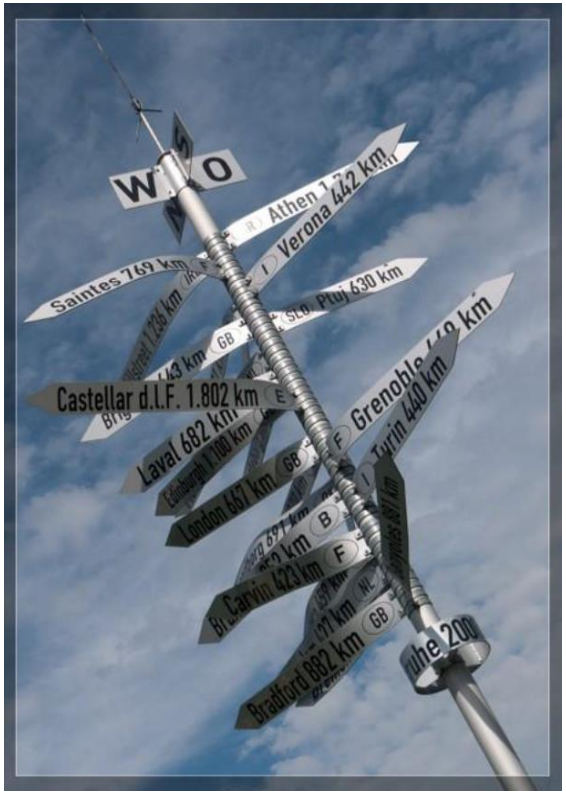
Goals

Students should be able to:

- motivate and define the process of motif detection.
- explain the different classes of anomaly detection approaches.
- distinguish between outlier, anomaly, novelty, and noise.
- describe and apply the most prominent novelty detection algorithms.

Agenda

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings

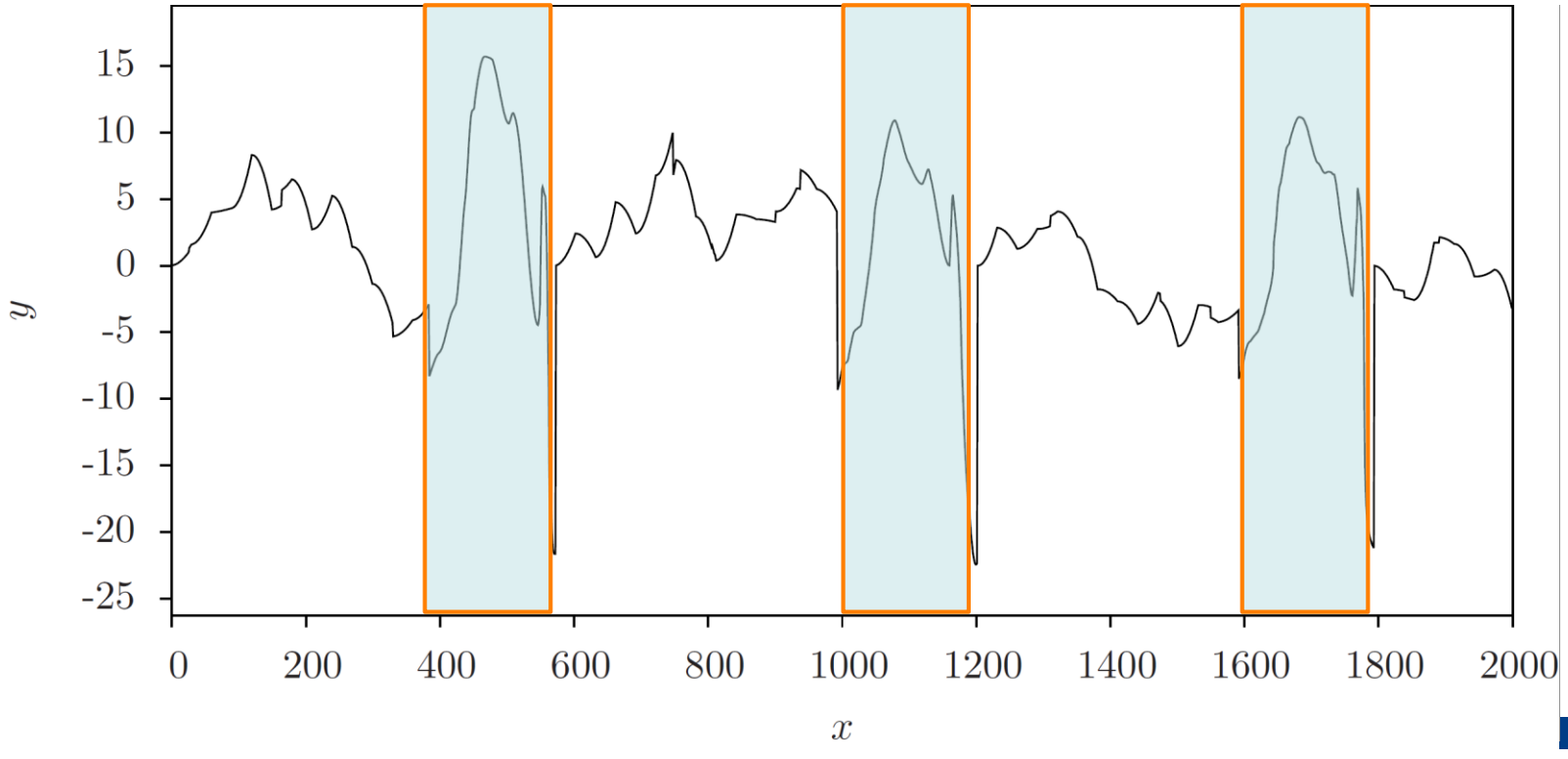


Motif Search

- Frequent task: finding interesting "patterns" in a lot of data
- Definition of "interesting" is very application-dependent
- Interesting can be anything that is useful for classification, clustering, prediction, etc.
- Many possible terms for such temporal patterns:
  - "temporal pattern"
  - "episode"
  - "similar sequence / trend / shape"
  - ...

Motif Search (2)

- Often "interesting" is correlated with "frequent
- The determination of frequently (repeatedly) recurring patterns is referred to as "frequent pattern mining" or "frequent episode discovery".



Motif Search (3)

Motif search / definition

- Repeatedly recurring patterns are also referred to as “motifs”.
- Term originally comes from bioinformatics: motives known on gene sequences
- With the information about repeated patterns or sequences, many different tasks can be solved, e.g.:
  - Determining the class affiliation of a pattern based on the presence of one or more motifs
  - Detection of disturbances or anomalies due to the absence of certain motifs
  - ...

Motif Search (4)

Motif search / definition (continued)

- First of all, a distinction must be made between motif definition and motif search.

Motif definition:

- Recognition of similarly recurring subsequences of a time series or a set of time series
- Various specifications possible, such as:
  - Length of the sequences,
  - Number of different types,
  - Degree of similarity,
  - etc.
- In addition to recognition, a suitable representation of the determined example sequences is also necessary.

Motif Search (5)

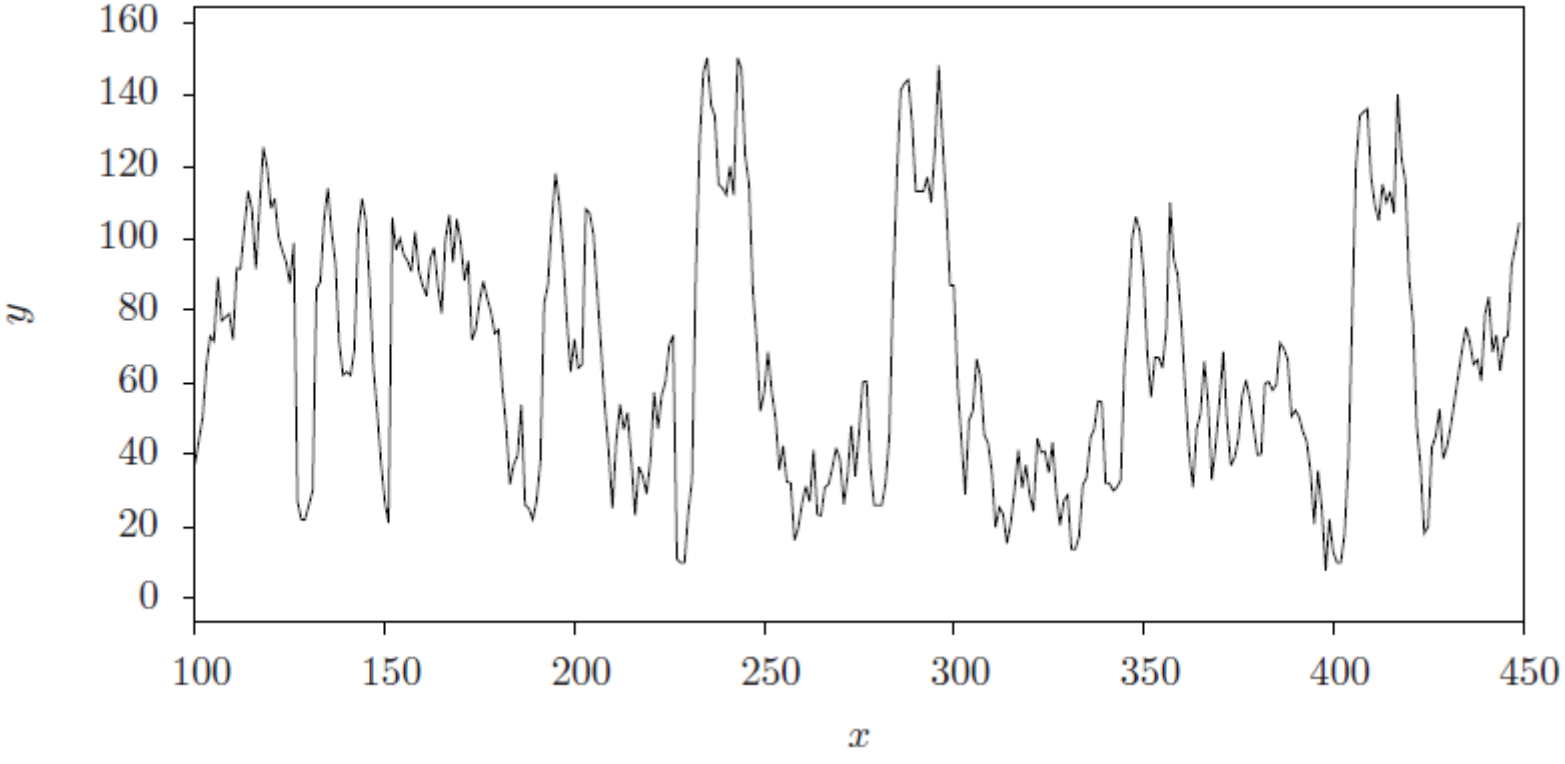
Motif search / definition (continued)

- Further subdivision of the motif definition into:
  - supervised and
  - unsupervised
- Supervised here means that secondary knowledge (e.g. labels or timestamps) determines the areas in which interesting sequences can be found.
- A supervised motif definition can be represented by a classification if the groups are suitably selected.

Motif Search: Example

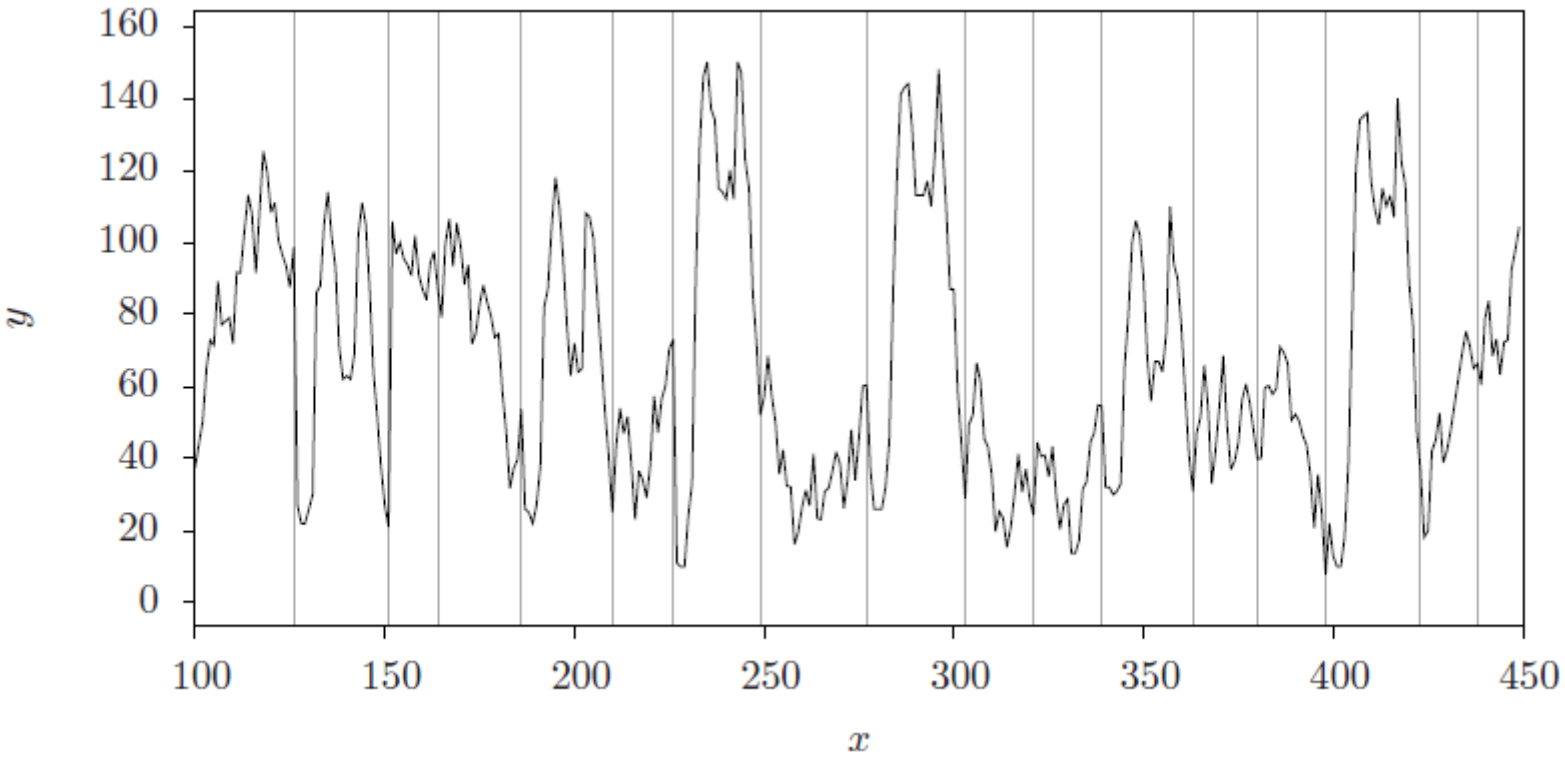
Christian-Albrechts-Universität zu Kiel

Example:



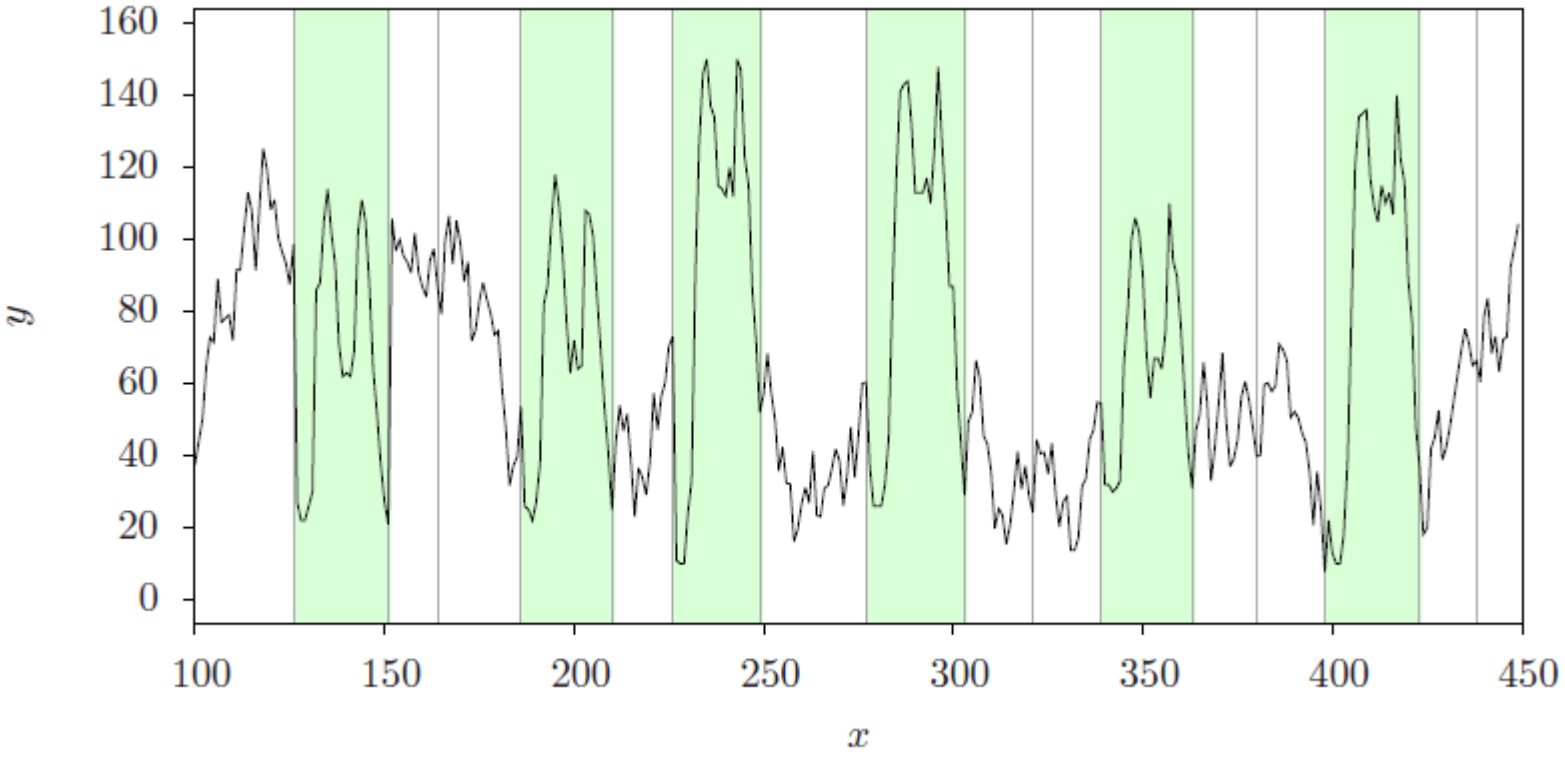
Motif Search: Example (2)

Example: Result of segmentation process



Motif Search: Example (3)

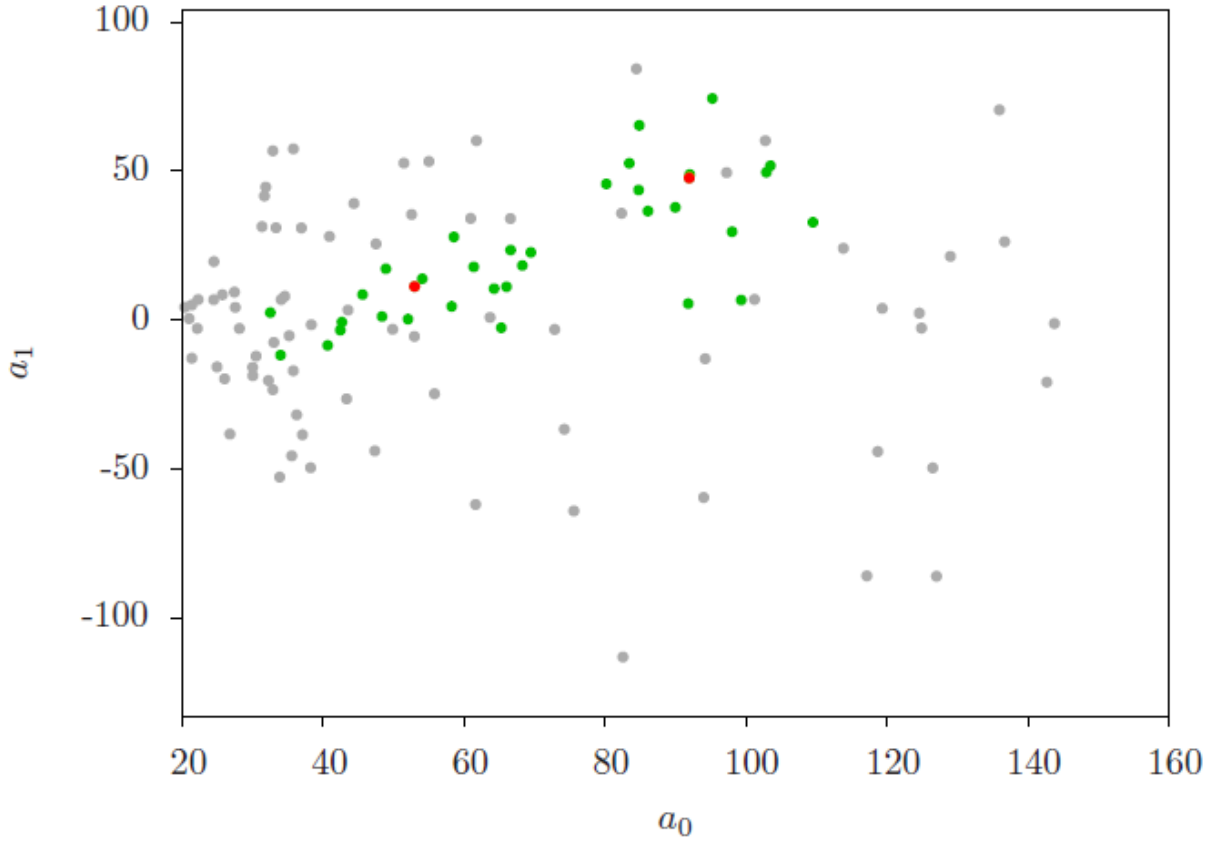
Example: Expert knowledge about interesting segments



Motif Search: Example (4)

Christian-Albrechts-Universität zu Kiel

Example: Scatterplot of trend aspects a_0 (average) and a_1 (slope)



Motif Search

Advantage:

- Actually relevant sections are used, not just randomly repeated (irrelevant) areas
- It should be noted that, despite the information about the beginning and end of relevant sections, the specified limits should not be used, but an appropriately selected segmentation should be used.

Disadvantage:

- If several types of examples exist, representation can be difficult (clustering may be necessary).

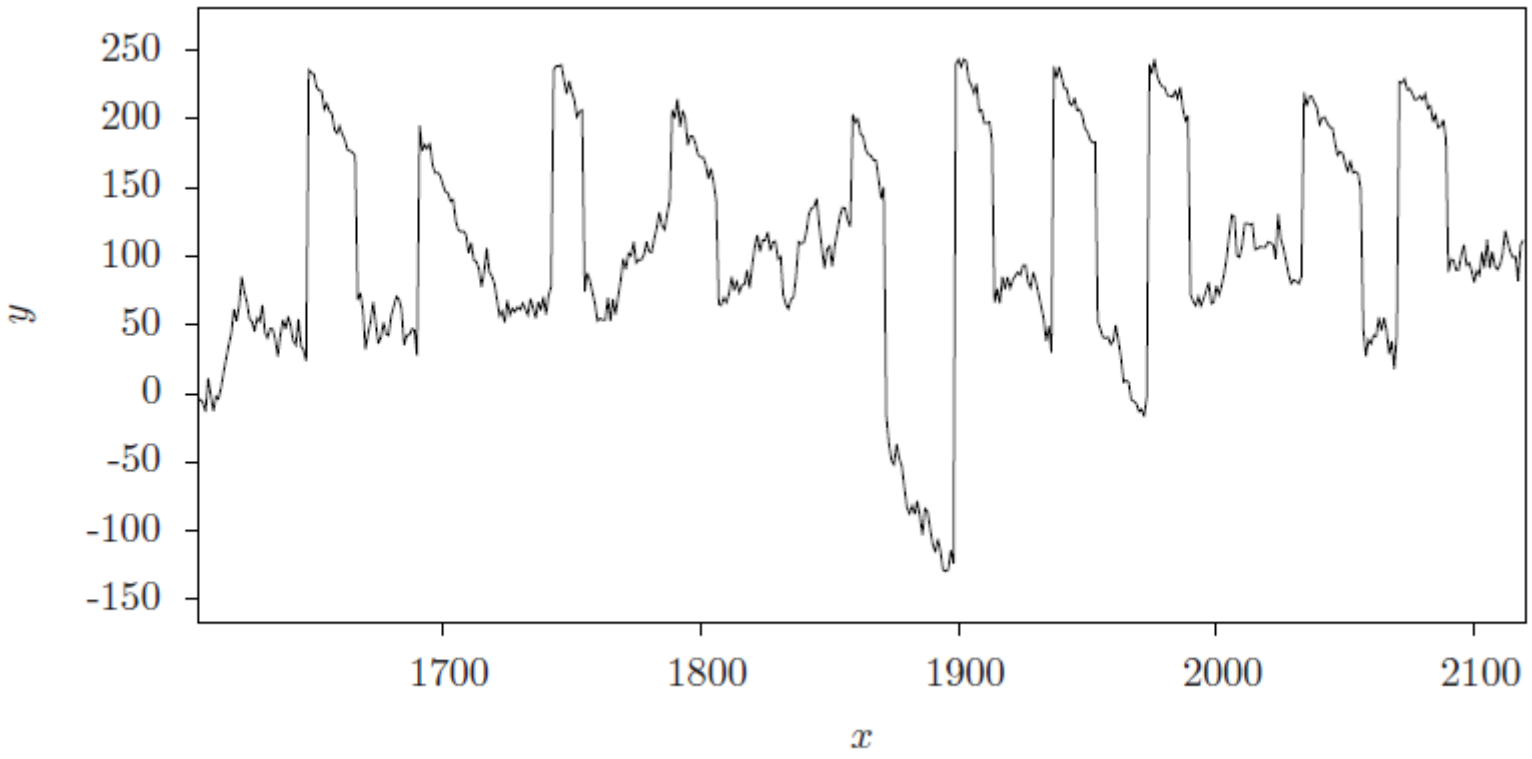
Unsupervised Motif Search

Unsupervised motif definition

- An unsupervised motif definition must get along without corresponding secondary knowledge.
- Many different approaches possible
  - Hashing of patterns
  - Use of suffix trees
  - Detection of clusters using index structures (R* trees, iSAX, ...)
  - Clustering of segments of time series
  - ...
- Some methods require a distinction between searching for a repeated sequence in a time series and a similar sequence in different time series (e.g. autocorrelation, cross-correlation, etc.).

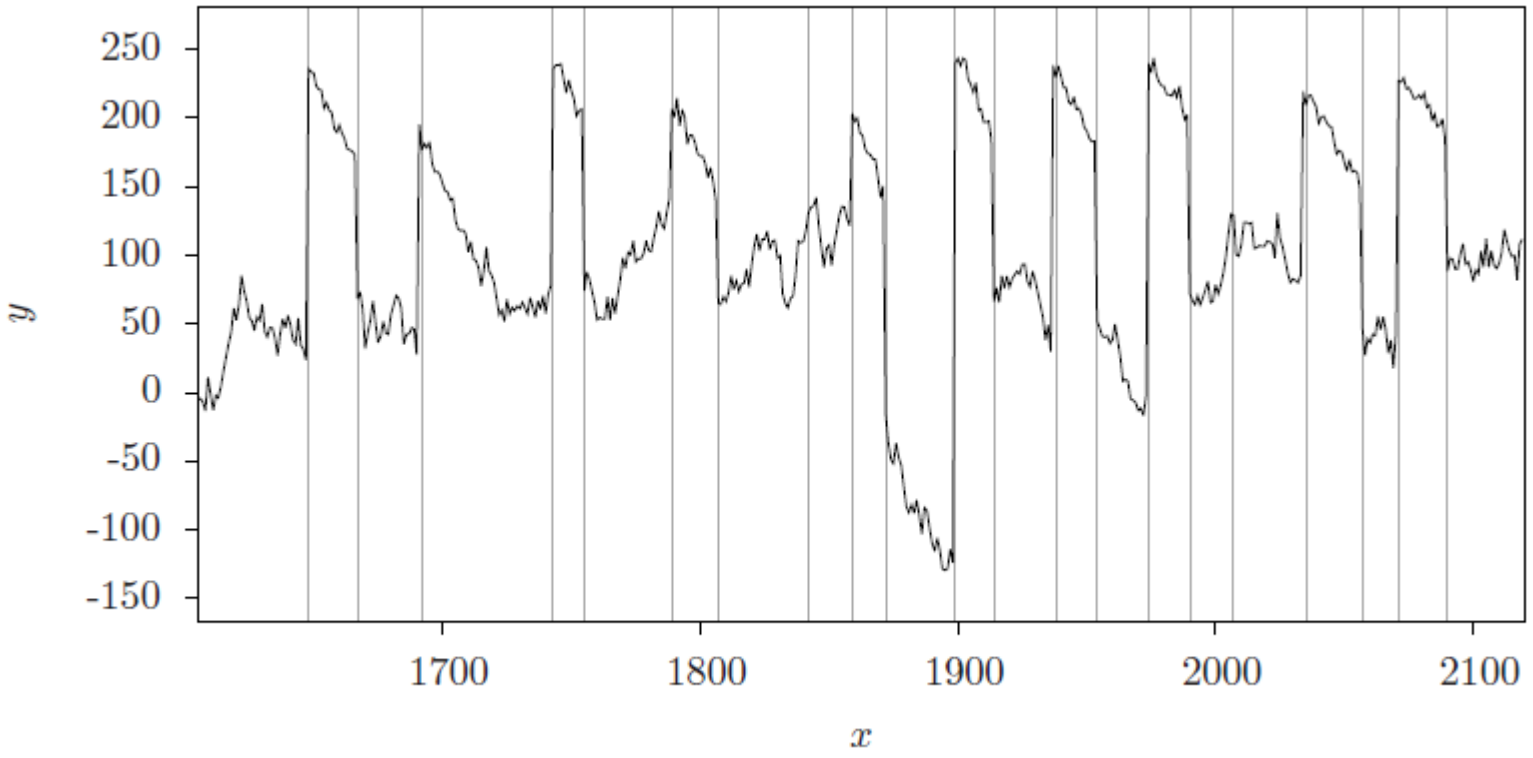
Unsupervised Motif Search (2)

Example: Unsupervised motif detection



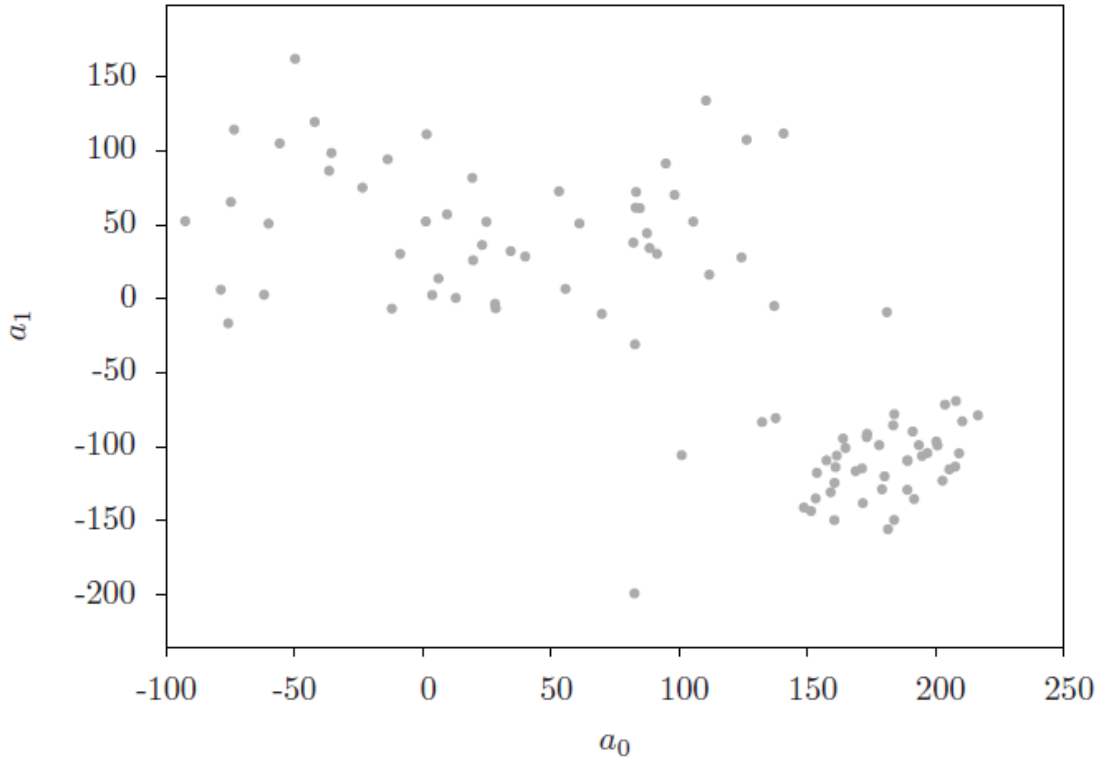
Unsupervised Motif Search (3)

Example for unsupervised motif detection: Result of segmentation



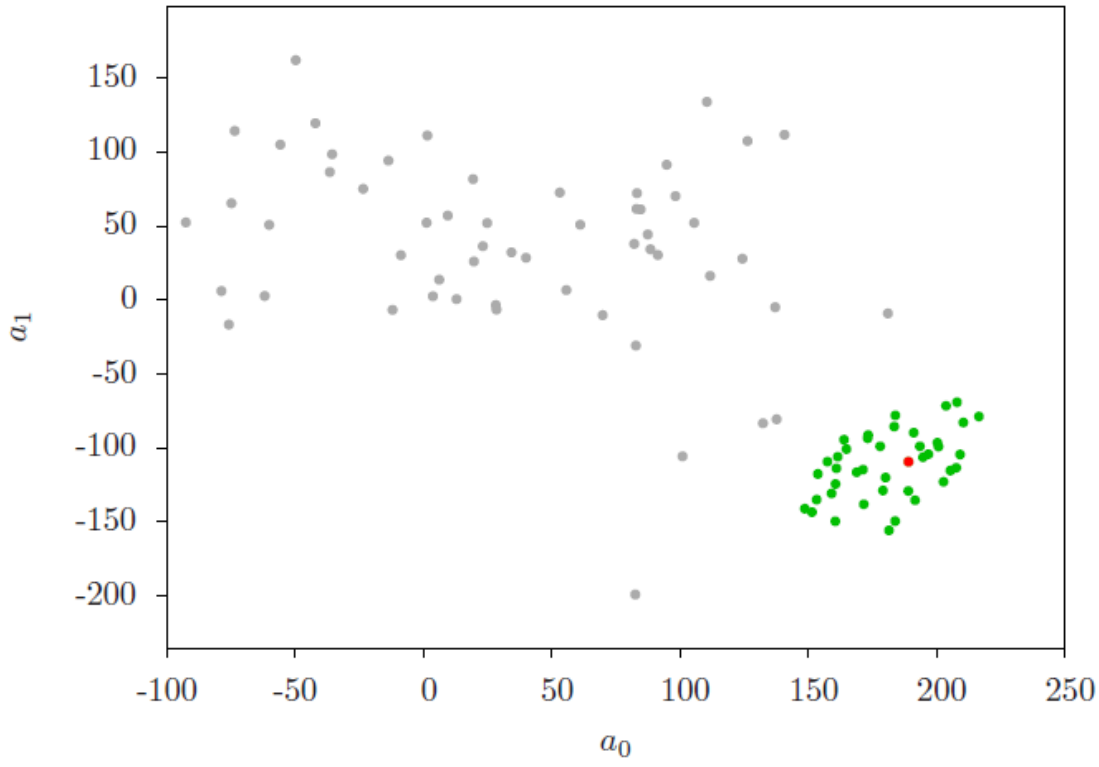
Unsupervised Motif Search (4)

Example for unsupervised motif detection: Scatterplot of trend aspects a_0 (average) and a_1 (slope)



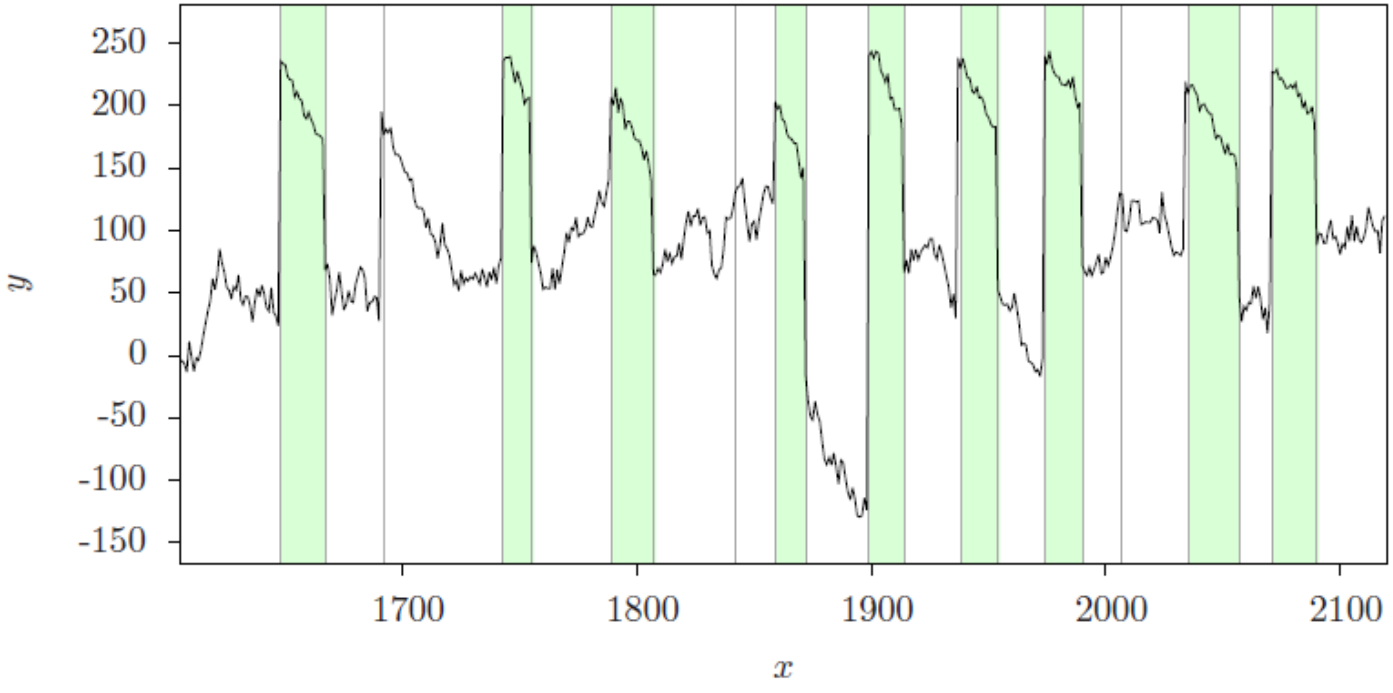
Unsupervised Motif Search (5)

Example for unsupervised motif detection: Clustering result



Unsupervised Motif Search (6)

Example for unsupervised motif detection: Transfer of clustering results to segments



Unsupervised Motif Search (7)

Unsupervised motif definition

- Advantage: Multiple similarly occurring segments will be recognised and used.
- Disadvantage: Result may not be directly usable, e.g. if the same motif appears in all classes during a classification task.

Motif search

- Once an interesting segment has been identified, it must be appropriately represented, e.g. by representing the clustering result.
  - Selection of a representative (partial) set, such as the medoid of each cluster
  - Additional information about size (number of data points) and extension (scatter)
- Modelling of relevant segments using models (e.g. HMM)
- Use a classifier that can distinguish relevant sections from irrelevant ones (e.g. one-class SVM).
- ...

Motif search (2)

Motif search (continued)

- Additionally necessary: Distance or similarity measure between the representation of the interesting segments and a single (unknown) segment.
- Basic procedure:
  - Unknown test time series is segmented in the same way as the training time series.
  - Each test segment is checked against the defined motifs and matches (including degree of match, if applicable) are determined.
  - Depending on the representation, threshold values for minimum similarity or maximum distance can be determined from training data.
  - Decisions can then be made on the basis of the presence and/or absence of motifs.

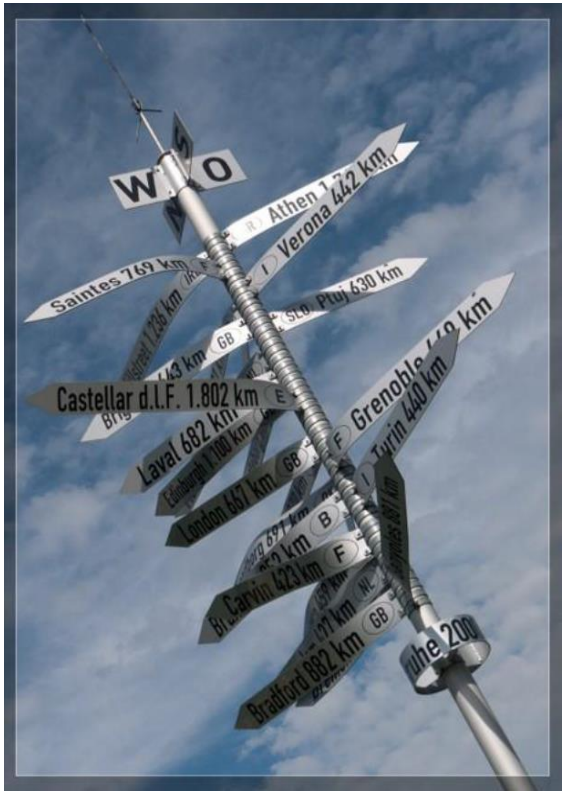
Motif search (3)

Motif as most similar pair

- In addition to the "classical" view, a simplified definition of a motif as the most similar pair of two time series exists in a database.
- If all possible partial sequences of a given length are extracted from all time series, the most similar pair among all subsequences can be called "Subsequence Motif".
- Complete search very inefficient: quadratic in the number of time series or subsequences
- Efficient and exact solution available, see e.g.:
  [Mueen, Keogh, Zhu, Cash, Westover, Exact Discovery of Time Series Motifs, 2009]

Agenda

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings



Anomaly detection: Motivation

Anomaly or novelty detection

- Can be understood as unknown patterns in time series or unexpected behaviour of some processes generating data (unexpected means that is does not fit to the model that the system has been derived from previous observations).
- Is closely related to the detection of repeating patterns (motif detection).
- Basic idea is to model everything that is known (or has been observed before) and use this knowledge to detect behaviour that is not covered by these training data.
- It depends largely on the particular application how such an anomaly looks like or how it is distinguished from the extreme cases of normal behaviour.
- Similar to the detection of outliers, a threshold for the detection of non-normal behaviour is needed next to the basic model of expected behaviour.

Anomaly detection: Motivation (2)

Basic process for time series data:

1. Model all known behaviour using training data
   - Ideally: Fully cover the input space with existing data
   - Within the initial model, we should consider that outlier, noise, and disturbances are available in the data
   - A model of the “normal conditions” could – but does not necessarily need to – cover explicitly the existing data
   - Only needed is a complete explanation of the existing data (e.g. thresholds are chosen in a way that normal variations or disturbances or deviations are not considered as anomalies)
2. Determine boundaries or thresholds from the training data to assess normal conditions and the maximally accepted deviation.
3. Apply online and search for anomalies
   - I.e. use in real system with novel data

Anomaly detection: Motivation(3)

Several different approaches known in literature:

1. One-Class Support Vector Machines, . . .
2. Probabilistic approaches, e.g. based on Gaussian Mixture Models (GMMs)
   - Use GMM to model your observations (at design-time under test and/or at runtime, e.g. at system start-up in a supervised mode)
   - Use divergence measures to compare distributions (and consider thresholds of acceptable deviations between these distributions)

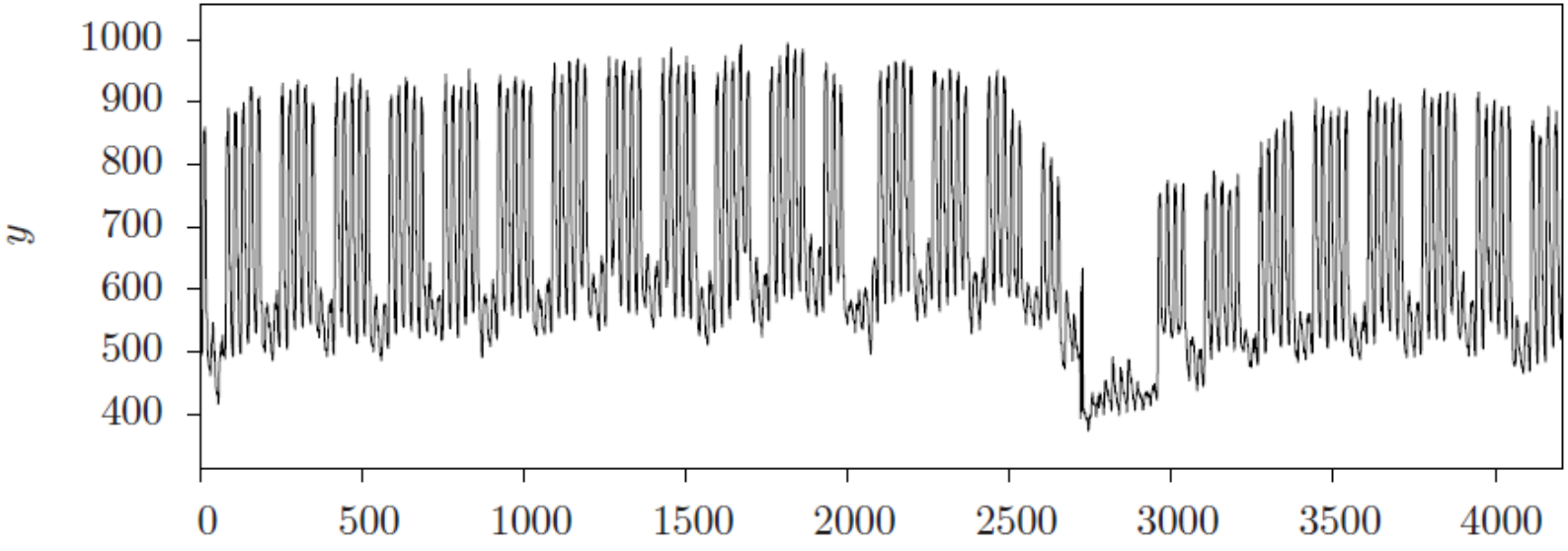
In most cases, we distinguish between two different goals:

- Search for anomalies in comparison to normal data (e.g. threshold based approaches / automatically defined by One-Class SVM)
- Search for patterns that are maximally dissimilar to the normal behaviour (e.g. top-10 dissimilar sub-sequences on time series)

Anomaly detection: Motivation(4)

Christian-Albrechts-Universität zu Kiel

Example 1: Search for anomalies in the energy demand of a building



Source: Prechelt, *PROBEN 1 - a set of benchmarks and benchmarking rules for neural network training algorithms*, 1994

Anomaly detection: Motivation (5)

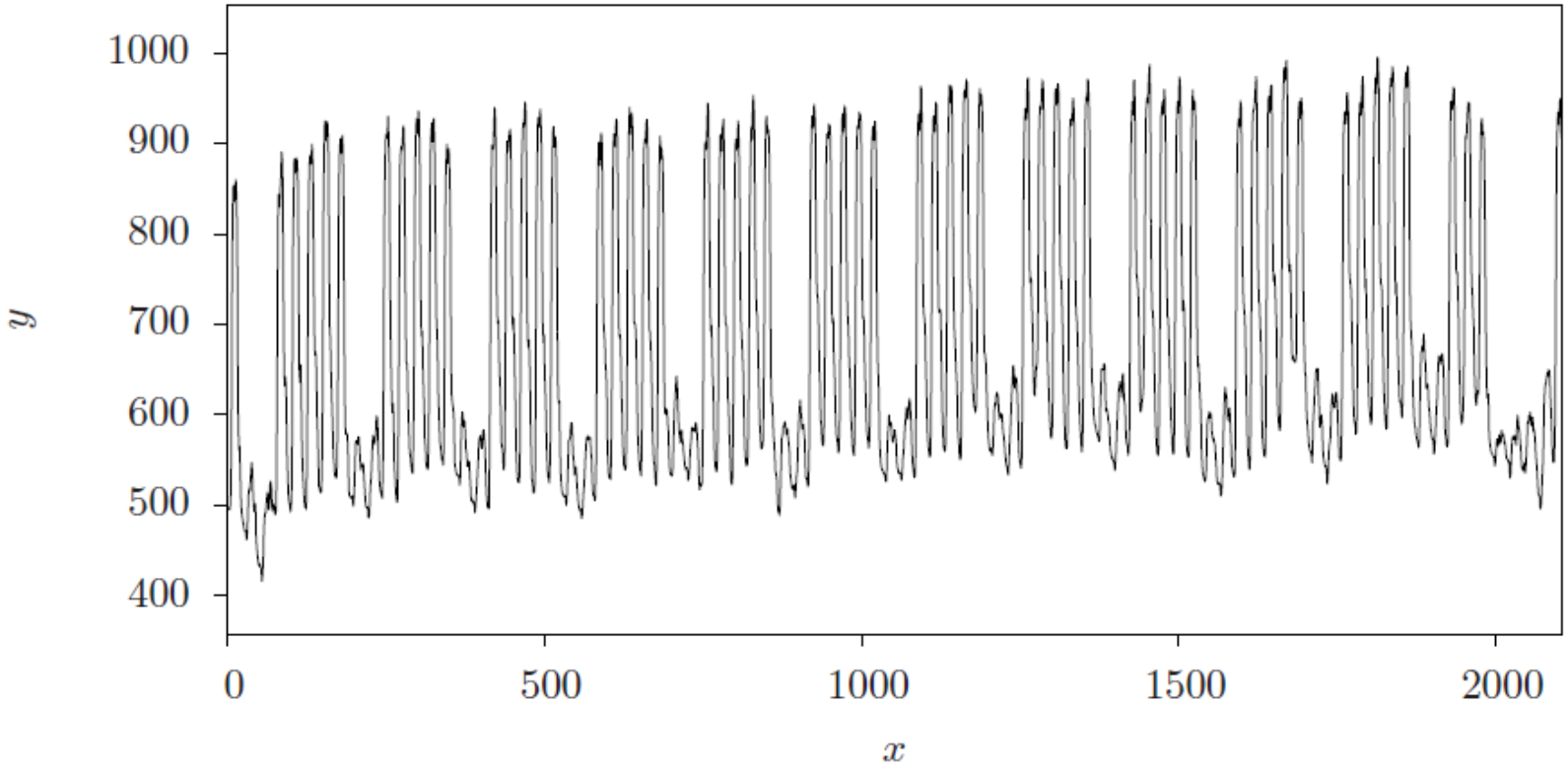
Christian-Albrechts-Universität zu Kiel

Approach

- Splitting the time series into sections for training and testing (here 50:50)
- Creation of models for weekdays and weekends (public holidays) from training data
- Search for abnormal days in test data
- Sort test days by similarity to known behaviour

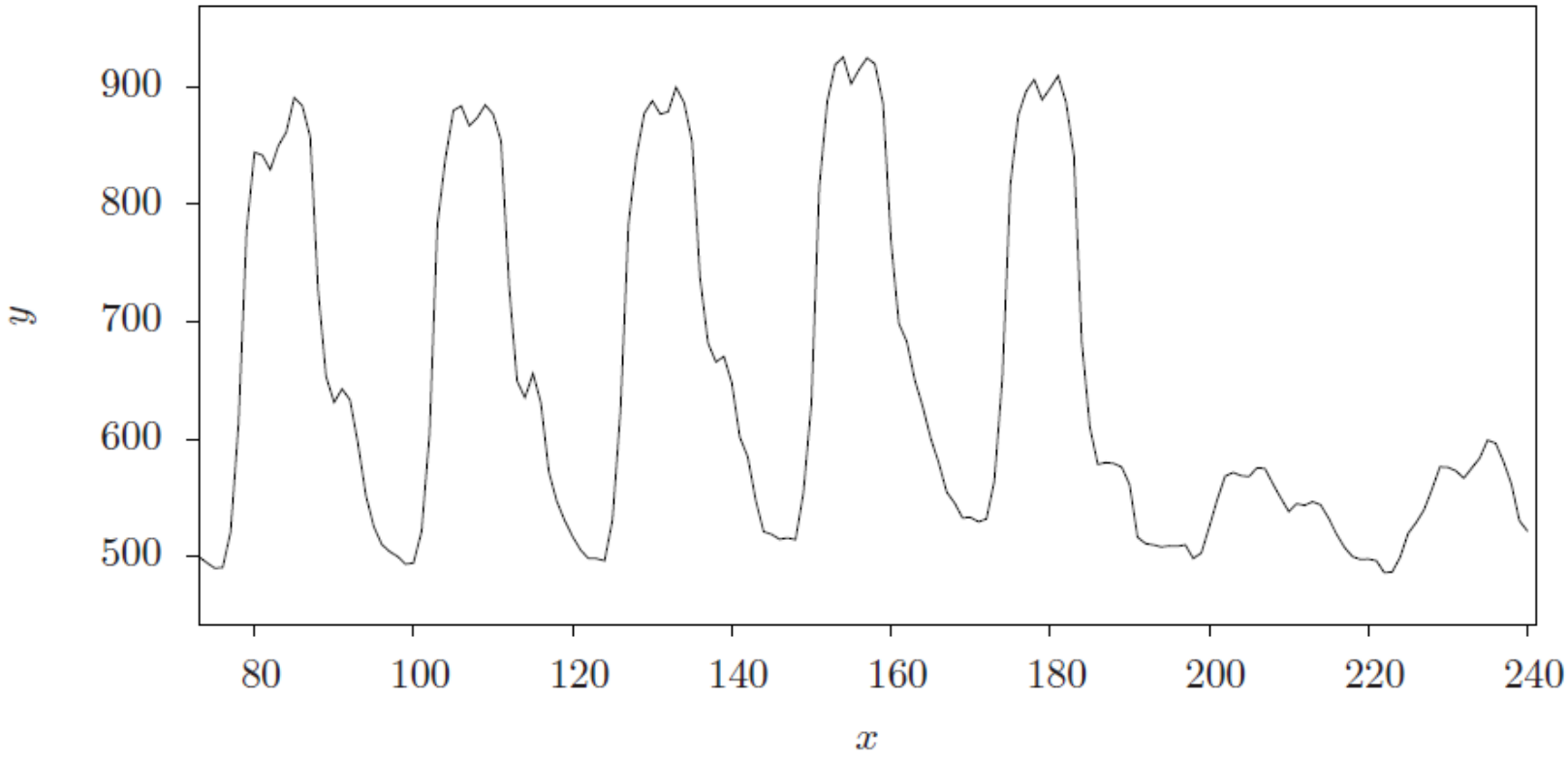
Anomaly detection: Motivation (6)

Example: Training data



Anomaly detection: Motivation (7)

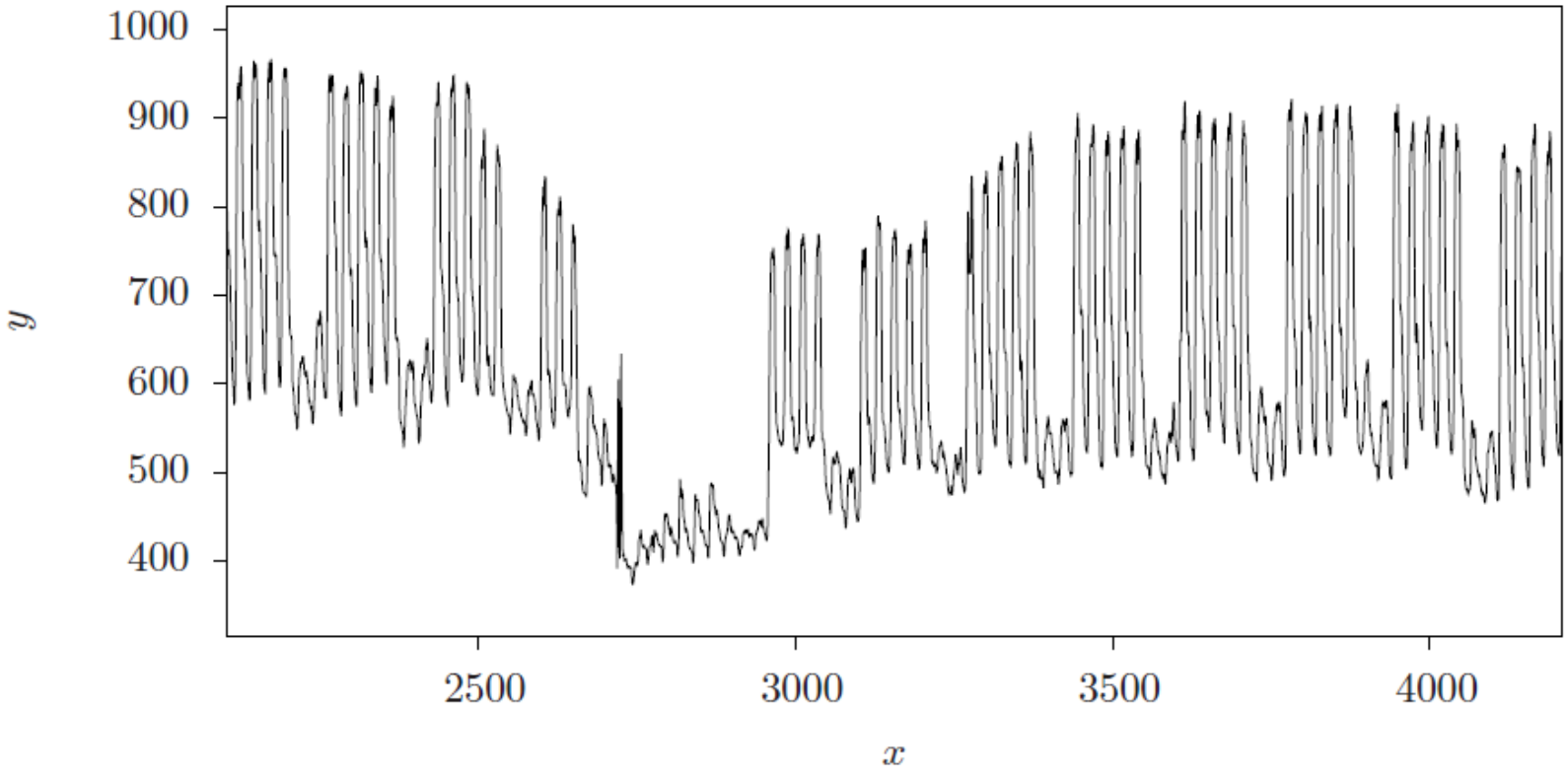
Example: Training data (zoom-in view of seven days)



Anomaly detection: Motivation (8)

Christian-Albrechts-Universität zu Kiel

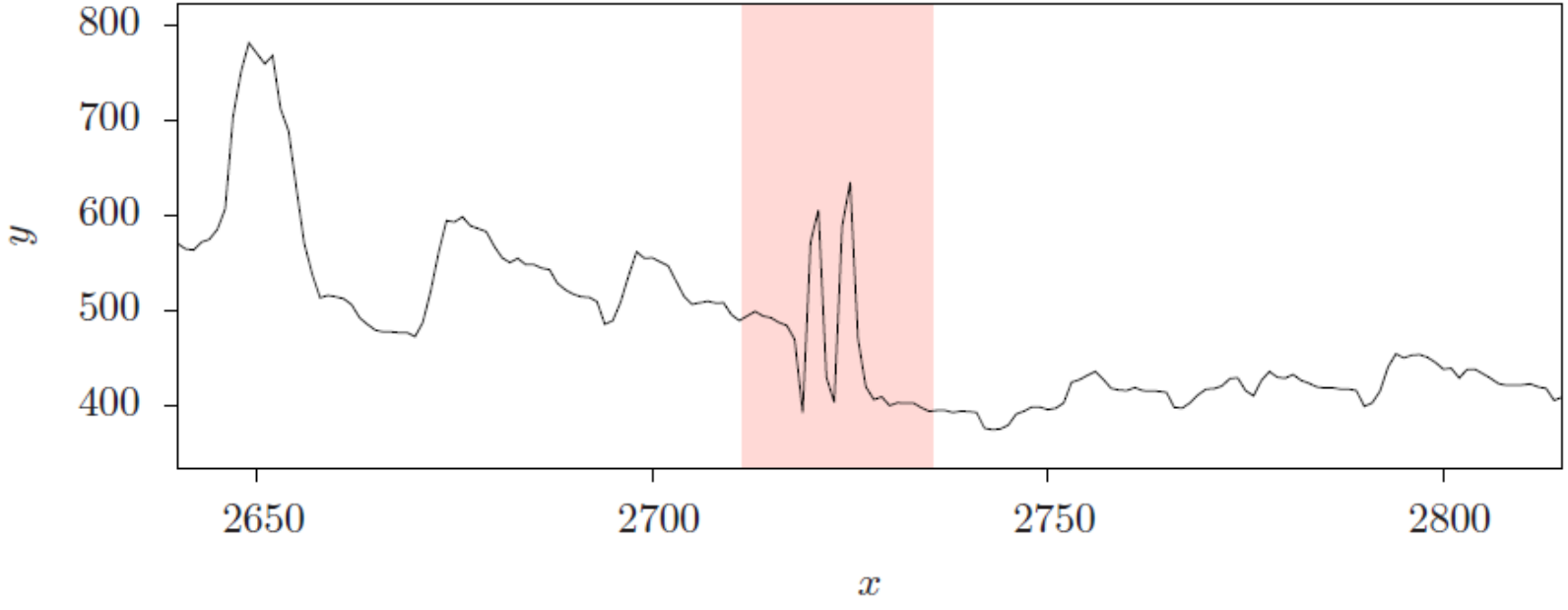
Example: Test data



Anomaly detection: Motivation (9)

Christian-Albrechts-Universität zu Kiel

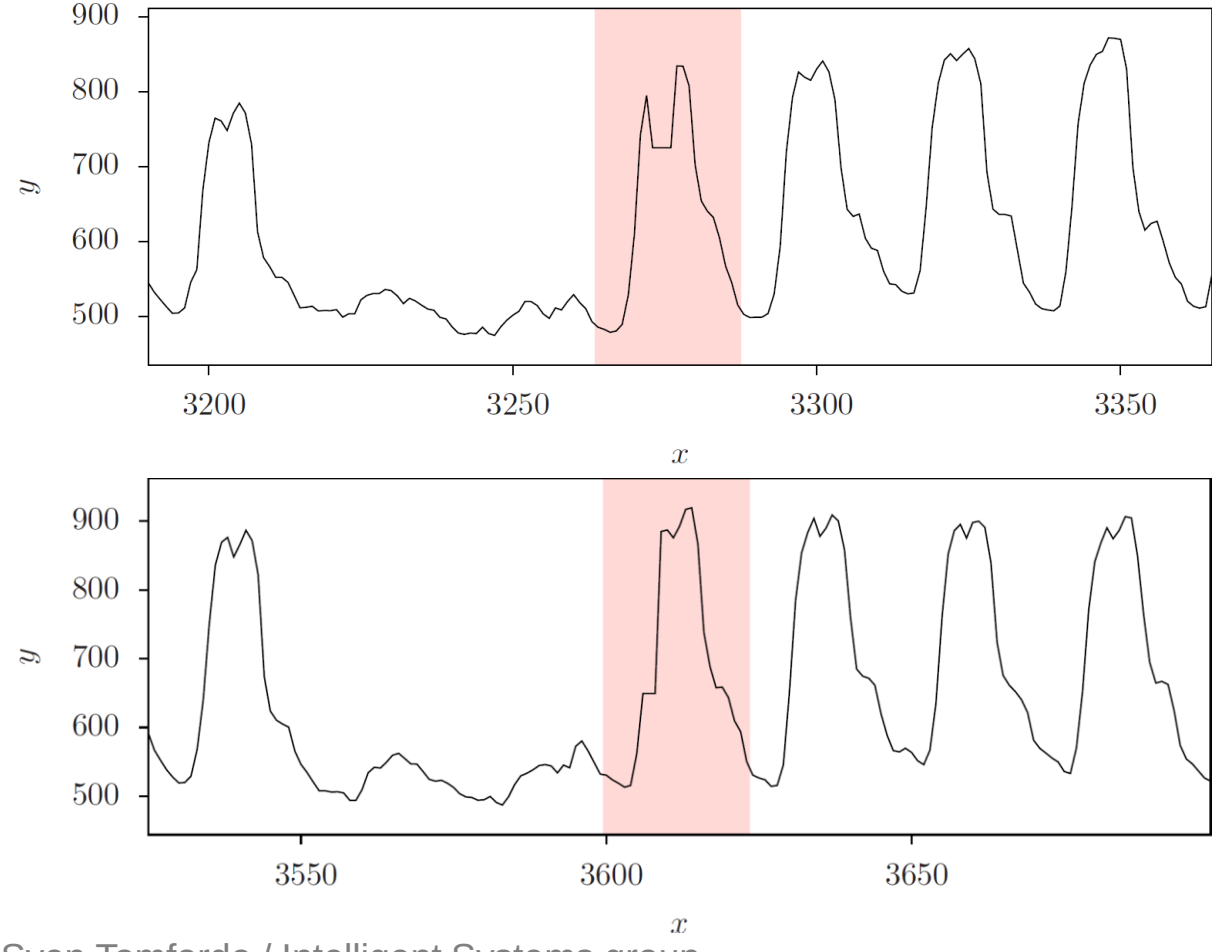
Example: Detected anomaly



Anomaly detection: Motivation (10)

Christian-Albrechts-Universität zu Kiel

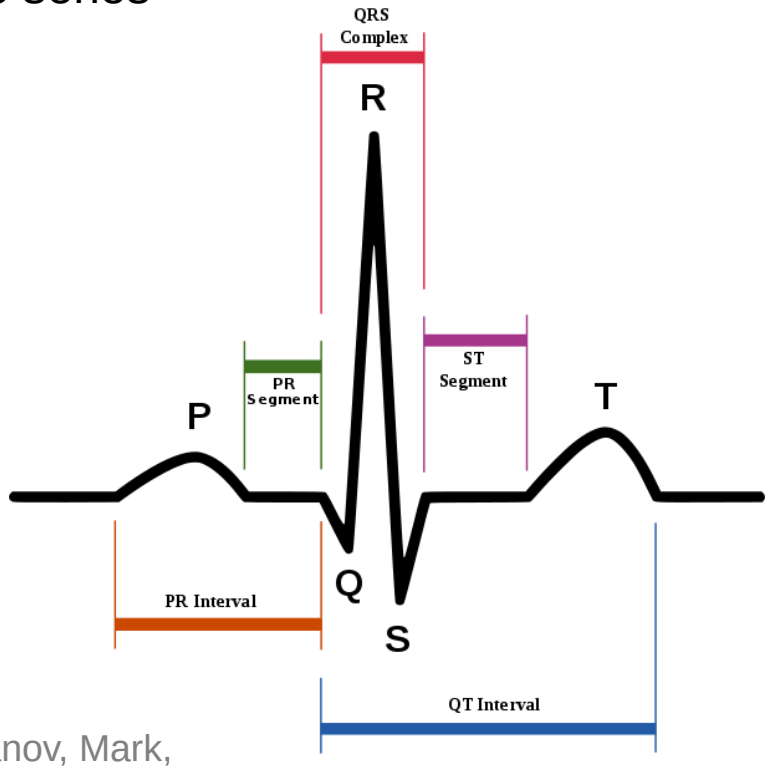
Example: Further detected anomalies



Anomaly detection: Motivation (11)

Example 2: Search for anomalies in ECG time series

- Challenges:
  - No fixed length
  - Several (harmless) disorders caused by the person's movements



Source: Goldberger, Amaral, Glass, Hausdorff, Jeffrey, Ivanov, Mark, Mietus, Moody, Peng, Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, 2000

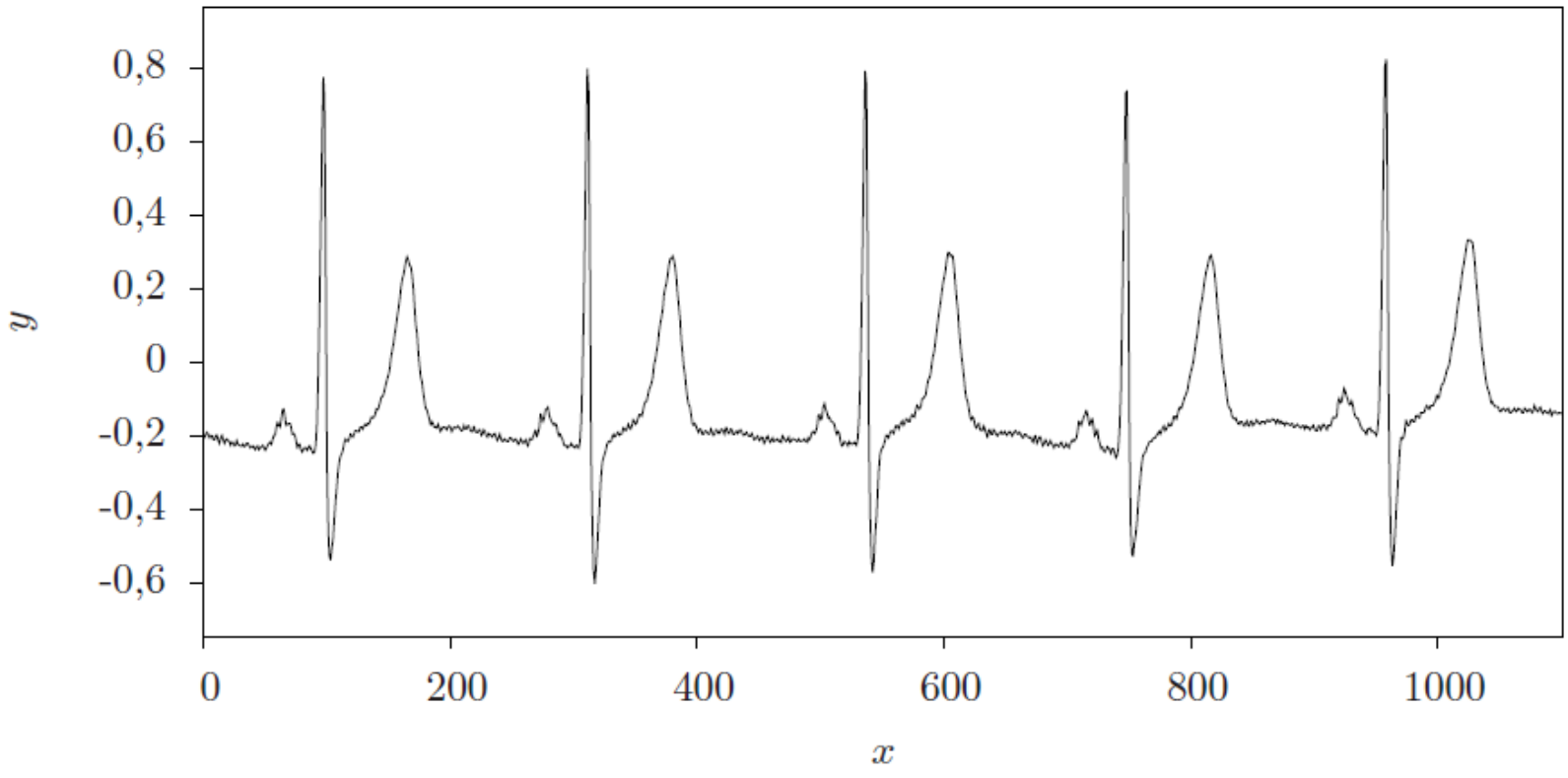
Anomaly detection: Motivation (12)

Example 2: Approach

- Consistent segmentation of each phase into multiple sections
  - Segmentation into one section each is trivial
  - Segmentation into several sections enables finer modelling or search for deviations
- Creation of models of each segment of the training time series
- Classification of each segment of the test time series
- Detection of deviations from expected sequence
- Problem: any misclassification leads to the detection of an anomaly
- Solution: strong deviation required, e.g. more than 8 errors within 10 consecutive segments

Anomaly detection: Motivation (13)

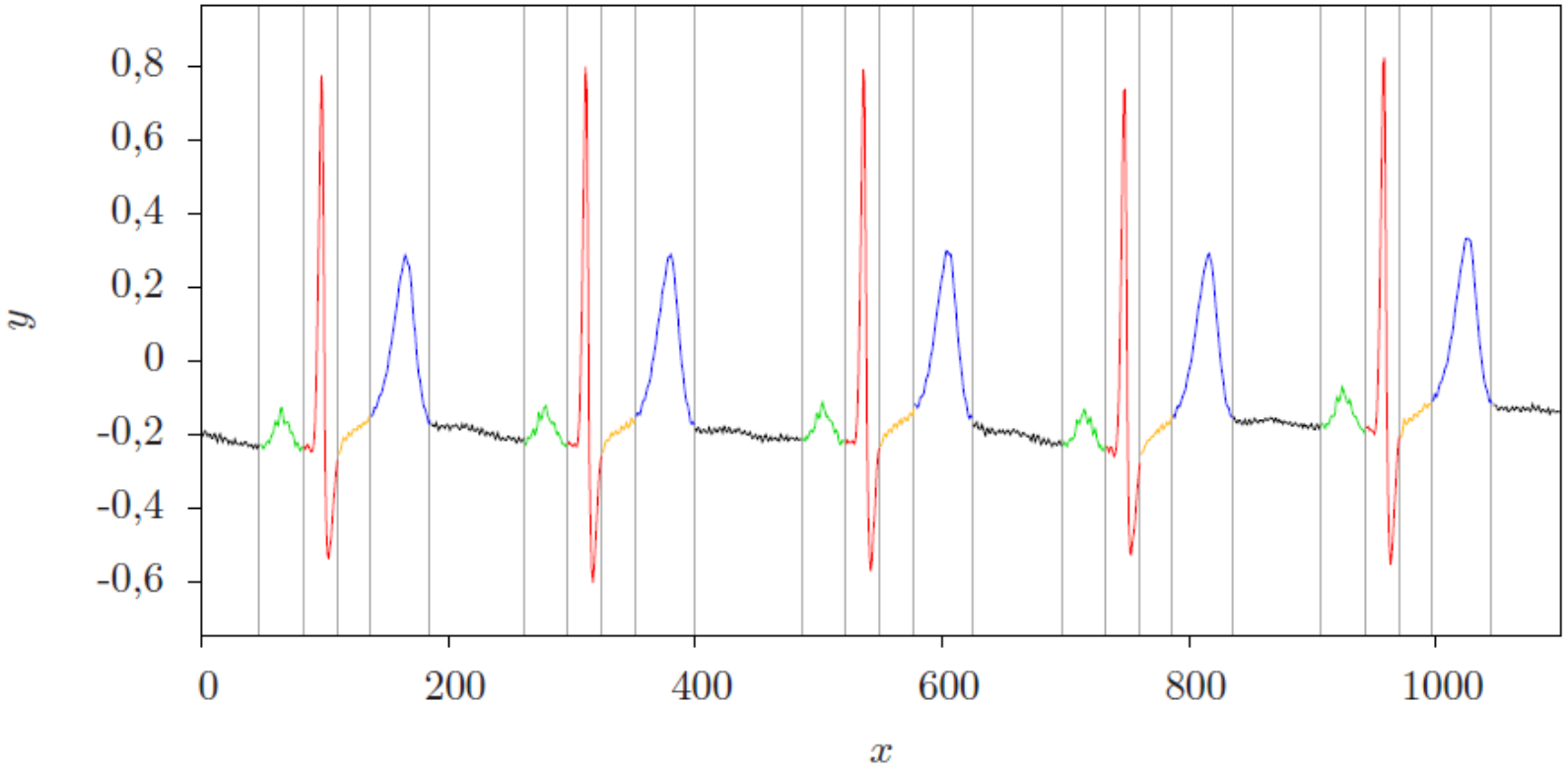
Example 2: Training data parts



Anomaly detection: Motivation (14)

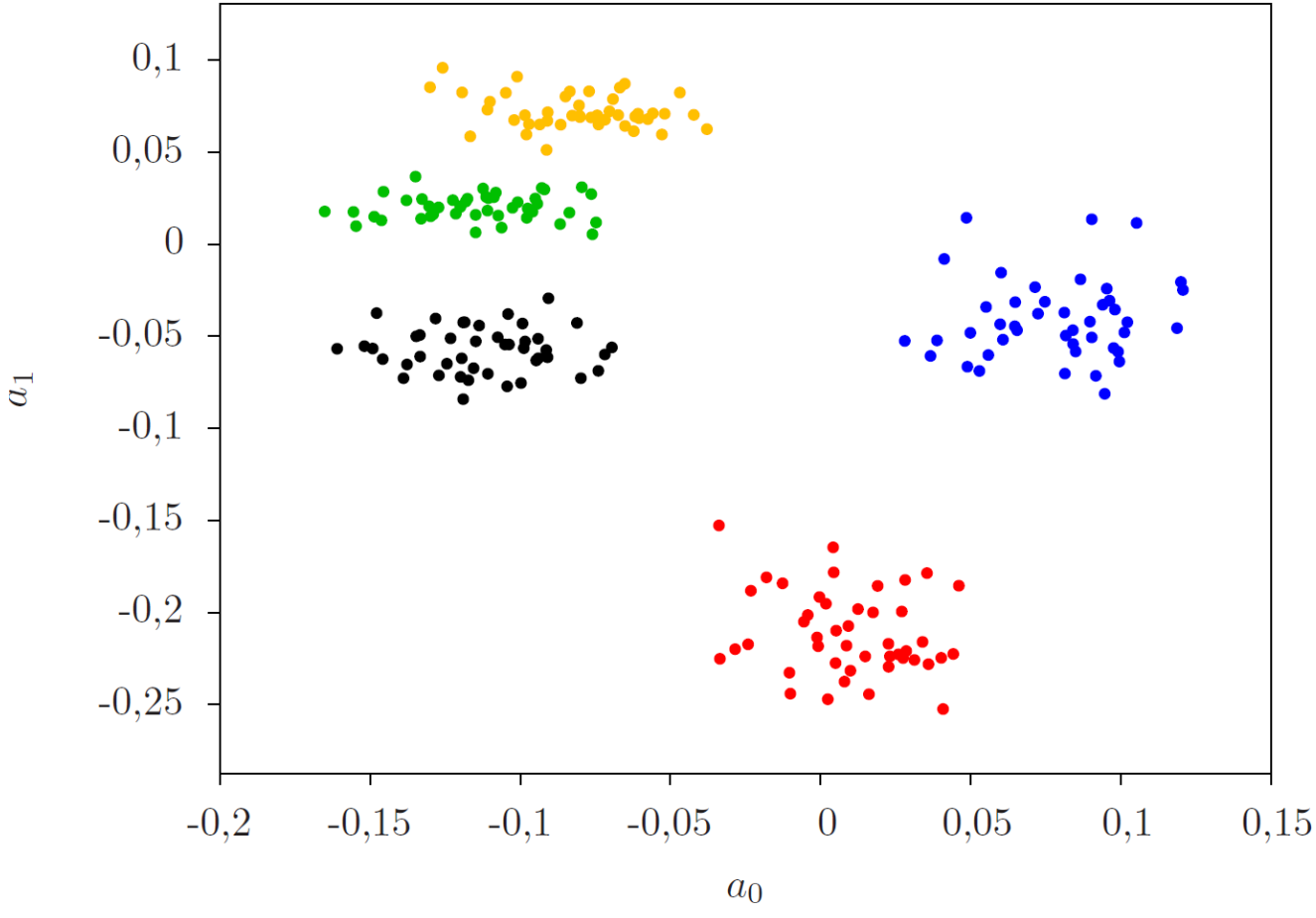
Christian-Albrechts-Universität zu Kiel

Example 2: Result after segmentation



Anomaly detection: Motivation (15)

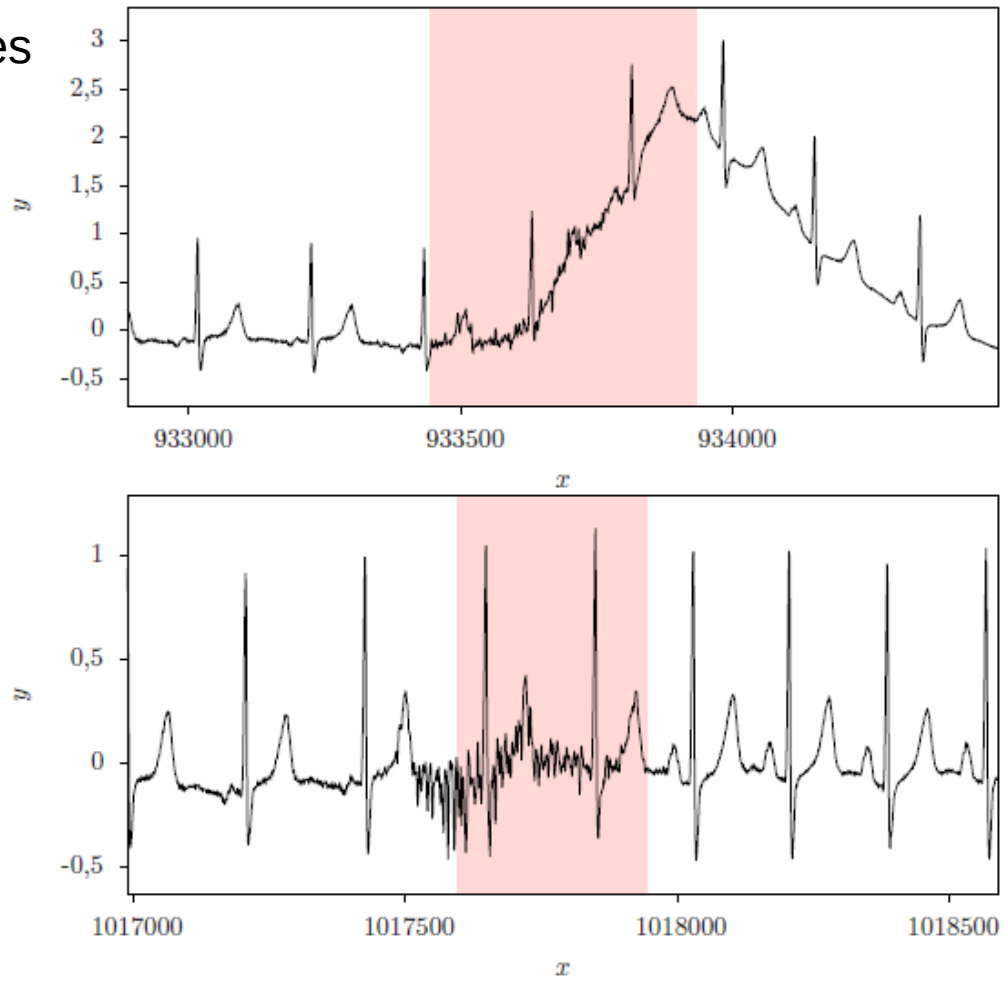
Example 2: Clustering of segments



Anomaly detection: Motivation (16)

Christian-Albrechts-Universität zu Kiel

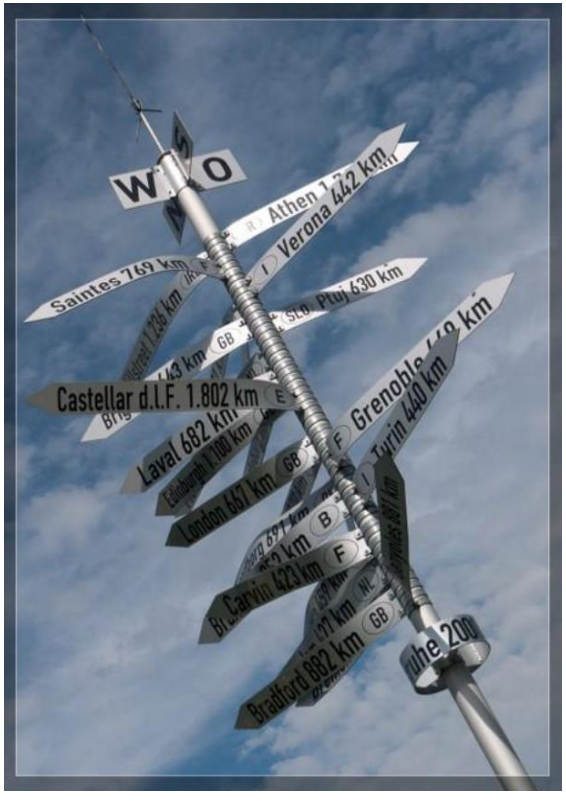
Example 2: Detected anomalies



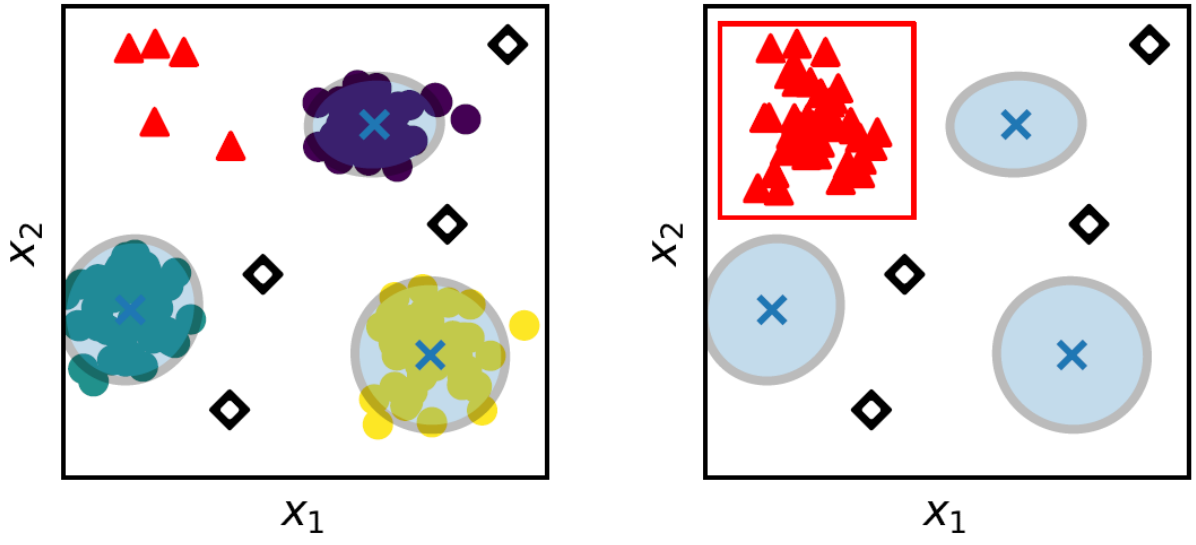
Agenda

Christian-Albrechts-Universität zu Kiel

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings



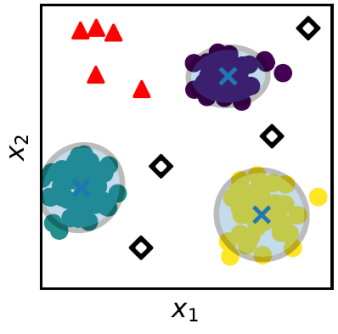
Anomaly and novelty detection

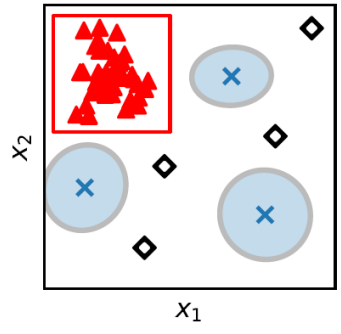


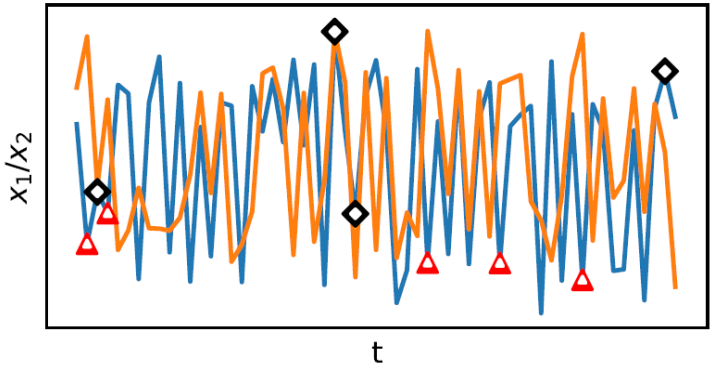
Anomaly and novelty detection (2)

Terminology

- Outlier
  - Usually removed to obtain clean training set.
  - Here, coloured samples beyond the ellipses.
- Anomaly
  - Samples not explained by the model.
  - Different from expected samples.
  - Here, depicted as red ▲.
- Noise
  - Random influences in input space (e.g., interference)
  - Here, depicted as black ◇.
- Novelty (our definition for this course)
  - Agglomeration of anomalies▲.
  - But also change of distribution.
  - Here, samples within red square.







Anomaly and novelty detection (3)

Terminology

- The term “novelty” is often used as synonym for anomaly
- Anomaly and outlier detection techniques often similar, but different reactions:
  - Outlier: Should be removed to cleanse data set.
  - Anomaly: Should trigger an action, e.g., detection of a fault.
- Anomalies, outliers, and noise are essentially indistinguishable.
- *Novelty* detection build *on-top* of detected anomalies.
  - Consider relations, e.g., spatial distance, or density between anomalies.
  - Short-term memory, e.g., *sliding-window* or *ring-buffer*.

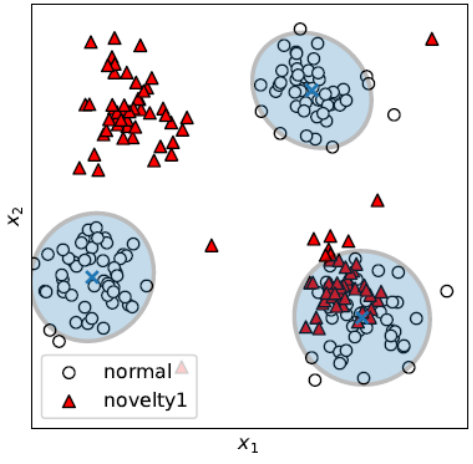
Anomaly and novelty detection (4)

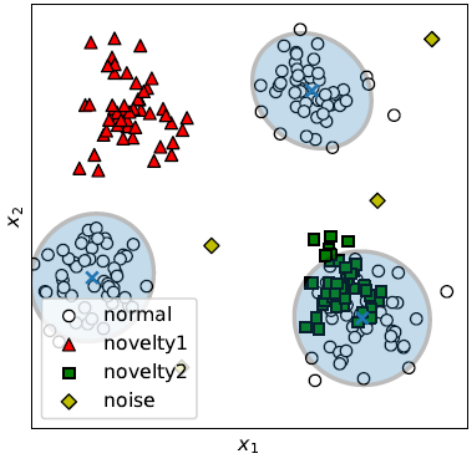
- *Offline*:
  - All *test samples* are available at once.
  - If samples are ordered, predecessors and successors are known and available.
  - For classification of a single sample all other samples are available.
- *Online*:
  - *Test samples* are processed one after another, or in batches.
  - Samples are ordered.
  - Successors are unknown.
  - Predecessors are known, but must be remembered (e.g., sliding-window)
  - Real-time applications possible.

Anomaly and novelty detection (5)

Novelty Detection as classification task

- Binary classification:
  - Two classes: *normal* and *anomaly*
  - Suitable for simple anomaly detection
- Multiclass classification:
  - Multiple classes: *normal* $c_1, \dots, c_n$ *novelty* $n_1, \dots, n_m$, and *noise* n_0
  - Better suited for novelty detection (novelties are distinguishable)
  - Intelligent Systems: more information, *what* changed in contrast to *something* changed.





Anomaly and novelty detection (6)

Requirements for novelty detection in intelligent systems:

- *Online*: Observations are processed when they appear.
  - *Data streams*, potentially infinitely many samples.
- *Multiclass*: single normal and multiple novelty classes (including *noise*).
- The i.i.d. assumption is usually too strong, i.e., temporal system has inherent dependencies.

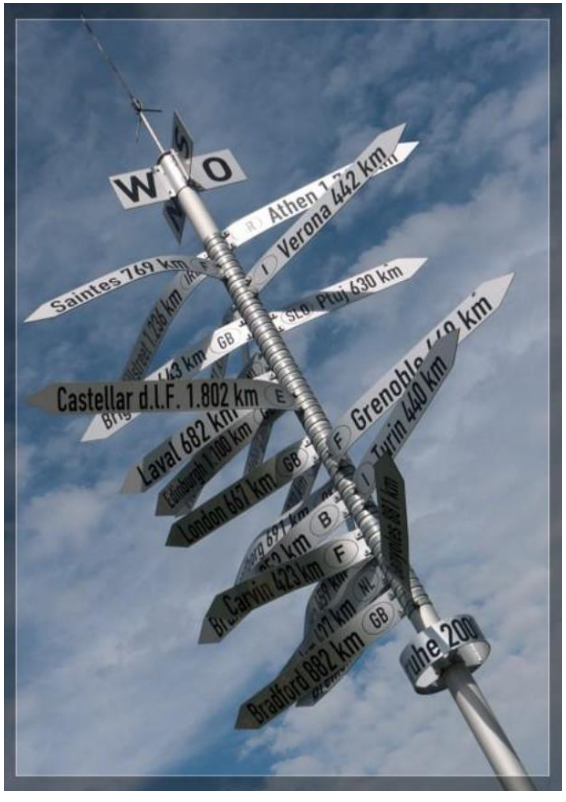
Anomaly and novelty detection (2)

Assumptions

- To detect *novelties* in data streams, certain assumptions must be made a-priori to select suitable models.
- Is the data *independent and identically distributed* (iid)?
- Which kinds of dynamics are expected?
  - Various operational states → latent variables, HMM.
  - More complex dependencies in underlying *processes*.
- Is *concept drift* an issue?
  - Example: Mean changes over time, but variance does not.
- **Note:** In most real world applications, the processed data do not fully comply to the made assumptions. Wrong assumption can lead to high error rates (type I and II).

Agenda

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings



Novelty detection algorithms

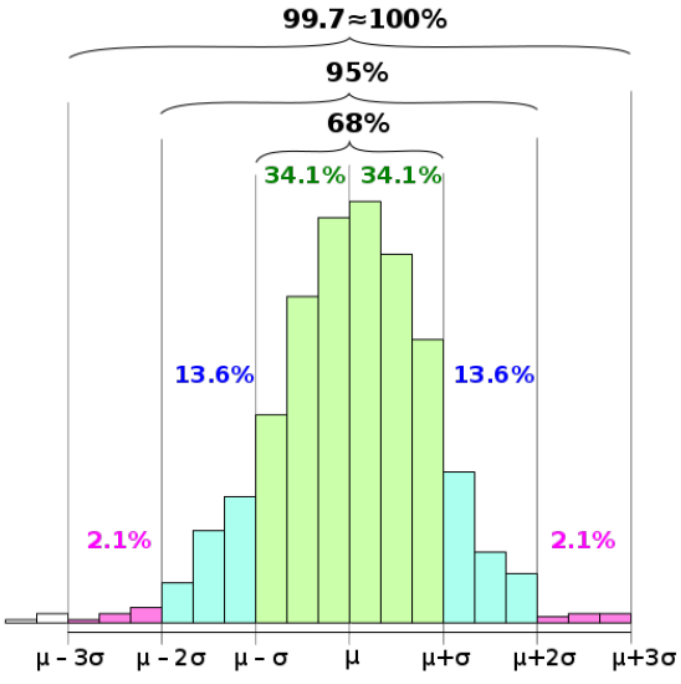
In the following, we discuss approaches and algorithms using a certain line of argumentation:

- Distance-based approach
- Density-based approaches
  - DBSCAN
  - OPTICS
  - Local Outlier Factor
- Probability-based density approaches
- One-Class SVM
- High- and low-density regions
- Distribution testing

Novelty detection algorithms (2)

Distance-based Approach

- Naïve approach: Euclidean distance
  - Nearest neighbour and threshold
- Derive threshold from probability distribution:
  - Three-sigma rule (also: 68–95–99.7 rule) for Gaussians
  - A fraction of 99.7% has a distance $\leq 3\sigma$ to μ
  - Expected false-positive rate $\sim 0.3\%$
- Density-based
  - Considers neighbourhood of samples for point estimate
  - In this case, not to be confused with probability density



Novelty detection algorithms (3)

Example: DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Density-based spatial cluster analysis with noise
- Basic idea is close to nearest neighbour clustering: Neighboured objects are summarised to one cluster
- However: In addition, the number of neighboured objects is taken into consideration and single (isolated) objects at the edge are explicitly handled as noise (or recognised as outlier)
- Related to Parzen-Window/Kernel-Density-Estimation

Novelty detection algorithms (4)

DBSCAN

- DBSCAN contains two parameters

1. ε defines the length of the neighbourhood. This refers to the maximal distance in which two objects are called ε –reachable. In other words: Radius of sphere to contain $minPts$.
2. $minPts$ minimum size of ε -neighbourhood for core-points (how many objects must be ε -reachable for an object becoming core object?).

Novelty detection algorithms (5)

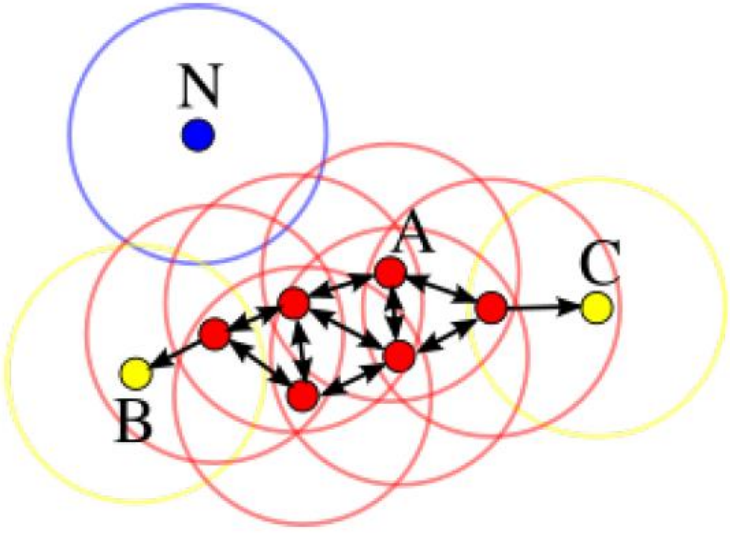
DBSCAN: different classes of objects

- Core objects possess at least $minPts$ objects that are ε-reachable
- Directly density-reachable objects are objects that are ε-reachable from core objects
- Density-reachable objects are connected through a chain of directly density-reachable objects
- Two objects are density-connected if they are density-reachable from a third object
- Noise objects are objects that are not density-reachable

Novelty detection algorithms (6)

DBSCAN: Example

- A and points close to A are core objects
- Points B and C are density-reachable from A and correspondingly density-connected (and belong to the same cluster)
- Point N is neither a core object, nor density-reachable – means: noise
- ($minPts = 3$ or $minPts = 4$)



Novelty detection algorithms (7)

DBSCAN: Remarks

- The algorithm determines cluster *C* (that are subsets of the total set of data objects *X*) for that the following statements hold:
  - Maximality: $\forall p, q \in X$: *C* and *q* density-reachable from $p \Rightarrow q \in C$
  - Connectedness: $\forall p, q \in C$: *p* is density-connected with *q*
- Advantages:
  - Number of cluster is not required in advance
  - Shape of cluster is arbitrary
  - Utilisation of arbitrary similarity measurements
  - No need for calculating cluster centres that may not correspond to valid samples

Novelty detection algorithms (8)

DBSCAN: Algorithm

```
DBSCAN(D, eps, MinPts)
  C = 0
  for each unvisited point P in dataset D
     mark P as visited
     N = D.regionQuery(P, eps)
     if sizeof(N) < MinPts
        mark P as NOISE
     else
        C = next cluster
        expandCluster(P, N, C, eps, MinPts)
```

Novelty detection algorithms (9)

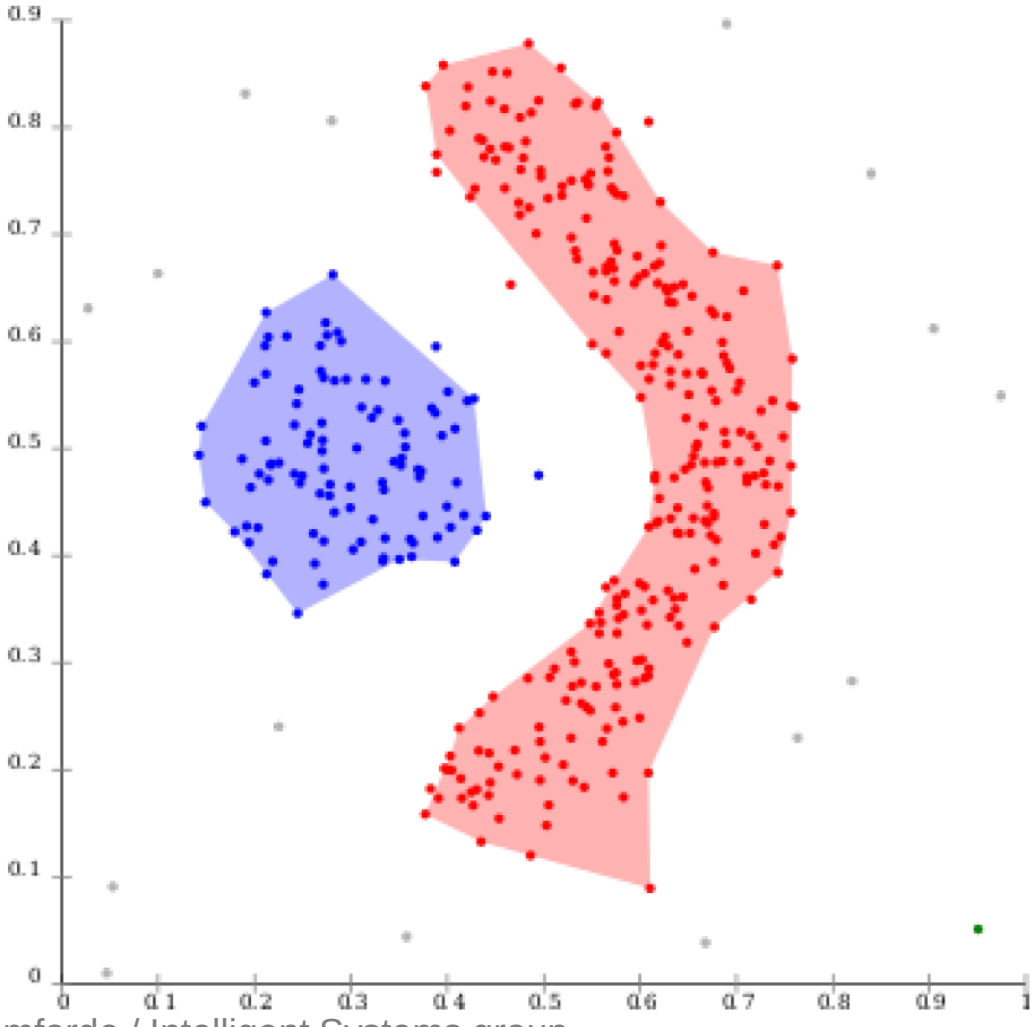
DBSCAN: Algorithm (ctd.)

```
expandCluster(P, N, C, eps, MinPts)
  add P to cluster C
  for each point P' in N
     if P' is not visited
        mark P' as visited
        N' = D.regionQuery(P', eps)
        if sizeof(N') >= MinPts
           N = N joined with N'
     if P' is not yet member of any cluster
        add P' to cluster C
```

Novelty detection algorithms (10)

Christian-Albrechts-Universität zu Kiel

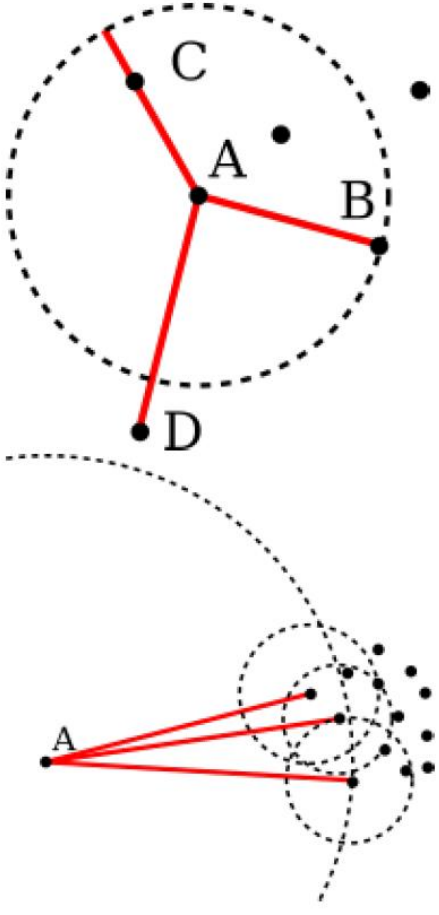
DBSCAN: Example



Novelty detection algorithms (11)

Local Outlier Factor (LOF)

- Based on ideas of DBSCAN (core-points, reachability, etc.)
- Locality of samples
  - Local density is estimated ...
  - ... and compared to density of neighbours.
  - Adaptive score to identify outliers
- Typical offline algorithm
  - k-nearest neighbours (k-nn) not known at runtime.
  - However: Buffered extension possible for online purposes??



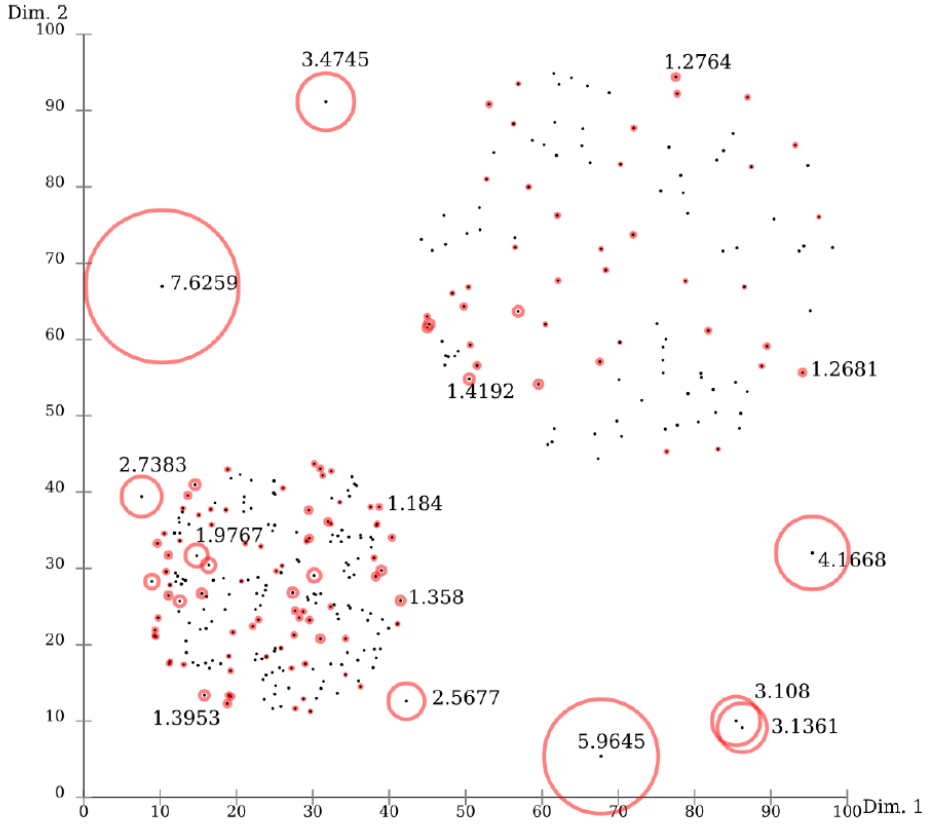
Novelty detection algorithms (11)

LOF – Equations

- $k =$ number of neighbours considered
- $kdist(x) =$ distance to k-th neighbour
- $N_k(x) =$ ordered set of k-nearest neighbours of x
- $reachability\_dist_k(x,y) = \max\left(kdist(y), dist(x,y)\right)$
- $lrd_k(x) = \frac{k}{\sum_{y \in N_k(x)} reachability\_dist_k(x,y)}$ = local reachability density
- $LOF_k(x) = \frac{\sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{k}$

Novelty detection algorithms (12)

- LOF score is evaluated to identify *anomalies/outliers*:
  - $LOF_k \sim 1 \Rightarrow$ similar density as neighbours
  - $LOF_k \ll 1 \Rightarrow$ higher density as neighbours (*inlier*)
  - $LOF_k \gg 1 \Rightarrow$ lower density as neighbours (*outlier*)
- Parameters:
  - Number of neighbours *k*
  - Threshold factor
- Computational cost:
  - reachability-dist*k* is not symmetric
  - *k*-nn based

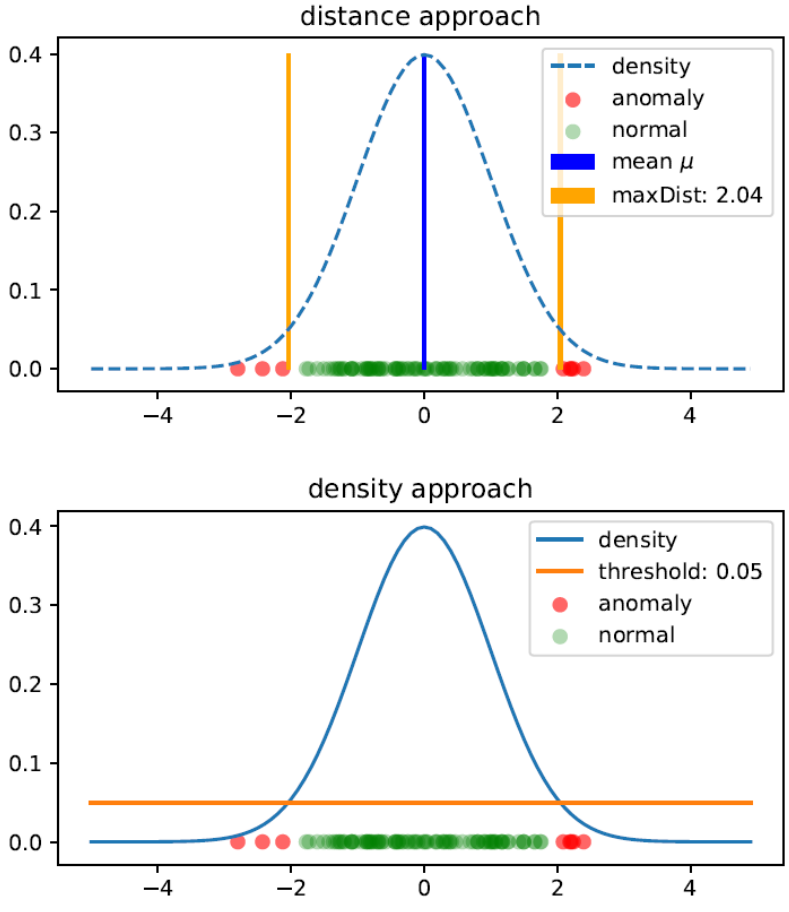


Novelty detection algorithms (13)

Christian-Albrechts-Universität zu Kiel

Probability density-based approach

- More general approach
- Requires a density function to evaluate samples
- Either parametric or non-parametric. For instance:
  - *GMM*
  - *Parzen-Window*
- Density of sample tested against threshold
- Density below threshold → *anomaly*



Novelty detection algorithms (14)

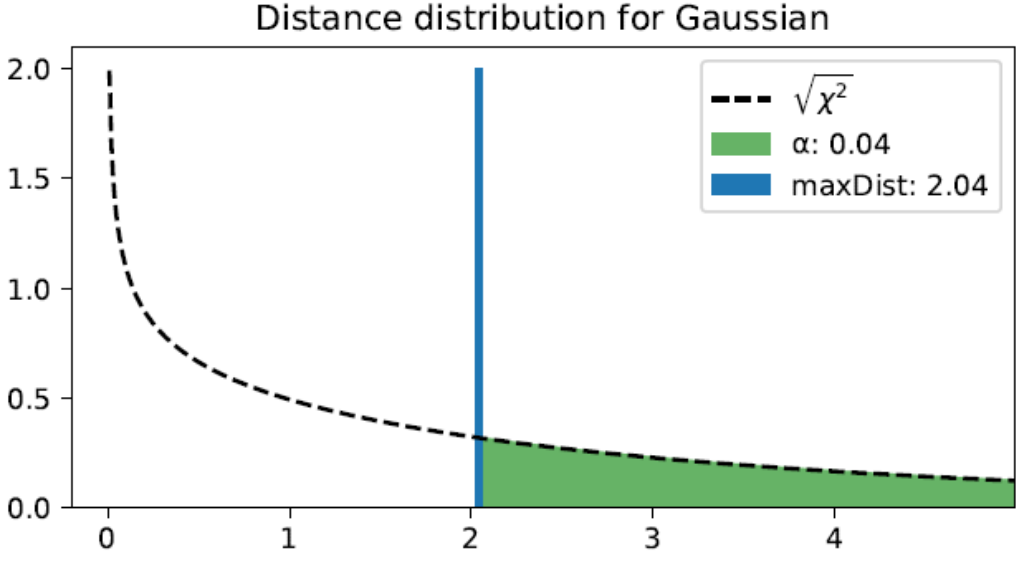
Probability density-based approach

- For Gaussians:
  Density directly dependent on distance between x and mean μ
- Threshold selection?
  - Borrow ideas from statistical testing
  - α false-positive rate (also: type I error rate)
- The squared distances of a of a Gaussian are χ_D^2 distributed
- Evaluate (Mahalanobis) distance and compare to precomputed threshold ρ
  - Threshold independent from shape
  - Less computationally expensive

Novelty detection algorithms (15)

Probability density-based approach

- Gaussians: $\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$
- Threshold: $\rho^2 = F^{-1}_{\chi^2_D}(1-\alpha)$
- $\rho \leq |x-\mu|$

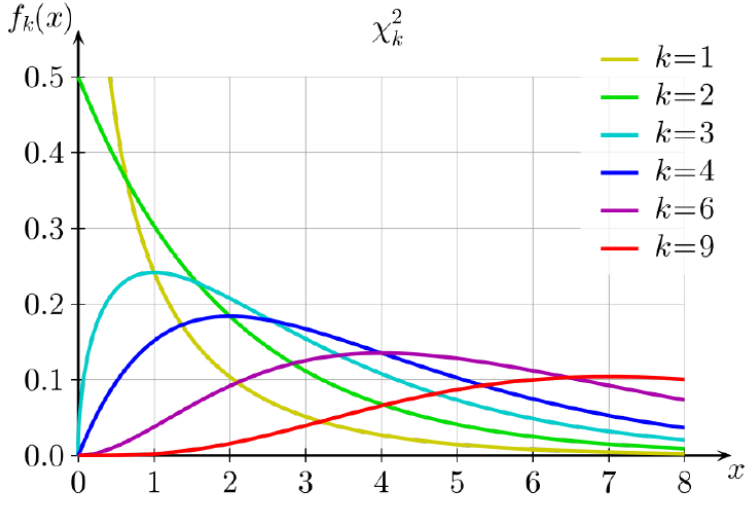


Novelty detection algorithms (16)

Chi-Square Distribution $\mathcal{X}_D^2$

- Is a continuous probability distribution based on the set of non-negative real values.
- It comes with a single parameter: the number of degrees of freedom n
- Can be derived from normal distributions:
  - Basis: n random variables Z_i (that are iid)
  - Chi-Square distribution with n degrees of freedom is defined as distribution over the sum of the squared random variables: $Z_1^2 + \cdots + Z_n^2$

Densities of Chi-Square distribution with varying number of degrees of freedom:



Novelty detection algorithms (17)

Utilisation of Chi-Square

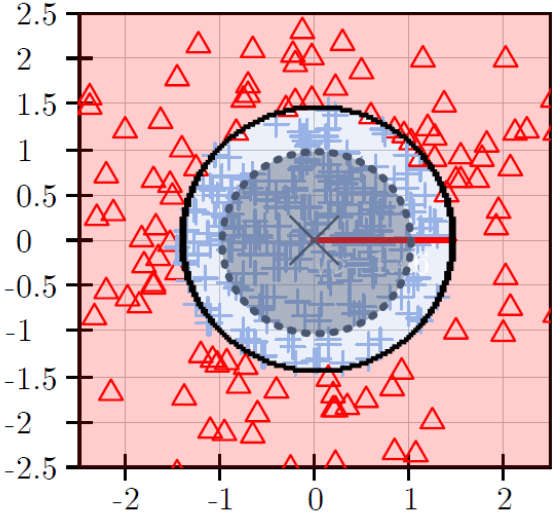
- Sums of squared random variables appear in estimation functions
- Chi-Square allows for assessing the compatibility with an expected functional relation (depending on temperature, time, etc) with empirically determined measurements
- E.g.: Do we need a linear, a logarithmic or an exponential model?
- The model with the lowest Chi-Square is the one with the best accuracy
- Major advantage in this context: Allows for estimation of “trusted interval”, i.e. encapsulating the unknown value with a given certainty

Novelty detection algorithms (18)

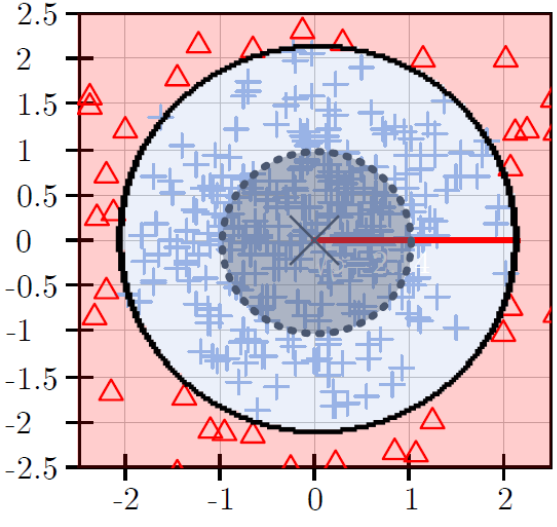
Christian-Albrechts-Universität zu Kiel

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\rho^2 = F^{-1}_{\chi^2_2}(1-\alpha)$$



$\alpha = 0.33 \rightarrow \rho = 1.47$

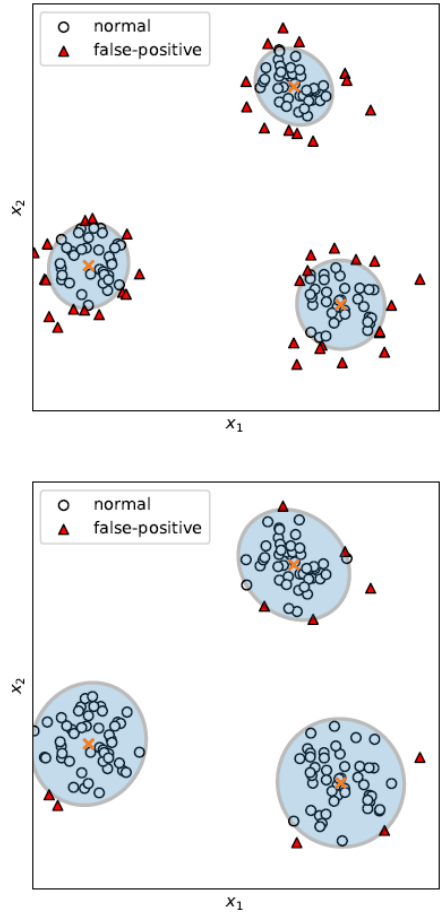


$\alpha = 0.10 \rightarrow \rho = 2.15$

Novelty detection algorithms (19)

Probability density-based approach

- Alphadetector (α-detector) for GMM
- Select α value
- Pre-compute threshold distance ρ
- Test new sample x' against all component means μ_j
- Anomaly?: $\bigwedge_{j=1}^{J} \Delta(x', \mu) \geq \rho$
- Due to potentially overlapping components type I error rate is at most $= \alpha$

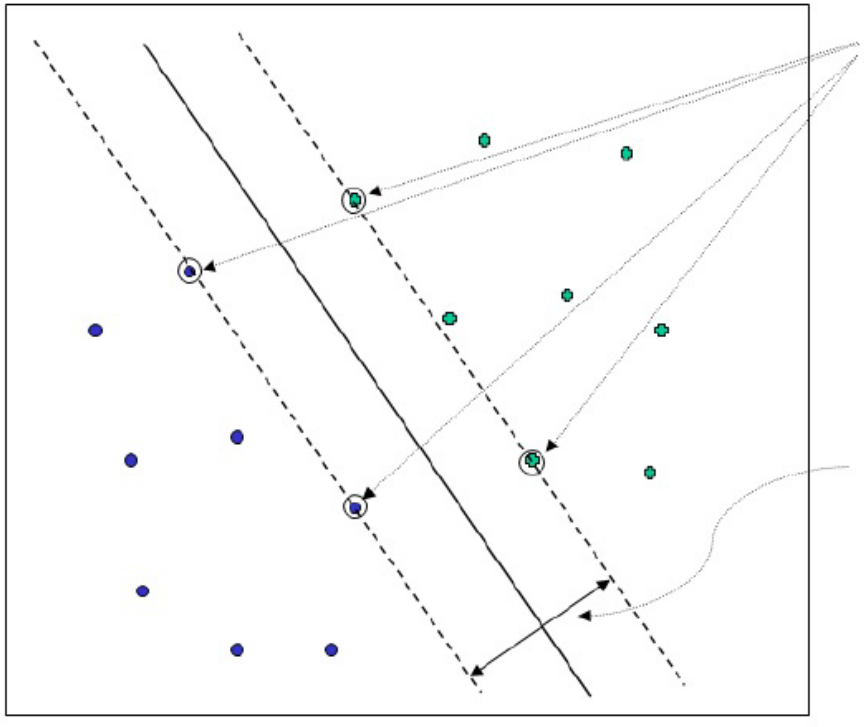


Novelty detection algorithms (20)

Christian-Albrechts-Universität zu Kiel

One-Class SVM

- Reminder: SVM



Support Vectors:
Those samples (data points) with the shortest distance to the separating hyper plane.

Margin:
Twice the distance between support vectors and the separating hyper plane.

Novelty detection algorithms (21)

One-Class C-SVM

•Only one class is given (sample of a class in the learning task).

•The goal is to determine a decision-making function that

1.provides positive values for those regions of the input space in which the samples are predominantly located, and

2.negative values for all other areas of the input space.

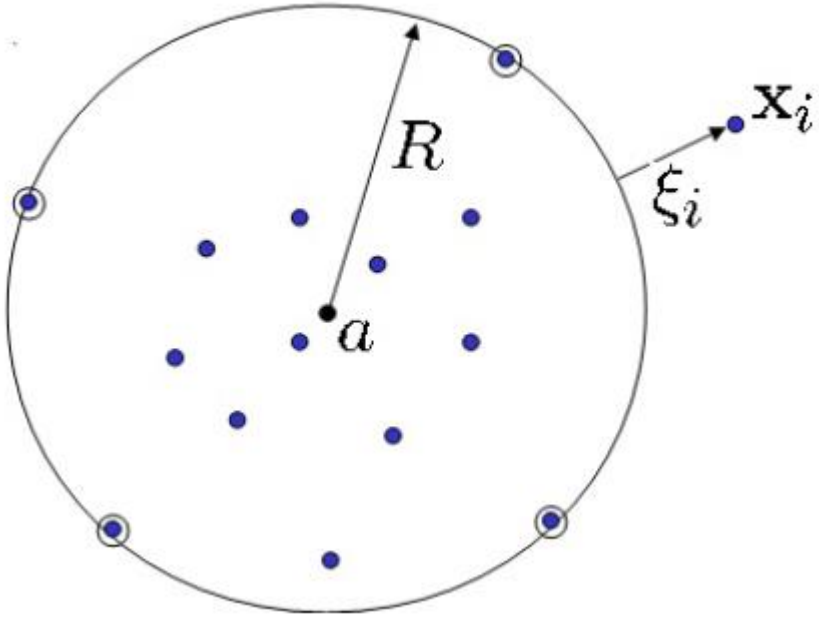
•Application: Novelty Detection (detection of new or abnormal instances of the class)

•Prominent areas of application (examples): condition monitoring or medical diagnosis

Novelty detection algorithms (22)

One Class C-SVM: Intuitive approach

- Find a hypersphere with minimal radius R and center α, so that most patterns lie within the hypersphere.
- Reduce the influence of outliers by slip variables



Novelty detection algorithms (23)

Objective function of the One-Class SVM

- The aim is to simultaneously minimise the hypersphere and the number of samples outside the hypersphere.
- Minimise

$$\mathcal{E} = R^2 + \frac{1}{N\nu}\sum_{n=1}^{N}\xi_n$$

- Under the constraints

$$||\mathbf{x}_n - \mathbf{a}||^2 \leq R^2 + \xi_n \quad \forall_n$$

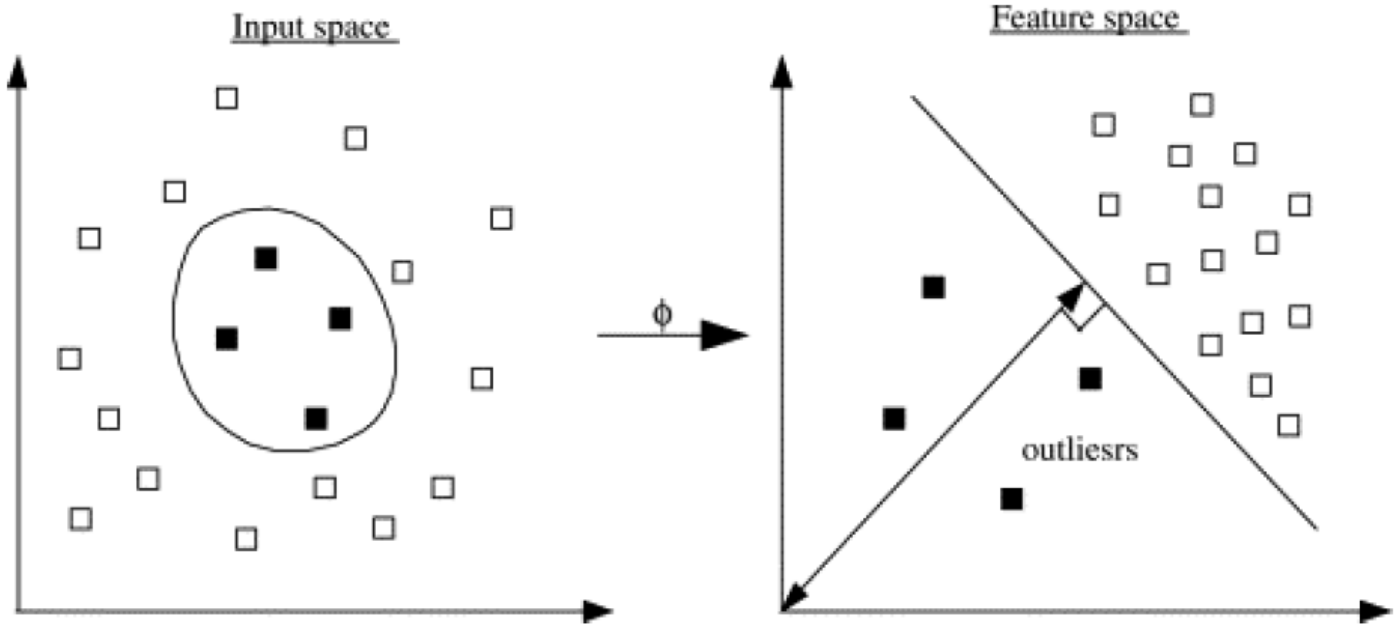
$$\xi_n \geq 0 \quad \forall_n$$

- v can be seen as a barrier to the proportion of patterns outside the hypersphere.

Novelty detection algorithms (24)

One-Class SVM

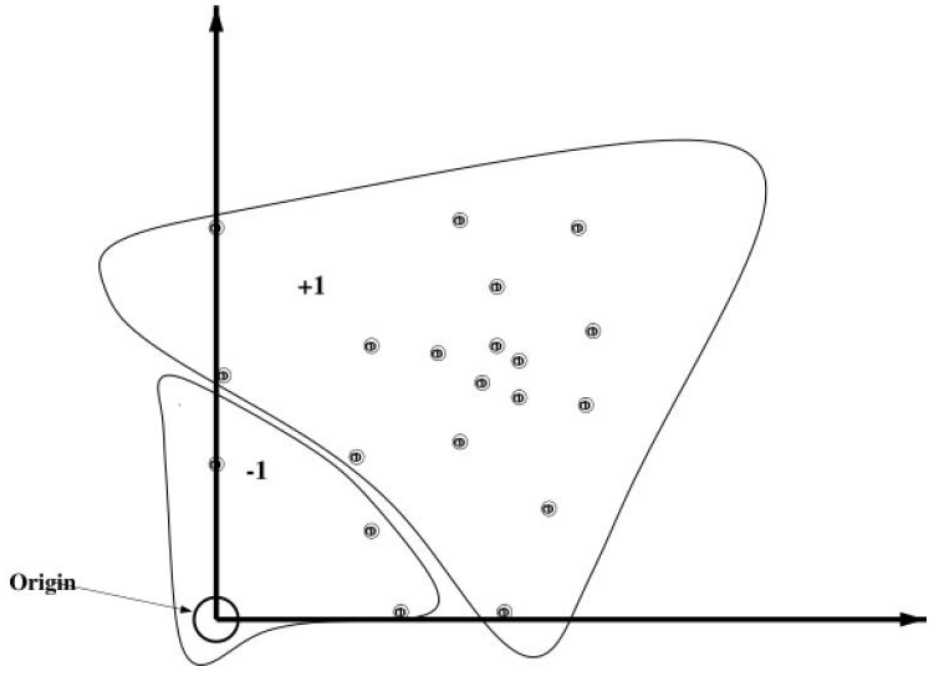
- Kernel-trick: Implicit transformation into (high dimensional) feature space
- Goal: Data set becomes linearly separable



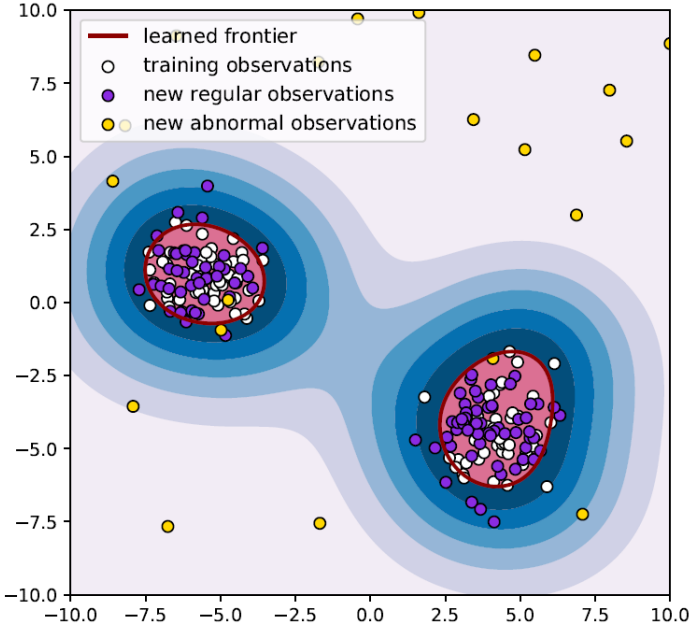
Novelty detection algorithms (25)

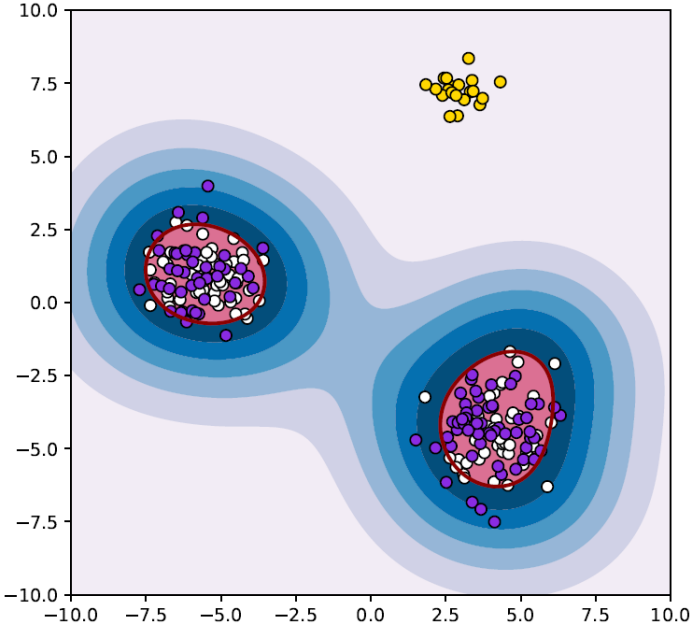
One-Class SVM

- Support estimation of (high dimensional) distributions
- Origin is only original sample belonging to the anomaly class (-1)
- Relaxation parameters to separate image from origin (cf. C-SVM, ν-SVM)



Novelty detection algorithms (26)

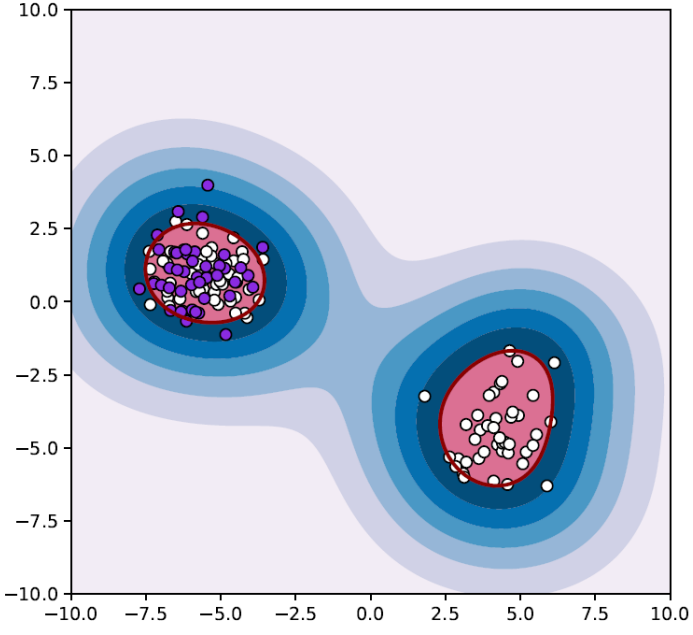


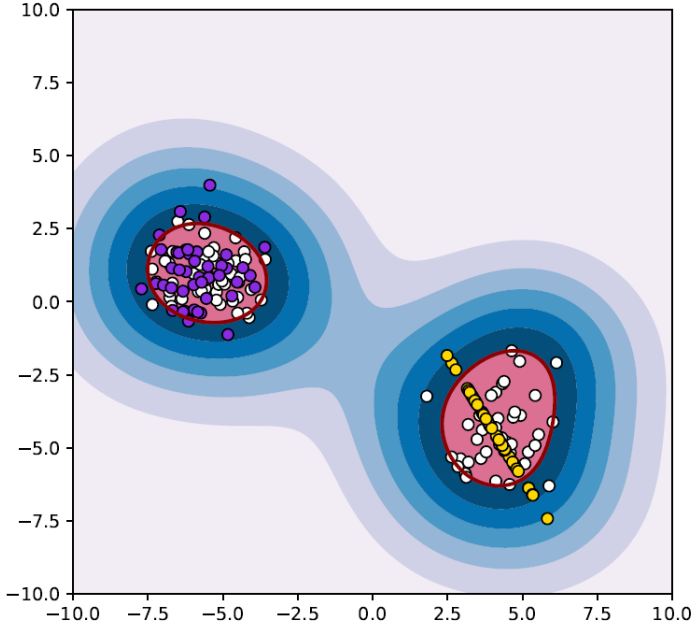


Novelty detection algorithms (27)

Christian-Albrechts-Universität zu Kiel

Negative examples: Obsoleteness and distribution change





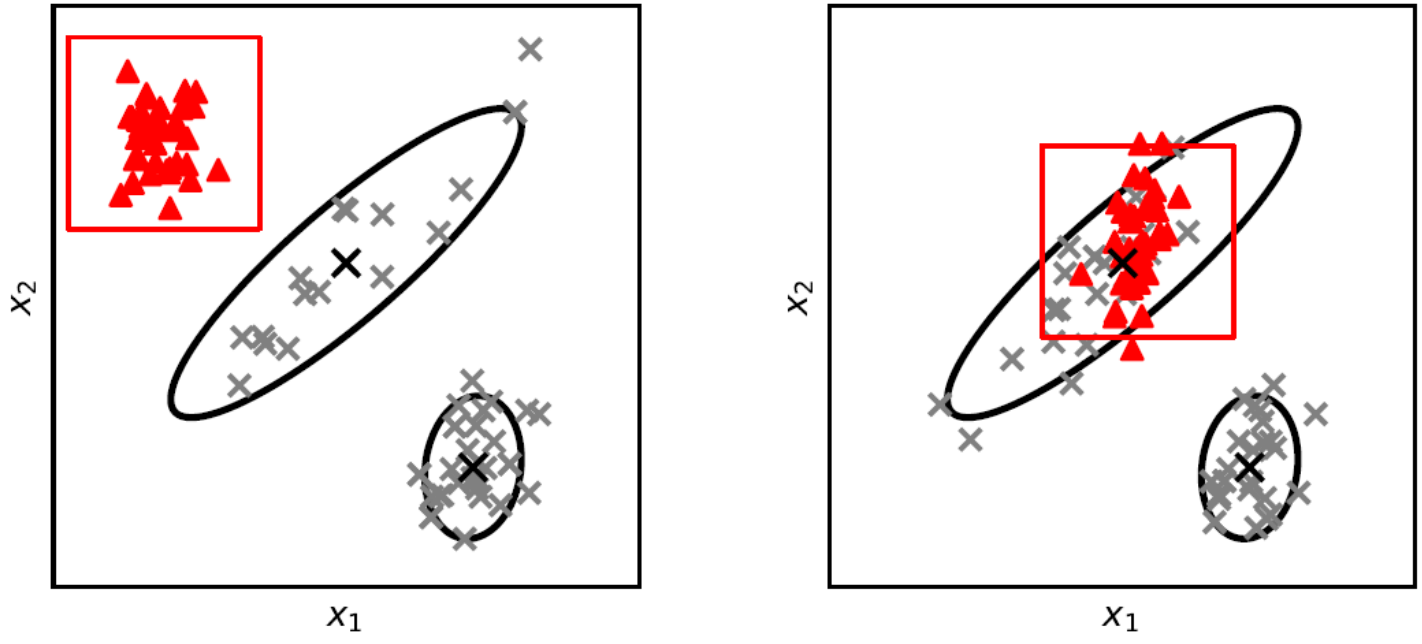
Novelty detection algorithms (28)

One Class SVM

- Only able to detect anomalies residing beyond the estimated support
  - Anomalies in low density region
- Samples are assumed to be i.i.d.
  - No detection of relations between anomalies
- Support estimation ≠ density distribution
  - No detection in high density regions

Novelty detection algorithms (29)

High density regions (HDR) and low density regions (LDR)

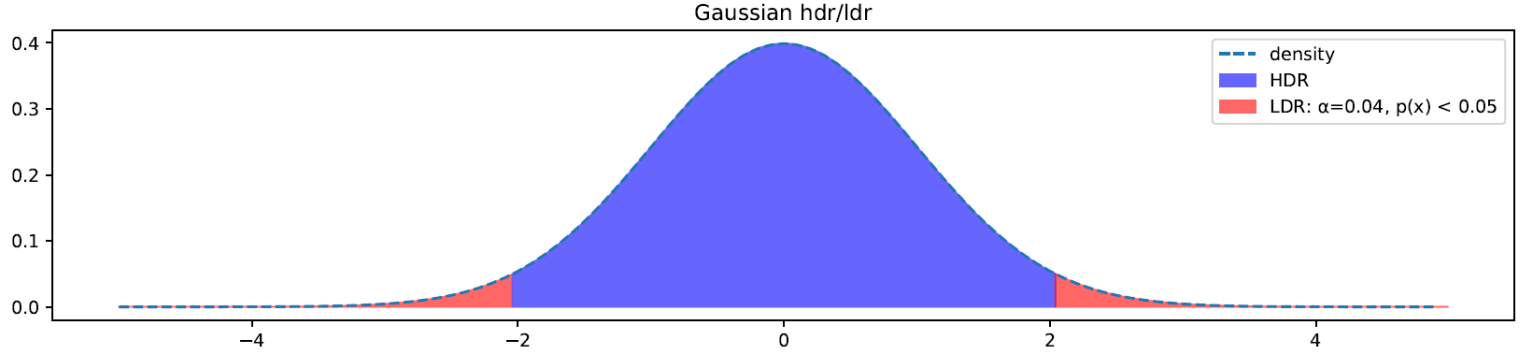


Novelty detection algorithms (30)

Masking of regions

Novelty detection algorithms (31)

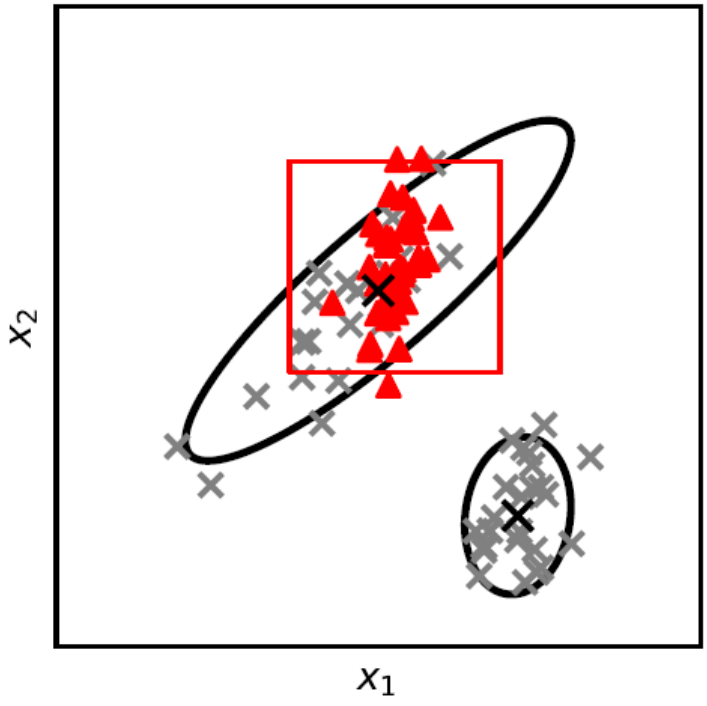
High and low density regions

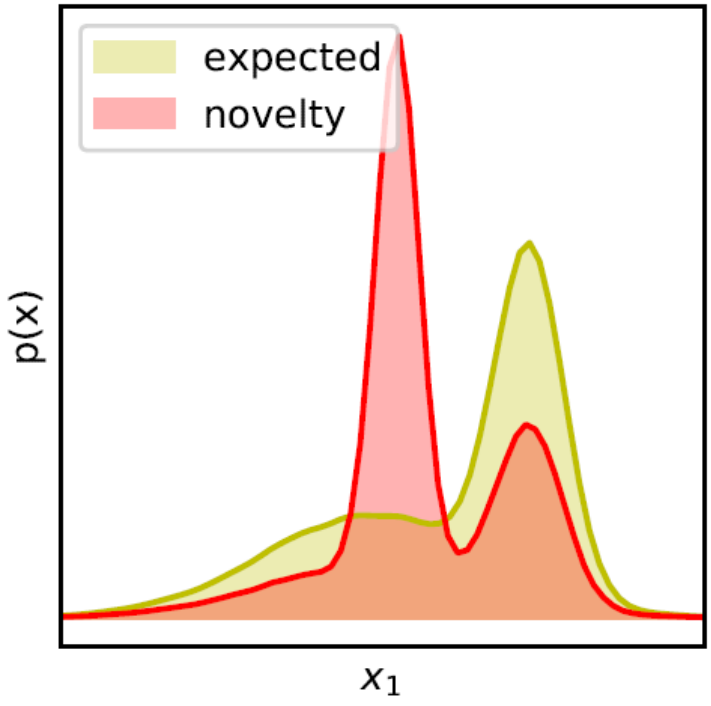


- HDR: Future observations are expected to appear in these regions
  - Multiple HDRs, e.g., each component of a GMM
  - Different novelty types as in LDR (for example, changing mean and covariance)
- LDR: Future observations are not expected to appear here
  - Observations in LDR ! anomalies
  - Usually one large LDR
- Transition between HDR and LDR is fluid (user chosen)

Novelty detection algorithms (32)

Detection of HDR novelties





Novelty detection algorithms (33)

Christian-Albrechts-Universität zu Kiel

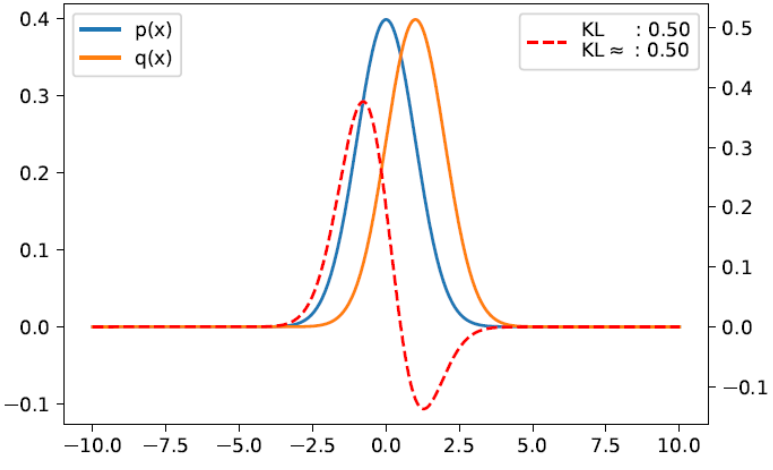
Divergence

- "Distance" between two probability distributions
- Here: Kullback-Leibler

$$KL(p||q) = -\int p(x) \mathrm{ld} \frac{p(x)}{q(x)} \mathrm{d}x$$

$$KL(p||q) \neq KL(q||p)$$

- $KL(p||q)$: Number of required bits to encode q with p



Novelty detection algorithms (34)

Kullback-Leibler is not symmetric

- How can we easily transform KL into a symmetric variant?

- Solution:

Solution

Novelty detection algorithms (35)

Kullback-Leibler divergence and novelty detection:

- How to choose an appropriate threshold to quantify *significant* divergence?
  - Divergence will decrease with increasing dimensionality.
- What is the ideal domain to sample from?
  - KL($p||q$) is only increased in areas where $p(x)$ has support.
  - KL2(p, q) doubles computational cost.
- Locality: local evaluation of divergence, solves also sampling domain problem

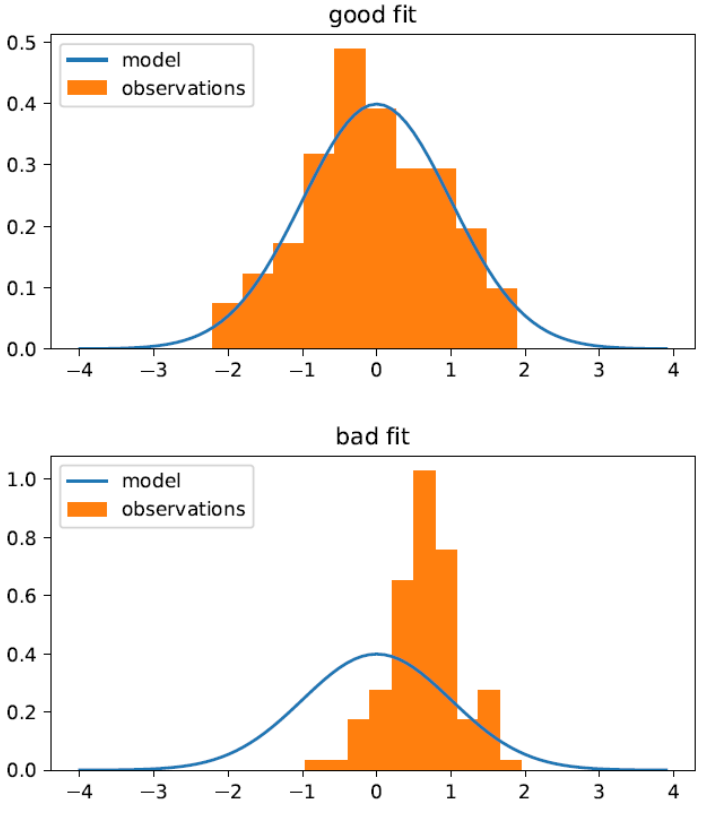
Novelty detection algorithms (36)

Goodness of fit

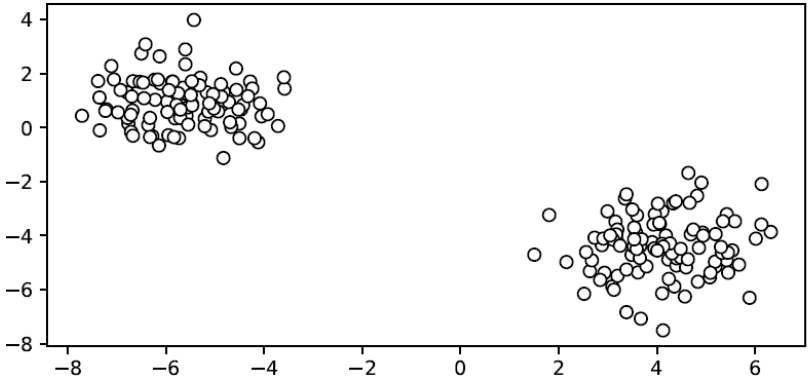
- Test how well a model (e.g., GMM) fits a set of observations
- Locality: affiliate observations with HDR (component of a GMM)
  - Responsibilities

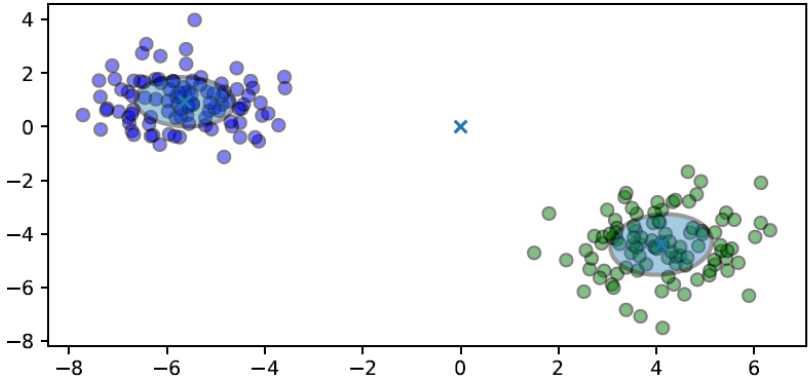
$$\gamma_j(x') = \frac{\pi_j \mathcal{N}(x' | \mu_j, \Sigma_j)}{\sum_{j'=1}^{J} \pi_{j'} \mathcal{N}(x' | \mu_{j'}, \Sigma_{j'})}$$

  - Component j with highest responsibility is affiliated with x'
  - Test if HDR has changed → HDR novelty detection



Novelty detection algorithms (37)





Novelty detection algorithms (38)

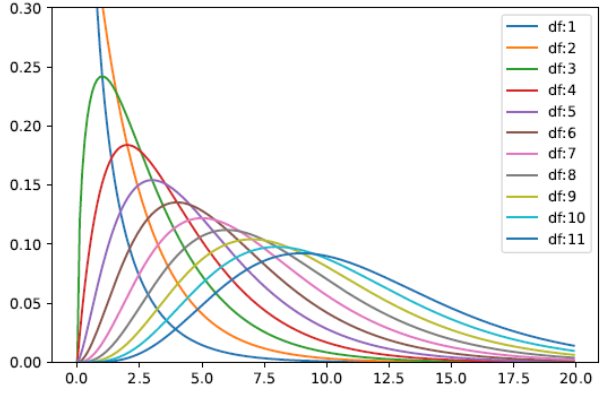
Chi-Squared ($\mathcal{X}^2$) Test

- Goodness of fit test for discrete distributions
- Applicable for continuous distributions by discretisation (e.g. histogram)
- Test statistic ($\boldsymbol{X}^2$) is weighted sum of deviations from the expectation
  - o_i number of actual observations of event i
  - m_i expectation for the number of events
- Test statistic$\boldsymbol{X}^2$ is $\mathcal{X}^2$distributed if H_0 (no difference between distributions) is true
- α is the *significance level* (*type I error* rate)

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - m_i)^2}{m_i}$$

$$m_i = p_i \cdot \sum_{i=1}^{k} o_i$$

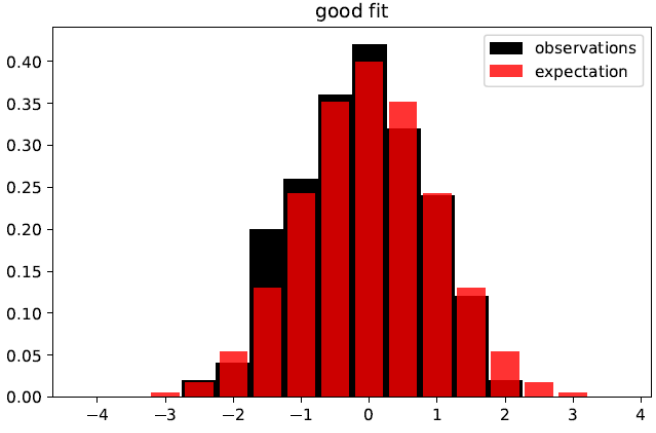
$$X^2 > F^{-1}_{\chi^2_\nu}(1-\alpha)$$

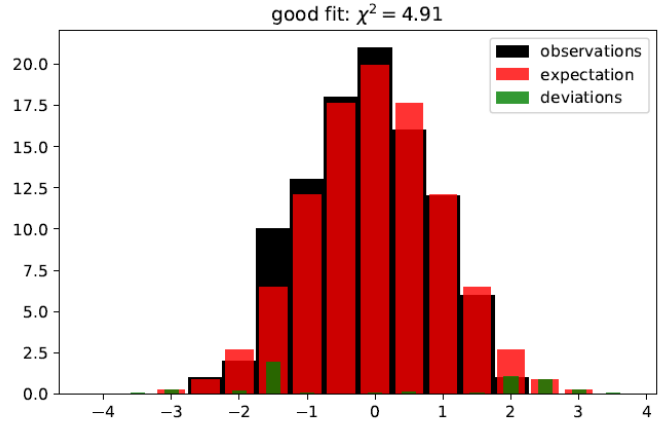


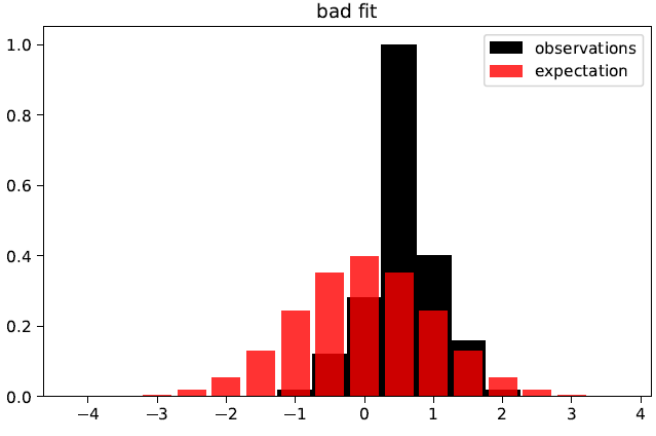
Novelty detection algorithms (39)

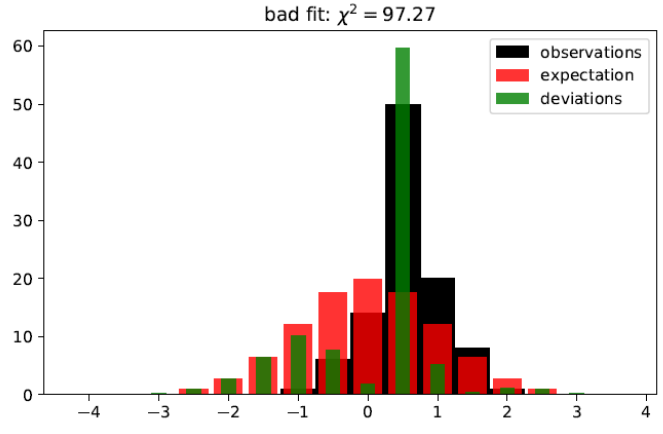
Christian-Albrechts-Universität zu Kiel

Chi-Squared Test





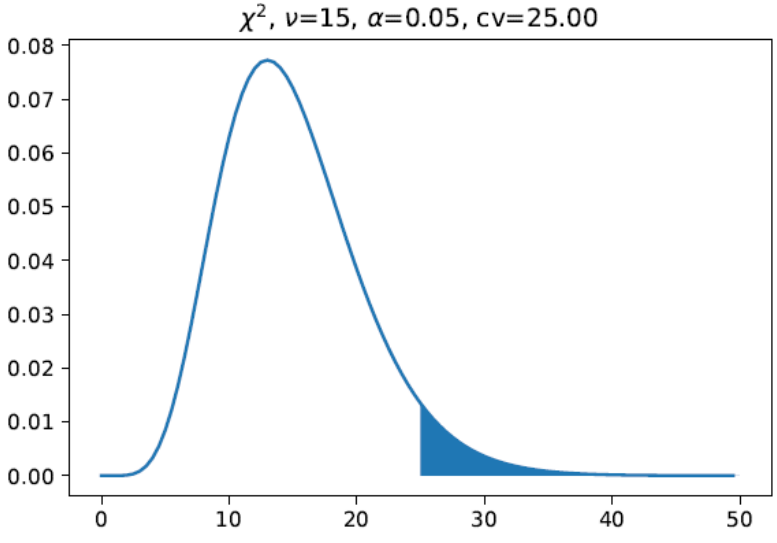




Novelty detection algorithms (40)

Chi-Squared Test

- Reject H_0 if X² exceeds a critical value cv
  - $cv = F^{-1}_{\chi^2_\nu}(1-\alpha)$
  - ν degrees of freedom, usually the number of events reduced by the number of estimated parameters plus 1
- Test becomes unreliable if $m_i < 5$ for more than 20% of the bins



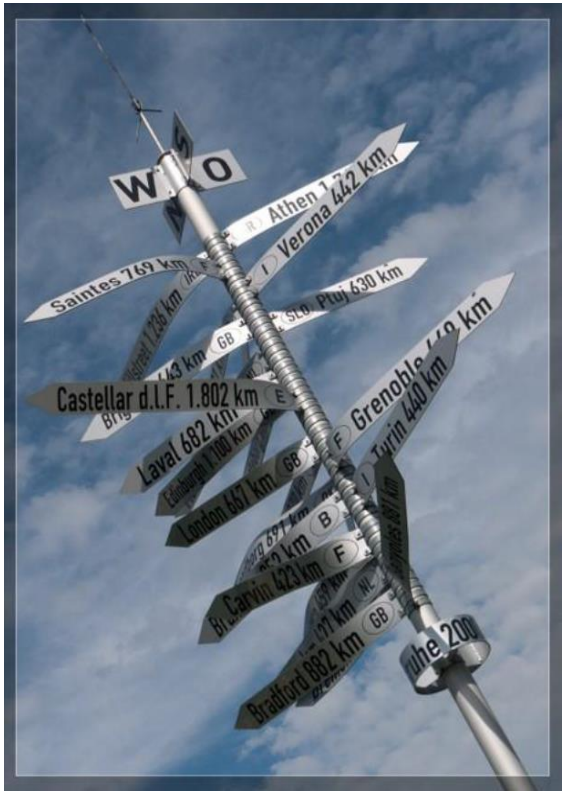
Novelty detection algorithms (41)

Distribution testing for novelty detection

- Sensitive to samples in LDR (e.g. noise)
- Requires discretisation or dimension reduction
- Suitable for HDR detection
  - Exploitation of locality
  - Compare model against sliding window
  - Rejection of H_0 → indicator for novelty
- Use moving average if window is small and type I error rate too high

Agenda

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings



About this Chapter

Christian-Albrechts-Universität zu Kiel

This chapter described:

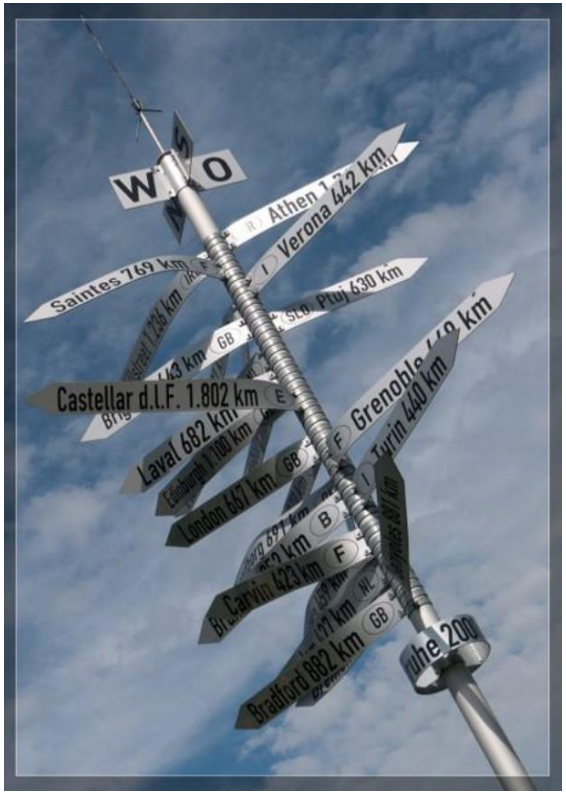
- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings

Students should now be able to:

- motivate and define the process of motif detection.
- explain the different classes of anomaly detection approaches.
- distinguish between outlier, anomaly, novelty, and noise.
- describe and apply the most prominent novelty detection algorithms.

Agenda

- Motif search and detection
- Anomaly detection: Motivation
- Anomaly/novelty detection: Formalisation
- Novelty detection algorithms
- Conclusion
- Further readings



Further readings

- Mueen, Keogh, Zhu, Cash, Westover, *Exact Discovery of Time Series Motifs*, 2009
- Prechelt, PROBEN 1 - a set of benchmarks and benchmarking rules for neural network training algorithms, 1994
- Goldberger, Amaral, Glass, Hausdorff, Jeffrey, Ivanov, Mark, Mietus, Moody, Peng, Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, 2000

End

Christian-Albrechts-Universität zu Kiel

- Questions….?