Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection

Björn Barz, Erik Rodner, Yanira Guanche Garcia, and Joachim Denzler, *Member, IEEE*

Abstract—Automatic detection of anomalies in space- and time-varying measurements is an important tool in several fields, e.g., fraud detection, climate analysis, or healthcare monitoring. We present an algorithm for detecting anomalous regions in multivariate spatio-temporal time-series, which allows for spotting the interesting parts in large amounts of data, including video and text data. In opposition to existing techniques for detecting isolated anomalous data points, we propose the "Maximally Divergent Intervals" (MDI) framework for unsupervised detection of coherent spatial regions and time intervals characterized by a high Kullback-Leibler divergence compared with all other data given. In this regard, we define an unbiased Kullback-Leibler divergence that allows for ranking regions of different size and show how to enable the algorithm to run on large-scale data sets in reasonable time using an interval proposal technique. Experiments on both synthetic and real data from various domains, such as climate analysis, video surveillance, and text forensics, demonstrate that our method is widely applicable and a valuable tool for finding interesting events in different types of data.

Index Terms—Anomaly detection, time series analysis, spatio-temporal data, data mining, unsupervised machine learning

1 Introduction

Many pattern recognition methods strive towards deriving models from complex and noisy data. Such models try to describe the prototypical normal behavior of the system being observed, which is hard to model manually and whose state is often not even directly observable, but only reflected by the data. They allow reasoning about the properties of the system, predicting unseen data, and assessing the "normality" of new data. In such a scenario, any deviation from the normal behavior present in the data is distracting and may impair the accuracy of the model. An entire arsenal of techniques has therefore been developed to eliminate abnormal observations prior to learning or to learn models in a robust way not affected by a few anomalies.

Such practices may easily lead to the perception of anomalies as being intrinsically bad and worthless. Though that is true for random noise and erroneous measurements, there may also be anomalies caused by rare events and complex processes. Embracing the anomalies in the data and studying the information buried in them can therefore lead to a deeper understanding of the system being analyzed and to the insight that the models hitherto employed were incomplete or—in the case of non-stationary processes—outdated. A well-known example for this is the discovery of the correlation between the *El Niño* weather phenomenon and extreme surface pressures over the equator by Gilbert Walker [1] during the early 20th century through the analysis of extreme events in time-series of climate data.

Thus, the use of anomaly detection techniques is not limited to outlier removal as a pre-processing step. In contrast, anomaly detection also is an important task *per se*, since only the deviations from normal behavior are the actual object of interest in many applications. Besides the scenario of knowledge discovery mentioned above, fraud detection (e.g., credit card fraud or identity theft), intrusion detection in cyber-security, fault detection in industrial processes, anomaly detection in healthcare (e.g., monitoring patient condition or detecting disease outbreaks), and early detection of environmental disasters are other important examples. Automated methods for anomaly detection are especially crucial nowadays, where huge amounts of data are available that cannot be analyzed by humans.

In this article, we introduce a novel unsupervised method called "Maximally Divergent Intervals" (MDI), which can be employed to point the expert analysts to the interesting parts of the data, i.e., the anomalies. In contrast to most existing anomaly detection techniques (e.g., [2], [3], [4], [5]), we do not analyze the data on a point-wise basis, but search for contiguous intervals of time and regions in space that contain the anomalous event. This is motivated by the fact that anomalies driven by natural processes rather occur over a space of time and, in the case of spatio-temporal data, in a spatial region rather than at a single location at a single time. Moreover, the individual samples making up such a so-called *collective anomaly* do not have to be anomalous when considered in isolation, but may be an anomaly only as a whole. Thus, analysts will intuitively be searching for anomalous *regions* in the data

- *B. Barz, Y. Guanche Garcia, and J. Denzler are with the Computer Vision Group, Department of Mathematics and Computer Science, Friedrich Schiller University Jena, Jena 07737, Germany.
E-mail: {bjoern.barz, yanira.guanche.garcia, joachim.denzler}@uni-jena.de.*
- *E. Rodner is with the Corporate Research and Technology, Carl Zeiss AG, Oberkochen, Germany. E-mail: Erik.Rodner@uni-jena.de.*

*Manuscript received 16 Nov. 2017; revised 12 Mar. 2018; accepted 29 Mar. 2018. Date of publication 29 Apr. 2018; date of current version 10 Apr. 2019.
(Corresponding author: Björn Barz.)
Recommended for acceptance by H. Kjellstrom.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2018.2823766*

instead of anomalous points and the algorithm assisting them should do so as well.

We achieve this by searching for anomalous *blocks* in multivariate spatio-temporal data tensors, i.e., regions and time frames whose data distribution deviates most from the distribution of the remaining time-series. To this end, we compare several existing measures for the divergence of distributions and derive a new one that is invariant against varying length of the intervals being compared. A fast novel interval proposal technique allows us to reduce the computational cost of this procedure by just analyzing a small portion of particularly interesting parts of the data. Experiments on climate data, videos, and text corpora will demonstrate that our method can be applied to a variety of applications without major adaptations.

Despite the importance of this task across domains, there has been very limited research on the detection of anomalous intervals in multivariate time-series data, though this problem has been known for a couple of years: Keogh et al. [6] have already tackled this task in 2005 with a method they called “HOT SAX”. They try to find anomalous subsequences (“discords”) of time-series by representing all possible sub-sequences of length d as a d-dimensional vector and using the euclidean distance to the nearest neighbor in that space as anomaly score. More recently, Ren et al. [7] use hand-crafted interval features based on the frequency of extreme values and search for intervals whose features are maximally different from all other intervals. However, both methods are limited to univariate data and a fixed length of the intervals must be specified in advance.

The latter is also true for a multivariate approach proposed by Liu et al. [8] who compare two consecutive intervals of fixed size in a time-series using the Kullback-Leibler or the Pearson divergence for detecting *change-point anomalies*, i.e., points where a permanent change of the distribution of the data occurs. This is a different task than finding intervals that are anomalous with regard to *all* the remaining data. In addition, their method does not scale well for detecting anomalous intervals of *dynamic* size and is hence not applicable for detecting other types of anomalies, for which a broader context has to be taken into account.

The task of detecting anomalous intervals of dynamic size has recently been tackled by Senin et al. [9], who search for typical and anomalous patterns in time-series by inducing a grammar on a symbolic discretization of the data. As opposed to our approach, their method cannot handle multivariate or spatio-temporal data.

Similar to our approach, Jiang et al. [10] search for anomalous blocks in higher-order tensors using the Kullback-Leibler divergence, but apply their method to discrete data only (e.g., relations in social networks) and use a Poisson distribution for modeling the data. Since their search strategy is very specific to applications dealing with graph data, it is not applicable in the general case for multivariate continuous data dealt with in our work.

Regarding spatio-temporal data, Wu et al. [11] follow a sequential approach for detecting anomalies first spatially, then temporally and apply a merge-strategy afterwards. However, the time needed for merging grows exponentially with the length of the time-series and their divergence measure is limited to binary-valued data. In contrast to this, our approach is able to deal with multivariate real-valued data efficiently and treats time and space jointly.

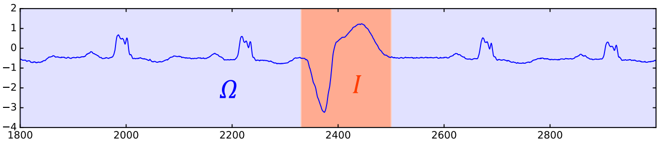


Fig. 1. Schematic illustration of the principle of the MDI algorithm: The distribution of the data in the inner interval I is compared with the distribution of the remaining time-series in the outer interval Ω.

The remainder of this article is organized as follows: Section 2 will introduce our novel “Maximally Divergent Intervals” algorithm for off-line detection of collective anomalies in multivariate spatio-temporal data. Its performance will be evaluated quantitatively on artificial data in Section 3 and its suitability for practical applications will be demonstrated by means of experiments on real data from various different domains in Section 4. Section 5 will summarize the progress made so far and mention directions for future research.

2 Maximally Divergent Intervals

This section formally introduces our MDI algorithm for off-line detection of anomalous intervals in spatio-temporal data. After a set of definitions that we are going to make use of, we start by giving a very rough overview of the basic idea behind the algorithm, which is also illustrated schematically in Fig. 1. The subsequent sub-sections will go into more detail on the individual aspects and components of our approach.

Our implementation of the MDI algorithm is available as open source at: https://cvjena.github.io/libmaxdiv/

2.1 Definitions

Let $\mathfrak{X} \in \mathbb{R}^{T \times X \times Y \times Z \times D}$ be a multivariate spatio-temporal time-series given as 5th-order tensor with 4 contextual attributes (point of time and spatial location) and D behavioral attributes for all $N := T \cdot X \cdot Y \cdot Z$ samples. We will index individual samples using 4-tuples $i \in \mathbb{N}^4$ like in $\mathfrak{X}_i \in \mathbb{R}^D$.

The usual interval notation $[\ell, r)$ will be used in the following for discrete intervals $\{t \in \mathbb{N} | \ell \le t < r\}$. Furthermore, the set of all intervals with size between a and b along an axis of size n is denoted by

$$\mathfrak{I}^n_{a,b} := \{[\ell, r) \,|\, 1 \le \ell < r \le n+1 \wedge a \le r - \ell \le b\}\,. \tag{1}$$

The set of all sub-blocks of a data tensor $\mathfrak{X}$ complying with given size constraints $A = (a_t, a_x, b_y, b_z), B = (b_t, b_x, b_y, b_z)$ can then be defined as

$$\begin{aligned}\mathfrak{I}_{A,B} := \{I_t \times I_x \times I_y \times I_z \,|\, I_t \in \mathfrak{I}^T_{a_t,b_t} \wedge I_x \in \mathfrak{I}^X_{a_x,b_x} \wedge \\ I_y \in \mathfrak{I}^Y_{a_y,b_y} \wedge I_z \in \mathfrak{I}^Z_{a_z,b_z}\}.\end{aligned} \tag{2}$$

In the following, we will often omit the indices for simplicity and just refer to it as $\mathfrak{I}$.

Given any sub-block $I \in \mathfrak{I}_{A,B}$, the remaining part of the time-series excluding that specific range can be defined as

$$\Omega(I) := ([1, T] \times [1, X] \times [1, Y] \times [1, Z]) \setminus I, \tag{3}$$

and we will often simply refer to it as Ω if the corresponding range I is obvious from the context.

2.2 Idea and Algorithm Overview

The approach pursued by the MDI algorithm to compute anomaly scores for all intervals $I \in \mathfrak{I}$ can be motivated by a long-standing definition of anomalies given by Douglas Hawkins [12] in 1980, who defines an anomaly as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". In analogy to this definition, the MDI algorithm assumes that there is a sub-block $I \in \mathfrak{I}$ of the given time-series that has been generated according to "a different mechanism" than the rest of the time-series in Ω (cf. the schematic illustration in Fig. 1). The algorithm tries to capture these mechanisms by modelling the probability density p_I of the data in the inner interval I and the distribution p_Ω in the outer interval Ω. We investigate two different models for these distributions: Kernel Density Estimation (KDE) and multivariate normal distributions (Gaussians), which will be explained in detail in Section 2.3.

Moreover, a measure $\mathfrak{D}(p_I, p_\Omega)$ for the degree of "deviation" of p_I from p_Ω has to be defined. Like some other works on collective anomaly detection [8], [10], we use—among others—the *Kullback-Leiber (KL) divergence* for this purpose. However, Section 2.5 will show that this is a sub-optimal choice when used without a slight modification and discuss alternative divergence measures.

Given these ingredients, the underlying optimization problem for finding the most anomalous interval can be described as

$$\hat{I} = \underset{I \in \mathfrak{I}_{A,B}}{\operatorname{argmax}} \ \mathfrak{D}\big(p_I, p_{\Omega(I)}\big). \tag{4}$$

Various possible choices for the divergence measure $\mathfrak{D}$ will be discussed in Section 2.5.

In order to actually locate this "maximally divergent interval" $\hat{I}$, the MDI algorithm scans over all intervals $I \in \mathfrak{I}_{A,B}$, estimates the distributions p_I and p_Ω and computes the divergence between them, which becomes the anomaly score of the interval I. The parameters A and B, which define the minimum and the maximum size of the intervals in question, have to be specified by the user in advance. This is not a severe restriction, since extreme values may be chosen for these parameters in exchange for increased computation time. But depending on the application and the focus of the analysis, there is often prior knowledge about reasonable limits for the size of possible intervals.

After the anomaly scores have been obtained for all intervals, they are sorted in descending order and non-maximum suppression is applied to obtain non-overlapping intervals only. For large time-series with more than 10k samples, we apply an approximative non-maximum suppression that avoids storing all interval scores by maintaining a fixed-size list of currently best-scoring non-overlapping intervals.

Finally, the algorithm returns a ranking of intervals, so that a user-specified number of top k intervals can be selected as output.

2.3 Probability Density Estimation

The divergence measure used in (4) requires the notion of the distribution of the data in the intervals I and Ω. We will hence discuss in the following, which models we employ to estimate these distributions and how this can be done efficiently.

2.3.1 Models

The choice of a specific model for the distributions p_I and p_Ω imposes some assumptions about the data which may not conform to reality. However, since the MDI algorithm estimates the parameters of those distributions for all possible intervals in the time-series, the use of models that can be updated efficiently is crucial. One such model is Kernel Density Estimation (KDE) with

$$p_{\mathfrak{S}}(\mathfrak{X}_i) = \frac{1}{|\mathfrak{S}|} \sum_{j \in \mathfrak{S}} k(\mathfrak{X}_i, \mathfrak{X}_j), \qquad \mathfrak{S} \in \{I, \Omega\}, \tag{5}$$

using a Gaussian kernel

$$k(x, y) = \left(2\pi\sigma^2\right)^{-\frac{D}{2}} \cdot \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right). \tag{6}$$

On the one hand, KDE is a very flexible model, but on the other hand, it does not scale well to long time-series and does not take correlations between attributes into account. The second proposed model does not expose these problems: It assumes that both the data in the anomalous interval I and in the remaining time-series Ω are distributed according to multivariate normal distributions (*Gaussians*) $\mathcal{N}(\mu_I, S_I)$ and $\mathcal{N}(\mu_\Omega, S_\Omega)$, respectively.

2.3.2 Efficient Estimation with Cumulative Sums

Both distribution models described above involve a summation over all samples in the respective interval. Performing this summation for multiple intervals is redundant, because some of them overlap with each other. Such a naïve approach of finding the maximally divergent interval has a time complexity of $\mathcal{O}(N^2 \cdot L^2)$ with KDE and $\mathcal{O}(N \cdot L \cdot (N + L)) \subseteq \mathcal{O}(N^2 \cdot L)$ with Gaussian distributions. This is due to the number of $\mathcal{O}(N \cdot L)$ intervals (with $L = (b_t - a_t + 1) \cdot (b_x - a_x + 1) \cdot (b_y - a_y + 1) \cdot (b_z - a_z + 1)$ being the maximum volume of an interval), each of them requiring a summation over $\mathcal{O}(L)$ samples for the evaluation of one of the divergence measures described later in Section 2.5. For KDE, $\mathcal{O}(N)$ distance computations are necessary for the evaluation of the probability density function for each sample, while for Gaussian distributions a summation over all $\mathcal{O}(N)$ samples has to be performed for each interval to estimate the parameters of the distributions.

This would be clearly infeasible for large-scale data. However, these computations can be sped up significantly by using cumulative sums [13]. For the sake of clarity, we first consider the special case of a non-spatial time-series $(x_t)_{t=1}^n, x_t \in \mathbb{R}^D$. With regard to KDE, a matrix $C \in \mathbb{R}^{n \times n}$ of cumulative sums of kernelized distances can be used:

$$C_{t,t'} = \sum_{t''=1}^{t'} k(x_t, x_{t''}) \ . \tag{7}$$

This matrix has to be computed only once, which requires $\mathcal{O}(n^2)$ distance calculations, and can then be used to estimate the probability density functions of the data in the intervals $I = [a, b)$ and $\Omega = [1, n] \setminus I$ in constant time:

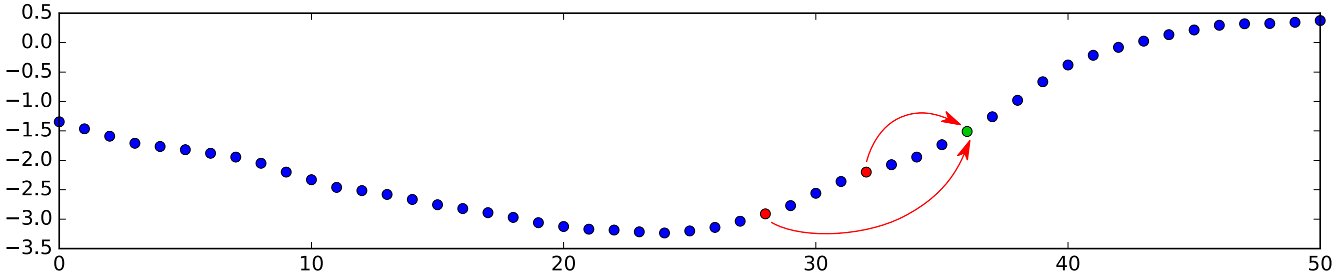


Fig. 2. Illustration of time-delay embedding with $\kappa = 3, \tau = 4$. The attribute vector of each sample is augmented with the attributes of the samples 4 and 8 time steps earlier.

$$
\begin{aligned}
p_I(x_t) &= \frac{C_{t,b-1} - C_{t,a-1}}{|I|} \quad , \\
p_\Omega(x_t) &= \frac{C_{t,n} - C_{t,b-1} + C_{t,a-1}}{n - |I|} \quad .
\end{aligned} \tag{8}
$$

In analogy, a matrix $C^\mu \in \mathbb{R}^{D\times n}$ of cumulative sums over the samples and a tensor $C^S \in \mathbb{R}^{D\times D\times n}$ of cumulative sums over the outer products of the samples can be used to speed up the estimation of the parameters of Gaussian distributions:

$$
C^\mu_t = \sum_{t'=1}^{t} x_{t'}, \quad C^S_t = \sum_{t'=1}^{t} x_{t'} \cdot x_{t'}^\top , \tag{9}
$$

where C^μ_t and C^S_t are the t^{th} column of C^μ and the t^{th} $D \times D$ matrix of C^S, respectively. Using these matrices, the mean vectors and covariance matrices can be estimated in constant time.

This technique can be generalized to the spatio-temporal scenario using higher order tensors for storing the cumulative sums. The reconstruction of a sum over a given range from such a cumulative tensor follows the *Inclusion-Exclusion Principle* and the number of summands involved in the computation grows, thus, exponentially with the order of the tensor, being 16 for a 4^{th}-order tensor, compared to only 2 summands in the non-spatial case. The exact equation describing the reconstruction in the general case of an M^{th}-order tensor is given in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2018.2823766.

Thanks to the use of cumulative sums, the computational complexity of the MDI algorithm is reduced to $\mathcal{O}(N^2 + N \cdot L^2)$ for the case of KDE and to $\mathcal{O}(N \cdot L^2)$ for Gaussian distributions.

2.4 Incorporation of Context

The models used for probability density estimation described in the previous section are based on the assumption of independent samples. However, this assumption is almost never true for real data, since the value at a specific point of time and spatial location is likely to be strongly correlated with the values at previous times and nearby locations. To mitigate this issue, we apply two kinds of embeddings that incorporate context into each sample as pre-processing step.

2.4.1 Time-Delay Embedding

Aiming to make combinations of observed values more representative of the hidden state of the system being observed, *time-delay embedding* [14] incorporates context from previous time-steps into each sample by transforming a given time-series $(x_t)_{t=1}^n, x_t \in \mathbb{R}^D$, into another time-series $(x'_t)_{t=1+(\kappa-1)\tau}^n$, $x'_t \in \mathbb{R}^{\kappa D}$, given by

$$
x'_t = \begin{pmatrix} x_t^\top & x_{t-\tau}^\top & x_{t-2\tau}^\top & \cdots & x_{t-(\kappa-1)\cdot\tau}^\top \end{pmatrix}^\top , \tag{10}
$$

where the *embedding dimension* κ specifies the number of samples to stack together and the *time lag* τ specifies the gap between two consecutive time-steps to be included as context. An illustrative example is given in Fig. 2.

This method is often motivated by Takens' theorem [15], which, roughly, states that for a certain embedding dimension $\bar{\kappa}$ the hidden state of the system can be reconstructed given the observations of the last $\bar{\kappa}$ time-steps.

2.4.2 Spatial-Neighbor Embedding

Correlations between nearby spatial locations are handled similarly: In addition to time-delay embedding, each sample of a spatio-temporal time-series can be augmented by the features of its spatial neighbors (cf. Fig. 3) to enable the detection of spatial or spatio-temporal anomalies. This pre-processing step, which we refer to as *spatial-neighbor embedding*, is parametrized with 3 parameters $\kappa_x, \kappa_y, \kappa_z$ for the embedding dimension along each spatial axis and 3 parameters τ_x, τ_y, τ_z for the lag along each axis.

Note that, in contrast to time-delay embedding, neighbors from both directions are aggregated, since spatial context is bilinear. For example, $\kappa_x = 3$ would mean to consider 4 neighbors along the x-axis, 2 in each direction.

Spatial-neighbor embedding can either be applied before or after time-delay embedding. As opposed to many spatio-temporal anomaly detection approaches that perform temporal and spatial anomaly detection sequentially (e.g., [11], [16], [17]), the MDI algorithm in combination with the two embeddings allows for a joint optimization. However, it implies a much more drastic multiplication of the data size.

2.5 Divergences

A suitable measure for the deviation of the distribution p_I from p_Ω is an essential part of the MDI algorithm. The following sub-sections introduce several divergence measures we have investigated and propose a modification to the well-known Kullback-Leibler (KL) divergence that is necessary for being able to compare divergences of distributions estimated from intervals of different size.

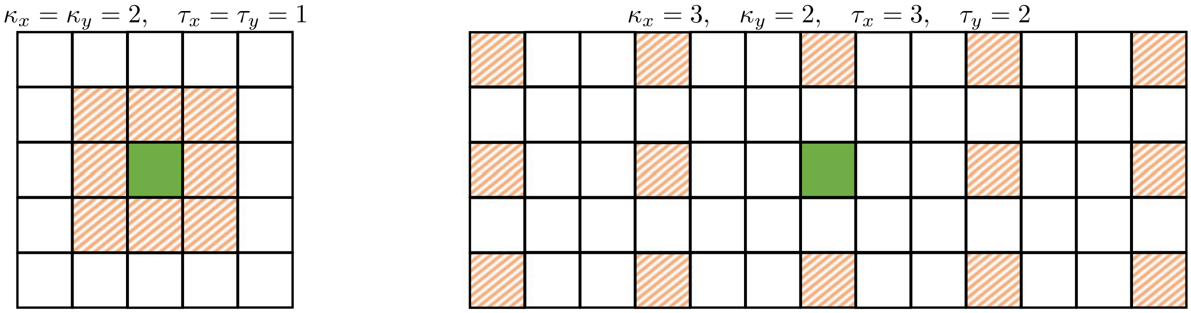

$\kappa_x = \kappa_y = 2, \quad \tau_x = \tau_y = 1$
$\kappa_x = 3, \quad \kappa_y = 2, \quad \tau_x = 3, \quad \tau_y = 2$

Fig. 3. Exemplary illustration of spatial-neighbor embedding with different parameters. The attribute vector of the sample with a solid fill color is augmented with the attributes of the samples with a striped pattern.

2.5.1 Cross Entropy

Numerous divergence measures, including those described in the following, have been derived from the domain of *information theory*. Being one of the most basic information theoretic concepts, the *cross entropy* between two distributions given by their probability density functions p and q may already be used as a divergence measure:

$$\mathfrak{D}_{\mathrm{CE}}(p,q) := \mathrm{H}(p,q) := \mathbb{E}_p[-\log q]\ . \tag{11}$$

Cross entropy measures how surprising a sample drawn from p is, assuming that it would have been drawn from q, and is hence already eligible as a divergence measure, since the unexpectedness grows when p and q are very different.

Since the MDI algorithm assumes, that the data in the intervals $I \in \mathfrak{I}$ and Ω have been sampled from the distributions corresponding to p_I and p_Ω, respectively, the cross entropy of the two distributions can be approximated empirically from the data:

$$\widetilde{\mathfrak{D}_{\mathrm{CE}}}(I,\Omega) = \frac{1}{|I|}\sum_{i \in I} \log p_\Omega(\mathfrak{X}_i). \tag{12}$$

This approximation has the advantage of having to estimate only one probability density, $p_\Omega(x_t)$, explicitly. This is particularly beneficial, since the possibly anomalous intervals I often contain only few samples, so that an accurate estimation of the probability density p_I is difficult.

2.5.2 Kullback-Leibler Divergence

The *Kullback-Leibler (KL) divergence* is a popular divergence measure that builds upon the fundamental concept of cross entropy. Given two distributions p and q, the KL divergence can be defined as follows:

$$\mathfrak{D}_{\mathrm{KL}}(p,q) := \mathrm{H}(p,q) - \mathrm{H}(p,p) = \mathbb{E}_p\left[-\log \frac{p}{q}\right]. \tag{13}$$

As opposed to the pure cross entropy of p and q, the KL divergence does not only take into account how well p is explained by q, but also the intrinsic entropy $\mathrm{H}(p,p) =: \mathrm{H}(p)$ of p, so that an interval with a stable distribution would get a higher score than an oscillating one if they had the same cross entropy with the rest of the time-series.

Like cross entropy, the KL divergence can be approximated empirically from the data, but in contrast to cross entropy, this requires estimating the probability densities of both distributions, p_I and p_Ω:

$$\begin{aligned}\widetilde{\mathfrak{D}_{\mathrm{KL}}}(I,\Omega) &= \frac{1}{|I|}\cdot\sum_{i\in I}\log\left(\frac{p_I(\mathfrak{X}_i)}{p_\Omega(\mathfrak{X}_i)}\right)\\ &= \frac{1}{|I|}\cdot\sum_{i\in I}\log\left(p_I(\mathfrak{X}_i)\right) - \log\left(p_\Omega(\mathfrak{X}_i)\right).\end{aligned} \tag{14}$$

When used in combination with the Gaussian distribution model, the KL divergence comes with an additional advantage from a computational point of view, since there is a known closed-form solution for the KL divergence of two Gaussians [18]:

$$\begin{aligned}\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega) = \frac{1}{2}\Bigg(&(\mu_\Omega - \mu_I)^\top S_\Omega^{-1}(\mu_\Omega - \mu_I)\\ &+ \mathrm{trace}\left(S_\Omega^{-1}S_I\right) + \log\frac{|S_\Omega|}{|S_I|} - D\Bigg).\end{aligned} \tag{15}$$

This allows evaluating the KL divergence in constant time for a given interval, which reduces the computational complexity of the MDI algorithm using the KL divergence in combination with Gaussian models to the number of possible intervals: $\mathcal{O}(N \cdot L)$.

Given this explicit solution for the KL divergence and the closed-form solution for the entropy of a Gaussian distribution [19] with mean vector μ and covariance matrix S, which is given by

$$\mathrm{H}(\mathcal{N}(\mu,S)) = \frac{1}{2}(\log|S| + d + d\cdot\log(2\pi)), \tag{16}$$

one can easily derive a closed-form solution for the cross entropy of those two distributions as well:

$$\begin{aligned}&\mathrm{H}(p_I,p_\Omega)\\ &= \mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega) + \mathrm{H}(p_I)\\ &= \frac{1}{2}\Bigg(\mathrm{trace}\left(S_\Omega^{-1}S_I\right) + \log|S_\Omega| + d\cdot\log(2\pi)\\ &\quad + (\mu_\Omega - \mu_I)^\top S_\Omega^{-1}(\mu_\Omega - \mu_I)\Bigg).\end{aligned} \tag{17}$$

Compared with the KL divergence, this does not assign extremely high scores to small intervals I with a low variance, due to the subtraction of $\log|S_I|$. This may be an explanation for the evaluation results in Section 3, where cross entropy in combination with Gaussian models is often superior to the KL divergence, although it does not account for intervals of varying entropy.

However, in contrast to the empirical approximation of cross entropy in (12), this requires the estimation of p_I.

2.5.3 Polarity of the KL Divergence and Its Effect on MDI

It is worth noting that the KL divergence is not a metric and, in particular, not symmetric: $\mathfrak{D}_{\mathrm{KL}}(p,q) \neq \mathfrak{D}_{\mathrm{KL}}(q,p)$. Some authors use, thus, a symmetric variant [8]:

$$\mathfrak{D}_{\mathrm{KL-SYM}}(p,q) = \frac{1}{2}\ \mathfrak{D}_{\mathrm{KL}}(p,q) + \frac{1}{2}\ \mathfrak{D}_{\mathrm{KL}}(q,p). \tag{18}$$

This raises the question whether $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$, $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$, or the symmetric version $\mathfrak{D}_{\mathrm{KL-SYM}}$ should be used for the detection of anomalous intervals. Quantitative experiments with an early prototype of our method [20] have shown that neither $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$ nor $\mathfrak{D}_{\mathrm{KL-SYM}}$ provide good performance, as opposed to $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$.

A visual inspection of the detections resulting from the use of $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$ with the assumption of Gaussian distributions shows that all the intervals with the highest anomaly scores have the minimum possible size specified by the user and a very low variance. An example is given in Fig. 4. The scores of the top detections in that example are around 100 times higher than those yielded by $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$.

This bias of $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$ towards small low-variance intervals can also be explained theoretically. For the sake of simplicity, consider the special case of a univariate time-series. In this case, the closed-form solution for $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$ assuming Gaussian distributions given in (15) reduces to

$$\frac{1}{2}\left(\frac{\sigma_\Omega^2}{\sigma_I^2} + \frac{(\mu_I-\mu_\Omega)^2}{\sigma_I^2} + \log\sigma_I^2 - \log\sigma_\Omega^2 - 1\right), \tag{19}$$

where μ_I, μ_Ω are the mean values and σ_I^2, σ_Ω^2 are the variances of the distributions in the inner and in the outer interval, respectively. It can be seen from (19) that, due to the division by σ_I^2, the KL divergence will approach infinity when the variance in the inner interval converges towards 0. And since the algorithm has to estimate the variance empirically from the given data, it assigns high detection scores to intervals as small as possible, because smaller intervals have a higher chance of having a low empirical variance. The term $\log\sigma_I^2$ cannot counterbalance this effect, though it is negative for $\sigma_I < 1$, since its absolute value grows much more slowly than that of σ_I^{-2}, as can be seen from the fact that $\forall_{\sigma_I<1}\left(-\log\sigma_I^2 = \log\sigma_I^{-2}\ <\ \sigma_I^{-2}\right)$, since $\forall_{\sigma_I<1}\left(\sigma_I^{-2}\ >\ 1\right)$.

In contrast, $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$, where the roles of I and Ω are swapped, does not possess this deficiency, since σ_Ω^2 is estimated from a much larger portion of data and, thus, is a more robust estimate.

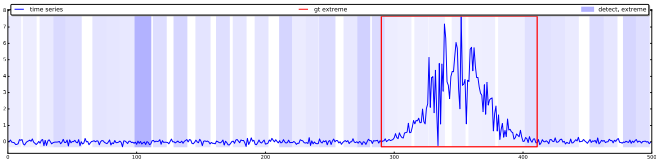


Fig. 4. Example for the bias of $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$ detections towards small intervals with low empirical variance on a synthetic time-series. The intensity of the fill color of the detected intervals corresponds to the detection scores. The ground-truth anomalous interval is indicated by a red box.

The symmetric version $\mathfrak{D}_{\mathrm{KL-SYM}}(p_I,p_\Omega)$ is useless as well, since the scores obtained from $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$ will just be absorbed by the much higher scores of $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$.

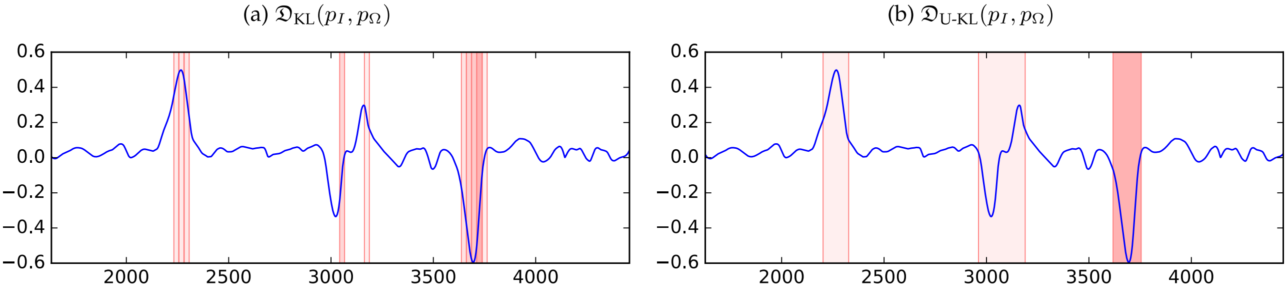
2.5.4 Statistical Analysis and Unbiased KL Divergence

Though $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$ does not overestimate the anomalousness of low-variance intervals as extremely as $\mathfrak{D}_{\mathrm{KL}}(p_\Omega,p_I)$ does, the following theoretical analysis will show that it is not unbiased either. In contrast to the previous section, this bias is not related to the data itself, but to the length of the intervals: smaller intervals systematically get higher scores than longer ones. This harms the quality of interval detections, because anomalies will be split up into multiple contiguous small detections (see Fig. 5a for an example).

Recall that $\mathfrak{I}_{m,m}^n$ denotes the set of all intervals of length m in a time-series with n time-steps. Furthermore, let $\vec{0}^d, d \in \mathbb{N}$, denote a d-dimensional vector with all coefficients being 0 and $\mathbb{I}_d$ the identity matrix of dimensionality d.

When applying the MDI algorithm to a time-series $(x_t)_{t=1}^n, x_t \sim \mathcal{N}(\vec{0}^d, \mathbb{I}_d)$, sampled independently and identically from plain white noise, an ideal divergence is supposed to yield constant average scores for all $\mathfrak{I}_{m,m}, m = a,\ldots,b$ (for some user-defined limits a,b), i.e., scores independent from the length of the intervals.

For simplicity, we will first analyze the distribution of those scores using the MDI algorithm with Gaussian distributions with the simple, but for this data perfectly valid assumption of identity covariance matrices. In this case, the KL divergence $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$ of two Gaussian distributions with the mean vectors $\mu_I, \mu_\Omega \in \mathbb{R}^d$ in some intervals $I \in \mathfrak{I}_m, \Omega = [1,n] \setminus I$ for some arbitrary m is given by $\frac{1}{2}\left\|\mu_\Omega - \mu_I\right\|^2$. Moreover, since all samples in the


(a) $\mathfrak{D}_{\mathrm{KL}}(p_I,p_\Omega)$

(b) $\mathfrak{D}_{\text{U-KL}}(p_I,p_\Omega)$

Fig. 5. (a) Top 10 detections obtained from the KL divergence on a real time-series and (b) top 3 detections obtained from the unbiased KL divergence on the same time-series. This example illustrates the phenomenon of several contiguous minimum-size detections when using the original KL divergence (note the thin lines between the single detections in the left plot). The MDI algorithm has been applied with a time-delay embedding of $\kappa = 3, \tau = 1$ and the size of the intervals to analyze has been limited to be between 25 and 250 samples.

time-series are normally distributed, so are their empirical means:

$$\mu_I = \frac{1}{m}\sum_{t\in I} x_t \sim \mathcal{N}(\vec{0}^d, m^{-1}\cdot \mathbb{I}_d)\ ,$$
$$\mu_\Omega = \frac{1}{n-m}\sum_{t\notin I} x_t \sim \mathcal{N}(\vec{0}^d, (n-m)^{-1}\cdot \mathbb{I}_d).$$

Thus, all dimensions of the mean vectors are independent and identically normally distributed variables. Their difference is, hence, normally distributed too:

$$\mu_\Omega - \mu_I \sim \mathcal{N}\left(\vec{0}^d, \left(\frac{1}{m}+\frac{1}{n-m}\right)\cdot \mathbb{I}_d\right).$$

Thus, $(\mu_\Omega - \mu_I)/\sqrt{\frac{1}{m}+\frac{1}{n-m}} \sim \mathcal{N}(\vec{0}^d, \mathbb{I}_d)$ is a vector of independent standard normal random variables and

$$\begin{aligned}&\mathfrak{D}_{\mathrm{KL}}(p_I, p_\Omega)\\ &= \frac{1}{2}\left(\frac{1}{m}+\frac{1}{n-m}\right)\sum_{i=1}^{d}\left(\frac{(\mu_\Omega-\mu_I)_i}{\sqrt{\frac{1}{m}+\frac{1}{n-m}}}\right)^2 \\ &\sim \frac{1}{2}\left(\frac{1}{m}+\frac{1}{n-m}\right)\cdot \chi^2_d\end{aligned} \tag{20}$$

is the sum of the squares of d independent normal variables and, hence, distributed according to the chi-squared distribution with d degrees of freedom, scaled by half the variance of the variables. The mean of a χ^2_d-distributed random variable is d and the mean of the $\mathfrak{D}_{\mathrm{KL}}(p_I, p_\Omega)$ scores for all intervals in $\mathfrak{I}_m$ is, accordingly, $\frac{d}{2}\left(\frac{1}{m}+\frac{1}{n-m}\right)$, which is inversely proportional to the length of the interval m. Thus, the KL divergence is systematically biased towards smaller intervals.

When the length n of the time-series is very large, the asymptotic scale of the chi-squared distribution is $\lim_{n\to\infty}\frac{1}{2}\left(\frac{1}{m}+\frac{1}{n-m}\right) = \frac{1}{2m}$ and the estimated parameters μ_Ω, S_Ω of the outer distribution converge towards the parameters of the true distribution of the data. Thus, if the restriction of the Gaussian model to identity covariance matrices is weakened to a global, shared covariance matrix S, the above findings also apply to the case of long time-series with correlated variables and, hence, also when time-delay embedding is applied. Because in this case, the KL divergence reduces to $\frac{1}{2}(\mu_I-\mu_\Omega)^\top S^{-1}(\mu_I-\mu_\Omega)$ and the subtraction of the true mean μ_Ω followed by the multiplication with the inverse covariance matrix can be considered as a normalization of the time-series, transforming it to standard normal variables with uncorrelated dimensions.

For the general case of two unrestricted Gaussian distributions, the test statistic

$$\begin{aligned}\lambda := {}& dm(\log(m)-1) + m(\mu_I-\mu_\Omega)^\top S_\Omega^{-1}(\mu_I-\mu_\Omega)\\ &+ \mathrm{trace}\left(mS_IS_\Omega^{-1}\right) - m\cdot\log\left|mS_IS_\Omega^{-1}\right|\end{aligned} \tag{21}$$

has been shown to be asymptotically distributed according to a chi-squared distribution with $d+\frac{d(d+1)}{2}$ degrees of freedom [21]. This test statistic is often used for testing the hypothesis that a given set of samples has been drawn from a Gaussian distribution with known parameters [22]. In the scenario of the MDI algorithm, the set of samples is the data in the inner interval I and the parameters of the distribution to test that data against are those estimated from the data in the outer interval Ω. The null hypothesis of the test would be that the data in I has been sampled from the same distribution as the data in Ω. The test statistic may then be used as a measure for how well the data in the interval I fit the model established based on the data in the remainder of the time-series.

After some elementary reformulations, the relationship between this test statistic λ and the KL divergence becomes obvious: $\lambda = 2m\cdot\mathfrak{D}_{\mathrm{KL}}(p_I, p_\Omega)$. This is exactly the normalization of the KL divergence by the scale factor identified in (20). Thus, we define an *unbiased KL divergence* as follows:

$$\mathfrak{D}_{\mathrm{U-KL}}(p_I, p_\Omega) := 2\cdot|I|\cdot\mathfrak{D}_{\mathrm{KL}}(p_I, p_\Omega). \tag{22}$$

The distribution of this divergence applied to asymptotically long time-series depends only on the number d of attributes and not on the length m of the interval any more. However, this correction may also be useful for time-series of finite length. An example of actual detections resulting from the use of the unbiased KL divergence compared with the original one can be seen in Fig. 5.

A further advantage of knowing the distribution of the scores is that this knowledge can also be used for normalizing the scores with respect to the number of attributes, in order to make them comparable across time-series of varying dimensionality. Moreover, it allows the selection of a threshold for distinguishing between anomalous and nominal intervals based on a chosen significance level. This may be preferred in some applications over searching for a fixed number of top k detections.

Interestingly, Jiang et al. [10] have derived an equivalent unbiased KL divergence $(m\cdot\mathfrak{D}_{\mathrm{KL}}(p_I, p_\Omega))$ from a different starting point based on the assumption of a Poisson distribution and the inverse log-likelihood of the interval as anomaly score.

2.5.5 Jensen-Shannon Divergence

A divergence measure that does not expose the problem of being asymmetric is the *Jensen-Shannon (JS) divergence*, which builds upon the KL divergence:

$$\mathfrak{D}_{\mathrm{JS}}(p,q) = \frac{1}{2}\ \mathfrak{D}_{\mathrm{KL}}\left(p, \frac{p+q}{2}\right) + \frac{1}{2}\ \mathfrak{D}_{\mathrm{KL}}\left(q, \frac{p+q}{2}\right). \tag{23}$$

where p and q are probability density functions. $\frac{p+q}{2}$ is a mixture distribution, so that a sample is drawn either from p or from q with equal probability (though a parametrized version of the JS divergence accounting for unequal prior probabilities exists as well, but will not be covered here).

The JS divergence possesses some desirable properties, which the KL divergence does not have: most notably, it is symmetric and bounded between 0 and $\log 2$ [23], so that anomaly scores cannot get infinitely high.

Like the KL divergence, the JS divergence can be approximated empirically from the data in the intervals I and Ω. However, there is no closed-form solution for the JS divergence under the assumption of a Gaussian distribution (as opposed to the KL divergence), since $\frac{p_I+p_\Omega}{2}$ would then

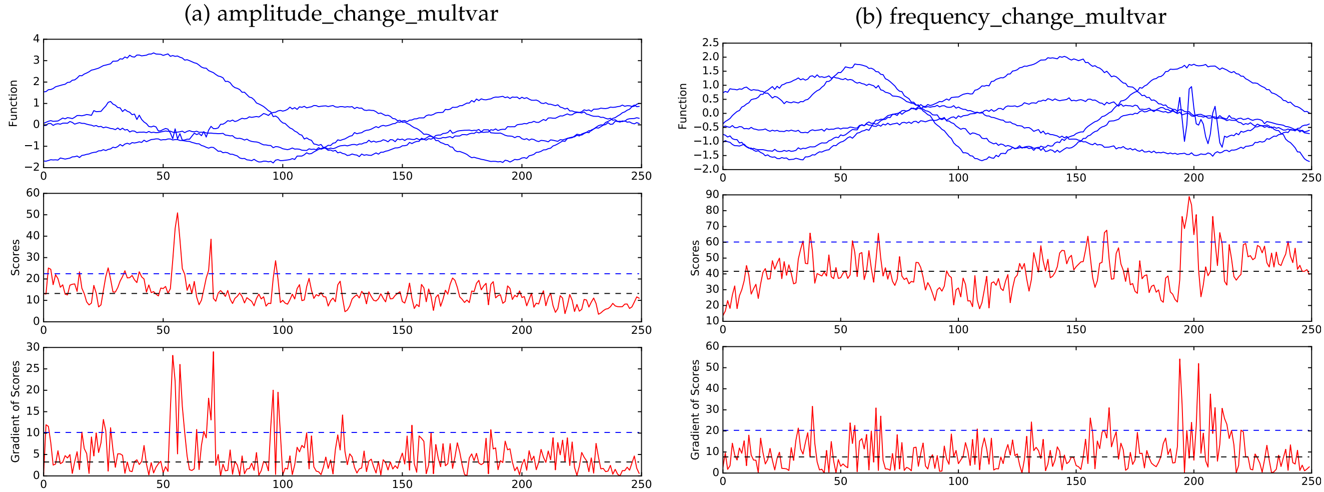


Fig. 6. Two exemplary synthetic time-series along with the corresponding Hotelling's T^2 scores and their gradients. The dashed black line indicates the mean of the scores and the dashed blue line marks a threshold that is 1.5 standard deviations above the mean. Time-delay embedding with $\kappa = 3, \tau = 1$ was applied before computing the scores.

be a Gaussian Mixture Model (GMM). Though several approximations of the KL divergence of GMMs have been proposed, they are either computationally expensive or abandon essential properties such as positivity [24]. This lack of a closed-form solution is likely to be the reason why the JS divergence was clearly outperformed by the KL divergence in our quantitative experiments in Section 3 when the Gaussian model is used, despite its desirable theoretic properties.

2.6 Interval Proposals for Large-Scale Data

Exploiting cumulative sums and a closed-form solution for the KL divergence, the asymptotic time complexity of the MDI algorithm with a Gaussian distribution model could already be reduced to be linear in the number of intervals (see Section 2.3.2). If the maximum length of an anomalous interval is independent from the number of samples N, the run-time is also linear in N. However, due to high constant-time requirements for estimating probability densities and computing the divergence, the algorithm is still too slow for processing large-scale data sets with millions of samples.

Since anomalies are rare by definition, many of the intervals analyzed by a full scan will be uninteresting and irrelevant for the list of the top anomalies detected by the algorithm. In order to focus on the analysis of non-trivial intervals, we employ a simple proposal technique that selects interesting intervals based on point-wise anomaly scores.

Simply grouping contiguous detections of point-wise anomaly detection methods in order to retrieve anomalous intervals is insufficient, because it will most likely lead to split-up detections. However, it is not unreasonable to assume that many samples inside of an anomalous interval will also have a high point-wise score, especially after applying contextual embedding. Fig. 6, for example, shows two exemplary time-series from the synthetic data set introduced in Section 3.1 along with the point-wise scores retrieved by applying the Hotelling's T^2 method [4], after time-delay embedding has been applied to the time-series. Note that even in the case of the very subtle amplitude-change anomaly, the two highest Hotelling's T^2 scores are at the beginning and the end of the anomaly. The idea is to apply a simple threshold operation on the point-wise scores to extract interesting points and then propose all those intervals for detailed scoring by a divergence measure whose first and last samples are among these points if the interval conforms to the size constraints.

This way, the probability density estimation and the computation of the divergence have to be performed for a comparatively small set of interesting intervals only and not for all possible intervals in the time-series. The interval proposal method is not required to have a low false-positive rate, though, because the divergence measure is responsible for the actual scoring. Instead, it has to act as a high-recall system so that truly anomalous intervals are not excluded from the actual analysis.

Since we are only interested in the beginning and end of the anomalies, the point-wise scores are not used directly, but the centralized gradient filter $[-1 \quad 0 \quad 1]$ is applied to the scores for reducing them in areas of constant anomalousness and emphasizing changes of the anomaly scores.

The evaluation in Section 3.3 will show that the interval proposal technique can speed-up the MDI algorithm significantly without impairing its performance.

3 Experimental Evaluation

In this section, we evaluate our MDI algorithm on a quantitative basis using synthetic data and compare it with other approaches well-known in the field of anomaly detection.

3.1 Data Set

In contrast to many other established machine learning tasks, there is no widely used standard benchmark for the evaluation of anomaly detection algorithms; not for the detection of anomalous intervals and not even for the very common task of point-wise anomaly detection. This is mainly for the reason that the notion of an "anomaly" is not well defined and varies between different applications and even from analyst to analyst. Moreover, anomalies are, by definition, rare, which makes the collection of large-scale data sets difficult. However, even if a large amount of data were available, it would

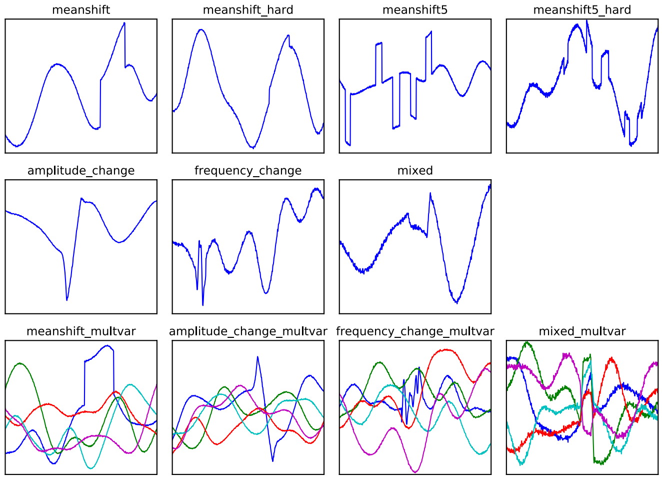


Fig. 7. Examples from the synthetic test data set.

be nearly impossible to annotate it in an intersubjective way everyone would agree with. But accurate and complete ground-truth information is mandatory for a quantitative evaluation and comparison of machine learning techniques. Therefore, we use a synthetic data set for assessing the performance of different variants of the MDI algorithm.

All time-series in that data set have been sampled from a Gaussian process $\mathcal{GP}(m, K)$ with a squared-exponential covariance function $K(x_t, x_{t'}) = (2\pi\ell^2)^{-1/2} \cdot \exp\left(-\frac{\|x_t - x_{t'}\|^2}{2\ell^2}\right) + \sigma^2 \cdot \delta(t, t')$ and zero mean function $m(x) = 0$. The *length scale* of the GP has been set to $\ell^2 = 0.01$ and the noise parameter to $\sigma^2 = 0.001$. $\delta(t, t')$ denotes Kronecker's delta. Different types of anomalies have then been injected into these time-series, with a size varying between 5 percent and 20 percent of the length of the time-series:

meanshift: A random, but constant value $\gamma \in [3, 4]$ is added to or subtracted from the anomalous samples.

meanshift_hard: A random, but constant value $\gamma \in [0.5, 1]$ is added to or subtracted from the anomalous samples.

meanshift5: Five `meanshift` anomalies are inserted into the time-series.

meanshift5_hard: Five `meanshift_hard` anomalies inserted into the time-series.

amplitude_change: The time-series is multiplied with a Gaussian window with standard deviation $L/4$ whose mean is the centre of the anomalous interval. Here, L is the length of the anomalous interval and the amplitude of the Gaussian window is clipped at 2.0. This modified time-series is added to the original one.

frequency_change: The time-series is sampled from a non-stationary GP, whose covariance function $K(x_t, x_{t'}) = (\ell^2(t) \cdot \ell^2(t'))^{1/4} \cdot \left(\frac{\ell^2(t)+\ell^2(t')}{2}\right)^{-1/2} \cdot \exp\left(-\frac{\|x_t - x_{t'}\|^2}{\ell^2(t)+\ell^2(t')}\right) + \sigma \cdot \delta(t, t')$ uses a reduced length scale $\ell^2(t) = \begin{cases} 10^{-2} & \text{if } t \notin [a, b), \\ 10^{-4} & \text{if } t \in [a, b) \end{cases}$ during the anomalous interval $I = [a, b)$, so that correlations between samples are reduced, which leads to more frequent oscillations [25].

mixed: The values in the anomalous interval are replaced with the values of another function sampled from the Gaussian process. 10 time-steps at the borders of the anomaly are interpolated between the two functions for a smooth transition. This rather difficult test case is supposed to reflect the concept of anomalies as being "generated by a different mechanism" (cf. Section 2.2).

The above test cases are all univariate, but there are as well similar multivariate scenarios `meanshift_multvar`, `amplitude_change_multvar`, `frequency_change_multvar`, and `mixed_multvar` with 5-dimensional time-series. Regarding the first three of these test cases, the corresponding anomaly is injected into one of the dimensions, while all attributes are replaced with those of the other time-series in the `mixed_multvar` scenario, which is also a property of many real time-series.

This results in a synthetic test data set with 11 test cases, a total of 1100 time-series and an overall number of 1900 anomalies. Examples for all test cases are shown in Fig. 7.

3.2 Performance Comparison

Since the detection of anomalous regions in spatio-temporal data is rather a *detection* than a *classification* task, we do not use the *Area under the ROC Curve (AUC)* as performance criterion like many works on point-wise anomaly detection do, but quantify the performance in terms of *Average Precision (AP)* with an Intersection over Union (IoU) criterion that allows an overlap between 50 and 100 percent.

Hotelling's T^2 [4] and Robust Kernel Density Estimation (RKDE) [3] are used as baselines for the comparison. For RKDE, a Gaussian kernel with a standard deviation of 1.0 and the Hampel loss function are used. We obtain interval detections from those point-wise baselines by grouping contiguous detections based on multiple thresholds and applying non-maximum suppression afterwards. The overlap threshold for non-maximum suppression is set to 0 in all experiments to obtain non-overlapping intervals only. To be fair, MDI also has to compete with the baselines on the task they have been designed for, i.e., point-wise anomaly detection, by means of AUC. The interval detections can be converted to point-wise detections easily by taking the score of the interval a sample belongs to as score for that sample.

Fig. 8 shows that the performance of the MDI algorithm using the Gaussian model is clearly superior on the entire synthetic data set compared to the baselines by means of Mean AP and even on the task of point-wise anomaly detection measured by AUC. The $\mathfrak{D}_{\mathrm{KL}}(p_I, p_\Omega)$ polarity of the KL divergence has been used in all experiments following the argumentation in Section 2.5.3. In addition, the performance of the unbiased variant $\mathfrak{D}_{\mathrm{U-KL}}(p_I, p_\Omega)$ is reported for the Gaussian model. The parameters of time-delay embedding have been fixed to $\kappa = 6, \tau = 2$ which we have empirically found to be suitable for this data set. For KDE, we used a Gaussian kernel with bandwidth 1.0.

While MDI KDE is already superior to the baselines, it is significantly outperformed by MDI Gaussian, which improves on the best baseline by 286 percent. This discrepancy between the MDI algorithm using KDE and using Gaussian models is mainly due to time-delay embedding, which is particularly useful for the Gaussian model, because it takes correlations of the variables into account, as opposed to KDE. As can be seen in Fig. 9, the Gaussian model would be worse than KDE and on par with the baselines without time-delay embedding.

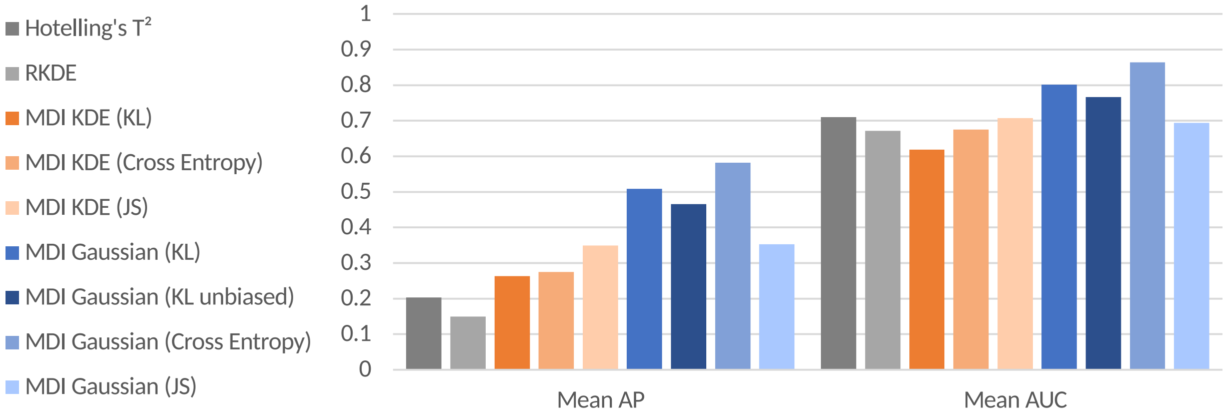


Fig. 8. Performance comparison of different variants of the MDI algorithm and the baselines on the synthetic data set.

Considering the Mean AP on this synthetic data set, the unbiased KL divergence did not perform better than the original KL divergence. However, on the test cases `meanshift5`, `meanshift5_hard`, and `meanshift_hard` it achieved an AP twice as high as that of $\mathfrak{D}_{\text{KL}}(p_I, p_\Omega)$, which was poor on those data sets (see Fig. 10). Since real data sets are also likely to contain multiple anomalies, we expect $\mathfrak{D}_{\text{U-KL}}$ to be a more reliable divergence measure in practice.

Another interesting result is that cross entropy was the best performing divergence measure. This shows the advantage of reducing the impact of the inner distribution p_I, which is estimated from very few samples. However, it may perform less reliably on real data whose entropy varies more widely over time than in this synthetic benchmark.

The Jensen-Shannon divergence performed best for the KDE method, but worst for the Gaussian model. This can be explained by the lack of a closed-form solution for the JS divergence, so that it has to be approximated from the data, while the KL divergence of two Gaussians can be computed exactly. This advantage of the combination of the KL divergence with Gaussians models is, thus, not only beneficial with respect to the run-time of the algorithm, but also with respect to its detection performance.

The differences between the results in Fig. 8 are significant on a level of 5 percent according to the permutation test.

3.3 Interval Proposals

In order not to sacrifice detection performance for the sake of speed, the interval proposal method described in Section 2.6 has to act as a high-recall system proposing the majority of anomalous intervals. This can be controlled to some degree by adjusting the threshold $\theta = \mu + \vartheta \cdot \sigma$ applied to the point-wise scores, where μ and σ are the empirical mean and standard deviation of the point-wise scores, respectively. To find a suitable value for the hyperparameter ϑ, we have evaluated the recall of the proposed intervals for different values of $\vartheta \in [0, 4]$ using the usual IoU measure for distinguishing between true and false positive detections. The results in Fig. 11a show that time-delay embedding is of a great benefit in this scenario too. Based on these results, we selected $\vartheta = 1.5$ for subsequent experiments, which still provides a recall of 97 percent and is already able to reduce the number of intervals to be analyzed in detail significantly.

The processing of all the 1100 time-series from the synthetic data set, which took 216 seconds on an Intel Core$^{\text{TM}}$ i7-3930K with 3.20 GHz and eight virtual cores using the Gaussian model and the unbiased KL divergence after the usual time-delay embedding with $\kappa = 6, \tau = 2$, could be reduced to 5.2 seconds using interval proposals. This corresponds to a speed-up by more than 40 times.

Though impressive, the speed-up was expected. What was not expected, however, is that the use of interval proposals also increased the detection performance of the entire algorithm by up to 125 percent, depending on the divergence. The exact average precision achieved by the algorithm on the synthetic data set with a full scan over all intervals and with interval proposals is shown in Fig. 11b. This improvement is also reflected by the AUC scores not reported here and is, hence, not specific to the evaluation criterion. A possible explanation for this phenomenon is that some intervals that are uninteresting but distracting for the MDI algorithm are not even proposed for detailed analysis.

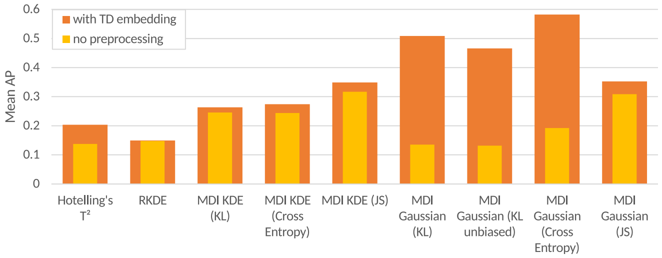


Fig. 9. Effect of time-delay embedding with $\kappa = 6, \tau = 2$ on the performance of the MDI algorithm and the baselines on the synthetic data set.

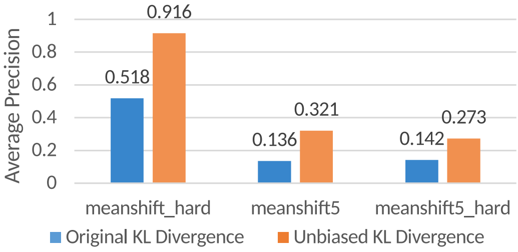


Fig. 10. Performance of the original and the unbiased KL divergence on test cases with multiple or subtle anomalies.

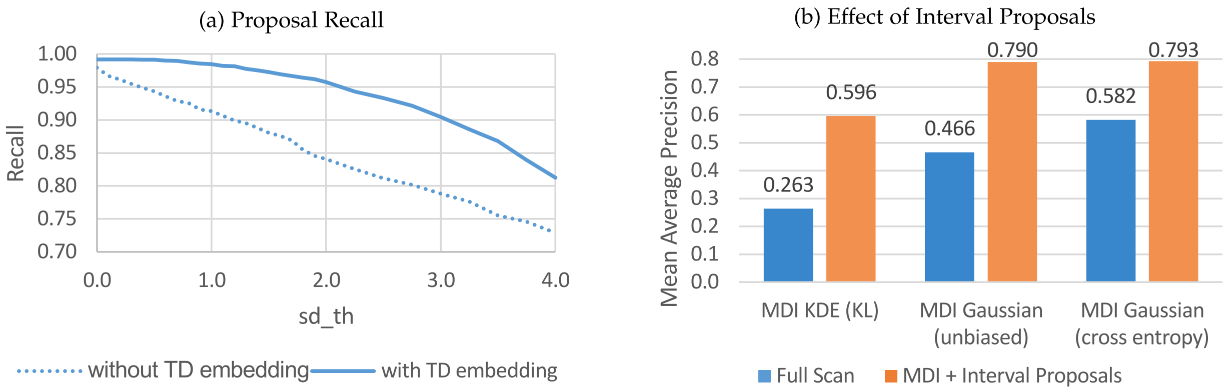


Fig. 11. (a) Recall of interval proposals without time-delay embedding and with $\kappa = 6, \tau = 2$ on the synthetic data set for different proposal thresholds. (b) Effect of interval proposals on the Mean Average Precision of different variants of the MDI algorithm on the synthetic data set.

4 Application Examples on Real Data

The following application examples on real data from various different domains are intended to complement the quantitative results presented above with a demonstration of the feasibility of our approach for real problems.

4.1 Detection of North Sea Storms

To demonstrate the efficiency of the MDI algorithm on long time-series, we apply it to storm detection in climate data: The coastDat-1 hindcast [26] is a reconstruction of various marine climate variables measured at several locations over the southern North Sea between 51° N, 3° W and 56° N, 10.5° E with an hourly resolution over the 50 years from 1958 to 2007, i.e., approximately 450,000 time steps. Since measurements are not available at locations over land, we select the subset of the data between 53.9° N, 0° E and 56° N, 7.7° E, which results in a regular spatial grid of size 78×43 located entirely over the sea (cf. Fig. 12). Because cyclones and other storms usually have a large spatial extent and move over the region covered by the measurements, we reduce the spatio-temporal data to purely temporal data in this experiment by averaging over all spatial locations. The variables used for this experiment are significant wave height, mean wave period and wind speed.

We apply the MDI algorithm to that data set using the Gaussian model and the unbiased KL divergence. Since North Sea storms lasting longer than 3 days are usually considered two independent storms, the maximum length of the possible intervals is set to 72 hours, while the minimum length is set to 12 hours. The parameters of time-delay embedding are fixed to $\kappa = 3, \tau = 1$.

28 out of the top 50 and 7 out of the top 10 detections returned by the algorithm can be associated with well-known historic storms. The highest scoring detection is the so-called "Hamburg-Flut" which flooded one fifth of Hamburg in February 1962 and caused 340 deaths. Also among the top 5 is the "North Frisian Flood", which was a severe surge in November 1981 and lead to several dike breaches in Denmark.

A visual inspection of the remaining 22 detections revealed, that almost all of them are North Sea storms as well. Only 4 of them are not storms, but the opposite: they span times of extremely calm sea conditions with nearly no wind and very low waves, which is some kind of anomaly as well.

A list of the top 50 detections and animated heatmaps of the three variables during the detected time-frames can be found in the supplemental material, available online and at: http://www.inf-cv.uni-jena.de/libmaxdiv_applications.html.

Processing this comparatively long time-series using 8 parallel threads took 27 seconds. This time can be reduced further to half a second by using interval proposals without changing the top 10 detections significantly. This supports the assumption, that the novel proposal method does not only perform well on synthetic, but also on real data.

4.2 Spatio-Temporal Detection of Low Pressure Areas

As a genuine spatio-temporal use-case, we have also applied the MDI algorithm to a time-series with daily sea-level pressure (SLP) measurements over the North Atlantic Sea with a much wider spatial coverage than in the previous experiment. For this purpose, we selected a subset of the NCEP/NCAR reanalysis [27] covering the years from 1957 to 2011. This results in a time-series of about 20,000 days. The spatial resolution of 2.5° is rather coarse and the locations are organized in a regular grid of size 28×17 covering the area between 25° N, 52.5° W and 65° N, 15° E.

Again, the MDI algorithm with the Gaussian model and the unbiased KL divergence is applied to this time-series to detect low-pressure fields, which are related to storms. Regarding the time dimension, we apply time-delay embedding with $\kappa = 3, \tau = 1$ and search for intervals of size between 3 and 10 days. Concerning space, we do not apply any embedding for now and set a minimum size of

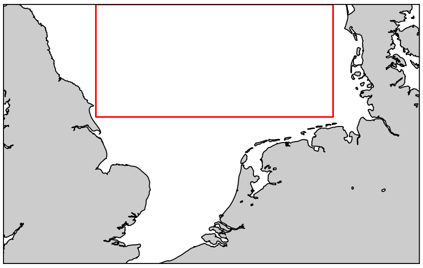


Fig. 12. Map of the area covered by the coastDat dataset. The highlighted box denotes the area from which data have been aggregated for our experiment.

$7.5^\circ \times 7.5^\circ$, but no maximum. 7 out of the top 20 detections could be associated with known historic storms.

A visual inspection of the results shows that the MDI algorithm is not only capable of detecting occurrences of anomalous low-pressure fields over time, but also their spatial location. This can be seen in the animations in the supplemental material, available online or on our web page: http://www.inf-cv.uni-jena.de/libmaxdiv_applications.html.

It is not necessary to apply spatial-neighbor embedding in this scenario, since we are not interested in spatial outliers, but only in the location of temporal outliers. We have also experimented with applying spatial-neighbor embedding and it led to the detection of some high-pressure fields surrounded by low-pressure fields. Since high-pressure fields are both larger and more common in this time-series, they are not detected as temporal anomalies.

Since we did not set a maximum spatial extent of anomalous regions, the algorithm took 4 hours to process this spatio-temporal time-series. This could, however, be reduced to 22 seconds using our interval proposal technique, with only a minor loss of localization accuracy.

4.3 Stylistic Anomalies in Texts of Natural Language

By employing a transformation from the domain of natural language to real-valued features, the MDI algorithm can also be applied to written texts. One important task in Natural Language Processing (NLP) is, for example, the identification of paragraphs written in a different language than the remainder of the document. Such a segmentation can be used as a pre-processing step for the actual, language-specific processing.

In order to simulate such a scenario, we use a subset of the *Europarl* corpus [28], which is a sentence-aligned parallel corpus extracted from the proceedings of the European Parliament in 21 different languages. The 33,334 English sentences from the *COMTRANS* subset of *Europarl*, which is bundled with the Natural Language Toolkit (NLTK) for Python, serve as a basis and 5 random sequences of between 10 and 50 sentences are replaced by their German counterparts to create a semantically coherent mixed-language text.

We employ a simple transformation of sentences to feature vectors: Since the distribution of letter frequencies varies across languages, each sentence is represented by a 27-dimensional vector whose first element is the average word length in the sentence and the remaining 26 components are the absolute frequencies of the letters “a” to “z” (case-insensitive). German umlauts are ignored since they would make the identification of German sentences too easy.

The MDI algorithm using the unbiased KL divergence is then applied in order to search for anomalous sequences of between 10 and 50 sentences in the mixed-language text after sentence-wise transformation to the feature space. Because the number of features is quite high in relation to the number of samples in an interval, we use a global covariance matrix shared among the Gaussian models and do not apply time-delay embedding.

The top 5 detections returned by the algorithm correspond to the 5 German paragraphs that have been injected into the English text. The localization is quite accurate, though not perfect: on average, the boundaries of the detected paragraphs are off by 1.4 sentences from the ground-truth. The next 5 detections are mainly tables and enumerations, which are also an anomaly compared with the usual dialog style of the parliament proceedings.

For this scenario, we had designed the features specifically for the task of language identification. To see what else would be possible with a smaller bias towards a specific application, we have also applied the algorithm to the 1st Book of Moses (Genesis) in the King James Version of the bible, where we use `word2vec` [29] for word-wise feature embeddings. `word2vec` learns real-valued vector representations of words in a way, so that the representations of words that occur more often in similar contexts have a smaller euclidean distance. The embeddings used for this experiment have been learned from the Brown corpus using the continuous skip-gram model and we have chosen a dimensionality of 50 for the vector space, which is rather low for `word2vec` models, but still tractable for the Gaussian probability density model. Words which have not been seen by the model during training are treated as missing values.

The top 10 detections of sequences of between 50 and 500 words according to the unbiased KL divergence are provided in the supplemental material, available online. The first five of those are, without exception, genealogies, which can indeed be considered as anomalies, because they are long lists of names of fathers, sons and wives, connected by repeating phrases. The 6th detection is a dialog between God and Abraham, where Abraham bargains with God and tries to convince him not to destroy the town Sodom. This episode is another example for stylistic anomalies, since the dialog is a concatenation of very similar question-answer pairs with only slight modifications.

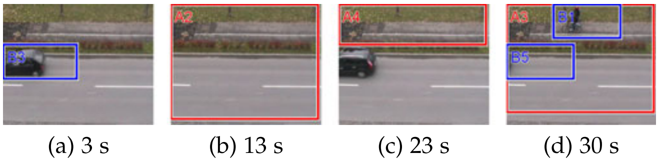
Due to the rather wide limits on the possible size of anomalous intervals, the analysis of the entire book Genesis, a sequence of 44,764 words, took a total of 9 minutes, where we have not yet used interval proposals.

4.4 Anomalies in Videos

The detection of unusual events in videos is another important task, e.g., in the domain of video surveillance or industrial control systems. Though videos are already represented as multivariate spatio-temporal time-series with usually 3 variables (RGB channels), a semantically more meaningful representation can be obtained by extracting features from a Convolutional Neural Network (CNN).

In this experiment, we use a video of a traffic scene from the ViSOR repository [30]. It has a length of 60 seconds (1495 frames) and a rather low resolution of 360×288 pixels. The video shows a street and a side-walk with a varying frequency of cars crossing the captured area horizontally in both directions. At one point, a group of two pedestrians and one cyclist appears on the side-walk and crosses the area from right to left at a low speed. Another sequence at the end of the video shows a single cyclist riding along the side-walk in the opposite direction at a higher speed. Altogether, 26 seconds of the video contain moving objects and 34 seconds just show an empty street. The nominal state of the scene hence is not unambiguous.

We extract features for each frame of the video from the `conv5` layer of CaffeNet [31], which reduces the spatial resolution to 22×17, but increases the number of feature



(a) 3 s (b) 13 s (c) 23 s (d) 30 s

Fig. 13. Snapshots from the example video with corresponding detections. Regions detected using the unbiased KL divergence start with the character “A”, those detected by cross entropy start with “B”. The full video can be found in the supplemental material, available online or on our web page: http://www.inf-cv.uni-jena.de/libmaxdiv_applications.html.

dimensions to 256. This rather large feature space is then reduced to 16 dimensions using PCA and the MDI algorithm is applied to search for anomalous sub-blocks with a minimum spatial extent of 10×5 cells and a length between 3 and 12 seconds. The time-delay embedding parameters are fixed to $\kappa = 3, \tau = 4$ for capturing half a second as context without increasing the number of dimensions too much. We apply the MDI algorithm with both the unbiased KL divergence and cross entropy as divergence measures. The Gaussian distribution model is employed in both cases.

The results (some snapshots are shown in Fig. 13) exhibit an interesting difference between the two divergence measures: The KL divergence detects a sub-sequence of approximately 10 seconds where absolutely no objects cross the captured area. Thus, car traffic is identified as normal behavior and long spans of time without any traffic are considered as anomalous, because they have a very low entropy and the KL divergence penalizes the entropy of all other intervals, as opposed to cross entropy which does not take the entropy of the detected interval into account. Another detection occurs when the group of pedestrians enters the area. The localization, however, is rather fuzzy and spans nearly the entire frame. Cross entropy, on the other hand, seems to identify the state of low or no traffic as normal behavior and yields two detections at the beginning and the end of the video where the frequency of cars is higher than in the rest of the video. It detects the pedestrians too, but with a better localization accuracy. This detection, however, does not cover the entire side-walk, since the pedestrians are moving from right to left and the algorithm is not designed for tracking moving anomalies.

Without using interval proposals, the comparatively high number of features combined with the large spatial search space would result in a processing time of 13 hours for this video. This can be reduced to 5 minutes using our novel interval proposal technique.

5 Summary and Conclusions

We have introduced a novel unsupervised algorithm for anomaly detection that is suitable for analyzing large multivariate time-series and can detect anomalous *regions* not only in temporal but also in spatio-temporal data from various domains. The proposed MDI algorithm outperforms existing anomaly detection techniques, while being comparatively time efficient, thanks to an efficient implementation and a novel interval proposal technique that excludes uninteresting parts of the data from in-depth analysis. Moreover, we have exposed a bias of the Kullback-Leibler (KL) divergence towards smaller intervals and proposed an unbiased KL divergence that is superior when applied to real data. We have also investigated other divergence measures and found that the use of cross entropy can result in improved performance for data with a low variability of entropy.

Various experiments on data from different domains, including climate analysis, natural language processing and video surveillance, have shown that the algorithm proposed in this work can serve as a generic, unsupervised anomaly detection technique that can facilitate tasks such as process control, data analysis and knowledge discovery. These application examples emphasize the importance of interval-based anomaly detection techniques, and we hope that our work is able to motivate further research in this area.

For processing data with a large spatial extent or a high number of dimensions, a full scan over all possible sub-blocks of the data would be prohibitively time-consuming. To this end, we have introduced a novel interval proposal technique that can reduce computation time significantly. However, interval proposals usually lead to less accurate detections, which is particularly noticeable with regard to spatial dimensions. Future work might hence investigate applying in-depth analysis not only to the proposed intervals themselves, but also to their neighborhood. An alternative might be a hierarchical approach of successive refinement.

Other open problems to be addressed in the future include efficient probability density estimation in the face of high-dimensional data, the automatic determination of suitable parameters for time-delay embedding, and tracking anomalies moving in space over time. Furthermore, it is often necessary to convince the expert analyst that a detected anomaly really is an anomaly. Thus, future work will include the development of an attribution scheme that can explain which variables or combinations of variables caused a detection and why.

Acknowledgments

The support of the project EU H2020-EO-2014 project BACI “Detecting changes in essential ecosystem and biodiversity properties-towards a Biosphere Atmosphere Change Index”, contract 640176, is gratefully acknowledged.

References

[1] G. Walker, “World weather,” *Quarterly J. Royal Meteorological Society*, vol. 54, no. 226, pp. 79–87, 1928.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, vol. 29, no. 2, 2000, pp. 93–104.

[3] J. Kim and C. D. Scott, “Robust kernel density estimation,” *J. Mach. Learn. Res.*, vol. 13, pp. 2529–2565, 2012.

[4] J. F. MacGregor and T. Kourti, “Statistical process control of multivariate processes,” *Control Eng. Practice*, vol. 3, no. 3, pp. 403–414, 1995.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[6] E. Keogh, J. Lin, and A. Fu, “Hot sax: Efficiently finding the most unusual time series subsequence,” in *Proc. IEEE Int. Conf. Data Mining*, 2005, Art. no. 8.

[7] H. Ren, M. Liu, X. Liao, L. Liang, Z. Ye, and Z. Li, “Anomaly detection in time series based on interval sets,” *IEEJ Trans. Electr. Electron. Eng.*, vol. 13, pp. 757–762, 2018.

[8] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, “Change-point detection in time-series data by relative density-ratio estimation,” *Neural Netw.*, vol. 43, pp. 72–83, 2013.

[9] P. Senin, J. Lin, X. Wang, T. Oates, S. Gandhi, A. P. Boedihardjo, C. Chen, and S. Frankenstein, "Grammarviz 3.0: Interactive discovery of variable-length time series patterns," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 1, 2018, Art. no. 10.

[10] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "A general suspiciousness metric for dense blocks in multimodal data," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 781–786.

[11] E. Wu, W. Liu, and S. Chawla, "Spatio-temporal outlier detection in precipitation data," in *Knowledge Discovery from Sensor Data*. Berlin, Germany: Springer, 2010, pp. 115–133.

[12] D. M. Hawkins, *Identification of Outliers*. Berlin, Germany: Springer, 1980, vol. 11.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[14] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Physical Rev. Lett.*, vol. 45, no. 9, 1980, Art. no. 712.

[15] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*. Berlin, Germany: Springer, 1981, pp. 366–381.

[16] A. Kut and D. Birant, "Spatio-temporal outlier detection in large databases," *CIT. J. Comput. Inf. Technol.*, vol. 14, no. 4, pp. 291–297, 2006.

[17] T. Cheng and Z. Li, "A multiscale approach for spatio-temporal outlier detection," *Trans. GIS*, vol. 10, no. 2, pp. 253–263, 2006.

[18] J. Duchi, "Derivations for linear algebra and optimization," Berkeley, California, 2007, https://web.stanford.edu/~jduchi/projects/general_notes.pdf

[19] N. A. Ahmed and D. Gokhale, "Entropy expressions and their estimators for multivariate distributions," *IEEE Trans. Inf. Theory*, vol. 35, no. 3, pp. 688–692, May 1989.

[20] E. Rodner, B. Barz, Y. Guanche, M. Flach, M. Mahecha, P. Bodesheim, M. Reichstein, and J. Denzler, "Maximally divergent intervals for anomaly detection," in *Proc. ICML Workshop Anomaly Detection*, 2016, pp. 1–5.

[21] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 1962.

[22] T. Kanungo and R. M. Haralick, "Multivariate hypothesis testing for Gaussian data: Theory and software," Tech. Rep. ISL-TR-95-05, 1995.

[23] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[24] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, vol. 4, Art. no. IV–317.

[25] C. Paciorek and M. Schervish, "Nonstationary covariance functions for Gaussian process regression," *Advances Neural Inf. Process. Syst.*, vol. 16, pp. 273–280, 2004.

[26] *Helmholtz-Zentrum Geesthacht, Zentrum für Material- und Küstenforschung GmbH. (2012) coastdat-1 waves north sea wave spectra hindcast (1948–2007)*. World Data Center for Climate (WDCC). (2012). [Online]. Available: https://doi.org/10.1594/WDCC/coastDat-1_Waves

[27] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al., "The ncep/ncar 40-year reanalysis project," *Bulletin Amer. Meteorological Society*, vol. 77, no. 3, pp. 437–471, 1996.

[28] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. Conf. Mach. Trans. Summit*, pp. 79–86, vol. 5, 2005.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781 2013.

[30] R. Vezzani and R. Cucchiara, "Video surveillance online repository (visor): An integrated framework," *Multimedia Tools Appl.*, vol. 50, no. 2, pp. 359–380, 2010. [Online]. Available: http://www.openvisor.org/video_details.asp?idvideo=339

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM int. conf. Multimedia*, 2014, pp. 675–678.

Björn Barz received the BSc and MSc degrees in computer science with honours from Friedrich Schiller University Jena, Germany, in 2014 and 2016, respectively. He is currently working towards the PhD degree at the Computer Vision Group of Joachim Denzler, University of Jena. His research interests include the field of machine learning, visual object detection, content-based image retrieval, and natural language processing.

Erik Rodner received the diploma (Hons.) degree in computer science from the Friedrich Schiller University Jena, Germany, in 2007, and the PhD degree with summa cum laude for his work on learning with few examples, which was done under supervision of Joachim Denzler from the Computer Vision Group, University of Jena, in 2011. From 2012 to 2013, he joined UC Berkeley as a postdoctoral researcher. He was senior researcher and lecturer with the Computer Vision Group, University of Jena from 2013 to 2016 and is now researcher at Carl Zeiss AG. His research interests include domain adaptation, deep learning, visual object discovery, active and continuous learning, and scene understanding.

Yanira Guanche Garcia received the MSc degree in coastal and ports engineering, in 2010, and the PhD degree from Universidad de Cantabria, Spain, in 2013. From 2014 to 2015, Yanira joined IFREMER and BRGM in France as a postdoctoral researcher. Since 2015, she is a postdoctoral researcher with the computer vision group of Joachim Denzler, Friedrich Schiller University, Jena, and research coordinator of the Michael Stifel Center for Data-Driven and Simulation Science, Jena.

Joachim Denzler received the "Diplom-Informatiker", "Dr.-Ing." and "Habilitation" degrees from the University of Erlangen, Germany, in 1992, 1997, and 2003, respectively. Currently, he holds a position as full professor for computer science and is head of the Computer Vision Group with the Friedrich Schiller University Jena, Germany. He is also director of the Michael Stifel Center for Data-Driven and Simulation Science, Jena. His research interests comprise the automatic analysis, fusion, and understanding of sensor data, especially development of methods for visual recognition tasks and dynamic scene analysis. He contributed in the area of active vision, 3D reconstruction, as well as object recognition and tracking. He is author and co-author of more than 300 journal and conference papers as well as technical articles. He is a member of the IEEE, IEEE Computer Society, DAGM, and GI.