Outlier detection for multivariate time series

Jorge L. Serras
jorge.serras@tecnico.ulisboa.pt

Alexandra M. Carvalho
alexandra.carvalho@tecnico.ulisboa.pt

Susana Vinga
susanavinga@tecnico.ulisboa.pt

***Abstract*—Outliers can be defined as observations which are suspected of not have been generated by data's underlying processes. Many applications require a way of identifying interesting or unusual patterns in multivariate time series (MTS), however, most outlier detection methods focus solely on univariate series, providing analysts with strenuous solutions. We propose a complete outlier detection system covering problems since pre-processing that adopts a dynamic Bayesian network modeling algorithm. The latter encodes optimal inter and intra time-slice connectivity of transition networks capable of capturing conditional dependencies in MTS datasets. A sliding window mechanism is employed to gradually score each MTS transition given the model. Simultaneously, complete MTS are evaluated. Two score-analysis strategies are studied to assure an automatic boundary classifying anomalous data. The proposed approach is first validated through simulated data, demonstrating the system's performance. Comparison with an assembled probabilistic suffix tree method is available displaying the leverage of the proposed multivariate approach. Further experimentation is made on real data, by uncovering anomalies in distinct scenarios such as electrocardiogram series, mortality rate data and written pen digits. The developed system proved beneficial in capturing unusual data resulting from temporal contexts, being suitable for any MTS scenario. A widely accessible web application employing the complete system is made available jointly with a tutorial.**

***Index Terms*—multivariate time series, outlier detection, dynamic Bayesian networks, sliding window algorithm, score analysis, web application.**

I. Introduction

In recent times, the machine learning community has boomed coupled with the always expanding desire to acquire maximum benefit from collected data, apparent in sectors like biomedicine, socio-economics and industry. In the current study, an outlier is described as a data element or segment which there is no explanation for it to stand out. Hence, being suspected of not have been generated by the data's underlying processes. Outliers have the capability to mislead analysts to altogether different insights. However, their discovery is crucial in acquiring a better understanding of the data's underlying nature leading to the development of more efficient methods. *Without deviation from the norm, progress is not possible* [1].

Longitudinal data, also known as multivariate time series (MTS) is defined as a set of observations measured along time. Each observation represents a collection of variables which the combined evolution over time is object of analysis. Open-mindedly, an outlier detection system is assembled from scratch with the aim of providing an intuitive and effective alternative for discovering abnormal entities among real-world MTS datasets. Such are typically not found in existing literature, being the latter only concerned with univariate time series (TS) [2, 3]. Even with a sophisticated univariate system, many contextual outliers can not be disclosed without contemplating inter-variable relations, limiting thus univariate strategies.

The approach proposed in the current thesis resides in the statistical paradigm, providing a probabilistic graphical model representing a normality standard. MTS datasets are scored according to a sliding window mechanism capable of capturing compelling patterns encoded by temporal dependencies amidst variables, absent in existing literature.

Temporal dependencies within and between discrete variables can be modeled using dynamic Bayesian networks (DBN) which extend traditional Bayesian networks to temporal processes. These are graphical statistical methods capable of encoding conditional relationships of complex MTS structures. A modeling technique known as tree-augmented DBN (tDBN) [4] is used to provide a network possessing optimum inter/intra-slice connectivities for each transition network, verified to outperform existing literature. Both stationary and non-stationary DBNs are studied. The model provides a normality standard for anomaly detection.

The proposed approach measures data's level of discrepancy with respect to the model. Such is accomplished by scoring each observation using stored conditional probabilities. A dataset is comprised by a set of MTS also known as subjects. A transition is regarded as a subset of a subject which is comprised by the observations responsible for a time-stamp, thus possessing the observations of the given time slice including previous time-stamps. These are known as windows and have an associated transition network depending on the parameters of the model. A sliding window mechanism is proposed to uncover transition scores in MTS datasets. The likelihood of each transition is computed. Low scores depict time instances which are not explained by its own and previous observations according to the model. Likewise, whole subjects are scored using the average of all its transition scores. Hence, the proposed approach is adapted to detect anomalous portions or entire MTS, fitting into numerous scenarios. A score-analysis phase is available to classify each score. Two main strategies are studied, being Tukey's Method [5, 6] and Gaussian Mixture Models (GMM) [7]. A threshold is automatically selected to determine the outlierness disclosure boundary.

A complete anomaly detection system is assembled, portioned in 4 phases. The latter are pre-processing, modeling, scoring and score-analysis. Pre-processing is comprised by a discretization and dimensionality reduction method known as SAX [8], which is employed prior to modeling when handling continuous MTS. Additionally, a review of existing outlier detection algorithms on temporal and sequential data as well as a rundown on Bayesian modeling is available. The

proposed system is validated through synthetic and real data sets. Furthermore, a multivariate probabilistic suffix tree (PST) technique is built and contrasted with the proposed approach.

Due to the increasing demand of data science related appliances aspiring not only promptness but also easily adaptable mechanisms, the current system is made completely free and accessible through a web application [9]. The latter does not require any download and is accompanied by a tutorial video. Furthermore, a paper based on the current thesis is currently under preparation for submission in a journal.

A general perspective of outlier detection is explored in Sec. II, together with existing literature. Theoretical background regarding Bayesian modeling is made available in Sec. III prior to the description of each phase of the proposed system from pre-processing to score-analysis in Sec. IV. Experimental validation and conclusions are available respectively in Sec. V and Sec. VI.

II. Outlier detection

Anomaly detection has consistently been present in a vast range of applications [10, 11] with an ever evolving research aspiration by data science communities. Without a slowdown in sight, outlier detection mechanisms are becoming more and more extensive with the aim of tackling each scenario particularities as well as the varied challenges they face. Outlier detection is engaged across multiple data types, being data's nature of great importance. Data can subsist of solely one attribute (univariate) or multiple attributes (multivariate). Attributes can be binary, categorical, continuous among others. Mechanisms used to categorize relationships within data are influenced by data's nature. An example are spatial and temporal relationships. Temporal outlier detection arises from frameworks handling sensor data, biomedical data among others. Temporal trends play a crucial role in anomaly discovery. Data patterns are not assumed to change abruptly through time, which is the main reasoning behind most of existing techniques. This is normally not a reasonable assumption when considering temporal data.

One of the first steps towards performing outlier detection is defining outlier itself. Anomalies have been defined by Grubbs [12] as "*outlying observations, or outliers, that appear to deviate markedly from other members of the sample in which they occur*". These observations can be single or collective instances independent of context. Differently, contextual outliers are data instances whose abnormality is caused by specific contexts rather than solely their value. Furthermore, frequency outliers correspond to patterns whose frequency is abnormal.

Outlier detection algorithms can be organized in three groups, analogous to classification problems. Unsupervised methods determine outliers without prior knowledge, assuming these can be separated from the rest of the data. Supervised algorithms use pre-labeled data to model both normality and abnormality. Semi-supervised methods require only one of the classes to be labeled, typically modeling solely normal data.

At first glance outlier detection seems trivial, however, we quickly realize the breadth of existing challenges. Due to a vast variety of techniques, choosing and adapting an existing method to a specific problem is a lingering activity. Some models are even impossible in some problem formulations. Furthermore, different domains have different notions of anomaly, being these continuously changing. Data depicted as abnormal today may not be tomorrow. Additionally, outliers are often blended with normal data which worsens if the anomalies are intentional. Such is common in security applications.

Multiple outlier detection algorithms have been considered for temporal data [13]. An example are regression methods, extensively used in TS data. They adapt equations or models to datasets considering outliers as high residual entities. An example are AutoRegressive (AR) models [11] decreeing that data points linearly depend on previous values and on a stochastic term, representing a random process. AR models can be made more robust by simultaneously considering outlier scores of previous observations, autoregressive moving-average (ARMA) [14] models are built.

In distance-based techniques, the distance or similarity between data instances is computed. Outliers are believed to be isolated from the rest of the data. A k-th nearest neighbor (KNN) technique has the objective of, for each test instance, determine the distance to its kth nearest neighbor in the training set, being these used as anomaly scores [10]. The same reasoning can be applied to discrete sequences [15] using the length of the longest common subsequence as a measuring distance, and easily adapted to TS data.

Classification and clustering methods provide additional insight on outlier detection problems. An example are one-class support vector machines (SVM) [16] which learn a region containing training data. A boundary is computed around the region of normal behavior. Each data instance is evaluated according to its position regarding the boundary. Expectation-maximization (EM) methods are capable of finding parameter estimations which determine the distribution of data. The latter are commonly used when considering Gaussian mixture models (GMM), which are considered in the score-analysis phase of the proposed approach. Created clusters can then be used to classify test instances. Anomalous samples are expected to deviate themselves from the clusters. A robust EM algorithm used in outlier detection is established in [17].

Statistical methods fit a statistical model to data observations in order to shape normal behavior. Tests are performed to determine if the probability of an instance is appropriate to be considered normal. Techniques based on dynamic Bayesian networks have proved to benefit anomaly detection [18] and gene expression data modeling [19].

III. Theoretical background

The implemented application is designed to handle discrete MTS. Other types of data like symbolic sequences can be loaded as well if in the right format.

A univariate TS $X_i[t]$ is seen as a set of observations along a consistent time rate represented as

$$X_i[t] = x_i[1], \ldots, x_i[T] \quad i \in \mathbb{N}, \tag{1}$$

expressing variable i along T time-stamps. For the case of MTS, each row of a data-set represents a subject identified by its row index. Subjects are MTS with a common number of variables n and width T. Each column depicts a certain variable at a certain time stamp. In other words, a subject S_j of length T and n variables is represented as

$$S_j = x_1[1], \ldots, x_n[1], \ldots, x_1[T], \ldots, x_n[T], \tag{2}$$

where $x_i[t]$ depicts the discrete value of variable $i \in 1, \ldots, n$ at time stamp $t \in 1, \ldots, T$ of subject S_j with row index j. A subject is thus a combination of univariate TS, each represented by Eq. (1). Columns are sorted according to time steps rather than variable indexes. Many appliances require anomaly detection to be engaged in TS descendant from sensor devices. These are normally not discretized and can also reveal unwanted offsets. To tackle such issues, a pre-processing phase is employed formerly to the outlier detection mechanism.

1) ***Bayesian networks****:* A Bayesian network (BN) is a probabilistic graphical model which encodes conditional relationships among variables. It is composed by two components, a directed acyclic graph (DAG) encoding dependencies between random variables and a set of parameters defining the local conditional distributions of the attributes. Random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ are assumed to be discrete with a finite domain. The DAG represents the set of variables (nodes) and their conditional dependencies (edges). Two unconnected random variables X and Y are conditionally independent given a third random variable Z, $X \perp Y|Z$, if and only if $P(X \cap Y|Z) = P(X|Z)P(Y|Z)$. A node X_p is pronounced as a parent of node X_s if there is a connection directed from X_p to X_s. A variable is independent of all its non-descendant nodes given its parents. Each node is associated to a local probability distribution storing a set of parameters encoding each probability of a possible configuration of a node X_i given its parents $p_a(X_i)$,

$$P(X_i = x_{ik}|p_a(X_i) = m_{ij}), \quad i \in \{1, \ldots, n\} \tag{3}$$

where x_{ik} is the $k\text{-}th$ possible value from the domain of X_i and m_{ij} the $j\text{-}th$ configuration of the set of variables $p_a(X_i)$. The set of conditional probabilities associated to each node is denotes the BN parameters.

The network's joint probability distribution is composed of the several local probability distributions associated to each variable and can be used to compute the probability of an evidence set. By evoking the Markov property, it comes

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|p_a(X_i)). \tag{4}$$

Learning the structure of a BN [20] can be summarized as finding the DAG which may have generated training data. A network's goodness of fit is measured using a decomposable scoring function, such as the log-likelihood (LL). Network parameters are computed by counting the number of observations of each variable given its parents.

MTS contain inter-variable as well as temporal dependencies. DBNs [21] are BNs which relate variables over adjacent time steps. These can model probability distributions over time. Unlike standard BNs, first-order DBNs are composed by multiple networks known as transition networks B_{t-1}^{t} over attributes from slices $t-1$ and t. A transition network possess two types of connectivity among variables noted as inter and intra-slice connectivities. The latter refers to variable dependencies at the same time frame, just like standard BNs. Inter-slice connectivity is responsible for the temporal aspect which relates attributes of different time slices. Nodes $X_i[t]$ have an associated time-stamp, possessing time-dependent parameters. Temporal associations among entities are obtained from consecutive data, since these are typically highly correlated in time. Concepts defined for standard BNs can thus be generalized to DBNs by considering variables at different time steps as distinct attributes.

A DBN can be stationary, possessing transition networks B_{t-1}^{t} which are invariant over time. A single transition structure is unrolled to every frame. Non-stationary DBNs are formed by a set of B_{t-1}^{t} particular of each transition. The latter adapt easier to time signatures. Furthermore, DBNs possess a prior network B_0 denoting the distribution of initial states.

The order associated to a DBN influences the structure behind it. An increase of the order L associated to a DBN means that a certain attribute $X_i[t]$ at time frame t can possess parent nodes from $t-L$ to t. Transition networks require information from $L+1$ time frames when considering a lag of L. A transition $t-1 \rightarrow t$ can thus be seen as a window D_{t-1}^{t} which is compromised by observations of the attributes belonging to the respective transition. The current thesis assembles a sliding approach to gradually capture windows from training data, the latter are scored by how likely they could have been generated by the model. A LL scoring criterion is used.

Learning a DBN is essentially disclosing the set of transition networks which better match a training set. In the proposed approach, an optimal tDBN structure learning algorithm [4] is used in the modeling phase. The latter provides optimum inter/intra-slice connectivities for each transition network, contrary to existing literature. An attribute node at a certain time-slice can only possess at most one parent at that same slice. Furthermore, in each node, the maximum number of parents from preceding time slices is bounded by a parameter p.

IV. Proposed method

A. Pre-processing

Real-world datasets have a tendency to be continuous and of high dimension. Un-discretized variables foment an heavy and over-fitted model which ends up behaving poorly. The same can be said to over-sampled data. Being the trained model a DBN, pre-processing is required.

A representation known as Symbolic Aggregate approXimation (SAX) [8] is enforced on each input TS prior to the modeling phase. SAX has already been practiced in anomaly detection scenarios [22] providing discretization and dimensionality reduction. The procedure is applied to each univariate TS separately. Each series is then combined to form a discrete MTS dataset, each with an alphabet $\Sigma = \sigma_1, ..., \sigma_{|\Sigma|}$.

The pseudo-code for the SAX pre-processing mechanism is available in Algorithm 1 and specified next:

1) **Normalization**: Every TS is normalized to present zero mean and a standard deviation of one, such is achieved by employing Z-normalization. The mean of a TS is subtracted from every data point. The result is then divided by the TS' standard deviation.
2) **Dimensionality Reduction**: Each TS of length T can be compressed into equivalent sequences of size $m \ll T$. Such can be assured by Piecewise Aggregate Approximation (PAA). The latter subdivides a normalized TS into m equally sized windows. The mean of each window of size T/m is computed replacing all its values. The m means of each window serve as the new TS.
3) **Symbolic Discretization**: Normalized TS typically have Gaussian distributions [23]. Hence, their domain can be divided into $|\Sigma|$ equiprobable regions according to a Gaussian distribution $N(0,1)$, where $|\Sigma|$ denotes the size of the alphabet. Regions are identified by boundaries, known as breakpoints β_i. The goal is to resolve in which of the regions each TS point resides. A value falling in interval $[\beta_{i-1}, \beta_i[$ is associated to symbol σ_i.

Note that PAA can be overlooked, being normalized TS directly discretized. Additionally, variable selection can be performed prior to SAX to reduce ambiguous variables.

Algorithm 1 Data pre-processing

Input: Multiple continuous MTS S_j with a fixed length T and format. These are considered not to have missing values.
Output: The corresponding input MTS discretized with a fixed alphabet size $|\Sigma_i|$, $1 \leq i \leq n$ for each variable X_i and reduced to a length $m \ll T$.

procedure SAX
  for each MTS S_j **do**
    for each variable X_i of S_j **do**
      $Norm_i \leftarrow z\_Norm(X_i)$
      function PAA($Norm_i, m$)
        for each section of size T/m **do**
          $\widehat{x}[i] \leftarrow (m/T) \sum_{j \in section} Norm_i[j]$
          $\widehat{N}orm_i \leftarrow \widehat{N}orm_i.Append(\widehat{x}[i])$
      function DISCRETIZATION($\widehat{N}orm_i, |\Sigma_i|$)
        $\beta \leftarrow SegmentGaussianDistrib(|\Sigma_i|)$
        for each value i in $\widehat{N}orm_i$ **do**
          $Discrete_i^j \leftarrow ToSymbolic(i, \beta)$

B. DBN Outlier Detection

A statistical outlier is defined as being a single or set of observations belonging to a subject which is suspicious of not being generated by the data's main underlying processes, being typically caused by qualitatively distinct mechanisms. After performing the pre-processing and modeling phases, the scoring phase is discussed.

A window is defined as a sample of a discrete MTS with n variables at a specified time interval, having a size equal to $n \cdot (L+1)$. The DBN's order is represented by L. A window is described as

$$D_{t-L}^t = x_1[t-L], \ldots, x_n[t-L], \ldots, x_1[t], \ldots, x_n[t], \quad (5)$$

where $L \leq t \leq T$ identifies the last time frame present in the window. Attributes of a certain time slice t are conditioned by nodes no later than L prior time frames, being such dependent on the DBN structure. All the information mentioned is akin when considering both stationary and non-stationary DBNs.

Conditional probabilities for each attribute are kept in matrices within the algorithm. The model is composed by a prior network B_0 and a transition network B_{t-1}^t for each transition $t-1 \rightarrow t$ in the case of a first-order DBN. A Transition is associated to a window D_{t-1}^t composed by the observed attributes required to compute its corresponding anomaly score. A transition score is computed as

$$Score_{t-L}^t = \sum_{i=1}^{n} \log(P(x_i[t] | p_a(x_i[t]))), \quad (6)$$

where $P(x_i[t]|p_a(x_i[t]))$ is the probability of attribute $x_i[t]$ conditioned by its parent nodes' values, represented by $p_a(x_i[t])$. The latter can be attributes from the same time frame or prior ones according to the transition network. Equation (6) portrays the log-likelihood (LL) of the transition, which indicates the probability of the observations of time frame t given the window's observations. The LL score is applicable to any type of DBN, with the difference between stationary and non-stationary models being the transition network considered for the computation of the conditional probabilities. It is worth noting that a transition score represents the outlierness of the transition and not the time slice.

However, if the evidence possesses an unseen pattern, the probability of at-least an attribute is zero, nullifying the LL score associated to it. A technique known as probability smoothing is thus employed to prevent score disruption for unseen patterns. Probabilities are transformed according to

$$P_i = (1 - |\Sigma_i| y_{\min}) p_i + y_{\min}, \quad (7)$$

where p_i is a conditional probability $P(x_i[t]|p_a(x_i[t]))$, $y_{\min}$ a parameter expressing the degree of probability uncertainty and $|\Sigma_i|$ the granularity of the variable in question. This means that when p_i is zero, the new probability will be equal to y_{min}, instead of zero. Additionally, Eq. (7) ensures that probabilities of one are decreased according not only to y_{min} but also the size of the alphabet related to that attribute, reducing thus overfitting. *Doubt is not a pleasant condition, but certainty is absurd* [24]. Consequently, the LL score is computed using the smoothed probabilities.

C. Sliding Window

To acquire the outlierness of every MTS transition, a sliding window is employed. The latter can be seen as a sub-network of a DBN over consecutive time slices. The mechanism gradually captures all equally sized windows D_{t-L}^t of a subject

to compute the LL scores $Score^t_{t-L}$ for each transition. Since the trained model possesses an initial network B_0, time frames $t \leq L$ can not be explained by windows of size $n \cdot (L+1)$. Hence, according to the model's order, only transitions from slice $L+1$ forward are captured. However, the initial frames influence the scores of the next consecutive windows which include them, having the ability of inducing anomalies. The whole procedure is depicted in Algorithm 2.

Furthermore, subject scoring is made available, offering the detection of anomalous MTS. A straightforward approach is considered, being a subject outlierness equal to the mean of every transition score of that subject. A subject j is scored as

$$SubjectScore_j = \frac{1}{T-L} \sum_{t=L+1}^{T} Score^t_{t-L}, \tag{8}$$

using smoothed probabilities, where $Score^t_{t-L}$ represents transition scores belonging to subject j.

The next step is to apply score analysis to discern the final decision boundary between normal and anomalous scores. Scores below a specified value are classified as outliers.

Algorithm 2 Transition outlier detection

Input: A DBN model with stored conditional probabilities for each transition network B^t_{t-L}, a MTS dataset containing subjects S_j and a threshold $tresh$ to discern abnormality.
Output: The LL scores for every transition $t-L \rightarrow t$ below the specified threshold.

```
procedure
    for each time slice t do
        for each subject S_j do
            D^t_{t-L} ← x_1[t-L]...x_n[t-L]...x_1[t]...x_n[t]
            X_t ← x_1[t], ..., x_n[t]
            function SCORING(D^t_{t-L}, X_t, t)
                TNet ← DBN.getTransNet(t)          ▷ B^t_{t-L}
                for x_i in X_t do
                    pa_i ← TNet.getParents(x_i)
                    p_i ← TNet.getProbability(i, pa_i)
                    P_i ← (1-|Σ_i|y_min)p_i + y_min
            Score^t_{t-L} ← Σ_i log P_i
            if Score^t_{t-L} < tresh then
                outliers ← outliers.Append(Score^t_{t-L})
```

D. Score Analysis

Two score-analysis strategies are studied to elect an optimum threshold for outlier disclosure amidst scores.

*1) **Tukey's strategy**:* One can define abnormal scores as values that are too far away from the norm, presuming the existence of a cluster comprising normality. The current technique has inspiration in John Tukey's method [5, 6], which determines the score's interquartile range (IQR) as

$$IQR = Q3 - Q1, \tag{9}$$

where $Q1$ and $Q3$ are the first and third quartiles respectively. The IQR measures statistical dispersion, depicting that 50% of the scores are within $\pm 0.5 \cdot IQR$ of the median. By ignoring the scores' mean and standard deviation, the impact of extreme scores does not influence the procedure. The IQR is hence a measure of variability robust to the presence of outliers.

Tukey uses the notion of *fences* [6], frontiers which separate outliers from normal data. The proposed approach typically generates negatively skew score distributions. Hence, a lower fence computed as $Q1(1.5 \cdot IQR)$ is used. The reason behind choosing $1.5 \cdot IQR$ is that for most cases, a value of IQR labels too many outliers (too exclusive) while $2 \cdot IQR$ begins to classify extreme values as normal (too inclusive), being such value fruit of conducted experiments [6]. Transition and subject scores are classified as abnormal if their value subsists below their respective lower fence, since these are low likelihood entities. Thus, scores s_i holding inequality

$$s_i \leq Q1 - (1.5 \cdot IQR) \tag{10}$$

are considered abnormal, being $Q1-(1.5 \cdot IQR)$ the threshold.

Tukey's procedure prefers symmetric score distributions with a low ratio of outliers having a breakdown at about 25% [25]. In scenarios with absence of anomalies, this mechanism is capable of completely eliminating false positive occurrences, since fences are not forced to be in the data's observed domain. Outliers may be formed by specific phenomena which can cause the appearance of detached clusters in the score's histogram. The latter may be considered as normal. Such is typically noticeable in the presence of a large outlier ratios.

*2) **Gaussian mixture model**:* To handle disjoint score distributions, a method based on a Gaussian Mixture Model (GMM) [7] is employed. Commonly used in classification and clustering problems, GMMs are probabilistic models that assume data is generated from a finite mixture of Gaussian distributions with unknown parameters. Most real-world phenomena has Gaussian like distributions [23].

A score distribution is modeled as a mixture of two Gaussian curves. Labeling each score becomes a classification problem among two classes C_1 and C_2, representing normality and abnormality respectively. Such is interpreted as uncovering the value of $P(C_1, C_2|y)$ for each score value y, which can be obtained by employing Bayes Rule

$$P(C_i|y) = \frac{P(y|C_i) \cdot P(C_i)}{P(y)}, \quad i \in 1, 2 \tag{11}$$

where $P(y|Ci)$ is the likelihood of score y belonging to class C_i, $P(C_i)$ the priors for each class and $P(y)$ the evidence. The threshold is the boundary that better separates both curves, which describes the point of maximum uncertainty. Evidence $P(y)$ for each score is calculated according to

$$P(y) = P(y|C_1)P(C_1) + P(y|C_2)P(C_2). \tag{12}$$

Combining Eq. (11) and Eq. (12) leads to the conclusion that for a score y be classified as anomalous, $P(y|C_1)P(C_1) > P(y|C_2)P(C_2)$. Such is known as the Bayes Classification Rule (BCR) that provides the desired boundary.

To discover the parameters of each Gaussian distribution. The GMM is defined as the sum of the two Gaussian distributions, $\alpha_1 N(Y|\mu_1, \sigma_1^2) + \alpha_2 N(Y|\mu_2, \sigma_2^2)$. A Expectation

Maximization algorithm [26] is used to determine the values of parameters α_i, μ_i and σ_i^2. In the current thesis, GMM is employed with the aid of the available R package mclust [27].

The GMM strategy can handle discontinued score distributions, however, it assumes the existence of an outlier cluster. Thus, Tukey's and GMM strategies expect distinct scenarios.

V. Experimental Results

To outline the performance of the proposed approach, several experiments are conducted using simulated data as well as real-world datasets from distinct sources and applications.

A. Simulated data

The performance and consistency of the proposed approach is validated using simulated data. More specifically, the present experiments subsist on training two separate DBNs, one for generating normal data DBN_N and another to produce outliers DBN_O. All data is mixed together in a single dataset and fed to the system. A DBN is trained using the combined dataset, with the aim of locating subjects generated by DBN_O.

To evaluate the performance of each experiment, the number of true positives (TP), false positives (FP) and false negatives (FN) is measured. Such are used to determine the Positive Predictive Value (PPV), representing precision, and True Positive Rate (TPR), representing recall. These metrics assume a value from 0 to 1 and are computed as

$$PPV = \frac{TP}{TP + FP} \quad \text{and} \quad TPR = \frac{TP}{TP + FN}. \tag{13}$$

To conjointly consider both metrics, a measure denoted as F_1 score is computed along with the accuracy (ACC) of each test

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} \quad \text{and} \quad ACC = \frac{TP + TN}{P + N}, \tag{14}$$

where $P+N$ represents the sum of all positives and negatives.

Experiments are identified by their outlier ratio compared to normal data and the anomalous model, DBN_B or DBN_C, used. Transition networks for the anomalous networks are displayed in Fig. 1 together with their dissimilarities with respect to the normal model. The higher the outlier ratio, the more the trained model admits them. The same is said for anomalies generated by structures similar to DBN_N. This results in a scoring procedure which reduces the score of the deviants while possibly increasing the number of FN and FP.

Each test is identified using $\text{DBN}_O\text{\_P}_O\text{\_N}_s$. The latter indicates which anomalous network DBN_O, DBN_B or DBN_C is employed, the percentage of subjects that belong to the said anomalous network P_O and the total number of subjects in the dataset N_s. Columns N_O store the number of anomaly subjects for each experiment. Every value is rounded down to two decimal places and represent an average among 5 trials. Since subject scores are determined by an average of all the corresponding transition scores, it is possible to observe which transitions contributed for the classification of each subject, validating thus both disclosures simultaneously.

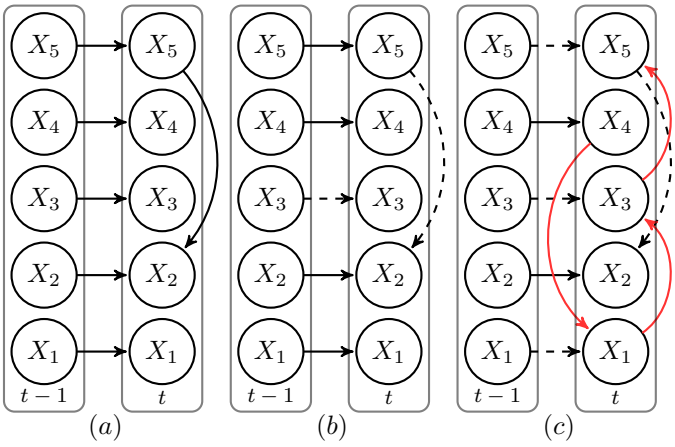

Fig. 1. Transition networks of stationary first-order DBNs. The left network (a) represents the transition network of DBN_N which generates normal subjects. Both networks (b) and (c) represent DBN_B and DBN_C respectively which generate anomalous subjects. Dashed connections represent links which are removed with respect to the normal network (a), while red links symbolize added dependencies. Non-dashed black edges are connections which are common with respect to (a).

The proposed approach is tested against simulated data. Experiments are divided in two groups due to the two possible strategies in the score-analysis phase. Such means that both strategies are validating the proposed approach. Experiments considering a different approach are carried out afterwards.

*1) **Tukey's score-analysis**:* Results employing Tukey's method in the score-analysis phase are shown in Table I, it comes that datasets with only 100 subjects perform generally poorly. Such is explained by the fact that these do not possess enough information about the data's underlying processes. Replicating the results of an experiment with only 100 subjects is difficult, since these are very sensible to each observation. Accuracy as well as F_1 scores tend to decline with the increase of outlier ratios, which is expected due to less normal data available for a correct modeling phase. The latter is observed by the decrease of TPR measurements. The computed thresholds converge to more stable values with the increase of data causing oscillations to tend to zero. Hence outputting more reliable values for every performance measure.

Discussing the impact of outlier ratios, Tukey's method is recognized to be more effective in the presence of lower anomaly percentages. The aforementioned arises from the fact that score distributions start to be increasingly asymmetric with the increase of outlierness, being the latter confirmed in existing literature [28]. Moreover, when the ratio is high enough and the majority of outliers are generated by a common process, the score distribution of abnormal data becomes visible. Such phenomena explains why for the same ratio, F_1 scores may decrease with the increase of subjects when employing Tukey's threshold. The breakpoint of Tukey's method [25] prevents favorable results when in the presence of abundance outlierness. However, FP tend to disappear, reflecting high precision measurements. Such is explained by

TABLE I
Subject outlier detection of the proposed approach on simulated data using Tukey's strategy

Experiment	N_O	Threshold	PPV	TPR	ACC	F_1	Experiment	N_O	Threshold	PPV	TPR	ACC	F_1
B_5_100	5	-8,72	0.88	0.70	0.98	0.78	C_5_100	5	-8.61	0.89	0.73	0.98	0.80
B_5_1000	50	-5.56	0.93	0.96	0.99	0.94	C_5_1000	50	-5.46	0.91	0.98	0.99	0.94
B_5_10000	500	-5.47	0.95	0.98	0.99	0.96	C_5_10000	500	-5.32	0.94	1.00	0.99	0.97
B_10_100	10	-9.85	0.96	0.38	0.94	0.54	C_10_100	10	-8.72	0.89	0.73	0.97	0.80
B_10_1000	100	-5.61	0.99	0.87	0.99	0.93	C_10_1000	100	-5.57	0.97	0.87	0.98	0.92
B_10_10000	1000	-5.56	0.99	0.91	0.99	0.95	C_10_10000	1000	-5.55	0.99	0.87	0.98	0.93
B_20_100	20	-10.01	1.00	0.19	0.83	0.32	C_20_100	20	-8.71	0.90	0.22	0.84	0.35
B_20_1000	200	-6.08	1.00	0.20	0.84	0.33	C_20_1000	200	-6.17	1.00	0.37	0.87	0.54
B_20_10000	2000	-6.09	1.00	0.16	0.83	0.28	C_20_10000	2000	-6.01	1.00	0.29	0.86	0.45

the fact that Tukey's thresholds avoid the main distribution also causing FN to rise.

Comparing experiments from both anomalous networks DBN_B and DBN_C, accuracy is in general higher in experiments with DBN_C. Such is expected, since the latter has fewer connections in common with DBN_N, resulting in a more dissimilar structure. However, such is not always visible true, since asymmetric distributions disturb Tukey's analysis.

Control experiments performed using data sets solely comprised by normal subjects demonstrated favorable results when employing Tukey's score-analysis, with the detection of few FP. On the hand, the GMM strategy divided the distribution in two classes creating an high number of FP. It is worth noting that despite FP being unwanted, these are normal to appear in the present experiments. DBN_N generates sequences with a low probability due to its parameters. When creating a large number of transitions, these are doom to occur.

Advantages and disadvantages of both score-analysis strategies are discussed next using the same experiments.

*2) **Gaussian mixture model**:* Inspecting results using Tukey's analysis, the performance of experiments with larger anomaly ratios bear low F_1 values due to high counts of FN. The main reason for the aforementioned is the presence of an outlier curve in the scores' distribution. The latter occurs due to the high proportion of outliers formed by a specific mechanism, in this case an abnormal DBN. GMM score-analysis is thus employed in the same experiments as Tukey's method, being the score-analysis phase the only difference affecting threshold computation.

Results are available in Table II, where it is noticeable the considerable increase in recall for experiments with higher proportions of outliers when compared with results from Table I. Such is confirmed in existing literature [28]. In general, the number of FP is higher when employing GMM. Such is caused by the GMM's assumption of the existence of an abnormal model even in its absence. Due to similarities between DBNs, especially when considering DBN_B, scores from both networks tend to mix together around the threshold being thus difficult to discern them. the GMM approach has typically higher recall but lower precision with thresholds smaller in module. The latter is more noticeable in higher outlier ratios, since in the presence of fewer anomalies Tukey's method displays higher F_1 scores.

Although the present synthetic experiments endorse the use of GMM score-analysis, one should note that both normal and abnormal data are generated according to two defined models which can by some degree be separated. Such is a favorable scenario for GMM. When considering other scenarios, Tukey's method is not so susceptible to the presence of well defined curves being thus always a strategy to consider. With that said, it is adviced for the analyst to apply as much knowledge as possible with the experimentation of both strategies.

*3) **Comparison with probabilistic suffix trees**:* To compare the proposed system, an additional multivariate outlier detection mechanism is built. The latter adopts PSTs [29] with the aim of mining abnormal values in MTS. These methods are only capable of modeling univariate TS. To tackle MTS, each dataset is divided into multiple sets, each one containing data concerning solely one variable. Every set is used to model a PST, where subjects are seen as a set of separate sequences s_j for each variable j associated to its corresponding PST. Thus, subjects with 5 variables model 5 trees, which do not interchange any information. Each PST computes an univariate score $logloss(s_j)$ for all subjects considering its variable, according to $logloss(s_j) = \frac{1}{l}\log_2 P(s_j)$ [30], where l is the sequence maximum length. Scores from all the separate PSTs belonging to a common subject are stored in an array, being its mean the multivariate score for the MTS subject.

An existing PST modeling software [30] is engaged. Each experiment is compared with the proposed approach. Likewise, score-analysis is employed posterior to scoring, selecting one of the two considered strategies. Every PST is modeled with maximum possible depth and a *nmin* of 30. Pruning using the G1 function [30] with a C of 1.2 is used.

Experiments performed with solely 100 subjects did not provide satisfactory results, since the latter do not grant sufficient information. Tests are thus focused on larger datasets, and are available in Table III. Results demonstrate the low performance of the PST approach when discerning anomalies generated by DBN_B. Such is explained by the fact that this network is much similar to DBN_N when compared with DBN_C. Furthermore, since inter-variable relations are not considered, subjects become identical when seen by the PSTs. Hence, the resulting score distributions display a single curve blending both classes. One exception are experiments considering 5% of anomalies, which indicate that with the

TABLE II
Subject outlier detection of the proposed approach on simulated data using GMM's strategy

Experiment	N_O	Threshold	PPV	TPR	ACC	F_1	Experiment	N_O	Threshold	PPV	TPR	ACC	F_1
B_5_100	5	-8.78	0.82	0.70	0.98	0.76	C_5_100	5	-7.51	0.64	1.00	0.96	0.78
B_5_1000	50	-5.71	0.91	0.97	0.99	0.94	C_5_1000	50	-5.40	0.86	0.99	0.99	0.92
B_5_10000	500	-5.45	0.95	0.98	0.99	0.96	C_5_10000	500	-5.49	0.98	1.00	0.99	0.99
B_10_100	10	-8.55	0.77	0.68	0.93	0.72	C_10_100	10	-8.64	0.92	0.78	0.97	0.84
B_10_1000	100	-5.43	0.94	0.96	0.99	0.95	C_10_1000	100	-5.24	0.89	0.97	0.98	0.93
B_10_10000	1000	-5.27	0.91	0.98	0.99	0.94	C_10_10000	1000	-5.28	0.93	0.96	0.99	0.95
B_20_100	20	-8.18	0.66	0.49	0.85	0.56	C_20_100	20	-7.17	0.75	0.58	0.88	0.65
B_20_1000	200	-5.18	0.86	0.89	0.94	0.87	C_20_1000	200	-5.24	0.91	0.92	0.96	0.92
B_20_10000	2000	-5.13	0.86	0.94	0.96	0.90	C_20_10000	2000	-5.11	0.93	0.94	0.97	0.94

TABLE III
PST results using Tukey (black) and GMM (violet) strategies on simulated data

Experiment	N_O	Thresh	PPV	TPR	ACC	F_1
B_5_10000	500	-1.55	0.96	0.73	0.98	0.83
C_5_10000	500	-1.61	0.96	0.94	0.99	0.95
B_10_10000	1000	-1.65	0.70	0.02	0.90	0.04
C_10_10000	1000	-1.63	0.98	0.39	0.94	0.56
B_20_10000	2000	-1.74	0.42	0.00	0.80	0.00
C_20_10000	2000	-1.73	1.00	0.03	0.81	0.06
B_5_10000	500	-1.47	0.86	0.88	0.99	0.87
C_5_10000	500	-1.59	0.94	0.95	0.99	0.94
B_10_10000	1000	-1.27	0.20	0.87	0.65	0.33
C_10_10000	1000	-1.55	0.88	0.68	0.96	0.77
B_20_10000	2000	-1.30	0.25	0.67	0.53	0.36
C_20_10000	2000	-1.45	0.763	0.883	0.92	0.82

increase of outlier ratios, the few dissimilarities among classes are modeled, causing outliers to fit each PST. Regarding experiments with DBN_C, the latter output better results when compared to DBN_B. Such is mainly due to the differences present in the intra-variable relationships between DBN_C and the normal model DBN_N. In Fig. 2 a comparison between

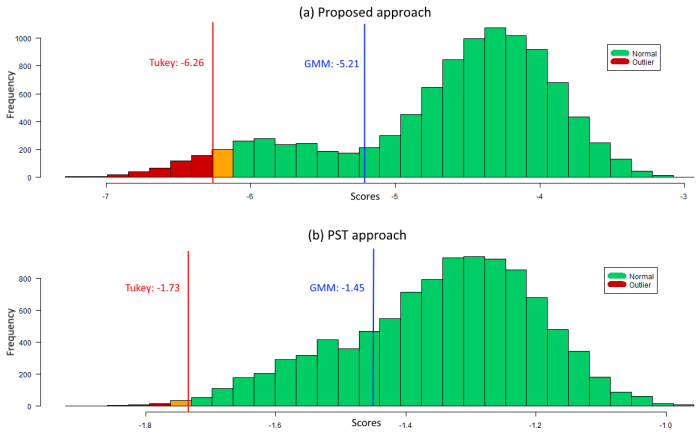

Fig. 2. Subject outlierness using the proposed approach (a) and the PST approach (b) for a same experiment C_20_10000. Both histograms display thresholds using both score-analysis strategies. Scores below the threshold are classified as abnormal (in red) while the rest are classified as normal (in green). The presented color representation is considering Tukey's thresholds.

the proposed and the PST approaches for an experiment with 20% of outliers from DBN_C is shown. The PST system can not separate both classes as well as the DBN approach, demonstrating the importance of inter-variable relationships present in DBN_C and the robustness of the proposed approach. In general, the PST approach scales poorly with the increase of outlier ratios and never demonstrated to outperform the proposed approach in the experiments conducted.

B. ECG

A common application of anomaly detection in medical scenarios is in Electrocardiogram (ECG) alert systems [22]. These have the capability of detecting unusual patterns in signals measured from patients. Data is usually un-discretized and present expected patterns when under normal circumstances.

An ECG dataset, available at [31], is composed by 200 MTS. The location of the ventricular contraction peaks typically occur at time frames 3 and 10. Tests are performed using non-stationary DBNs since specific phenomena occurs in particular time instances, peculiar of ECG signals with common phase. The experiments have the objective of testing the system's behavior to noisy and unstable data.

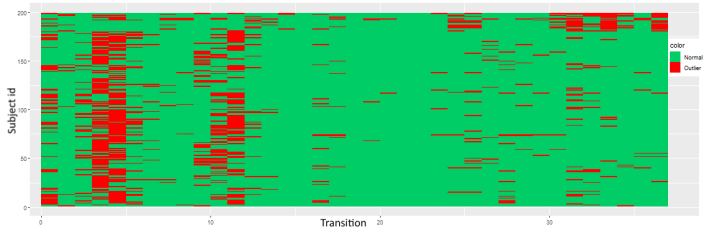

Fig. 3. ECG dataset transitions arranged by subject. A non-stationary second-order DBN model is used together with Tukey's score-analysis. Flipped subjects are associated to the highest subject ids. Data is discretized using SAX with an alphabet Σ of 5 symbols. Transitions displayed in red are classified as abnormal while green ones are classified as normal.

A conducted experiment, available in Fig. 3, shows that the current approach has difficulty evaluating time slices with higher variance. To further test the aforementioned, 10% of the subjects are flipped and mixed together in the original set. Fig. 3 demonstrate the detection of such transitions present on subjects with higher id. Series are discretized with a SAX alphabet of 5 and modeled using a second-order DBN, which demonstrates favorable results.

The system has the ability of detecting unusually behaved sections in ECGs which coincide with high variance portions.

The latter is due to not existing a predominant pattern in the location of the peaks, observable as vertical red stripes, since these vary intensively from subject to subject contrary to more advanced slices. SAX discretization offers low definition in such locations. Score distributions are negatively skew, accommodating the use of Tukey's thresholds.

C. Mortality

A cell wise outlier detection scenario where a multivariate dataset is confined in a data matrix is considered. Such is studied in [32], where the suggested approach *DetectDeviatingCells* classifies cell-wise as well as row-wise anomalies. One of the tested experiments [32] refers to a dataset comprising male mortality in France from $19th$ century forward, extracted from [33]. The aim is to discover outlying years, representative of the main iconic events in France's history.

Data is structured as a matrix where the only apparent variables are the mortality rate for each age group. To adapt it to the present approach, the several ages are regarded as variables X_i, meaning that correlations among mortality rates of different ages can be modeled. Due to the excessive number of ages, only a sub set is selected. The year in which measurements were obtained are regarded as time instants t for each variable X_i. With longitudinal data assembled, SAX pre-processing is applied to each series. It is worth noting that with all the transformations performed, the data set is reduced to a single subject which portrays a MTS. Attributes $X_i[t]$ are thus mortality rates of specific age groups at particular years.

Being data only comprised of a single subject, non-stationary models can not be employed. These would renounce the drawn aspirations due to overfitting, since only a single observation of each time slice is available.

Two experiments are presented in Fig. 4. A data set comprising France's male mortality rates from 1841 to 1987 is used. In the first experiment, 5 variables are selected, being ages 20, 30, 40, 60 and 80. Each variable is discretized with an alphabet size of 5 and all tests employ Tukey's method in the score analysis phase. The objective is to determine unusual events such as wars and epidemics. The trained model involves a stationary third-order DBN. Nodes are allowed to have at-most 1 parent from previous slices. The reasoning behind the parameter choice is purely experimental. The problem exhibits a preference of attributes establishing connections with previous nodes which are not consecutive with themselves. It's worth recalling that having an order of 3 does not mean that every or even any relation has such lag, it just offers such possibility.Results confirm major events which shaked France's history. These are displayed in Fig. 4, representing both world wars, the influenza pandemic, the Franco-Prussian War and the European revolutionary wave of 1848. France was a belligerent in several conflicts as well as colonization wars in the 1850s.

In the second experiment, a variable is added to the first set. The new age represents the male mortality rate of children aged 10 years old. The aim is to capture the impact of youth mortality in the outputted years. Results are similar, being the differences observed in the 1860s and around the Spanish flu confirming that youth is more susceptible to epidemics.

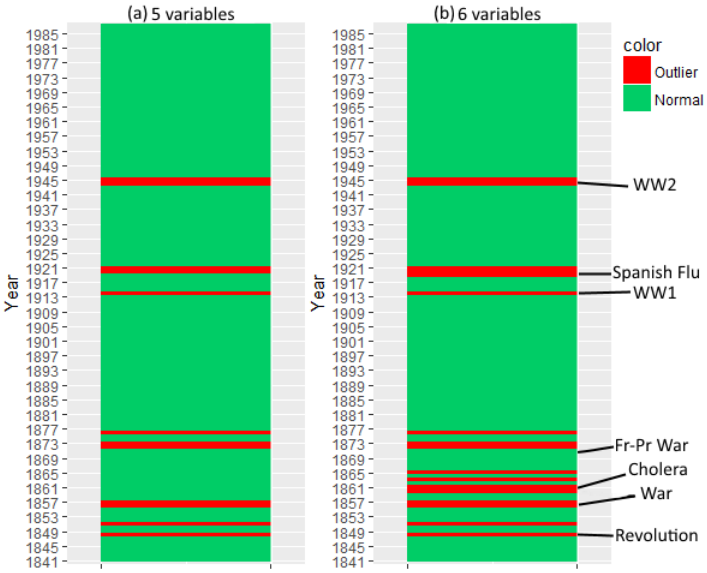

Fig. 4. Transition outlierness for mortality datasets of 5 (a) and 6 (b) variables using a third-order DBN. Dataset (a) is comprised by 5 variables representing male mortality rates of males with ages 20, 30, 40, 60 and 80. Dataset (b) includes the same variables as (a) with the addition of a variable representing the mortality rate of males aged 10. Transitions are arranged by year and classified as anomalous (red) and normal (green). Major wars and epidemics which affected France in the selected years are exhibited.

D. Pen-digits

A distinct application is the recognition of drawn digits. With measurements taken along time from each drawing phase, longitudinal data is built. Data is available at [34] and studied in [35]. Handwriting samples are captured using a sensitive tablet which outputs the x and y coordinates of the pen at fixed time intervals. The goal is to model the system to a certain character being simultaneously unwanted digits amidst the data. A set comprising 1143 MTS along 8 time frames representing digit 1 is assembled from 44 different writers. The original MTS are discretized with an alphabet size of 8. The dataset is injected with 130 subjects belonging to a different digit, roughly 10%. The aim is to detect the aforementioned and subsequently understand similarities between digits. Data conceals different types of features even for the same digit, since 44 writers are present, existing no pre-processing specific to the scenario, contrary to existing literature. Raw sensory data is directly applied to the system.

Results are present in Table IV, where D_i represents the anomaly digit i introduced. A first-order non-stationary DBN is modeled, demonstrating that a pair of coordinates is more easily explained by its immediate precedent. Every attribute can possess at-most one parent from its preceding slice. Thresholds are selected manually. The objective is not only to capture the performance of the outlier detection system but further understand which digits are more commonly resembled with digit 1. Results show that distinguishing digit 7 from 1

TABLE IV
Results of pen digits outlier detection experiments

Experiment	TP	FP	TN	FN	Tresh	PPV	TPR	ACC	F_1
D_7	24	41	1102	106	-3.0	0.37	0.18	0.88	0.25
D_8	98	45	1098	32	-2.5	0.69	0.75	0.94	0.72
D_9	90	42	1101	40	-3.0	0.68	0.69	0.94	0.69

is difficult due to their similarity, proved by the low F_1 score obtained. Such reflects the blending of both class distributions. Digits 8 and 9 proved to be more easily discerned from 1 with favorable performance measures.

VI. Conclusions

The developed system utilizes a sliding window mechanism to uncover portions as well as entire MTS, presenting an adaptable outlier detection system previously nonexistent. A diverse set of applications have proved to benefit from the former. It ranges from pre-processing to score-analysis including a DBN modeling phase versatile for longitudinal data. A widely available web application [9] is deployed to assist any analyst in their specific endeavour along with an user-friendly interface and tutorial. A complete MTS outlier detection system is provided capable of mining contexts with both temporal and inter-variable dependencies, oblivious in most existing literature. The developed approach shows particularly good results when employed in discretized data with strong inter-variable correlations. Although not being built for MTS with an high number of variables, most scenarios are analyzed successfully considering a smaller subset of variables, being thus easily adapted to such cases. Other types of temporal data are proven to accommodate the current system, being the pre-processing phase specially of great importance. Stationary models allow confronting data which does not vary its behavior over time, in opposition to non-stationarity. Parameter choices are capable of likewise change the perception of outlierness.

The developed system can be enhanced with an iterative training phase. Outliers are detected and removed in a first step, leaving normal data for further modeling. Such reduces the influence of anomalies in the obtained model, which can be further employed. In the score analysis phase, more complex and domain specific methods can be considered such as GMM with more than two components, regarding more than two classes. Furthermore, emphasizing specific scores or performing score analysis separately for each transition in the case of non-stationarity can enhance outlier disclosure. The modeling phase can further be augmented with the employment of change-point mechanisms in such cases. The latter can model changes in the underlying processes through time instead of considering each transition separately.

Different discretization and dimensionality reduction mechanisms should be studied and compared with the aim of better capturing the main features of data before modeling a DBN. In the current approach only SAX is considered. The enhancement of the multivariate PST mechanism is additionally encouraged. By providing an algorithm capable of modeling a tree encoding multivariate configurations, an efficient MTS variable length Markov model can surge.

References

[1] F. Zappa and P. Occhiogrosso. *Real Frank Zappa Book*. Simon and Schuster, 1989.

[2] J. López-de Lacalle. tsoutliers: Detection of Outliers in Time Series. *R package version 0.6-6. https://CRAN.R-project.org/package=tsoutliers*, 2017.

[3] D. V. Matt Dancho. anomalize: Tidy Anomaly Detection. *R package version 0.1.1. https://CRAN.R-project.org/package=anomalize*, 2018.

[4] J. L. Monteiro, S. Vinga, and A. M. Carvalho. Polynomial-time algorithm for learning optimal tree-augmented dynamic bayesian networks. In *UAI*, pages 622–631, 2015.

[5] J. W. Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[6] D. C. Hoaglin. John w. tukey and data analysis. *Statistical Science*, pages 311–318, 2003.

[7] G. McLachlan. Finite mixture models. *Annual Review of Statistics and Its Application*, 5(1), 2018.

[8] J. Lin, E. J. Keogh, S. Lonardi, and B. Y. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD*, pages 2–11. ACM, 2003.

[9] J. L. F. Serras. Dynamic Bayesian outlier detection. https://jorgeserras.shinyapps.io/outlierdetection/, October 2018.

[10] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[11] C. C. Aggarwal. *Outlier analysis*. Springer International Publishing Switzerland, 2017.

[12] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[13] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. *IEEE Trans. Knowl. Data Eng.*, 26(9):2250–2267, 2014.

[14] P. Galeano, D. Peña, and R. S. Tsay. Outlier detection in multivariate time series by projection pursuit. *J. Am. Stat. Assoc*, 101(474):654–669, 2006.

[15] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Trans. Knowl. Data Eng.*, 24(5):823–839, 2012.

[16] J. Ma and S. Perkins. Time-series novelty detection using one-class support vector machines. In *IJCNN*, volume 3, pages 1741–1745. IEEE, 2003.

[17] K. R. Koch. Robust estimation by expectation maximization algorithm. *Journal of Geodesy*, 87(2):107–116, 2013.

[18] D. J. Hill, B. S. Minsker, and E. Amir. Real-time bayesian anomaly detection in streaming environmental data. *Water Resources Research*, 45(4), 2009.

[19] K. Murphy, S. Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

[20] N. Friedman. The bayesian structural EM algorithm. In *UAI*, pages 129–138. Morgan Kaufmann, 1998.

[21] N. Friedman, K. P. Murphy, and S. J. Russell. Learning the structure of dynamic probabilistic networks. In *UAI*, pages 139–147. Morgan Kaufmann, 1998.

[22] E. Keogh, J. Lin, and A. Fu. HOT SAX: Finding the most unusual time series subsequence: Algorithms and applications. In *ICDM*, pages 440–449. Citeseer, 2004.

[23] R. J. Larsen, M. L. Marx, et al. *An introduction to mathematical statistics and its applications*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1986.

[24] Voltaire. Letter to Frederick II of Prussia, 6 April. https://en.wikiquote.org/wiki/Voltaire, 1767. Accessed: 23 September 2018.

[25] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, 88(424):1273–1283, 1993.

[26] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):381–396, 2002.

[27] C. Fraley, A. Raftery, L. Scrucca, T. B. Murphy, M. Fop, and M. L. Scrucca. Package mclust. 2017.

[28] P. R. Jones. A note on detecting statistical outliers in psychophysical data. *bioRxiv*, page 074591, 2016.

[29] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Mach. Learn.*, 25(2-3):117–149, 1996.

[30] A. Gabadinho and G. Ritschard. Analyzing state sequences with probabilistic suffix trees: the pst r package. *JSS*, 72(3):1–39, 2016.

[31] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[32] P. J. Rousseeuw and W. V. D. Bossche. Detecting deviating data cells. *Technometrics*, 60(2):135–145, 2018.

[33] Human Mortality Database. *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).*, Available at www.mortality.org or www.humanmortality.de (data downloaded on 18 September 2018).

[34] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[35] F. Alimoglu and E. Alpaydin. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In *TAINN*, 1996.