Journal of Machine Learning Research 17 (2016) 1-21 Submitted 3/16; Revised 9/16; Published 9/16

One-class classification of point patterns of extremes

Stijn Luca STIJN.LUCA@KULEUVEN.BE
KU Leuven - Technology Campus Geel
Department of Electrical Engineering
Kleinhoefstraat 4, 2440, Geel, Belgium

David A. Clifton DAVIDC@ROBOTS.OX.AC.UK
University of Oxford
Department of Engineering Science
Old Road Campus Research Building
Roosevelt Drive, Oxford, OX3 7DQ, UK

Bart Vanrumste BART.VANRUMSTE@KULEUVEN.BE
KU Leuven - Technology Campus Geel
Department of Electrical Engineering
Kleinhoefstraat 4, 2440, Geel, Belgium

Editor: Amos Storkey

Abstract

Novelty detection or one-class classification starts from a model describing some type of 'normal behaviour' and aims to classify deviations from this model as being either novelties or anomalies.

In this paper the problem of novelty detection for point patterns $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset \mathbb{R}^d$ is treated where examples of anomalies are very sparse, or even absent. The latter complicates the tuning of hyperparameters in models commonly used for novelty detection, such as one-class support vector machines and hidden Markov models.

To this end, the use of extreme value statistics is introduced to estimate explicitly a model for the abnormal class by means of extrapolation from a statistical model X for the normal class. We show how multiple types of information obtained from any available extreme instances of S can be combined to reduce the high false-alarm rate that is typically encountered when classes are strongly imbalanced, as often occurs in the one-class setting (whereby 'abnormal' data are often scarce).

The approach is illustrated using simulated data and then a real-life application is used as an exemplar, whereby accelerometry data from epileptic seizures are analysed - these are known to be extreme and rare with respect to normal accelerometer data.

Keywords: Sequence classification; novelty detection; extreme value theory; class imbalance; asymptotic theory

1. Introduction

Novelty detection is a particular example of pattern recognition that addresses the problem of identifying new patterns in data that are previously unseen. It shares many similarities with anomaly detection where one also wishes to detect abnormalities, but where in the latter these may not necessarily be entirely novel; i.e. a small amount of the training data

may contain outliers or anomalies. Novelty detection has a broad range of applications ranging from intrusion detection in computer related systems; industrial damage detection; to healthcare (Pimentel et al., 2014). All these applications have in common the fact that data describing failure conditions (or other abnormal behaviour) are rare or even absent, such that traditional classification methods may perform suboptimally. Novelty detection provides an alternative approach that starts from a model of normal behaviour and then detects deviations from this model (Bishop, 1994). It is for this reason that novelty detection is also termed one-class classification where there is no explicit model for 'abnormal behaviour'. It may also be described in terms of a hypothesis test, in which the null-hypothesis is described by the model of normality.

This article considers one-class classification of 'point patterns', defined as sets of vectors $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, $k \in \mathbb{N}_0$ located in data space $\mathbb{R}^d$ where each $\mathbf{x}_i$ is a realization of a random variable X[1]. We propose a statistical approach that starts from a probability density function (PDF) $y = p(\mathbf{x})$ associated with X that models the normal behaviour described by a dataset $\mathcal{D} \subset \mathbb{R}^d$. Novelty detection then addresses the question of whether a set S of vectors is drawn from the distribution X or not.

In this article the use of the use of extreme value theory (EVT) is introduced to tackle classification of sets S (Embrechts et al., 1997). The Poisson point process (PPP) characterization of EVT is used to extract count data describing the number of times measurements in S fall in low-density regions defined by X. Furthermore, asymptotic results are provided in this article that allow us to unify this count information with the mean and maximal excess in $p(S)$ with respect to a low threshold e^{-u}. The method is evaluated using synthetic as well as real-world data, and is compared with commonly used algorithms for outlier detection such as one-class support vector machines (OCSVMs) and hidden Markov models (HMMs).

In contrast to existing novelty detection methods, EVT enables us to define a model for the abnormal class, where data are sparse or even unobserved. This enables us to circumvent the optimization of hyperparameters that is typically encountered in using one-class classifiers and which often requires data from the abnormal class. In essence, the use of EVT relies on extrapolation from the normal class, providing a class of models for low-density regions; the latter are particularly beneficial for novelty detection, because the decision boundary is expected to be situated in regions where data are sparse.

The remainder of this paper is organized as follows. Section 2 is devoted to related work on sequence classifications and provides an introduction to EVT. Subsequently, Section 3 introduces the EVT-based one-class classifier. In Section 4, the method is evaluated and its limitations are discussed.

2. Related work and EVT

This section starts with a short review of related work on sequence classification. The necessary background of EVT is then reviewed.

1. The common convention in statistics is used that applies capital letters to refer to population attributes and lower-case letters to refer to sample attributes.

2.1 Related work

The problem setting in this article is an example of a collective novelty detection problem where the individual instances within a set S are not classified with respect to the distribution X. Instead, the entire set S of vectors is considered to be one single instance that is assigned a single label. This contrasts with conventional one-class classification, in which every element of S is classified independently. Closely related to this problem is that of sequential learning. However in the latter each instance of the set S is given a different label. Widely-used machine learning techniques for sequential learning, such as HMMs and conditional random fields (CRFs), are not able to learn from one class only (Bishop, 2006; Sutton and McCallum, 2011). A commonly-used technique to tackle sequence classification is to concatenate the separate labels that are obtained by applying a one-class classifier (e.g., an OCSVM) to each instance $\mathbf{x}_i$ separately. The mean novelty score of all instances, for example, can be used to decide whether or not S is novel (Dietterich, 2002). This latter approach, however, is more naturally expressed by taking a point-wise approach where, from a statistical point of view, a number (k) of hypothesis tests are considered:

H_0 *:* $\mathbf{x}_i$ *is a realization of* X
H_1 *:* $\mathbf{x}_i$ *is novelty with respect to* X*,*

where H_0 denotes the so-called null-hypothesis and H_1 the alternative hypothesis. Due to the multiple hypothesis-testing problem, the number of false alarms can increase considerably for $k > 1$. Indeed, while each hypothesis test is chosen to have a small type-I error α (i.e., the probability of wrongly classifying $\mathbf{x}_i$ as being novel, which is a false positive), the error of making at least least one type-I error among the k hypothesis tests corresponds to $\overline{\alpha} = 1 - (1 - \alpha)^k$; e.g., when $\alpha = 5\%$ and $k = 6$, $\overline{\alpha} = 26\%$.

To obtain the correct decision boundary corresponding to the significance level α, Clifton et al. (2011) considered the univariate distribution over the probability density values $p(\mathbf{x})$ on the image $\mathrm{Im}(p) = \{p(\mathbf{x}) \mid \mathbf{x} \in \mathcal{D}\}$ by reducing the multivariate analysis of the multidimensional data set $\mathcal{D}$ to an univariate analysis on $\mathrm{Im}(p)$. The PDF $y = p(\mathbf{x})$ can be obtained, for example, using a kernel density estimator (Scott, 1992). The distribution Y of these densities is strongly related to that of X, with a density defined by:

$$q(y) = \frac{\mathrm{d}Q}{\mathrm{d}y}(y) \quad \text{and} \quad Q(y) = \int_{p^{-1}(]0,y])} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1}$$

As will be made clear in the following section, univariate EVT can then be used to describe sets $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, which have a typical minimal density with respect to $y = p(\mathbf{x})$. In this way, a distribution is obtained for the most 'extreme' vectors that possibly occur in (truly 'normal') samples S drawn from X. A new set S it then evaluated by comparing its most extreme vector w.r.t. this model of extremes. Although this approach enables one to obtain a correct statistical type I-error α in testing S, its main drawback is that is captures limited information concerning the set S (Luca et al., 2014b). Indeed, only the single most extreme element in S is used to obtain a decision, while (non-extreme) information contained in the remaining part of the set is discarded. In this article we show how EVT can be used to include information contained in the remaining part of the pattern S while maintaining the correct statistical type I-error when testing S.

2.2 An introduction to EVT

EVT is a statistical discipline where the objective is to model the stochastic behavior of a univariate process at unusally large (or small) levels. It has already been used for many applications ranging from biomedical engineering, structural health monitoring, meteorology, and risk assessment in financial domains (Embrechts et al., 1997).

The central result in EVT is the Fisher-Tippett theorem concerning the limiting distribution of maxima of a sequence of independent and identically distributed (i.i.d.) random variables $X_1, \ldots X_k$ according to a common distribution X:

$$M_k = \max\{X_1, \ldots, X_k\},$$

as $k \to +\infty$. It states that when the following convergence in distribution appears:

$$P\left(\frac{M_k - c_k}{d_k} \leq x\right) \to G_\xi(x), \text{ as } k \to +\infty \tag{2}$$

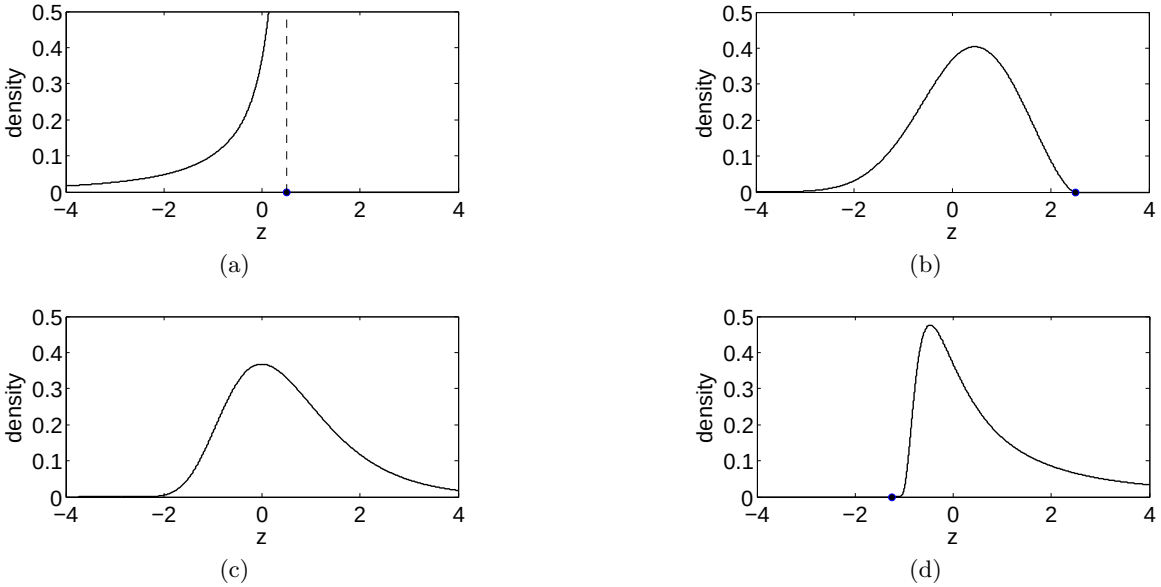


Figure 1: Different members of the GEV family in Eq. (3), with different values of the shape parameter ξ. The dot in the figures indicates the abscis $z = -\frac{1}{\xi}$, where the density is zero, (a) $\xi = -2$ where we see that when $\xi \leq -1$ a short tail with an upper bound is described (b) $\xi = -0.4$ where we see that when $-1 < \xi < 0$ maxima have an upper bound (c) $\xi = 0$ where the maxima have no upper- or lower bound. Finally, (d) $\xi = 0.8$ where we see that for $\xi > 0$ the maxima have a lower bound.

for some normalizing constants c_k, d_k, the limiting distribution $G_\xi(x)$ is a member of the so-called family of *generalized extreme value (GEV) distributions*:

$$G_\xi(x) = \begin{cases} \exp\left\{-\left[1+\xi x\right]^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp(-x)\right\}, & \xi = 0. \end{cases} \tag{3}$$

For $\xi \neq 0$ the domain of the distribution is restricted to the set $\{x \mid 1+\xi x > 0\}$. When the shape parameter ξ is negative, zero, or positive, the subset of members of the family correspond to the *Weibull, Gumbel* and *Fréchet* families respectively. The shape parameter thus determines the behaviour in the tail of the distribution of X, as shown in Figure 1.

The normalizing constants in (2) prevent a degenerate limit of the distribution of M_k, because clearly:

$$\lim_{k\to+\infty} P(M_k \leq x) = \lim_{k\to+\infty} \prod_{i=1}^{k} P(X_i \leq x)$$

which approaches zero for each $x < x_+$, where x_+ (possible $+\infty$) denotes the rightmost endpoint of the support of X.

The GEV family provides a model for block maxima, obtained by blocking (or windowing) the training data into blocks of equal length, and then fitting the GEV to the obtained

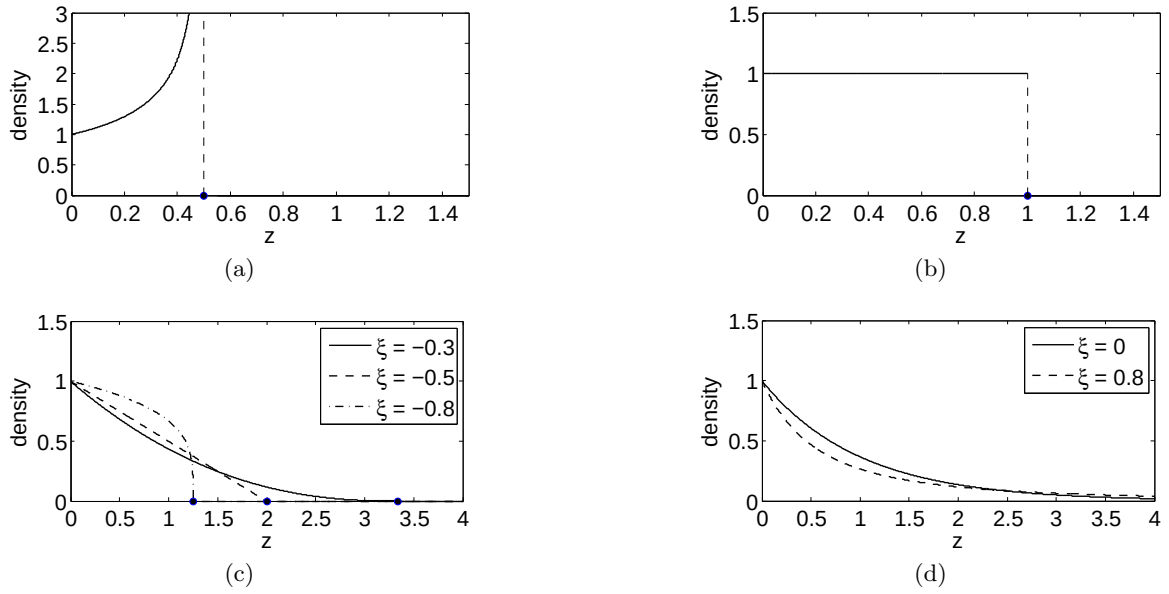


Figure 2: Different members of the GPD family in Eq. (3). The dot in the figures indicates the abscis $z = -\frac{1}{\xi}$, where the density is zero, (a) $\xi = -2$, where $\xi < -1$, an asymptote occurs at $z = -\frac{1}{\xi}$. (b) $\xi = -1$ corresponds to an uniform distribution of excesses. (c) Different types of behaviour for $-1 < \xi < 0$ corresponding to excesses with an upperbound. (d) For $\xi > 0$ the density has an intercept at $(0, 1)$.

set of block maxima. However, when these block are relatively large, this leads to using only a few block maxima, which can bias the estimation process. An alternative approach to overcome this problem is the so-called peaks over threshold (POT) method. In this approach, complete tails of a distribution X are modelled, defined as those measurements X_i of a sequence $X_1, X_2, \ldots$ that fall above some threshold u. A basic result of EVT states that when (2) holds for some member $G_\xi(x)$ of the GEV-family, the distribution of the exceedances $X - u$, conditional on $X > u$, satisfies the limiting property:

$$\lim_{u \uparrow x_+} P\left(\frac{X-u}{a(u)} < x \mid X > u\right) = H_\xi(x) \tag{4}$$

for some appropriate scaling factor $a(u)$ and

$$H_\xi = \begin{cases} 1 - (1+\xi x)^{-1/\xi} & \text{if} \quad \xi \neq 0 \\ 1 - e^{-x} & \text{if} \quad \xi = 0 \end{cases} \tag{5}$$

denotes the family of generalized Pareto distributions (GPDs) where $x \geq 0$ for $\xi \geq 0$ and $0 \leq x \leq -\frac{1}{\xi}$ for $\xi \leq 0$, as shown Figure 2. For the Gumbel case $\xi = 0$, the scaling factor $a(u)$ is given by $E(X - u|X > u)$.

2.3 Poisson point processes and EVT

An elegant way to describe extremes, and one that unifies the block and POT approaches is based on Poisson point processes (PPPs). Any inference made using one of both above approaches could equally be made using the PPP model because it can be parametrized in terms of the GEV- and GPD- parameters. In this way, no extra computational effort is needed when using the PPP model.

Generally a point process $\mathcal{P}$ on a subset $U \subset \mathbb{R}^d$ is a stochastic model for which any one realization consists of a set of points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}$ that are randomly located in U and of which the number N is a random variable. The point processes closely related to EVT are the point processes of exceedances and consider those observations from sequences of random variables $X_1, \ldots X_k$ which exceed a threshold u.

In particular, for a fixed choice of $k \in \mathbb{N}$, the point process of exceedances $\mathcal{P}_k$ is defined on regions of the form $U =]0,1[\times]u, +\infty[$ and considers those points that are situated in the intersection:

$$\mathcal{P}_k(\omega) = \left\{ \left(\frac{i}{k+1}, \frac{X_i(\omega) - c_k}{d_k}\right) \mid 1 \leq i \leq k \right\} \cap]0,1[\times]u, +\infty[, \tag{6}$$

where c_k and d_k are normalizing constants and ω denotes the stochastic event corresponding to a realization $\mathcal{P}_k(\omega)$ of the point process of exceedances. The indices are divided by the factor $k+1$ to rescale the process to the interval $]0,1[$, as illustrated in Figure 3. The point processes $\mathcal{P}_k$ can be characterised by *random counting measures*, which assign to each subset of the form $A = [t_1, t_2] \times]u + x, +\infty[\subset]0,1[\times]u, +\infty[$ a random variable N_A describing the number of points of a realization that fall in region A:

$$N_A^k : \quad \omega \quad \mapsto \quad \text{“number of points of } \mathcal{P}_k(\omega) \text{ in A”}$$

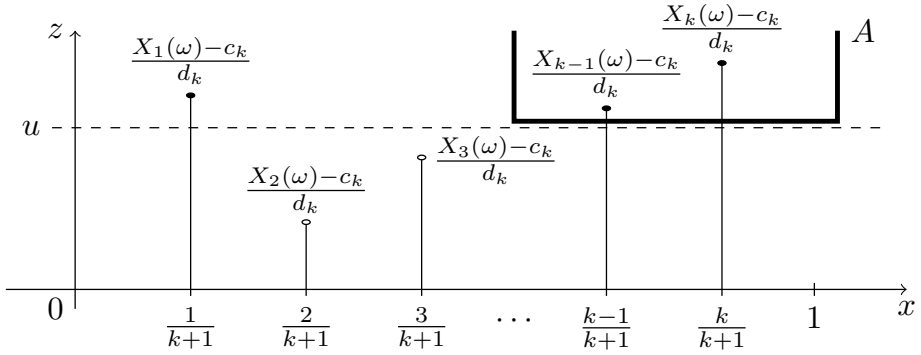


Figure 3: A realization $\mathcal{P}_k(\omega)$ of a point process of exceedances with $N_A^k(\omega) = 2$.

Indeed the values of these counting measures N_A^k for all subsets A give sufficient information to reconstruct completely those X_i that fall above a threshold of value $c_k + d_k u$. In fact, setting $A = \{\frac{i}{k+1}\} \times]z, +\infty[$, $N_A^k(\omega) > 0$ only applies when $X_i(\omega) > c_k + d_k z$.

The point process characterization of EVT is obtained by letting $k \to +\infty$. It is known (Embrechts et al., 1997) that when (2) holds for some normalization constants c_k and d_k, then the corresponding point processes of exceedances $\mathcal{P}_k$ will converge to a PPP $\mathcal{P}$ for $u > x_-$ where x_- denotes the leftmost endpoint of the support of the GEV-distribution in (2). This means that the following convergence of distributions holds:

$$N_A^k \xrightarrow{d} \text{Poi}\left[\Lambda(A)\right] \text{ as } k \to +\infty \tag{7}$$

on sets $A =]t_1, t_2[\times]u + x, +\infty[\subset U$ and where the distributions of N_A^k on non-overlapping sets A are mutually independent; i.e., the occurence of a point at a location should not influence the probability of the occurence of other points at other locations. In the limiting case, the rate parameter of the Poisson distribution $\Lambda(A)$ depends on the set A and is called the *intensity measure* $\Lambda(A)$ of the PPP. The fact that the PPP-characterization of extremes unifies the block and POT approach is due to the fact that the values of $\Lambda(A)$ in (7) can be written as a function of ξ (Embrechts et al., 1997):

$$\Lambda(A) = (t_2 - t_1)\left(1 + \xi(u + x)\right)^{-1/\xi} = (t_2 - t_1)\lambda\left(1 + \xi\lambda^{\xi} x\right)^{-1/\xi} \tag{8}$$

with $\lambda = (1 + \xi u)^{-1/\xi}$. Therefore any inference made using the PPP limit of extremes yields immediately the shape parameter ξ in (2) and (21). In this way EVT describes three equivalent limiting properties (2), (4), and (7).

3. Learning from sparse data regions

In this article, a learning algorithm is proposed that explores the link between the three representations of extremes as introduced in the previous section. For this purpose so-called EVT-based features will be introduced in section 3.1 that describe characterizing measures of a set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ of vectors independently drawn from a distribution X. In Section 3.2, a joint asymptotic distribution of these features is calculated as $k \to +\infty$. Subsequently,

analytical expressions of cumulative scores with respect to this distribution are obtained that will be used as novelty scores to evaluate the novelty of S with respect to X for large k.

3.1 EVT-based features

Consider a d-dimensional random variable X with PDF $y = p(\mathbf{x})$. The transformation $Z = -\log p(X)$ allows us to study multivariate low-density regions $\{\mathbf{x} \mid p(\mathbf{x}) < e^{-u}\}$, with u some large real number, as a convex univariate region $\{z \mid z > u\}$. Associated with a sequence of i.i.d. random variables $X_1, \dots X_k$, we define the following associated features based on the log-transformed sequence $Z_1, \dots Z_k$, $Z_i = -\log p(X_i)$:

1. The *number of exceedances* among $Z_1, \dots Z_k$ above some threshold u_k:

$$N_k = \sum_{i=1}^{k} \mathbb{I}_{\{Z_i > u_k\}},$$

   where $\mathbb{I}_{\{Z_i > u_k\}}$ denotes an indicator function taking the value 1 when $Z_i > u_k$ and zero otherwise. This feature describes the number of multivariate points from a sequence $\{X_1, \dots, X_k\}$ that are situated in a low density region $\mathcal{R}_k = \{\mathbf{x} \mid p(\mathbf{x}) < e^{-u_k}\}$.

2. The *mean exceedance* among $Z_1, \dots Z_k$ above some threshold u_k:

$$V_k = \frac{1}{N_k} \sum_{i=1}^{k} (Z_i - u_k) \mathbb{I}_{\{Z_i > u_k\}}$$

   A high value of V_k indicates that, on average, the points of the sequence $X_1, \dots X_k$ are outlying with respect to the locus of the training data while a low value indicates that the sequence is situated near the locus of the training data.

3. The *maximal exceedance* among $Z_1, \dots Z_k$ above some threshold u_k:

$$M_k = \max_{1 \le i \le k} \{Z_i - u_k \mid Z_i > u_k\}$$

   corresponding to the most outlying point of $X_1, \dots X_k$ with respect to to the training data.

Note that the mean exceedance V_k and the maximal exceedance M_k are only well-defined when $N_k \geq 1$. The features above provide a natural way to summarize the extent to which densities of observations falling in low-density regions exceed some low threshold e^{-u_k}. Therefore when a set $S = \{\mathbf{x}_1, \dots \mathbf{x}_k\}$ of k observations is novel with respect to the distribution X, it is expected that the corresponding features v_S, m_S, and n_S of the sample S have a higher cumulative score given their respective distributions V_k, M_k, and N_k. Hence these features allow us to summarize the information contained in the tail of a d-dimensional distribution X (that can be of arbitrarily high dimension) in a 3-dimensional distribution. To determine the joint distribution of these EVT-based features, the PPP characterization (7) is applied to the univariate random variable Z whose tail describes the

multivariate points X that are lying in low-density regions. In the next section we will determine the joint distribution of these 3 features to fuse the information from each.

To apply the PPP characterization to Z, we consider the sequence of point processes $\mathcal{P}_k$ on $\mathbb{R}^2$ associated with $Z = -\log p(X)$:

$$\mathcal{P}_k = \left\{ \left(\frac{i}{k+1}, Z_i \right) \mid 1 \leq i \leq k \right\}.$$

From the limiting property (7), the point processes $\mathcal{P}_k$ will converge to a PPP as $k \to +\infty$ on regions of the form $]0,1[\times]u_k, +\infty)$, with $u_k = c_k + ud_k$, $u \in \mathbb{R}$, and with c_k, d_k being the normalizing constants as in (6). Block maxima of Z_i are not bounded from above or below, and so the Gumbel distribution is the only possible limiting EVT distribution for this one-class formulation; i.e., $\xi = 0$ in the limiting property (7). For the Gumbel case it is known that the normalizing constants can be chosen as (Embrechts et al., 1997)[2]:

$$c_k = \inf \left\{ z \mid P(Z \leq z) \geq 1 - \frac{1}{k} \right\} \text{ and } d_k = E(Z - c_k | X > c_k). \tag{9}$$

The intensity measure of the limiting PPP can be obtained by letting $\xi \to 0$ in (8):

$$\Lambda(A) = (t_2 - t_1)e^{-(x+u)} = (t_2 - t_1)\lambda e^{-x}, \quad \text{with } \lambda = e^{-u} \tag{10}$$

and where the parameter λ is given by the expected number of exceedances of Z above $u_k(x) = c_k + (u + x)d_k$. We can now state the following theorem that is proved in Appendix A.1 and that characterizes the distribution of the EVT features defined above.

Theorem 1 *Consider the random variables N_k, V_k and M_k associated with sets S of k observations $\{X_1, \ldots, X_k\}$ drawn from a d-dimensional random variable X. Denote $y = p(\mathbf{x})$ the PDF of X and suppose $Z = -\log p(X)$ satisfies the following limiting property:*

$$\lim_{w \to +\infty} P\left(\frac{Z - w}{a(w)} > x \mid Z > w \right) = e^{-x}, \quad \forall x \in \mathbb{R}^+ \tag{11}$$

where $a(w) = E(X - w | X > w)$. Denoting, for $u \geq 0$, the following sequence of thresholds:

$$u_k = c_k + ud_k, \text{ with } c_k = \inf \left\{ z \mid P(Z \leq z) \geq 1 - \frac{1}{k} \right\}, \quad d_k = a(c_k),$$

the following limiting properties hold as $k \to +\infty$:

(i) The distribution N_k of the number of observations among k of X that fall in regions $\{\mathbf{x} \mid p(\mathbf{x}) < e^{-u_k}\}$ converges to a Poisson distribution with a rate $\lambda = e^{-u}$:

$$\lim_{k \to +\infty} P(N_k = n) = \frac{\lambda^n}{n!} e^{-\lambda} \tag{12}$$

2. The operator inf in (9) refers to the infimum or greatest lower bound.

(ii) After normalization, the distribution of the maximal exceedance M_k above threshold u_k converge in distribution to a Gumbel member of the GEV family with $\mu = \log\lambda$ that is conditioned on the positive real line; i.e.,

$$\lim_{k\to+\infty} P\left(\frac{M_k}{d_k} \leq m | N_k \geq 1\right) = \frac{\exp\left\{-\exp\left[-(m-\log\lambda)\right]\right\} - e^{-\lambda}}{1-e^{-\lambda}} \tag{13}$$

(iii) After normalization, the mean exceedance V_k above u_k converges in distribution to a random variable distributed according to a cumulative distribution function:

$$\lim_{k\to+\infty} P\left(\frac{V_k}{d_k} \leq v | N_k \geq 1\right) = 1 - \frac{1}{e^{\lambda}-1}\left(\sum_{l=1}^{+\infty}\sum_{j=0}^{l-1}\frac{\lambda^l}{l!j!}(lv)^j e^{-lv}\right) \tag{14}$$

Figure 4 illustrates the limiting properties obtained in Theorem 1 based on a two-dimensional distribution X given by a Gaussian mixture model (GMM) of two standard normal distributions centred at the origin and $(1,1)$ respectively. The constants c_k and d_k were estimated by an empirical estimation of (9) based on a simulated sample of length 5×10^6 from the mixture. Setting $u=0$, the empirical distributions of N_k, M_k and V_k were estimated based on 5×10^3 sets of lengths $k\in\{5,20,50\}$ and compared with the analytical expression obtained in Theorem 1. The figure shows that the distributions are approximating the limiting case more closely as k increases, while for $k\geq20$ this approximation may already be seen to be satisfactory.

3.2 EVT-based one-class classifier

A joint distribution is here calculated to fuse the information from the EVT-based features M_k, N_k, and V_k, as introduced in Section 3.1. For this purpose, we suppose that at least one exceedance of $-\log p(X_i)$ above u_k is observed in a sequence $S=\{X_1,\ldots,X_k\}$ of length $|S|=k$. The proof of the following theorem can be found in Appendix A.2.

Theorem 2 *Consider the random variables N_k, V_k, and M_k associated with sets S of k observations $\{X_1,\ldots,X_k\}$ drawn from a d-dimensional random variable X. Denote $y = p(\mathbf{x})$ the density function of X and suppose $Z=-\log p(X)$ satisfies the following limiting property:*

$$\lim_{w\to+\infty} P\left(\frac{Z-w}{a(w)} > x \mid Z > w\right) = e^{-x}, \quad \forall x\in\mathbb{R}^+ \tag{15}$$

where $a(w) = E(Z-w|Z>w)$. After normalization, the joint cumulative distribution function of (N_k, V_k, M_k) conditioned on $N_k\geq1$ and related to the sequence of thresholds u_k as in Theorem 1:

$$F_k(v,m,n) = P\left(\frac{V_k}{d_k}\leq v, \frac{M_k}{d_k}\leq m, N_k\leq n \mid N_k\geq1\right), \tag{16}$$

converges on $D=\{(v,m,n) \mid \frac{m}{n}\leq v\leq m\}$ to a mixture of translated chi-squared distribution as k tends to infinity:

$$F(v,m,n) = \lim_{k\to+\infty} F_k(v,m,n) = \sum_{l=1}^{n}\frac{\lambda^l e^{-\lambda}}{l!(1-e^{-\lambda})}\sum_{i=0}^{r}(-1)^i\binom{l}{i}e^{-im}\chi_{2l}\left(2(lv-im)\right) \tag{17}$$

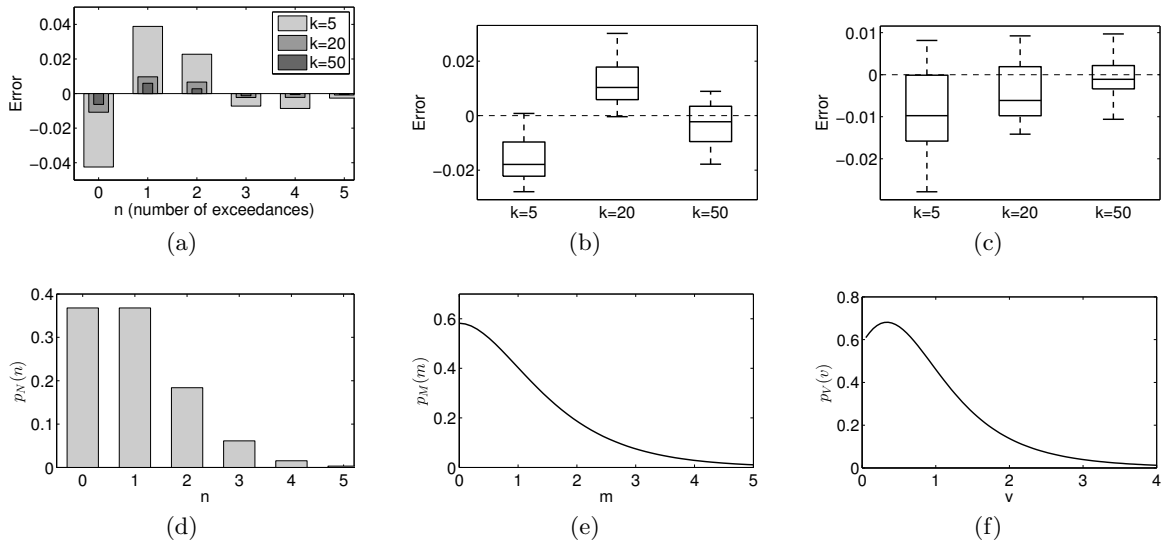


Figure 4: Comparison between limiting distribution as $k \to +\infty$ and empirical distribution functions for $k \in \{5, 20, 50\}$ using simulated data from a GMM when $u = 0$. (a) - (c) Differences between empirical distribution and asymptotic distribution for N_k, M_k, and V_k respectively. (d) - (f) Limiting PDFs p_N, p_M, and p_V from Eqns. (12) - (14) as $k \to +\infty$ for N_k, M_k, and V_k respectively.

where $r = \lfloor \frac{lv}{m} \rfloor$ *(i.e.* $\frac{lv}{m} \in [r, r+1[$*, for* $0 \leq r \leq l-1$*),* χ_p *denotes the cumulative chi-squared distribution function with* p *degrees of freedom and* $\lambda = e^{-u}$ *is the exceedance rate of the limiting Poisson distribution of* N_k *as in Theorem 1-(i).*

Note that the term in (17) for $l = 1$ has the identity line $m = v$ as its domain and the expression reduces to $\frac{\lambda e^{-\lambda}}{1-e^{-\lambda}}(1 - e^{-m})$. The corresponding limiting joint density function of (N_k, V_k, M_k) on D can be found by partial derivation of formula (17):

$$f(v, m, n \mid n \geq 1) = \begin{cases} e^{-nv} \sum_{i=1}^{\lfloor \frac{nv}{m} \rfloor} c_{in}(nv - im)^{n-2} & , n \geq 2 \\ \frac{\lambda}{e^{\lambda}-1} e^{-m} \mathbb{I}_{v=m} & , n = 1 \end{cases} \tag{18}$$

where c_{in} are constants defined for $1 \leq i \leq n$ as:

$$c_{in} = -\frac{n\lambda^n}{(e^{\lambda} - 1)\Gamma(n)\Gamma(n-1)}(-1)^i \binom{n-1}{i-1}.$$

and where $\mathbb{I}_{v=m}(v, m)$ is an indicator function taking the value 1 when $v = m$, and which is zero elsewhere.

To apply Theorem 2, note that (15) implies that an exponential approximation of the exceedances is valid from some high threshold u_0:

$$P(Z - u_0 > x | Z > u_0) \approx e^{-\frac{x}{\sigma}} \tag{19}$$

with $\sigma = a(u_0) = E(Z - u_0|Z > u_0)$ and $\sigma \approx d_k$. Then, based on Theorems 1 and 2, a novelty score of a sequence S with corresponding EVT features (v_S, m_S, n_S) can be defined:

$$\chi_S = \begin{cases} P(N_k < n_S) + P(V_k \leq v_S, M_k \leq m_S, N_k = n) & \text{when} \quad n_S > 0 \\ P(N_k = 0) & \text{when} \quad n_S = 0 \end{cases}$$

and for large k this is approximated by:

$$\chi_S \approx \begin{cases} \left(\sum_{l=0}^{n_S - 1} \lambda^l \frac{e^{-\lambda}}{l!} \right) + F(\frac{v_S}{\sigma}, \frac{m_S}{\sigma}, n_S) - F(\frac{v_S}{\sigma}, \frac{m_S}{\sigma}, n_S - 1) & \text{when} \quad n_S > 0 \\ e^{-\lambda} & \text{when} \quad n_S = 0 \end{cases} \tag{20}$$

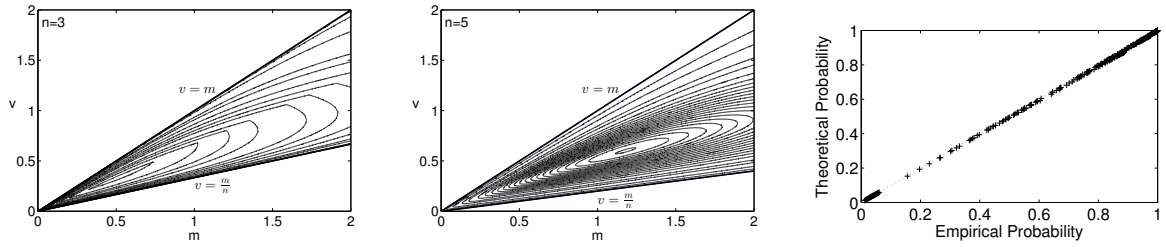
These novelty scores quantify the 'extremity' of a sequence S by cumulatively summing the probability of having fewer than n_S exceedances, while the mean and maximal exceedances with respect to the threshold u_0 do not exceed v_S and m_S respectively. There is a valid probabilistic interpretation to χ_S making it a risk metric that quantifies the risk that S is novel; i.e., that S has some distribution other than X.

The choice of u_0 in the approximation (19) can be assessed by means of a *mean excess plot* which is a graphic diagnostic in which the sample means of the excesses $(Z - u)$ are plotted against a range of thresholds along with the confidence intervals (Embrechts et al., 1997). The threshold is chosen to be the lowest level where all the higher threshold-based sample mean excesses are consistent with a horizontal line. Alternatively an empirically driven rule-of-thumb can be chosen that specifies the tail fraction which satisfies the approximation in (19) and where u_0 is estimated as the quantile at $1 - \frac{n^{2/3}}{n \log \log(n)}$ of a sample of length n of the distribution (Scarrot and MacDonald, 2012). The parameters σ and λ can then be estimated by means of maximum likelihood estimation (Falk et al., 2011).

Figure 5(a)-(b) illustrates the limiting joint PDF (18) on the domain D conditioned on the number of exceedances for $n = 3$ and $n = 5$ for a GMM X of two standard normal distribution centred at $(0, 0)$ and $(1, 1)$. As the number of exceedances increases, the mode of the distributions moves diagonally upwards. Figure 5(c) shows a probability-probability (P-P) plot assessing the limiting property (17) for $k = 20$. For this purpose a sample of 5×10^3 sets of length $k = 20$ were simulated from X to estimate the cumulative probabilities $F_k(v, m, n)$ empirically, on a grid of $(v, m, n) \in [0, 10] \times [0, 10] \times \{2, 3, 5\}$ consisting of 300 vertices and compare these estimations with $F(v, m, n)$.

4. Experiments

In this section, the validity of our proposed method is illustrated using both artificial and real-world data sets. The novel EVT algorithm is compared with the conventional sequence classifiers HMMs and OCSVMs. To this end, 5-fold cross-validation is performed where in each run a random subset of the data from the normal class is used for training and the remainder of the data is split evenly between validation and test data. The randomized runs are kept the same across the different classifiers to allow a consistent comparison. The novelty score of a sequence with respect to a HMM or OCSVM is calculated as being the mean of the likelihoods assigned by the model to each individual instance of the sequence.



(a) (b) (c)

Figure 5: (a) - (b) Limiting joint PDF (18) on the domain D conditioned on $n = 3$ and $n = 5$ respectively, for a GMM X consisting of two standard normal distribution centred at $(0, 0)$ and $(1, 1)$. (b) A probability-probability (P-P)-plot comparing the joint empirical cumulative distribution of (V_k, M_k, N_k) for $k = 20$ with the limiting joint distribution.

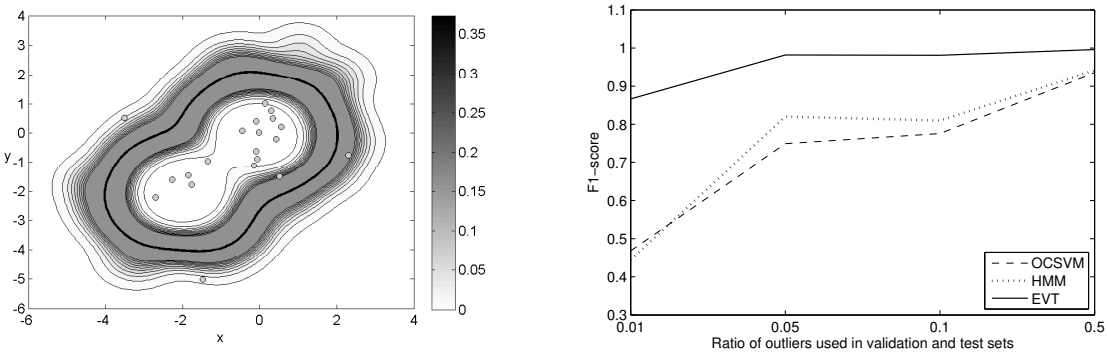
Both HMMs and OCSVMs depend on hyperparameters, the value of which are estimated using the validation sets by maximizing a cost-function. For the HMM, the number of states varies from $1 - 4$ (Rabiner and Murray, 1989), while for the OCSVM the standard hyperparameters (σ, ν) are optimized that respectively denote the kernel width of the Gaussian kernel that is used and an upper bound on the fraction of outliers (Schölkopf et al., 2001). The threshold on the novelty scores is optimized using the validation data.

For the EVT model, no validation step is performed and no data from the abnormal class are considered during training. A threshold of 95% is chosen on the novelty score (motivated from a probabilistic viewpoint). The density of the distribution X describing the normal class is estimated using a kernel density estimation with Gaussian kernels, and where the kernel width is estimated by minimization of the mean integrated squared error (Scott, 1992).

4.1 Synthetic data set

In order to validate the use of our EVT-based method a simulated dataset is constructed where data from the abnormal class are situated in the tail regions of a planar Gaussian mixture X consisting of two components located at $(-2, -2)$ and $(0, 0)$ repsectively with covariance matrix $\frac{1}{2}I_2$, with I_2 the identity matrix in $\mathbb{R}^{2\times 2}$. The training data of the normal class consisted of 100 sets of length $k = 20$ points drawn from X. Several experiments were performed where the proportion of abnormal instances in the validation and test sets varied in the range $p_a \in \{0.01, 0.05, 0.1, 0.5\}$. The abnormal class of patterns contained a mixture of normal instances from X and abnormal instances coming from the tail region where the density $p(\mathbf{x}) \leq 5 \times 10^{-4}$. In a 5-fold cross-validation experiment, the ability of the detection of these patterns between an OCSVM, a HMM, and our EVT model is compared.

Figure 6(a) shows the contours of the tail region obtained from applying the Gumbel model of M_k on the densities that are estimated using a kernel density estimation of X. The dark contour surrounding the central region indicates the tail region estimated by the Gumbel model. In this region, the dark contour corresponds to an empirical estimation of



(a) (b)

Figure 6: (a) The estimation of the tails of a Gaussian mixture X using a Gumbel model on the distribution of densities (1). The bold contour indicates the estimation of the EVT threshold e^{-u_k} for $k = 20$ on the likelihoods as defined in Theorem 1. (b) F1-scores averaged over the runs of a 5-fold-validation experiment across different ratios of available abnormalities.

the threshold $u_k = c_k + ud_k$ where u is set to zero (see Theorem 1). It is with respect to this threshold that the number of exceedances N_k and the maximal and mean exceedance M_k and V_k are calculated. Using our EVT-based method method, an abnormal sequence can be evaluated as a cumulative probability score (20) with respect to the joint distribution of the EVT-based features. For example, the sequence of gray points shown in Figure 6 contains three exceedances with respect to the threshold u_k and has a score $\chi_S = 98.97\%$ such that it is classified as being novel with respect to X. Figure 6(b) shows the F1-scores of the classifiers, averaged over the 5 folds in our cross-validation experiment. When the ratio of abnormal patterns in the training phase is 50% the classifiers perform equally well. EVT, however, is able to outperform the classifiers when data from the abnormal class become sparse, as is typically the case for novelty detection problems. When there is a lack of examples from the abnormal class, the optimization of the hyperparameters and the novelty threshold in a HMM and an OCSVM is suboptimal. EVT, on the other hand, provides a class of models for the tail region where training data are sparse and is able to estimate the threshold exactly by using a statistical distribution that is obtained by extrapolation from the normal class (where data are usually abundant).

4.2 Accelerometer data for the detection of epileptic seizures

In this section, a case study in the healthcare domain is considered using a set of acceleration data collected from movements of patients suffering from epilepsy (Cuppens et al., 2013). The acceleration data were recorded during several nights using four 3D acceleration sensors attached to the extremities of 7 children with hypermotor seizures, all between the age of 5 and 16 years. Hypermotor seizures are epileptic convulsions that are marked by a strong and uncontrolled movement of the arms and legs that can last from a couple of seconds to a number of minutes. Due to the exaggerated movement involved, the patient can injure

themselves during the seizure, which increases the need for an alarm system with high sensitivity to abnormality.

In a pre-processing phase, movement events E_s are extracted from the data set using an energy-based threshold. We denote the acceleration vectors in these events as

$$E_s = \{\vec{a}_{tl} | 1 \leq t \leq T, 1 \leq l \leq 4\}$$

where the indices refer to the time index and the limb respectively (1 = left arm, 2 = right arm, 3 = left leg, 4 = right leg). Cuppens et al. (2013) performed a feature analysis where 3 features were identified as being relevant to this application:

i) Movement length, $f_1 = |E_s| = T$

ii) Average energy in a movement:

$$f_2 = \frac{1}{T} \sum_{t,l} \|\vec{a}_{tl}\|^2$$

iii) The maximal energy in an arm movement:

$$f_3 = \max_{1 \leq t \leq T} \left\{ \|\vec{a}_{t1}\|^2, \|\vec{a}_{t2}\|^2 \right\}$$

The features are calculated within sliding windows containing 125 samples (Luca et al., 2014a) which are randomly subsampled to obtain sets $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ of fixed length $k = 20$ containing data instances $\mathbf{x}_i = (f_1^i, f_2^i, f_3^i) \in \mathbb{R}^3$ on which the EVT algorithm for sequence classification can be applied.

The data are highly unbalanced as may be seen in Table 1. Only three patient recordings contain more than 3 examples of seizures. For these patients, an OCSVM and HMM were trained in a 5-fold cross-validation experiment where in each fold the seizures are randomly split between validation and test sets to optimize the following cost-function (Cuppens et al., 2013):

$$C(\boldsymbol{\lambda}) = 2 \cdot SS(\boldsymbol{\lambda}) + PPV(\boldsymbol{\lambda})$$

with respect to the hyper-parameters $\boldsymbol{\lambda}$ of the model. Here, the weight of the sensitivity (SS) is higher than the weight of the positive predictive value (PPV), because missing a

Table 1: Overview of epileptic accelerometry data set.

Patient number	Nights of monitoring	Hypermotor seizures	Normal movements
pat 1	1	2	117
pat 2	2	9	287
pat 3	2	2	439
pat 4	1	2	239
pat 5	5	26	784
pat 6	2	7	381
pat 7	2	3	468
total	*15*	*51*	*2715*

Table 2: SS and PPV scores of different approaches used in the detection of epileptic seizures (a) OCSVM, (b) HMM, and (c) EVT. Mean and standard deviations (SD) are calculated over the folds in a 5-fold-cross-validation experiment.

OCSVM	SS		PPV		F1	
	mean	SD	mean	SD	mean	SD
pat2	100.0	0.00	48.03	13.19	64.07	11.62
pat5	64.62	18.53	34.08	2.68	43.59	2.22
pat6	100.00	0.00	31.85	7.20	47.96	8.14

(a)

HMM	SS		PPV		F1	
	mean	SD	mean	SD	mean	SD
pat2	70.00	20.92	89.33	15.35	76.83	14.09
pat5	56.92	8.77	46.57	16.75	49.71	10.40
pat6	80.00	29.81	85.00	13.69	77.43	15.53

(b)

EVT	SS		PPV		F1	
	mean	SD	mean	SD	mean	SD
pat2	100.0	0.00	69.65	21.38	80.63	14.75
pat5	35.38	10.32	21.80	2.73	26.80	4.95
pat6	100.0	0.00	48.21	15.15	64.05	12.34
pat1	100.0	0.00	19.68	9.55	32.05	13.08
pat3	100.0	0.00	56.67	25.28	70.00	18.26
pat4	100.0	0.00	48.33	30.28	61.33	23.64
pat7	100.0	0.00	66.67	31.18	76.67	22.36

(c)

seizure is more costly than generating a false-positive classification for this type of seizure. Tables 2(a) and 2(b) show the mean performance scores calculated over the different test sets in the runs for three patients of which more than 3 examples of seizures were available for the training of these models. As there are at most 3 seizures present for the remaining patients, at most two seizures could be used in the validation set when training the HMMs and OCSVMs. In this way at most one of the seizures could be held out and detected by the algorithms during the different cross-validation experiments.

Table 2(c) shows performance scores related to the EVT approach. In contrast to the OCSVM and HMM, performance scores could easily be obtained for all patients without the need for optimization using validation data. As hypermotor seizures are marked by strong and uncontrolled movements, the use of EVT is very suitable in this application to recognize this type of 'extremity' from the class of normal movement events. In contrast to an OCSVM our EVT-based method was able to improve PPV values in patients 2 and 6 (averaged over the folds, a decrease of 3 false alarms while testing 50 normal movements was obtained) while the SS scores remained 100%. The OCSVM was able to outperform the EVT method for patient 5. This is mainly due to (i) the seizures for this patient are less extreme than in the rest of the data (Cuppens et al., 2013); (ii) a sufficient amount of seizures is present giving the OCSVM the ability to perform a thorough optimization of the hyperparameters during the training phase. A HMM was not able to detect all seizures and obtained better PPV values compared to our EVT-based method.

5. Conclusion

This article focuses on the problem of novelty detection, where data instances from the normal class are abundant but where examples from the abnormal class are sparse. In particular a new approach is introduced that is based on the use of EVT and which is particularly well-suited to detecting outliers that present 'extreme' behaviour with respect to a statistical model X. It is shown how EVT can be adapted to define a model over

regions where data are sparse (or even unavailable) circumventing the need for optimization of hyperparameters as otherwise occurs when using conventional OCSVMs or HMMs. This leads to a more robust and exact estimation of the support of X when abnormal data are limited in availability.

One of the main challenges in novelty detection is to improve the PPV. Indeed, when classes are highly unbalanced, an unusually high accuracy is required to overcome a high false-alarm rate. Therefore rich models that combine several types of information in a natural way are needed to increase the PPV of a novelty detector. An estimation procedure from EVT is proposed that encodes the three different types of EVT-based information for a sequence S. Given a treshold u and an estimation $y = \hat{p}(\mathbf{x})$ of the density of X, the following types of information were fused: (i) the maximal exceedance of $-\log p(S)$ above u; (ii) the mean exceedance of $-\log p(S)$ above u; and (iii) the number of exceedances of $-\log p(S)$ above u.

We have demonstrated the use of this method on both artificial data and a real-world set of acceleration data collected from movements of patients that suffer from epilepsy. By applying the proposed method, it was shown that SS scores and PPV scores could be improved compared to the use of conventional HMMs and OCSVMs, especially when examples from the abnormal class are sparse.

Acknowledgments

Special thanks go to Peter Karsmakers for the fruitful discussions concerning the validation steps and preprocessing study of the epileptic seizure data. This data set is collected in collaboration with the Pulderbos rehabilitation Center for Children and Youth in Zandhoven (Pulderbos), Belgium and the assistance of Berten Ceulemans, Lieven Lagae, Anouk Van de Vel and Sabine Van Huffel in the framework of an IWT TBM project 100404. The authors would also like to acknowledge networking support by the ICT COST action IC1303 (AAPELE). David A. Clifton is funded by the Royal Academy of Engineering and an EPSRC Healthcare Technologies Challenge Award.

Appendix A. Proofs

In this appendix we prove the results obtained in Section 3.

A.1 Proof of Theorem 1

Proof In terms of the normalized sequence of random variables $\frac{Z-c_i}{d_i}$, it can be shown that (11) is equivalent to:

$$\lim_{i\to+\infty} P\left(\frac{Z-c_i}{d_i} < u+x \middle| \frac{Z-c_i}{d_i} > u\right) = 1-e^{-x} \tag{21}$$

with $u \in \mathbb{R}$ and $x \geq 0$ (Falk et al., 2011, p.21). The statements (i)-(iii) can now be proven as follows.

(i) This result follows by applying the link between the limiting properties (2), (4) and (7)

on the transformed variable $Z = -\log(p(X))$ as discussed in Sections 2.2 and 2.3. The exceedance rate of the PPP can be found by calculating the limit:

$$\begin{aligned}
\lim_{k\to+\infty} kP(Z \geq c_k + d_k u) &= \lim_{k\to+\infty} \frac{P(Z \geq c_k + d_k u)}{P(Z > c_k)}, \quad \text{as } P(Z \leq c_k) = 1 - \frac{1}{k} \\
&= \lim_{k\to+\infty} P(Z \geq c_k + d_k u | Z > c_k) \\
&= \lim_{k\to+\infty} P\left(\frac{Z - c_k}{a(c_k)} \geq u | Z > c_k\right), \quad \text{as } d_k = a(c_k) \\
&= \lim_{w\to+\infty} P\left(\frac{Z - w}{a(w)} \geq u | Z > w\right), \quad \text{as } \lim_{k\to+\infty} c_k = +\infty \\
&= e^{-u}
\end{aligned}$$

(ii) The limiting distribution of the maximal exceedance M_k conditioned on the number of exceedances $N_k \geq 1$ is obtained as:

$$\begin{aligned}
\lim_{k\to+\infty} P\left(\frac{M_k}{d_k} \leq m | N_k = l\right) &= \lim_{k\to+\infty} P\left(\frac{Z - u_k}{d_k} \leq m \middle| Z > u_k\right)^l \\
&\underset{u_k = c_k + u d_k}{=} \lim_{k\to+\infty} P\left(\frac{Z - c_k}{d_k} - u \leq m \middle| \frac{Z - c_k}{d_k} > u\right)^l \\
&= (1 - e^{-m})^l. \qquad (22)
\end{aligned}$$

where we used (21). The distribution of M_k is found by marginalization over the number of excesses $1 \leq l \leq k$ conditionned on $N_k \geq 1$. From (i) one finds:

$$\begin{aligned}
\lim_{k\to+\infty} P\left(\frac{M_k}{d_k} \leq m | N_k \geq 1\right) &= \lim_{k\to+\infty} \sum_{l=1}^{k} P\left(\frac{M_k}{d_k} \leq m | N_k = l\right) P(N_k = l | N_k \geq 1) \\
&= \frac{1}{1 - e^{-\lambda}} \sum_{l=1}^{+\infty} (1 - e^{-m})^l \left(\frac{\lambda^l}{l!} e^{-\lambda}\right)
\end{aligned}$$

Further simplification leads to:

$$\begin{aligned}
\lim_{k\to+\infty} P\left(\frac{M_k}{d_k} \leq m | N_k \geq 1\right) &= \frac{e^{-\lambda}}{1 - e^{-\lambda}} \sum_{l=1}^{+\infty} \frac{(\lambda(1 - e^{-m}))^l}{l!} \\
&= \frac{e^{-\lambda}}{1 - e^{-\lambda}} \left[\exp\left\{\lambda - \lambda e^{-m}\right\} - 1\right] \\
&= \frac{\exp\left\{-\exp\left[-(m - \ln\lambda)\right]\right\} - e^{-\lambda}}{1 - e^{-\lambda}}
\end{aligned}$$

which is the cumulative distribution function of a Gumbel member of the family (3) located at $\mu = \ln\lambda$ and conditioned on the positive real line.

(iii) From (21) it follows that the excesses $\frac{Z - c_i}{d_i} - u$ converge in distribution to an exponential distribution as $i \to +\infty$. Therefore, from the continuous mapping theorem (stating that

convergence is preserved by continuous transformation (Embrechts et al., 1997, p. 561)), the mean of n such independent excesses converges to the distribution of a mean of n independent variables that are distributed according to an exponential distribution. Thus the limiting distribution conditioned on $N_k = l \geq 1$ is given by an Erlang distribution with shape-parameter l and rate parameter l (Feller, 1971, p. 11) with a cumulative distribution function:

$$\lim_{k\to+\infty} P\left(\frac{V_k}{d_k} \leq v | N_k = l\right) = 1 - \sum_{j=0}^{l-1} \frac{1}{j!} (lv)^j e^{-lv}$$

Marginalisation over the number of exceedances leads to:

$$\begin{aligned}
\lim_{k\to+\infty} P\left(\frac{V_k}{d_k} \leq v | N_k \geq 1\right) &= \lim_{k\to+\infty} \sum_{l=1}^{k} P\left(\frac{V_k}{d_k} \leq v | N_k = l\right) P(N_k = l | N_k \geq 1) \\
&= \sum_{l=1}^{+\infty} \left(1 - \sum_{j=0}^{l-1} \frac{1}{j!} (lv)^j e^{-lv}\right) \left(\frac{\lambda^l}{l!} \frac{e^{-\lambda}}{1 - e^{-\lambda}}\right) \\
&= \frac{1}{e^{\lambda} - 1} \left((e^{\lambda} - 1) - \sum_{l=1}^{+\infty} \sum_{j=0}^{l-1} \frac{\lambda^l}{l!j!} (lv)^j e^{-lv}\right) \\
&= 1 - \frac{1}{e^{\lambda} - 1} \left(\sum_{l=1}^{+\infty} \sum_{j=0}^{l-1} \frac{\lambda^l}{l!j!} (lv)^j e^{-lv}\right)
\end{aligned}$$

■

A.2 Proof of Theorem 2

Proof Convergence in distribution is expressed in terms of the joint (cumulative) distribution of the features V_k, M_k and N_k conditioned on $N_k \geq 1$:

$$F_k(v, m, n) = P\left(\frac{V_k}{d_k} \leq v, \frac{M_k}{d_k} \leq m, N_k \leq n \mid N_k \geq 1\right). \tag{23}$$

Clearly, the mean v of a sequence of n positive numbers is situated between $\frac{m}{n}$ and m such that the support of F_k is situated in $D = \{(v, m, n) \mid \frac{m}{n} \leq v \leq m\}$. The conditioned joint distribution (23) can be written as:

$$\begin{aligned}
F_k(v, m, n) &= \sum_{l=1}^{n} \frac{P\left(\frac{V_k}{d_k} \leq v, \frac{M_k}{d_k} \leq m, N_k = l\right)}{1 - P(N_k = 0)} \\
&= \sum_{l=1}^{n} \frac{P\left(\frac{V_k}{d_k} \leq v \mid \frac{M_k}{d_k} \leq m, N_k = l\right) P\left(\frac{M_k}{d_k} \leq m \mid N_k = l\right) P(N_k = l)}{1 - P(N_k = 0)}
\end{aligned} \tag{24}$$

The limiting distribution of (23) can be obtained by considering the limit of each factor in the nominators of the terms in (24) as $k \to +\infty$. Firstly, from Theorem 1-(i), it follows

that:

$$\lim_{k\to+\infty} P(N_k = l) = \frac{\lambda^l e^{-\lambda}}{l!}, \qquad \lambda = e^{-u}. \tag{25}$$

Secondly, the limiting distribution of $P\left(\frac{M_k}{d_k} \leq m \mid N_k = l\right)$ is given by (22). Thirdly, the distribution $P\left(\frac{V_k}{d_k} \leq v \mid \frac{M_k}{d_k} \leq m, N_k = l\right)$ corresponds to the distribution of the mean of l independent exceedances that each converge in distribution to an exponential distribution truncated at m:

$$\begin{aligned}\lim_{k\to+\infty} P\left(\frac{Z-u_k}{d_k} \leq v \mid \frac{Z-u_k}{d_k} \leq m\right) &= \lim_{k\to+\infty} P\left(\frac{Z-c_k}{d_k} - u \leq v \mid \frac{Z-c_k}{d_k} - u \leq m\right) \\ &= \frac{1-e^{-v}}{1-e^{-m}}.\end{aligned}$$

Therefore, according to the continuous mapping theorem (Embrechts et al., 1997), the distribution of lV_k converge in distribution to the sum of l truncated exponential distributions such that (Bain and Weeks, 1964):

$$\lim_{k\to+\infty} P\left(\frac{V_k}{d_k} \leq v \mid \frac{M_k}{d_k} \leq m, N_k = l\right) = \frac{1}{(1-e^{-m})^l} \sum_{i=0}^{r} (-1)^i \binom{l}{i} e^{-im} \chi_{2l}\left(2(lv - im)\right)$$

for $r = \lfloor \frac{lv}{m} \rfloor$. Substituting the latter expression together with (22) and (25) in the factorisation (24) gives the desired result. ■

References

L.J. Bain and D.L. Weeks. A note on the truncated exponential distribution. *The Annals of Mathematical Statistics*, 35(3):1366–1367, 1964.

C.M. Bishop. Novelty detection and neural network validation. In *Proceedings of the IEEE Conference on Vision, Image and Signal Processing*, volume 141, pages 217–222. IEE, London, 1994.

C.M. Bishop. *Pattern Recognition and machine learning*. Springer, New York, USA, 2006.

D.A. Clifton, S. Hugueny, and L. Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of Signal Processing Systems*, 65:371–389, 2011.

K. Cuppens, P. Karsmakers, A. Van de Vel, B. Bonroy, M. Milosevic, S. Luca, B. Ceulemans, L. Lagae, S. Van Huffel, and B. Vanrumste. Accelerometer based home monitoring for detection of nocturnal hypermotor seizures based on novelty detection. *IEEE Journal of Biomedical and Health Informatics*, In Press, 2013.

T.G. Dietterich. Machine learning for sequential data: A review. In *Proceedings of the Joint International Workshop on Structural Syntactic and Statistical Pattern Recognition*, pages 15–30. Springer-Verlag, Londen, 2002.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, 1997.

M. Falk, J. Hüsler, and R.-D. Reiss. *Laws of small numbers: Extremes and rare events*. Birkhäuser, 3rd edition, 2011.

W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 2*. Wiley, New York, 2nd edition, 1971.

S. Luca, P. Karsmakers, K. Cuppens, T. Croonenborghs, A. Van de Vel, B. Ceulemans, L. Lagae, S. Van Huffel, and B. Vanrumste. Detecting rare events using extreme value statistics applied to epileptic convulsions in children. *Journal of Artificial Intelligence In Medicine*, 60(2):89–96, 2014a.

S. Luca, P. Karsmakers, and B. Vanrumste. Anomaly detection using the Poisson process limit for extremes. In R. Kumar, H. Toivonen, J. Pei, Zhexue H., and X. Wu, editors, *IEEE International Conference on Data Mining*, pages 370–379, 2014b.

Marco A. F. Pimentel, D.A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215 – 249, 2014.

L.R. Rabiner and H. Murray. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257 – 286. IEEE, 1989.

C. Scarrot and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical journal*, 10(1):33–60, 2012.

B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley and Sons, New York, 1992.

C. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2011.