Anomaly detection using the Poisson process limit for extremes

Stijn Luca, Peter Karsmakers and Bart Vanrumste
Department of Electrical Engineering
iMinds Medical Information Technology Department
KU Leuven, Kasteelpark Arenberg 10 B-3001 Leuven, Belgium
Email: stijn.luca@kuleuven.be

***Abstract*—Anomaly detection starts from a model of normal behavior and classifies departures from this model as anomalies. This paper introduces a statistical non-parametric approach for anomaly detection that is based on a multivariate extension of the Poisson point process model for univariate extremes. The method is demonstrated on both a synthetic and a real-world data set, the latter being an unbalanced data set of acceleration data collected from movements of 7 pediatric patients suffering from epilepsy that is previously studied in [1]. The positive predictive values could be improved with an increase up to 12.9% (and a mean of 7%) while the sensitivity scores stayed unaltered. The proposed method was also shown to outperform an one-class SVM classifier. Because the Poisson point process model of extremes is able to combine information on the number of excesses over a fixed threshold with that on the excess values, a powerful model to detect anomalies is obtained that can be of high value in many application.**[1]

***Keywords*-anomaly detection; extreme value statistics; Poisson point process; unbalanced data; semi-supervised**

I. Introduction

Anomaly detection is a particular example of pattern recognition that attacks the problem of identifying patterns in data that do not conform the expected (or normal) behavior. Anomaly detection has a broad range of applications. To name a few: intrusion detection in a computer related system (e.g. [2]), industrial damage detection (e.g. [3]), healthcare (e.g. [4] and [1]). All these applications have in common that data describing failure conditions or abnormal behavior are rare. Hence the data set is unbalanced such that traditional classification methods like support vector machines perform suboptimal.

In [5] anomaly detection techniques are categorized as belonging to one of the following classes: classification based, nearest neighbor-based, clustering based, statistical (parametric and nonparametric), information theoretic and spectral. An important aspect of an anomaly detection technique is the necessity of data labels that denote whether an instance is normal or anomalous. Based on the extent to which labels are needed to train a classification model, anomaly detection techniques can be classified as supervised, semi-supervised or unsupervised.

In this article the focus lies on a statistical nonparametric approach that can be used in a semi-supervised context. Such approach models the normal behavior using a kernel density estimator (KDE) that places a Gaussian kernel on each data point and then sums the local contributions from each kernel over the entire data set [6]. This kernel estimation is then used to predict whether a test sample comes from the same distribution or not. Such approach is based on the assumption that anomalies occur in low probability regions of the stochastic model. The use of extreme value theory (EVT) is proposed to obtain a novel method for multivariate anomaly detection.

A. Related work

It is common in literature to identify anomaly detection with outlier detection. The concepts outlier and outlier detection have been known quite a long time and are extensively studied in the statistics community [7]. An outlier in a data set is defined as an observation (or subset of observations) which appear to be inconsistent with the remainder of that set of data. Anomalies however can be viewed more generally as patterns in data that do not conform to a well defined notion of normal behavior [5].

In [8] a nonparametric methodology is developed to identify outliers in multivariate data that is based on data depth. Data depth is a way of measuring how deep or central a given data point $\boldsymbol{x} \in \mathbb{R}^d$ is w.r.t. to a given distribution. It can be quantified using the probability density $y = p(\boldsymbol{x})$ of the distribution, also known as the likelihood depth of $\boldsymbol{x}$. The use of EVT for anomaly detection in multivariate data was first approached in [9], where models of normality were given by Gaussian models such that the multivariate problem could be reduced to an univariate problem using the Mahalanobis radius. Clifton et al. [10] introduced a statistical non-parametric approach using EVT by modeling those data points that have a minimal density with respect to the KDE that models normal behavior.

B. Contributions

In EVT literature there are three approaches used to model univariate extremes: block extrema models, peaks over threshold (POT) models and the point process modeling framework [11]. In [10] block extrema models are used to classify multivariate anomalies and in [12] the use of POT models is introduced to classify multivariate anomalies.

In this paper the use of the point process characterization of univariate extremes is explored in order to classify multivariate anomalies. For this purpose count data is extracted from a multivariate density distribution modeling normal behavior that describe the number of times densities fall below some boundary. The Poisson process approach of extremes is then used to unify this count information with that on the excess values to obtain new ways of defining anomaly scores that are able to fully capture all information provided by the EVT models. The method is evaluated using as well simulated as real world data and compared with an one-class support vector machine (SVM) approach. For the real world data this will lead to an increase up to 12.9% of the positive predictive values while the sensitivity scores stay unaltered in comparison with the results obtained in [1].

The paper is organized as follows. To start with, the problem setting is stated in section II. In section III the existing EVT approaches for anomaly detection on multivariate data are reviewed. Subsequently section IV introduces the use of the point process approach of EVT for anomaly detection leading to the introduction of new anomaly scores in section V. In section VI the method is evaluated using as well simulated as real world data. Finally the paper ends with a conclusion that discusses the methodology and results.

II. Problem setting

Denote $\mathcal{D} \subset \mathbb{R}^d$ a d-dimensional data set describing normal behavior. Mostly data is gathered by a sensor system. As analysis of sensor signals in their original form usually yields poor results a typical anomaly study is preceded by a preprocessing phase and a feature extraction. In that case vectors $\boldsymbol{x} \in \mathcal{D}$ will present d-dimensional feature vectors describing signals coming from normal behavior (e.g. EEG signals from normal brain activity, signals of normal heart rates etc.)

Statistically, the vectors $\boldsymbol{x} \in \mathcal{D}$ are assumed to be independent realizations of a stochastic variable X that is distributed according to a probability distribution function (PDF) $y = p(\boldsymbol{x})$. Throughout the paper, this model of normal behavior is assumed to be static, rather than dynamic. Anomalies in time series can be modeled using dynamic models such as Hidden Markov models or Kalman filters [13].

In order to be able to apply EVT a set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subset \mathbb{R}^d$ of k feature vectors is considered and anomaly detection addresses the question whether such set S of vectors is drawn from the distribution X or not. In practice S will present the last vector and the $k-1$ vectors observed before it such that information of the last k measurement can be combined using EVT. As will be further discussed in the next section the choice of k will imply a trade-off between bias and variance of the estimated parameters of the EVT model.

In terms of statistical hypothesis testing the problem setting can be stated as:

H_0 : S is a set of vectors drawn from the population X
H_1 : S is an anomalous set with respect to X.

where H_0 denotes the so-called null-hypothesis and H_1 the alternative hypothesis. The probability of wrongly classifying a normal sample S as anomalous (known as a type I-error) is given by the significance level of the test denoted as α (typically $\alpha = 0.05$ or $\alpha = 0.01$).

From the point of view of hypothesis testing, it is clear that for $k > 1$ the problem is related to one of multiple testing. Indeed, for $k = 1$ a threshold can be chosen on the likelihood $p(\boldsymbol{x}_1)$ of the unique sample in S to obtain a classification model [14]. For $k > 1$ the probability to make at least one type-I error while testing each $\boldsymbol{x}_i \in S$ is given by:

$$\text{P(type I-error)} = 1 - (1-\alpha)^k > \alpha,$$

e.g. when $\alpha = 5\%$, then P(Type-I Error)=26% for $k = 6$. The problem of multiple (hypothesis) testing refers to testing more than one hypothesis at a time and is a well known statistical problem [15].

III. EVT and anomaly detection

We review the recent methodologies of the use of EVT for anomaly detection in case the model of normal behavior is static and that are recently introduced in [10], [12]. The proposed methods in literature starts from block models or POT models.

For this the univariate distribution over the probability density values $p(\boldsymbol{x})$ on $\mathrm{Im}(p) = \{p(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{D}\}$ is considered reducing the multivariate analysis of the multidimensional data set $\mathcal{D}$ to an univariate analysis on $\mathrm{Im}(p)$. The distribution Y of densities $y = p(\boldsymbol{x})$ is strongly related to that of X with a PDF defined by:

$$q(y) = \frac{dQ}{dy}(y) \quad \text{and} \quad Q(y) = \int_{p^{-1}(]0,y[)} p(\boldsymbol{x}) d\boldsymbol{x} \qquad (1)$$

A. Block models

Univariate EVT can be used to describe sets:

$$S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$$

which have a typical minimal density with respect to the PDF $y = p(\boldsymbol{x})$. In this way the most 'rare' vectors are described that possibly occur in samples S drawn from X .

In particular, one considers for each set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ the corresponding sample of density values in $\mathrm{Im}(p) \subset \mathbb{R}$

given by $(p(\boldsymbol{x}_1), \ldots, p(\boldsymbol{x}_k))$. Using EVT, one can determine the distribution of minima of samples from Y:

$$y^e := \min_{1 \leq i \leq k}(p(\boldsymbol{x}_i)). \tag{2}$$

The distribution of the population Y^e tells us where the minimal densities of samples S of length k are expected to lie. According to EVT, the distribution for minimal densities y^e, with k large enough, can be approximated by the Weibull distribution that possesses a lower bound at zero [11].

In practice, we find abnormal samples with an abnormal small minimal density with respect to the Weibull distribution. These are all pushed in a small interval close to zero. A standard logarithmic transform $z = -\log(y^e)$ overcomes this skewness. The short tail near zero of the Weibull distribution is stretched out as the right tail of a Gumbel distribution for maxima, where extremes can be shown in a clearer way. The corresponding cumulative distribution function reads:

$$G(z) = \exp(-\exp(-\frac{z-\mu}{\sigma})) \tag{3}$$

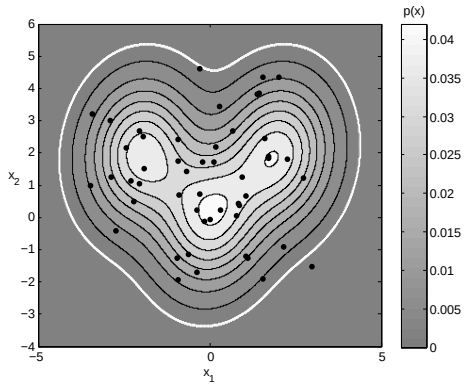
where (μ, σ) describe location and scale respectively. Together with the shape parameter which is zero for a Gumbel distribution, they are called the generalized extreme value (GEV)-parameters of the so-called *block (maxima) model*.

It's well known in EVT that the choice of *block size* k implies a trade-off between bias and variance. A too large block size results in large estimation variance and in high CPU-times for practical implementations. A too small block size results in a poor estimation of the model parameters.

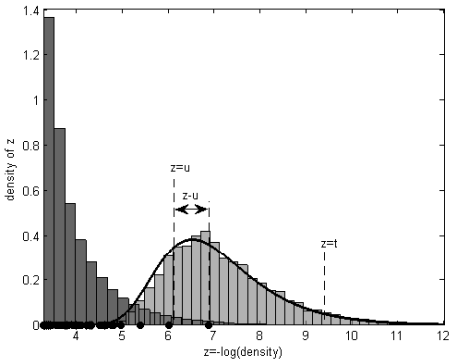
As an illustration the estimation of the distribution $Z = -\log(Y^e)$ is performed based on a trimodal Gaussian distribution X shown in Fig. 1(a). The estimation of Z is accomplished by simulating 6000 blocks of $k = 50$ densities from X. Based on the minimal densities in each of these blocks a Gumbel distribution can be fit after a logarithmic transformation using maximum likelihood estimation (MLE). Fig. 1(b) shows a histogram of $-\log(p(\boldsymbol{x}))$ for the samples $\boldsymbol{x}$ drawn from X, together with the Gumbel model of the right tail of this distribution. The dots in the figures correspond with one sample of length 50 drawn from the distribution of X. Indeed, the value of $z = -\log(y^e)$, which correspond to the rightmost dot on the x-axis in Fig. 1(b), is situated near the mode of the Gumbel distribution. The cumulative probability (3) increases as the samples drawn from X are situated further in the tail. Therefore the cumulative probabilities:

$$\chi_g(S) := G(y^e) \tag{4}$$

can be used to indicate the anomalous character of the set S. From a probabilistic point of view it is natural to threshold these cumulative probabilities at e.g. 95%. In the following these scores will be referred to as *Gumbel scores*.



(a)



(b)

Figure 1. (a) Contourplot of a tri-modal Gaussian mixture X with means at $(0,0)$, $(-2,2)$ and $(2,2)$ and covariance matrix $\begin{pmatrix} 1 & 0.1 \\ 0.1 & 2 \end{pmatrix}$ together with a sample of length 50. The white contour corresponds with a density u chosen using Van Kerms rule of thumb. (b) The corresponding distribution of densities $q(y)$ after a logarithmic transformation. The dots on the x-axis correspond with the log-densities $z = -\log(p(\boldsymbol{x}))$ of the sample from X. The tail is modeled using a Gumbel distribution of block maxima with block size $k = 50$. The threshold t correspond with the 95% threshold border on $G(z)$.

B. POT models

Alternatively one can use the peaks-over-threshold (POT) technique [11], also known as the threshold model. In this approach complete tails of a distribution are modeled. In particular a generalized Pareto distribution (GPD) can be used to model log-densities $z = -\log p(\boldsymbol{x})$ that fall above some threshold u. The parameters of the GPD are closely related to the GEV-parameters. When the block maxima of Z follow a Gumbel distribution with parameters (μ, σ), the corresponding GPD of excesses $V = Z - u$ is given by:

$$H(v) = P(Z - u \leq v \mid Z > u) = 1 - \exp\left(-\frac{v}{\sigma}\right) \tag{5}$$

The threshold u can be chosen using standard methods as mean-excess plots [11] or can be based on a rule of thumb (e.g. Van Kerms rule of thumb [16], see Fig. 1(a) and 1(b)). As explained in [12], this method gives the possibility to model the tail and the non-extremal data with

different distributions. In this way one is able to model that which actually exists in the tail rather than relying on an extrapolation of the model via simulations.

IV. Starting from Poisson processes

Rather than starting from the block model, the so-called *point process characterization* of extremes is used to start with. A Poisson point process (for short a Poisson process) is a particular case of a point process that basically models the number of times log-densities $z = -\log p(\boldsymbol{x})$ fall above some boundary u. Because the Poisson process model can be parametrized in terms of the GEV- and GPD- parameters a model is obtained that unifies the block model and POT model. Therefore the point process characterization leads to several advantages in the context of anomaly detection:

(i) A Poisson processes model enables to capture the information of the number of times densities fall below some boundary e^{-u}. This information can be very important in the distinction between anomalies and normal data points.

(ii) The Poisson process model can be parametrized in terms of the GEV- and GPD- parameters. Therefore the information of the number of times densities fall below some boundary can be combined with the amounts by which the threshold is exceeded resulting in a powerful model for anomaly detection.

(iii) In comparison with the POT model, the parametrization of the Poisson process model is invariant to threshold choice. This could be beneficial to include co-variate effects in the parameters [11].

Moreover any inference made using the classical extreme value model could equally be made using the Poisson process 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 Poisson process model.

Let us start with some general notions on point processes. For a more comprehensive description we refer to [17]. A (spatial)point process N on a subset $U \subset \mathbb{R}^n$ is a stochastic model for which any one realization consists of a random configuration of points in U. A point process can be denoted as a random set $N = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\} \subset U$ consisting of all points in the process and where the number of points in N is random or deterministic. Alternatively one can refer to N as a random counting measure associating to each (measurable) subset $A \subset U$ a random variable $N(A) = N_A$ describing the number of points in A:

$$N_A : \quad \omega \quad \mapsto \quad \text{“number of points in A”}$$

In this definition ω denotes a stochastic event presenting some random configuration of points in U.

The intensity measure Λ of a point process is defined as $\Lambda(A) = E(N_A)$, where E denotes the expectation operator. The intensity (density) function $\Psi(x), x \in \mathbb{R}^n$ is defined as the derivative function (provided it exists) of this measure:

$$\Lambda(A) = \int_A \Psi(\boldsymbol{x}) d\boldsymbol{x}$$

A particular case of a point process is given by a (non-homogeneous) Poisson process with a (non-constant) intensity function $\Psi(.)$ such that the associated random variables N_A follow a Poisson distribution $N_A \sim \text{Poi}(\Lambda(A))$ with the property that for disjunct subsets $F_1, \ldots F_k$ the random variables $(N_{F_1}, N_{F_2}, \ldots, N_{F_k})$ are independent. In particular the occurence of a point at a location in U should not influence the probability of the occurence of other points at other locations.

Let us explain how the point process framework of EVT can be used to determine whether a set $S = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_k\} \subset \mathbb{R}^d$ is drawn from a population X with a density function $y = p(\boldsymbol{x})$. In particular the Poisson process limit of univariate extremes is used to estimate the number of points of S that are expected to lie beyond a boundary in $\mathbb{R}^n$ determined by $p(\boldsymbol{x}) = e^{-u}$. To this end, consider the sequence of i.i.d. variables $(X_i), 1 \leq i \leq k$ of which x_i are realizations and denote $Z_i = -\log(Y_i), 1 \leq i \leq k$ where for each i, Y_i denotes the random variable associated to the distribution of densities of X_i as defined in (1). One supposes that the Gumbel approximation is valid, meaning that

$$P\left(\max\{Z_1, \ldots, Z_k\} \leq z\right) \approx G(z), \quad \text{as} \quad k \to +\infty$$

where $G(z)$ is the Gumbel distribution as in (3) depending on the GEV-parameters (μ, σ). It is known from EVT literature that the sequence of point processes N_k on $\mathbb{R}^2$:

$$N_k = \{\left(\frac{i}{k+1}, Z_i\right) \mid 1 \leq i \leq k\}$$

can be approximated by a Poisson process on regions of the form $U = (0,1) \times [u, +\infty)$ for sufficiently large u. The intensity measure of this Poisson process limit is given by:

$$\Lambda(A) = (t_2 - t_1) \exp\left(-\frac{z-\mu}{\sigma}\right)$$

on $A = (t_1, t_2) \times (z, +\infty) \subset U$ and where (μ, σ) are the GEV-parameters.

The parameters of the Poisson process are obtained by MLE. For this purpose, densities $z_i = -\log(p(\boldsymbol{x}_i))$ are simulated from the distribution defined in (1). The Poisson process model is based on those z_i that fall above some threshold u. The threshold u is chosen in front and can be based on a rule of thumb (e.g. Van Kerm's rule of thumb [16]). If one observes n excesses $\{z_1 - u, \ldots z_n - u\}$ from kT simulated densities, the corresponding log-likelihood is given by:

$$l(\lambda, \sigma) = n \log \lambda - \lambda T - n \log \sigma - \sum_{i=1}^{n} \frac{z_i - u}{\sigma} \qquad (6)$$

where $\lambda = \lambda(\mu, \sigma) = \exp\left(-\frac{u-\mu}{\sigma}\right)$ is given by the intensity measure of the Poisson process on $(0,1) \times (u, +\infty)$. The parameters (μ, σ) correspond with the GEV-parameters as in (3) with block size k.

It is clear that the Poisson process limit can be used to determine the number of log-densities z_i of S that are expected to lie above the boundary $Z = u$. Moreover because the likelihood (6) is parametrized in terms of the GEV-parameters and POT-parameters, one immediately obtains information about the amount of excesses above this threshold u. In this way the point process approach provides us a way to unify all information provided by the EVT-models.

As an illustration the estimation is performed on the trimodal Gaussian distribution, shown in Fig. 1(a). The simulated counts of exceedances follow a Poisson distribution with intensity λ, see Fig. 2(a). The quality of the fit can be assessed by a quantile-quantile (QQ) plot. The x-axis shows the quantiles obtained from the GPD in (5) after estimating the parameters using the likelihood in (6). The y-axis shows the empirical quantiles obtained from the data. If the model approximates well during simulation, the points on the graph are expected near the diagonal $y = x$ shown as a dashed line. While the fit follows the data well over the majority of the range, the quality of the fit is lower for higher quantiles. This divergence of the fit from the more extremal data is known in the EVT literature and is also noticed in [12].

V. Defining anomaly scores

A new test set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ of feature vectors can be evaluated after a model of normal behavior is estimated. For this purpose so-called anomaly scores are defined combining the information obtained from the GEV-, the GPD and the Poisson process model.

Associated to the event S is a sample from the point process:

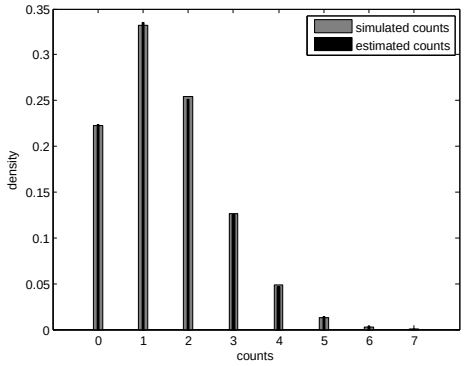
$$N_k = \{\left(\frac{i}{k+1}, Z_i\right) \mid 1 \leq i \leq k\}$$

where k denotes the number of vectors in S and Z_i is defined as in the previous section. Denote $\{z_1, \ldots z_{n_S}\}$ as the log-densities $z_i = -\log(p(\boldsymbol{x}_i))$ that fall above the threshold u. A first anomaly score can be based on the Poisson process model itself:

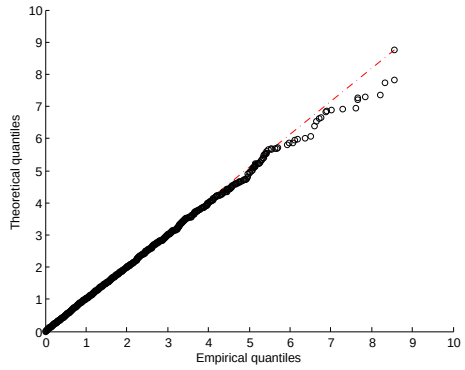
$$\chi_p(S) = P(N_k \leq n_S) = \sum_{i=0}^{n_S} \frac{\lambda^i}{i!} e^{-\lambda} \tag{7}$$

with $\lambda = \exp\left(-\frac{u-\mu}{\sigma}\right)$. Let us call this score, the *Poisson score* of S. The Poisson score increases with an increasing number of times that the log-densities of S exceed the threshold u.

A second score that we can define is based on the excess model. This score will inform us about the amounts the log-densities of S exceed the threshold u. Denote



(a)



(b)

Figure 2. (a) Number of exceedances counted in the simulated sample and compared with the estimated counts given by the Poisson distribution Poi(λ). (b) The fit of parameters obtained from the likelihood (6) is assessed using a QQ-plot.

$\{z_1 - u, \ldots, z_{n_S} - u\}$ as the excess values of the log-densities of S. One way to capture the extent of exceedeness is by considering the mean excess associated to S:

$$m_S = \frac{1}{n_S} \sum_{i=1}^{n_S} (z_i - u)$$

Because the excesses come from an exponential distribution defined in (5), this mean is distributed according to an Erlang distribution with shape-parameter n_S and rate parameter $\frac{n_S}{\sigma}$:

$$M_S \sim \Gamma(n_S, \frac{n_S}{\sigma})$$

The Gamma-notation comes from the fact that an Erlang distribution is a special case of a Gamma distribution with an integer shape parameter n_S and a rate parameter $\frac{n_S}{\sigma}$. We now define an anomaly score $\chi_e(S) = P(M_S < m_S)$:

$$\chi_e(S) = \int_0^{m_s} \frac{\left(\frac{n_S}{\sigma}\right)^{n_S}}{(n_S - 1)!} x^{n_S - 1} e^{-\frac{n_S}{\sigma} x} dx \tag{8}$$

representing the extent of mean excess of S. We will call this score the *Gamma score* of S.

The anomaly scores (7) and (8) can be combined together with the Gumbel scores (4) in one score using a generalized

mean:

$$\overline{\chi}_r(S) = \left(\frac{1}{3}(\chi_p(S)^r + \chi_e(S)^r + \chi_g(S)^r)\right)^{1/r} \quad (9)$$

Note that the Gumbel scores only reflects whether or not the most extreme feature vector of S will exceed some predefined threshold and therefore contains limited information. The scores in (9) succeed in combining all information obtained from the three approaches from EVT to model extremes.

Depending on the application one can choose an appropriate r. When $r \mapsto 0$ one obtains a geometric mean:

$$\overline{\chi}_0(S) = \sqrt[3]{\chi_p(S)\chi_e(S)\chi_g(S)}$$

For $r \mapsto -\infty$ and $r \mapsto +\infty$ one gets respectevely:

$$\begin{aligned} \overline{\chi}_{-\infty}(S) &= \min\{\chi_p(S), \chi_e(S), \chi_g(S)\} \\ \overline{\chi}_{+\infty}(S) &= \max\{\chi_p(S), \chi_e(S), \chi_g(S)\} \end{aligned}$$

The following inequality holds for $r < s$:

$$\overline{\chi}_{-\infty}(S) < \overline{\chi}_r(S) < \overline{\chi}_s(S) < \overline{\chi}_{+\infty}(S)$$

Depending on the choice of r the sensitivity of the novelty system is influenced. A choice of $r = +\infty$ leads to a novelty system that gives an alarm when at least one novelty score exceeds a threshold and therefore implies maximal sensitivity (SS) scores but possible lower positive predictive values (PPV). For $r = -\infty$ all novelty scores has to exceed a threshold implying less false alarms and mostly lower SS scores. All other choices are situated between these two extremes.

VI. Experimental results

To illustrate the validity of the method results on both artificial and real-world data sets are given in this section.

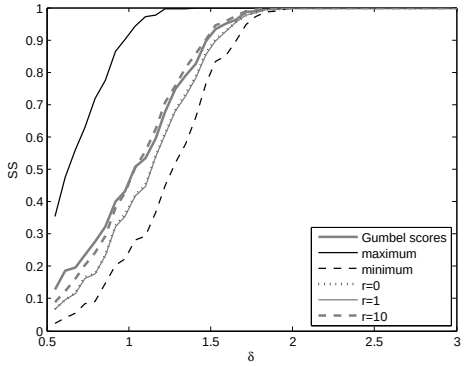
A. Synthetic Data set

In this section a validation is presented on data simulated from a bi-variate normal distribution $N(\boldsymbol{\mu}_0, \Sigma_0)$ centered at $\boldsymbol{\mu}_0 = (0, 0)$ and with a covariance matrix given by:

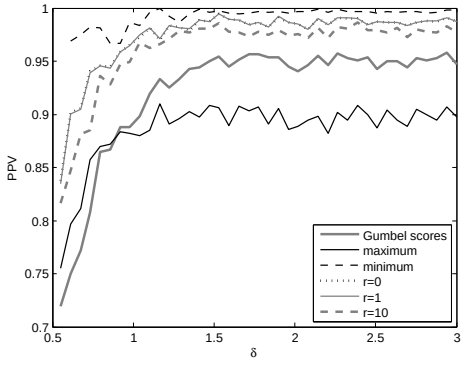
$$\Sigma_0 = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 0.3 \end{pmatrix}$$

Considering this distribution as representing normal behavior the SS of the proposed method is tested when a shift in the x-direction occurs. For this purpose test sets are constructed containing 100 sets of length $k = 50$ drawn from $N(\boldsymbol{\mu}_0, \Sigma_0)$ and 100 sets from $N(\boldsymbol{\mu}_1, \Sigma_0)$ where $\boldsymbol{\mu}_1 = \mu_0 + \delta(1, 0)$ with the amount of shift δ varying between 0.5 and 3 (corresponding to 2 standard deviations in the x-direction).

The proposed method is tested using (9) at some typical values $r \in \{-\infty, 0, 1, +\infty\}$ corresponding respectively to the minimum, arithmetic mean, geometric mean and maximum of the Gumbel, Gamma and Poisson scores as defined in the previous section. For each shift $\delta \in [0.5, 3]$ (with a step size of 0.03) PPV and SS scores are calculated by using a 10-fold randomization for training and testing. Fig. 3(a) and 3(b) show the mean of these scores taken over the 10 runs and plotted with respect to the amount of shift δ. The results are compared with the EVT method proposed in [10] that was only based on the Gumbel scores (4).



(a)

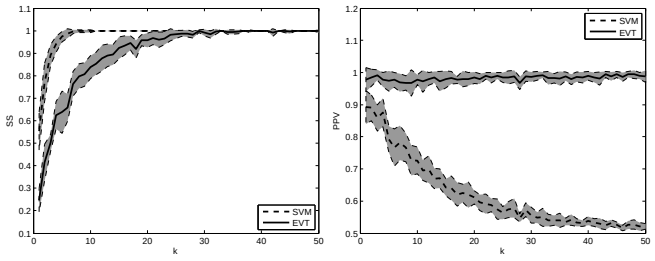


(b)

Figure 3. Mean of SS (a) and PPV (b) scores calculated using a 10-fold randomization and plotted with respect to the shift δ.

The SS and PPV scores for $r = 0$ and $r = 1$ are very similar. By increasing r to $r = 10$ the SS scores could be further improved to closely match the SS scores obtained using the Gumbel scores. At the same time the PPV scores only decrease slightly resulting in an overall better performance than the approach using the Gumbel sores only. Note that, as already expected from section V, the choice of $r = +\infty$ leaded to the lowest PPV scores but the highest SS scores. For $r = -\infty$ the method resulted in the lowest amount of false alarm but was less sensitive for anomalies. The curves for $r = 0, 1$ and $r = 10$ are situated between these two extremes.

The EVT-methods are compared with an one-class support vector machine (SVM) algorithm [18]. Given a prior probability ν and a kernel width γ the support of the training can be estimated such that the probability that a sample point lies outside the support is bounded from above by ν. As ν is an upper bound on the fraction of type-I errors it is set to



(a) (b)

Figure 4. Mean of SS (a) and PPV (b) scores calculated using a 10-fold randomization and plotted with respect to the block size k. The gray zones indicate the standard deviation of the scores.

0.05 in accordance to the threshold of 95% on the anomaly scores. Moreover a Gaussian kernel was used that has the advantage that the data are always separable from the origin in feature space.

An SVM support is expected to perform well when sample points are individually evaluated. However, for $k > 1$ the PPV scores will decrease significantly due to the problem of multiple testing inducing an increase in false alarm rate as can be depicted in Fig. 4(b). This effect is independently of the choice of kernel width, which is here chosen to be $\gamma = 0.1$ for illustrative purposes. The use of EVT enables us to keep the PPV scores at a high level by using a correct probabilistic model for samples of length $k > 1$. Moreover, Fig. 4(a) shows that the SS scores for the SVM and the EVT approach are not further apart than 1% for values of $k > 20$. The gray zones in figures 4(a) and 4(b) show the standard deviation of the scores obtained from a 10-fold randomization method. Note that the standard deviation of the scores decrease as k increases. This is a consequence of the increase of precision of the approximation of the Poisson process limit of EVT as k increases. At the same time a larger k implies a higher alarm rate for the SVM approach such that the SS scores naturally approach 100% with a variance of zero. This increase in alarm rate results at the same time in PPV scores that approach 50% (the percentage of abnormal points present in the test set) with a variance of nearly zero.

B. Detection of epileptic convulsions

In this section a case study in healthcare is considered using a data set of acceleration data collected from movements of patients suffering from epilepsy [4]. The acceleration data was recorded during several nights using four 3D acceleration sensors that are attached to the extremities of 7 patients 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 some minutes. Due to the heavy movement, the patient can injure himself during the seizure, which increases the need for an alarm system (with high SS scores).

In [1] a classification algorithm for seizure detection based on the Gumbel scores as defined in (4) was evaluated. Because the study was only based on the Gumbel scores, it is plausible that performance scores of this classification system can be improved by incorporating more information into the model using the proposed point process approach as indicated in section IV.

Let us first briefly review the necessary aspects of the study in [1]. The classification algorithms start with a feature extraction using 50% overlapping sliding windows containing 125 samples. In this way a set of 3-dimensional feature vectors is obtained per movement event. Based on fixed-length subsets S (of length $k = 50$) of these feature sets EVT is used to judge whether movement events are related to normal behavior or to rare events including hypermotor seizure. The extracted features are the same as the ones that are computed in [1]. They are common features that are widely used in accelerometer detection [19] and epilepsy research [20]. The results of the study based on Gumbel scores are presented in table I. The SS and PPV scores are calculated using fixed partitions of the data in training and test sets in a 10-fold randomization (which is patient specific). The threshold of 95% on the Gumbel scores is chosen firstly and is a typical choice from a probabilistic point of view. However it was shown that it can be optimized patient specifically. Moreover the choice of k in the preprocessing phase did not influence the performance scores significantly from $k = 50$. As noticed in [1] the PPV values may be on the low end and they do seem to have higher ranges. Therefore, further research is needed to optimize the method in order to integrate the methodology in an alarm system that can be used in the clinical routine. At the moment however it can be seen as a screening tool that can be used to monitor patients during night time and analyze seizures that were not observed by the caregivers.

In this section it will be shown that the proposed extension of the classical use of EVT leads to higher PPV scores. To this end r is set to $-\infty$ such that the combined anomaly score (9) is at its minimum. When applying the point process approach the same preprocessing steps as in [1] are followed keeping the partitions of training and test sets and the threshold of 95% on the anomaly scores the same to make a consistent comparison. The performance scores are presented in table II. In comparison with table I, the PPV score of 6 patients increases up to 12.9% with a mean increase of 9%. At the same time the sensitivity scores stay unaltered, except for patient 5. The decrease of performance scores for patient 5 is explained by the fact that seizures of this patient are less extreme and hence they are less detectable. This is discussed in detail in [1]. Using $\overline{\chi}_{-\infty}(S)$ as anomaly score will result in a general decrease of the anomaly scores and hence to lower sensitivity scores for patient 5. The decrease in PPV score indicates a decrease in

Table I
SENSITIVITY AND PPV FOR PATIENTS 1-7 USING χ_g AS ANOMALY SCORE THRESHOLDED AT 95%. MEANS AND STANDARD DEVIATIONS ARE CALCULCATED OVER 10 RUNS IN A 10-FOLD RANDOMIZATION (SCORES ARE COPIED FROM [1]).

PATIENT	SS			PPV		
1	100.0	±	0.0	49.1	±	37.4
2	100.0	±	0.0	60.0	±	20.0
3	100.0	±	0.0	56.3	±	17.8
4	70.0	±	25.81	31.8	±	25.2
5	27.8	±	12.0	20.8	±	10.0
6	100.0	±	0.0	56.7	±	17.3
7	100.0	±	0.0	44.0	±	9.8

Table II
SENSITIVITY(SS) AND PPV FOR PATIENTS 1-7 USING THE POINT PROCESS APPROACH WITH $\chi_{-\infty}$ AS ANOMALY SCORE THRESHOLDED AT 95%. MEANS AND STANDARD DEVIATIONS ARE CALCULCATED OVER 10 RUNS IN A 10-FOLD RANDOMIZATION.

PATIENT	SS			PPV		
1	100.0	±	0.0	52.8	±	35.9
2	100.0	±	0.0	71.8	±	18.9
3	100.0	±	0.0	64.7	±	21.5
4	70.0	±	25.8	40.5	±	32.2
5	13.3	±	11.5	15.8	±	13.1
6	100.0	±	0.0	69.6	±	24.6
7	100.0	±	0.0	52.6	±	12.4

true positives.

The improvement in performance scores is due to the fact that a Poisson process model enables to combine the information of the amount of excess values with the number of times accelerations are extremely high while a block model only models the fact whether or not an extreme threshold is exceeded. It is indeed plausible that a typical epileptic convulsion does not result in one very high exceedance in the acceleration data but to multiple exceedances with a high mean excess. To illustrate this fact, consider the two movements of patient 2 shown in Fig. 5. Both movements are presented by a set S of feature vectors $\boldsymbol{x}_i$, $1 \leq i \leq 50$ for which $z = \max_{1 \leq i \leq 50} \{-\log(p(\boldsymbol{x}_i))\}$ exceeds the threshold determined by the Gumbel model indicated by the dashed line at height t. Both movements will trigger an alarm when using a block model. The Gumbel scores of the normal movement and seizure are respectively given by 96.02 and 99.99 and both exceed the threshold of 95%. However the movements in the seizure are clearly more violent than the normal movement as the densities in Fig. 5 exceed the threshold t many times and with high excess values while the densities of the normal movement only exceeds the threshold t once with an excess value that is nearly zero (note that the dot graphically appearing on the dashed line falls below the threshold border t).

The Gamma scores and Poisson scores use the threshold

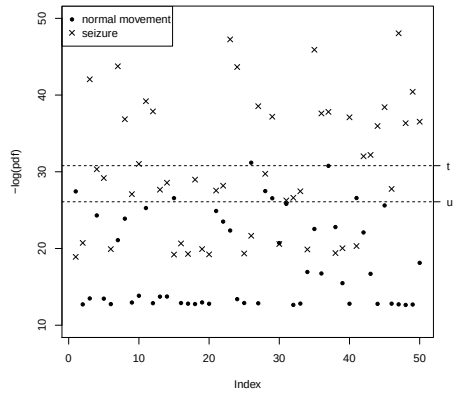


Figure 5. Plot of the log-densities $-\log(p(\boldsymbol{x}_i)), 1 \leq i \leq 50$ of a normal movement and a seizure. The threshold t corresponds with the 95% quantile of the trained Gumbel distribution. The threshold u is chosen using Van Kerms rule of thumb to estimate the parameters of a Poisson process.

u as reference line. Because the number of densities that fall above the threshold u is high for each movement the Poisson scores both exceed 99%. However there is a clear difference between the Gamma scores, given by 80.47% and 99.99% for the normal movement and seizure respectively. Therefore because the excess values of the normal movement are not high with repect to our model of normal behavior an alarm is not triggered by using $\overline{\chi}_{-\infty}(S)$ as anomaly score.

In this application the choice of r was straightforward for the majority of the patients, but this does not have to be in all applications. For patients 1-3 and 6-7 it is clear that we obtain an optimum by choosing $r = -\infty$ as the anomaly score $\chi_r(S)$ is at it minimum implying higher PPV values while in the meantime the sensitivity scores stay 100% in this application. For patients 4 and 5 the dependency of the sensitivity scores and PPV values is illustrated in Fig. 6. For patient 4 a further optimization of the sensitivity is only possible at the expense of a lower PPV. For patient 5 the sensitivity could be optimized without lowering the PPV considerably by choosing a higher r.

QQ-plots are verified during all runs of the 10-fold randomization process of all patients. All plots were very similar and showed that the models were satisfactory. As an illustration Fig. 7 shows a QQ-plot of one run of patient 2.

Finally the results are compared with an one-class SVM classifier [18]. To this end, feature are extracted from complete movements such that each movement is represented by 1 feature vector. To make a consistent comparison with the EVT-method the same features and randomizations during the 10-fold cross validation are chosen. There are several possible criterions for selecting γ [21]. Here the parameter ν was set to 0.05 (in accordance with the 95% threshold on the anomaly scores based on the EVT-method) and performance scores were optimized with respect to the kernel width γ

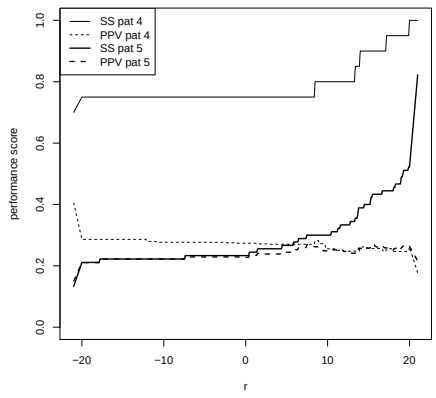


Figure 6. Dependency of the sensitivity scores (SS) and PPV values in function of the choice r in (9) for patient 4 and 5. The endpoints of the curve correspond with $r = -\infty$ and $r = +\infty$.

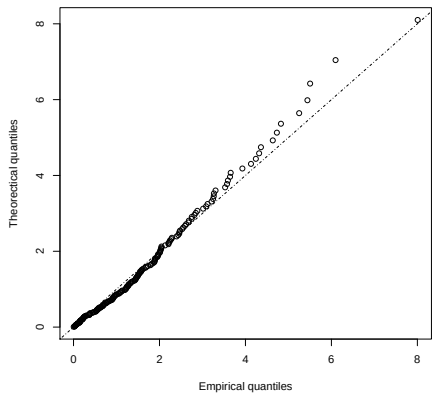


Figure 7. QQ-plot of the model during a run of patient 2.

varying over the range $]0, 10]$ with a step size of 0.01. Results are shown in table III. The PPV scores of patients $1-3$ and $6-7$ can be maximized while the SS scores are 100%. Note that during this optimization step the labeling of the seizures is needed, which is strictly not allowed in a semi-supervised context. Nonetheless the EVT-method is able to outperform the SVM approach in 5 of the 7 patients with a mean increase in PPV of 24.5%. For patient 4 a kernel width is chosen that maximizes the PPV score while maintaining a SS score of 100%. For any gamma the PPV scores do not exceed 26% while the sensitivity fluctuates between 60% and 100%. Although the SS scores are better for patient 4, the PPV scores undergo a global decrease in comparison with the EVT-method. For patient 5 it is possible to obtain a higher SS score and PPV score in comparison with our EVT-method by setting $\gamma = 0.81$. For this patient the SVM method was able to outperform the EVT method, although a parameter tuning is used.

Table III
SENSITIVITY(SS) AND PPV FOR PATIENTS 1-7 USING AN ONE-CLASS SVM CLASSIFIER. MEANS AND STANDARD DEVIATIONS ARE CALCULATED OVER 10 RUNS IN A 10-FOLD RANDOMIZATION. PERFORMANCE SCORES ARE OPTIMIZED WITH RESPECT TO THE KERNEL WIDTH PARAMETER γ.

PAT.	SS			PPV			γ
1	100.0	±	0.0	31.66	±	16.08	0.01
2	100.0	±	0.0	37.90	±	10.22	0.01
3	100.0	±	0.0	40.19	±	11.17	0.14
4	100.0	±	0.0	17.62	±	5.33	0.56
5	64.44	±	10.21	19.12	±	36.94	0.81
6	100.0	±	0.0	39.04	±	24.40	0.01
7	100.0	±	0.0	40.07	±	17.03	0.09

VII. CONCLUSION

This paper proposes a method for anomaly detection that extends the point process model of univariate extremes to multivariate use by considering the exceedances of $-\log(p(\boldsymbol{x}))$ above some threshold u, where $y = p(\boldsymbol{x})$ denotes a multivariate statistical distribution modeling normal behavior. Because the point process model can be parametrized in terms of the GEV- and GPD- parameters information obtained from the block model and POT model can be combined resulting in a powerful model for anomaly detection. In particular the information of the number of times densities fall below some boundary can be combined with that of the size of the excess values.

Given a test set $S = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_k\}$, anomaly detection addresses the question whether the set S is drawn from the distribution $y = p(\boldsymbol{x})$. For this anomaly scores are associated to S, defined as cumulative probabilities under the block-, POT- and Poisson model. These scores can be combined into one single score using a generalized mean with a parameter r (9). A new test set S that corresponds to normal behavior is expected to have a low anomaly score, while test sets corresponding to abnormal behavior will probably have a high anomaly score. From a probabilistic point of view a threshold of 95% is a typical threshold choice on these scores when applying this method in an unsupervised context.

We have demonstrated this method on artificial data and a real-world data set of acceleration data collected from movements of patients that suffer from epilepsy. Using artificial data it was shown that the proposed extension resulted in higher PPV values than the classical method using Gumbel scores. Furthermore, when more than one data point is observed an one-class SVM classifier is not appropriate due to the problem of multiple hypothesis testing. The real-world data set was highly unbalanced because seizures occur seldomly compared to normal movements. In comparison to the classical EVT method using Gumbel scores, the PPV values of 6 patients could be improved with an increase up to 12.9% (and a mean of 9%) while the sensitivity scores stayed unaltered. The proposed method was also shown to outperform the one-class SVM method (even when this is

optimized with respect to the parameters γ and ν).

Note that in the application of epileptic seizure detection the choice of r was straightforward for the majority of the patients, but this doesn't have to be in all application. The dependency of the performance scores on r are illustrated. As expected the sensitivity scores increase with increasing choices of r, while the PPV values decrease. For one patients the sensitivity score could be optimized significantly by choosing $r = +\infty$ while the PPV values stay roughly the same. However such an optimal choice of r is unlikely to be obtained in practice. In an semi-supervised setting where no labeled data of seizures is available, the dependency of the performance scores cannot be studied in advance. Therefore it would be interesting to study in future research how to automatically optimize the model parameter r. Moreover, we would like to validate this approach on other data sets.

ACKNOWLEDGMENTS

The 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. Also a thank you is send out to Kris Cuppens who helped us with the feature extraction. Finally, the authors would also like to acknowledge networking support by the ICT COST action IC1303 (AAPELE).

REFERENCES

[1] 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*, vol. 60, no. 2, pp. 89–96, 2014.

[2] D. Yeung and Y. Ding, "Host-based intrusion detection using dynamic and static behaviorial models," *Pattern Recognition*, vol. 36, no. 1, pp. 229–243, 2003.

[3] C. Bishop, "Novelty detection and neural network validation." in *Proceedings of the IEEE Conference on Vision, Image and Signal Processing*, vol. 141. IEE, London, 1994, pp. 217–222.

[4] 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*, vol. In Press, 2013.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[6] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, London, 1986.

[7] F. Edgeworth, "On discordant observations," *Philosophical Magazine*, vol. 23, no. 5, pp. 364–375, 1887.

[8] R. Liu, J. Parelius, and K. Singh, "Multivariate analysis by data depth: Descriptive statistics, graphics and inference," *The Annals of Statistics*, vol. 27, no. 3, pp. 783–858, 1999.

[9] S. Roberts, "Novelty detection using extreme value statistics," *IEE Proceedings on Vision, Image and Signal processing*, vol. 146, no. 3, pp. 124–129, 1999.

[10] D. Clifton, S. Hugeny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *Journal of Signal Processing Systems*, vol. 65, pp. 371–389, 2011.

[11] S. Coles, *An Introduction to Statistical modeling of Extreme Values*. Springer Verlag, London, 2001.

[12] D. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko, "An extreme function theory for novelty detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 28–37, 2013.

[13] P. Hayton, S. Utete, D. King, S. King, P. Anuzis, and L. Tarassenko, "Static and dynamic novelty detecton methods for jet engine health monitoring," *Phil. Trans. R. Soc. A*, vol. 365, no. 1851, pp. 493–514, 2007.

[14] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," in *Proceedings of 4th Internation Conference on Artificial Neural Networks*, F. Fogelman, Ed., vol. 4. IEE, London, 1995, pp. 442–447.

[15] P. Shaffer, "Multiple hypothesis testing," *Annual Review of Psychology*, vol. 46, pp. 561–584, 1995.

[16] A. Alfons and M. Templ, "Estimation of social exclusion indicators from complex surveys: The R package laeken," *Journal of Statistical Software*, vol. 54, no. 15, pp. 1–25, 2013.

[17] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley and Sons, Chichester, 2008.

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

[19] W. Dargie, "Analysis of time and frequency domain features of accelerometer measurements," in *Computer Communications and Networks, 2009 (ICCCN 2009)*, ser. Proceedings of 18th Internatonal Conference on Computer Communications and Networks, K. Gopalan, Ed. IEEE, New York, 2009, pp. 1–6.

[20] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 464–473, 2011.

[21] N. Cristianini and S.-T. J., *An introduction to Support Vector Machines an Other Kernel-based Learning Methods*. Cambridge University Press, 2000.