Point process models for novelty detection on spatial point patterns and their extremes

Stijn E. Luca[a,*], Marco A.F. Pimentel[b], Peter J. Watkinson[c], David A. Clifton[b]

[a] *KU Leuven, Campus Geel, Department of Electrical Engineering, Kleinhoefstraat 4, B-2440 Geel, Belgium*
[b] *University of Oxford, Department of Engineering science, Old Road Campus Research Building, Roosevelt Drive, Oxford, OX3 7DQ, UK.*
[c] *Oxford University Hospitals NHS Foundation Trust, Nuffield Department of Clinical Neurosciences, Headley Way, Oxford, OX3 9DU, UK.*

Abstract

Novelty detection is a particular example of pattern recognition identifying patterns that departure from some model of "normal behaviour". This article considers the classification of point patterns $\tilde{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ defined as sets of N observations of a multivariate random variable X and where the value N follows a discrete stochastic distribution. The use of point process models is introduced that allow us to describe the length N as well as the geometrical configuration in data space of such patterns $\tilde{\mathbf{x}}$. It is shown that such infinite dimensional study can be translated into a one-dimensional study that is analytically tractable for a multivariate Gaussian distribution. Moreover, for other multivariate distributions, an analytic approximation is obtained, by the use of extreme value theory, to model point patterns that occur in low-density regions as defined by X. The proposed models are demonstrated on synthetic and real-world data sets.

Keywords: novelty detection, point processes, extreme value theory, one-class classification, process monitoring

1. Introduction

Novelty detection is the task of recognising test data that differ in some respect from the data that were available during training [1]; it is typically used when there is a large quantity of "normal" data available, but an insufficient quantity of "abnormal" data, thus preventing accurate estimation of the "abnormal" class in a two-class classification setting [2]. Closely related to novelty detection is anomaly or outlier detection, where one also wish to detect abnormalities but where these may not necessarily be entirely novel with respect to the training data. A probabilistic approach starts with a statistical model describing the "normal" state and then detects deviations from this model. The majority of such work deals with a point-wise approach where the novelty of individual points $\mathbf{x}$ is evaluated. However when multiple points are evaluated, this can lead to a large number of misclassifications due to the multiple-hypothesis testing problem [3, 4].

This article considers the more general problem of classifying point patterns $\tilde{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ defined as sets of N observations of a multivariate random variable X and where the value N follows a discrete stochastic distribution. This problem setting is of particular importance when one has to deal with sparse data, defined as data in which segments of data are missing such that the number of observed measurements varies with time, space or some other index variable.

The method introduced in this article is based on the development of point process models (PPMs) and provides a valid probabilistic interpretation of the degree of novelty of an entire point pattern $\tilde{\mathbf{x}}$. This approach enables us to prevent those misclassifications induced by the multiple hypothesis testing problem. We show that the probabilistic assessment of novelty is analytical tractable in closed form when X has a multivariate normal distribution. Furthermore, we provide an analytic approximation using extreme value theory (EVT) to model those point patterns that are situated in regions where the density defined by X is low (the 'extremes' of the pattern). These regions are of particular importance because the decision boundary for novelty detection is typically situated at the edge of the support of X, where

*Corresponding author
Email address: stijn.luca@kuleuven.be (Stijn E. Luca)
URL: www.kuleuven.be/advise (Stijn E. Luca)

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

the density is low. Moreover, in several applications it is not the distribution of the bulk of data that is of interest, but rather the behaviour of extremes, e.g., earth science or financial applications [5].

Traditional point anomaly detection techniques often use a single distribution to describe deviations from a model of normal behaviour. A PPM is based on infinitely many random variables that fully characterise the configuration of a pattern $\tilde{\mathbf{x}}$ in its data space. Unlike existing approaches, a PPM follows-up the length N of patterns as well as the spatial configuration of the values $\mathbf{x}_i$ in data space. Furthermore, PPMs have the additional advantage that they can be adopted to follow-up the extremes within a pattern as well using extreme value theory (EVT).

The remainder of the paper is structured as follows. In Section 2, the problem setting is described, followed by an overview of related work in Section 3. In Section 4, an introduction to the theory of PPMs and EVT is given. Section 5 introduces our main novel approaches to novelty detection. In Section 6, the method is illustrated on synthetic and real-world data sets and compared with existing models that are commonly used for novelty detection. Conclusions are presented in Section 7.

2. Problem setting

In this article, the problem is addressed to determine whether or not a realized pattern of vectors:

$$\tilde{\mathbf{x}} = \{\mathbf{x}_1, \dots \mathbf{x}_n\}, \tag{1}$$

is anomalous, where as well the length n is evaluated with respect to a discrete distribution N as the locations $\mathbf{x}_i$ are evaluated with respect to a distribution X on $\mathbb{R}^d$. When $n = 0$, the pattern is treated as empty.

The problem setting is an example of a collective novelty detection problem where neither the individual instances within a pattern $\tilde{\mathbf{x}}$ nor the length n itself are classified. Instead, the entire pattern $\tilde{\mathbf{x}}$ of vectors is considered to be one single instance that is assigned a single label. In terms of statistical hypothesis testing, the problem can be stated as:

H_0 : $\tilde{\mathbf{x}}$ is a set of vectors drawn from X and n is drawn from N
H_1 : $\tilde{\mathbf{x}}$ is an anomalous pattern with respect to X and N

where H_0 denotes the null-hypothesis and H_1 the alternative hypothesis. The probability of wrongly classifying a 'normal' pattern $\tilde{\mathbf{x}}$ 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 the problem is related to one of multiple testing. The problem of multiple (hypothesis) testing refers to testing more than one hypothesis at a time and is a well known statistical problem [3]. In this article, we will show how PPMs can be applied to obtain the correct boundary of normality corresponding with the significance level α.

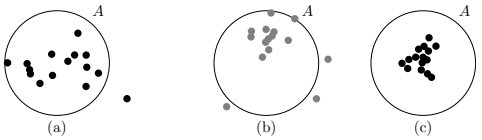


Figure 1: An illustration of samples with three different configurations with respect to some decision boundary A.

The use of a PPM will allow us to model the spatial configuration of the locations as well as the stochastic length of the pattern. To illustrate this, Figure 1 shows three artificial samples that are anomalous with respect to a standard 2-dimensional Gaussian distribution. The configuration of these points with respect to the boundary of the region A clearly differs as well as the number of point that are situated within the region. While the first sample (Figure 1(a)) contains one point that is situated far beyond the boundary defined by A, the second example (Figure 1(b)) contains multiple points near the boundary indicating that there is probably a a shift in the underlying process. The third sample (Figure 1(c)), however, indicates an accumulation of points near the centre which probably indicates that the variance of the underlying process is decreased.

We will assume in this article that the locations $\mathbf{x}_i$ of the pattern $\tilde{\mathbf{x}}$ are independent and identically distributed (i.i.d.). However, results may be extended to their use on time series data as well by considering the residuals after detrending. The latter will be demonstrated in Section 6.3 using a real-world data set.

3. Related work

In this section, an overview of related work is given for the main subjects treated in this work: (i) novelty detection, (ii) EVT, and (iii) PPMs.

3.1. Novelty detection

Most of the literature of novelty detection deals with a point-wise approach classifying individual points $\mathbf{x}_i$ and therefore only gives an answer to our problem setting in the case when $N = 1$. Widely-used examples include

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

the one-class support vector machine (OCSVM) [6]; active outlier (AO) [7], and local outlier factor (LOF) models [8]. For a complete review of the literature on novelty detection techniques, we refer to [9].

Closely related to our problem setting is sequence classification in which the point pattern $\tilde{\mathbf{x}}$ is considered as one instance that is assigned a single classification label. A commonly-used strategy in the literature is a sequential learning approach, where each point $\mathbf{x}_i$ within a pattern $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with a fixed length $N = k$ is given a label; the labels for all points in a pattern are then combined to yield a single classification for the whole pattern; e.g., this could be the mean of the individual novelty scores learned by an OCSVM. Similarly, a hidden Markov model (HMM) or a conditional random field (CRF) can be used to decide whether a pattern of data points is novel or not [10]. This approach, however is much in line with a point-wise approach where the number of false alarms can increase considerably due to the multiple hypothesis testing problem [3].

Alternatively, times series cluster models have been used to cluster sequences where the instances $\mathbf{x}_i$ are time dependent. Such approaches, however, heavily depend on the similarity metric and alignment method that are used [11]. Furthermore, such methods are not suited to incorporate the stochastic properties of the spatial configuration of a pattern $\tilde{\mathbf{x}}$. Group anomaly detection on the other hand aims to detect interesting aggregate behaviours of data points among several groups [12].

All these approaches, however, lack a joint model for the stochastic properties of the lengths and those of the spatial configuration in data space $\mathbb{R}^d$. A suitable PPM will be applicable to an arbitrary bags of points of which sequences of a fixed length is a special case. Moreover, due to its link with EVT which will be explained in Section 4, PPMs will allow us to follow-up the extremes within a pattern as well.

3.2. Extreme value theory

In many applications, it is not the distribution of the bulk of data that is of interest, but rather the behaviour of the extremes. Modelling the stochastic behaviour of such extremes is the subject of EVT, which has already been used for many applications ranging from biomedical engineering, structural health monitoring, meteorology, to risk assessment in financial domains [13].

In [14, 15, 4], the use of univariate EVT is proposed to classify patterns $\tilde{\mathbf{x}}$ of fixed length $N = k$ based on their extremes. The proposed EVT approaches were based on the so-called block model and peaks over threshold (POT) model. In such approach, only the single most extreme element in $\tilde{\mathbf{x}}$ (i.e. the vector where the density defined by a PDF $y = p(\mathbf{x})$ of a variable $\mathbf{X}$ is lowest) is used to obtain a decision. However, the most extreme element is expected to capture limited information about the tails of X that are defined as those regions where the density $y = p(x)$ of the variable X is below a (low) threshold e^{-u}. In [16, 17], it is shown how EVT can be used to include information contained in the number of exceedances and the average amount of exceedances present in:

$$\tilde{y} = p(\tilde{\mathbf{x}}) = \{p(\mathbf{x}_1), \ldots, p(\mathbf{x}_n)\},$$

with respect to a low threshold e^{-u} on the densities $p(\mathbf{x}_i)$, $1 \leq i \leq n$.

In this work a PPM is proposed that is able to fully capture the spatial configuration that is hidden in the instances of $\tilde{y}$ where the density defined by some PDF is lower than a threshold e^{-u}. Such PPM of exceedances will unify the existing EVT approaches discussed above. By working directly with PPM, the higher-order information arising from the configuration of exceedances can be incorporated efficiently.

3.3. Point process models

PPMs are random processes that describe the geometrical structure of patterns formed by objects that are randomly distributed in a multidimensional space. They are well-studied in probability theory and are mainly used to model and analyse spatial data. PPMs are applied in fields as diverse as astronomy, agriculture field trials, epidemiology, and computational neuroscience [18].

The use of PPMs for machine learning and pattern recognition applications is relatively new. PPMs show some links with random fields that are often applied in pattern recognition (e.g., conditional and Markov random fields [19]). Where a random field $\{Z(\mathbf{x})\}$ on $\mathbb{R}^d$ is a family of random variables having values in all $\mathbf{x}$ of $\mathbb{R}^d$, a PPM on $\mathbb{R}^d$ describes values occurring in random locations in $\mathbb{R}^d$. The use of Poisson and determinantal PPMs have recently been introduced in machine learning tasks as image search, tracking and text summarisation [20, 21]. In this article, the use of cluster PPMs and PPM of exceedances are introduced for novelty detection applications.

4. Point processes on a Euclidean space

In this section, general concepts for PPMs are reviewed [22]. After starting with an informal definition in Section 4.1, some typical characteristics of PPMs are given in Section 4.2. In Sections 4.3 and 4.4, two general classes of PPMs are given that will be of high practical use in the application of novelty detection.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

4.1. Informal definition

Informally, a PPM on a Euclidean space can be viewed as a random variable $\mathcal{M}$ with a distribution over all possible point patterns (or "point configurations") in some subspace $\mathcal{D}$ of $\mathbb{R}^d$, $d \in \mathbb{N}_0$, and where a point pattern is given by:

$$\tilde{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}, \quad \forall i : \mathbf{x}_i \in \mathcal{D}. \tag{2}$$

PPMs are able to describe the stochastic properties of the number of points of $\tilde{\mathbf{x}}$ as well as their location in space $\mathbb{R}^d$. The set-theoretical notation in (2) indicates that (i) the ordering of the points in a point pattern is irrelevant and that (ii) the points are different and thus do not coincide (a property referred to as *simplicity*[1]).

PPMs are often characterised by *counting measures*. The latter are random variables that map each configuration $\tilde{\mathbf{x}}$ of the PPM to the number of points falling in a bounded subset[2] $A \subset \mathbb{R}^d$:

$$N_A(\tilde{\mathbf{x}}) = \sum_{i \geq 1} \varepsilon_{\mathbf{x}_i}(A) < \infty \tag{3}$$

where:

$$\varepsilon_{\mathbf{x}_i}(A) = \begin{cases} 1 & \text{if} \quad \mathbf{x_i} \in A \\ 0 & \text{if} \quad \mathbf{x_i} \notin A. \end{cases}$$

A PPM is defined such that each N_A is a finite random variable, implying that only configurations $\tilde{\mathbf{x}}$ are considered that are *locally finite*, meaning that they contain a finite number of points in each bounded subset A. In fact, the values of these counting measures N_A for all subsets A give sufficient information to reconstruct completely the positions of a configuration $\tilde{\mathbf{x}}$. Indeed $N_{\{\mathbf{x}\}}(\tilde{\mathbf{x}}) > 0$ only applies for those $\mathbf{x} \in \tilde{\mathbf{x}}$.

4.2. Distribution and intensity measure

The distribution of a PPM is defined by a measure $\mathbb{P}$ that enables us to calculate probabilities of events $\mathcal{X}_0$:

$$\mathbb{P}(\mathcal{M} \in \mathcal{X}_0).$$

This describes the probability that the realisation of the PPM $\mathcal{M}$ belongs to a set $\mathcal{X}_0$ of point patterns. For example, setting $\mathcal{X}_0$ as the set of point patterns with a predefined length k that fall in some bounded subset $A \subset \mathcal{D}$ gives:

$$\mathbb{P}(\mathcal{M} \in \mathcal{X}_0) = P(N_A = k), \tag{4}$$

where P denotes the probability measure associated with N_A. It is clear from (4) that the distribution $\mathbb{P}$ of the PPM completely defines the distribution of the random variables N_A. Conversely it can be shown that the finite-dimensional distributions

$$(N(A_1), \ldots, N(A_n)), \quad n \in \mathbb{N} \text{ and } A_i \subset \mathcal{D} \text{ bounded}$$

characterise the distribution $\mathbb{P}$ of the PPM [18].

A fundamental concept related to the distribution of the counting measures N_A are their expected values. The *intensity measure* of a PPM is defined as:

$$\Lambda(A) = E(N_A)$$

and is a deterministic function operating on sets. The derivative function (provided it exists) of this measure is the so-called *intensity function* $\lambda(\mathbf{x}), \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^d$ and satisfies:

$$\Lambda(A) = \int_A \lambda(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{5}$$

4.3. Finite point processes

PPMs are called finite when each realised point pattern $\tilde{\mathbf{x}}$ almost surely consists of a finite number of points. A well-known class of such finite point processes is the class of independent and identically distributed (i.i.d.) *cluster models* [22]. These are PPMs such that the point patterns $\tilde{\mathbf{x}}$ consist of a finite number of points that are i.i.d. distributed according to some PDF $y = f(\mathbf{x})$. In particular, an i.i.d. cluster model $\widetilde{X}$ on $\mathcal{D} \subset \mathbb{R}^d$ associated with a random variable X is uniquely defined by:

(i) A random variable N describing the total number N of points in a point pattern, and which is distributed according to a discrete distribution on $\mathbb{N}$:

$$P(N = n) \quad \text{with} \quad \sum_{n=0}^{+\infty} P(N = n) = 1.$$

(ii) The random variable X on $\mathcal{D} \subset \mathbb{R}^d$ with a PDF $y = f(\mathbf{x})$ generating the locations of the points in the Euclidean space.

Conditioned on the length N of a pattern $\tilde{\mathbf{x}}$, the number of points N_A of the pattern $\tilde{\mathbf{x}}$ that fall in a subset $A \subset \mathcal{D}$ follows a binomial distribution $B(N, p_A)$ with $p_A = \int_A f(\mathbf{x}) d\mathbf{x}$:

$$P(N_A = k | N = n) = \binom{n}{k} p_A^k (1 - p_A)^{n-k},$$

[1] The general theory of PPMs also considers models with multiple coincident points.

[2] Formally, A is a Borel set. Supplemental material is associated with this article in which the formal definition of a PPM is given in more detail.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

Unconditionally the distribution of N_A can be found by marginalisation:

$$\begin{aligned}\eta_k &:= P(N_A = k)\\ &= \sum_{n=0}^{+\infty} P(N = n)\binom{n}{k} p_A^k (1-p_A)^{n-k}. \end{aligned} \tag{6}$$

The density function associated with the probability measure $\mathbb{P}$ of a cluster PPM $\widetilde{X}$ can be found by calculation of the probability of an event $\mathcal{X}_A$ of point patterns falling in a subset $A \subset \mathcal{D}$:

$$\begin{aligned}\mathbb{P}(\widetilde{X} \in \mathcal{X}_A) &= \sum_{k=0}^{+\infty} P(N_A = k) P((X_i)_{i=1}^k \in A | N_A = k)\\ &= \sum_{k=0}^{+\infty} \eta_k \prod_{i=1}^{k} \int_A f(\mathbf{x_i}) \mathrm{d}\mathbf{x_i}\\ &= \sum_{k=0}^{+\infty} \int_A \frac{1}{k!} \tilde{f}(\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}) \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x_k} \end{aligned} \tag{7}$$

where we have defined:

$$\tilde{f}(\tilde{\mathbf{x}}) = k!\, \eta_k \prod_{i=1}^{k} f(\mathbf{x_i}) \tag{8}$$

for $k > 0$ and $\tilde{f}(\{\emptyset\}) = \sum_{n=0}^{+\infty} P(N = n)(1 - p_A)^n$, when $k = 0$. The density function $\tilde{f}$ associated with the probability measure $\mathbb{P}$ of the PPM $\widetilde{X}$ is also called the *Janossy density function* of the PPM $\widetilde{X}$.

There is one final point to be noted here. As PPMs are treated as a theory of unordered point patterns (2), realisation of the random variable $\widetilde{X}$ can be considered as a point in a quotient space[3] in which point patterns are determined up to permutations. To be consistent, the density function $\tilde{f}$ is likewise considered on this quotient space, yielding the additional factorial $k!$ in (7).

4.4. Point processes of exceedances

PPMs are closely related to the study of exceedances in EVT. To see this, the PPM of exceedances must be considered studying those observations from a sequence of i.i.d. univariate random variables $W_1, \ldots W_n$ which exceed a given threshold u.

A basis result of EVT, termed peaks over threshold (POT), models complete tails of a univariate distribution W, defined as those measurements that fall above some threshold u. In [15] the use of the POT approach is extended to its multivariate use. For this purpose, those measurement $\mathbf{x}$ of a multivariate variable X are described for which $z = -\log(f(\mathbf{x}))$ falls above some threshold u:

$$A_u = \{\mathbf{x} \mid f(\mathbf{x}) < e^{-u}\}$$

It can be shown that when, the distribution of the exceedances $Z - u$, conditional on $Z > u$, satisfies the limiting property:

$$\lim_{u \to +\infty} P(\frac{Z-u}{\sigma} < z | Z > u) = F(z). \tag{9}$$

for some scaling factor σ, then:

$$F(z) = 1 - e^{-z}.$$

The limiting distribution in (9) is an exponential distribution which belongs to the family of generalized Pareto distributions.

For a fixed choice of $n \in \mathbb{N}$, the PPM of exceedances associated to Z is then defined on regions of the form $]0, 1[\times]u, +\infty[$:

$$\widetilde{Z}_n^e = \left\{(\frac{i}{n+1}, Z_i) | 1 \le i \le n\right\} \cap\,]0, 1[\times]u, +\infty[,$$

where we use the superscript e to denote exceedances. The indices are divided by the factor $n + 1$ to rescale the process to the interval $]0, 1[$ as illustrated in Figure 2.

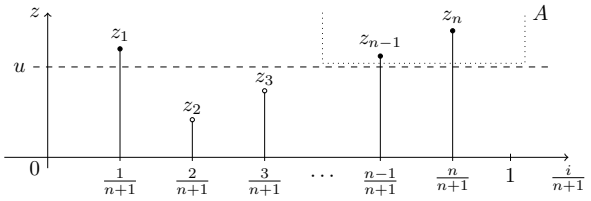


Figure 2: A realization $\widetilde{Z}_n^e$ of a point process of exceedances with $N_A^n(\omega) = 2$.

The link between PPMs and EVT is obtained by letting $n \to +\infty$ and $u \to +\infty$ [17, 5]. It can be shown that when the limit in (9) holds for the random variable Z for some scale parameter σ, the corresponding sequences of PPMs of exceedances $\widetilde{Z}_n^e$ will converge to a Poisson point process (PPP) for large u, meaning that the corresponding sequence of counting measures N_A^n associated with $\widetilde{Z}_n^e$ converge in distribution to a Poisson distribution:

$$N_A^n \xrightarrow{d} \mathrm{Pois}(\Lambda(A)) \text{ as } n \to +\infty,$$

on sets

$$A =]t_1, t_2[\times]w, +\infty[, \quad w > u, \quad]t_1, t_2[\subset]0, 1[$$

[3] A space where points are identified with respect to some equivalence relation.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

and where the intensity measure $\Lambda(A)$ can be parametrised in terms of the scale parameter σ and a rate parameter λ:

$$\Lambda(A) = (t_2 - t_1) \exp\left(-\frac{w-u}{\sigma}\right)\lambda.$$

Thus, for large u and n, the PPM of exceedances on $]0,1[\times]u,+\infty[$ can be approximated by a PPP where the number of exceedances is distributed according to a Poisson distribution with a rate parameter λ and where the locations of the exceedances are distributed according to an exponential distribution with scale σ, whenever the limit in (9) holds. The choice of u for this approximation to be valid is well-studied in the field of EVT and can be assessed by means of a *mean excess* plot. The latter is a graphic diagnostic tool in which the sample means of the excesses $(Z-u)$ are plotted against a range of thresholds [13]. When the approximation is valid for $u > u_0$, this plot should be linear for $u > u_0$. Alternatively, an empirical rule-of-thumb can be chosen that specifies the tail fraction of exceedances above u. One commonly-used choice is to set u as the quantile at $1 - \frac{n^{2/3}}{n\log\log(n)}$ of a sample of length n of the distribution Z [23]. The parameters σ and λ may then be estimated by means of maximum likelihood estimation.

5. Novelty detection for point patterns

In this section, we treat the problem of classifying patterns as introduced in Section 2:

$$\tilde{\mathbf{x}} = \{\mathbf{x}_1, \dots \mathbf{x}_N\}, \tag{10}$$

with respect to a distribution X modelling location in space $\mathbb{R}^d$ and a distribution N modelling the stochastic behaviour of the length of the pattern. To tackle this novelty detection problem in its full generality, a PPM is proposed where $\tilde{\mathbf{x}}$ is viewed as a realised point pattern of the i.i.d. cluster process $\widetilde{X}$ associated with the random variable X.

In Section 5.1 the infinite-dimensional study of the distribution $\widetilde{X}$ on the quotient space of configurations is translated into a one-dimensional study by considering a distribution of Janossy densities $\tilde{f}(\tilde{\mathbf{x}})$. In Section 5.2 it is shown that for a normal distribution X, this distribution is analytically tractable. In Section 5.4, PPMs of multivariate exceedances are introduced that completely characterise the low-density regions of a point pattern and this yields a model that unifies the methods previously introduced in [14, 15, 17].

5.1. Distributions of Janossy densities

Consider a finite i.i.d. cluster PPM $\widetilde{X}$, as introduced in Section 4.3, defined by a random variable N describing the length of the point patterns and a multivariate distribution X with a PDF $y = f(\mathbf{x})$ on $\mathcal{D} \subset \mathbb{R}^d$ describing the locations of the points. In this section a numerical method is proposed to evaluate patterns realised by $\widetilde{X}$ that fall in a subset $A \subset \mathcal{D}$.

For this purpose consider a given subset $A \subset \mathcal{D}$ and denote $\widetilde{X}_A$ as being the PPM describing those points within the point patterns $\tilde{\mathbf{x}}$ that fall in A:

$$\tilde{\mathbf{x}}_A \;=\; \tilde{\mathbf{x}} \cap A \;=\; \{\mathbf{x} \in \tilde{\mathbf{x}} | \mathbf{x} \in A\}.$$

The length of these patterns is governed by a discrete distribution $\eta_k = P(N_A = k)$ as given in (6) depending on the random variable N. We remark that $\widetilde{X}_{\mathcal{D}} = \widetilde{X}$. The distribution of all possible point patterns $\tilde{\mathbf{x}}_A$ on $A \subset \mathcal{D}$ is impossible to visualise. Therefore it can be very useful to reduce the analysis of a PPM to the study of a univariate variable V describing the Janossy densities of point patterns $v = \tilde{f}(\tilde{\mathbf{x}}_A)$ distributed according to some cumulative distribution function (CDF) $G(v)$. This distribution will make it possible to evaluate configurations of point patterns $\tilde{\mathbf{x}}_A$ by using novelty scores that have a suitable probabilistic meaning.

The corresponding CDF describes the probability that we might observe a point pattern with a Janossy density that is smaller than some density v. In particular the following event in the probability space of $\widetilde{X}$ is considered:

$$\begin{aligned}
\mathcal{X}_v &= \left\{\tilde{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2 \dots) \middle| (\tilde{f}(\tilde{\mathbf{x}}) \le v) \wedge (\mathbf{x}_i \in A)\right\} \\
&= \bigcup_{k\ge 0} \left\{\tilde{\mathbf{x}} = (\mathbf{x}_1, \dots, \mathbf{x}_k) \middle| (\tilde{f}(\tilde{\mathbf{x}}) \le v) \wedge (\mathbf{x}_i \in A)\right\} \\
&= \bigcup_{k\ge 0} A_v^k,
\end{aligned} \tag{11}$$

where we introduced the disjoint sets:

$$A_v^k = \left\{(\mathbf{x}_1, \dots, \mathbf{x}_k) | \tilde{f}(\mathbf{x}_1, \dots, \mathbf{x}_k) \le v\right\}, \quad k \ge 1. \tag{12}$$

and

$$A_v^0 = \begin{cases} \{\emptyset\} & \text{when} \quad v \ge \eta_0 \\ \emptyset & \text{when} \quad v < \eta_0, \end{cases}$$

with $\eta_0 = \tilde{f}(\{\emptyset\})$ being the probability for $\widetilde{X}_A$ to generate an empty point pattern. The univariate distribution $G(v)$ is now given by the probability of $\mathcal{X}_v$; i.e., $G(v) = \mathbb{P}(\mathcal{X}_v)$ or:

$$G(v) = \sum_{k\ge 0} \mathbb{P}(A_v^k).$$

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

Applying the same reasoning used in (7) for the events A_v^k and using the abbreviation $d\tilde{\mathbf{x}}$ for $d\mathbf{x}_1 \cdots d\mathbf{x}_k$, one finds:

$$\begin{aligned} G(v) &= \eta_0 H(v-\eta_0) + \sum_{k\geq 1} \int_{A_v^k} \frac{1}{k!} \tilde{f}(\{\mathbf{x}_1, \ldots, \mathbf{x}_k\})\, \mathrm{d}\tilde{\mathbf{x}} \\ &= \eta_0 H(v-\eta_0) + \sum_{k\geq 1} \eta_k \prod_{i=1}^{k} \int_A f(\mathbf{x_i}) \mathrm{d}\mathbf{x_i}, \quad (13) \end{aligned}$$

where we have used the Heavyside step function:

$$H(v) = \begin{cases} 1, & v \geq 0, \\ 0, & v < 0 \end{cases} \quad (14)$$

to describe the probability mass associated with an empty pattern. The distribution $G(v)$ can in any case be calculated numerically for a given PDF $y = f(\mathbf{x})$; e.g., obtained by a kernel density estimation (KDE) [24]. For this purpose, $G(v)$ can be approximated by an empirical CDF of Janossy densities of point patterns simulated from the cluster PPM $\widetilde{X}_A$.

5.2. Point patterns of multivariate normal distributions

In this section an analytic expression is derived for the CDF $G(v)$ and PDF $g(v)$ of Janossy densities of an i.i.d. cluster PPM $\widetilde{X}$ associated with a random variable X distributed according to a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In particular, in equation (13), one sets:

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d\, |\boldsymbol{\Sigma}|}} \exp\left(-\frac{M(\mathbf{x})^2}{2}\right) \quad (15)$$

with:

$$M(\mathbf{x}) = ((\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}))^{1/2}.$$

In the theorem below we will show that the distribution of V is one of mixed type [25]. Random variables of mixed type are neither discrete or continuous, but are a mixture of both. The discrete component results from the Janossy density η_0 of empty patterns which have a strictly positive probability mass $\eta_0 > 0$. This implies a discontinuity in the CDF that we will describe using the Heavyside step function (14).

Theorem 1 *Consider the random variable V describing Janossy densities $v = \tilde{f}(\tilde{\mathbf{x}})$ of random point patterns drawn from an i.i.d. cluster PPM $\widetilde{X}$ of normal distributed variables $X_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The distribution of V is one of mixed-type with CDF:*

$$G(v) = \eta_0 H(v-\eta_0) + \sum_{k\geq 1} \eta_k \left[1 - F_{kd}\left(-2\log\left(\frac{c_k v}{\eta_k}\right)\right)\right]$$

where $c_k = \frac{\sqrt{(2\pi)^{kd}|\boldsymbol{\Sigma}|^k}}{k!}$ and for each $k > 0$, F_{kd} denotes the CDF of a chi-squared distribution with kd degrees of freedom. Furthermore, conditioned on the non-empty patterns, the random variable V has a PDF given by:

$$g(v|\tilde{\mathbf{x}} \neq \emptyset) = \sum_{k\geq 1} \frac{c_k}{\overline{\eta}_0 \Gamma(\frac{kd}{2}) 2^{\frac{kd}{2}-1}} \left(-2\log\left(\frac{c_k}{\eta_k} u\right)\right)^{\frac{kd-2}{2}}$$

with $\overline{\eta}_0 = 1 - \eta_0$.

Proof We proceed by the derivation of the distribution of the random variable $W = -2\log(V)$ whereafter a transformation $V = e^{-\frac{W}{2}}$ recovers the original distribution. From (8), one finds for $k > 0$:

$$\begin{aligned} w &= -2\log\left(k!\eta_k \prod_{i=1}^{k} f(\mathbf{x}_i)\right) \\ &= -2\log(k!\eta_k) - 2\sum_{i=1}^{k} \log f(\mathbf{x}_i) \\ &= -2\log(k!\eta_k) + 2\log\left(\sqrt{(2\pi)^{kd}\,|\boldsymbol{\Sigma}|^k}\right) + \sum_{i=1}^{k} M(\mathbf{x}_i)^2 \\ &= 2\log\left(\frac{c_k}{\eta_k}\right) + \sum_{i=1}^{k} M(\mathbf{x}_i)^2 \end{aligned}$$

where $c_k = \frac{\sqrt{(2\pi)^{kd}|\boldsymbol{\Sigma}|^k}}{k!}$ and clearly $w > 2\log\left(\frac{c_k}{\eta_k}\right)$. Conditioned on the length $k > 0$ of a non-empty pattern, the random variable $W - 2\log\left(\frac{c_k}{\eta_k}\right)$ is given by a sum of k Mahalanobis distances which is distributed according to a chi-squared distribution with k degrees of freedom [25]. Therefore the CDF $\overline{G}(w)$ associated with $W = -2\log(V)$ is given by:

$$\begin{aligned} \overline{G}(w) &= P(W < w) \\ &= \sum_{k\geq 0} \eta_k P(W < w | N = k) \\ &= \eta_0 H(w + 2\log\eta_0) + \sum_{k\geq 1} \eta_k P(W < w | N = k) \\ &= \eta_0 H(w + 2\log\eta_0) + \sum_{k\geq 1} \eta_k F_{kd}\left(w - 2\log\frac{c_k}{\eta_k}\right). \end{aligned}$$

To recover the distribution of V we transform back by means of $V = e^{-\frac{W}{2}}$:

$$\begin{aligned} G(v) &= 1 - \overline{G}(-2\log v) \\ &= \eta_0 H(v-\eta_0) + \sum_{k\geq 1} \eta_k \left[1 - F_{kd}\left(-2\log\frac{c_k v}{\eta_k}\right)\right] \end{aligned}$$

where we have used the fact that $\sum_{k\geq 0} \eta_k = 1$. This yields us the expression for the CDF $G(v)$.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

For the conditional PDF $g(v|\tilde{\mathbf{x}} \neq \emptyset)$, we proceed by calculating the conditional probability

$$G_0(v) = P(V \leq v | N > 0),$$

which is given by:

$$\begin{aligned} G_0(v) &= \frac{P(V \leq v, N > 0)}{P(N > 0)} \\ &= \sum_{k \geq 1} \frac{\eta_k}{\overline{\eta}_0} \left[1 - F_{kd}\left(-2 \log \frac{c_k v}{\eta_k} \right) \right] \end{aligned}$$

where $\overline{\eta}_0 = 1 - \eta_0$. Using the expression of the PDF of a chi-squared distribution one finds:

$$G_0(v) = \sum_{k \geq 1} \frac{\eta_k}{\overline{\eta}_0 \Gamma\left(\frac{kd}{2}\right) 2^{\frac{kd}{2}}} \int_{\alpha_k(v)^2}^{+\infty} u^{\frac{kd}{2}-1} e^{-u/2} \mathrm{d}u.$$

where $\alpha_k(v)^2 = -2 \log\left(\frac{c_k}{\eta_k} v\right)$ for $v > \frac{c_k}{\eta_k}$ and zero elsewhere. Using the substitution $u = \frac{\eta_k}{c_k} e^{-\rho_k^2/2}$ with inverse $\rho_k = \sqrt{-2 \log(\frac{c_k}{\eta_k} u)}$, this integral simplifies to:

$$G_0(v) = \sum_{k \geq 1} \frac{c_k}{\overline{\eta}_0 \Gamma(\frac{kd}{2}) 2^{\frac{kd}{2}-1}} \int_0^v \left(-2 \log(\frac{c_k}{\eta_k} u) \right)^{\frac{kd-2}{2}} \mathrm{d}u \quad (16)$$

yielding the expression of the conditional PDF $g(v)$ given earlier. ■

5.3. Examples and special cases

As an example, Figure 3(a) shows the CDFs of log-transformed Janossy densities w drawn from multi-dimensional standard normal distributions where the number of dimensions is respectively given by d = 2,3, and 4. The length N of the patterns is distributed according to a binomial $B(n, p)$ with parameters $n = 20$ and $p = 0.7$. At $w = -2 \log \eta_0$ a discontinuity appears as is shown in Figure 3(b) for $d = 3$. It is clear that the Janossy densities are generally decreasing with increasing dimension. This is a consequence of increasing dimensionality implying that the mass of the Gaussian distribution moves away from the mode [24].

The special case in which the length of the point patterns is fixed and known is worth studying in more detail. In this case the random variable N describing the length of patterns is deterministic meaning that $\eta_k = 1$ for some $k > 0$ and all other $\eta_i = 0$ such that:

$$G(v) = \int_0^v \frac{c_k}{\Gamma\left(\frac{kd}{2}\right) 2^{\frac{kd}{2}-1}} \left(-2 \log(\frac{c_k}{\eta_k} u) \right)^{\frac{kd-2}{2}} \mathrm{d}u$$

For $k = 1$, one obtains the distribution of densities of samples $\mathbf{x}$ drawn from a normal distribution with PDF:

$$g(v) = |\mathbf{\Sigma}|^{1/2} \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^v (-2 \log(c_1 u))^{\frac{d-2}{2}} \mathrm{d}u, \quad (17)$$

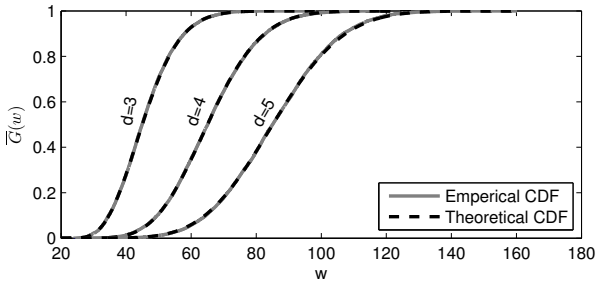
obtaining a result that was previously found for the special case in [14].

5.4. Multivariate point patterns of exceedances

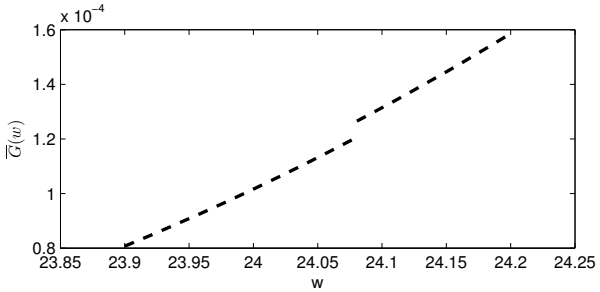
As mentioned in Section 5.1 the distribution of Janossy densities (13) is not analytically tractable for general random variables. However, in this section it is shown that the distribution of Janossy densities of point patterns that occur in low-density regions of a multivariate distribution can be analytically approximated using EVT.

Consider a i.i.d. cluster PPM $\widetilde{X}$ associated with the random variable of X distributed according to a PDF $y = f(\mathbf{x})$ on $\mathcal{D} \subset \mathbb{R}^d$ such that point patterns (almost surely) have a minimal length $n_{\min}$. In this section a PPM of exceedances (PPM-ex) $\widetilde{X}_u^e$ is defined describing patterns in low-density region $A_u (u \in \mathbb{R})$ w.r.t. $y = f(\mathbf{x})$; i.e.,

$$A_u = \{\mathbf{x} | f(\mathbf{x}) < e^{-u}\}. \quad (18)$$



(a)



(b)

Figure 3: (a) Distribution of the log-transformed Janossy densities of point patterns drawn from a i.i.d. cluster PPM associated with a multidimensional standard normal distribution where the number of dimensions is respectively given by 3,4 and 5. (b) The discontinuity that appears at $-2 \log \eta_0$ for $d = 3$.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

A realisation of $\widetilde{X}_u^e$ is a pattern of exceedances, defined as those points of a point pattern $\tilde{\mathbf{x}}$ realized by $\widetilde{X}$ that fall in the low-density region:

$$\tilde{\mathbf{x}}^e = \{\mathbf{x} \in \tilde{\mathbf{x}} | f(\mathbf{x}) < e^{-u}\}.$$

In the theorem that follows, an analytic approximation of the distribution of the PPM-ex is found when $n_{\min}$ and u are sufficiently large.

Using the transformation $z : \mathbb{R}^d \mapsto \mathbb{R} : \mathbf{x} \mapsto -\log f(\mathbf{x})$ the PPM $\widetilde{X}$ is transformed into an i.i.d. cluster process $\widetilde{Z}$ associated with the random variable Z of NLLs. An observed non-empty point pattern of exceedances:

$$\tilde{\mathbf{x}}^e = \{\mathbf{x}_1, \dots \mathbf{x}_N\} \cap \{\mathbf{x} | f(\mathbf{x}) < e^{-u}\}$$

is transformed into a point pattern of univariate exceedances of $Z_i = -\log f(X_i)$ above u:

$$\tilde{\mathbf{z}}^e = \{z_1, \dots z_N\} \cap]u, +\infty[= \{z_1^e, \dots z_K^e\}, \qquad (19)$$

where K denotes the counting variable associated with the PPM-ex $\widetilde{X}_u^e$. From Section 4.4, we know that for large $n_{\min}$ and u, the distribution of K, conditioned on the length of the underlying point pattern $\tilde{\mathbf{x}}$ of $\widetilde{X}$, is Poisson with a rate λ:

$$P(K = k | N = n) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad (20)$$

and the exceedances are distributed according to an exponential distribution with scale σ, whenever the distribution of Z satisfies the limiting property in (9). Therefore, the Janossy density of an observed point pattern of $K = k$ extremes (19) can be approximated by:

$$\tilde{f}(\tilde{\mathbf{z}}^e) = k! \alpha_k \prod_{i=1}^{k} \frac{1}{\sigma} \exp\left(-\frac{z_i^e - u}{\sigma}\right), \qquad (21)$$

where we have introduced $\alpha_k = P(K = k)$ and by using (20):

$$\alpha_k = \sum_{n \geq n_{\min}} \eta_n P(K = k | N = n) = \sum_{n \geq n_{\min}} \eta_n \frac{\lambda^k}{k!} e^{-\lambda}. \qquad (22)$$

For $k = 0$ one obtains the likelihood of an empty point pattern of extremes:

$$\alpha_0 = \sum_{n \geq n_{\min}} \eta_n e^{-\lambda}. \qquad (23)$$

In the following theorem an analytical expression is obtained for the distribution of Janossy densities $f(\tilde{\mathbf{z}}^e)$ of the univariate point patterns of exceedances $\{z_1^e, \dots z_K^e\}$.

Theorem 2 *Consider the random variable V^e describing Janossy densities $v^e = f(\tilde{\mathbf{z}}^e)$ as defined in* (21) *of point patterns of exceedances that are distributed according to an exponential distribution with scale parameter σ. The CDF of V^e is given by:*

$$G^e(v^e) = \alpha_0 H(v^e - \alpha_0) + \sum_{k \geq 1} \alpha_k \left[1 - F_{2k}\left(-2 \log\left(\frac{\sigma^k}{k! \alpha_k} v^e\right)\right)\right]$$

where F_{2k} denotes the CDF of a chi-squared distribution with $2k$ degrees of freedom.

Proof We proceed as in the proof of Theorem 1 and first determine the distribution of $W^e = -2 \log V^e$. From (21), one finds for $k > 0$:

$$w^e = -2 \log\left(\frac{k! \alpha_k}{\sigma^k}\right) + \sum_{i=1}^{k} \frac{z_i^e - u}{\sigma}.$$

The rescaled exceedances $\frac{Z_i^e - u}{\sigma}$ are distributed according to an expontial with scale 1 which coincides with a chi-squared distribution with 2 degrees of freedom. Conditioned on a length $k > 0$, the random variables $W^e - 2 \log\left(\frac{\sigma^k}{k! \alpha_k}\right)$ are therefore distributed according to a chi-squared distribution with $2k$ degrees of freedom and thus the CDF $\overline{G}^e(w^e)$ associated with W^e is given by:

$$\begin{aligned} \overline{G}^e(w^e) &= \sum_{k \geq 0} \alpha_k P(W^e < w^e | K = k) \\ &= \alpha_0 H(w^e + 2 \log \alpha_0) + \sum_{k \geq 1} \alpha_k F_{2k}\left(w^e - 2 \log\left(\frac{\sigma^k}{k! \alpha_k}\right)\right) \end{aligned}$$

Transforming back to the original distribution, one obtains the desired result for $G^e(v^e)$. ■

Figure 4 (a)-(c) shows the analytic approximation obtained in Theorem 2 of point patterns of exceedances drawn from an i.i.d cluster PPM associated with a Gaussian mixture model (GMM) X with 4 components. The value of $G^e(w^e)$ corresponding to the pattern of exceedances $\tilde{\mathbf{z}}_0^e$ is given by 84%. Therefore, at a significance level of 5%, one cannot reject the null hypothesis that the sample $\tilde{\mathbf{x}}_0$ was generated from X.

The approximation obtained in Theorem 2 holds whenever the PPM of exceedances associated with Z is approximately a PPP for large u and $n_{\min}$ as noted in Section 4.4. For Gaussian mixture models X, a minimum length of $n_{\min} = 10$ may already lead to valid approximations [16, 14].

6. Experimental results

In this section the finite PPM introduced in Sections 5.1 and 5.2 and the PPM-ex model introduced in Sec-

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

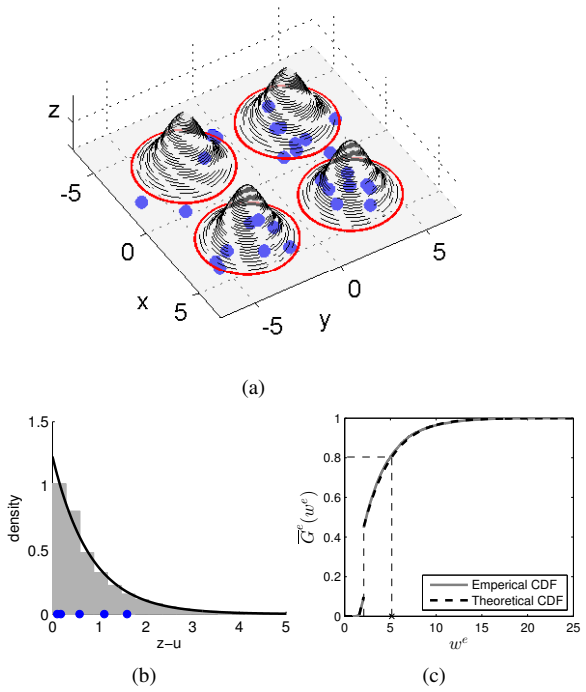


Figure 4: (a) A GMM of 4 components centered at $(\pm 3, \pm 3)$ and with $\boldsymbol{\Sigma} = \boldsymbol{I}$. The dots depict a pattern $\tilde{\mathbf{x}}_0$ with a length that is distributed according to a binomial distribution $B(50, 0.7)$. For this pattern 5 points exceed the contour corresponding to a threshold u on the NLLs determining the tail of the distribution. (b) Histogram of exceedances above u of NLLs drawn from the GMM and the point pattern of exceedances $\tilde{\mathbf{z}}_0^e$ corresponding to $\tilde{\mathbf{x}}_0$. For large u the PDF is approximately exponential. (c) Emperical CDF of Janossy densities of 10^4 patterns of exceedances generated from the GMM after the transformation $w = -2\log(v)$ together with the analytic approximation obtained via Theorem 2. The transformed Janossy density $w^e = 5.14$ of the pattern of exceedances corresponding to $\tilde{\mathbf{x}}_0$ is indicated with a cumulative probability of 84%.

tion 5.4, are demonstrated using three datasets. In Section 6.1, the responsiveness of the PPM to changes of the variance of the components in GMMs is studied using artificial data. Section 6.2 examines the detection of adverse outcomes in patients during their stay in a post-operative ward, using vital-sign data acquired in a clinical trial at the Oxford University Hospitals [26]. Finally, in Section 6.3, an online novelty detection problem is considered that is based on the PPM-ex model to monitor the extremes of a time-series. Calculations[4] were performed in Matlab® and R 3.4.2 [27, 28].

[4]A Matlab® & R library and accompanying data sets supporting the results of Sections 6.1 and 6.3 are available on `www.kuleuven.be/advise`

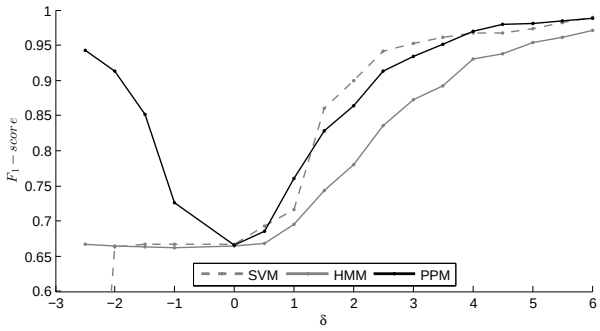
6.1. Capability of industrial processes

In the field of statistical process control, a vital part of an overall quality-improvement program of a manufacturing process is capability analysis [29]. One way to express process capability is by the ratio of the natural or inherent variance a process experiences and the range of the specification limits in which it is allowed to operate. When the natural variance of the process increases, the capability ratio decreases indicating a decrease in the overall quality of the manufacturing process. Conversely a decrease in natural variance can indicate unnecessarily precision of the process, which may be too expensive to maintain in practice.

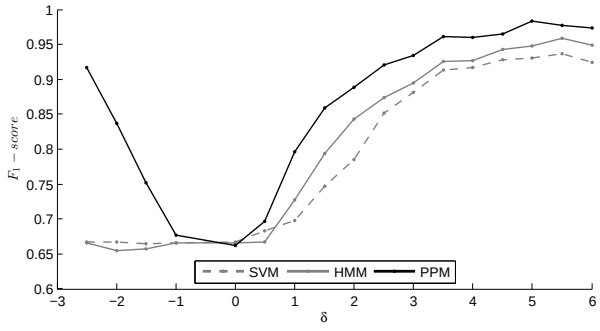
In this section, two experiments are considered where data from the normal class are generated via GMMs. In a first experiment, a multivariate normal distribution is considered centred at the origin with a covariance matrix given by $\boldsymbol{\Sigma} = 3\boldsymbol{I}_3$, where $\boldsymbol{I}_3$ denotes the three- dimensional identity matrix. In a second experiment a GMM is considered with two modes centered at respectively the origin and $(3, 3, 3)$ with covariance matrices that are respectively given by $3\boldsymbol{\Sigma} = 3\boldsymbol{I}_3$ and $\boldsymbol{I}_3$. In both experiments, the performance of a one-class classifier is studied when a change of $\delta\boldsymbol{I}_3$ is adapted to the covariance matrix $3\boldsymbol{I}_3$ of the component centred at the origin, where δ ranges from -2.5 to 3 with a step size of $\Delta\delta = 0.5$.

Training data consists of 400 patterns with a length governed by a binomial distribution with a probability parameter $p = 0.8$ and $n = 25$ trials. For each change of the covariance matrix 320 patterns were simulated that constitute the abnormal class. (The choice of 320 ensures that test and validation sets are balanced in 5-fold cross-validation). Training involves a 5-fold cross-validation process, where in each run, the data set is randomly partitioned into 5 subsets; one subset is used for training; two subsets are used to optimize the hyperparameters in a validation step; and two subsets are hold out for testing. F_1 scores on the test data are used to compare performances of our models, which is the harmonic mean of precision and recall [31]. The PPMs are trained using a KDE with isotropic Gaussian kernels (i.e., the covariance matrix $\boldsymbol{\Sigma} = \sigma\boldsymbol{I}_n$, is given by a scalar multiple of the identity matrix in n variables) to estimate a distribution X of the training data. Performance scores are compared with a OCSVM algorithm (ν-SVM with a Gaussian kernel [6]) and a HMM [30], which are commonly-used methods for sequence classification, as described earlier [10]. During training, the kernel width of the OCSVM and the KDE was varied in the range $[10^{-3}, 0.3]$. The number of states of the HMM was varied between 3, 6, and 9, where each state

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019



(a)



(b)

Figure 5: F_1 scores for the PPMs and one-class classifiers: the OCSVM and the HMM. (a) F_1 scores when the covariance matrix of a multivariate normal distribution is changing from $3\boldsymbol{I}_3$ to $(3+\delta)\boldsymbol{I}_3$. (b) F_1 scores where a GMM model consisting of two components at $(0,0,0)$ and $(3,3,3)$ with covariance matrices $3\boldsymbol{I}_3$ and $\boldsymbol{I}_3$ is perturbed by changing the covariance matrix of the first component to $(3+\delta)\boldsymbol{I}_3$.

presented a Gaussian mixture with a number of components that varied between 1 and 2. The novelty scores for each of the points within a sequence that are obtained by applying the OCSVM method are combined by calculating the mean score.

Figure 5(a) shows the results of the simulation based on the unimodal multivariate normal distribution. In this experiment, Theorem 1 is used to obtain novelty scores of patterns in the test sets. Performance of the models are compared using the F_1 score which considers both sensitivity (SS) and positive predictive value (PPV) [32]. When the variance increases, the PPMs are competitive with the OCSVM and outperform a HMM. However, none but the finite PPM is able to detect a decrease in variance. For this a PPM, $\widetilde{X}_A$ is considered for the region A that contains 50% of the training data and that is estimated using a KDE. When variance decreases, sequences situated in these high-density regions are indeed expected to be longer inducing a decrease in the probabilities $\eta_k = P(N_A = k)$, (6), and hence in their Janossy density.

Figure 5(b) shows the results of the simulation performed using a GMM consisting of two components. Novelty scores of patterns in the test sets are obtained numerically by simulation. In this experiment, one sees that the finite PPM is able to outperform each method indicating that a change in the spatial configuration of patterns drawn from the perturbed GMM is more prominent than a change in the boundary of the normal class.

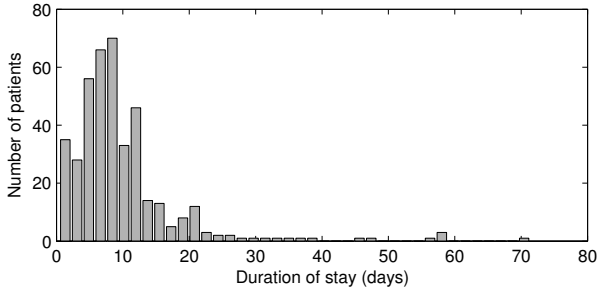
6.2. Predictive monitoring of patients

In this section a clinical data set is considered coming from a study that is carried out at the Oxford Cancer Hospital in the Oxford University Hospitals NHS Trust (Oxford, UK) [26]. During this study, 407 patients were monitored using bedside monitors during a stay in a post-operative ward after an upper-gastrointestinal cancer sugery. The data set involves manual observations of heart rate (HR), systolic blood pressure (BP), and respiratory rate (RR) that are taken at regular times with a mean of 3.5 hours between two observations (but which can rise to as long as 14 hours). The novelty detection problem in this section addresses the prediction of physiologically deterioration resulting in adverse outcomes (such as readmission to the intensive care unit, or death) by detection of novelties in the observation sequence of a patient. The data set is highly unbalanced, where only 13% of the patients suffered from post-surgical complications at some point in their stay.

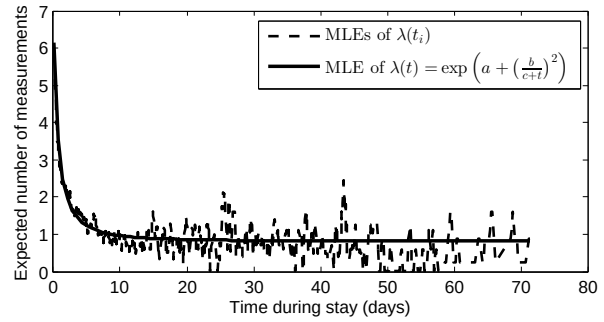
Clinical guidance in the UK recommends the use of an early warning score (EWS) system in combination with vital-sign measurements [33]. This system applies a univariate scoring to each vital sign and warns of patient risk when any of the scores, or the sum of all scores, exceed some threshold. However this current standard of care treats each vital sign independently and thus disregards the correlations between vital signs. Furthermore, it is expected that information is contained in the length of observation sequences of patients, as a higher frequency of measurement indicates increased concern of the clinician taking the measurements [34]. A PPM will allow us to combine information obtained from the values of the vital signs with that obtained from the measurement frequency.

As is clear from Figure 6(a), the length of the stay in the post-operative ward varies substantially from patient to patient. The distribution of the durations is skewed with a maximum of 71 days and a mean of 10 days. It is expected that shortly after the surgery (at $t = 0$) care is given at a higher level and so more measurements are expected to take place during the start of each patients' stay on the ward. Figure 6(b) shows the expected number of measurements taken in the past 5 hours as a func-

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019



(a)



(b)

Figure 6: (a) Histogram of the length of stays of 407 patients staying in the post-operative ward after an upper-gastrointestinal cancer sugery. (b) The expected number of measurements as a function of time t taken in the past 5 hours in the post-operative ward.

tion of time t (the length of this window is empirically chosen to be 5 hours for illustration). The expectation (dashed curve) and its confidence interval is obtained by fitting a Poisson distribution to the number of measurements at each time t_i using a maximum likelihood estimator (MLE). To obtain a continuous estimate at each time t, a model for $\lambda(t)$ was obtained:

$$\lambda(t, a, b, c) = \exp\left(a + \left(\frac{b}{c+t}\right)^2\right)$$

and plugged into the likelihood of all Poisson counts N_{t_i} taken at available times t_i:

$$L(a, b, c) = \prod_{i=1}^{k} \frac{\lambda(t)^{N_{t_i}}}{N_{t_i}!} e^{-\lambda(t_i)}.$$

Maximising L leads to an overall estimate of the parameters (a, b, c) given by $(23.27, 3.62, 0.90)$, as shown in Figure 6(b).

The fit of $\lambda(t)$ was used to apply the PPM. For this purpose, a multivariate normal distribution was fitted to the measurements after an appropriate Box-Cox transformation [29]. To mimic the unbalanced nature of the data set in our experiments, unbalanced test sets were constructed containing 20 abnormal patients and 142 normal patients (noting that only 13% of the patients had a readmission to the intensive care unit or death at the end of their stay). Accounting for the underlying prior probability of readmissions allows us to obtain realistic estimates of how the system will perform in practice. Unlike previous sections, the PPMs were trained in a semi-supervised manner, meaning that the covariance matrix of the Gaussian kernel in a KDE on the training data was chosen based on data from the 'normal class' only (i.e. data from patients not having any post-surgical complication during their stay). This is a pragmatic assumption, as in practice no (or only few data) from the abnormal class are available, making it difficult to optimise parameters using a validation step with cross-validation. The covariance matrix of the KDE was estimated using a minimum covariance determinant estimator which is robust to outliers [35]. Using Theorem 1, risk score for each patient were obtained at each time t by considering patterns of measurements from the past 5 hours. This risk score was threshold at $\overline{G}^e(w^e) = 95\%$ corresponding to a significance level $\alpha = 5\%$.

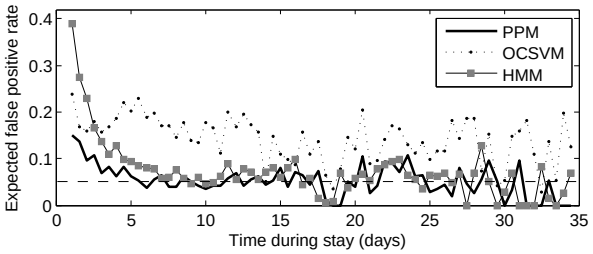
The selection of the patients was randomised in a 5-fold cross-validation experiment to train a HMM and a OCSVM on the point patterns. The same partitions in training and test set were used to calculate the performance scores of a PPM and the EWS system to allow for a consistent comparison. As in Section 6.1, the scores obtained from the OCSVM for each of the points within a pattern were combined by calculating the mean score. For the EWS scores, a pattern was classified as anomalous when one of the scores exceeded the threshold defined by the EWS system [33]. The kernel width of the OCSVM was optimized over the range $[10^{-4}, 10^{-2}]$. The number of states of the HMM was varied between 3, 6, and 9, where each state presented a Gaussian mixture with a number of components that varied between 1 and 2. Table 1 shows performance scores for the various classifiers considered. It may be seen that the PPM outperformed both the HMM and OCSVM. Compared to the clinically-recommended EWS score, the PPM was able to increase the PPV scores without decreasing the SS score, which is one of the challenges when working with unbalanced data sets [2].

Figure 7 shows empirical estimates of the expected FP rates invoked by each classifier during the stay of patients that had no post-surgical complications. Estimates were obtained by averaging rates over the different patients on time intervals of a 0.5 day. The number of stays with a duration longer than 35 days was too small to obtain reliable estimates (see Figure 6(a)). A higher number of FPs is expected during the start of

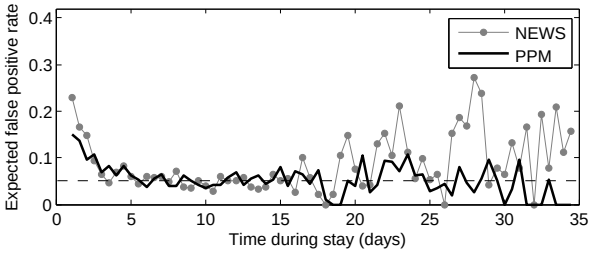
Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

Table 1: F_1, SS, and PPV scores for the prediction of readmissions of patients staying in a post-operative ward. Means and standard deviations are calculated over a 5-fold cross-validation.

Patient	F1	SS	PPV
EWS	55.00 ± 1.70	82.00 ± 3.74	41.48 ± 1.34
PPM	**72.27 ± 1.88**	**82.00 ± 3.39**	**65.00 ± 2.57**
OCSVM	59.44 ± 2.28	65.00 ± 4.74	56.31 ± 3.96
HMM	62.54 ± 1.92	80.20 ± 3.54	51.61 ± 2.11



(a)



(b)

Figure 7: Empirical estimates of the FP rates invoked by the various classifiers considered during the stay of patients that had no post-surgical complications. The dashed line indicates the 5% level. (a) The FP rate of the PPM compared with the HMM and OCSVM. (b) The FP rate of PMM compared with the EWS method.

each stay as during that period more measurements are expected to take place. However, the PPM is able to reduce this increase in FP rate considerably compared to the other classifiers. This is because a PPM enables to prevent those misclassifications induced by the multiple hypothesis testing problem. The OCSVM and HMM classifier both show an increased FP rate during the complete period. Compared to the EWS method, the FP rate of the PPM shows a lower variability and fluctuates around 5% which corresponds to the choice of the 95% threshold on the Janossy likelihoods.

6.3. Online novelty detection

In this section, a public available time series data set is used that consists of the global mean land-ocean temperature index from 1880 to 2016[5]. In particular the data are deviations from measures in degrees centigrade, from the 1951 – 1980 average. To study climate changes, often it's not the distribution of the bulk of deviations that is of interest, but rather the behaviour of patterns in larger deviations [37].

As noted in previous studies, there is an apparent upward trend in the data as indicated by the linear regression fit in Figure 8. In the latter part of the twentieth century, however, there is an increase in the upward trend starting in the period 1980 – 1990 and which may be used as an argument for an acceleration of the global warming hypothesis [38]. We will show in this section how the PPM-ex introduced in Section 5.4 can be used to test whether this increase during the last decennia is statistically significant with respect to the overall increasing trend that is observed. Furthermore there is a levelling off at about 1940 that may be interesting to investigate further.

We will apply an online novelty detector based on the use of a Kalman filter model (KFM) and the PPM-ex approach and will compare its performance with the EVT approach introduced in [14]. In particular, an online linear regression model $y_t = \beta_0 + \beta_1 t + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma^2)$ is fitted to the data through the use of a KFM with the following state space form:

$$\begin{cases} y_t &= H_t^T \xi_t + \epsilon_t, \\ \xi_{t+1} &= \xi_t, \end{cases} \tag{24}$$

[5]Data was downloaded from `https://data.giss.nasa.gov/gistemp/`

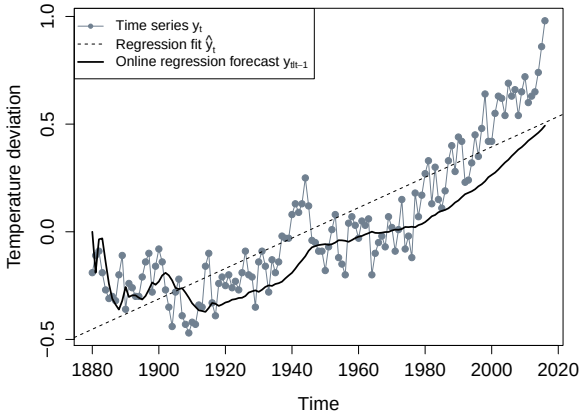


Figure 8: Time series of the global warming dataset[3] consisting of temperature deviations from the 1951 – 1980 average. The dashed line shows a linear regression model. The bold line shows the forecasts $\hat{y}_{t|t-1}$ of an online linear regression model using a KFM where the coefficients are updated at each time step t given the previous data up to time step $t-1$.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

where y_t denotes the temperature deviation at time t, H_t the observation matrix consisting of the explanatory data $H_t = (1\ t)^T$ and ξ_t the state vector consisting of the unknown regression coefficients $\xi_t = (\beta_0\ \beta_1)^T$.

The KFM allows at each time step t an update of the estimated regression coefficients $\hat{\xi}_{t|t-1}$ given the past data. This update can be calculated by a set of well-known recursive linear estimation steps:

$$\hat{\xi}_{t+1|t} = \hat{\xi}_{t|t-1} + P_{t|t-1}H_t(H_t^T P_{t|t-1}H_t + \sigma^2)^{-1}e_t,$$

where $P_{t|t-1}$ denotes the variance-covariance matrix of the estimated coefficients $\hat{\xi}_{t|t-1}$:

$$P_{t+1|t} = P_{t|t-1} - P_{t|t-1}H_t(H_t^T P_{t|t-1}H_t + \sigma^2)^{-1}H_t^T P_{t|t-1}$$

and e_t denotes the error $y_t - y_{t|t-1}$ on the forecast $y_{t|t-1} = H_t^T\hat{\xi}_{t|t-1}$ of y_t at time t given the past observations. The distribution of the random variable Y_t conditioned on the past observations $Y_1, \dots Y_{t-1}$ is given by a normal distribution:

$$p(y_t|y_1, \dots y_{t-1}) \sim N(H_t^T\hat{\xi}_{t|t-1}, S_t),$$

with

$$S_t = \sigma^2 + H_t P_{t|t-1}H_t^T.$$

The first 30 observation are used to find the unknown variance σ^2 and the initial state vector $\xi_{1|0}$ by a maximum likelihood criterion [39]. The matrix $P_{1|0}$ is set to a diffuse prior stating that there is no prior knowledge available about the true regression coefficients $\xi_t = (\beta_0\ \beta_1)^T$.

For each time t the window of the past $n = 10$ measurements $\tilde{y} = \{y_{t-n+1}, \dots, y_t\}$ is considered together with the pattern $\tilde{e} = \{e_{t-n+1}, \dots, e_t\}$ of i.i.d. standardized errors:

$$e_{t-i} := S_t^{-1/2}\left(y_{t-i} - y_{t-i|t-i-1}\right) \sim N(0, 1). \tag{25}$$

The patterns $\tilde{e}$ may now be evaluated by considering the corresponding patterns of exceedances $\tilde{e}^e$ together with their Janossy likelihoods v^e and corresponding cumulative probabilities $G^e(v^e)$ as obtained from Theorem 2. The exponential fit on the exceedances $e_i - u$ is estimated using a MLE and a mean excess plot [13] using the first 30 observations. The scale parameter σ of the exponential model and Poisson rate λ of the Poisson model on the number of exceedances were estimated as $(\hat{\sigma}, \hat{\lambda}) = (3.35, 1)$ for $\hat{u} = 3.5$. The fit can be assessed by a quantile-quantile (QQ) plot that shows he empirical quantiles versus the theoretical quantiles obtained from the fitted exponential distribution, see Figure 9. 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 [4].

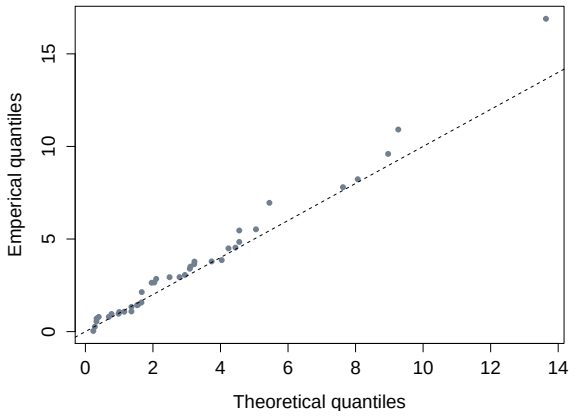


Figure 9: QQ-plot of emperical quantiles of exceedances $e_i - u$ versus quantiles of the exponential fit. If the exponential model approximates well, the points on the graph are expected near the diagonal $y = x$, shown as a dashed line.

Figure 10(a) shows the cumulative probabilities $G^e(v^e)$ at each time step $t \geq n$ associated with the pattern $\tilde{y} = \{y_{t-n+1}, \dots, y_t\}$ of the past $n = 10$ measurements. The scores identify the levelling off starting at 1940 and the increase in the upward trend starting at 1990 by setting a probabilistic threshold of 95%. Figure 10(b) shows the NLLs of the observations y_t with respect to the normal distributions $N(H_t^T\hat{\xi}_{t|t-1}, \sigma^2)$. Commonly used methods for novelty detection rely on a threshold for this likelihoods, e.g. by setting a threshold as the 95% quantile that is estimated during a run-in period of e.g. the first 30 observations. However, even though novelties could be detected, the threshold should be further optimized to overcome the false alarms during the period $1900 - 1920$, which is undesirable for a novelty detection problem. Furthermore, the likelihoods oscillate over the whole period while the cumulative probabilities $G^e(v^e)$ are much easier to interpret as they are low where they have to be and show steep peaks where patterns become less likely. Figure 10(c) show novelty scores that are defined by using a Gumbel model for the most extreme measurement as proposed in [40]. In this approach, at each time t, only the most extreme standardized error is used to evaluate a pattern $\tilde{y} = \{y_{t-n+1}, \dots, y_t\}$ of the past $n = 10$ measurements. In particular, for each time t, the maximum M_n of the NLLs of the standardized errors $\{e_{t-n+1}, \dots, e_t\}$ as defined in (25) is modelled. These maxima will follow ap-

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

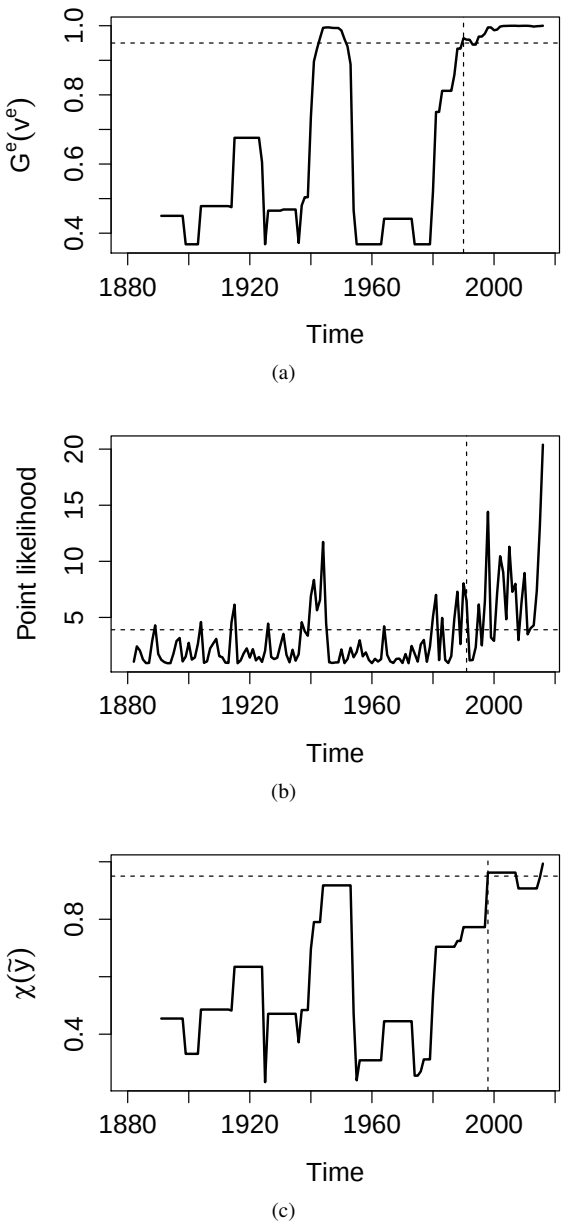


Figure 10: (a) Cumulative probabilities $G^e(v^e)$ of patterns of measurement evaluated with respect to the KFM (24). (b) Likelihoods of observations y_t with respect to the normal distributions $N(H_t^T \hat{\xi}_{t|t-1}, S_t)$. (c) Novelty scores $\chi(\tilde{y})$ of the most extreme errors. Thresholds are shown as dashed lines and correspond to 95% quantiles.

proximately a Gumbel distribution and can be evaluated using:

$$\chi(\tilde{y}) = \exp\left(-\exp\left(-\frac{M_n - \mu}{\sigma}\right)\right),$$

where $\mu = u + \sigma \log \lambda$. As can be seen from figure 10(c), these scores show an increase at the levelling off at about 1940, but do no exceed the probabilistic 95% threshold border. Moreover, the score $G^e(\tilde{v}^e)$ is able to detect the increase in upward trend in the latter part of the twentieth century earlier. The score $\chi(\tilde{y})$ only includes information about the most extreme error and is not able to recognize the change in pattern that occurs before the year 1998.

7. Conclusion

This paper is concerned with the problem of identifying novel point patterns $\tilde{\mathbf{x}} = \{\mathbf{x}_1, ...\mathbf{x}_N\}$ with respect to a statistical distribution X and with a length governed by a discrete random variable N. PPMs explore the spatial configuration contained in $\tilde{\mathbf{x}}$ while jointly modelling the random length of $\tilde{\mathbf{x}}$.

It is shown that the complex distribution of a PPM for a space of point patterns can be translated to a univariate formulation (in Janossy densities) that is readily analysed. For multivariate normal data, our formulation is exact. Moreover, for other multivariate distributions, point patterns occurring in regions of lower density can be evaluated by an analytic approximation that is obtained by the use of EVT. This model is of particular importance when not the bulk of data is of interest, but rather the behaviour of extremes.

We have demonstrated the use of our PPMs using multiple synthetic and real-world data sets, and showed that for these data our models can outperform commonly-used methods for sequence classification, such as HMMs and OCSVMs. For a synthetic data set it was shown how PPMs can detect reductions in variances in the components of Gaussian mixture models by monitoring the expected number of instances in high density regions. The PPM can easily be trained when few or no data of the abnormal class are available in a novelty detection setting. This is in contrast with OCSVMs and HMMs that need the tuning of several hyperparameters. We demonstrated this on a real-world data set consisting of vital signs of patients staying in a postoperative hospital ward. The data set was highly unbalanced as readmission was infrequent compared to the normal state present in the data. For this particular application and data set, the proposed models matched clinical best-practice for sensitivity, while substantially improving PPV in comparison to HMMs and OCSVMs. Furthermore, a real-world time series data set was used to illustrate the use of the method when observations are not independent. The combination of a KFM and the PPM-ex model allowed us to define an online approach to detect novelties in the extremes of time series. In contrast to other EVT methods, the PPM was able to

monitor the spatial configurations of exceedances leading to models that were able to detect changes in patterns rather than detecting point anomalies.

There are several interesting extensions of the method possible for future research. Firstly, an extension of the calculation of the distribution of Janossy densities can be considered for random variables that are not multivariate normally distributed. Secondly, we would like to study how this method can be used to model dependent sequences of point patterns including a temporal variation into the model.

Acknowledgements

This work was partially performed during a research stay at the Institute of Biomedical Engineering in the University of Oxford and was supported by travel grants from COST Action IC1303 (AAPELE) and the Flanders Research Foundation (FWO). David A. Clifton is funded by the Royal Academy of Engineering and an EPSRC Grand Challenge Award.

References

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

[2] C. Nitesh, Data mining for imbalanced datasets: an overview, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer, US, 2005, pp. 853–867.

[3] P. Shaffer, Multiple hypothesis testing, Annual Review of Psychology 46 (1995) 561–584.

[4] D. Clifton, L. Clifton, S. Hugueny, D. Wong, L. Tarassenko, An extreme function theory for novelty detection, IEEE Journal of Selected Topics in Signal Processing 7 (2013) 28–37.

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

[6] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, R. Williamson, Estimating the support of a high-dimensional distribution, Neural Computation 13 (2001) 1443–1471.

[7] N. Abe, B. Zadrozny, J. Langford, Outlier detection by active learning, in: Knowledge discovery and data mining: Proceedings of the 12th ACM SIGKDD international conference, Association for Computing Machinery (ACM), 2006, pp. 504–509.

[8] M. Breunig, H.-P. Kriegel, T. Raymond, J. Sander, LOF : Identifying density-based local outliers, in: Management of data: Proceedings of the 2000 ACM SIGMOD international conference, Association for Computing Machinery (ACM), 2000, pp. 93–104.

[9] M. Pimentel, D. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Processing 99 (2014) 215–249.

[10] T. Dietterich, Machine learning for sequential data: A review, in: Proceedings of the Joint International Workshop on Structural Syntactic and Statistical Pattern Recognition, Springer-Verlag, Londen, 2002, pp. 15–30.

[11] S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering a decade review, Information Systems 53 (2015) 16–38.

[12] K. Muandet, B. Schölkopf, One-class support measure machines for group anomaly detection, in: A. Nicholson, P. Smyth (Eds.), Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI), AUAI Press, Virginia, 2013, pp. 449–458.

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

[14] D. Clifton, S. Hugueny, L. Tarassenko, Novelty detection with multivariate extreme value statistics, Journal of Signal Processing Systems 65 (2011) 371–389.

[15] D. Clifton, L. Clifton, S. Hugueny, L. Tarassenko, Extending the generalised pareto distribution for novelty detection in high-dimensional spaces, Journal of Signal Processing Systems (2013) 1–17.

[16] S. Luca, D. Clifton, B. Vanrumste, One-class classification of point patterns of extremes, Journal of Machine Learning Research 17 (2016) 1–21.

[17] S. Luca, P. Karsmakers, B. Vanrumste, Anomaly detection using the Poisson process limit for extremes, in: R. Kumar, H. Toivonen, J. Pei, Z. H., X. Wu (Eds.), IEEE International Conference on Data Mining, 2014, pp. 370–379.

[18] A. Gelfand, P. Diggle, M. Fuentes, P. Guttorp, Handbook of Spatial Statistics, Chapman & Hall, CRC Press, 2010.

[19] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.

[20] A. Kulesza, B. Taskar, Determinantal point processes for machine learning, Foundations and Trends® in Machine Learning 5 (2012) 123–286.

[21] R. Streit, Poisson Point Processes: Imaging, Tracking, and Sensing, Wiley, New York, 2010.

[22] D. Daley, D. Vere-Jones, An introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods, Springer Verlag, New York, 2003.

[23] M. Falk, J. Hüsler, R.-D. Reiss, Laws of small numbers: Extremes and rare events, 3rd ed., Birkhäuser, 2011.

[24] C. Bishop, Pattern Recognition and machine learning, Springer, New York, USA, 2006.

[25] T. Soong, Fundamentals of probability and statistics for engineers, Wiley, UK, 2004.

[26] M. Pimentel, D. Clifton, L. Clifton, P. Watkinson, L. Tarassenko, Modelling physiological deterioration in post-operative patient vital sign data, Medical & Biological Engineering & Computing 51 (2013) 869–877.

[27] MATLAB, version 7.10.0 (R2013a), The MathWorks Inc., Natick, Massachusetts, 2013.

[28] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[29] D. C. Montgomery, Introduction to statistical quality control, 7th ed., Johan Wiley & Sons, USA, 2013.

[30] L. Rabiner, H. Murray, A tutorial on hidden Markov models and selected applications in speech recognition., Proceedings of the IEEE 77 (1989) 257 – 286.

[31] D. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation, Journal of Machine Learning Technologies 2 (2011) 37–63.

[32] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and F-score, with implication for evaluation, in: D. E. Losada, J. M. Fernández-Luna (Eds.), Advances in Information Retrieval: 27th European Conference on IR Research, Springer, Berlin, Heidelberg, 2005, pp. 345–359.

[33] Royal College of Physicians, National Early Warning Score (NEWS), Standardising the assessment of acute illness severity in the NHS., Technical Report, Royal College of Physicians, 2012. Report of a working party, London.

Preprint submitted at Journal of Computational Statistics & Data Analysis. The final publication is available at
doi.org/10.1016/j.csda.2018.03.019

[34] D. Clifton, L. Clifton, D. Sandu, G. Smith, L. Tarassenko, S. Vollam, P. Watkinson, "Errors" and omissions in paper-based early warning scores: The association with changes in vital signs - a database analysis, British Medical Journal Open 5 (2015) 1–7.

[35] S. Verboven, M. Hubert, LIBRA : a matlab library for robust analysis, Chemometrics and intelligent Laboratory Systems 75 (2005) 127–136.

[36] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves., in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233–240.

[37] D. Cooley, Extreme value analysis and the study of climate change, Climatic Change 97 (2009) 77–83.

[38] R. Shumway, D. Stoffer, Time Series Analysis and its Applications with R examples, 3rd ed., Springer, Berlin, 2011.

[39] S. Särkkä, Bayesian Filtering and Smoothing, Cambridge University Press, UK, 2013.

[40] H. Lee, S. Roberts, On-line novelty detection using the kalman filter and extreme value theory, in: IEEE, 19th International Conference on Pattern Recognition (ICPR), volume 19, IEEE, 2008, pp. 1–4.