J Sign Process Syst (2011) 65:371–389
DOI 10.1007/s11265-010-0513-6

Novelty Detection with Multivariate Extreme Value Statistics

David Andrew Clifton · Samuel Hugueny · Lionel Tarassenko

Received: 15 January 2010 / Revised: 5 May 2010 / Accepted: 20 July 2010 / Published online: 13 August 2010

Abstract Novelty detection, or one-class classification, aims to determine if data are "normal" with respect to some model of normality constructed using examples of normal system behaviour. If that model is composed of generative probability distributions, the extent of "normality" in the data space can be described using Extreme Value Theory (EVT), a branch of statistics concerned with describing the tails of distributions. This paper demonstrates that existing approaches to the use of EVT for novelty detection are appropriate only for univariate, unimodal problems. We generalise the use of EVT for novelty detection to the analysis of data with multivariate, multimodal distributions, allowing a principled approach to the analysis of high-dimensional data to be taken. Examples are provided using vital-sign data obtained from a large clinical study of patients in a high-dependency hospital ward.

Keywords Novelty detection · Extreme value theory · Patient monitoring · Multivariate statistics · Multimodal statistics · Biomedical engineering

1 Introduction

1.1 Novelty Detection

Novelty detection, alternatively termed *one-class classification* or *anomaly detection*, classifies test data as "normal" or "abnormal" with respect to a model of normality. This approach is particularly well suited to problems in which a large quantity of examples of "normal" behaviour exist, such that a model of normality may be constructed, but where examples of "abnormal" behaviour are rare, such that a multi-class approach cannot be taken. For this reason, novelty detection has become popular in the analysis of data from high-integrity systems, such as hospital patients [10, 24, 25], jet engines [4, 11], manufacturing processes [6], or power-generation facilities [27], which spend the majority of their operational life in a "normal" state, and which exhibit few, if any, failure conditions.

Furthermore, the complexity of such systems is usually very high, because they are comprised of large numbers of mutually interacting subsystems, often including very large numbers of components. The number of possible modes of failure in such systems is very large, making explicit modelling of expected failure modes difficult, and the variability in observed behaviour between systems of the same type (such as between jet engines of the same type, or hospital patients of similar ages) can be significant, making it difficult to use fault information obtained from other systems. Thus, the novelty detection approach can be used to best exploit our prior knowledge that the "normal" class is better represented, and more readily defineable, than the large number of ill-defined classes of "abnormality" that may exist.

This work was supported by the NIHR Biomedical Research Centre, Oxford; the EPSRC LSI Doctoral Training Centre, Oxford; and the Centre of Excellence in Personalised Healthcare funded by the Wellcome Trust and EPSRC under grant number WT 088877/Z/09/Z.

D. A. Clifton (✉) · S. Hugueny · L. Tarassenko
Institute of Biomedical Engineering,
Department of Engineering Science,
University of Oxford,
Old Road Campus, Roosevelt Drive,
Oxford OX3 7DQ, UK
e-mail: davidc@robots.ox.ac.uk

1.2 Overview

We first consider existing approaches to novelty detection, in Section 2, and describe the disadvantages associated with such approaches. EVT is introduced in Section 3 as being a principled method of avoiding such disadvantages.

A review of existing work in the use of EVT for novelty detection follows in Section 4, which also describes how existing methods are inappropriate for novelty detection in data which have multivariate or multimodal distributions.

Section 5 proposes a numerical solution for using EVT for novelty detection in multivariate, multimodal applications. This is based on the insight that novelty detection in multivariate data is equivalent to performing novelty detection in the probability space of the model of normality.

A closed-form solution is proposed in Section 6 for novelty detection in multivariate, *unimodal* problems, which avoids the requirement of sampling in the numerical method proposed in Section 5.

An application of the latter proposed technique is given in Section 7, where novelty detection is performed in the monitoring of vital signs obtained from patients in a high-dependency hospital ward.

Conclusions are drawn in Section 8, where potential extensions to the work proposed in this paper are discussed.

2 Existing Work

In much of the existing work on novelty detection, an assumption is made that "normal" data $\{\mathbf{x}_1 \ldots \mathbf{x}_N\}$ are i.i.d.,[1] and distributed according to some underlying generative distribution $f_n(\mathbf{x})$, which is a probability distribution function[2] (pdf) over an n-dimensional data space, $\mathcal{D}$; i.e., $\mathbf{x} \in \mathcal{D} = \mathbb{R}^n$, where n is the number of *features* in the data (or the *dimensionality* of each *feature vector* $\mathbf{x}$). Typically, the underlying distribution f_n is multivariate and multimodal; it could be, for example, approximated using a Gaussian mixture model (GMM), a Parzen window estimator, or some other mixture of components [2]. If we define a null hypothesis H_0 that test data $\mathbf{x}$ are generated from f_n, then novelty detection evaluates the hypothesis that H_0 is true; if H_0 holds with probability $P < \kappa$ for some threshold probability κ, then the null hypothesis is rejected, and $\mathbf{x}$ is classified "abnormal" w.r.t. f_n. The problem then becomes one of setting the novelty threshold.

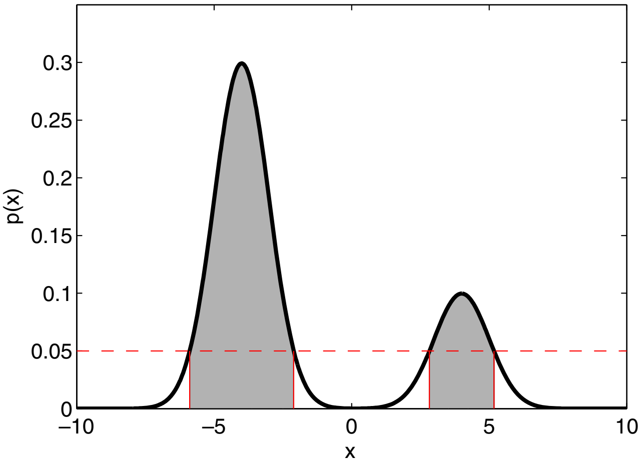

Figure 1 Integrating a bimodal probability distribution $p(x)$ to the contour $p(x) = 0.05$. The contour is shown as a *horizontal dashed line*, intercepting the distribution function at the locations shown by the *vertical straight lines*. The area corresponding to the integration is shown in *grey*.

In previous work [10, 16, 17, 26], a heuristic novelty threshold has been set on the pdf $f_n(\mathbf{x}) = \kappa$, such that $\mathbf{x}$ is classified "abnormal" if $f_n(\mathbf{x}) < \kappa$. Such thresholds are set with no principled probabilistic interpretation: $f_n(\mathbf{x})$ is used simply as a novelty score, and the threshold is set such that separation between "normal" and any "abnormal" data is maximised on a validation dataset. Some authors [10, 14] have interpreted f_n probabilistically, by considering the cumulative probability F_n associated with f_n. That is, they find the probability mass obtained by integrating f_n over the region $\mathcal{R}$ where f_n exceeds the novelty threshold; i.e., the region $\mathcal{R} = \{\mathbf{x} \in \mathcal{D} | \, f_n(\mathbf{x}) \geq \kappa\}$:

$$F_n(\kappa) = \int_{\mathcal{R}} f_n(\mathbf{x}) \, d\mathbf{x} \qquad (1)$$

An example is shown in Fig. 1, in which the distribution f_n is univariate and multimodal, and which has been approximated using a GMM with two components of equal variance ($\sigma_1^2 = \sigma_2^2 = 1$) and with prior probabilities[3] $P(\mu_1) = 0.75$, $P(\mu_2) = 0.25$. A novelty threshold is shown at $f_n(\mathbf{x}) = \kappa = 0.05$ in the figure, where the probability mass P enclosed by that threshold is shaded. The probability mass will fall in the range $0 \leq P \leq 1$. In this example, the shaded areas from each Gaussian

[1] independent and identically distributed

[2] Note that this probabilistic approach contrasts with another popular method of novelty detection, that of one-class support vector machines (SVMs) [21].

[3] alternatively known as *mixing coefficients* or *weights*

component do not significantly overlap, and it is possible to approximate closely the integration in Eq. 1 in closed form, treating each component distribution independently.

For a unimodal pdf, Eq. 1 corresponds to integrating from the mode of f_n to the pdf contour defined by the novelty threshold $f_n(\mathbf{x}) = \kappa$, which can usually be determined in closed form. However, for a multimodal pdf, the integration may need to be performed using Monte Carlo techniques [17]. Setting the novelty threshold using the above method, and then determining the probability mass F_n enclosed by that $f_n(\mathbf{x}) = \kappa$ contour, allows a probabilistic interpretation: if we were to draw one sample from f_n, we would expect it to lie outside the novelty threshold with probability $1 - F_n$. Thus, we could set the novelty threshold $f_n(\mathbf{x}) = \kappa$ such that F_n is some desired probability mass; e.g., $F_n(\kappa) = 0.99$. We note in passing that this approach is equivalent to the *high-density region* (HDR) approach of [14], which has been used for novelty detection [15, 18].

However, setting a novelty threshold using F_n has associated disadvantages for novelty detection. In order to examine these disadvantages, we must first consider a different method for determining the location of novelty thresholds: that of EVT.

3 Classical Extreme Value Theory

EVT is a branch of statistics concerned with modelling the distribution of very large or very small values (*extrema*) w.r.t. a generative distribution f_n. Here, we consider "classical" EVT as previously used in novelty detection [19, 20, 23, 28], in contrast to a method commonly used in estimating financial risks, often termed the peaks-over-threshold (POT) technique [8].

3.1 Extreme Value Distributions

Consider a set of m i.i.d. data $\mathbf{X} = \{x_1, x_2, \ldots, x_m\}$, which are univariate ($n = 1$) for classical EVT, and are distributed according to some pdf $f_1(x)$, with maximum $x_{\max} = \max(\mathbf{X})$. We define the cumulative distribution function (cdf) for $x_{\max}$ to be $H^+(x_{\max} \leq x)$; i.e., H^+ models our belief in where the maximum of m data generated from distribution f_1 will lie.[4] This is directly applicable to novelty detection, in which we wish to determine where the boundary of "normality" lies under normal conditions. We here term H^+ the *extreme value distribution* (EVD), because it describes the expected location of the extremum of m data generated from f_n.

According to the Fisher–Tippett theorem [9] upon which classical EVT is based, H^+ must belong to one of the following three families of distributions,[5] no matter what the form of f_1:

$$\text{Gumbel,} \quad H_1^+(y) = \exp(-\exp(-y)) \tag{2}$$

$$\text{Fréchet,} \quad H_2^+(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ \exp(-y^{-\alpha}) & \text{if } y > 0 \end{cases} \tag{3}$$

$$\text{Weibull,} \quad H_3^+(y) = \begin{cases} \exp(-(-y)^{\alpha}) & \text{if } y \leq 0 \\ 1 & \text{if } y > 0 \end{cases} \tag{4}$$

for $\alpha \in \mathbb{R}^+$, and where y is a transformation[6] of x, $y = (x - c_m)/d_m$, for location and scale parameters c_m and d_m, respectively. In classical EVT, these parameters for the EVD corresponding to the univariate Gaussian are dependent only on the number of data m drawn from the underlying distribution f_1 [8]:

$$c_m = \sqrt{2 \ln m} - \frac{\ln \ln m + \ln 4\pi}{2\sqrt{2 \ln m}}, \qquad d_m = \frac{1}{\sqrt{2 \ln m}} \tag{5}$$

In this paper, we are primarily concerned with mixtures of Gaussian distributions, for which the limiting distribution of maxima is the Gumbel distribution H_1^+. We label this limiting cdf $F_n^e = H_1^+$, which has a corresponding pdf f_n^e (the EVD).[7] Here, the superscript e is used to denote the fact that these are the cdf and pdf of the *extremum* of our dataset $\mathbf{X}$, where that dataset $\mathbf{X}$ has its own cdf F_n and pdf f_n, as described previously.

Note that in the case of $m = 1$ (i.e., when we observe only one data-point from f_n), the EVD is the original generative distribution $f_n^e = f_n$; i.e., the original distribution describes where a single sample drawn from f_n will lie, because the extremum of a singleton set $\{\mathbf{x}\}$ is simply $\mathbf{x}$.

In Section 6.2, we are interested in the Weibull-type EVD of minima, or minimal Weibull, for which the attractor is [3]:

$$H_3^-(x, \alpha) = \begin{cases} 0, & x < 0 \\ 1 - \exp(-(x)^{\alpha}), & x \geq 0 \end{cases}. \tag{6}$$

[4] Noting that the superscript '+' refers to the distribution of maxima.

[5] Noting that this is an asymptotic relationship, which is true as the number of data $m \to \infty$.

[6] termed the *reduced variate*

[7] The EVD f_n^e is implicitly parameterised by the number of observations in the dataset, m. Thus, each value of m will yield a different EVD f_n^e.

3.2 Disadvantages of Using F_n

With a method of describing where we expect the extremum of a set of m samples generated from f_n to occur, we can examine the disadvantages associated with setting a novelty threshold using F_n, as was introduced in Section 2. For the purpose of illustration, suppose that $f_1 = N(0, 1)$, the standard univariate Gaussian distribution.

The upper plot in Fig. 2 shows the generative distribution f_1, and the EVDs f_1^e for increasing numbers m of observed data $\mathbf{x}$. As more data are observed from f_1, the expected location of their maximum increases on the x-axis. This matches our intuition: if we generate large numbers of random data, the extremum of those random data is likely to be more extreme than if we generate small numbers of data. Similarly, in the case of novelty detection, if we observe more "normal" data, distributed according to f_n, then we expect their extremum to be more extreme than if smaller numbers of data are observed.

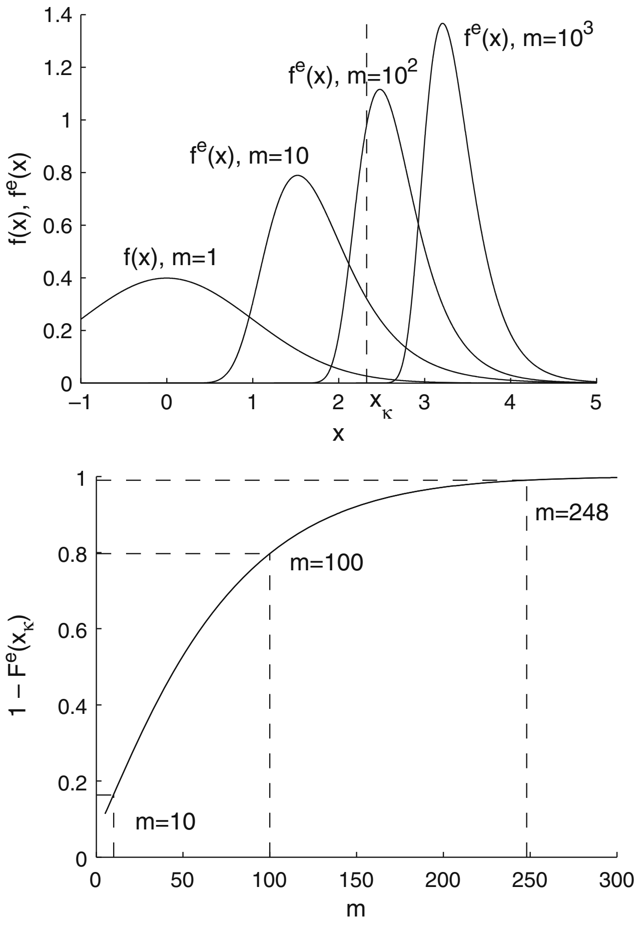

Figure 2 EVDs $f_1^e(\mathbf{x})$ and probability $1 - F_1^e(\kappa)$ in the upper and lower plots, respectively, for increasing number m of data $\mathbf{x}$.

A novelty threshold has been set at $F_1(\kappa) = 0.99$ in the figure, which occurs at $x = x_\kappa$, and which is shown in the upper plot of the figure as a dashed line. We anticipate that this novelty threshold will be exceeded with probability $1 - F_1(\kappa) = 1 - 0.99 = 0.01$ when generating a *single sample* (i.e., $m = 1$) from f_1. However, when generating multiple samples ($m > 1$) from f_1, the probability that the novelty threshold will be exceeded is given instead by the extreme cdf, $1 - F_1^e(\kappa)$, and *not* by the original distribution f_n. This is a key point in the use of EVT for novelty detection: only EVT provides the correct probability distribution of the boundary of "normality" when more than one data-point has been observed.

The lower plot in Fig. 2 shows the probability that the novelty threshold set at $F_1(\kappa) = 0.99$ will be exceeded for increasing numbers m of observed data $\mathbf{x}$. After $m = 10$ observations, the threshold will have been exceeded with probability $P = 0.163$, and after $m = 100$, the threshold will have been exceeded with probability $P = 0.797$. The threshold is exceeded with probability $P = 0.99$ after $m = 248$ samples have been generated. That is, if we observe 248 data from f_1, then the novelty threshold is almost certain to be exceeded ($P = 0.99$), rather than being exceeded with probability $P = 0.01$ that we would expect if we were to use the cdf F_1.

Thus, setting a novelty theshold using F_n only has a valid probabilistic interpretation when $m = 1$; i.e., for classification tasks in which a single entity is being compared with a model of normality. An example of this is when comparing a single mammogram to a model constructed using "normal" mammogram data [26].

4 EVT for Novelty Detection

Existing work on the use of EVT for novelty detection has been limited to [4, 5, 19, 20, 23, 28]. We here consider the use of EVT for novelty detection with (i) multivariate and (ii) multimodal data, and identify problems with existing approaches in both cases.

4.1 EVT in Multivariate Novelty Detection

Multivariate extrema defined in the EVT literature [3, 8], also termed *component-wise extrema*, are those n-dimensional data $\mathbf{x}_n$ that are maxima or minima in one or more dimensions of n. For novelty detection, we require extremes w.r.t. our multivariate model of normality, rather than considering extrema in each dimension independently. EVT was first used for novelty detection in multivariate data in [19, 20], where models of normality were represented by mixtures of Gaussian distributions. In multivariate space, the Gaussian distribution describes a hyperellipsoid with $f_n(\mathbf{x})$ varying along a radius r according to the univariate Gaussian (scaled by a normalisation factor dependent on dimensionality n). That is, to determine the probability density $f_n(\mathbf{x})$ at any point in the hyperellipsoid, the problem is reduced to a univariate case $f_1(r)$, in Mahalanobis radius r. Roberts [19, 20] uses this assumption to reduce the problem of determining the EVD for a multivariate Gaussian kernel to a corresponding univariate case. As illustrated in Fig. 3, the EVD for a single Gaussian distribution along a radius r varies according to a univariate Gumbel distribution.

Existing work uses classical EVT to estimate the parameters c_m, d_m of this Gumbel cross-section f_n^e in multivariate n-space, as defined in Eq. 5. Figure 4 shows the estimates of c_m, d_m given by classical EVT compared to maximum likelihood estimates (MLEs) obtained using the unimodal, univariate method of [3] for increasing dimensionality n of the Gaussian distribution. For the univariate case f_1, it may be seen from the figure that classical EVT correctly estimates the Gumbel parameters. For $n > 2$, the location parameter c_m of f_n^e is significantly underestimated, and this error becomes greater with increasing dimensionality n.

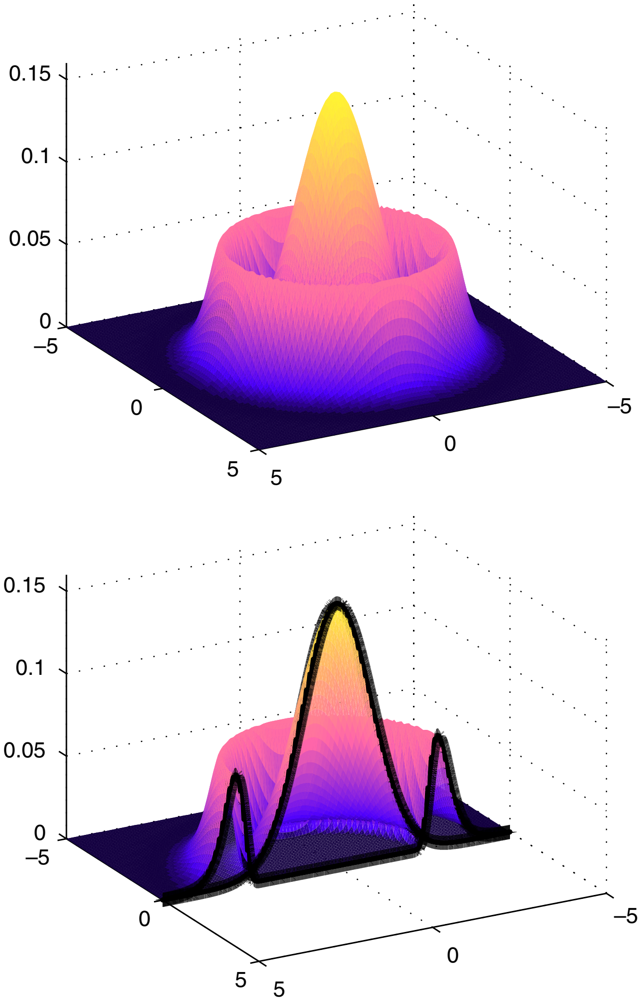

Figure 3 The bivariate Gaussian f_2 and its EVD f_2^e (shown in the *upper plot*), which is a torus with Gumbel cross-section (shown in the *lower plot*).

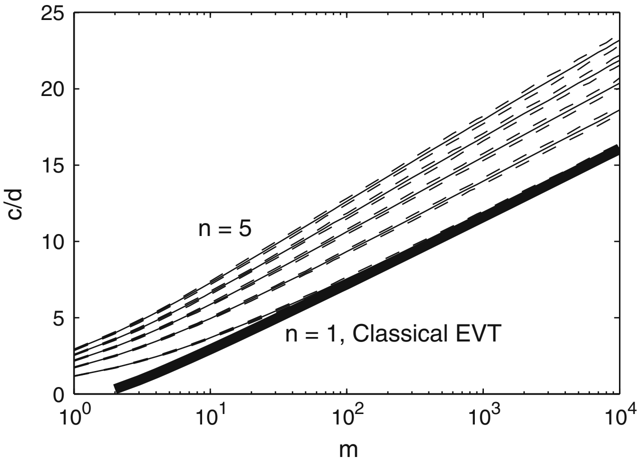

Figure 4 Ratio c/d of the MLE parameters for the multivariate Gumbel cross-section as a function of m for increasing dimensionality $n = 1 \ldots 5$ (*thin lines*), compared to c/d estimated by classical EVT (*thick line*). MLEs were obtained from $N = 10^6$ experiments for each value of m and n. Confidence intervals are shown at two standard deviations from the mean.

Figure 5 shows the error between the actual parameters c_m, d_m of f_n^e and varying estimates $\hat{c_m}, \hat{d_m}$ for $n = 2$

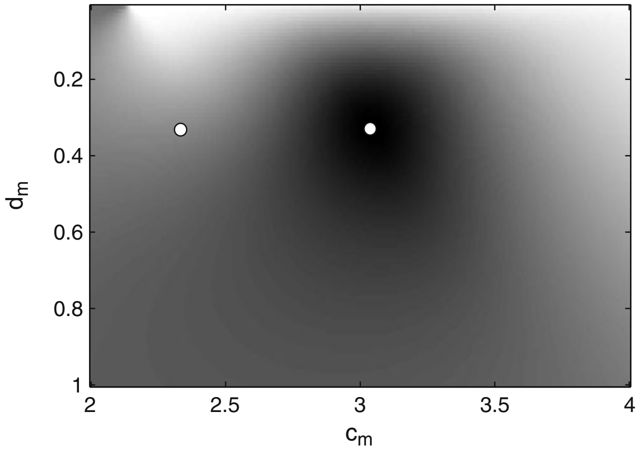

Figure 5 Error in estimates of Gumbel parameters c_m, d_m for f_2^e, for varying c_m and d_m. The MLE is shown by the marker at $c_m = 3.04, d_m = 0.32$; the classical EVT estimates are shown by the marker at $c_m = 2.37, d_m = 0.33$, which is significantly far from the desired MLE values.

and $m = 100$, where it may be seen that classical EVT fails to estimate c_m correctly.

Thus, we conclude that for *multivariate* distributions f_n, we cannot use classical EVT to estimate the EVD, f_n^e.

4.2 EVT in Multimodal Novelty Detection

EVT defines an "extreme value" to be that which is either a minimum or maximum of a set of observations $\mathbf{x}$. This is due to the conventional use of EVT [7] for determining events of extremely large or small magnitude, such as extreme financial events, extreme meteorological events, etc. In novelty detection, when considering the extrema of unimodal distributions, as is the focus of most previous work in the field of EVT [5, 23, 28], this existing definition of "extreme value" is sufficient for univariate f_1. For multivariate data, providing that f_n is *unimodal*,[8] the existing definition may be taken to mean the minimum or maximum radius r from the single mode of f_n (reducing the multivariate problem to a univariate problem in r, as we have described in Section 4.1).

However, for multimodal f_n, whether uni- or multivariate, the notion of minimum or maximum value is no longer sufficient, because there is no single mode from which distance may be defined.[9] The upper plot in Fig. 6 shows a univariate bimodal pdf. While the minimum or maximum should be treated as extrema (e.g., $x = 1$ or $x = 28$), values between the two modes (such as $x = 10$) should similarly be taken to be extreme in terms of probability density $f_n(x)$, because they are just as improbable as the minimum or maximum on the x-axis.

This is a key point of departure in our work with EVT for novelty detection, in which we are interested in determining extremely *unlikely* events, whereas conventional EVT is interested in determining events of extremely large or small *magnitude*. As illustrated in Fig. 6, while events of extremely large or small magnitude may be events that are extremely unlikely, not all events that are extremely unlikely are events of extremely large or small magnitude.

Similarly, the lower plot of Fig. 6 shows a multivariate, multimodal distribution. For bivariate data $\mathbf{x} = (x_1, x_2)$, data-points which represent a minimum or maximum in either dimension x_1 or x_2 should be considered "extreme",[10] but it is also desireable that similarly improbable areas of data space, such as the origin $\mathbf{x} = (0, 0)$ in this example, should be considered "extreme" for the purposes of novelty detection.

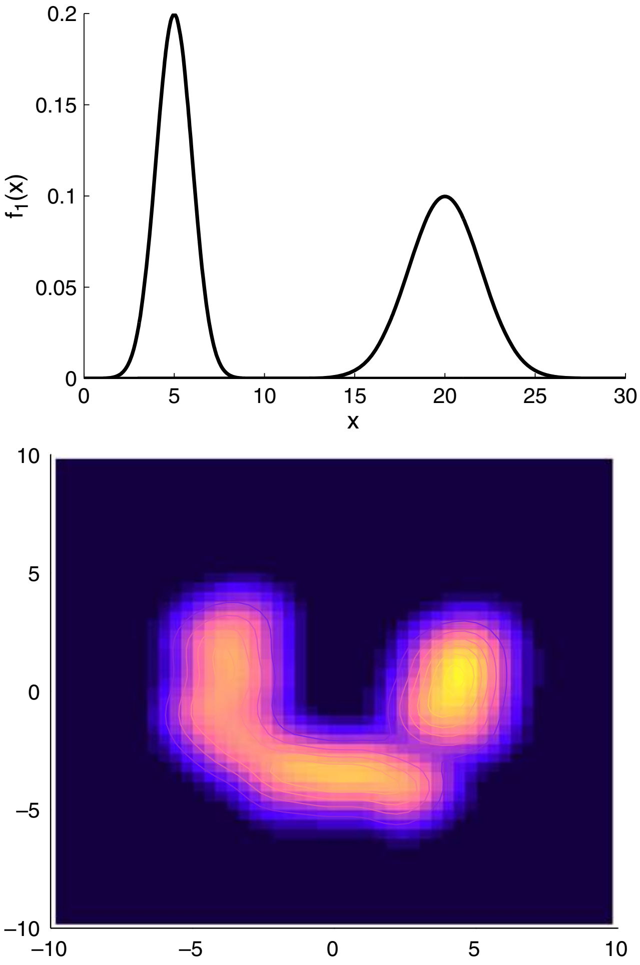

Figure 6 Multimodal distributions require redefinition of the term "extreme value". The *upper plots* shows a bimodal distribution f_1, where extrema could fall between the two modes, and so we must consider more than "minimum" and "maximum" values on the x-axis. The *lower plot* shows a multimodal Parzen windows model formed from $k = 60$ components, where extrema could fall in the "horseshoe" between clusters of modes.

Given that the goal of using EVT for novelty detection is to identify *improbable* events w.r.t. f_n, rather than events of extreme absolute magnitude, we redefine "extreme value" in terms of probability:

Definition 1 For novelty detection, the "most extreme" of a set of m samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$

[8] and radially symmetric

[9] and where there is no global symmetry

[10] This is a component-wise extremum, as described earlier.

distributed according to pdf $f_n(\mathbf{x})$ is that which is most improbable with respect to the distribution; i.e., $\mathrm{argmin}_{\mathbf{x}\in\mathbf{X}}\left[f_n(\mathbf{x})\right]$.

The conventional use of "extrema" to mean minimum or maximum values with respect to a *unimodal* distribution becomes a special case of the above definition: because the pdf f_n typically decreases monotonically with increasing distance from the single mode, selecting the samples furthest from that mode (i.e., the minimum or maximum of a set of samples) is equivalent to selecting the samples for which f_n is minimised.

This selection of extrema based on minimising f_n is equivalent to selecting extrema by maximising F_n using Eq. 1. Our definition satisfies the condition of [14], which states that the probability of observing data inside a novelty boundary should be at least as large as the probability of observing data outside the novelty boundary. That is, the novelty boundary is a lower bound on f_n for "normal" data.

Definition 1 provides us with the mechanism we require to determine the extent of data space that is considered "normal": if we observe m "normal" data generated from a model of normality f_n, the EVD f_n^e now describes where the *least probable* of those m normal data will lie. Thus, we can use the EVD to set a novelty threshold, as will be described, and perform novelty detection in a principled manner.

As with multivariate novelty detection, little work exists in using EVT for novelty detection with multimodal f_n [19, 20]. In this work, the multimodal distribution represented by a mixture of Gaussian component distributions was reduced to a single-component problem: to find the value of the EVD f_n^e at some location $\mathbf{x}$, the closest component distribution (determined using Mahalanobis distance) was assumed to dominate f_n^e, and thus the EVD is based on the Gumbel distribution corresponding to that closest component (using radius r from that component's centre, as described in Section 4.1). Here, the contribution of other components to f_n^e is assumed to be negligible, and they are ignored.

4.2.1 Problems Arising due to Differing Component Variances

Figure 7 shows the EVD f_1^e, determined using the existing method, corresponding to the univariate, bimodal distribution f_1 from Fig. 6, which is a mixture of two Gaussian component distributions. Here, the prior probabilities of each component are equal, $P(\mu_1) = P(\mu_2) = 0.5$, and the kernel variances are $\sigma_1^2 = 1$, $\sigma_2^2 = 4$. Figure 7 also shows the corresponding cdf F_1^e, from which it may be seen that each component is responsible for 0.5 of the total probability mass. The EVD modes for component μ_2 have a maximum value of $f_1^e(x) = 0.15$, half of the maximum value of the modes for component μ_1, because the modes for μ_2 are spread out twice as far ($\sigma_2 = 2\sigma_1$).

In Fig. 7, the circles correspond to a histogram of $N = 10^6$ extrema generated from *each component distribution independently*. That is, component μ_1 has been responsible for $0.5N$ extrema and component μ_2 has been responsible for $0.5N$ extrema. This is *not* the

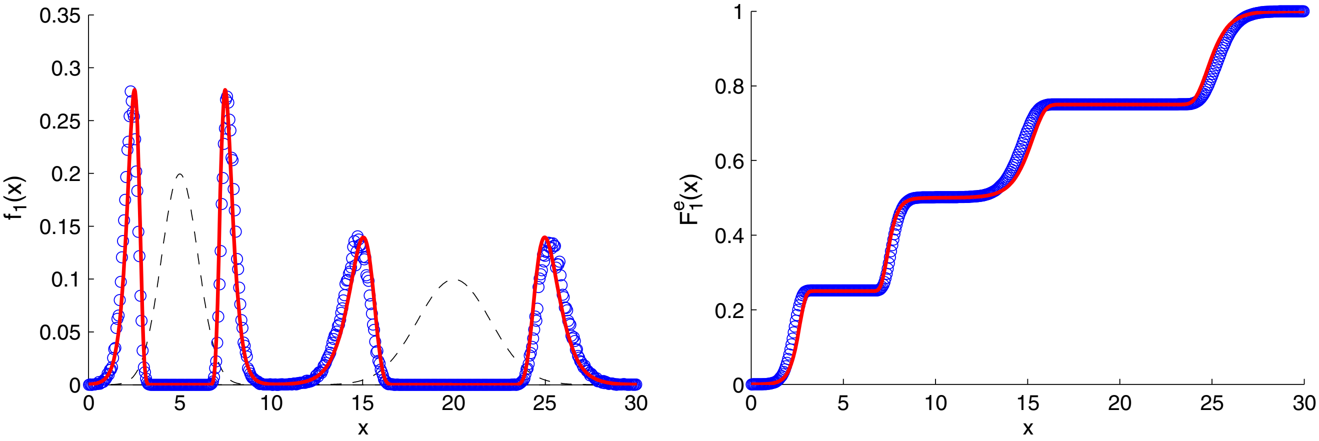

Figure 7 Comparison of EVD determined using the existing method (*solid line*) and histograms of 10^6 extrema (*circles*) drawn from bimodal distribution $f_1(x)$ (*dashed line*). The left- and right-hand plots show the pdf f_1^e and cdf F_1^e, respectively. Extrema are drawn from each component of $f_1(x)$ independently, according to their prior, $P(\mu_1) = P(\mu_2) = 0.5$, and so each component generates half of the samples. This closely matches the EVD determined by the existing method, but does not represent the process of drawing m samples from $f_1(x)$.

same as drawing N extrema directly from the mixture f_1, which is what we wish to perform.

Figure 8 shows the histogram obtained when $N = 10^6$ extrema are generated *correctly* from the mixture; i.e., where m samples are drawn from the mixture $f_1(x) = \frac{1}{2}\sum_i P(\mu_i)p(x|\mu_i)$, and the most extreme is retained, for each of $N = 10^6$ iterations. The cdf F_1^e in Fig. 8 shows that component μ_2 is responsible for more than 0.5 of the probability mass (in fact, two-thirds of it).

This example shows that, when samples are generated directly from the model of normality, component μ_2, having wider variance and thus taking lower $f_1(x)$ values than component μ_1, is responsible for more extrema than we would expect from the kernels' equal prior probabilities. Though in each of $N = 10^6$ iterations, we generate, on average, $0.5m$ samples from each component (in accordance with their equal prior probabilities), we have a greater chance of retaining those samples from μ_2 because the absolute values of their probability densities $f_1(x)$ are generally lower, due to the increased variance of that component. This unintuitive phenomenon was observed by [14] when considering conventional integration of mixtures of Gaussian distributions.

Thus, the assumption that only the closest component distribution to $\mathbf{x}$ need be considered when determining the EVD $f_n^e(\mathbf{x})$ cannot generally be made. Though the effect of other components on $f_n(\mathbf{x})$ may be negligible, because of their distance from $\mathbf{x}$, their effect on $f_n^e(\mathbf{x})$ may be significant due to the relative differences in variances between kernels, as shown in the example in Fig. 8. This typically occurs with Gaussian mixture models, in which components usually have differing variances.

4.2.2 Problems Arising due to Overlapping Components

The existing method results in a piecewise-hyperspherical EVD, because f_n^e for all $\mathbf{x}$ is determined using only a single component (the closest to $\mathbf{x}$). This is illustrated in Fig. 9, in which a mixture of two component distributions with equal variance and significant overlap is shown. The figure shows the actual EVD (shown by the outer solid line) evaluated at $F_2^e(\kappa) \leq 0.999$, evaluated using Monte Carlo methods. The resulting equiprobable contour follows the contours of the underlying distribution $f_2(\mathbf{x})$. The figure also shows the equivalent contour of the EVD obtained using the existing method (shown by the inner solid line), in which the resultant contour is piecewise-circular. It may be seen that areas of data space that should be considered "normal" (i.e., lying within the outer line, if the contour is used as a novelty threshold) would be incorrectly considered "abnormal" by the existing method.

The areas of data space thus misclassified as "abnormal" by the existing method will increase with increasing overlap between components. For models of normality comprising large numbers of components, such as those constructed using Parzen windows estimation, the misclassified areas of data space could be large due to the typically considerable overlap between

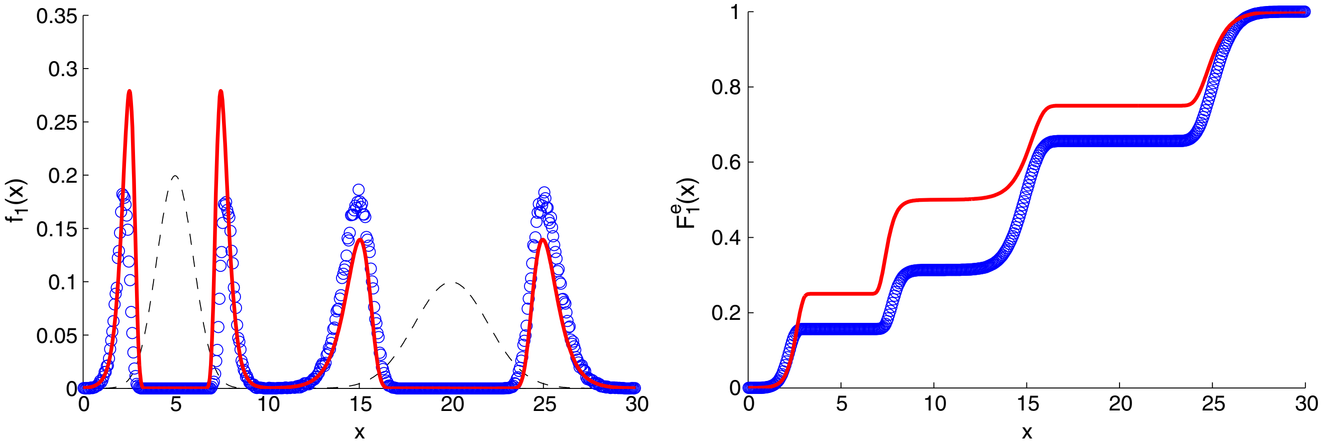

Figure 8 Comparison of EVD determined using the existing method (*solid line*), as before, and histograms of 10^6 extrema (*circles*) drawn from bimodal distribution $f_1(x)$ (*dashed line*). The left- and right-hand plots show the pdf f_1^e and cdf F_1^e, respectively. Here, the extrema are correctly drawn from the mixture $f_1(x)$ (i.e., m samples are drawn from $f_1(x)$ and the most extreme of those m samples is retained). The EVD determined by the existing method does not fit the actual observed extrema.

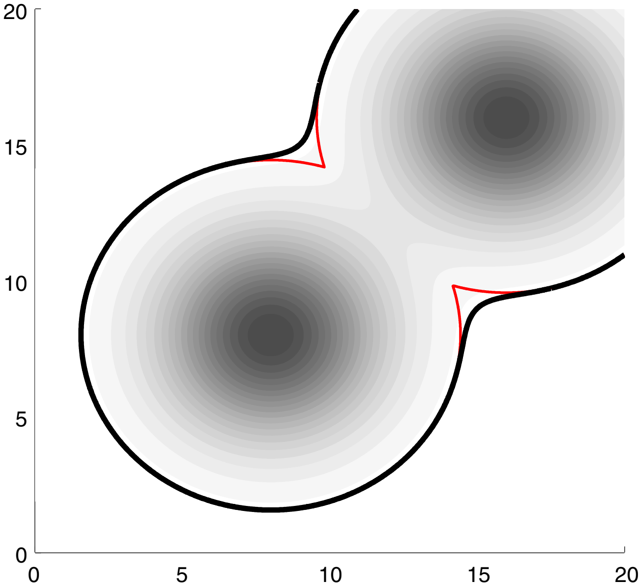

Figure 9 A mixture of two component distributions with equal variance and significant overlap. The true EVD integrated to $F_2^e(\mathbf{x}) \leq 0.999$ is shown by the outer solid line. The existing method results in an equiprobable contour $F_2^e(\mathbf{x}) \leq 0.999$ that is piecewise circular, shown by the *inner solid line*. Part of the overlapped regions of the components (*between the inner and outer solid lines*) would incorrectly be classified "abnormal" by the existing method, using this contour as a novelty threshold.

components and the typically large numbers of components used in such models.

We note that with increasing data dimensionality n, the overlapped hypervolume of data space between neighbouring components (as a proportion of the total probability mass) decreases. Thus, for high-dimensional models with smaller regions of overlapped data space, we expect that these problems will be less significant.

5 Understanding the EVD

We have shown that existing approaches to EVT are unsuitable for novelty detection where the generative data distribution is multivariate or multimodal. This section presents a new method of understanding the EVD, which we require in order to estimate the EVD for multivariate, multimodal f_n.

5.1 The EVD as a Transformation of f_n

The EVD f_n^e for a distribution f_n follows the probability contours of that distribution. This is a consequence of using Definition 1, where extrema are defined in terms of minimising $f_n(\mathbf{x})$ for a set $\mathbf{X}$ of m samples (or, equivalently, maximising F_n).

It is convenient to consider the EVD as a transformation of equiprobable contours on f_n. Figure 10 shows equiprobable contours for a bivariate model of normality f_2 represented by a mixture of three Gaussian components with full (non-diagonal) covariance matrices. The EVD f_2^e is shown for $m = 100$. Equiprobable contours of the EVD f_2^e occur at equiprobable contours of f_2, and thus we may consider the EVD to be a weighting function of the contours of f_n,

$$f_n^e(\mathbf{x}) = g\left[f_n(\mathbf{x})\right] \tag{7}$$

for some weighting function g. With the EVD thus defined in terms of f_n, we have the facility to accurately determine f_n^e for complex, multimodal, multivariate distributions, if we can find the form of g.

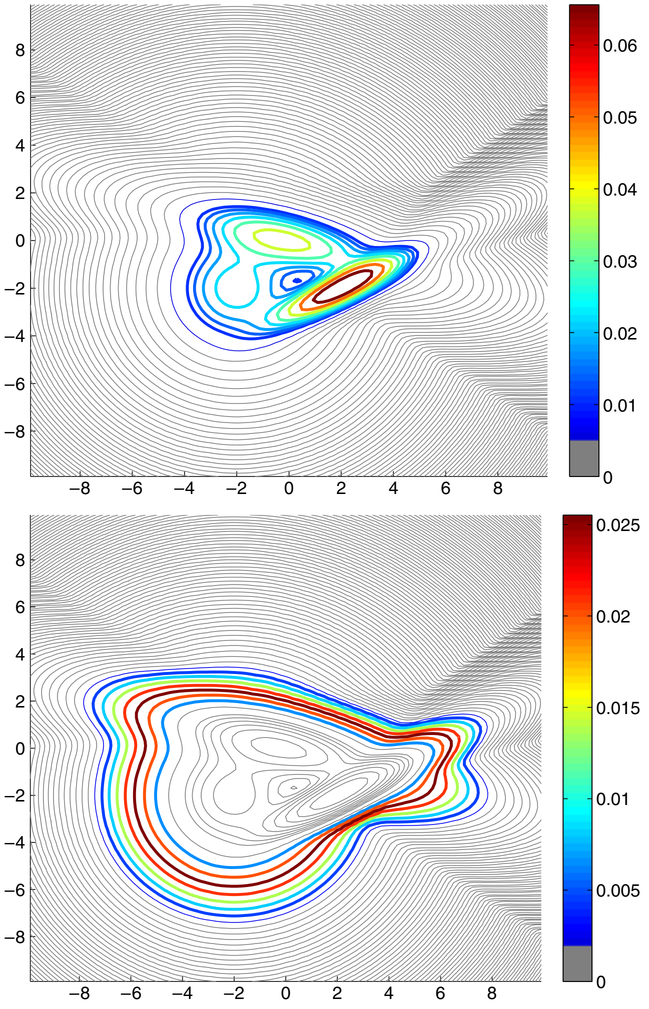

Figure 10 Probability contours on f_n for a tri-modal GMM (*upper plot*) with full (*non-diagonal*) covariance matrices and corresponding EVD f_n^e (*lower plot*).

5.2 The Ψ-Transform

For a standard Gaussian distribution,

$$f_n(x) = (2\pi)^{-n/2} \exp(-r^2/2) \tag{8}$$

and so, rearrangement gives

$$r = (-2 \ln f_n(x) - n \ln 2\pi)^{1/2} . \tag{9}$$

We define a transform of the extrema $\mathbf{x}$,

$$\Psi \left[f_n(\mathbf{x}) \right] = \begin{cases} (-2 \ln f_n(\mathbf{x}) - n \ln 2\pi)^{1/2} & \text{if } f_n(\mathbf{x}) < K \\ 0 & \text{if } f_n(\mathbf{x}) \geq K \end{cases} \tag{10}$$

where $K = (2\pi)^{-n/2}$. If f_n is a (unimodal) Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the Ψ-transform would map the $f_n(\mathbf{x})$ values back onto r, the radii of $\mathbf{x}$ from $\boldsymbol{\mu}$, which we know are distributed according to the Gumbel distribution (as shown in Section 4.1). The Ψ-transform maps the distribution of $f_n(\mathbf{x})$ values back into a space into which a Gumbel distribution can be fitted, having observed that $f_n(\mathbf{x})$ for extrema are distributed similarly for mixtures of negative exponentials of varying number of kernels, priors, and covariances [4].

The upper plot in Fig. 11 shows a normalised histogram of $N = 10^6$ extrema generated from the example mixture of three Gaussian components f_2 in Fig. 10, for $m = 100$. The distribution is highly skewed towards $f_2(\mathbf{x}) = 0$, as is expected for extrema. The lower plot in the figure shows the Ψ-transform of the histogram of extrema, which may be seen to be distributed according to the Gumbel (after normalisation such that its area is unity). The MLE Gumbel distribution fitted in Ψ-space using the univariate, unimodal method of [3] is shown, which is $f_2^e\big(\Psi[f_2(\mathbf{x})]\big)$. Thus, all locations $\mathbf{x}$ in the original data space $\mathcal{D} = \mathbb{R}^n$ may be evaluated w.r.t. $f_n^e\big(\Psi[f_n(\mathbf{x})]\big)$, as was shown in Fig. 10, and we have successfully found the multivariate, multimodal EVD which is a transformation of f_n, as required.

We may determine the location of a novelty threshold on f_n^e by equating the corresponding cdf F_n^e (which is univariate in Ψ-space) to some probability mass; e.g., $F_n^e[f_n(\mathbf{x})] = 0.99$, as shown in Fig. 11. Thus, we have defined a contour in data space $\mathcal{D}$ that describes where the most extreme of m "normal" samples generated from f_n will lie, to some probability (e.g., 0.99).

This is the key insight that allows us to determine the EVD for data of arbitrarily large dimensionality n: we have reduced the problem of analysis in multivariate data space to a simpler, but equivalent, problem in univariate probability space.

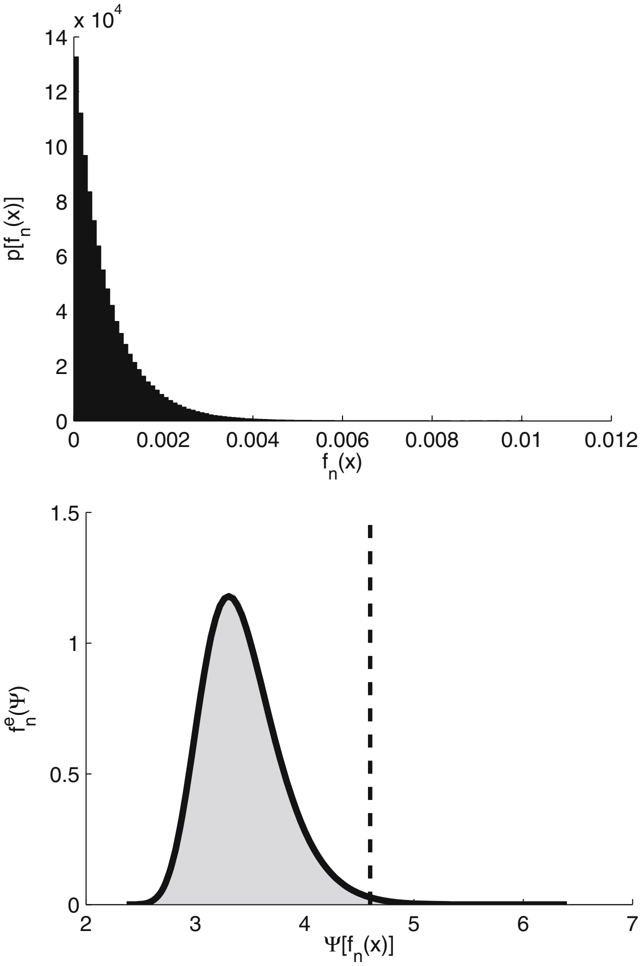

Figure 11 Normalised histogram of $f_2(\mathbf{x})$ values for $N = 10^6$ extrema generated from trimodal GMM with $m = 100$ (*upper plot*). Histogram of the Ψ-transformed $f_2(\mathbf{x})$ values shown in *grey*, with the corresponding MLE Gumbel distribution fitted in Ψ-space, shown in *black* (*lower plot*). A novelty threshold at $F_2^e = 0.99$ is shown as a *dashed line*.

Note finally that, though the novelty threshold set using our proposed method occurs at some contour $f_n(\mathbf{x}) = \kappa$ (due to Definition 1), it is not heuristic: the threshold is set such that generating m samples from f_n will exceed the threshold with probability $1 - F_n^e(\kappa) = 1 - 0.99 = 0.01$; that is, the final novelty threshold has a valid probabilistic interpretation provided by EVT.

5.3 Investigating the EVD

The method described in the previous sub-section scales with dimensionality n and the number of components in the mixture f_n. Figure 12 shows the EVD for

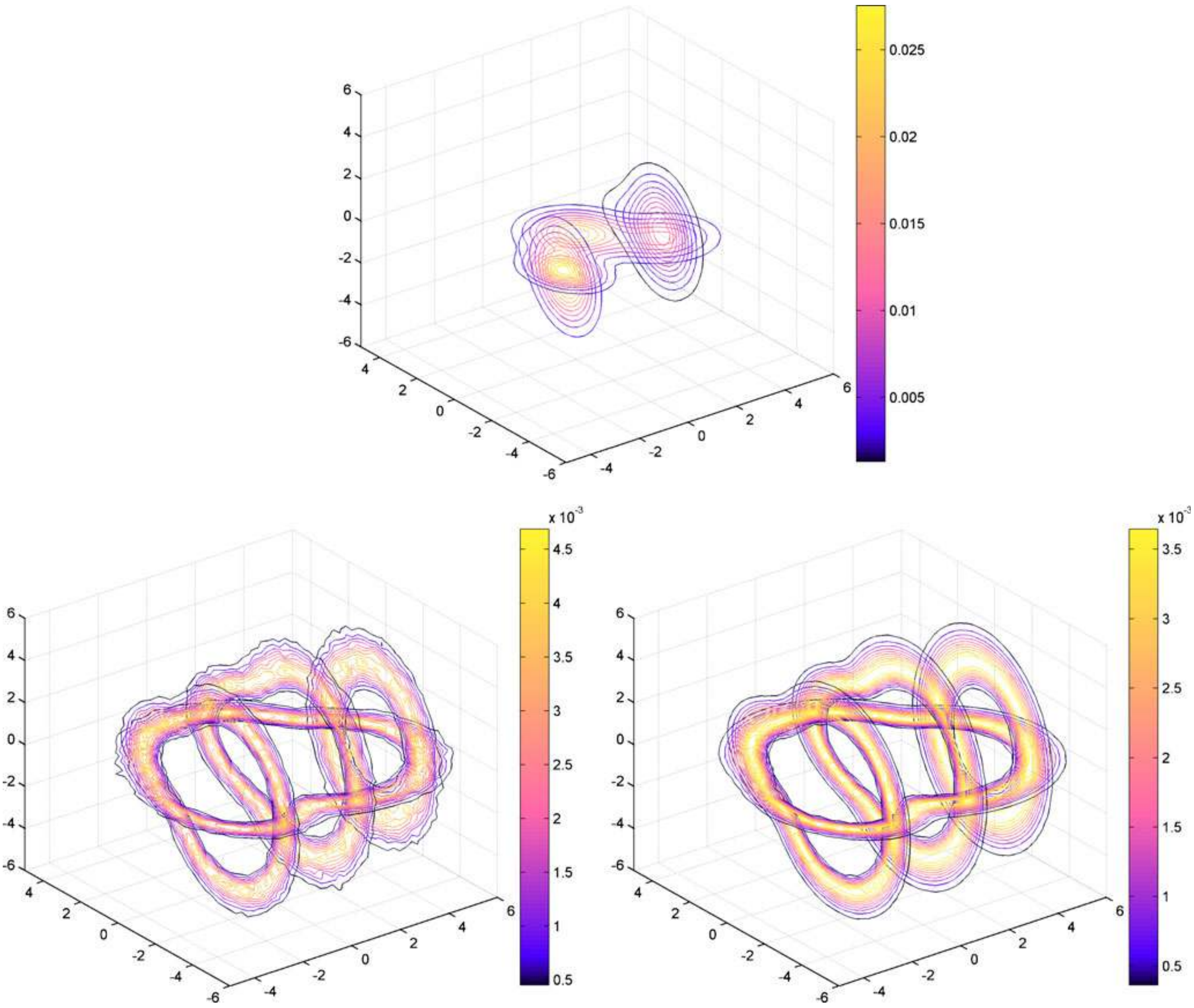

Figure 12 EVD f_3^e shown in data space $\mathcal{D}$ estimated using the proposed Ψ-transform method for a trivariate, trimodal f_3 (*upper plot*). The EVD determined using the proposed method (*lower-right plot*) closely matches the histogram of experimentally-obtained extrema (*lower-left plot*).

a trivariate model f_3 (projected onto four planes for visualisation), which closely matches the EVD observed from experimentally-obtained extrema. Figure 13 also shows the Ψ-transform for a 6-dimensional mixture f_6 of 15 Gaussian components with full (non-diagonal) covariance, where it may also be seen that the EVD closely matches that of experimentally-obtained extrema.

We note that this Ψ-transform method is numerical, requiring the sampling of extrema from f_n, and then fitting the MLE Gumbel distribution after application of the Ψ-transformation. Future work aims to find closed-form solutions (or approximations) for multivariate, multimodal distributions f_n. Currently, closed forms have been obtained for multivariate, *unimodal* distributions, which are presented in the next section.

6 Closed-Form EVT for Multivariate, Unimodal Novelty Detection

In the previous section, we argued the need for a multivariate Extreme Value Theory (mEVT) in machine learning, identified some of the limitations of the existing approach, and proposed a numerical scheme for multimodal, multivariate estimation.

In this section, we offer an analytical approach to mEVT, restricted to single Gaussian distributions; i.e.,

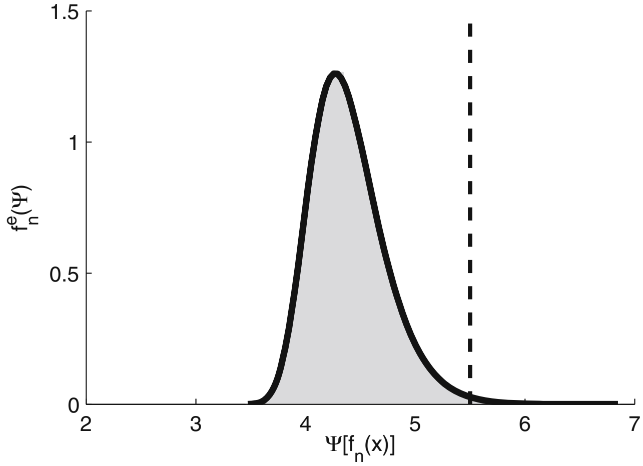

Figure 13 Histogram of Ψ-transformed extrema (shown in *grey*) and MLE Gumbel (shown in *black*) for a 6-dimensional f_6 mixture of 15 components. A novelty threshold has been set at $F^e_6 = 0.99$, shown as a *dashed line*.

for dimensionality $n \in \mathbb{N}^*$, distributions F_n with probability density functions of the form:

$$f_n(\mathbf{x}) = \frac{1}{C_n} \exp\left(-\frac{M(\mathbf{x})^2}{2}\right) \tag{11}$$

where

$$M(\mathbf{x}) = \left((\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{1/2} \tag{12}$$

is the Mahalanobis distance,

$$C_n = (2\pi)^{n/2}|\Sigma|^{1/2} \tag{13}$$

is the normalisation coefficient, $\boldsymbol{\mu}$ the centre, and Σ the covariance matrix. We term $\mathcal{D} = \mathbb{R}^n$ the data space, and $\mathcal{P} = f_n(\mathcal{D}) = \left]0, \frac{1}{C_n}\right]$, the associated probability space. A unimodal approach is of interest because the existing method [19] yields significant errors when estimating EVD parameters, as shown previously, which can be solved directly by the use of analytically-derived estimates. Furthermore, the numerical approach presented in the previous section requires that we generate extrema from a multivariate distribution, which can be time-consuming as the sample size is increased. While a fully analytical mEVT may not be possible, the analytical study presented here paves the way to more elaborate numerical schemes with no need for sampling extrema.

6.1 Probability Distribution of Probability Density Values

6.1.1 Sampling in Data Space is Equivalent to Sampling in Probability Space

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ be samples (vectors of $\mathcal{D}$) drawn from the distribution F_n for $k \in \mathbb{N}$ and $f_n(\mathbf{x}_1)$, $f_n(\mathbf{x}_2), \ldots,$ $f_n(\mathbf{x}_k)$ be the corresponding pdf values of these samples. The probability of obtaining a given value in the probability space by drawing a sample in the data space is strongly related to the form of f_n. Assuming that X is a random variable distributed according to F_n, our aim in this section is to determine the form of the distribution function (df) G_n according to which $f_n(X)$ is distributed on $\mathcal{P}$. That is, we wish to determine the probability distribution over probability density values, noting from the previous section that novelty detection in the probability space defined by the model of normality is equivalent to novelty detection in the data space.

6.1.2 Distribution Function Over $f_n(X)$

We define a distribution over $Y = f_n(X)$ as follows:

$$\forall y \in \mathcal{P}, \quad G_n(y) = \int_{f_n^{-1}(]0,y])} f_n(\mathbf{x})d\mathbf{x} \tag{14}$$

where $f_n^{-1}(]0, y])$ is the preimage of $]0, y]$ under f_n. G_n is the complementary function of that in Definition 1 where κ is now y.

To take advantage of the ellipsoidal symmetry of the problem we rewrite f_n in a Mahalanobis n-dimensional spherical polar coordinate system. Then, $x_v = (r, \boldsymbol{\theta})$ such that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{n-1})$, $r = M(\mathbf{x})$, $\theta_i \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right[$ for $i \leq n-2$ and the base angle θ_{n-1} ranges over $[0, 2\pi]$.

The Jacobian of the transformation [22] is

$$|J| = |\Sigma|^{1/2} r^{n-1} \prod_{i=0}^{n-3} (\cos\theta_i)^{n-i} \tag{15}$$

Thus, we may now expand Eq. 14, to find the distribution over the pdf f_n,

$$G_n(y) = \int_{f_n^{-1}(]0,y])} \frac{1}{C_n} \exp\left(-\frac{M(\mathbf{x})^2}{2}\right) d\mathbf{x}, \tag{16}$$

$$= \int_{f_n^{-1}(]0,y])} \frac{|J|}{C_n} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r\mathrm{d}\boldsymbol{\theta}\,, \tag{17}$$

$$= \Omega \int_{M^{-1}(y)}^{+\infty} \frac{r^{n-1}}{(2\pi)^{n/2}} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r, \tag{18}$$

$$= \Omega |\Sigma|^{1/2} \int_0^y \left[-2\ln(C_n u)\right]^{(n-2)/2} \mathrm{d}u. \tag{19}$$

In the above, $M^{-1}(y) = \sqrt{-2\ln(C_n y)}$ is the unique Mahalanobis distance associated with the pdf value y.

Equation 17 is obtained by rewriting Eq. 16 in the spherical polar coordinate system. Integrating out the angles yields Eq. 18, where $\Omega = \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})}$ is the total solid angle subtended by the unit n-sphere. Equation 19 is obtained after making the substitution $u = \frac{1}{C_n}\exp\left(-\frac{r^2}{2}\right)$. Note that Eq. 18 and Eq. 19 hold for $n = 1$, such that they may be applied to univariate distributions as well as multivariate distributions.

The integrand in Eq. 19 is the pdf of G_n, which we aim to find,

$$g_n(y) = \Omega_n |\Sigma|^{1/2} \left[-2\ln(C_n y)\right]^{n-2/2} \tag{20}$$

In the univariate case $n = 1$, the integration of Eq. 18 yields

$$G_1(y) = \operatorname{erfc}\left(\sqrt{-\ln(C_1 y)}\right) \tag{21}$$

where erfc (.) is the complementary error function.

For the multivariate case $n \geq 2$, the integration of Eq. 19 is possible using a recursive integration by parts, which yields two cases:

$$G_{2p}(y) = y \sum_{k=0}^{p-1} A_{2p}^k \left(-2\ln\left(C_{2p} y\right)\right)^{(p-k-1)} \tag{22}$$

$$G_{2p+1}(y) = y \sum_{k=0}^{p-1} A_{2p+1}^k \left(-2\ln\left(C_{2p+1} y\right)\right)^{p-k-1/2} + \operatorname{erfc}\left(\sqrt{-\ln\left(C_{2p+1} y\right)}\right) \tag{23}$$

for all $p \in \mathbb{N}^*$, where

$$A_{2p}^k = \Omega_{2p} |\Sigma|^{1/2} \frac{2^k (p-1)!}{(p-1-k)!} \tag{24}$$

and

$$A_{2p+1}^k = \Omega_{2p+1} |\Sigma|^{1/2} \frac{(2p-1)!(p-k)!}{2^{k-1}(p-1)!(2p-2k)!}. \tag{25}$$

G_n and g_n are plotted for $n = 1$ to 5 in Fig. 14, together with simulated data. Perhaps counter-intuitively, we observe that, relative to the right endpoint of G_n, the probability mass shifts towards 0 as the dimensionality n increases, which indicates that the probability mass in the data space moves away from the centre of the distribution, as is noted by Bishop in [1] (example 1.4, page 29).

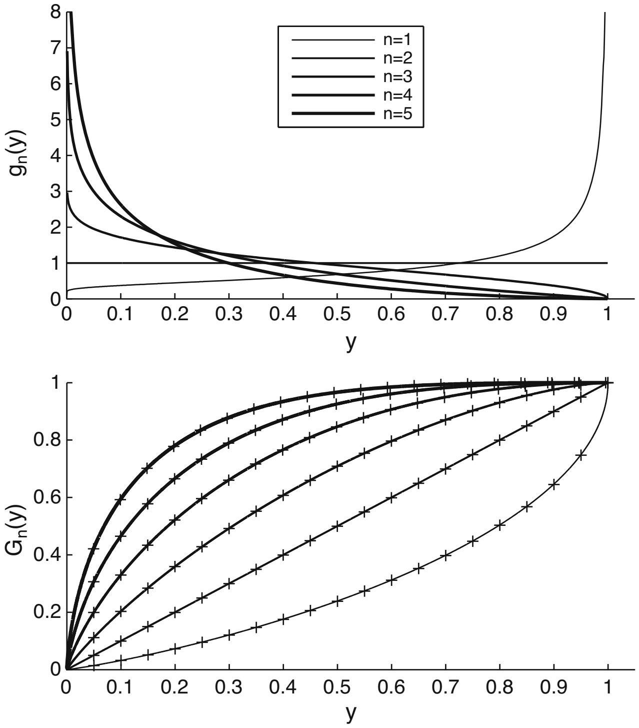

Figure 14 Analytical and simulated G_n and g_n for various values of n. The x-axis is scaled so that all distributions have the same right endpoint. *Crosses* are the result of drawing 10^6 samples in the data space and computing the cumulative histograms of their probabilities. F_n is the multivariate standard normal distribution.

This is the key reason that classical EVT cannot accurately estimate the EVD parameters for multivariate data, as was discussed in Section 4.1; as dimensionality of the data increases, the probability mass tends towards regions of low probability density, whereas classical EVT can only estimate the case for $n = 1$, in which the probability mass is clustered around the mode of f_1, as shown in Fig. 14.

6.2 Finding the EVD for Probability Density Values

Our approach is based on the idea that pdf values of the EVD in the data space must be equal on a level set of F; i.e., the EVD is obtained by applying a weighting function to the level sets of F, as was shown in Section 5. Consequently, determining the EVD in the data space can be performed by determining the EVD of G in the probability space. We therefore reduce an n-dimensional problem (finding an EVD in $\mathcal{D}$) to a simpler one-dimensional case (finding an EVD in $\mathcal{P}$).

In this section, our aim is to determine the extreme value distribution of minima for G_n and estimate its parameters, using classical EVT.

6.2.1 Maximum Domain of Attraction of the Weibull Distribution

The Fisher–Tippett theorem effectively defines a three-class (Gumbel, Fréchet, and Weibull) equivalence relation on the set of non-degenerate univariate distributions. Embrechts et al. [8] gives the characterizations for each class.

Theorem 3.3.12 in [8] characterizes the maximum domain of attraction (MDA) of the maximal Weibull distribution. We adapt it to the MDA of the minimal Weibull distribution, H_3^-:

Theorem 1 (Maximum domain of attraction of H_3^-) *The df F belongs to the maximum domain of attraction of the minimal Weibull distribution* ($\alpha > 0$)*, if and only if $x_F > -\infty$ and $F(x_F + x^{-1}) = x^{-\alpha} L(x)$ for some slowly varying function L. If $F \in MDA(H_3^-)$, then $c_m^{-1}(E_m - x_F) \overset{d}{\to} H_3^-$, where the norming constants c_m, d_m can be chosen to be $c_m = x_F + F^{\leftarrow}(m^{-1})$ and $d_m = x_F$, and where E_m is the extremum of m data.*

In the above, x_F is the left endpoint of the df F, $F^{\leftarrow}(p)$ is the p-quantile of F, and L is a slowly varying function at ∞; i.e., a positive function that obeys

$$\forall t > 0, \lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1. \tag{26}$$

From Eq. 26, it may be seen that $y \mapsto -\ln(1/y)$, $y > 1$ is slowly varying, as is $y \mapsto -\ln(1/y)^{\beta}$, $y > 1$, for all $\beta \in \mathbb{R}$. Therefore $y\, G_{2p}(1/y)$ is a sum of slowly varying functions, which is itself slowly varying. Theorem 1 can therefore be applied to G_{2p}. A similar process can be followed to show that G_{2p+1} is in the MDA of H_3^-. Consequently, G_n is in the MDA of H_3^- for all values of n.

6.3 Parameter Estimation

If G_n is in the MDA of H_3^-, Theorem 1 gives the minimal Weibull parameters:

$$d_m = 0, \quad \alpha_m = 1, \tag{27}$$

$$c_m = G_n^{\leftarrow}\left(\frac{1}{m}\right). \tag{28}$$

The scale parameter can be easily estimated numerically from Eq. 28 to arbitrary accuracy, as G_n is a strictly increasing function over finite support.

Figure 15 shows that Eq. 28 is a very close approximation to the values of the scale parameter d_m obtained via maximum likelihood estimation. However, the value of the shape parameter α_m, although theoretically guaranteed to converge to 1 in the limit $m \to \infty$ seems to decrease significantly as the dimensionality of the data space increases, and it is overestimated even for large values of m.

To address this issue, we note that the class of equivalence of H_3^- contains all the distributions with a power law behaviour at the finite left endpoint [8]. Therefore the tail of G_n is, in the limit $y \to 0$, equivalent to a power law; i.e., $G_n(y) \sim Ky^s$. Here, s can be estimated locally by noting that, in this case, $g_n(y) \sim sKy^{s-1}$; i.e., $s = y\frac{G_n(y)}{g_n(y)}$.

We therefore propose the following formula for the shape parameter,

$$\alpha_m = c_m \frac{g_n(c_m)}{G_n(c_m)} \tag{29}$$

Figure 15 shows that Eq. 29, although still inaccurate for very small values of m, gives values closer to the MLE estimates as m and n increase.

Finally, the EVD of G_n is:

$$G_n^e(y) = 1 - \exp\left(-\left(y/c_m\right)^{\alpha_m}\right), \tag{30}$$

where c_m and α_m are given by Eqs. 28 and 29, respectively.

6.4 Novelty Scores

In novelty detection, extrema are regarded as potentially abnormal data. Assuming a distribution for the normal data, if we observe m samples for which the extremum has pdf value y_m, the probability of drawing an extremum of lower probability is given by $G_n^e(y_m)$. Therefore, the probability of drawing an extremum of higher probability is $1 - G_n^e(y_m)$ and our extremum is abnormal with probability $1 - G_n^e(y_m)$.

We define a novelty score in the data space as being the probability of obtaining an extemum closer to the centre of the distribution (in the Mahalanobis sense):

$$F_n^e(\mathbf{x}) = 1 - G_n^e(f_n(\mathbf{x})), \tag{31}$$

$$= \exp\left(-\left(\frac{1}{C_n c_m} e^{-\frac{M(\mathbf{x})^2}{2}}\right)^{\alpha_m}\right). \tag{32}$$

In the limit $m \to \infty$, $F_n^e(\mathbf{x})$ can be interpreted as a Mahalanobis-radial cdf of extrema.

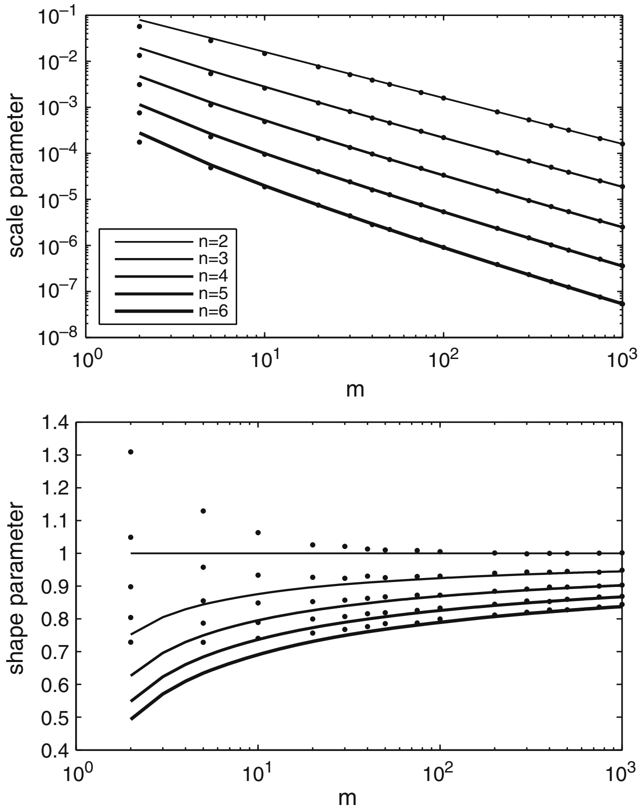

Figure 15 Comparison between results of maximum likelihood estimation of the scale parameter c_m (*top*) and the shape parameter α_m (*bottom*) parameter, and values obtained using formulae 29 and 28 for increasing m. F_n is again the multivariate standard normal distribution. The *dots* are obtained by taking the means of 10 MLEs, each with 10^4 simulated extrema. Error bars are too small to be visible at this scale.

7 Application to a Vital-Sign Monitoring Problem

7.1 Introduction to Patient Monitoring

In this section, we present an application of our proposed mEVT method to a vital-sign monitoring problem. Continuous real-time patient monitoring in hospital is usually based on single vital-sign channel alarms, which yield an unusably large number of false alarms. Recently, efforts have been made to take advantage of the correlation between vital-sign channels. In [10, 25], the authors adopt a novelty detection approach, whereby the model of normality is a 4-dimensional pdf constructed using data from a high-risk adult population. The authors then test for abnormal data by comparing the probability densities $f(\mathbf{x})$ of new measurements obtained in real-time to a predefined threshold.

Here, we introduce the use of mEVT to address the same problem, limited to the use of two vital-sign channels, heart rate (HR) and breathing rate (BR). The data set was collected during the first phase of a trial conducted at the University of Pittsburgh Medical Center [10, 12, 13], which is composed of the recordings of 332 high-risk adult patients, totalling over 32,000 hours of data. Vital-sign measurements are available every second for all patients. Recordings are labeled with "crisis events"; i.e., events that should have caused an emergency call to clinical staff to be made on the patient's behalf, as they are indicative of potentially adverse events. The cause of each event (high/low HR, high/low BR, etc.) is given with its start-time and end-time. 46 of 113 events are caused by an abnormally high or low heart rate or breathing rate (approximately 19 hours of data).

7.2 Method

We split patients into three groups: a *test* group, composed of the 28 patients who suffered at least one cardio-respiratory crisis over the course of their stay (approximately 1000 hours of data), a *training* group, and a *control* group, these latter two each composed of 154 randomly-assigned patients who did not suffer a cardio-respiratory crisis (approximately 15,500 hours of data each).

Figure 16 shows histograms of the training data. A bivariate Gaussian distribution F_2 is fitted to the data

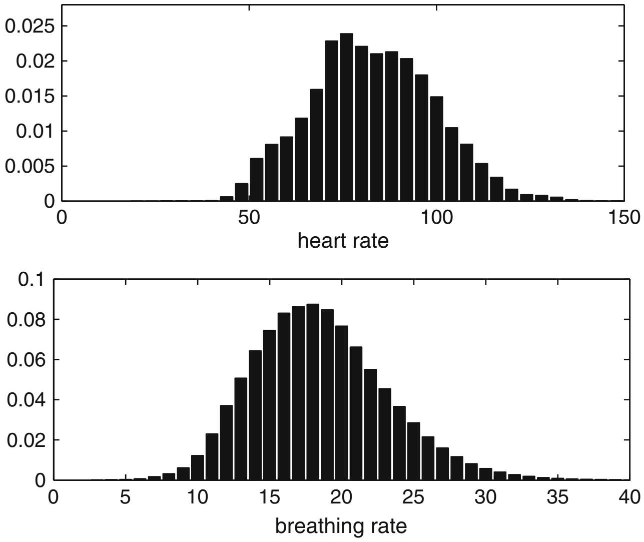

Figure 16 Normalised histograms of the heart rate and breathing rate values for all patients in the training group. The means and standard deviations are, 84.41 and 18.45 bpm for the heart rate, and 16.39 and 4.60 rpm for the breathing rate. The covariance is 14.75.

of the control group. To ease their graphical interpretation, we define novelty scores to be:

$$Z_1(\mathbf{x}_M) = -\ln\left(1 - F_2^e(\mathbf{x}_m)\right). \tag{33}$$

where $\mathbf{x}$ is the extremum of m samples and F_2^e is defined by Eq. 32. Note that $Z_1(\mathbf{x})$ takes low values if $\mathbf{x}$ is close to the centre of the distribution and increases as $\mathbf{x}$ becomes more and more "abnormal". Novelty scores are subsequently assigned to the entire recordings of the patients in all groups. At time t, the highest novelty score of the last m HR and BR measurements is returned. The value of the parameter m is empirically chosen to be 30 for illustration.

We compare the mEVT-based approach with the classical thresholding approach described in Section 2; i.e., setting a threshold in the probability space to which each measurement is individually compared.

To make the comparison with the mEVT-based method easier, we deduce from Eq. 17 that

$$P\left(f_2(\mathbf{x}) \geq f_2(\mathbf{x}_0)\right) = 1 - \exp\left(-\frac{M(\mathbf{x}_0)^2}{2}\right) \tag{34}$$

and therefore define the novelty score for the classical thresholding approach to be:

$$Z_2(\mathbf{x}_0) = -\ln\left(1 - f_2(\mathbf{x}_0)\right) = \frac{M(\mathbf{x}_0)^2}{2}. \tag{35}$$

For a sample $\mathbf{x}$, Z_2 answers the question: "What was the probability of drawing a sample of smaller magnitude?". Z_1 answers the question: "Considering $\mathbf{x}$ and the m-1 samples observed before it, what was the probability of drawing m samples with a more probable extremum?"

7.3 Results

For both approaches, we define $\tau_{\text{training}}(q)$, $\tau_{\text{control}}(q)$, $\tau_{\text{test}}(q)$ and $\tau_{\text{crisis}}(q)$ to be the fraction of the total recording time that novelty scores are higher than the threshold q for the training group, the control group, the test group in the absence of a crisis, and the test group during crises, respectively.

Figure 17 shows the evolution of these fractions as q is increased for the mEVT-based method and the thresholding method. The heterogeneity of the measurements in the crisis windows means that we cannot expect a $\tau_{\text{crisis}} = 100\%$. However, it is important to detect as much of the crisis data as possible to avoid false negatives (where the novelty score is below the threshold during a crisis). If we allow τ_{crisis} to be 90%, then τ_{training}, τ_{control}, τ_{test} equal 1.02%, 1.92%, 5.70% for the mEVT method, respectively, and 4.29%, 4.67%, and 10.25% for the thresholding method.

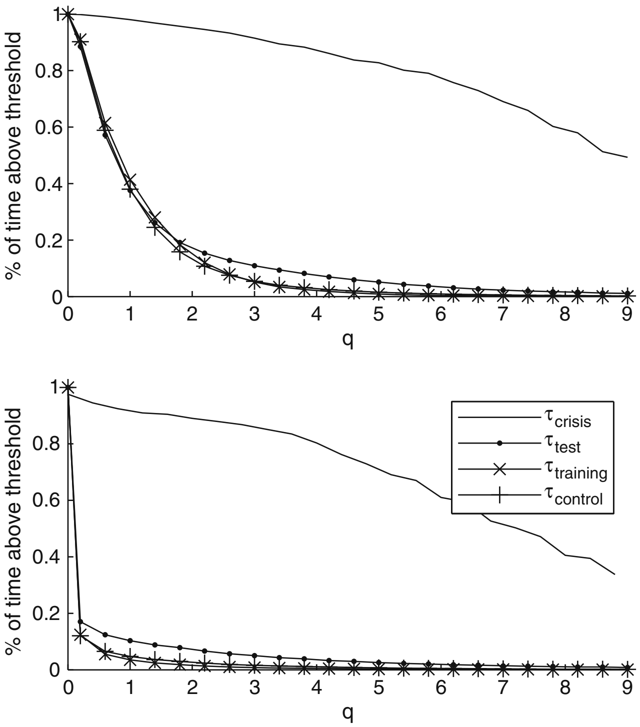

Figure 17 τ_{training}, τ_{control}, τ_{test} and τ_{crisis} for the thresholding method (*upper plot*) and the mEVT-based method (*lower plot*). A warning system should aim to trigger an alarm only during crises; i.e., we must set the threshold so that for an acceptable value of τ_{crisis}, τ_{control} is minimal.

This is a 58.1% reduction of false-positive alert time (where the novelty score is above the threshold in the absence of a crisis) for the control group. This reduction becomes 56.1%, 59.3%, and 50.4% for $\tau_{\text{crisis}} = 85\%$, 80% and 70%, respectively.

Therefore, the mEVT-based method consistently brings a significant improvement to the false-positive time while yielding the same true-positive detection rate.

8 Conclusion

8.1 Discussion

Novelty detection can benefit from a comprehensive multivariate Extreme Value Theory. We have motivated the use of EVT for novelty detection, showing that novelty thresholds set on the generative cdf F_n

have a valid probabilistic interpretation only for single-point classifications; i.e., $m = 1$. We must use EVT to describe where the most extreme of $m > 1$ samples will lie.

We have described existing methods for the estimation of multivariate and multimodal EVDs f_n^e w.r.t. some mixture distribution f_n, and have shown that the estimation of f_n^e is inaccurate for both multivariate and multimodal cases.

We have proposed a numerical method of accurately determining the EVD of a multivariate, multimodal distribution f_n, which is a transformation of the probability density contours of the generative distribution, and have termed this the Ψ-transform. This allows EVDs for mixture models of arbitrary complexity to be estimated by finding the MLE Gumbel distribution f_n^e in the transformed Ψ-space. A novelty threshold may be set on the corresponding univariate cdf F_n^e in the transformed Ψ-space, which describes where the most extreme of m samples generated from f_n will lie.

Furthermore, we have proposed a solution for multivariate, *unimodal* models of normality. Here, by giving an alternate definition of extrema and closed-form solutions for the distribution function over pdf values, we show that we can obtain accurate estimates of the EVDs of multivariate Gaussian kernels.

We applied our formulae to actual patient vital-sign data and showed that the use of EVD is a significant improvement over the conventional thresholding method.

Obtaining these formulae relies on our ability to proceed from Eq. 16 to Eq. 19; i.e., our ability to parameterise the level sets of the generative probability distribution and integrate the resulting parameterisation. While this is relatively easy for multivariate Gaussian distributions, it is not possible for arbitrarily complex, non-symmetrical distributions and fully-analytical closed-form extreme value distributions are not to be expected. However, depending on our ability to estimate the distribution function over the pdf values, accurate estimates of its minimal EVD can be obtained without any sampling of extrema, which would be a great improvement over the existing numerical approach presented in Section 5.

8.2 Future Work

While we have proposed solutions in closed form for multivariate, unimodal models of normality, it should be possible to determine either closed-form solutions, or good approximations to those solutions, for fully multivariate, multimodal distributions. This paper also proposed a numerical solution to estimating the EVDs from such distributions, which was shown to agree closely with experimentally-obtained extrema from such models, but a more light-weight version (that avoids the sampling requirement of the proposed method) would be beneficial for applications in which model training is performed on-line, and where processing resources are constrained.

EVT considers the distribution of the extremum of a set of observed data, motivated by the fact that we wish to consider the extent of “normality” w.r.t. some normal model. However, by considering the additional information contained in the distributions of other order statistics, not just the distributions of the extrema, we may increase our capacity to perform novelty detection.

References

1. Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.
2. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.
3. Castillo, E., Hadi, A. S., Balakrishnan, N., & Sarabia, J. M. (2005). *Extreme value and related models with applications in engineering and science*. New York: Wiley.
4. Clifton, D. (2009). *Novelty detection with extreme value theory in jet engine vibration data*. Ph.D. thesis, University of Oxford.
5. Clifton, D., McGrogan, N., Tarassenko, L., King, S., Anuzis, P., & King, D. (2008). Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *Proceedings of IEEE aerospace* (pp. 1–11). Montana, USA.
6. Clifton, D., Tarassenko, L., Sage, C., & Sundaram, S. (2008). Condition monitoring of manufacturing processes. In *Proceedings of condition monitoring 2008* (pp. 273–279). Edinburgh, UK.
7. Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Berlin: Springer-Verlag.
8. Embrechts, P., Kluppelberg, C., & Mikosch, T. (2008). *Modelling extremal events for insurance and finance* (4th Ed.). Berlin: Springer.
9. Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest members of a sample. *Proceedings of the Cambridge Philosophical Society, 24*, 180–190.
10. Hann, A. (2008). *Multi-parameter monitoring for early warning of patient deterioration*. Ph.D. thesis, University of Oxford.
11. Hayton, P., Tarassenko, L., Schölkopf, B., & Anuzis, P. (2000). Support vector novelty detection applied to jet engine vibration spectra. In *Proceedings of NIPS* (pp. 946–952). London.
12. Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M., & Pinsky, M. (2008). Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system. *Archives of Internal Medicine, 168*(12), 1300–1308.
13. Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M., & Pinsky, M. (2008). Impact of electronic integrated

monitoring system upon the incidence and duration of patient instability on a step down unit. In *Proceedings of 4th international medical emergency team conference*. Toronto.

14. Hyndman, R. (1996). Computing and graphing high density regions. *The American Statistician, 50*(2), 120–126.
15. Ilonen, J., Paalanen, P., Kamarainen, J., & Kälviäinen, H. (2006). Gaussian mixture PDF in one-class classification: Computing and using confidence values. In *Proceedings of the 18th international conference on pattern recognition* (pp. 577–580).
16. Lauer, M. (2001). A mixture approach to novelty detection using training data with outliers. In *Proceedings of the 12th European conference on machine learning, ECML. Lecture notes in computer science* (Vol. 2167, pp. 300–311).
17. Nairac, A., Corbett-Clark, T., Ripley, R., Townsend, N., & Tarassenko, L. (1997). Choosing an appropriate model for novelty detection. In *Proceeding of the 5th IEE international conference on artificial neural networks* (pp. 227–232). Cambridge.
18. Paalanen, P., Kamarainen, J., Ilonen, J., & Kälviäinen (2006). Feature representation and discrimination based on Gaussian mixture model probability densities—practices and algorithms. *Pattern Recognition, 39*, 1346–1358.
19. Roberts, S. J. (1999). Novelty detection using extreme value statistics. *IEE Proceedings on Vision, Image and Signal Processing, 146*(3), 124–129.
20. Roberts, S. J. (2000). Extreme value statistics for novelty detection in biomedical signal processing. *IEE Proceedings on Science, Technology and Measurement, 47*(6), 363–367.
21. Schölkopf, B., Williamson, R., Smola, A. J., Shawe-Taylor, J., & Platt, J. (1999). Support vector method for novelty detection. In *Proceedings of NIPS* (pp. 582–588).
22. Scott, D. (1992). *Multivariate density estimation. Theory, practice and visualization*. New York: Wiley.
23. Sohn, H., Allen, D. W., Worden, K., & Farrar, C. (2005). Structural damage classification using extreme value statistics. *Journal of Dynamic Systems, Measurement, and Control, 127*(1), 125–132.
24. Tarassenko, L., Hann, A., Patterson, A., Braithwaite, E., Davidson, K., Barber, V., et al. (2005). Biosign: Multiparameter monitoring for early warning of patient deterioration. In *Proceedings of the 3rd IEEE international seminar on medical applications of signal processing* (pp. 71–76).
25. Tarassenko, L., Hann, A., & Young, D. (2006). Integrated monitoring and analysis for early warning of patient deterioration. *British Journal of Anaesthesia, 98*(1), 149–152.
26. Tarassenko, L., Hayton, P., Cerneaz, N., & Brady, M. (1995). Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th IEE international conference on artificial neural networks* (Vol. 4, pp. 442–447). Perth, Australia.
27. Tarassenko, L., Nairac, A., Townsend, N., Buxton, I., & Cowley, P. (2000). Novelty detection for the identification of abnormalities. *International Journal of Systems Science, 31*(11), 1427–1439.
28. Worden, K., Manson, G., & Allman, D. (2003). Experimental validation of a structural health monitoring methodology: Part I. Novelty detection on a laboratory structure. *Journal of Sound and Vision, 259*(2), 323–343.

David A. Clifton is the W.K. Kellogg Junior Research Fellow in Biomedical Engineering at the Institute of Biomedical Engineering in the University of Oxford, and a College Lecturer in Engineering Science at Pembroke College. He received an MEng degree in Engineering Mathematics from the University of Bristol and a DPhil degree in Machine Learning from the University of Oxford, for which he was awarded prizes from the IEEE, the IOP, the IET, the UK Engineering Council, the European Commission, Hewlett-Packard, Shell, Abbey-Santander, and the UK Engineering Prize from the House of Commons. His research interests include machine learning for condition monitoring of biomedical and industrial systems, and his work has resulted in patented monitoring systems used in the engines of the Eurofighter, the Airbus A380, and the Boeing 787 "Dreamliner"; in engine manufacturing facilities at Rolls-Royce PLC.; and in patient vital-sign monitors.

Samuel Hugueny received a Master's Degree in Computing Engineering from the École Nationale Supérieure de Techniques Avancées (ENSTA-Paritech, Paris) in 2006. He is currently reading a DPhil at the University of Oxford, as a member of both the Institute of Biomedical Engineering (Department of Engineering Science), and the EPSRC Life Sciences Interface Doctoral Training Centre. His research areas of interest include medical imaging, signal processing, novelty detection, extreme value statistics and patient monitoring.

Lionel Tarassenko was born in Paris, France, in 1957. He received the B.A. degree in engineering science in 1978, and the Ph.D. degree in medical engineering in 1985, both from Oxford University, Oxford, U.K.. After graduating, he worked for Racal Research Ltd. on the development of digital signal processing techniques, principally for speech coding. He then held a number of positions in academia and industry, before taking up a University Lecturership at Oxford in 1988. Since then, he has devoted most of his research effort to the development of neural network techniques and their application to signal processing, diagnostic systems, and parallel architectures. He has held the Chair in Electrical Engineering at Oxford University since October 1997. He was elected to a Fellowship of the Institution of Electrical Engineers (IEE) in 1996, when he was also awarded the IEE Mather Premium for his work on neural networks, and to a Fellowship of the Royal Academy of Engineering (RAE) in 2000.