NOVELTY DETECTION WITH MULTIVARIATE EXTREME VALUE THEORY, PART I: A NUMERICAL APPROACH TO MULTIMODAL ESTIMATION

David A. Clifton, Samuel Hugueny†, Lionel Tarassenko*

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Roosevelt Drive, Oxford, OX3 7DQ, UK
{davidc, lionel}@robots.ox.ac.uk, samuel.hugueny@eng.ox.ac.uk

ABSTRACT

Extreme Value Theory (EVT) describes the distribution of data considered extreme with respect to some generative distribution, effectively modelling the tails of that distribution. In novelty detection, or one-class classification, we wish to determine if data are "normal" with respect to some model of normality. If that model consists of generative distributions, then EVT is appropriate for describing the behaviour of extremes generated from the model, and can be used to determine the location of decision boundaries that separate "normal" areas of data space from "abnormal" areas in a principled manner. This paper introduces existing work in the use of EVT for novelty detection, shows that existing work does not accurately describe the extrema of multivariate, multimodal generative distributions, and proposes a novel method for overcoming such problems. The method is numerical, and provides optimal solutions for generative multivariate, multimodal distributions of arbitrary complexity. In a companion paper, we present analytical closed-form solutions which are currently limited to unimodal, multivariate generative distributions.

1. INTRODUCTION

Novelty detection, or one-class classification, 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 jet engines, manufacturing processes, or power-generation facilities, which spend the majority of their operational life in a "normal" state, and which exhibit few, if any, failure conditions.

*Supported by the NIHR Biomedical Research Centre, Oxford.
†Supported by the EPSRC LSI Doctoral Training Centre, Oxford.

1.1. Existing Work

In much of the existing work on novelty detection, an assumption is made that "normal" data $\mathbf{x}$ are i.i.d., and distributed according to some underlying generative distribution $f_n(\mathbf{x})$, which is a probability distribution over an n-dimensional data space, $\mathcal{D}$; i.e., $\mathbf{x} \in \mathcal{D} = \mathbb{R}^n$. Typically, f_n is multivariate and multimodal; it could be, for example, a Gaussian mixture model (GMM), Parzen window estimator, or other mixture of kernels. 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.

In previous work, a heuristic novelty threshold has been set on the unconditional pdf $f_n(\mathbf{x}) = \kappa$, such that $\mathbf{x}$ is classified "abnormal" if $f_n(\mathbf{x}) < \kappa$. Such thresholds have 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 [1, 2] have interpreted f_n probabilistically, by considering the cumulative probability $F_n(\mathbf{x})$ associated with $f_n(\mathbf{x})$. 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 $\{\mathbf{x} \in \mathcal{R} | f_n(\mathbf{x}) \geq \kappa\}$:

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

For a unimodal pdf, this 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. 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

978-1-4244-4948-4/09/$25.00 © 2009 IEEE

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., $F_n = 0.99$. 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.

2. 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 [3, 4, 5, 6], in contrast to a method commonly used in estimating financial risks, often termed the peaks-over-threshold technique [7].

2.1. Extreme Value Distributions

Consider a set of m i.i.d. data, which we here assume to be univariate ($n = 1$) for the purposes of simplifying the introduction, $\mathbf{X} = \{x_1, x_2, \ldots, x_m\}$, distributed according to some function $f_1(x)$, 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.

According to the Fisher-Tippett theorem [8] upon which classical EVT is based, H^+ must belong to one of the following three families of distributions, 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 of x, termed the *reduced variate*, $y = (x - c_m)d_m^{-1}$, 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 [7]:

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

In this paper, we are primarily concerned with mixtures of Gaussian distributions, for which the limiting distribution is the Gumbel distribution H_1^+. This limiting cdf $F_n^e = H_1^+$ has a corresponding pdf f_n^e, which we term the *extreme value distribution* (EVD). Note that in the case of $m = 1$, 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.

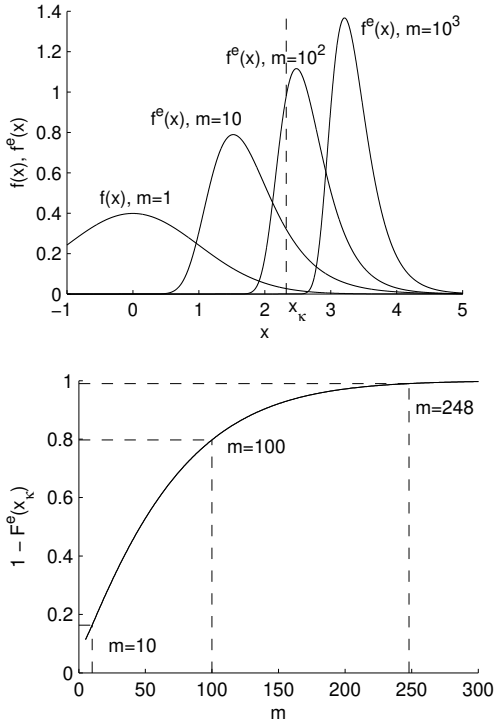

Fig. 1. EVDs $f_1^e(\mathbf{x})$ and probability $1 - F_1^e(\mathbf{x}_\kappa)$ in the upper and lower plots, respectively, for increasing m.

2.2. Disadvantages of Using $F_n(\mathbf{x})$

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 1.1. For the purpose of illustration, suppose that $f_1 = N(0, 1)$, the standard univariate Gaussian distribution.

The upper plot in Figure 1 shows the generative distribution f_1, and the EVDs f_1^e for increasing numbers of samples m. As more samples are generated from f_1, the expected location of their maximum increases on the x-axis. A novelty threshold has been set at $F_1 = 0.99$, which occurs at $x = x_\kappa$, and which is shown in the figure as a dashed line. We anticipate that this novelty threshold will be exceeded with probability $1 - F_1 = 1 - 0.99 = 0.01$ 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, $1 - F_1^e(x_\kappa)$.

The lower plot in Figure 1 shows the probability that the novelty threshold set at $F_1 = 0.99$ will be exceeded for increasing numbers of samples m. After $m = 10$ samples, the threshold will have been exceeded with probability 0.163, and after $m = 100$, the threshold will have been exceeded with probability 0.797. The threshold is exceeded with probability 0.99 after $m = 248$ samples have been generated.

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 [9].

3. EVT FOR NOVELTY DETECTION

Existing work on the use of EVT for novelty detection has been limited to [3, 4, 5, 6, 10]. We here consider the use of EVT for multivariate and multimodal novelty detection, and identify problems with existing approaches in both cases.

3.1. EVT in Multivariate Novelty Detection

Multivariate extrema exist in the EVT literature [7, 11], are also termed *block extrema*, and 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 model of normality, rather than considering block extrema. EVT was first used for novelty detection in multivariate data in [3, 4], where models of normality were represented by mixtures of Gaussian kernels. In multivariate space, the Gaussian kernel 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. [3, 4] uses this assumption to reduce the problem of determining the EVD for a multivariate Gaussian kernel to a corresponding univariate case. As illustrated in Figure 2, the EVD for a single Gaussian kernel 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 (5).

Figure 3 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 [11] for increasing dimensionality n of the Gaussian kernel. 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.

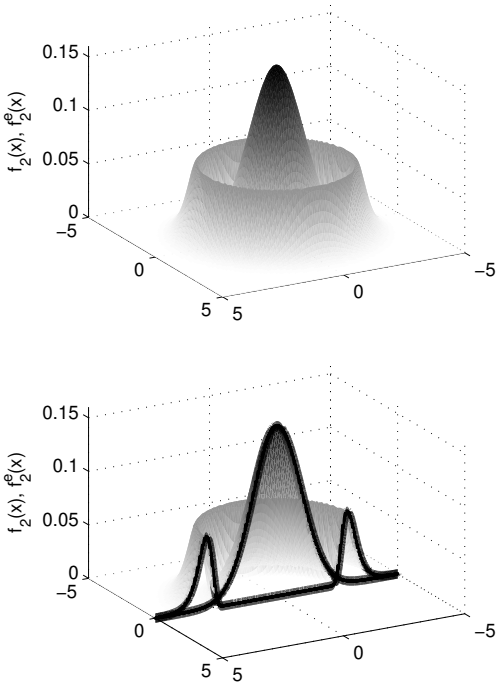

Fig. 2. The bivariate Gaussian f_2 and its EVD f_2^e (shown in the upper plot), which is a torus with Gumbel cross-section (shown in the lower plot).

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

3.2. EVT in Multimodal Novelty Detection

EVT defines an "extreme value" to be that which is either a minimum or maximum of a set. This is due to the conventional use of EVT 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, 6, 10], this existing definition of "extreme value" is sufficient for univariate f_1. For multivariate data, providing that f_n is *unimodal*, 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 3.1).

However, for multimodal f_n, whether uni- or multivariate, the notion of minimum or maximum value is longer sufficient, because there is no single mode from which distance may be defined. 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

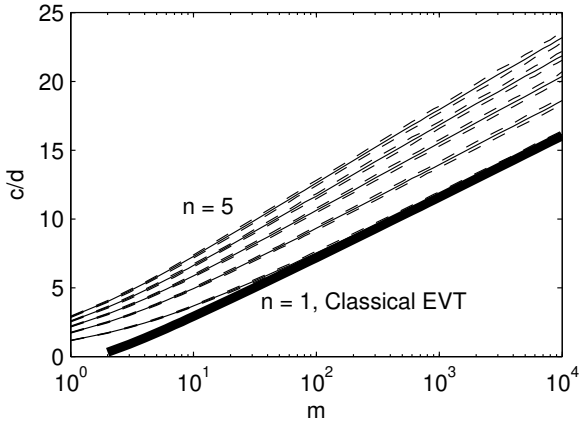

Fig. 3. Ratio c/d of the MLE parameters for the multivariate Gumbel cross-section as a function of m for increasing dimensionality $n = 1 \ldots 5$ (thin lines), compared to c/d estimated by classical EVT (thick line). MLEs were obtained from $N = 10^6$ experiments for each value of m and n. Confidence intervals are shown at 2 standard deviations from the mean.

redefine "extreme value" in terms of probability:

Definition 1. *For novelty detection, the "most extreme" of a set of m samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ generated from a distribution $f_n(\mathbf{x})$ is that which is most improbable with respect to the distribution; i.e.,* $\operatorname{argmin}_{\mathbf{x}} f_n(\mathbf{x})$.

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 $f_n(\mathbf{x})$ is equivalent to selecting extrema by maximising $F_n(\mathbf{x})$ using (1). Our definition satisfies the condition of [1], which states that the probability of observing data $f_n(\mathbf{x})$ 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 $f_n(\mathbf{x})$ 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 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.

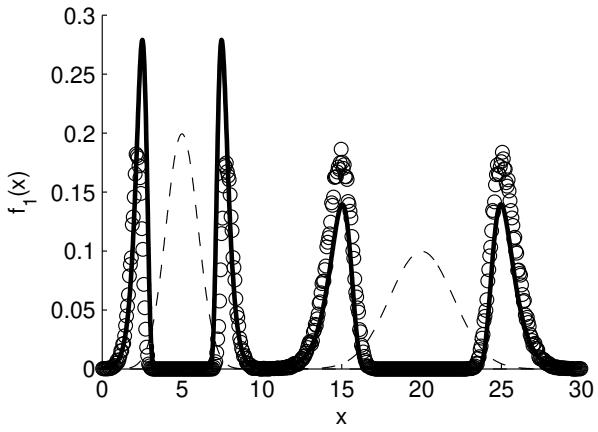

Fig. 4. Bimodal generative pdf f_1 (dashed line) with EVD f_1^e predicted by existing methods (solid line). The histogram for $N = 10^6$ experimentally-obtained extrema for $m = 100$ is shown by the circles.

As with multivariate novelty detection, [3, 4] is the only existing work in using EVT for novelty detection with multimodal f_n. In this work, the multimodal distribution represented by a mixture of Gaussian kernels is reduced to a single-kernel problem: to find the value of the EVD f_n^e at some location $\mathbf{x}$, the closest kernel (determined using Mahalanobis distance) is assumed to dominate $f_n^e(\mathbf{x})$, and thus the EVD is based on the Gumbel distribution corresponding to that closest kernel (using radius r from that kernel centre, as described in Section 3.1). Here, the contribution of other kernels to $f_n^e(\mathbf{x})$ is assumed to be negligible, and they are ignored.

However, as shown in Figure 4, this existing method produces poor estimates for the actual EVD f_n^e when compared to the distribution of experimentally-obtained extrema.

4. UNDERSTANDING THE EVD

This section presents a new method of understanding the EVD, which we require in order to estimate the EVD for multivariate, multimodal f_n.

4.1. 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 5 shows equiprobable contours for a bivariate model of normality f_2 represented by a mixture of three Gaussian kernels with full

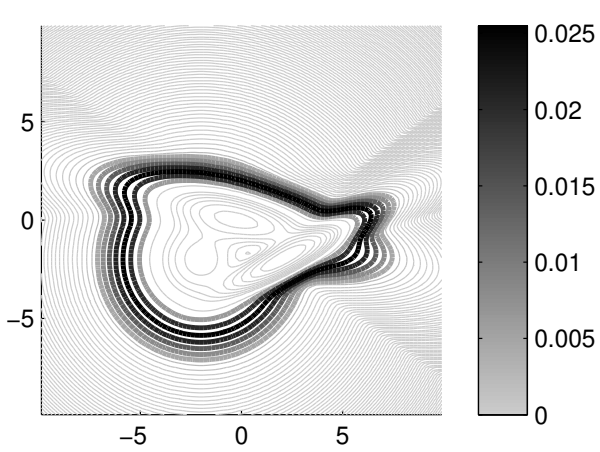

Fig. 5. Probability contours on f_n (light grey) for a trimodal GMM with full (non-diagonal) covariance matrices and corresponding EVD f_n^e.

(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,

$$f_n^e(\mathbf{x}) = g\left[f_n(\mathbf{x})\right] \tag{6}$$

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.

4.2. The Ψ-Transform

For a standard Gaussian, $f_n(x) = (2\pi)^{-n/2}\exp\left(-r^2/2\right)$, and so $r = \left(-2\ln f_n(x) - n\ln 2\pi\right)^{1/2}$. We define a transform,

$$\Psi\left[f_n(\mathbf{x})\right] = \begin{cases} \left(-2\ln f_n(\mathbf{x}) - n\ln 2\pi\right)^{1/2} & \text{if } f_n(\mathbf{x}) < K \\ 0 & \text{if } f_n(\mathbf{x}) \geq K \end{cases} \tag{7}$$

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 is distributed according to the Gumbel distribution (as shown in Section 3.1). 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 [12].

The upper plot in Figure 6 shows a normalised histogram of $N = 10^6$ extrema generated from the example mixture of three Gaussian kernels f_2 in Figure 5, for $m = 100$. The

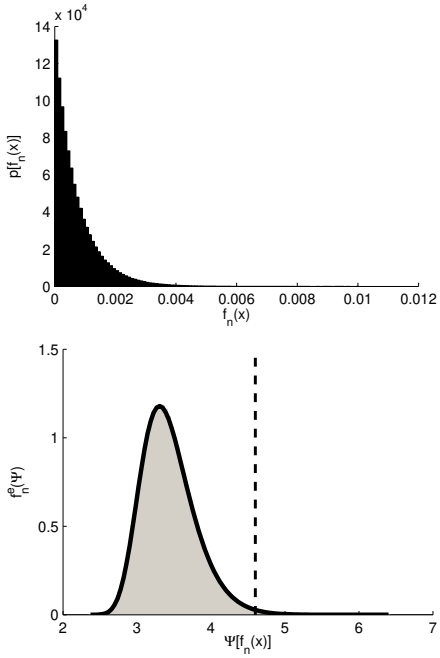

Fig. 6. Normalised histogram of $f_2(\mathbf{x})$ values for $N = 10^6$ extrema generated from trimodal GMM with $m = 100$ (upper plot). Histogram of the Ψ-transformed $f_2(\mathbf{x})$ values shown in grey, with the corresponding MLE Gumbel distribution fitted in Ψ-space, shown in black (lower plot).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

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 extrema, which may be seen to be distributed according to the Gumbel. The MLE Gumbel distribution fitted in Ψ-space using the univariate, unimodal method of [11] 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 Figure 5, 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., $F_n^e = 0.99$, as shown in Figure 6. 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).

Note finally that, though the novelty threshold set using our proposed method occurs at some contour $f_n = \kappa$ (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 $1 - F_n^e = 1 - 0.99 = 0.01$; that is, the final novelty threshold has a valid probabilistic interpretation provided by EVT.

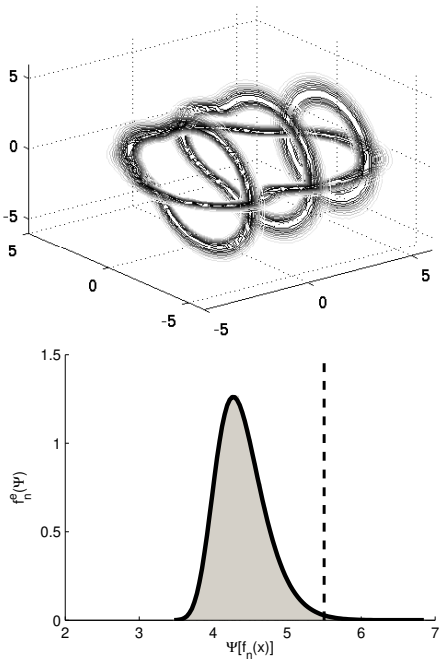

Fig. 7. EVD f_3^e shown in data space $\mathcal{D}$ estimated using the proposed Ψ-transform method for a trivariate, trimodal f_3 (upper plot). Histogram of Ψ-transformed extrema (shown in grey) and MLE Gumbel (shown in black) for a 6-dimensional f_6 mixture of 15 components. A novelty threshold has been set at $F_6^e = 0.99$, shown as a dashed line.

5. CONCLUSION

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 method of accurately determining the EVD of a multivariate, multimodal distribution f_n, which is a transformation of the probability 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.

This method scales with dimensionality n and the number of kernels in the mixture f_n. Figure 7 shows the EVD for a trivariate model f_3 (projected onto four planes for visualisation), which closely matches the EVD observed from experimentally-obtained extrema. The figure 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.

This 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 for multivariate, multimodal distributions f_n. Currently, closed forms have been obtained for multivariate, unimodal distributions, which are presented in a companion paper [13].

6. REFERENCES

[1] R.J. Hyndman, "Computing and graphing high density regions," *The American Statistician*, vol. 50, no. 2, pp. 120–126, 1996.

[2] A. Hann, *Multi-parameter monitoring for early warning of patient deterioration*, Ph.D. thesis, University of Oxford, 2008.

[3] S. J. Roberts, "Novelty Detection Using Extreme Value Statistics," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 146, no. 3, pp. 124–129, 1999.

[4] S. J. Roberts, "Extreme Value Statistics for Novelty Detection in Biomedical Signal Processing," *IEE Proceedings on Science, Technology and Measurement*, vol. 47, no. 6, pp. 363–367, 2000.

[5] K. Worden, G. Manson, and D. Allman, "Experimental Validation of a Structural Health Monitoring Methodology: Part I. Novelty Detection on a Laboratory Structure," *Journal of Sound and Vision*, vol. 259, no. 2, pp. 323–343, 2003.

[6] H. Sohn, D. W. Allen, K. Worden, and C.R. Farrar, "Structural damage classification using extreme value statistics," *Journal of Dynamic Systems, Measurement, and Control*, vol. 127, no. 1, pp. 125–132, 2005.

[7] P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin, 4 edition, 2008.

[8] R. A. Fisher and L. H. C. Tippett, "Limiting Forms of the Frequency Distributions of the Largest or Smallest Members of a Sample," *Proceedings of the Cambridge Philosophical Society*, vol. 24, 1928.

[9] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," in *Proceedings of the 4th IEE International Conference on Artificial Neural Networks*, Perth, Australia, 1995, vol. 4, pp. 442–447.

[10] D.A. Clifton, N. McGrogan, L. Tarassenko, S. King, P. Anuzis, and D. King, "Bayesian extreme value statistics for novelty detection in gas-turbine engines," in *Proceedings of IEEE Aerospace, Montana, USA*, 2008, pp. 1–11.

[11] E. Castillo, A. S. Hadi, N. Balakrishnan, and J. M. Sarabia, *Extreme Value and Related Models with Applications in Engineering and Science*, John Wiley and Sons, New York, 2005.

[12] D.A. Clifton, *Novelty Detection with Extreme Value Theory in Jet Engine Vibration Data*, Ph.D. thesis, University of Oxford, 2009.

[13] S. Hugueny, D.A. Clifton, and L. Tarassenko, "Novelty detection with multivariate extreme value theory, part I: An analytical approach to unimodal estimation," in *Proceedings of IEEE Machine Learning in Signal Processing, In press.*, 2009.