Extreme Value Theory for Novelty Detection in Vital-Sign Monitoring

August 21, 2009

Samuel Hugueny
Balliol College

Supervised by

Professor Lionel Tarassenko & Dr. David A. Clifton
Submitted: Michaelmas 2009

This PRS Transfer is Submitted to the Department of Engineering Science
Contents

1 Project Summary 4

2 Literature Review 6
   2.1 The clinical need: in-hospital patient monitoring 6
   2.2 State of the art in patient monitoring: the Visensia project 7
   2.3 Novelty detection and Extreme Value Theory 9

3 Multivariate Approach to Extreme Value Theory for Novelty Detection 13
   3.1 Key results of classical Extreme Value Theory 13
      3.1.1 Classical EVT results 13
   3.2 Redefining extrema 16
   3.3 Distribution of probability density values 18
   3.4 Novelty scores 19

4 Analytical Extreme Value Distributions for Multivariate Gaussian Distributions 22
   4.1 Analytical Distribution of PDVs 22
   4.2 Maximum domain of attraction of the weibull distribution 24
   4.3 Parameter estimation 24
   4.4 Novelty scores 27

5 Extreme Value Distributions for Mixture of Gaussians 29
   5.1 The Ψ-transform method 29
   5.2 Histogram fitting via least-squares optimisation 31
   5.3 Validation on simulated data 34

6 EVT for Novelty Detection in Time Series 37
   6.1 Scheme 37
   6.2 Proof of principle 37

7 Application to a Vital-Sign Data Set 42
   7.1 The data set 42
   7.2 Method 43

8 Discussion and Future Work 46
   8.1 Exact Extreme Value Distributions 46
   8.2 Limitations of current approaches 47
   8.3 Future work 50
      8.3.1 Order statistics for novelty detection 50
      8.3.2 Patient-specific modelling 52
      8.3.3 Dynamic Analysis 53
List of abbreviations and symbols

\[ \alpha \]  
Weibull or Fréchet shape parameter.

\[ \text{BR} \]  
brathing rate.

\[ c_m \]  
scale parameter of extreme value distributions.

\[ c \]  
classical univariate extreme value theory.

\[ d_m \]  
location parameter of extreme value distributions.

\[ \text{cdf} \]  
cumulative distribution function.

\[ \text{df} \]  
distribution function.

\[ \text{EVD} \]  
extreme value distribution.

\[ \text{EVT} \]  
extreme value theory.

\[ f_n \]  
distribution function and distribution of an n-variate random variable.

\[ F_n^{-} (p) \]  
p-quantile of \( F_n \).

\[ g_n \]  
probability density function.

\[ G_n \]  
distribution of probability density values associated with \( F_n \).

\[ \text{GMM} \]  
Gaussian mixture model.

\[ H^{+}_1 \]  
maximal Gumbel distribution.

\[ H^{-}_1 \]  
minimal Gumbel distribution.

\[ H^{+}_2 \]  
maximal Fréchet distribution.

\[ H^{-}_2 \]  
minimal Fréchet distribution.

\[ H^{+}_3 \]  
maximal Weibull distribution.

\[ H^{-}_3 \]  
minimal Weibull distribution.

\[ \text{HR} \]  
heart rate.

\[ \text{iid} \]  
independent and identically distributed.

\[ m \]  
number of samples from which an extremum is drawn; this is the primary parameter of EVT.

\[ M_m \]  
maximum of a set of random samples \( X_1, \ldots, X_m \).

\[ \text{MDA} \]  
maximum domain of attraction.

\[ \text{mEVT} \]  
multivariate extreme value theory.

\[ \text{MEWS} \]  
modified early warning score.

\[ n \]  
dimensionality of the sample (data) space.

\[ \text{ROC} \]  
receiving operating characteristic.

\[ \text{rv(s)} \]  
random variable(s).

\[ \text{pdf} \]  
probability density function.

\[ \text{pdv} \]  
probability density value.

\[ \text{PSI} \]  
patient status index.

\[ \text{SDA} \]  
systolic diastolic average, a measure of blood pressure.

\[ \text{SpO}_2 \]  
saturation of peripheral oxygen, a measure of blood oxygen saturation.

\[ d_m \rightarrow A \]  
\( A_n \rightarrow A \): convergence in distribution.

\[ F \sim G \]  
tail-equivalence: \( F \sim G \) means that the distributions \( F \) and \( G \) are tail equivalent.
1 Project Summary

Studies indicate that a large number of in-hospital cardiac events could be prevented, provided that the preceding physiological abnormalities are identified and acted upon. Continuous monitoring of patient physiological condition is not a task that can reasonably be assigned to the nursing staff, but one that could benefit from the use of an automated system. To this end, an early warning system based on a data-driven novelty detection scheme has been designed by Tarassenko et al. [44]. In this system, vital sign data (such as heart rate, respiration rate, etc.) collected from clinical trials are fused to create a multi-patient ‘model of normality’ in the form of an unconditional multivariate probability density function, using data from a population of ‘high-risk’ adult patients. A novelty threshold is set on this pdf and used to discriminate between normal and abnormal test data, with respect to the model. The accuracy of using this methodology to classify clinically reviewed data as normal or abnormal is reported in [19].

Despite performing well when used to classify vital sign data, the existing method uses a heuristic approach to the assignment of novelty scores to test data, and the manner in which resulting scores are post-processed. In this report, we introduce developments in novelty detection based on Extreme Value Theory, a branch of statistics which describes the tails of univariate probability distributions. For the purpose of this work, we redefine ‘extrema’ to include all regions of low probability and show that novelty scores and thresholds can be assigned in a more principled manner; i.e., in relation to meaningful probabilities. Classical Extreme Value Theory is a theory applicable to univariate data, and thus a generic methodology for the study of multivariate distributions is proposed. If the model of normality is assumed to be a multivariate Gaussian distribution, extreme value distributions can be determined in a fully analytical manner. If the model of normality is multimodal but can be approximated accurately by a mixture of Gaussian distributions, a semi-numerical scheme is derived and demonstrated.

The question of the application of our proposed multivariate Extreme Value Theory to time-series of data is then examined. EVT deals with statistical samples that are independent and identically distributed, an assumption that rarely holds in real life and certainly does not hold for vital sign
data, in which consecutive observations are very clearly correlated. The straightforward application of our EVT-based method to a sliding window of time-series data is tested on simulated and real-life data, yielding better results than the existing density thresholding method used in [19].

However, in order to overcome the assumption that vital sign data for a patient are iid, work in the immediate future will focus on the investigation of novelty detection methods for dependent observations, and on increasing the amount of information used by considering not only the ‘extremum’ of a window but all the order statistics.

Much of the work presented here has already been reported in four conference papers [8, 28, 7, 27]: the first two have been accepted for presentation at the 2009 IEEE International Workshop on Machine Learning for Signal Processing (Grenoble, France); the third at the 2009 IEEE Workshop on Statistical Signal Processing (Cardiff, UK); and the fourth was submitted for review at Biosignals 2010: Third International Conference on Bio-inspired Systems and Signal Processing (Valencia, Spain).
2 Literature Review

2.1 The clinical need: in-hospital patient monitoring

Every year in the UK, at least 20,000 hospital patients have unplanned transfers to an Intensive Care Unit, and at least 23,000 have cardiac arrests or die unexpectedly. In a number of cases, timely intervention by clinicians could prevent these adverse events, provided that the preceding patient deterioration was identified and acted upon \cite{22,32}. Although studies \cite{38,40} have reported that most patients who have one of these events show one or more significant physiological abnormalities within the 24-hour period preceding it, such deterioration in patient condition can go unnoticed, due to the infrequency of observations made by the nursing staff, or failure to identify the warning signs of patient deterioration. To address this issue, hospital-wide strategies to detect and treat patients before adverse consequences arise, known as Rapid Response Systems (RRSs), have been developed. These systems are based on the premise that early recognition of physiological abnormalities coupled with the rapid intervention of suitably trained staff should result in an improvement in patient outcome or mortality rate. RRSs are typically referred to as Medical Emergency Teams (METs) in the USA and Critical Care Outreach Teams (CCOT) in the UK. Although their capabilities differ, their role is essentially to provide surveillance, identify deteriorating patients and begin an ICU-level of care at the bedside. In the case of METs, calling criteria have been established, such that any member of the nursing staff may trigger a call of the MET to the patient’s bedside. These rely either on alerting limits based on single vital sign parameters, or on a multi-parameter scoring system, such as the Modified Early Warning Score (MEWS, Table 2.1).

<table>
<thead>
<tr>
<th>SBP (mmHg)</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR (bpm)</td>
<td>&lt; 70</td>
<td>71-80</td>
<td>81-100</td>
<td>101-199</td>
<td>200+</td>
<td>201+</td>
<td>300+</td>
</tr>
<tr>
<td>BR (rpm)</td>
<td>&lt; 9</td>
<td>9-14</td>
<td>15-29</td>
<td>≥ 30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temp. (°C)</td>
<td>≤ 35</td>
<td>35-38.5</td>
<td>≥ 38.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVPU</td>
<td>Alert</td>
<td>Reacting to Voice</td>
<td>Reacting to Pain</td>
<td>Unresponsive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Modified Early Warning Score (MEWS). A score greater than 5 is associated with increased risk of death at 60 days, ICU admission and high-dependency unit (HDU) admission. The AVPU score is a measure of level of consciousness.
The effect of METs on patient outcome and length of stay has generally been deemed positive by observational studies [23, 1], supporting the premise that patient care benefits from frequent monitoring and specific training of intervention teams. To date, there has only been one randomised controlled trial of METs, the MERIT study [21]. Although it failed to show a significant benefit in hospitals where a MET was introduced despite an increased number of emergency calls, its authors noted that more than half of the patients who satisfied the MET calling criteria had no call made on their behalf - indicating a problem with the MET activation protocol - and conclude that “these findings suggest the need for improved intensive monitoring of patients in general wards; frequent and rigorous documentation of patients’ condition; and increased attention to education to ensure a timely response by appropriately trained clinicians”. In the UK, the National Institute of Health and Clinical Excellence has formulated similar recommendations. They state that heart rate, respiratory rate, systolic blood pressure, level of consciousness and oxygen saturation observations “should be part [...] of the routine” and that the frequency of monitoring should increase when abnormal physiology is detected. They also identify the implementation of “physiological track and trigger systems [...] to monitor all adult patients in acute hospital settings” as a key priority [17].

2.2 State of the art in patient monitoring: the Visensia project

From such evidence presented by the studies described above, it appears that current care of critically ill patients is suboptimal and could benefit from an accurate automated system. The Visensia (formerly BioSign) project was developed by Tarassenko et al. [44] to fulfill this need. Designed for use as a real-time early warning system at the bedside for triggering the intervention of METs or CCOTs, the Visensia system addresses the problem of abnormal vital sign detection as a novelty detection one. It fuses measurements of vital signs (heart rate, breathing rate, systolic-diastolic pressure, oxygen saturation and in the pilot studies, temperature) and compares the result to a model of normality constructed using data from a population of high-risk surgical and medical patients. Test data are compared to this model, and assigned a univariate representation of patient
status, the Patient Status Index (PSI). The data fusion method adopted is a multidimensional probabilistic model of normality (where the number of dimensions equals the number of vital signs monitored), learnt from the data acquired from high-risk adult patients during previous studies. The PSI is thus derived from a data-driven approach, as opposed to the rule-based approach adopted by heuristic clinical systems. Scoring systems such as the aforementioned MEWS are either based on clinical experience or related to the prevalence of individual abnormal signs in the general ward population. The predictive value of such rules for specific adverse events remains hard to validate, whereas the performance of a data-driven model can be assessed using case-examples reviewed by clinicians.

The model of normality in the Visensia system is an estimate of the unconditional probability density function of the training data $\hat{p}(x)$, where $x$ is the vector of normalised vital-sign parameters. The pdf is then modelled using an $N$-kernel Parzen Windows estimator (where typically $N = 400$, see [34]). The PSI is defined to quantify departures from normality, understood to be observations with low values of $\hat{p}(x)$:

$$\text{PSI} = -\ln \left[ \hat{p}(x) \right]$$

An alert is generated when this index increases above a threshold value for a fixed period (e.g., 4 minutes out of the previous 5 minutes). This threshold must then be set so as to maximise the detection of clinically-validated event data, while minimising the occurrence of false detections on normal data.

Using a model constructed using over 3,500 hours of data, Tarassenko et al. report in [45] that 95% of all alerts generated were ‘true positives’, as assessed by clinicians (i.e., a positive predictive value of 0.95. This value was 0.098 in the MERIT study) and shows that the Visensia system is capable of detecting critical events in advance of single-channel alerts. A later clinical study enabled Hann [19] to define a ‘gold standard’ of critical events. An improved model of normality is retrained with a much larger training set and validated using Receiving-Operating-Characteristic (ROC) analysis against this gold standard. The best performing model achieves a sensitivity on an event basis of 76.7% and a specificity on a patient basis of 76.7% (50.4% and 93.3% in the
2.3 Novelty detection and Extreme Value Theory

The task of vital-sign abnormality detection could be considered to be a two-class classification problem, using the two classes ‘normal’ and ‘abnormal’. A model could then be constructed using a training set of patient data containing instances from both classes, using a supervised learning scheme (such as support vector machines, artificial neural networks, etc.). Such an approach would rely heavily on the usual assumption of supervised classification, which is that the training set contains sufficient information regarding all classification hypotheses. In other words, supervised learning assumes that the training procedure has seen a significant number of examples all possible classes. This assumption breaks down under two major conditions:

1. if at least one important class is under-represented, it becomes increasingly difficult to make decisions regarding this class as the number of examples decreases. In a two-class classification problem, for instance, systematically classifying data as being from the more populous class may yield a smaller error rate than performing actual classification. This is often the case in medical problems: data from normal states are inexpensive to acquire, whereas abnormal data are comparatively fewer in number. Meta-algorithms such as bootstrapping or boosting are designed to construct classifiers when classes are as inconveniently imbalanced.

2. if the training set contains information regarding all the defined classes, but the number of classes itself has been underestimated; i.e., data from some previously unseen class are observable in the test set. Such a classification system will classify this unaccounted-for data erroneously into one of the defined classes, based on its knowledge of the training data.

Both situations are applicable to systems that have a high degree of complexity. As system complexity increases, so does the number of possible states and the influence of parameters on observable data becomes harder to describe. Therefore we must be able to differentiate between known and unknown data during testing, prior to performing a classification task. For this purpose, novelty detection theory is based on constructing a model of the normal data and then testing
for novel data against this model. Departures from normal behaviour are classified as being novel events. A large variety of techniques have been proposed to perform novelty detection including density function thresholding, Gaussian Mixture Models, hypothesis testing, Hidden Markov Models, Artificial Neural Networks and Extreme Value Theory. These techniques have been applied to a number of domains: fault detection, radar target detection, computer-aided diagnosis, hand-written digit recognition, e-commerce and structural health monitoring, amongst others. (See [31] for a complete review).

The approach used in the Visensia system is popular in the literature: it relates the concept of novelty to the probability of observing data which do not belong to a distribution characterising normal data. Fundamental to this approach is the assumption that normal data are generated from an underlying data distribution, and that this distribution may be estimated from training data. Estimation of the underlying distribution is based on the use of a density estimation technique (in Visensia, this is an approximation by a mixture of Gaussians). The resultant distribution \( \hat{p}(x) \) is then thresholded to define the boundaries of ‘normal’ areas of data space.

This the novelty threshold on \( \hat{p}(x) \) was set such that a large fraction of the probability mass lies within the bounds of the novelty threshold. This empirically-determined value was then validated during a later clinical trial. This conventional method is well-suited for single-sample classification since any single sample drawn from the distribution has a very low probability of exceeding the novelty threshold. However, the probability of exceeding the novelty threshold increases rapidly as more samples are drawn from \( \hat{p}(x) \). EVT tells us where we should expect the largest (or smallest) of \( m \) samples drawn from \( \hat{p}(x) \) to lie, and thus allows us to set our novelty thresholds accordingly. EVT may therefore be better suited for a dynamic situation such as continuous vital-sign monitoring (or time-series analysis in general). Ultimately, the aim of EVT is also to set a novelty threshold in the probability space, but it does so in a principled fashion, while providing a less heuristic scoring system than was used for the Patient Status Index in [19].

EVT is the branch of statistics concerned with the distributions of ‘abnormally’ low or high values in the tails of some data-generating distributions by forming representations of their tails. In
classical data analysis tasks, extremes are labeled as outliers and often ignored. In the case of novelty detection, these events are of significance and must be treated accordingly. However rare phenomena which mainly lie outside the range of available observation are, by definition, observed infrequently, and thus it is difficult to reason about them using conventional means. Efforts in the fields of finance, weather prediction, internet traffic modelling, structural reliability and biomedical analyses have helped to discover distributions that can be used to model extreme events. However, the literature on the subject deals mostly with univariate distributions [18, 14] which are inadequate for modelling of multivariate data as typically arise in the analysis of complex systems. Most of the literature on the subject focuses on component-wise maxima [48, 49] or simultaneously extreme components of a multivariate sample [12, 10]. ‘Extreme’ is here understood as a maximum or minimum value in terms of absolute magnitude, assuming a metric over the sample space (which is independent of the distribution).

In the context of novelty detection, observations of abnormally large magnitude, if rare, are of interest, but so are observations of less extreme absolute magnitude that are equally or more improbable. In other words, all regions of the sample space over which a model of normality has low probability density values are of interest for determining the presence of extremes. Although it is not formulated explicitly as such, this is the approach adopted by Roberts in a pair of papers [35, 36], in which models of normality are represented by mixtures of Gaussian kernels. In multivariate space, the Gaussian kernel describes a hyperellipsoid whose pdf $f_n(x)$ varies along a radius $r$ according to the univariate one-sided Gaussian (scaled by a normalisation factor dependent on dimensionality $n$). That is, all regions of low probability density values are considered potentially extreme. To determine the probability density $f_n(x)$ at any point in the hyperellipsoid, the problem is reduced to a univariate case $f_1(r)$, in Mahalanobis radius $r$. The work presented in [35, 36] uses this assumption to reduce the problem of determining the EVD for a multivariate Gaussian kernel to a corresponding univariate case ($r$) and assumes that formulae derived for univariate case apply to the multivariate one. When the model of normality is composed of more than one Gaussian kernel, Roberts approximates the distribution of extrema by assuming that, at location $x$ the Gaussian kernel closest to $x$ in the Mahalanobis sense dominates the extreme value distribution,
and thus the EVD is based on the Gumbel distribution corresponding to that closest kernel. Contribution of other kernels are ignored. We hereafter refer to this technique as the ‘winner-takes-all method’. Roberts’ work is highly relevant to our investigation, because a number of the data sets used are biomedical in nature (epilepsy, anaesthesia, vigilance and MRI data), and also because, to our knowledge, it is the only attempt made to extend EVT to the multivariate case for novelty detection which understands extrema as meaning extrema in probability. However, not all the simplifying assumptions made are sound or necessary, as is shown in [8] and discussed later in this report.
3 Multivariate Approach to Extreme Value Theory for Novelty Detection

3.1 Key results of classical Extreme Value Theory

3.1.1 Classical EVT results

Univariate Extreme Value Theory is based on key results by Fisher and Tippett [16]: consider a set \( X_m = (x_1, x_2, \ldots, x_m) \) of \( m \) independent and identically distributed random variables (iid rvs), \( x_i \in \mathbb{R} \) (univariate), drawn from a distribution function \( F \), and let \( M_m = \max(X_m) \) be the largest element observed in \( X_m \). First, if \( H \), the distribution function over \( M_m \), is to be stable in the limit \( m \to +\infty \), then it must weakly converge under a positive affine transform of the form:

\[ M_m \to c_m X + d_m \quad (3.1) \]

for appropriate sequences of norming parameters \( c_m \in \mathbb{R}^+ \) (the scale parameter) and \( d_m \in \mathbb{R} \) (the location parameter). The following theorem is considered the basis of EVT and gives the possible distribution of \( X \):

**Theorem 1. (Fisher-Tippett theorem - Theorem 3.2.3 in [14], p.121)**

Let \( (X_m) \) be a sequence of iid rvs. If there exist norming constants \( d_m \in \mathbb{R} \), \( c_m > 0 \) and some non-degenerate distribution function \( H \) such that

\[ c_m^{-1}(M_m - d_m) \overset{d}{\to} H, \quad (3.2) \]
then $H$ belongs to the type of one of the following three distribution functions:

\[
\begin{align*}
\text{Gumbel,} & \quad H_1^+(y) = \exp(-\exp(-y)) \\
\text{Fréchet,} & \quad H_2^+(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ \exp(-y^{-\alpha_m}) & \text{if } y > 0 \end{cases} \\
\text{Weibull,} & \quad H_3^+(y) = \begin{cases} \exp(-(y)^{\alpha_m}) & \text{if } y \leq 0 \\ 1 & \text{if } y > 0 \end{cases}
\end{align*}
\]

where $y = c_m^{-1}(x - d_m)$ is termed the reduced variate.

A 'degenerate' distribution is a distribution whose support consist of only one value (which can be $\pm\infty$). The superscript '+' refer to the fact that these are distributions of maxima. $H_1^+, H_2^+, H_3^+$ are referred to as the maximal Gumbel distribution, the maximal Fréchet distribution and the maximal Weibull distribution, respectively. Since minima of $\{X_m\}$ are maxima of $\{-X_m\}$, the theorem can easily be rewritten for the EVDs of minima: they are the same as EVDs of maxima, with a reverse axis. The minimal Weibull distribution in particular will be of particular interest to this work:

\[
H_3^-(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp(-(x)^{\alpha_m}), & x \geq 0 \end{cases}
\]

Theorem I can be seen as being equivalent, for the distribution of maxima of random variables, to the central limit theorem for sums of random variables, with the difference that in the case of maxima of random variables, there are three limit forms. The theorem essentially defines a three-class (Fréchet, Weibull and Gumbel) equivalence relation over the set of all non-degenerate univariate distributions. Consequently the distribution of maxima from any distribution is in the maximum domain of attraction of one of the three limit forms, defined as:

**Definition 1.** (Maximum domain of attraction - Definition 3.3.1 in [14], p. 128)

We say that the random variable $X$ (the distribution function $F$ of $X$, the distribution of $X$) belongs to the maximum domain of attraction of the extreme value distribution $H$ if there exist
<table>
<thead>
<tr>
<th>EVT Type</th>
<th>Distribution</th>
<th>Probability density function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fréchet</td>
<td>Cauchy</td>
<td>$f(x) = (\pi(1 + x^2))^{-1}, x \in \mathbb{R}$</td>
</tr>
<tr>
<td></td>
<td>Pareto</td>
<td>$f(x; k; x_m) = k x_m^k x^{-k-1}, x_m, k &gt; 0, x &gt; x_m$</td>
</tr>
<tr>
<td></td>
<td>Loggamma</td>
<td>$f(x; \alpha, \beta) = \frac{\alpha^\beta}{\Gamma(\beta)} (\ln x)^{\beta-1} x^{\alpha-1}, x &gt; 1, \alpha, \beta &gt; 0$</td>
</tr>
<tr>
<td>Weibull</td>
<td>Uniform</td>
<td>$f(x; a, b) = (b - a)^{-1}, a &lt; x &lt; b$</td>
</tr>
<tr>
<td></td>
<td>Beta</td>
<td>$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} x^{a-1} (1 - x)^{b-1}, 0 &lt; x &lt; 1, a, b &gt; 0$</td>
</tr>
<tr>
<td>Gumbel</td>
<td>Exponential</td>
<td>$f(x; \lambda) = \lambda \exp(-\lambda x), \lambda, x &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Weibull</td>
<td>$f(x; \lambda) = \exp(-(-x)^\lambda), x &lt; 0, \alpha &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Gamma</td>
<td>$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x &gt; 0, \alpha, \beta &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Gaussian</td>
<td>$f(x; \mu, \sigma) = (2\pi \sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}, x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>$f(x; \mu, \sigma) = (2\pi \sigma^2 x^2)^{-1/2} e^{-(\ln x - \mu)^2/2\sigma^2}, x &gt; 0, \mu \in \mathbb{R}, \sigma &gt; 0$</td>
</tr>
</tbody>
</table>

Table 3.1: Type of extreme value distribution for the maxima of common distributions

Constants $d_m \in \mathbb{R}$ and $c_m > 0$, such that:

$$\frac{1}{c_m} (M_m - d_m) \xrightarrow{d} H. \quad (3.7)$$

We write $X \in \text{MDA}(H) \ (F \in \text{MDA}(H))$.

Table 3.1 gives the attractor of some common distributions.

Given a distribution $F$, we need to know which MDA $F$ is in. Authors [14] give various characterizations of the MDA of each EVD type and possible choices for the sequences of norming parameters $(d_m)$ and $(c_m)$. We give here the parameter sequences for the standard Gaussian distribution:

$$\forall m \in \mathbb{N}, m > 1, \ c_m = (2 \ln (m))^{-1/2} \quad (3.8)$$

$$d_m = (2 \ln (m))^{1/2} - \frac{\ln (\ln (m)) + \ln (4\pi)}{2(2 \ln (m))^{1/2}} \quad (3.9)$$
In [35, 36], the following formulae are used for the one-sided Gaussian distribution:

\[
\forall m \in \mathbb{N}, \ m > 1, \ c_m = (2 \ln (m))^{-1/2} \tag{3.10}
\]

\[
d_m = (2 \ln (m))^{1/2} - \frac{\ln (\ln (m)) + \ln (2\pi)}{2(2 \ln (m))^{1/2}} \tag{3.11}
\]

We show in [8] that eqs. 4.4 and 3.11 do not agree with maximum likelihood estimates as the dimensionality of the model is increased. A more accurate method of estimating the EVD of a multivariate Gaussian distribution is derived in section 4.

Theorem 3.3.12 in [14], which characterizes the maximum domain of attraction of the maximal Weibull distribution, is also of interest. We adapt it here to the MDA of the minimal Weibull distribution, which we will need in section 4:

**Theorem 2. (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 \( c_m^{-1}(M_m - x_F) \xrightarrow{d} H^+_1 \), where the norming constants \( c_m, d_m \) can be chosen to be \( c_m = x_F + F^{-\alpha}(-1) \) and \( d_m = x_F \).

\( M_m \) is here the minimum of \( m \) iid rvs. \( x_F \) is the left endpoint of the df \( F \), \( F^-_p \) is the \( p \)-quantile of \( F \), and \( L \) a slowly varying function at \( \infty \); i.e., a positive function which obeys

\[
\forall t > 0, \lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1. \tag{3.12}
\]

### 3.2 Redefining extrema

Classical univariate EVT (uEVT) cannot be directly applied to the estimation of multivariate EVDs. For the multivariate case, we no longer wish to answer the question “how is the sample of greatest magnitude distributed?”, which would require choosing an ad-hoc multivariate distance, but rather “how is the most improbable sample distributed?”. To this end, we use the alternate definition of extremum proposed in [6]:
Definition 2. Let $m \in \mathbb{N}^*$ and $\{X_m\}$ be a sequence of (possibly multivariate) iid rvs, drawn from a distribution $F$ with probability density function $f$. We define the extremum to be the random variable $E_m = \text{argmin}\{f(X_1), \ldots, f(X_m)\}$.

Figure 3.1: Empirical distributions of minima (left) and maxima (right) for the standard normal distribution. Histograms (top row) and cumulative histograms (middle) are represented. The case $m = 1$ gives the generative distribution. As $m$ increases, distributions of extrema move further away from 0, the distribution mean. Bottom row: same cumulative histograms (except the case $m = 1$) compared to the Gumbel distributions (grey curves) obtained using eq. 3.8 and 3.9.

Figure 3.1 shows the type of distribution that classical EVT aims to approximate; i.e., the distribution of minima and maxima of the standard univariate Gaussian distribution. Figure 3.2 shows the multivariate distributions of the type we are actually interested in modelling for the work
described in this report: the top left plot shows the distribution of extrema of the standard uni-
ivariate Gaussian distribution. The top right plot shows that in the case of a multimodal model of
normality, extrema can and do appear between the modes of the kernels. The remaining four plots
reinforce this point, while showing that definition \[2\] is straightforwardly extended to multivariate
situations.

This alternative perspective on EVT can be related to classical EVT in two ways, as is underlined
in [6]:

1. away from the distribution modes, the pdf monotonically decreases with increasing distance
to the modes. Extrema in magnitude are therefore also minima in probability density values,
2. selecting the most improbable sample with respect to \( f \) is equivalent to selecting the sample
   of minimal magnitude with respect to the df over \( f(X) \). uEVT can therefore be applied to
   samples drawn in the probability space.

However, using this definition is better-suited to problems where all improbable events are of
interest, including those that could occur between modes of \( f \).

For the remainder of our description, \( F_n \) will refer to a continuous multivariate distribution over
a sample space (or data space) \( D \) of dimensionality \( n \in \mathbb{N}^* \) (usually \( \mathbb{R}^n \)). \( P \) will be the associated
probability space (usually a subset of \( \mathbb{R}^+ \)), and \( f_n \) will be its probability density function. We also
define \( m\)-extremum to mean the sample of smallest density value out of \( m \) iid samples \( cEVT \) (for
classical EVT) will refer to the univariate approach to EVT based on magnitude, and \( mEVT \) (for
multivariate EVT) will refer to the approach to EVT which uses definition \[2\].

### 3.3 Distribution of probability density values

Let \( \mathbf{x} \) be a sample drawn from a (possibly multivariate) distribution \( F_n \) and \( f_n \) its associated
probability density function (pdf). The probability of obtaining a specific value in the probability
space (a probability density value, or pdv) by drawing a sample in the data space is entirely defined
by the form of \( f_n \). Consequently, if \( F_n \) is the distribution over the random variable \( X \), there exists
a distribution $G_n$ over the random variable $Y = f_n(X)$. $G_n$ is defined as follows:

$$\forall y \in \mathcal{P}, \quad G_n(y) = \Pr(Y \leq y)$$

$$= \Pr(f_n(X) \leq y)$$

$$= \int_{f_n^{-1}([0,y])} f_n(x) \, dx$$

(3.13)

where $f_n^{-1}([0,y])$ is the preimage\(^1\) of $[0,y]$ under $f_n$. In other words, $G_n(f_n(x))$ is the probability of drawing a random sample of smaller pdf than that of $x$. Note that $G_n(f_n(\cdot))$ takes identical values on a level set\(^2\) of $F_n$. For the purpose of our analysis, the advantage of working with $G_n$ rather than $F_n$ is that $G_n$ is univariate. Obtaining an analytical form of $G_n$ relies on our ability to parameterize the level sets of $f_n$ and integrate the resulting parameterization. This is, in general, not possible for an arbitrary $F_n$. Section 4.1 studies the useful case where $F_n$ is a multivariate Gaussian distribution.

### 3.4 Novelty scores

Our approach is based on the assumption that sufficiently unlikely observations with respect to a model characterize ‘novelty’; i.e., the more improbable an observation is, the more certain we are of observing a new behaviour of the system with respect to our knowledge of normal system behaviour. Setting a lower threshold on $G_n$, and then classifying as ‘novel’ observations that break this threshold is strictly equivalent to setting a lower threshold on $f_n$, which is similar to the conventional density-thresholding method at first sight. However, unlike the value of threshold on density, the value of a threshold set on $G_n$ has an actual probabilistic interpretation: it represents the probability mass below this threshold; i.e., the expected false-positive rate under normal circumstances. This feature is desirable, for instance, if the dimensionality of observations is variable (in the case of missing or corrupt data), and where some dimensions must thus be temporarily ignored (integrated out). In this case, models of variable dimensionalities may be used, and know-

---

\(^1\)Let $f$ be a function from $A$ to $B$. The preimage of set $C \subseteq B$ under $f$ is defined by $f^{-1}[C] = \{x \in A | f(x) \in C\}$; i.e. the set of all elements of $A$ whose image under $f$ is in $C$.

\(^2\)A level set of a real-valued function $f$ is a set where the function takes on a given constant value.
ing the value of an acceptable false-positive rate for the system makes switching dimensionalities easier.

More importantly, using our new approach classical EVT becomes applicable to the univariate function $G_n$. If we define $G^e_n(\cdot; m)$ to be the minimal extreme value distribution associated with $G_n$ and $m$ (the distribution of $m$-extrema of $G_n$), then $G^e_n(y; m)$ is the probability of drawing an $m$-extremum of lower density value. This is interpreted here as being the probability for the extremum to be normal with respect to the model. $1 - G^e_n(y; m)$, conversely is the probability of observing an $m$-extremum of higher density value; i.e., the probability of observing a more probable $m$-extremum. This is interpreted as the probability that the $m$-extremum is novel with respect to the model.

As it is desirable for novelty scores to take low values for normal data and higher values for increasingly abnormal, we define the novelty score function as follows:

$$q(y) = \phi [1 - G^e_n(y; m)],$$  \hspace{1cm} (3.14)

where $\phi$ is a monotonically increasing function with domain $[0, 1]$. In general $\phi$ will be the identity function when we want to relate observations to actual probabilities, and $\phi(.) = -\ln(1 - .)$ for representation or visualisation purposes.

The next two chapters aim at estimating $G^e_n$ when $F_n$ is, first, a multivariate Gaussian, and, second, a mixture of Gaussians.
Figure 3.2: Examples of distributions of interest. Top row: histograms of extrema for single kernel and two-kernel univariate generative distributions. Values of $m$ are (from darker to lighter) 1, 2, 5, 10, 30, 50, 100. Middle row: bivariate standard normal distribution (left) and histogram of extrema for $m = 10$ (right). Bottom row: bivariate 4-kernel mixture of Gaussians (left) and histogram of extrema for $m = 10$ (right).
4 Analytical Extreme Value Distributions for Multivariate Gaussian Distributions

In this section, let us consider the multivariate Gaussian distribution $F_n$ for dimensionality $n \in \mathbb{N}^*$, of which the pdf is:

$$f_n(x) = \frac{1}{C_n} \exp \left( -\frac{M(x)^2}{2} \right) \tag{4.1}$$

where $M(x) = ((x - \mu)^T \Sigma^{-1} (x - \mu))^{1/2}$ is the Mahalanobis distance, $C_n = (2\pi)^{n/2} |\Sigma|^{1/2}$ is the normalisation coefficient, $\mu$ is the centre, and $\Sigma$ is the covariance matrix. $\mathcal{D} = \mathbb{R}^n$ and $\mathcal{P} = f_n(\mathcal{D}) = \left[ 0, \frac{1}{C_n} \right]$.

Estimating the EVDs of multivariate Gaussians is doubly interesting: first, Gaussians are suitable models of normality for a number of problems; second, accurately estimating their EVDs is an improvement to the winner-takes-all method described in section 2.3, having previously stated that the parameter estimates used in [35, 36] do not agree with maximum likelihood estimates as the dimensionality of the sample space increase.

The ellipsoidal symmetry of $f_n$ and its relatively simple form make it possible to obtain analytical closed-form expressions of its minimal EVDs. This is done here in three steps: (i) calculating an analytical form of the distribution $G_n$ associated with $F_n$, (ii) showing that the resulting $G_n$ is for the distribution of its minima in the maximum domain of attraction of the minimal Weibull distribution, (iii) estimating the parameters of that minimal Weibull distribution.

4.1 Analytical Distribution of PDVs

To take advantage of the ellipsoidal symmetry of the problem, we rewrite $f_n$ in a Mahalanobis $n$-dimensional spherical polar coordinate system. Then, $x = (r, \theta)$ such that $\theta = (\theta_1, \ldots, \theta_{n-1})$, $r = M(x)$, $\theta_i \in \left[ -\frac{\pi}{2}, \frac{\pi}{2} \right]$ for $i \leq n - 2$ and the base angle $\theta_{n-1}$ ranges over $[0, 2\pi]$ (or $x = r$ if $n = 1$). The Jacobian of the transformation is $|J_n| = |\Sigma|^{1/2} r^{n-1} \prod_{i=0}^{n-3} (\cos \theta_i)^{n-i}$ (see, for instance,
This yields

\[ G_n(y) = \int_{f_n^{-1}(y,y)} \frac{1}{C_n} \exp \left( - \frac{M(x)^2}{2} \right) dx, \quad (4.2) \]

\[ = \int_{f_n^{-1}(0,y)} \frac{|J_n|}{C_n} \exp \left( - \frac{r^2}{2} \right) dr \theta, \quad (4.3) \]

\[ = \Omega_n \int_{M^{-1}(y)}^{+\infty} r^{n-1} \frac{(2\pi)^{n/2}}{\Gamma(n/2)} \exp \left( - \frac{r^2}{2} \right) dr, \quad (4.4) \]

\[ = \int_0^y \Omega_n |\Sigma|^{1/2} [-2 \ln (C_n y)]^{(n-2)/2} du. \quad (4.5) \]

\[ M^{-1}(y) = \sqrt{-2 \ln (C_n y)} \] is the unique Mahalanobis distance associated with the pdf \( y \). Eq. \( 4.3 \) is obtained by rewriting eq. \( 4.2 \) in the spherical polar coordinate system. Integrating out the angles yields eq. \( 4.4 \), where \( \Omega_n = \frac{2\pi^{n/2}}{\Gamma(n/2)} \) is the total solid angle subtended by the unit \( n \)-sphere (and \( \Gamma \) is the Gamma function). Eq. \( 4.5 \) is obtained after making the substitution \( u = \frac{1}{C_n} \exp \left( - \frac{r^2}{2} \right) \).

The integrand in \( 4.5 \) is the pdf of \( G_n \), which we aim to find,

\[ g_n(y) = \Omega_n |\Sigma|^{1/2} [-2 \ln (C_n y)]^{(n-2)/2} \quad (4.6) \]

If \( n = 1 \), the integration of eq. \( 4.5 \) yields \( G_1(y) = \text{erfc} \left( \sqrt{-\ln (C_1 y)} \right) \) where \( \text{erfc}(.) \) is the complementary error function. If \( n = 2 \), \( G_2(y) = C_2 y \). If \( n \geq 2 \), the integration of \( 4.5 \) is possible using a recursive integration by parts\(^4\), which yields two cases:

\[ G_{2p}(y) = \sum_{k=0}^{p-1} A_{2p}^k (-2 \ln (C_{2p} y))^{p-k-1} \quad (4.7) \]

\[ G_{2p+1}(y) = \sum_{k=0}^{p-1} A_{2p+1}^k (-2 \ln (C_{2p+1} y))^{p-k-1/2} + \text{erfc} \left( \sqrt{-\ln (C_{2p+1} y)} \right) \quad (4.8) \]

for all \( p \in \mathbb{N^*} \), where \( A_{2p}^k = \Omega_{2p} |\Sigma|^{1/2} \frac{2^k (p-1)!}{(p-k)!} \) and \( A_{2p+1}^k = \Omega_{2p+1} |\Sigma|^{1/2} \frac{(2p-1)! (p-k)!}{2^k (p-1)! (2p-2k)!} \).

\(^3\)Details of the derivation can be found in \[26\]

\(^4\)Again, see \[26\]
$G_n$ and $g_n$ are plotted for $n = 1$ to 5 in figure 4.1 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 [3] (ex.1.4, p.29).

4.2 Maximum domain of attraction of the Weibull distribution

From eq. 3.12, 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 $yG_{2p}(1/y)$ is a sum of slowly varying functions, which is itself slowly varying. Theorem 2 can therefore be applied to $G_{2p}$. A similar process may 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 $n$.

4.3 Parameter estimation

If $G_n$ is in the MDA of $H_3^-$, theorem 2 gives possible choices for the minimal Weibull parameters:

\[
c_m = G_n^- (1/m), \quad d_m = 0, \quad \alpha_m = 1.
\]

The scale parameter can be easily estimated numerically to arbitrary accuracy, as $G_n$ is a strictly increasing function over a finite support. Figure 4.2 shows that the formula for the scale parameter $c_m$ is a very close approximation to maximum likelihood estimates. However, the value of the shape parameter $\alpha_m$, 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 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 [4]. 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
Figure 4.1: All graphs except that in bottom right: comparison between simulated histograms and eq. 4.6 (black curves). The dimensionality varies from 1 to 5. The normalised histograms are obtained by computing the probabilities of $10^6$ random samples drawn from the $n$-dimensional standard Gaussian distribution. The vertical gray lines indicate the greatest value of the multivariate pdf. Bottom right: 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.
that, in this case, \( g_n(y) \sim sK y^{s-1} \); i.e., \( s = y^{G_n(y)} \). We therefore propose to evaluate \( s \) at the value of the scale parameter \( c_m \); i.e., define:

\[
\alpha_m = c_m \frac{g_n(c_m)}{G_n(c_m)} \tag{4.12}
\]

Figure 4.2 shows that eq. 4.12 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:

\[
G_n^*(y; m) = 1 - \exp (- (y/c_m)^{\alpha_m}) \tag{4.13}
\]

where \( c_m \) and \( \alpha_m \) are given by eq. 4.9 and 4.12 respectively. Figure ?? shows the good agreement between eq. 4.13 and simulated cumulative histograms.
4.4 Novelty scores

Returning to consideration of the data space, the quantity

\[ F_n^e(x) = 1 - G_n^e(f_n(x); m), \]

\[ = \exp \left( - \left( \frac{1}{C_n \sigma_m} e^{-\frac{M(x)^2}{2}} \right)^\alpha_m \right), \]

is the probability of drawing an \( m \)-extremum whose pdv is higher than that of \( x \); i.e., in the case of a multivariate Gaussian, the probability of drawing an \( m \)-extremum closer to the centre of the distribution. If \( x \) is itself an \( m \)-extremum, we interpret \( F_n^e(x) \) as the probability of \( x \) being novel, as is discussed in section 3.4.

The novelty function over the data space (which we also term \( q \)) is therefore:

\[ q(x) = \phi \left[ 1 - G_n^e(f_n(x); m) \right]. \]

where \( \phi \) is a monotonically increasing function with domain \( [0, 1] \), as described in section 3.4.

Eqs. 4.13 and 4.15 allow us to describe extreme value distributions of multivariate Gaussian distribution both in the probability and the sample space with satisfactory accuracy as demonstrated in figure 4.3. At this point, no simplifying assumption has been made, contrary to what was done in [36, 36]. We argue that using eq. 4.3 instead of the Gumbel distribution with parameters given by eqs. and 3.11, which have been shown to be inaccurate, is an improvement to the winner-takes-all method. The other major assumption made by the winner-takes-all method is that the closest kernel of the mixture of Gaussians to a given location has a dominant contribution to the EVDs, and therefore that the contribution of other kernels can be ignored. Again, this is shown to be only true if the overlap between kernel is minimal [6]. The next chapter presents a method to estimate the EVDs of mixture of Gaussians, for which kernel overlap is important.
Figure 4.3: Comparison between eq. 4.13 (black curves) and simulated cumulative log-
arithmic histograms (crosses) for dimensionality 1 to 4. $F_n$ is in each case the standard
multivariate Gaussian distribution. $10^5$ m-extrema are used to generate the histograms.
5 Extreme Value Distributions for Mixture of Gaussians

In this section, let us consider the $k$-component Gaussian mixture model (GMM) $F_n$ for dimensionality $n \in \mathbb{N}^*$, of which the pdf is:

$$f_n(x) = \sum_{j=1}^{k} \pi_j \frac{1}{C_{n,j}} \exp \left( -\frac{M_j(x)^2}{2} \right)$$  \hspace{1cm} (5.1)

where, for $j = 1$ to $k$, $M_j(x) = ((x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j))^{1/2}$ is the Mahalanobis distance, $C_{n,j} = (2\pi)^{n/2} |\Sigma_j|^{1/2}$ is the normalisation coefficient, $\mu_j$ is the centre, $\Sigma_j$ is the covariance matrix and $\pi_j$ is the mixing weight of the $j^{th}$ component.

5.1 The $\Psi$-transform method

To estimate the distribution of $m$-extrema of a GMM, [6] proposes to use the so-called $\Psi$-transform to map the distribution of $f_n(x)$ back into a radial space into which a Gumbel distribution can be fitted, via maximum likelihood estimation, to a set of extrema.

For a standard Gaussian, $f_n(x) = (2\pi)^{-n/2} \exp (-r^2/2)$, and so $r = (-2 \ln f_n(x) - n \ln 2\pi)^{1/2}$. The $\Psi$-transform is defined as:

$$\Psi [f_n(x)] = \begin{cases} 
(-2 \ln f_n(x) - n \ln 2\pi)^{1/2} & \text{if } f_n(x) < K \\
0 & \text{if } f_n(x) \geq K
\end{cases}$$  \hspace{1cm} (5.2)

where $K = (2\pi)^{-n/2}$. If $f_n$ were a single Gaussian distribution $N(\mu, \Sigma)$, the $\Psi$-transform would map the $f_n(x)$ values back onto $r$, the radii of $x$ from $\mu$. The $\Psi$-transform maps the distribution of $f_n(x)$ values back into a space into which a Gumbel distribution can be fitted.

The left plot in Figure 5.1 shows a normalised histogram of $N = 10^6$ extrema generated from a mixture of three bivariate Gaussian kernels, for $m = 100$. The distribution is highly skewed towards 0, as is expected for extrema. The right plot in the figure shows the $\Psi$-transform of the $m$-extrema, which simulations show is distributed according to the Gumbel distribution.
Figure 5.1: Normalised histogram of $f_2(x)$ values for $N = 10^6$ extrema generated from trimodal GMM with $m = 100$ (left plot). Histogram of the $\Psi$-transformed $f_2(x)$ values shown in grey, with the corresponding MLE Gumbel distribution fitted in $\Psi$-space, shown in black (right plot). The vertical dashed line represents a 99% novelty threshold: 99% of the probability mass lies left of this threshold and data falling right of it will be considered novel.

While the $\Psi$-transform is theoretically sound, there is no actual need to transform the probability space into a $\Psi$-space other than for the clarity of representation. The types of EVDs are closely linked from a mathematical point of view. Indeed, one can immediately verify the following properties ([14, 41]):

$$X \in \text{MDA} \left( H_2^+ \right) \text{ with shape parameter } \alpha$$

$$\iff \ln(X)^{\alpha} \in \text{MDA} \left( H_1^+ \right)$$

$$\iff -X^{-1} \in \text{MDA} \left( H_3^+ \right) \text{ with shape parameter } \alpha$$

$$\iff X^{-1} \in \text{MDA} \left( H_3^- \right) \text{ with shape parameter } \alpha$$

etc . . . (5.3)

As a consequence, maximum likelihood estimation can be performed directly on $m$-extrema in the probability space to fit a minimal Weibull distribution. This means fitting a minimal Weibull distribution to the distribution in the left plot in figure 5.1 rather than using the $\Psi$-transform and then fitting a maximal Gumbel distribution.

All MLE methods require that $m$-extrema be sampled, which can be computationally costly as
the cost of sampling a vector increase with the dimensionality of the sample space. Moreover, an EVD fitted to a set of $m$-extrema is only valid for this particular value of $m$, meaning that testing for a range of values of $m$ requires drawing extrema for all values in range, which makes MLE methods even more computationally intensive. In the remainder of this section, we propose a novel semi-numerical method to estimate the EVDs of a GMM, which follows in the footsteps of the method adopted to estimate the EVDs of multivariate Gaussian distribution and described section 4. We show that under reasonable assumptions, we can obtain accurate estimates of EVDs for almost any value of $m$ with no need for sampling extrema, significantly reducing the complexity of the estimation process.

### 5.2 Histogram fitting via least-squares optimisation

In the general case, $f_n$ does not exhibit any remarkable symmetry and no closed-form expression of $G_n$ may be found. However, we can make the assumption that sufficiently far away from the modes of the distribution, a mixture of Gaussian kernels behaves approximately like a single Gaussian kernel. This assumption is useful because we are interested in the probability mass which lies in the tail of $F_n$ (where $f_n(x)$ is close to zero), not near its modes. This probability mass also lies in the tail of $G_n$, for which we wish to find the EVD.

Thus, for $y$ sufficiently close to zero, the $g_n$ of a GMM can be approximated by the $g_n$ of a single Gaussian kernel; i.e., by a function of the form:

$$k_n(y; \beta) = \Omega_n \beta \left[-2 \ln \left((2\pi)^{n/2} \beta y\right)\right]^{n-2/2}$$

for some positive value of $\beta$. Eq. 5.4 is the same as eq. 4.6 where $|\Sigma|^{1/2}$ has been replaced by the parameter $\beta$, hence the requirement that $\beta$ be positive. It is noteworthy that $k_n$ is only dependent on the dimensionality $n$ and the scalar parameter $\beta$. In other words, the form of $k_n$ does not depend on the precise form of the covariance matrix but only on its determinant, which is a scalar.
The expression of the cdf $K_n$ associated with $k_n$ is adapted from eq. 4.8:

$$K_{2p}(y; \beta) = y \sum_{k=0}^{p-1} B_{2p}^k (-2 \ln((2\pi)^p \beta y))^{(p-k-1)}$$

$$K_{2p+1}(y; \beta) = y \sum_{k=0}^{p-1} B_{2p+1}^k (-2 \ln ((2\pi)^{p+1/2} \beta y))^{p-k-1/2} + \text{erfc} \left( \sqrt{-\ln((2\pi)^{p+1/2} \beta y)} \right)$$

for all $p \in \mathbb{N}^*$, where $B_{2p}^k = \Omega_{2p} \beta^{2k(p-1)!}$ and $B_{2p+1}^k = \Omega_{2p+1} \beta^{(2p-1)!(p-k)!/(2p-2k)!}$.

The family of parametric distributions $k_n$ may be used to estimate $g_n$. To estimate the value of $\beta$ that best approximates the tail of our $g_n$, we propose to estimate $g_n$ using a simple histogram, and then find the value of $\beta$ that minimises the least-square error between $k_n$ and the histogram $g_n$ near zero. A full description of the algorithm is given in table 5.1. Note that because the probability mass is shifted towards zero in the probability space as $n$ increases, the LSE minimization gives more weight to the left-hand end of the histogram. This is seen as a desirable effect, as we are interested in modelling the left tail (i.e., where $f_n(x) \to 0$) as accurately as possible.

From 4.2, $k_n$ is known to be in the domain of attraction of the minimal Weibull EVD:

$$H^{-3}(y; d_m, c_m, \alpha_m, ) = 1 - \exp \left[ - \left( \frac{y - d_m}{c_m} \right)^{\alpha_m} \right],$$

therefore theorem 2 applies, and the location, scale, and shape parameters are given by:

$$c_m = K_n^{-1} \left( \frac{1}{m} \right), \quad d_m = 0, \quad \alpha_m = m \ c_m \ K_n[c_m],$$

where $K_n$ is the integral of $k_n$ and $K_n^{-1} \left( \frac{1}{m} \right)$ is the $1/m$ quantile of $K_n$.

After estimation of $\beta$, we can use eq. 5.8 to define the EVD of our $G_n$. 

32
Algorithm 1:
Let $F_n$ be a Gaussian mixture model, whose EVD we aim at estimating. Parameters of the algorithm are: $N \in \mathbb{N}^*$, the number of samples drawn from $F_n$, $B \in \mathbb{N}^*$, the number of bins in the histogram, and $B' \in \mathbb{N}^*$, $B' \leq B$, the number of bins considered to be in the tail of the histogram.

- Choose sensible values for parameters $N$, $B$ and $B'$. $B$ and $N$ should be chosen so that a big enough number of samples fall each of the bins of the left tail. $B'$ must be chosen such that $\frac{B'}{B \max(F_n)}$ is less than the minimal pdv in the convex hull of the modes of $F_n$, to ensure that the target function is monotonically decreasing over $\left]0, \frac{B'}{B \max(F_n)}\right]$.
- Draw samples $x_1, \ldots, x_N$ from the distribution $F_n$.
- Compute their pdvs: $f(x_1), \ldots, f(x_N)$.
- Construct a histogram $(h_1, \ldots, h_B)$ of these pdvs using $B$ equally spaced bins:
  \[
y_1 = \frac{1}{B \max(F_n)}, \quad y_1 = \frac{2}{B \max(F_n)} \ldots, y_B = \max(F_n).
\]
- Fit the tail of histogram by solving the following least square error minimization problem:
  \[
  \beta_{\min} = \arg\min_{\beta} \left( \sum_{b=1}^{B'} (h_j - k_n(y_j; \beta))^2 \right).
\]
- For any value of $m \in \mathbb{N}^*$, use eq. 5.8 and the quantiles of $K_n(\cdot; \beta_{\min})$ to compute the values of the Weibull parameters.
- The resulting minimal Weibull distribution is an approximation of the EVD of $K_n(\cdot; \beta_{\min})$.

Table 5.1: Description of the proposed algorithm to estimate the EVDs of a Gaussian mixture model.
5.3 Validation on simulated data

To validate our approach, we compare EVDs obtained using equations (4.9), (4.10), and (4.12) with the EVDs obtained using maximum likelihood estimates of the Weibull parameters, using simulated data.

For dimensionality $n = 1$ to 6, we define $F_n$ to be the $n$-dimensional mixture of Gaussians comprised of two multivariate standard Gaussian distributions with equal priors and a Euclidean distance between their centres equal to two.

In order to estimate the EVD using MLE, for each dimensionality $n = 2 \ldots 6$, and for increasing values of $m$, a large number of extrema (e.g., $N = 10^6$) must be sampled. Figure 5.2 shows estimates obtained using MLE for both the scale $c_m$ and shape $\alpha_m$ parameters of the EVD.

The scale parameter appears to be accurately estimated even for small values of $m$. However, the proposed method’s use of eq. (4.12) to estimate the shape parameter only matches the MLE estimate for $m > 15$. This was expected, as the Fisher-Tippett theorem tells us that the Weibull distribution is the EVD for asymptotically increasing $m$, and that actual EVDs are not expected not to match the Weibull distribution closely for small values of $m$.

Figure 5.3 presents a comparison between the cdfs of the corresponding distributions estimated using MLE and the method proposed in this section, for $n = 4$ and a range of values of $m$. Taking into account the logarithmic scale in $y$, we conclude that solutions obtained using the new method are a good match to the maximum likelihood estimates.

The main advantage of our approach is that it does not require sampling of extrema, which is a particularly intensive process. Assuming a model $F_n$, we only need to obtain $N$ samples from that model to build a histogram approximating $G_n$, and then we solve a simple least-squares estimation problem, before finally applying the closed-form eqs. 5.8 to obtain an estimate of the Weibull parameters for any value of $m$. On the other hand, the MLE (which in itself is more intensive than the least-square estimation problem) requires $m \times N$ samples to be drawn to obtain $N$ extrema, and this is for a single value of $m$. To test all values of $m$ between 1 and 100, for instance, our
Figure 5.2: Comparison of results of MLE estimates of the scale parameter $c_m$ (top) and the shape parameter $\alpha_m$ (bottom) parameter (shown as points in the plots), and values obtained using eq. 5.8 for $n = 2$ to 6 and increasing values of $m$ (shown as continuous lines). For each dimensionality $n$, the GMM $F_n$ is composed of two standard Gaussian kernels with equal priors, with a Euclidean distance between their centres equal to two. Error bars are too small to be visible at this scale.

algorithm requires up to 5,000 times less sampling, and none of the 100 iterations of the MLE algorithm.
Figure 5.3: Similar figure to fig. ?? showing the comparison between our original method and simulated cumulative logarithmic histograms (crosses) for dimensionality 1 to 4. $F_n$ is in each case the normalised sum of two standard multivariate Gaussian kernels with equal mixing weights and Euclidean distance between their centres equal to 4. $10^5$ $m$-extrema are used to generate the histograms.
6 EVT for Novelty Detection in Time Series

6.1 Scheme

In [28] and [7], we use the following algorithm to perform novelty detection in time-series data for which a single underlying generative process is assumed, and where the iid assumption holds:

- Choose an appropriate value for the parameter $m$.
- At a time $t$, consider the window of data containing the current sample and the $m-1$ samples preceding (provided that samples are obtained regularly), and compute their probability densities with respect to the assumed underlying generative model.
- Select the sample of minimum density.
- Compute a novelty score for that sample using one of the previously described methods. This score is the novelty score at time $t$.

Note that this means that a high-scoring sample can result in the novelty taking a high value for up to $m$ timesteps; i.e., as long as the extreme sample remains in the sliding window. Note also that this algorithm relies on the assumption that every set of $m$ consecutive samples has the same distribution as the generative process. This is reasonable under the i.i.d assumption and becomes moreso as the value of $m$ is increased. The next section shows experimental results obtained on artificial data.

6.2 Proof of principle

In this section, we aim to show the validity of an EVT-based approach to novelty detection in time-series data. A simulated time-series is generated that comprises 50 cycles. Each cycle is divided in two phases: a 1800-step ‘normal’ phase and an 200-step ‘event’ phase. For each phase, samples are generated using two different Gaussian distributions. All 50 ‘normal’ phases use the same Gaussian distribution, $N(\mu_n, \sigma_n)$. All 50 ‘event’ phases use the same Gaussian distribution,
$N(\mu_e, \sigma_e)$. The EVT parameters of the ‘normal distribution’ are computed and used together with
the scheme described in the previous section for various values of $m$. The classical thresholding
method described in section 2.2 is also applied to obtain novelty scores for comparison with the
EVT-based methods. Thresholds on all series of novelty scores are then varied to obtain the
specificities$^5$ and sensitivities$^6$ (sample-wise) needed to plot the ROC curves. We also define the
number of alerts at a given threshold as the number of times the threshold is crossed from a lower
value to a higher value. The number of alerts per cycle shown as a function of the specificity
is shown for all simulations. Figures 6.1 and 6.2 show results for two univariate simulations, one
where only the variance of the ‘event distribution’ is different from that of the ‘normal’ distribution,
and one where only the mean is shifted. Figure 6.3 shows results obtained for three multivariate
simulations.

\[
^5 \text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}
\]

\[
^6 \text{specificity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}
\]
Figure 6.1: Results of our novelty detection algorithm obtained when applied to artificial data. The top graph shows two cycles of randomly generated data. Values of the parameters are as follows: $\mu_n = 0$, $\sigma_n = 1$, $\mu_e = 0$, $\sigma_e = 2$. The bottom left graph shows the ROC curves for the thresholding method and different values of the parameter $m$ for the EVT-based method. The bottom right graph shows the number of 'alarms' as a function of the sensitivity.
Figure 6.2: The same as 6.1 with $\mu_n = 0$, $\sigma_n = 1$, $\mu_e = 2$, $\sigma_e = 1$. 
Figure 6.3: Results for the multivariate (bivariate) EVT approach. Each row corresponds to a different simulation. In all three simulations, $\mu_n$ is the origin and $\sigma_n$ is the identity matrix. Top row: $\mu_e$ is the origin and $\sigma_e$ is 4 times the identity matrix. Middle row: $\mu_e = (2, 2)$ and $\sigma_e$ is the identity matrix. Bottom row: $\mu_e = (0, 0)$ and $\sigma_e$ is $(1,0;0,4)$. 
7 Application to a Vital-Sign Data Set

7.1 The data set

The data set used comes from the first phase of three-phase trial conducted at the Presbyterian Hospital, University of Pittsburgh Medical Centre (UPMC, Pennsylvania, USA), between November, 2006 and August, 2007 [24]. Four parameters were recorded for each patient: heart rate (HR), breathing rate (BR), blood oxygen saturation (SpO\textsubscript{2}) and blood pressure (SDA, for systolic/diastolic average). The sampling rate was of the order of 30 seconds, except for the blood pressure, which was recorded approximately every 30 minutes during the day, and every hour at night to minimise patient discomfort. The data set contains 332 patients and has a total duration of approximately 18,692 hours.

Events satisfying the criteria of a Medical Emergency Team (Table 7.1) call were identified. They are termed Condition C’ or C events (‘C’ for ‘crisis’). Those events deemed to be due to measurement artefacts were discarded. A merging scheme was then applied to prevent multiple detection of overlapping events. Merged events were then reviewed by clinicians in order to verify that those which had been removed were indeed artefactual, and whether the vital sign data that caused the remaining alerts were non-artefactual. Condition C events which were generated as a result of non-artefactual vital signs, as validated by clinicians, were termed C’ events.

<table>
<thead>
<tr>
<th>Respiratory</th>
<th>- rate &lt; 8 or &gt; 36 breaths per minute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- new onset difficulty breathing</td>
</tr>
<tr>
<td></td>
<td>- new oxygen saturation reading less than 85% for more than 5 minutes (unless patient is known to have chronic hypoxaemia)</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>- &lt; 40 or ≥ 140 bpm with symptoms</td>
</tr>
<tr>
<td></td>
<td>- any rate &gt; 160 bpm</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>- &lt; 80 or &gt; 200 mmHg systolic</td>
</tr>
<tr>
<td></td>
<td>- &gt; 110 mmHg diastolic with symptoms</td>
</tr>
<tr>
<td>Acute neurological change</td>
<td>- acute loss of consciousness</td>
</tr>
<tr>
<td></td>
<td>- new onset lethargy or narcan use without immediate response</td>
</tr>
<tr>
<td></td>
<td>- seizure (outside of seizure monitoring unit)</td>
</tr>
<tr>
<td></td>
<td>- sudden loss of movement (or weakness) of face, arm or leg</td>
</tr>
<tr>
<td>Other</td>
<td>chest pain unresponsive to nitroglycerine or doctor unavailable colour change (of patient or extremity): pale, dusky, gray or blue unexplained agitation for more than 10 minutes suicide attempt uncontrolled bleeding</td>
</tr>
</tbody>
</table>

Table 7.1: Clinical criteria for activating a MET response at the UPMC Presbyterian Hospital [13]. Narcan is a tradename of the drug naloxone that is used to counter the effects of opioid overdose.
7.2 Method

Three methods are compared here:

- the traditional heuristic pdf thresholding method;
• the Ψ-transform method, which estimate EVDs using MLE on sampled extrema;

• our EVT-based analytical method (section 4), which does not require sampling of extrema.

We split the dataset into four subsets:

• training, consisting of data from 144 ‘normal’ patients (containing approximately 8000 hours of data);

• control, consisting of data from 144 ‘normal’ patients (approximately 8000 hours of data as well);

• test consisting of ‘normal’ data from the 44 patients who went on to have crisis events (approximately 2000 hours);

• crisis, those ‘abnormal’ data labelled by clinical experts (approximately 43 hours).

Vital-sign measurements of heart rate, respiration rate, and blood oxygen saturation were subsampled every second for all patients. A model of normality was constructed using the training set, where a single trivariate Gaussian kernel was used (with full covariance), noting that the ‘normal’ data were approximately distributed according to a Gaussian distribution in each of the $n = 3$ dimensions for this case study.

For the Ψ-transform and analytical EVT-based methods, we define a novelty score $q = -\ln[1 - P_e(x)]$. For the traditional heuristic method of defining a novelty threshold on the pdf $p(x)$, we define a novelty score $q = -\ln p(x)$. The proportion of data $\tau$ classified ‘abnormal’ from each of the four classes described above were determined for a range of thresholds on $q$ using each method, the results of which are shown in figure 7.1. The ideal classifier would result in all “abnormal” data (those from the event class) lying above the novelty threshold, and all ‘normal’ data (those from all other classes) lying below the novelty threshold.

In order to obtain a correct ‘abnormal’ classification of 90% for the crisis data, the conventional thresholding method misclassifies the test, control, and training data with rates 14.9%, 7.8%, and 5.9%, respectively. For the same 90% successful classification rate of crisis data, the misclassification rates decrease to 5.8%, 2.4%, and 2.0% when using the analytical EVT-based method, and
to 7.0%, 2.6%, and 1.8% when using the $\Psi$-transform method.

For varying $q$, while both EVT-methods provide an improvement in misclassification rates compared to the traditional method, it can be seen that the proposed analytical method performs at least as successfully as the $\Psi$-transform solution, which is optimal in the MLE sense.
8 Discussion and Future Work

8.1 Exact Extreme Value Distributions

Let $X$ be a random variable with pdf $f(x)$ and cdf $F(x)$. Let $(x_1, x_2, \ldots, x_m)$ be an iid sample drawn from $F(x)$. Then, their joint pdf is (see, for instance, [4]):

$$f(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} f(x_i), \quad (8.1)$$

and their cdf is

$$F(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} F(x_i). \quad (8.2)$$

The cdf of the maximum sample is therefore:

$$F_{\text{max}}(x) = \Pr(\text{max}(X_1, \ldots, X_m) \leq x) = \Pr(\text{all } X_i \leq x) = \prod_{i=1}^{m} \Pr(X_i \leq x) = \prod_{i=1}^{m} F(x) = [F(x)]^m. \quad (8.3)$$

The associated pdf is:

$$f_{\text{max}}(x) = mf(x) [F(x)]^{m-1}. \quad (8.4)$$

Similarly, the cdf and pdf of the minimum sample are:

$$F_{\text{min}}(x) = 1 - [1 - F(x)]^m \quad (8.5)$$

$$f_{\text{min}}(x) = mf(x) [1 - F(x)]^{m-1} \quad (8.6)$$

These are exact formulae for the minimal and maximal extreme value distribution. In the limit $m \to \infty$, these distributions are degenerate and the Fisher-Tippett theorem provides an efficient
asymptotic representation by mapping them to known standard distributions. However, for engineering purposes, and especially if we are interested in small values of the parameter \( m \), using exact distributions is more accurate, provided that the calculations involved do not exceed machine precision.

In figure 8.1, we compute the exact EVD associated with the distribution of densities of the standard normal distribution, and shows a comparison between the asymptotic distributions (Weibull) used so far in this report and the exact distributions defined in eqs. 8.5 and 8.6. It is clear that the exact EVDs fit the histograms better than the Weibull distributions, and hence more accurately describe the extrema of the generative distribution.

Figure 8.2 is similar, where we use the bivariate standard Gaussian distribution instead of the univariate one.

### 8.2 Limitations of current approaches

**EVT cannot detect all classes of novelty that a system can statistically manifest.**

EVT-based novelty scores have so far been interpreted as our belief that our set of observations comes from the assumed generative distribution. In reality, this test tells us very little about the \( m-1 \) samples that are not the extremum, because EVT tests only the extremum in a window of \( m \) samples: if the extremum is deemed to be an outlier, the other \( m-1 \) samples cannot be considered abnormal data, as they can all have high probability densities. nor can they be considered normal, as all observed densities could be close to that of the extremum.

This means that EVT is perfectly appropriate to detect an outlier, understood to most an improbable sample, but it is less well suited for the detection of a regime change in system behaviour. In section 6.2, we show examples where the ‘event’ data come from a distribution with either a shift in mean or an increased variance and show that we can identify such changes in regime. However a regime with identical mean (or slightly shifted) and a smaller variance than the distribution of normal data will be missed as all samples, including the extremum, become more likely to fall within the range of values taken by normal data.
Figure 8.1: Normalised histograms of $10^5$ extrema, exact EVDs (red) and asymptotic EVDs (green) of the standard normal distribution for $m = 2, 3, 5, 10, 20$ and $30$ (from upper left to lower right). The exact EVDs fit the histograms closely for all values of $m$, whereas the asymptotic EVDs are inaccurate for the smallest values of $m$. 
Figure 8.2: Normalised histograms of $10^5$ extrema, exact EVDs (red) and asymptotic EVDs (green) of the standard bivariate normal distribution for $m = 2, 3, 4, 10, 20$ and 30 (from upper left to lower right).
The iid assumption is problematic when dealing with time-series. All our approaches so far rely on the assumption that if normal, the data in the sliding window as similar in distribution to the normal data. We show here an example where this cannot be considered to be true using the UPMC phase I data. Considering two channels (HR, BR), the population is divided into three groups (training, control and test) as in [28]. Results are shown in figure 8.3. For each group, we build a bivariate estimate of the probability density function, which allows us to build a univariate distribution of densities (top left and black histogram in the top right graph). Now assuming that the data are iid for the training group, we compute the extreme value distribution corresponding to this distribution of densities. Separately, we select the most extreme sample in each sliding window for the training group, with respect to the distribution of densities for the training group. The lower left graph shows that this is dissimilar to that obtained under the iid assumption. The lower right graph shows the same distribution for the event data. This gives us an insight as to why we are able to obtain good results despite the iid assumption not holding. The event data are so extreme that we are able to detect it, but this would not be the case if the problem was harder, and where extreme data were ‘less extreme’. Thus, we require a method that could overcome these disadvantages.

8.3 Future work

8.3.1 Order statistics for novelty detection

One of the limitations of EVT is that it only gives information on the extremum of a statistical sample, and discards all the samples that are not the extremum. While we show that considering this extremum is an improvement on the pdf-thresholding method, we also note, for instance, that isolated outliers influence novelty scores up to the length of the window. If the data-generating process is noisy, this can result in an amplification of the noise, as the most noisy observation of a window will always be the one used to assign novelty scores. In our future work, we propose to address such issues by using a broader theory than EVT, that of order statistics, of which we give here a brief overview.
Figure 8.3: Empirical distributions for the UPMC phase I data set. Top left: empirical cdfs of densities for the training, control and test group. If the distribution in one group was a bivariate gaussian distribution, the corresponding curve would be a straight line. Top right: histogram of densities in the training group (black). If the distribution in the group was a bivariate Gaussian distribution, the histogram would be flat. In red, we show the simulated EVD of minima obtained by drawing extrema from the histogram considered as a discrete distribution. Bottom left: data-driven EVD obtained using the algorithm shown in §6.2. Note that it does not look at all like the red histogram in the top right graph. Bottom right: same as bottom left for the event data. The peaks observed at $y=2.1$ are the consequence of replacing missing data by the mean of the channel in the training set (as is done in [19] to cope with missing data).
Let \((X_1, X_2, \ldots, X_m)\) be a sample of size \(m\) drawn from a pdf \(f(x)\). Arrange \((X_1, X_2, \ldots, X_m)\) in an increasing order of magnitude and let \((X_{1:m}, X_{2:m}, \ldots, X_{m:m})\) be the ordered values. The \(k\)th element in this sequence, \(X_{k:m}\) is called the \(k\)th order statistic in the sample; i.e., the \(k\)th order statistic of a statistical sample is defined as its \(k\)th smallest value. The first and last order statistics, \(X_{1:m}\) and \(X_{m:m}\) are the minimum and maximum of \((X_1, X_2, \ldots, X_m)\) respectively, which means that classical extreme value theory is simply a branch of order statistic theory. In particular, the \(m\)-extremum as defined in section 3.2 is simply the first-order statistic of the distribution over probability density values.

As with EVT, it is possible (either analytically for simple processes or in a data-driven fashion if the data generating process is unknown) to find distributions for all order statistics of a statistical sample. In the context of time-series analysis, a statistical sample can be replaced by the content of a sliding window. Handling of multivariate data involves the sorting of a window of multivariate samples according to their probability density values, assuming an underlying model.

### 8.3.2 Patient-specific modelling

The current Visensia approach is not patient-specific, as it uses a population of high-risk patient to build its model of normality, and assumes all measurements to be independent. Additionally, it can be shown that data from one patient occupy only a fraction of the volume occupied by a multi-patient training set in the vital sign space. If ‘normal’ data can be collected from a patient for a sufficient amount of time (for instance, 24 hours after admission provided that no adverse event occur during the interval), a model of normality specific to this patient can be computed using the existing or an EVT-based methodology and used for the rest of their stay. Such an adjustment would be expected to yield a more accurate assessment of the patient’s condition. Preliminary work in this direction has been done and showed significant improvement so long as an stable ‘normal’ period is observed at the beginning of a patient’s recordings and can be used to construct a model. Of course, this is not always the case as we expect those patients who are going to be discharged to get better over the length of their stay. Dynamic models and on-line
learning are directions of research that we wish to explore in future work.

8.3.3 Dynamic Analysis
References


# List of lectures attended

<table>
<thead>
<tr>
<th>Lecture Title</th>
<th>Department/Session</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to CETL</td>
<td>DTC (Research skills session)</td>
<td>15/10/08</td>
</tr>
<tr>
<td>Settling into a research Environment</td>
<td>DTC (Research skills session)</td>
<td>23/01/09</td>
</tr>
<tr>
<td>Producing a Scientific Poster</td>
<td>DTC (Research skills session)</td>
<td>03/03/09</td>
</tr>
<tr>
<td>Managing your DPhil</td>
<td>Dept. of Engineering Science</td>
<td>20/10/08</td>
</tr>
<tr>
<td>Demonstrating Training Course</td>
<td>Dept. of Engineering Science</td>
<td>15/10/08</td>
</tr>
<tr>
<td>Tutorial Teaching Course (Part 1)</td>
<td>Dept. of Engineering Science</td>
<td>05/11/08</td>
</tr>
<tr>
<td>Scientific Computing for DPhil Students</td>
<td>Comlab</td>
<td>Michaelmas 2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12 lectures)</td>
</tr>
<tr>
<td>Applied Statistics</td>
<td>Dept. of Statistics (BS1b)</td>
<td>Trinity 2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8 lectures)</td>
</tr>
</tbody>
</table>