Novelty Detection with Extreme Value Theory in Vital-Sign Monitoring



Samuel Hugueny
Balliol College

Supervised by
Professor Lionel Tarassenko
Doctor David Clifton
Submitted: Trinity Term, 2013

This DPhil Thesis is submitted to the
Department of Engineering Science, University of Oxford,
in fulfillment of the requirements for the degree of Doctor of Philosophy.

Novelty Detection with Extreme Value Theory in Vital-Sign Monitoring

Samuel Hugueny
Balliol College

Doctor of Philosophy
Trinity Term, 2013

ABSTRACT

Every year in the UK, tens of thousands of hospital patients suffer adverse events, such as unplanned transfers to Intensive Therapy Units or unexpected cardiac arrests. Studies have shown that in a large majority of cases, significant physiological abnormalities can be observed within the 24-hour period preceding such events. Such warning signs may go unnoticed, if they occur between observations by the nursing staff, or are simply not identified as such. Timely detection of these warning signs and appropriate escalation schemes have been shown to improve both patient outcomes and the use of hospital resources, most notably by reducing patients' length of stay.

Automated real-time early-warning systems appear to be cost-efficient answers to the need for continuous vital-sign monitoring. Traditionally, a limitation of such systems has been their sensitivity to noisy and artefactual measurements, resulting in false-alert rates that made them unusable in practice, or earned them the mistrust of clinical staff. Tarassenko et al. (2005) and Hann (2008) proposed a novelty detection approach to the problem of continuous vital-sign monitoring, which, in a clinical trial, was shown to yield clinically acceptable false alert rates. In this approach, an observation is compared to a data fusion model, and its "normality" assessed by comparing a chosen statistic to a pre-set threshold. The method, while informed by large amounts of training data, has a number of heuristic aspects.

This thesis proposes a principled approach to multivariate novelty detection in stochastic time-series, where novelty scores have a probabilistic interpretation, and are explicitly linked to the starting assumptions made. Our approach stems from the observation that novelty detection using complex multivariate, multimodal generative models is generally an ill-defined problem when attempted in the data space. In situations where "novel" is equivalent to "improbable with respect to a probability distribution ", formulating the problem in a univariate probability space allows us to use classical results of univariate statistical theory. Specifically, we propose a multivariate extension to extreme value theory and, more generally, order statistics, suitable for performing novelty detection in time-series generated from a multivariate, possibly multimodal model.

All the methods introduced in this thesis are applied to a vital-sign monitoring problem and compared to the existing method of choice. We show that it is possible to outperform the existing method while retaining a probabilistic interpretation.

In addition to their application to novelty detection for vital-sign monitoring, contributions in this thesis to existing extreme value theory and order statistics are also valid in the broader context of data-modelling, and may be useful for analysing data from other complex systems.

Acknowledgments

While writing a DPhil thesis has felt, at times, like a solitary pursuit, this thesis would not have come into its full existence without the help and support of many friends, colleagues, and family members.

I am indebted, first, to the Life Sciences Interface Doctoral Training Centre at Oxford (or its current incarnation), and through it, to the EPSRC, for offering me a place on a stimulating and enjoyable graduate programme, as well as sponsoring me for a number of years. Special thanks go to Prof. David Gavaghan, Prof. Elspeth Garman, Dr. James Wakefield, Ms. Samantha Miles, and Ms. Maureen York.

Thanks are due to the members of the BSP research group, old and new, including, but not restricted to Dr. Susannah Fleming, Dr. Mark Larsen, Dr. Christina Orphanidou, Dr. Oliver Gibson, Dr. Lei Clifton, Dr. Tim Bonnici (the only "real" doctor in this list), Mr. Marco Pimentel, Dr. Yasmina Bohrani, Ms. Sarah Khalid, Dr. Ahmar Shah, Dr. Busi Vilakazi, and, of course, Ms. Val Keston-Hole. Special thanks should go to Mr. Louis Mayaud, for his inspiring work ethic and all around enthusiasm about all matters, and Dr. Dave Wong, for indulging me with a wide range of esoteric conversation topics, from Gaussian processes to the latest controversy in the world of professional cycling.
Thanks also to Dr. Roger Mark, Dr. Leo Celi, Dr. Pete Watkinson, and Dr. Duncan Young, who all, at one point or another, supervised my work.

Many thanks to my examiners, Prof. Stephen Roberts and Prof. Ian Nabney. The overall quality of this thesis would have been much less without their thorough reviewing and pertinent comments.

Finally, I am profoundly grateful to my two supervisors, Dr. David Clifton and Prof. Lionel Tarassenko. From the first day of what was, at the time, a rotation project, David has shown admirable passion for his work, a real talent for explaining difficult concepts in the clearest possible terms, and infinite patience when I still did not understand those. It has been a privilege working alongside him, and I shall forever admire his ability to transform ideas into peer-reviewed journal papers. And most of all, many thanks to Prof. Tarassenko who has been an exemplary supervisor and an inspiration to me. It has been an honour to study under such an influential figure in our world, and I am especially grateful to him for encouraging me to pursue my own research interests, while being a limitless source of guidance, optimism, and wisdom.

To you all, collaborators, friends, family, and supervisors, my most sincere *remerciements*.

Related Publications

Extending the Generalised Pareto Distribution for Novelty Detection in High-Dimensional Spaces
Clifton, D.A., Hugueny, S., Clifton, L., and Tarassenko, L.:
Journal of Signal Processing Systems [In press]

An Extreme Function Theory for Novelty Detection
Clifton, D.A., Clifton, L., Hugueny, S., Wong, D., and Tarassenko, L.:
IEEE Journal of Selected Topics on Signal Processing 7(1), 2013.

Probabilistic Patient Monitoring with Multivariate, Multimodal Extreme Value Theory
Hugueny, S., Clifton, D.A., and Tarassenko, L.:
Communications in Computer Science 127, 2011.

Novelty Detection with Multivariate Extreme Value Statistics
Clifton, D.A., Hugueny, S., and Tarassenko, L.:
Journal of Signal Processing Systems 65, 2011.

Pinning the Tail on the Distribution: A Multivariate Extension to the Generalised Pareto Distribution
Clifton, D.A., Hugueny, S., and Tarassenko, L.:
IEEE Machine Learning and Signal Processing, Beijing, China, 2011.

Understanding Vital-Sign Abnormalities in Critical Care Patients
Hugueny, S., Clifton, D.A., and Tarassenko, L.:
Critical Care 14(1), 2010.

Probabilistic Patient Monitoring Using Extreme Value Theory
Hugueny, S., Clifton, D.A., and Tarassenko, L.:
Biomedical Engineering Systems and Technologies, Valencia, Spain, 2010.

A Comparison of Approaches to Multivariate Extreme Value Theory for Novelty Detection
Clifton, D.A., Hugueny, S., and Tarassenko, L.:
IEEE Statistical Signal Processing, Cardiff, UK, 2009.

Novelty Detection with Multivariate Extreme Value Theory, Part II: An Analytical Approach to Unimodal Estimation
Hugueny, S., Clifton, D.A., and Tarassenko, L.:
IEEE Machine Learning in Signal Processing, Grenoble, France, 2009.

Novelty Detection with Multivariate Extreme Value Theory,Part I: A Numerical Approach to Multimodal Estimation
Clifton, D.A., Hugueny, S., and Tarassenko, L.:
IEEE Machine Learning in Signal Processing, Grenoble, France, 2009.

Contents

List of Figures **9**

List of Tables **11**

List of abbreviations and symbols **13**

1 Introduction **15**
1.1 The clinical need: in-hospital continuous vital-sign monitoring . . . 15
1.2 State of the art in patient monitoring: the Visensia project . . . 17
1.3 Novelty detection . . . 19
1.4 Extreme Value Theory . . . 22
1.5 Thesis outline . . . 24

2 UPMC data set **26**
2.1 The UPMC trial . . . 26
2.2 Extraction of clinical events and labelling . . . 30
2.3 Summary of UPMC Phase 1 data . . . 38
2.4 Summary of C' and C" events . . . 43
2.5 Problems with existing filtering . . . 50

3 Identifying and labelling events **56**
3.1 Regenerating MET events . . . 56
3.2 Label selection . . . 74
3.3 Summary of labels . . . 75
3.4 Comparison with historical labelling . . . 82
3.5 Future use of labels . . . 84

4 Univariate EVT for Novelty Detection **85**
4.1 Motivation . . . 86
4.2 Key results of classical Extreme Value Theory . . . 88
4.3 Extension to the multimodal case . . . 102
4.4 Novelty detection with distributions of probability densities . . . 106
4.5 Example: Gaussian distributions . . . 109
4.6 Novelty Detection with time-series . . . 113
4.7 Conclusions . . . 116

5 Multivariate EVT for Novelty Detection **118**

5.1 From univariate to multivariate data-space . . . 118
5.2 Existing approaches to multivariate EVT . . . 121
5.3 Empirical cdf interpolation . . . 128
5.4 DPD of the multivariate Gaussian . . . 131
5.5 Exact EVDs of the multivariate Gaussian . . . 137
5.6 Asymptotic EVDs of the multivariate Gaussian . . . 140
5.7 EVT for mixtures of Gaussians . . . 146
5.8 Conclusions . . . 154

6 Application of Multivariate EVT to the UPMC Data Set 159
6.1 Overview of this chapter . . . 159
6.2 Assessing performance . . . 160
6.3 Univariate Novelty Detection on the UPMC Phase 1 data set . . . 166
6.4 Trivariate HR/BR/SpO_2 model . . . 184

7 Novelty Detection with Order Statistics 200
7.1 Motivation . . . 200
7.2 Distributions of order statistics . . . 202
7.3 Order Statistics in multivariate space . . . 205
7.4 Application to the UPMC data set . . . 209

8 Conclusions 216
8.1 Summary of results . . . 216
8.2 Future work . . . 218

Bibliography 226

List of Figures

2.1 Labelling procedures for the UPMC Phase 1 data set. 31
2.2 Number of available vital-sign channels in the UPMC1 Phase 1 data 42
2.3 Univariate histograms of UPMC Phase 1 data. 44
2.4 C' and C" event durations. 45
2.5 Fractions of C" events. 45
2.6 Number of C" and C' events per patient. 46
2.7 Vital-sign channel completeness 47
2.8 Frequencies of C' and C" events 48
2.9 High breathing rate data. 52
2.10 Histograms of SpO_2 channel during hypoxaemic events. 54
2.11 Low SpO_2 data. 55

3.1 Legend for vital-sign plots. 64
3.2 Example of labelling inconsistency: UPMC1 Patient 284 (1) 65
3.3 Example of labelling inconsistency: UPMC1 Patient 235 66
3.4 Example of labelling inconsistency: UPMC1 Patient 284 (2) 68
3.5 Aggregate scores, labels and numbers of true MET events per patient. 72
3.6 Aggregate scores and labels on a patient-basis. 73
3.7 Distribution of tMET events across tMET patients. 76
3.8 Durations of tMET and aMET events. 77
3.9 Channel completeness for tMET and aMET events. 78
3.10 Acceptance rates of MET events by patient in the MET population. 80
3.11 Artefactual MET events for UPMC patient 284 81
3.12 True MET event for UPMC patient 59 82

4.1 Empirical distributions of minima and maxima 87
4.2 P-P plots of cdf vs EVD percentiles. 87
4.3 Exact distributions of minima for the univariate Gaussian 91
4.4 Exact distributions of maxima for the standard exponential distribution 92
4.5 Exact EVDs vs. Maximal Gumbel distributions. 101
4.6 Maximal Gumbel parameters for the standard Gaussian 102
4.7 Examples of distributions of extremes in probability 103
4.8 Natural vs. log-novelty scores 108
4.9 Pdfs, DPDs and novelty scores for univariate Gaussians 111
4.10 Distribution of extremes in data space for the standard Gaussian. 113
4.11 Novelty detection under variance shift. 115

5.1 Illustration of the novelty detection problem for multivariate time-series. 119
5.2 Empirical distributions of bivariate, multimodal extremes 120

5.3 Illustration of the Winner-Takes-All method. 125
5.4 Illustration of the Ψ-transform method. 127
5.5 Cdf estimation 130
5.6 DPDs of multivariate Gaussian distributions (pdfs). 135
5.7 DPDs of multivariate Gaussian distributions (cdfs). 136
5.8 EVDs in probability space for multivariate Gaussians. 138
5.9 Cross-sections of EVDs 139
5.10 Bivariate EVD in data space. 140
5.11 Weibull parameters for the multivariate Gaussian. 144
5.12 Comparison of asymptotic vs. exact EVDs for the multivariate Gaussian. 146
5.13 Example of multimodal mixture of Gaussians. 150
5.14 Tail-fitting under the AGD assumption 152
5.15 Comparison of 4 multivariate EVT methods for mixtures of Gaussians. 157

6.1 Five-fold cross validation results for the number of kernels in the HR model. . . 168
6.2 Empirical distribution of HR from two GMMs. 170
6.3 DPD of the univariate model of HR. 173
6.4 EVDs of G_1 and log-novelty in probability space. 173
6.5 Novelty and log-novelty scores in HR space. 175
6.6 UPMC Patient 1: Example of Novelty Detection on HR trend data 178
6.7 UPMC Patient 32: Example of Novelty Detection on HR trend data 179
6.8 Results of novelty detection in HR time-series 181
6.9 Empirical bivariate distributions of vital-sign observations 185
6.10 Five-fold cross validation results for the number of kernels in the trivariate model.186
6.11 Trivariate model marginal distributions. 187
6.12 DPD of the trivariate model 188
6.13 EVDs of G_2 and log-novelty in probability space. 190
6.14 Example of multivariate novelty detection in UPMC Phase 1 Patient 32 192
6.15 ROC curves for the procedure using multivariate EVT 193
6.16 Example of false alerts in UPMC Phase 1 Patients 271 & 288 196

7.1 Order Statistic Distributions: examples. 204
7.2 Novelty functions based on density-ordered statistics for the standard normal distribution. 207
7.3 Novelty maps based on density-ordered statistic for the standard two-dimensional Gaussian distribution. 208
7.4 ROC curves of the novelty detection procedure the 20% order statistics. 211
7.5 ROC curves of the novelty detection procedure with order statistics using the median. 212
7.6 ROC curves of the novelty detection procedure using 80% order statistics . . . 213

List of Tables

1.1 Modified Early Warning Score (MEWS) 16

2.1 Vital signs: units, methods and terminology. 28
2.2 Vital-sign limits for artefact rejection and single-channel MET alert limits. . . . 34
2.3 Breakdown of C, C', and C" events by primary cause. 34
2.4 Number of patients with C' and C" events broken down by primary cause. . . . 35
2.5 UPMC Phase 1 data statistics. 40
2.6 Frequencies of C and C" events. 49
2.7 Exceedances of existing physiological thresholds. 50

3.1 MET event primary causes. 62
3.2 Summary of reviewers' labelling by primary activation cause. 69
3.3 Consensus rates between combinations of reviewers. 70
3.4 Summary of reviewers' labels on an event- and patient-basis 71
3.5 Primary activation causes of tMET events using the absolute majority vote. . . 76
3.6 MET event durations and frequencies. 79
3.7 Overlap rates between types of events. 83

4.1 Type of EVDs for the maxima of common distributions. 98

5.1 Summary of multivariate EVT methods. 155

6.1 Summary of functions and notations for the HR model, f_1 172
6.2 AUCs of novelty detection in HR time-series 183
6.3 AUCs of novelty detection in HR/BR/SpO_2 time-series 195
6.4 AUCs of novelty detection in HR/BR/SpO_2 time-series with median filtering . . 198

7.1 AUCs of novelty detection in HR/BR/SpO_2 time-series using Order Statistics . 214

List of Algorithms

1.1 Vital-sign abnormality detection procedure used by Hann (2008) 18
3.1 Identification of MET events. 62
4.1 Univariate Novelty Detection with EVT 107
4.2 EVT-based Novelty Detection in time-series 114
5.1 Tail-fitting via least-square error minimisation for mixture of Gaussians. 151
5.2 Tail-fitting via probability-mass matching. 154
6.1 Novelty Detection procedure in HR time-series. 177
7.1 Tow-sided univariate novelty detection with order statistics 201
7.2 Novelty Detection procedure in HR/BR/SpO_2 time-series with density-ordered statistics 210

List of abbreviations and symbols

Abbreviations

AGD	asymptotic Gaussianity in density.
AUC	area under the (ROC) curve.
BR	breathing rate.
cdf	cumulative distribution function.
df	distribution function.
DPD	distribution of probability density.
EVD	extreme value distribution.
EVT	extreme value theory.
GMM	Gaussian mixture model.
GPD	generalized Pareto distribution
HR	heart rate.
i.i.d.	independent and identically distributed.
ITU	intensive therapy unit.
LSE	least-square error.
m-extreme	the observation of minimal density in a sample of size m.
MDA	maximum domain of attraction.
MET	Medical Emergency Team.
mEVT	multivariate extreme value theory.
MEWS	modified early warning score.
MG	mixture of Gaussians.
MLE	maximum likelihood estimation.
PSCA	persistent single-channel abnormality; i.e. a period of time during which a single-channel vital-sign is consistently exceeded.
ROC	receiving operating characteristic.
rv(s)	random variable(s).
pdf	probability density function.
pdv	probability density value.
PSI	patient status index.
SDA	systolic diastolic average, a measure of blood pressure.
SDU	Step-Down Unit.
SpO_2	saturation of peripheral oxygen, a measure of blood oxygen saturation.
UPMC	University of Pittsburgh Medical Center (Pittsburgh, PA, USA), or the data set resulting from this trial.

Symbols

α	Level of significance in an instance of hypothesis testing.
α_m	Weibull or Fréchet shape parameter.
c_m	scale parameter of extreme value distributions.
d	dimensionality of the sample (data) space.
d_m	location parameter of extreme value distributions.
e_m	instantiation of E_m
E_m	the extreme random variables of a set of m observations.
f	pdf of F.
F	distribution function and distribution of a random variable.
$F^{\leftarrow}(p)$	p-percentile of F.
$F^{-}(.;m)$	distribution of maxima of F.
$F^{+}(.;m)$	distribution of maxima of F.
$g(.;f)$	pdf of $G(.;f)$, the DPD of f
$G(.;f)$	DPD of f
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.
m	number of samples from which an extremum is drawn; this is the primary parameter of EVT.
M_m^-	minimum of a set of random samples $X_1, X_2 \ldots, X_m$.
M_m^+	maximum of a set of random samples $X_1, X_2 \ldots, X_m$.
$\mathbb{N}^*$	$\mathbb{N} \setminus \{0\}$; i.e, the set of non-zero natural integers.
$\mathcal{P}_{\text{MET}}$	all 150 MET patients
$\mathcal{P}_{\text{tMET}}$	all 97 tMET patients
$\mathcal{P}_{\varnothing}$	all 170 patients who did not have any events
$\mathbb{R}^*$	$\mathbb{R} \setminus \{0\}$; i.e, the set of non-zero real numbers.
$\mathbb{R}_+^*$	$\{x \in \mathbb{R} / x > 0\}$; i.e, the set of non-zero positive real numbers.
$\xrightarrow{d}$	$A_m \xrightarrow{d} A$: convergence in distribution.
$\overset{t}{\sim}$	tail-equivalence: $F \overset{t}{\sim} G$ means that the distributions F and G are tail equivalent.

Chapter 1

Introduction

1.1 The clinical need: in-hospital continuous vital-sign monitoring

Every year in the UK, at least 20,000 hospital patients have unplanned transfers to an Intensive Therapy Unit, and at least 23,000 have cardiac arrests or die unexpectedly (McQuillan et al. 1998, Findlay et al. 2012). 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 (Hodgetts et al. 2002, McGloin et al. 1999). Although studies (Schein et al. 1990, Smith & Wood 1998, Findlay et al. 2012) 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 ITU-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 1.1).

	3	2	1	0	1	2	3
SBP (mmHg)	< 70	71-80	81-100	101-199		≥ 200	
HR (bpm)		< 40	41-50	51-100	101-110	111-129	≥ 130
BR (rpm)		< 9		9-14	15-20	21-29	≥ 30
Temp. (^{o}C)		≤ 35		35-38.4		≥ 38.5	
AVPU				**A**lert	Reacting to **V**oice	Reacting to **P**ain	**U**nresponsive

Table 1.1: 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 total score is obtained by adding the individual component scores.

The effect of METs on patient outcome and length of stay has generally been deemed positive by observational studies (Hourihan et al. 1995, Bellomo et al. 2003), 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 (Hillman et al. 2005). 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 (NICE Short Clinical Guidelines Technical Team 2006).

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

From the evidence available in the literature, 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. (2005) as an answer to 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 one of novelty detection. 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).

Thus, the PSI is 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 $\hat{p}(\mathbf{x})$, where $\mathbf{x}$ is the vector of normalised vital-sign parameters. The pdf is modelled using an N-kernel Parzen Windows estimator (Parzen 1962), where typically $N = 400$. The PSI is defined to quantify departures from normality, understood to be

Algorithm 1.1 Summary of the abnormality detection procedure applied to newly-acquired time-series, used by Hann (2008), which we term the *pdf-thresholding method*

- Let $\hat{p}$ be the model of normality; i.e., the unconditional probability density function of normal training data, y a threshold in PSI-space, and $\{\mathbf{x}_t\}$ a newly-acquired vital-sign time-series, starting at time $t = 0$.
- As we loop through the time-series, we construct three functions A, B, and C.
- C is a binary function, such that $C(t) = 0$ if $\mathrm{PSI}(\mathbf{x}_t) \leq y$, and $C(t) = 1$ if $\mathrm{PSI}(\mathbf{x}_t) > y$.
- A and B are such that $A(t)$ and $B(t)$ represent the fraction of the last 5 minutes during which $C(\mathbf{x}_t) = 1$, and the fraction of the last 3 minutes during which $C(\mathbf{x}_t) = 0$, with $A(0) = B(0) = 0$.
- Finally, let S be the output of the algorithm, a function which takes value 0 if in an alerting state, and 1 otherwise, with $S(0) = 0$.
- At each time-step t in the time-series:
  1. Compute $\mathrm{PSI}(\mathbf{x}_t)$.
  2. Compute $C(t)$, then $A(t)$ and $B(t)$.
  3. If $A(t) > 0.8$, set $S(t)$ to 1 (start an alert).
  4. If $B(t) > \frac{2}{3}$, set $S(t)$ to 0 (end an alert).
  5. If neither of the above conditions are met, then S is unchanged, and $S(t)$ is assigned the value of S at the previous timestep.

observations with low values of $\hat{p}(\mathbf{x})$:

$$\mathrm{PSI}(\mathbf{x}) = -\ln[\hat{p}(\mathbf{x})] \tag{1.1}$$

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. A summary of the detection procedure is given in Algorithm 1.1.

Using a model constructed using over 3,500 hours of data, Tarassenko et al. (2006) report 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 (Hillman et al. 2005)) and shows that the Visensia system is capable of detecting critical events earlier than single-channel alerting systems. A later clinical study enabled Hann (2008) to define a 'gold standard' of crit-

ical events. An improved model of normality was retrained with a much larger training set and validated using Receiving-Operating-Characteristic (ROC) analysis against this gold standard. The best performing model is reported to achieve 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 MERIT study, both on an event basis). A more detailed review of the work by Hann (2008) is given in Chapter 2.

1.3 Novelty detection

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.

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, and a high degree of integrity. As system complexity increases, so does the number of possible states and the influence of parameters on observable data becomes harder to describe. An inevitable consequence is the emergence of a large number of possible "abnormal" modes, which makes conventional multi-class classification approaches unsuitable for these applications.

Novelty detection offers alternative approaches, in which a description of normality is learnt by constructing a model of observable data (essentially data observed under normal system behaviour) in the d-dimensional data space $\mathcal{D}$, where d is the size of the data vectors. A newly-acquired vector $\mathbf{x}$ can then be tested by comparing it to the model of normality, often resulting in some manner of novelty measure, or novelty score, which we note $z(\mathbf{x})$. Scores are then typically compared to a decision threshold κ to determine whether the test data are indicative of a departure from normal system behaviour. The hypersurface $z(\mathbf{x}) = \kappa$ defines a decision boundary in $\mathcal{D}$.

Novelty detection has been subjected to much research in application domains involving large datasets acquired from critical systems. These include the detection of mass-like structures in medical diagnostic problems (Quinn & Williams 2007), faults and failures detection in complex industrial systems (Tarassenko et al. 2009), structural damage (Surace & Worden 2010), intrusions in electronic security systems (Jyothsna et al. 2011, Patcha & Park 2007), video surveillance (Diehl & Hampshire 2002, Markou & Singh 2006), mobile robotics (Sofman et al. 2011, Vieira Neto & Nehmzow 2007), sensor networks (Zhang et al. 2010), astronomy catalogues (Dutta et al. 2007, Escalante 2005), and text mining (Basu et al. 2004). A variety of methods have been proposed to perform novelty detection, which we divide into three broad categories

1.3.1 Distance-based novelty detection

The simplest and most intuitive novelty detection method is to define $z(\mathbf{x})$ in terms of a distance from $\mathbf{x}$ to *prototypes* of normal data ξ_k, $k = 1, \ldots, K$. These prototypes can be all or a subset of the training data, or statistics of the training data, such as their mean. They can also

be determined using clustering methods. Examples of such distance-based methods include the Nearest Neighbour, or k-Nearest Neighbour algorithms adapted for novelty detection (Duda et al. 2012, Knorr & Ng 1998, Hautamaki et al. 2004). A test vector is labelled an outlier if it is located far from the nearest prototypes, according to a chosen distance. Popular choices of distances include the Euclidean distance and the Mahalanobis distance, whose parameters can be estimated using the normal training data. These choices of distance may however be problematic where the variables are heterogeneous in physical dimensions and units. Furthermore, basing the novelty boundary on a distance from normal prototypes implies that the resultant novelty boundary is piecewise-hyperspherical[1], which is unlikely to be an accurate representation of the true novelty boundary.

1.3.2 Domain-based novelty detection

Domain-based novelty detection methods aim at constructing the novelty boundary directly. Novelty of unknown data is then assigned by its location with respect to the boundary. Such approaches do not attempt to describe the density of the normal data, and are unconcerned with the exact normal density structure near the modes of the distribution, which therefore need not be known. Approaches using one-class Support Vector Machines (SVMs) have been proposed. Tax & Duin (1999) for instance, define the novelty boundary as being a hypersphere in a space transformed using the kernel trick which contains all of the training data. Schölkopf et al. (2000) define the novelty boundary to be that which separates the training data from the origin of the transformed space with maximum margin.

1.3.3 Probabilistic novelty detection

A probabilistic approach to novelty detection assumes that normal training data are generated from some underlying probability distribution F, and then uses the training data to estimate F. The estimate of F is thus a model of normality. We may then set a novelty threshold using F, such that it has probabilistic interpretation, or perform a statistical test which gives

[1] or hyperellipsoidal when using a Mahalanobis distance.

us some insight on whether new data is likely to have been generated by F. In univariate dataspace, choosing the decision boundary a number of standard deviations away from the distribution's mean is an easy choice, which maps to percentiles, if the distribution is symmetric. Alternatively, a level set of F may be chosen as the decision boundary, defined by $\kappa = f(\mathbf{x})$ (where f is the pdf of F). The novelty threshold on must be set such that a large fraction of the probability mass lies within the bounds of the novelty boundary. Areas of the data space where the density of $f(\mathbf{x}) < \kappa$ may then be classified as novel.

This is the approach used in the Visensia system (Tarassenko et al. 2006, Hann 2008): it relates the idea of vital-sign abnormality to the probability of observing data which do not belong to the distribution characterising stable patient data. Fundamental to this approach is the assumption that normal data are generated from an underlying data distribution. 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).

1.4 Extreme Value Theory

The pdf-thresholding 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 the sample size increases; i.e, as more observations are drawn from F. Extreme Value Theory (EVT) tells us where we should expect the largest (or smallest) of m samples drawn from F to lie, and thus allows us to set our novelty thresholds accordingly. EVT may therefore be better suited for dynamic situations such as continuous vital-sign monitoring (or time-series analysis in general), in which it is assumed that observations are repeatedly generated from a model distribution. 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 classical data analysis tasks, extremes are labelled 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 observations are, by definition, observed infrequently, and thus it is difficult to reason about them using conventional means. Efforts in the fields of finance (Embrechts et al. 1997, Longin & Solnik 2001, Poon et al. 2004), weather prediction (Cooley et al. 2007), internet traffic modelling (Choe & Shroff 1998), structural reliability (Hart et al. 1985) and biomedical analyses have helped to identify distributions that can be used to model extreme events. However, the literature on the subject deals mostly with univariate distributions (Gumbel 1935, Embrechts et al. 1997) which are inadequate for modelling multivariate data as typically arise in the analysis of complex systems. Most of the literature on multivariate EVT focuses on component-wise maxima (Tawn 1988, Tawn 1990) or simultaneously extreme components of a multivariate sample (De Haan & Resnick 1993, De Haan & de Ronde 1998). "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 (Roberts 1999, Roberts 2000), 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 the 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 (Roberts 1999, Roberts 2000) uses this assumption to reduce the problem of determining the EVD for a multivariate Gaussian kernel to a corresponding univariate case (where the only variable to be considered is r) and assumes that formulae derived for the 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 $\mathbf{x}$ the Gaussian kernel closest to $\mathbf{x}$ in the Mahalanobis sense dominates the extreme value distribution, and thus the EVD is based on the Gumbel distribution corresponding to that closest kernel. The contributions 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 extremes as meaning extremes in probability. However, not all the simplifying assumptions made are necessary, as is shown in (Clifton et al. 2009*b*) and discussed later in Chapter 5.

1.5 Thesis outline

In Chapter 2, we present a description of the UPMC data set used by Hann (2008), which we also use throughout this thesis. The procedures for acquiring the data and identifying the clinically relevant events are detailed. Inconsistencies in the resulting event labels are highlighted, which led us to design a new labelling procedure, described in Chapter 3.

Chapter 4 is concerned with setting a framework for novelty detection based on classical univariate Extreme Value Theory. We show that classical results used for modelling extremes in magnitude can be used to model extremes in probability, including in cases where the data-generating model is a multimodal probability distribution. In Chapter 5, existing approaches for performing novelty detection in multivariate space are reviewed, and a natural extension to the univariate framework described in the previous chapter is introduced.

A first case study using only the heart rate channel of the UPMC data set is described in Chapter 6. A second case study, using three vital-sign channels and our original multivariate framework, is presented in Chapter 7. From these case studies, we conclude that pure Extreme Value Theory methods suffer from a robustness issue, which makes them inadequate for practical use. A novelty detection method based on more general Order Statistics is introduced in Chapter 7.

Chapter 8 regroups our conclusions and directions for future work.

Chapter 2

UPMC data set

2.1 The UPMC trial

2.1.1 Trial phases

The main data set used in this thesis is the result of a three-phase clinical trial conducted in a 24-bed surgical trauma step-down unit (SDU) at the University of Pittsburgh Medical Center Presbyterian Hospital (UPMC, Pittsburgh, PA, USA) between November 2006 and August 2007. Patients involved in the trial were all adults (at least 18 years old) in need of a monitored bed. The SDU provides a level of care lower than that provided in the Intensive Care Unit. Tablet PCs were attached to the existing bedside monitors (Hewlett-Packard "Viridia 24"). Very little information is given by the manufacturer regarding the processing of the vital-sign signals produced by the monitor and we will assume that these signals are unprocessed.

The trial consisted of three phases:

- Phase 1: Data collection over an 8-week period. Existing bedside monitors were used according to the standard of care in the ward. No additional early-warning system was used. Calling criteria (Table 2.2) were used to activate the MET in the event of patient instability. The total number of patients included in Phase 1 was 332 patients. A detailed summary of their data is given in Section 2.3.

- Phase 2: In this 16-week phase, the Patient Status Index (PSI) yielded by Algorithm 1.1 was displayed on dedicated monitors installed at the bedside and used by nurses during the review of patients. Data collection continued during this period. However, as the caregivers were unblinded to the PSI, the resulting data were deemed inappropriate to use for setting a threshold on this PSI.

- Phase 3: During this 8-week phase, the nurses continued to activate the MET according to the existing standard of care, but also used the novelty alert (displayed on the tablet by the bedside and generated using the threshold determined in Phase 1) to activate the MET. Data from Phase 3 cannot be used to train data fusion algorithms, because they result from a standard of practice that already includes the use of a vital-sign-based early warning system, therefore are not suitable to measure the impact of introducing another such system.

The Phase 1 data, acquired under the normal level of care, without additional intervention or use of an early-warning system, were used "to determine the optimal threshold for detecting clinically significant critical events warranting MET activation".[1] The optimal threshold refers to the threshold on the Patient Status Index in the pdf-thresholding method described in Section 1.2. Phase 1 data were retrospectively analysed and labelled for artefactual events and clinically relevant events, as described in Section 2.2.1.

2.1.2 Monitored vital signs

In all three phases, the following vital signs were monitored:

Heart rate (HR)

The heart rate is the number of contractions of the heart in one minute, measured in beats per minute (bpm). It was derived from an electrocardiogram (ECG) but can alternatively be derived from photoplethysmography, using a pulse oximeter. An abnormally rapid heart is known as

[1] Section 7.1.1, page 122 in (Hann 2008).

Vital sign	Heart rate	Breathing rate	Oxygen saturation	Systolic blood pressure	Diastolic blood pressure
Abbreviation	HR	BR	SpO_2	SBP	DPB
Unit	beats per minute (bpm)	respirations per minute (rpm)	%	millimetres of mercury (mmHg)	
High value state	tachycardia	tachypnea	N/A	systolic hypertension	diastolic hypertension
Low value state	bradycardia	bradypnea	hypoxaemia	systolic hypotension	diastolic hypotension
Sampling type	"continuous"	"continuous"	"continuous"	intermittent	intermittent
UPMC meas. method	ECG	impedance pneumography	pulse oximetry	sphygmomanometry (cuff) and oscillometric methods	

Table 2.1: Vital signs: units, methods and terminology. "Continuous" refers to a sampling period of the order of a few seconds. "Intermittent" refers here to a longer sampling period, typically 1 to 4 hours.

tachycardia, while an abnormally slow heart rate is referred to as *bradycardia*.

Breathing rate (BR)

The breathing rate, or respiratory rate, is defined as the number of breathing cycles in one minute, measured in respirations per minute (rpm). In the UPMC trial, it was measured through electrical impedance pneumography (Folke et al. 2003). States of abnormally rapid or low breathing rates are called *tachypnea* and *bradypnea*, respectively.

Blood pressure (BP)

The blood pressure is the pressure exerted by the circulating blood upon the walls of blood vessels. When the ventricles of the heart contract (systole), there is a peak in blood pressure (systolic blood pressure, SBP). On relaxing (diastole), there is a minimum in blood pressure (diastolic blood pressure, DBP). Blood pressure measurements were made non-invasively using an inflatable cuff, or sphygmomanometer, placed around the upper arm. The cuff is inflated until the blood flow through the brachial artery is completely occluded. The air is then released and the oscillations in cuff pressure caused by the in-rushing blood are measured. A sphygmomanometer is designed to measure the Mean Arterial Pressure (MAP), defined as the average

arterial pressure during a single cardiac cycle. SBP and DBP are derived using oscillometric detection and empirical formulae (Geddes et al. 1982, Webster 1998). As a consequence, measurements of all the blood pressure parameters are always synchronous. In the existing analysis and labelling of the UPMC Phase 1 data, the Systolic-Diastolic Average (SDA), defined as the arithmetic mean of SBP and DPB, was substituted for MAP and used as the sole blood pressure parameter[2]. An elevated blood pressure is termed *hypertension*, with a low blood pressure being referred to as *hypotension*.

Peripheral oxygen saturation (SpO_2)

Oxygen saturation of the blood is defined as the ratio of oxyhaemoglobin, the molecule resulting from the combination of the haemoglobin molecule and oxygen, to the total concentration of haemoglobin present in the blood. Under normal physiological conditions, arterial blood is 97% saturated, whilst venous blood is 75% saturated. Peripheral oxygen saturation is an estimate of the arterial oxygen saturation level and was measured non-invasively using pulse oximetry.

A pulse oximeter is a sensor that is placed on a translucent part of the patient's body (usually a fingertip or earlobe). Light at two different wavelengths (usually red, 660 nm, and infrared, 940 nm) is passed through the patient tissues to a photodetector. Transmitted light varies as a function of oxygen saturation because of the difference in colours between oxyhaemoglobin and haemoglobin, and therefore in the absorption spectra of the two molecules. It is assumed that the attenuation of light can be split into three independent components: that due to arterial blood, that due to venous blood, and that due to tissues. It is further assumed that the increase in attenuation during the cardiac cycle is caused only by the inflow of arterial blood into the fingertip or earlobe. Since the inflow of arterial blood is synchronous with the heart rate, the attenuation due to arterial blood can be calculated by subtracting the DC component of the attenuation from the total attenuation. If the patient is breathing room air, their SpO_2 is considered normal if it is between 95% and 100%. If they receive supplemental oxygen (such as oxygen

[2]This was performed upon recommendation of a clinician to give equal weight to both systolic and diastolic blood pressures. This is not *a priori* justifiable, as there is no reason for systolic and diastolic blood pressure to be given equal weight in a data fusion algorithm which takes both into account as two contributing vital signs, albeit correlated ones.

therapy through the use of an oxygen mask), SpO_2 routinely reaches 100%. Abnormally low oxygen saturation is referred to as *hypoxaemia*.

An additional vital sign, temperature, was initially recorded, but was eventually dropped from further analyses because of the prevalence of artefacts in the data and the overall poor quality of the recorded signals due to the difficulty of measuring temperature continuously in a non-invasive manner. Units, methods of measurement and other terminology are summarised in Table 2.1.

2.2 Extraction of clinical events and labelling

A total of 27,000 hours of data were acquired during UPMC Phase 1, from which a number of abnormal events were extracted. Two analyses were then performed using the resulting labelling. The first, by Hann (2008) adopted an engineering approach, and aimed to determine automatically periods of physiological abnormality, whose clinical relevance was then to be assessed by clinical experts. These clinically validated events were then used to calibrate an automated early-warning system. The second analysis, by the UPMC medical team (Hravnak et al. 2008), adopted a clinical approach, and is summarised in Section 2.2.7. Sections 2.2.2 to 2.2.6 summarise the steps of Hann's analysis, and do not constitute original work, yet their presence in this thesis is essential as they form the basis from which novel work is developed. The complete labelling process, including pre-processing, event extraction and event rejection, is detailed in the following sections and illustrated in Figure 2.1.

2.2.1 Definition of clinical events

Hann (2008) defines a clinically significant critical event as "an event which triggers activation of the MET at the UPMC". At the UPMC, MET calls are termed "Condition A" if the patient is in cardiac or respiratory arrest requiring cardiopulmonary resuscitation, and "Condition C" if the patient is not in cardiac arrest, but their vital signs fulfill the MET calling criteria in Table 2.2. We concentrate on Condition C calls because they are meant to increase medical attention

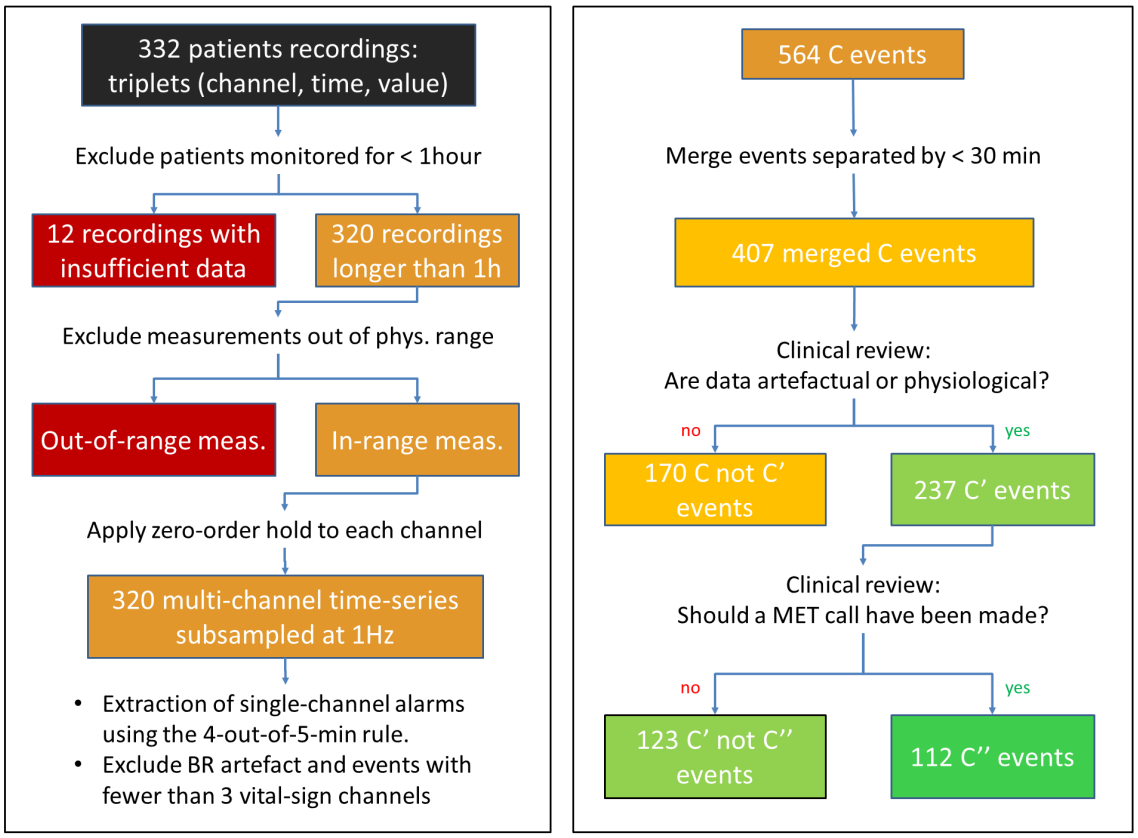


Figure 2.1: From raw data to clinical labels: left, flowchart of the pre-processing and event extraction procedures; right, procedure for event rejection and clinical labelling. The *C not C'* are periods of abnormal measurements resulting from artefactual measurements, as assessed by the clinical reviewers. The *C' not C''* events are periods of abnormal measurements resulting from trusted physiological measurements which were retrospectively deemed not to warrant a MET call to be made based on the patient's vital-sign data.

on those patients who are deteriorating, and stabilise them before they require a Condition A call.

During Phase 1, 10 MET calls were made by the ward staff. Of these, 3 were Condition C calls unrelated to vital-sign observations (caused by chest pain, or mental status changes without abnormal physiology), which an early-warning system based on vital signs cannot be expected to identify. This leaves 7 MET calls related to vital signs. The aim of an early-warning system is to detect gradual patient deterioration as soon as it sets in, thereby preventing the MET call. This would be achieved by identifying the periods of physiological abnormalities preceding the events, possibly occurring between intermittent observations made by the nursing staff.

2.2.2 Pre-processing

Raw data were acquired as series of triplets (vital-sign channel, time stamp, value). Due to the asynchronicity of the various measurement devices, these series of triplets were transformed into a regularly-sampled multivariate time-series, using the following two-step procedure:

1. Measurements outside the physiological "artefact rejection" limits shown in Table 2.2 were discarded.

2. For each vital-sign channel, a zero-order hold was applied: a measurement is held until a new measurement within the physiological limits is made, or it has been held for the maximum holding time allowed for this particular channel (40 seconds for HR, BR, and SpO_2; up to 30 minutes for blood pressure parameters), whichever comes first. If a new value has not been read after the maximum holding time has been exceeded, this channel is considered to be missing until a new valid value is acquired, or the recording for that patient ends.

In theory, this process enables us to have values of each vital sign channel at any point in time of a patient's recording, but for practical reasons, each vital-sign channel was resampled at 1 Hz.

2.2.3 Event generation

The pre-processed vital signs are then used to determine if vital-sign measurements occur outside the single-channel MET limits used to activate a MET call. A single-channel alert is generated if a vital sign takes values outside the MET alert thresholds for at least 80% of a 5-minute window (4 out of 5 minutes), termed the *activation window*. Alerts are deactivated if the vital-sign which triggered the alarm takes values inside the MET alert thresholds for at least 2 minutes out of 3-minute window, termed the *deactivation window*.[3] There is no MET rule associated with low diastolic blood pressure, which means that no event is generated from episodes of diastolic hypotension.

It should be noted that this event detection method is fundamentally univariate, which may be seen as fundamental limitation of the resulting labelled data. There may be challenging cases in which two or more vital-sign channels are jointly abnormal, but marginally normal, with respect to a normal population. Such cases would ideally be captured by an event generation procedure. We note, however, (i) that the standard of care at the time of data collection only included single-channel alert systems; (ii) that clinically adverse events, while they may be preceded by episodes where vital-sign are only jointly abnormal, systematically lead to episodes when at least one vital-sign channel is marginally abnormal. Conversely, assuming no artefactual data, all episodes in which one vital-sign is marginally abnormal must be examined. Episodes during which vital-sign are jointly abnormal but marginally normal, which would not be detected by traditional alert systems are poorly documented and fall outside of the scope of this thesis.

2.2.4 Event rejection

A visual inspection of the vital-sign time-series revealed the presence of a specific type of artefact in the breathing rate time-series, described as "a constant repeated very low breathing rate lasting for up to 15 minutes" (Hann 2008). Low breathing rate events were tested for the presence of 5 minutes of uninterrupted constant values in a 20-minute window starting 5 minutes before the start of the event. 32 events were removed from further analysis using this

[3] These 4-out-of-5 and 2-out-of-3 rules which were established empirically, are discussed in Section 3.1.1.

	Artefact rejection limits		Single-channel MET alert limits	
	Lower	Upper	Lower	Upper
HR (bpm)	30	300	40	140
BR (rpm)	3	45	8	36
SpO_2 (%)	60	-	80	-
SBP (mmHg)	-	-	80	200
DBP (mmHg)	-	-	-	110

Table 2.2: Vital-sign limits for artefact rejection and single-channel MET alert limits (Table 7.2 and 7.3 in Hann (2008) and Devita et al. (2004)). Using heart rate as an example, measurements less than 30 bpm or greater than 300 bpm are discarded as artefactual. Measurements between 30 and 40 bpm are considered to be indicative of bradycardia. Measurements between 140 and 300 bpm are considered to be indicative of tachycardia.

Vital Sign	HR		BR		SpO_2	SBP		DBP	Total
Type	High	Low	High	Low	Low	High	Low	High	
C event (before merging)	28	3	103	56	203	25	57	89	564
C event (after merging)	21	3	54	42	139	16	50	82	407
C' event	20	3	52	36	43	13	37	33	237
C" event	15	2	23	6	38	4	17	7	112

Table 2.3: Breakdown of C, C', and C" events by primary cause (Table 7.4 and 7.5 in (Hann 2008)). A C event results from the fulfillment of the MET calling criteria. A C' event results from the fulfilment of these criteria with non-artefactual measurements. A C" event is a clinically relevant C' event; i.e. a C' event which should have caused a MET call.

rule.

An event was also discarded "if there were fewer than three vital signs being recorded for the entirety of a 30-minute window centred around the start of [this] event" (Hann 2008). This was performed because the existing approach requires at least three vital signs to calculate novelty[4], but is not *a priori* clinically motivated.

After these two steps, we are left with 564 events representing periods of (potentially artefactual) vital-sign abnormality. These are termed *C events*. Patients who had at least one C event will hereafter be referred to as *C patients*. It was deemed that 564 was too large a number to be presented to clinicians for validation. Events were therefore merged if they occurred within 30 minutes of one another, which reduced the total number of events to 407. Consequently, a merged event can contain abnormal measurements relating to different vital-sign channels. The

[4]The FDA approval for the early warning system tested in the UPMC trial required that the system to use a minimum of three or more vital signs at any time.

Vital Sign	HR		BR		SpO_2	SBP		DBP	Total
Type	High	Low	High	Low	Low	High	Low	High	
Patients with C' events caused by ...	11	3	7	10	27	9	19	20	83
Patients with C" events caused by...	10	2	6	6	24	4	13	7	59

Table 2.4: Number of patients with C' and C" events broken down by cause. A patient can be a C' or C" patient as a result of multiple primary causes.

vital sign channel which first triggered an event combined with the type (high or low) of alerts will be referred to as the event's *primary cause*. Table 2.3 shows a breakdown of the primary causes by vital sign before and after merging.

2.2.5 Clinical validation

Merged events were independently presented to two UPMC clinicians for review[5]. This review resulted in the validation of 237 events for which Condition C criteria were fulfilled by non-artefactual data (whether or not a call was made in practice). These 237 events are considered to result from the fulfilment of the MET calling criteria by actual physiological data, and are termed *C' events*. Patients who had at least one C' event are termed *C' patients*. Compared to the number of MET calls made in practice (10), the number of fulfillments of the MET criteria over a period of 4 out of 5 consecutive minutes (237) is surprising at first sight. There are two major explanations for this large number of events: (i) most of these C' events were unobserved, because they occurred between observations made by the nursing staff; (ii) the (unknown) number of these events that were observed in practice will have either resulted in one of the ten MET calls made by the nursing staff, or were not deemed sufficiently abnormal to justify a MET call, a decision relying on clinical experience and knowledge of the patients.

The C' events were then assessed to determine whether, in hindsight, a MET call should have been made. The total number of events which would have necessitated an actual MET call,

[5]A comprehensive description of the review protocol can be found in Section of 7.7.3 of Hann (2008).

termed *C" events*, was 112. Given the 10 MET calls made in practice, and if it is assumed that observed C' events were correctly assessed by the nursing staff, it is clear that a large proportion of C" events were unobserved. As a consequence, assuming that C' events occur uniformly between observations, most C" events occur also between observations. We also concluded that only about half of all C' events appear to necessitate a MET call, indicating that the fulfillment of MET criteria by vital-sign measurements may have a low specificity when it comes to identifying the clinically relevant C" events. Patients who had at least one C" event are termed *C" patients*. Table 2.3 shows C' and C" events by primary cause.

Merged C events of hypoxaemia are the most likely to be discarded (97 out of 139, or 70%), while 60% of diastolic hypertensive events are also discarded. This suggests a low confidence in the readings for both vital signs. Going from C' to C" events (i.e., from non-artefactual events to truly clinically-relevant events), most blood pressure related events are discarded (55 out of 83, or 66%), as well as bradypneic (83%) and tachypneic (66%) events. These results reveal the belief of the clinical reviewers that the MET calling criteria for HR and BR can be fulfilled by patients unlikely to suffer from a critical cardiorespiratory event.

Table 2.4 gives the breakdown of C' and C" patients by primary event cause. Note that a patient can be a C' patient for different primary causes and the totals in the rightmost column are expected to be less than the sum of the values in their corresponding row.

2.2.6 Use of labels

Hann (2008) uses the events labelled C" to optimise the novelty threshold in the pdf-thresholding method described in Section 1.2. Assessing the performances of the system is difficult because the standard concepts of "true positive", "false positive", "true negative", and "false positive" are ambiguous when handling time-series containing discrete events. Hann uses event-based definitions for true positives and false negatives; i.e. a true positive or false negative refers to the classification of an individual C' or C" event, of which there may be several for a single patient. An alert is considered to be a true positive if it intersects with a 30-minute window centred around the start of the C' or C" event (termed the *diagnostic window*). A false negative

occurs when no alarm is triggered during this 30-minute window. These definitions allow the use of an event-based sensitivity defined as:

$$\text{Sen}_e = \frac{\text{TP}_e}{\text{TP}_e + \text{FN}_e}, \tag{2.1}$$

where TP_e and FN_e are the numbers of event-based true positives and false negatives, respectively.

Conversely, the definitions of "false positives" and "true negatives" are patient-based: a false positive is deemed to occur when at least one alarm is triggered during the entire recording of a patient for which no C' or C" event has been generated. A true negative is deemed to occur if no alarm is activated for this patient. These definitions allow the use of a patient-based specificity.

$$\text{Spe}_p = \frac{\text{TN}_p}{\text{TN}_p + \text{FP}_p}, \tag{2.2}$$

where TN_p and FP_p are the numbers of patient-based true negatives and false positives, respectively.

With these definitions, Hann obtains a sensitivity of 61.2% and a specificity of 61.8% of his novelty detection method, using C' events as the ground truth, and a sensitivity of 70.5% and a specificity of 71.0% using C" events as the ground truth.

2.2.7 Clinical analysis

Additional analysis of the UPMC Phase 1 data was conducted by the clinicians at the UPMC and reported by Hravnak et al. (2008). The C' and C" events were provided by Hravnak et al.[6], who used them to draw three conclusions:

- Despite pulse oximetry being routinely used as part of the existing standard of care in the ward, 30% of monitored hours did not have peripheral oxygen saturation data. Non-compliance was determined to have been caused by patients disliking probes, staff being

[6] respectively termed MET_{min} and MET_{full} by Hravnak et al. (2008).

unaware of the standard, and equipment unavailability.

- Most patients (75%) were stable during their entire stay in the SDU. Even patients who developed instability were stable for the majority of their stay, which has important implications for SDU bed use and SDU staffing. First, this is an indication that a number of SDU patients could be assigned to a lower-level ward. Second, continuous vigilance by nursing staff, either at a central station or at the bedside "reflects a highly inefficient use of human resources". Third, 5 out of 7 episodes of cardio-respiratory instabilities that were observed on the ward and required a MET call[7] were due to combinations of variables, rather than to a single vital sign. The authors conclude that attention to single-channel alarms also reflects sub-optimal use of monitoring technology and nurse time.

- Clinically significant cardio-respiratory instability and its warning signs frequently go unnoticed. The study established that an automated early-warning system has the potential to improve earlier recognition of instability. C" events were found to precede all 7 actual MET activations, with a mean lead time of 6.3 hours. However, the distribution of lead times was found to be bimodal with a large variance, and the small number of MET activations implies these results are not statistically significant. However, they emphasise the potential for improved early detection as a result of using automated alerting systems.

To gain a deeper understanding of the data set, we carried out new analyses of the UPMC Phase 1 patient population and sets of labels, which are reported in the remainder of this chapter.

2.3 Summary of UPMC Phase 1 data

2.3.1 Mean, standard deviation and channel completeness

Out of the 332 patients in Phase 1, 8 have no HR, BR, SpO_2 or BP data. An additional 4 have some vital-sign data, but their recordings are shorter than 1 hour. These 12 patients are excluded from all analyses in the remainder of this thesis, leaving 320 patients with HR, BR, SpO_2 and

[7] termed MET_{actual} by Hravnak et al. (2008).

BP data for a period longer than one hour. This group of 320 patients is hereafter referred to as the Phase 1 population.

There are three types of statistics in which we are interested:

- Statistics of the population of vital-sign measurements; i.e, statistics computed after appending all patient recordings end to end as if they formed one long single recording (called *all recordings* in Table 2.5). When computing such statistics, patients with longer recordings contribute more data, and are therefore given more weight.
- Statistics of features of individual patients' recordings (e.g., vital-sign channel mean,or vital-sign channel completeness for each patient), following which it is possible to compute averages and standard deviations across patients (*across-patients* or *per-patient* statistics). In such cases, each patient's recording contributes to the final statistic equally, regardless of its length.
- Statistics of patients who had measurements or events for a particular vital-sign channel or event type. These statistics are useful to determine sub-group characteristics. They are also necessary for cases in which the inclusion of the whole population is not appropriate, such as when calculating the mean of a vital-sign channel across patients (where only those patients who had measurements for this vital-sign channel should be included).

Table 2.5 reports a series of such statistics for the UPMC Phase 1 data:

- Line 1: most patients have data for each vital-sign channel. Breathing rate is the vital sign for which the smallest number of patients have recorded data (298 patients, or 93% of the population).
- Lines 2 and 3: the 3 "continuous" vital signs (HR, BR, SpO_2) were measured on average tens of thousands of time for each patient. However, the standard deviations across patients are of the order of the mean themselves, indicating a wide range of lengths of stay in the SDU. By nature less frequent than continuous measurements, BP measurements appear to occur 1000 times less frequently than HR, BR or SpO_2 measurements. This is confirmed by the mean sampling rate across patients and sampling rate across all

		HR	BR	SpO_2	SBP	DBP	SDA
1	Num. of patients with measurements	317	298	312	311	311	311
2	Num. of meas./patient	21,884±22,720	21,109±23,645	14,832±17,304	14.46±14.78	14.46±14.78	14.46±4.78
3	Num. of meas./patient with meas.*	22,091±22,727	22,667±23,771	15,213±17,359	14.87±14.78	14.87±14.78	14.87±14.78
4	Total num. of meas.	7,002,738	6,754,800	4,746,330	4,626	4,626	4,626
5	Mean sampling rate across patients (min)*	0.25±0.16	0.24±0.17	0.51±1.46	360.70±338.00	360.70±338.00	360.70±338.00
6	Total sampling rate (min)	0.22±6.03	0.21±6.18	0.31±12.57	323.74±345.00	323.74±345.00	323.74±345.00
7	Measurement mean across patients*	82.67±13.21	18.42±3.34	95.98±2.19	128.86±15.12	69.29±11.03	99.08±11.29
8	Measurement standard dev. across patients*	9.47±4.14	4.11±1.17	3.37±1.99	12.51±6.19	9.23±4.93	9.39±4.70
9	Measurement mean (all recordings)	83.11±16.86	18.49±5.37	95.96±4.36	128.51±20.68	68.06±14.56	98.28±15.18
10	Mean fraction of invalid meas. (%)	0.18±1.32	0.85±1.72	0.00±0.00	-	-	0.04±0.46
11	Fraction of invalid meas (%, all patients)	0.13	0.75	0.00	-	-	0.04
12	Monitoring length/patient, after zero-order hold (hours)	66.55±69.76	61.80 ±69.33	44.05±51.29	41.22±46.16	41.22±46.16	41.22±46.16
13	Total monitoring length, after zero-order hold (days)	887.29	824.04	587.29	549.60	549.60	549.60
14	Completeness/patient, after zero-order hold (%)	78.56±19.33	73.44±27.53	54.75±26.59	49.83±22.06	49.83±22.06	49.83±22.06
15	Completeness after zero-order hold (all patients, %)	80.18%	74.46%	53.07%	49.66%	49.66%	49.66%

Table 2.5: UPMC Phase 1 data statistics. An asterisk in the second column indicates that the values shown in the corresponding line included only the patients who had measurements for the corresponding vital-sign channel. Wherever possible, the format $\mu \pm \sigma$ is adopted, where μ and σ are the mean and the standard deviation of the corresponding statistic (in the second column), respectively. Line 8 therefore presents a mean of standard deviations across patients, and a standard deviation of patient-specific standard deviations. Data in lines 1 to 11 are computed using actual measurements, and are independent from the zero-order holds.

recordings given on line 5 and 6.

- Lines 7, 8 and 9: the average patient means (line 7) are comparable to the means across all recordings (line 9). The average patient standard deviations (line 8) are, however, significantly smaller than the standard deviations across all recordings, for all vital-sign channels. This indicates that the variability of a patient's vital-sign channel is smaller than that of the whole population.

- Lines 10 and 11: for all vital-sign channels, the percentage of measurements which fall outside the physiological limits in Table 2.2 (which are indicative of malfunctioning probes or extreme artefacts) is less than 1%, and less than 0.01% in the case of SpO_2. It is noteworthy than the lower limit for SpO_2 is set at an extremely low value ($SpO_2 = 10\%$, and a patient would be physiologically endangered long before their SpO_2 reaches that limit. The relatively large standard deviations shown in line 10 indicate that excursions outside the set limits do not occur in an i.i.d. fashion (i.e., randomly distributed between patients), but instead occur for a prolonged period of time in some patients, and not at all in most.

- Lines 12 and 13: monitoring lengths after filtering and zero-order hold show that SpO_2 and BP parameters are present approximately 30% less frequently than HR and BR. This phenomenon was mentioned in Section 2.2.7 and addressed by Hravnak et al. (2008). In particular, blood pressure parameters are measured less frequently and held for a shorter time than their average sampling rate, which implies significant periods of downtime.

- Lines 14 and 15: after the filtering and zero-order hold (in Section 2.2.2), vital-sign channel completeness is defined as the percentage of time data are available for each vital-sign channel. SpO_2 data are only available 53% of the time, while BP parameters are only available 50% of the time. Consequently, HR, BR, SpO_2, and BP cannot be obtained simultaneously more than 50% of the time, which highlights one of the main challenges of designing an automated early-warning system; i.e., coping with missing data and missing vital-sign channels, which are part of the reality of a clinical environment. The left-hand side plot in Figure 2.2 shows the fractions of the time n channels are simultaneously

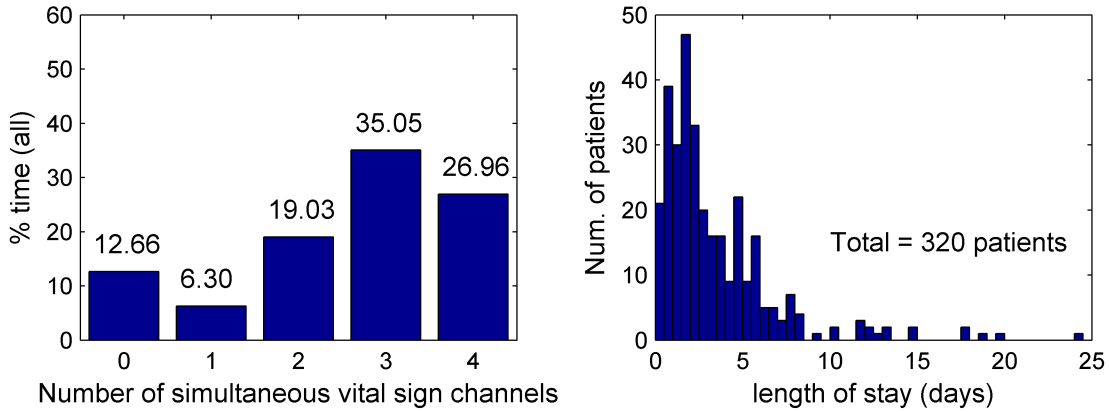


Figure 2.2: Left: number of available vital-sign channels (all recordings, in % of total time) after filtering. Blood pressure variables are counted as a single vital-sign channel as SBP, DBP, and SDA are all derived from the measurements of MAP. Overall, at least 3 vital-sign channels are available 62% of the time. Right: UPMC Phase 1 lengths-of-stay. Each vertical bar represent a 12-hour period. The height of the bar represents the number of patients whose monitoring finished during this 12-hour period.

available. Perhaps surprisingly, no vital-sign channel is available 12.7% of the time, while all 4 vital-sign channels are only available 27.0% of the time. While the limited simultaneous availability of vital-sign channels is not critical when using single-channel alerts, these numbers highlight the prevalence of missing data in a ward environment, and motivates the development of early-warning algorithms capable of returning consistent outputs, even at times when one or more vital-sign channels are unavailable.

2.3.2 Lengths-of-stay

The right-hand side plot in Figure 2.2 shows a histogram of lengths-of-stay (LOS) of Phase 1 patients. The mean LOS is 3.46 ± 3.41 days. A total of 205 patients (64%) have LOS shorter than 3.46 days, indicating than the distribution of LOS is positively skewed.

2.3.3 Univariate histograms

Figure 2.3 shows the empirical pdfs for HR, BR, SpO_2, and the three BP variables, obtained using all available measurements across all channels for a given vital-sign channel (with no weighting by patient length-of-stay). MET limits are represented by vertical lines. The his-

tograms of HR and BR data appear to be unimodal. The histogram of SpO_2 data is bimodal, with one peak around 97% (corresponding to SpO_2 measurements from stable patients who did not receive oxygen therapy) and a second one at 100% (corresponding to SpO_2 measurements made when the patient did receive oxygen therapy). Measurements from the monitors are quantised in their respective units (bpm, rpm, %, mmHg). However, normal and abnormal ranges in units of their respective standard deviations differ, which explains why the BR and SpO_2 histograms appear smoother (their quantisation bin sizes are approximately one-fifth of their respective standard deviations) than those for HR and BP (for which bin sizes are less than one-fifteenth of the respective standard deviations). Additionally, the histograms of the BP variables are also less smooth because they include 1,000 times fewer measurements than the "high rate" vital-sign channels.

2.4 Summary of C' and C" events

2.4.1 Event lengths

Figure 2.4 shows the superimposed histograms of durations for C' events and C" events. Since the C" events form a subset of the C' events, a red bar cannot be higher than the corresponding blue bar. The length ratio for a red and a blue bar is the proportion of C' events of given duration which were validated by clinicians as being clinically relevant (and therefore labelled as being C" events). From these graphs, there seems to be differences between the clinical relevance of shorter versus longer events. This hypothesis is seconded by the plot in Figure 2.5, in which the ratios of C" events to C' events of a given duration or shorter are represented, shown according to their primary cause for alerting. Including all primary causes (shown by the thicker black line), the percentage of C' events which are also C" events varies between 38 and 60% for C' events shorter than 10 minutes, which then slowly converges towards 47% (112 divided by 237). The lowest acceptance rates are observed for shorter durations for all primary causes, indicating that shorter events are less likely to be clinically relevant, either because they are transient and isolated physiological events, or because they are too short to be distinguished

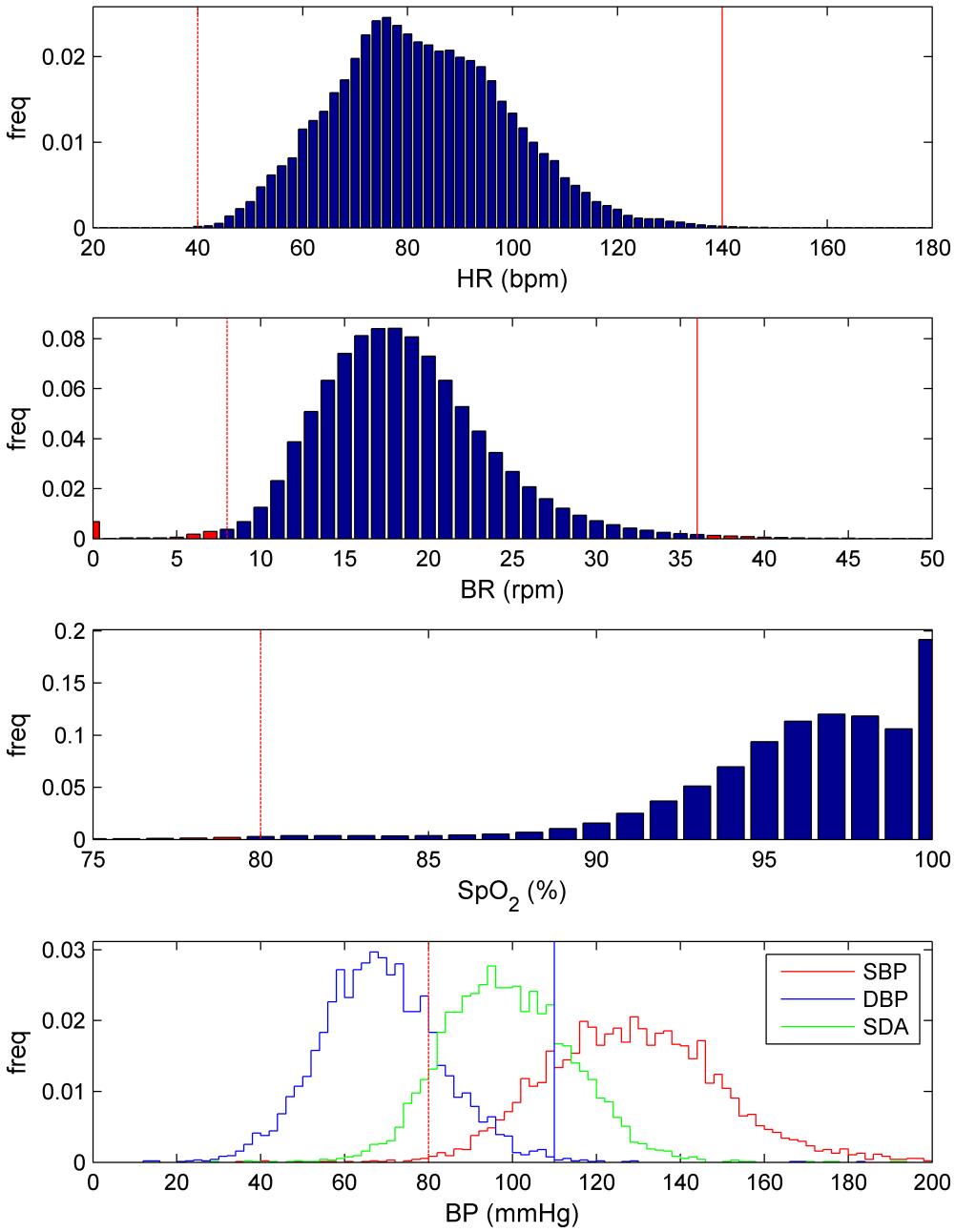


Figure 2.3: Empirical distributions of UPMC vital-sign data. For HR, BR, and SpO_2 measurements outside the MET limits in Table 2.2 are represented in red. In lower plot, which shows the distributions of BP variables, a vertical dotted line indicates the lower MET limit for the variable of the same colour, while a vertical continuous line indicates the upper MET limit for that variable. Note that there is only one upper and one lower MET limit set for the SBP and DBP, respectively.

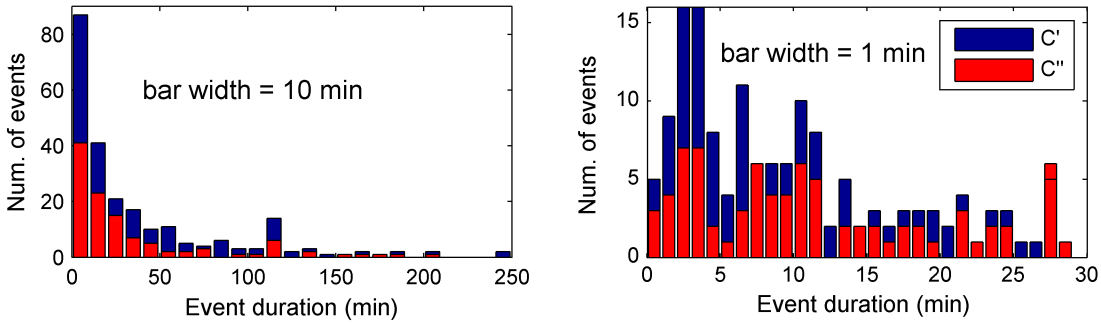


Figure 2.4: Histograms of C' (blue) and C" (red) event durations. Left: the whole range of durations is shown. A bar represents a 10-minute interval. Right: only events shorter than 30 minutes are shown. A bar represents a 1-minute interval.

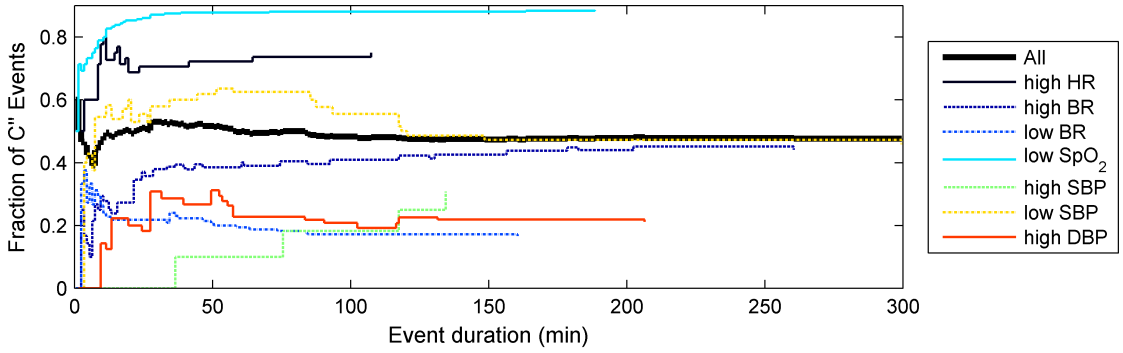


Figure 2.5: Fractions of C' events labelled C". At a given event duration (x-axis), the curves represent the fraction of C' events shorter than this duration which are also C" events, shown by primary abnormality (and where low breathing rate events, of which there are only 3, are excluded for clarity). A large fraction (y-axis) indicates that a large number of C' events were deemed clinically relevant. The thicker black line includes all types of abnormalities (all 407 C' events, and all 112 C" events). Each curve ends at the maximum length of the C' events to which it corresponds.

from short periods of artefactual measurements. The primary causes with the highest overall acceptance rates are low SpO_2 (88%), and high HR (75%). High DBP (21%) and low BR (17%) are the primary causes least accepted as being clinically relevant. In the case of the latter, there is a suspicion that the filtering technique of Section 2.2.2 did not suppress all the low breathing rate artefacts.

2.4.2 Number of events per patient

Figure 2.6 shows the number of C' and C" events by patient and alerting vital-sign. A total of 83 patients (37.2%) have at least one C' event. However, 59 patients (18.4%) have at least one

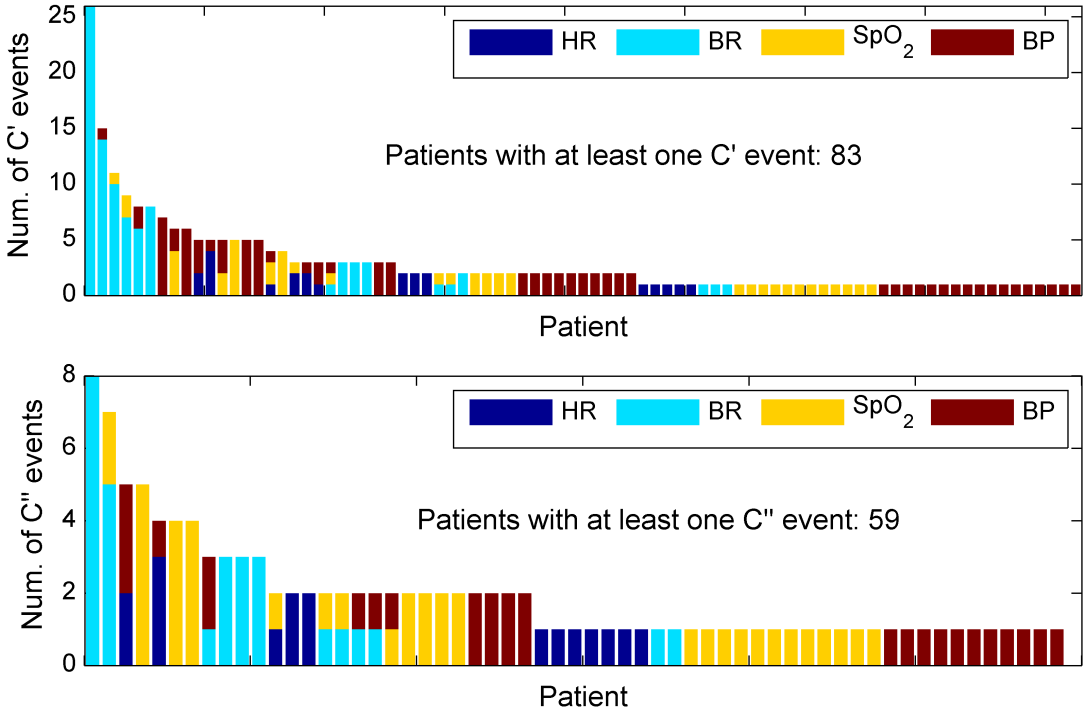


Figure 2.6: Number of C” (lower plot) and C’ (upper plot) events, shown by patient and alerting vital-sign. Each bar represents a patient. Its total height is the total number of events for this patient. The number of events caused by each vital-sign channel are colour-coded and shown as fractions of each bar. Patients are sorted in decreasing order of their total number of events.

C” events, indicating that a large proportion of patients in the study did not exhibit clinically significant vital-sign abnormality, and that 62.8% of patients did not have any non-artefactual vital-sign abnormality.

The numbers of patients who had C’ and C” events triggered by two distinct vital-sign channels are 14 (4% of all patients, 17% of C’ patients) and 10 (3% of all patients, 12% of C’ patients, 17% of C” patients) for three or more vital-sign channels.

2.4.3 Channel completeness

The left plot in Figure 2.4.3 shows that the completeness of every channel during C’ and C” events is greater than the corresponding average completeness reported on line 15 in Table 2.5 (e.g. 97% versus 80% on average for the heart rate channel). Similarly, channel measurements are simultaneously available significantly more often than they are in the entire set of recordings. Obviously, data need to be present for events to be detected, but these differences are

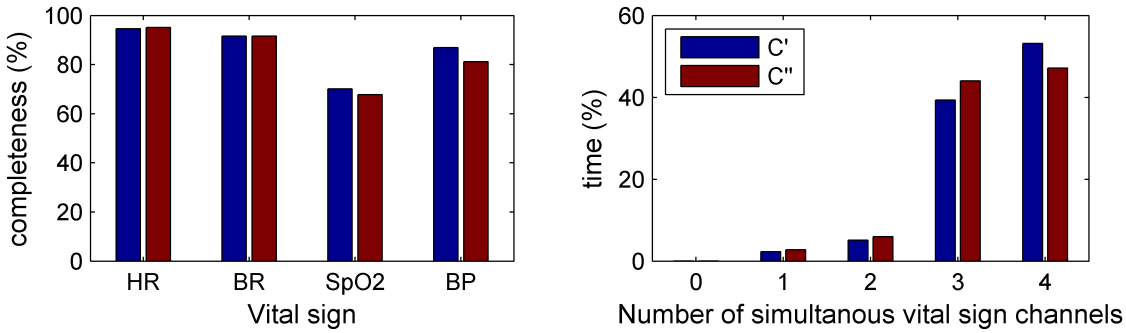


Figure 2.7: Vital-sign channel completeness (left) and simultaneous availability of measurements, expressed as a fraction of the total monitoring time, during C' (blue) and C" (red) events. The right plot is the equivalent of the left plot in Figure 2.2 for C' and C" events. BP variables are again counted as a single channel.

also explained by the rejection of events for which fewer than 3 channels had data in a 1-hour window centred around the start of a Condition C event.

2.4.4 Monitoring times

Figure 2.8 is plotted to estimate the frequencies of C' and C" events during monitoring. The upper plot shows the number of events per 24 hours of monitoring for both C' and C" events. Patients are sorted by decreasing number of C' events per monitoring day. The average number of C' events per monitoring day is 0.52 *in C' patients*, which means that even in these patients, events are few and far between. C' patients had an average of 0.24 C" events per 24 hours of monitoring. C" patients had on average 0.33 C" events per 24 hours of monitoring. Additionally, the number of C" events per C" patient monitoring day is 0.57 ± 1.24, indicating a large variability in these frequencies.

The lower plot shows the event time as a percentage of total monitoring time for all C' patients. The ordering of patients is the same as in the upper plot. The C' event time for all C' patients recordings is 1.43% of monitoring time. The C" event time for all C' patients recordings is 0.59% of monitoring time, and 0.78% of all C" patients recordings.

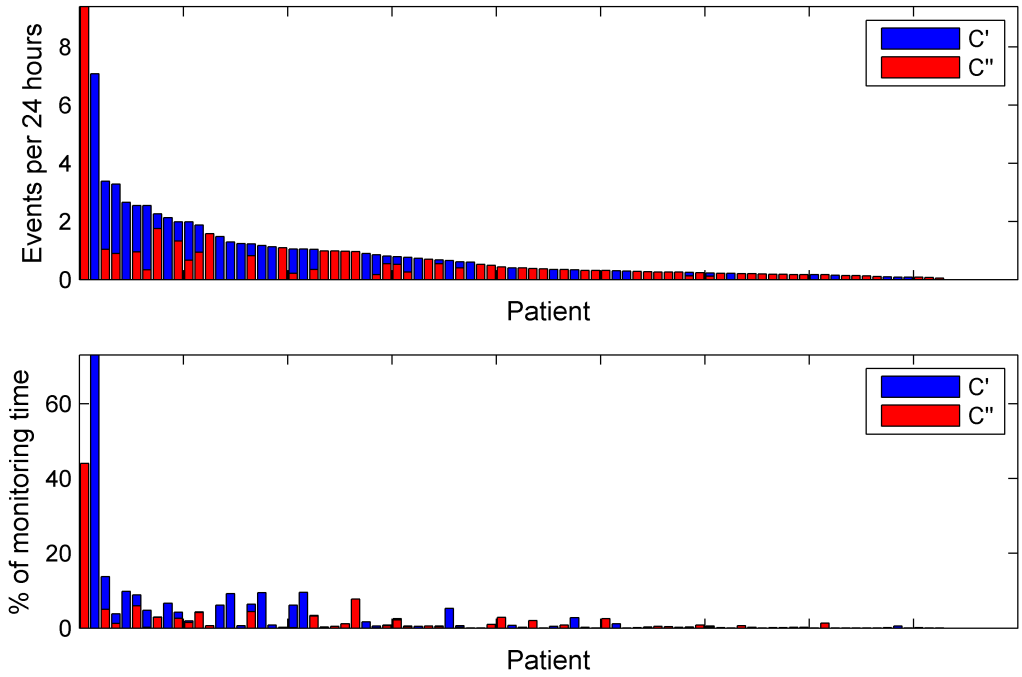


Figure 2.8: Frequencies of C' and C" events. In both graphs, patients are sorted by decreasing number of C' events. Each bar represents one of the 119 patients who had at least one C' event. Upper plot: number of events per 24 hours of monitoring. For a given patient, the total height of the bar is the number of C' events. The red part of the bar is the number of C' events which were labelled C". Lower plot: event time in percentage of monitoring time.

2.4.5 Summary

Early-warning of patient deterioration requires the identification of events which are rare at three levels:

- Few patients have clinically relevant events, but the majority of patients who do not have events have recordings which are as prone to artefacts and noise as the recordings of C' patients. Their presence in the UPMC Phase 1 population (and therefore in the SDU) could be the result of cautious triage, but may also be indicative of an inherent difficulty in assessing risks of cardio-respiratory complications, even when the assessment is made by experienced clinical staff.

- In patients who have non-artefactual vital-sign abnormality, such events remain rare, both in terms of numbers and duration. C" events occur on average 0.35 times per day per patient, and represent 0.24% of all monitoring time (an average of 3 min 27 s per monitoring day). Even if we assumed that we were able to isolate C' patients from the remainder of

	C" patients	C' patients	All patients
Number of C" events	1.91 ± 1.48	1.36 ± 1.51	0.35 ± 0.97
Number of C' events	3.19 ± 4.03	2.87 ± 3.57	0.74 ± 2.21
C" event time (all recordings)	0.77%	0.59%	0.24%
C' event time (all recordings)	1.26%	1.42%	0.58%
C" event time/patient	$1.80 \pm 5.82\%$	$1.27 \pm 4.96\%$	$0.33 \pm 2.58\%$
C' event time/patient	$2.32 \pm 6.01\%$	$3.00 \pm 7.37\%$	$0.78 \pm 3.96\%$

Table 2.6: Numbers and frequencies of C' and C" events in C" patients, C' patients, and all patients.

the SDU population, C" events would represent 0.46% of C' patient monitoring time (an average of 6 min 37 s per monitoring day). This information is summarised in Table 2.6.

- Finally, over half of the C' events (i.e., events corresponding to non-artefactual physiological abnormality) were not deemed clinically significant. Expert clinical judgement involves the use of more information than the mere observations of rare excursions in vital signs, and the gap between those potential precursors and the ultimate physiological instability that they precede represents a major challenge of in the design of automated early warning systems. Additionally, the procedure used to extract C' events relies on the persistence of abnormal unprocessed measurements, a process that does not incorporate any consideration of dynamics of, or correlations between vital signs. In other words, persistence of abnormal measurements might not be the best predictor of cardio-respiratory adverse events. This choice is motivated by and, to some extent, reflects the SDU's practice regarding MET calls, yet may not necessarily identify the most accurate warning signs of cardio-respiratory instability. While we address some methodological issues and motivate a relabelling of the UPMC Phase 1 data set in the following chapter, obtaining a "ground truth" on large time-series for events not formally defined is elusive. The crucial aim of labelling by experts is to identify cardio-respiratory adverse events requiring an unplanned transfer to the ICU or immediate resuscitation, which allows us to identify *a posteriori* the at-risk patients. We believe that Hann's labelling achieves this in capturing the most extreme periods of abnormality, at times when three vital-sign channels are available, and will be looking to extend our proposed labelling to all periods of monitoring.

	HR		BR		SpO_2	SDA	
	Low	High	Low	High	Low	Low	High
Threshold in units of standard deviations	-3.15	12.86	-2.88	7.95	-19.71	-5.15	5.38
Exceedances (per 10,000 meas)	13	0	74	4	0.00	0	4

Table 2.7: Number of threshold exceedances per 10,000 measurements for HR, BR, SpO_2 and SDA. The 0.00 in the SpO_2 column indicates that the threshold is exceeded, but less than once in one million measurements.

2.5 Problems with existing filtering

The thresholds listed in Table 2.2, which were used to discard measurements in the filtering stage, are intended to reflect the physiological impossibility of some vital-sign observations. After visually inspecting the time-series of vital signs and the univariate histograms in Figure 2.2, undesirable differences in the physiological significance of these thresholds were identified. At the filtering stage, thresholds should be conservative, as we do not want to risk discarding actual physiological measurements, and this is indeed the case for most of vital signs: as examples, the upper threshold for HR (300 bpm) and the lower threshold for SDA (20 mmHg) are never exceeded. There is *a priori* no reason why these thresholds should be exceeded at the same frequency for all vital-sign channels, because they are all derived from different probes using different physical concepts, technologies and algorithms. There is also *a priori* no reason why these thresholds should be set at the same number of standard deviations from the corresponding mean, because the distributions of each channel have different normal variance (spread), skewness (symmetry), and tails (extreme measurements). However, examination of the information provided by the univariate histograms, our knowledge of the various probes, and the data in Table 2.7 prompted us to evaluate the thresholds for BR and SpO_2.

2.5.1 Lower threshold for BR

According to Table 2.7, the BR lower threshold is the most frequently exceeded threshold despite being set at a low 3 respirations per minute (corresponding to 1 respiration cycle every 20

seconds). This is mostly because an average of 68 out of 10,000 measurements have BR$=0$, indicating probe disconnections, often at the end of a recording (where the probe was disconnected before the monitor was turned off). Accounting for probe disconnections, 7 measurements out of every 10,000 measurements are between BR$=1$ and BR$=3$, and must be considered artefactual, which is an acceptable rate, and there is no evidence justifying an increase in the value of the BR lower threshold.

2.5.2 Upper threshold for BR

The BR upper threshold was set to 45 in Hann (2008), and we choose to reexamine it because unlike a 300 bpm value for HR, a breathing rate of 45 rpm is not physiologically implausible (attainable, for instance, through intense physical activity), if only in conditions of stress, and need not correspond to a probe malfunction. The upper-left plot in Figure 2.9 shows the empirical upper-tail (BR>35 rpm) of the BR distribution. Blue and red bars represent measurements between 35 and 44, and above 45 rpm, respectively. Assuming that the blue bars represent the true BR distribution for BR<45, then the red bars form a plausible tail for that distribution. In other words, measurements above 45 could come from the same distribution as those under 45, and there is therefore no evidence to suggest that this physiologically plausible value should be the threshold for classifying measurements as artefactual. The other plots in Figure 2.9 are two-dimensional scatter plots of the pairs (HR,BR), (SpO_2,BR), and (SDA, BR), at times when BR is greater than 35 rpm in all patient recordings. Again, in every case, the red points cannot be assessed as not originating from the same distribution as the blue points. We therefore propose to increase the upper threshold for breathing rate to 60 rpm (12.44 standard deviations from the mean), above which the points in the figure appear sufficiently sparse and spread out not to come from each of the two-dimensional distributions. This new threshold is exceed 65 times (corresponding to a rate of < 1 exceedance per 10,000 BR measurements).

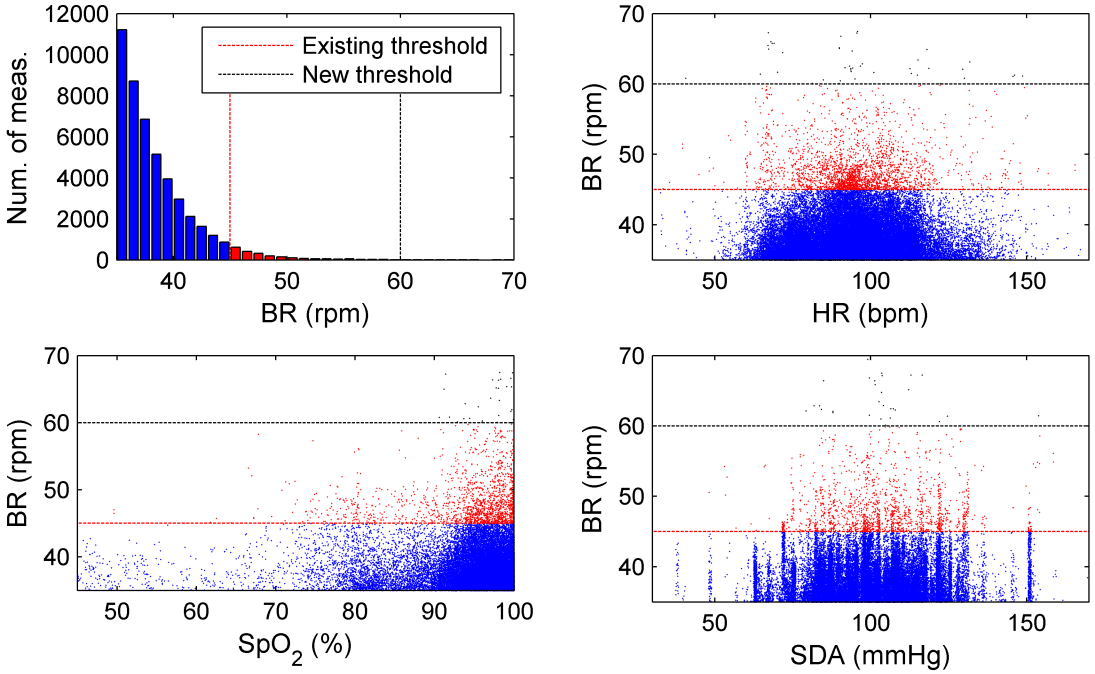


Figure 2.9: Top-left: histogram of the breathing rate measurements greater than 35 rpm. Red bars correspond to measurements above 45 rpm. The dotted black line is set at our proposed threshold of 60 rpm. Othe plots: Scatter plots of BR against HR, SpO_2 and SDA. Only points for which BR>35 rpm are represented. Patterns in red correspond to a breathing rate between 45 and 60 rpm; i.e, the patterns discarded as being artefactual using the existing threshold, but not using our proposed threshold. Points in black are those deemed artefactual using both thresholds. In these 3 plots, the locations of points are slightly perturbed (using a uniform distribution between -0.5 and 0.5) to suppress the visual effect of quantisation.

2.5.3 Lower threshold for SpO_2

The SpO_2 lower threshold was set to 10% in Hann (2008), which is extremely conservative, if the pulse oximetry measurements reflect true physiological values. There is however reasonable doubt that a significant portion of the UPMC measurements relate to true physiological values:

- One of the MET calling criteria (Table 2.2) is an "oxygen saturation reading less than 85% for more than 5 minutes (unless patient known to have chronic hypoxaemia)". Over 3% of the UPMC Phase 1 measurements of SpO_2 take values below 85% (43 minutes per 24 hours of SpO_2 monitoring), and should have been identified by existing monitors. The fact that these events did not result in MET calls indicates that the measurements are not trusted by the clinical staff.

- 4.73% of SpO_2 measurements take values below 88% (68 minutes per 24 hours of SpO_2 monitoring), which is the minimum level at which clinical guidelines recommend the onset of oxygen therapy (O'Driscoll et al. 2008, Kbar & Campbell 2006, Kallstrom 2002)[8], with the aim of stabilising SpO_2 above 94%. This again indicates a low level of trust in SpO_2 values by the clinical staff.

- The procedure used by manufacturers for calibrating pulse oximeters[9] means that pulse oximetry cannot be considered reliable well below the 80% threshold (Hill & Stoneham 2000, DeMeulenaere 2007). 0.22% of SpO_2 measurements lie below 70%, which is comparable to the percentage of C" event time.

- It is known that low values can be artefactually generated when there is imperfect contact between the finger and the probe.

The existing threshold set at $SpO_2 = 10\%$ is *a priori* the most conservative from the viewpoint of the physiology, yet this may be rendered necessary by the characteristics of the associated measurement technique. The histogram shown in the upper-left plot of Figure 2.11 has a lower-tail which is not physiologically plausible, and which is less smooth than the corresponding

[8]Recommended thresholds in these articles and guidelines are equal to or greater than 88%.

[9]empirical calibration curves developed from studies of healthy volunteers to calculate SpO_2, which therefore cannot be artificially brought down lower than 80% for ethical reasons.

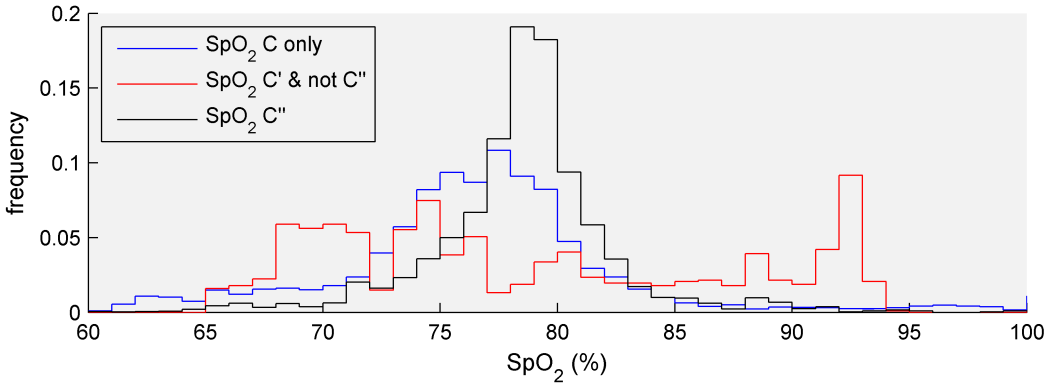


Figure 2.10: Histograms of SpO_2 channel during hypoxaemic events. In blue, only the 88 C events which were not converted into C' and C" are included. In red, only the 5 C' events which were not converted into C" events are included, such that all three curves refer to disjoint groups of events.

tail shown in Figure 2.9, indicating possible erratic measurements for SpO_2 at values as high as 60%. This is confirmed by the scatter plots in Figure 2.11 in which the two tails cannot plausibly originate from the rest of the distributions. Setting the threshold at some value greater than 50% would ensure that a large part of the artefactual measurements are discarded. There are however reasons not to set a different threshold and merely acknowledge that the existing threshold is arguably too conservative: (i) unlike the case of the upper BR threshold, no measurement of SpO_2 which may be a true measurement is discarded; (ii) measurements that are not true physiological values because they fall outside the pulse oximeter calibration range may still exhibit useful trends, as long as the order of measurements is preserved (i.e., as a smaller measurement corresponds to a smaller actual SpO_2 value). However, in doing so, we will affect the clinical clarity of classifiers trained using the available SpO_2 data (i.e., references to "SpO_2 values" should be understood as "Pulse oximeter outputs" rather than actual oxygen saturation levels).

In summary, the SpO_2 measurements have questionable accuracy and appear not to be trusted by the clinical staff at the UPMC. Any measurement below 70% should be considered with circumspection, although there are enough of these that discarding them would render the SpO_2 channel unusable. Having highlighted these points, we refrain from setting a different threshold on SpO_2 measurements.

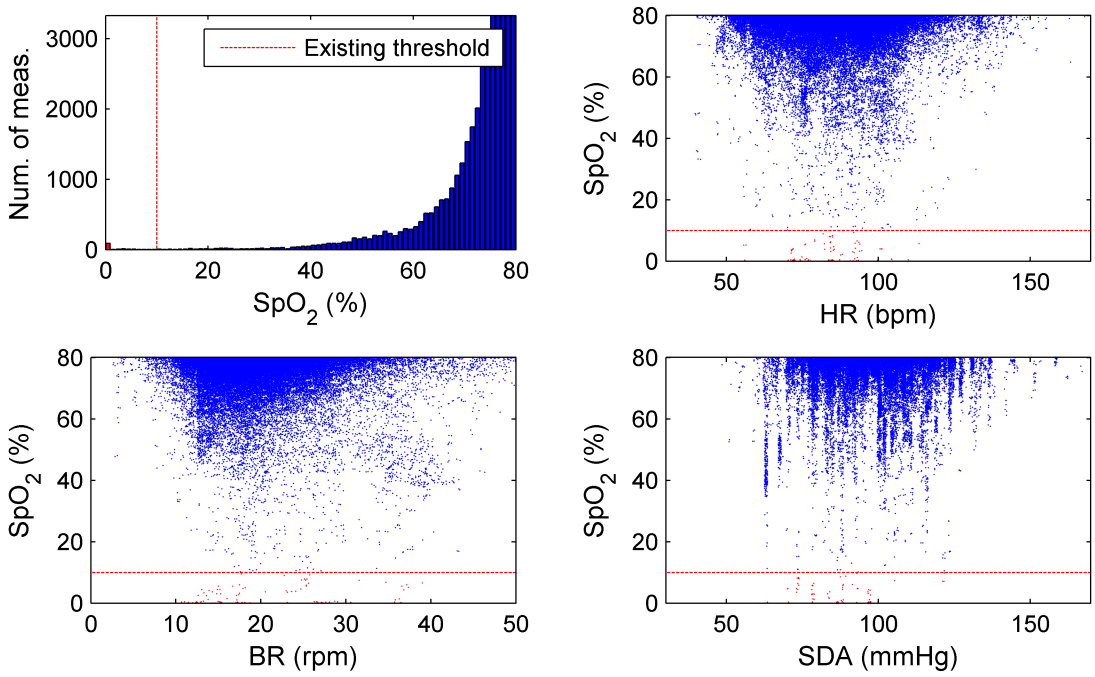


Figure 2.11: Top-left: histogram of the SpO_2 measurements below 80%. Red bars correspond to measurements less than 10%. Other plots: scatter plots of SpO_2 against HR, BR and SDA. Only points for which $SpO_2 < 80$ rpm are shown. Points in red correspond to a SpO_2 value below 10%. As in Figure 2.9, points are perturbed (using a uniform distribution between -0.5 and 0.5) for ease of visualisation.

2.5.4 Conclusion

Only the upper threshold for the BR channel was modified. Breathing rate measurements between 45 and 60 were retrieved and included in the pre-processed data set, and the zero-order hold procedure described in Section 2.3.1 was applied for consistency, thus slightly reducing the amount of missing data for the BR channel. C' and C" events generated using the former BR threshold have not been changed to allow comparison with previous work, and will be referred to as the *historical labels*. Instead, because of this threshold modification and additional reasons listed in Section 3.1 it was decided to generate an entirely new set of labels.

Chapter 3

Identifying and labelling events

3.1 Regenerating MET events

3.1.1 Motivation

This chapter is concerned with generating a new set of events and labels for the UPMC Phase 1 data set. We have already demonstrated in Section 2.5 the existence of inconsistencies in the pre-processing procedure used to generate the C' and C" events (hereafter referred to as the "historical labels"), specifically in the choices of upper physiological threshold for the breathing rate and lower threshold for oxygen saturation. The eventual aim of this research is to detect all warning signs of future physiological deterioration by taking maximal advantage of the information contained in the patient's recordings. There are two major ways in which the historical labelling is constrained.

Firstly, activation and deactivation windows have to be 100% complete; i.e, no data must be missing in a 5-minute window for it to be considered a potential activation window, or in a 3-minute window for it to be considered a potential deactivation window. The 4-out-of-5 rule means that the data in a window corresponding to 80% of the window length must be exceeding the corresponding threshold *and* data must be present the rest of the time (and therefore be in MET range, as defined in Table 2.2). This rule excludes as potential activation window any

5-minute window during which no vital-sign measurement is available for as little as 1 second. Since the observed missing data rates are over 50% for some channels (Table 2.5), and very short missing data periods are common, a large number of windows were discarded even if they were only missing a single measurement, thus reducing the number of events presented to clinicians for validation and limiting the historical labelling to the monitoring periods for which the highest signal quality was achieved. The 4-out of-5 rule and the 2-out-of-3 rule for deactivation windows are modified for the procedure described in Section 3.1.3 and we show in Section 3.4 the effect of these modifications.

Secondly, the 3-channel constraint enforced while generating the historical labelling is useful to demonstrate the ability of an early-warning system to make use of the inter-channel correlations in their decision making. However, our aim is to detect all warning signs of physiological deteriorations irrespective of the number of available vital-sign channels. When attempting to establish a *gold standard* labelling of a particular data set, it is not sound to enforce a 3-channel constraint at the event extraction stage. As an example, it would be incorrect to conclude that a patient for whom no event was identified under the 3-channel constraint did not exhibit any sign of physiological abnormality. Single- or two-channel events may have occurred, or three channels may never have been available simultaneously for this patient; however it would be incorrect to infer from the lack of 3-channel events that this patient was stable throughout their stay on the ward. Additionally, one of the expected features of a multivariate data fusion model for patient monitoring is the ability to identify abnormal combinations of vital-sign measurements which cannot be inferred from single-channel alarm systems. Since C' and C" events are single-channel events, alarms triggered by a data fusion system in a diagnostic window (such as the 30-minute window used in Section 2.2.6 preceding a C" event are indications that the data fusion system is indeed using correlations to identify precursors of single-channel events. Consistency between the labelling procedure and the event detection procedure in Algorithm 1.1 is retained by performing event detection only in those periods when at least three vital-sign channels are available. Since three or more channels are available simultaneously only 62% of the time [1], event detection is not performed on 38% of the total monitoring time. Potentially valid

[1] Exactly three channels are simultaneously available 35.05% of the time and four channels are available simultaneously 27.0% time as shown in Figure 2.2

single-channel activations occurring when only one or two channels are available are ignored, thus skewing conclusions regarding early warning and lead times. [2]

In the subsequent chapters, we develop novelty scoring systems which retain a consistent probabilistic interpretation with respect to the information available to them at the time of scoring. These systems achieve this by switching the dimensionality of the model to match the dimensionality of the input, and can therefore perform novelty detection even when a single channel is available.

Thirdly, we will aim to demonstrate that novelty scoring systems based on Extreme Value Theory or Order Statistics outperform the pdf-thresholding method even on univariate time-series. We therefore need to consider all single-channel alarms, regardless of the number of other channels available.

Finally, a number of inconsistencies in the labels themselves were found. These are better explained with respect to the set of labels regenerated in Section 3.1.2 and detailed in Section 3.1.4.

3.1.2 Related work

The availability of labelled data is a central problem in machine learning. Labelling is useful for training, and essential for testing and validating of automated techniques. In most real-life situations, labelling is a complicated and time-consuming task, especially when the labelling process, as is the case for medical data, requires a high level of expertise. In the absence of a categorical ground truth, recourse to the expertise of multiple experienced human reviewers is generally accepted as the next best option. When the data to be labelled are noisy, artefactual, or inherently ambiguous, intra-observer variability (the variability of one observer when reporting on the same data set more than once) and inter-observer variability (the variability between two or more observers reporting on the same material) can take significant values; a 26 percentage point range in self-consensus and a 43% consensus rate across 8 experts, for instance, was

[2] Hravnak et al. (2008) report a bimodal distribution of lead times between C" events (termed MET_{full}) and actual MET calls (MET_{actual}) for 7 patients (lead times under one hour versus lead times over 6 hours). A more inclusive novelty detection scheme is likely to identify additional activations and help inform these observations.

found by Culverhouse et al. (2003). Additionally, aggregating the opinion of multiple observers involves the need to adjudicate disagreements. A popular approach is to use majority voting, and to use an authoritative adjudicator to break ties. The decisions of the adjudicator serve as final decisions. Such adjudicated methods have been used with similar problems to that considered in this thesis:

- In Neamatullah et al. (2008), three clinical experts were asked to identify protected health information (PHI) in free-text medical records. This is an easier problem than that considered in this work, because consensus is not required for a word to be identified as being PHI. The difficulty of the task lies in locating *all* PHI, rather than deciding if a particular word is PHI or not. A single reviewer's opinion that a word is PHI means, in general, that a word is indeed a PHI, as indicated by the high positive predicted values reported (98% on average for individual observers). The range of sensitivities for individual clinicians (63% to 94%) is an indication that, even for what is essentially a retrieval task rather than a classification task, the performances of individual reviewers are open to large variations. The mean sensitivity for the union of two clinicians' decisions (84%, range: 89 to 98%), and the sensitivity for the union of all three clinicians' decisions (98%), indicate that, in this case, aggregating experts' opinions is beneficial to the overall sensitivity. Deciding if a given word is PHI is relatively easy; i.e, the two classes "PHI" and "not PHI" have (almost) no overlap, and a word's meaning in context is sufficient to determine the class it belongs to.

- The work of Aboukhalil et al. (2008) is more closely related to our labelling task, as it proposes to address the high false-alarm rates typically observed from existing patient monitors in the Intensive Care Unit. The authors presented a set of arterial blood pressure and ECG waveforms to two independent, experienced reviewers, who were asked to annotate each waveform as "event", "non-event" or "ambiguous". In case of disagreement between two reviewers, the event "was reviewed by one experienced physician or one experienced research engineer to provide a final adjudication". In this case, the overlap between classes is significant, and the observed features are inherently ambiguous. Consequently, over 25% of events required adjudication, either because of a disagreement

between reviewers, or because both reviewers annotated events as "ambiguous". This should be taken as anecdotal evidence that if the labelling problem is sufficiently hard, reviewers can disagree on a very significant proportion of cases.

3.1.3 Procedure

The events of interest are the periods of time when at least one vital sign takes values outside its MET range for a sustained period of time, which we term *persistent single-channel abnormality* (PSCA). Our new labelling procedure to detect PSCAs builds upon the approach described in Hann (2008), and aims to obtain a labelled set of time windows, termed *event windows* or simply *events*, for which we trust that:

1. the vital-sign data in each window meet at least one of a set of MET criteria (such as those in Table 2.2);
2. the vital-sign data in each window are non-artefactual; i.e., represent valid physiological measurements as assessed by biomedical signal processing experts.

Events which satisfy the former condition will be termed *MET events* and are equivalent to the unmerged Condition C events. The subset of MET events which satisfy both conditions will be termed *true MET events* (tMET) and are similar to C' events, with the restriction that they do not result from the merging of MET events (except those overlapping in time).[3] Events which satisfy the former condition but not the latter will be termed *artefactual MET events* (aMET). Patients who had at least one MET, tMET or aMET event will be termed *MET patients*, *tMET patients*, *aMET patients*, respectively.[4] Note that the non-artefactual nature of the tMET events is claimed only from a signal quality viewpoint, and may not reflect clinical relevance.

The procedure to detect events as MET events is close to that used to detect Condition C events. A MET event is activated if one vital-sign channel takes values outside the thresholds listed in Table 2.3 for 4 out of 5 consecutive minutes (the activation window). It is subsequently deac-

[3] C' events were merged if they occurred within 30 minutes of one another (Section 2.2.4).

[4] Note that aMET and tMET patients both form subset of MET patients, but not a partition, because a patient can have both aMET and tMET events. However, the union of the sets of aMET events and tMET events is the set of MET events

tivated if that channel takes values in its normal range for 2 out of 3 consecutive minutes (the deactivation window). We retain the approach using these hysteresis criteria, especially since doing so facilitates comparison between the existing labels and our new labels[5]. Differences with the procedure from Hann (2008) are as follows:

- The upper threshold for breathing rate is increased from 45 rpm to 60 rpm.
- Previously, if any data were missing in a 5-minute window, the window was deemed to be invalid as a potential activation window, and the next potential activation window would start at the time of the next available measurement. In our procedure, missing data are no longer ignored. The 4-out-of-5-minute rule is defined as: "MET limits are exceeded for that channel for 4 out of 5 minutes", even if no measurement is available during the remaining 20% of the time. Since this rule is less restrictive than that used to identify the condition C events, all C events are identified as MET events. The modified rule has the potential to capture more artefactual events, since missing data are often indicative of a equipment problem (e.g., a disconnected probe), but is introduced to identify true activations which would have been invalidated by the presence of a few seconds of artefactual measurements.
- A MET event is defined to last from the start of its activation window to the start of its deactivation window. This is defined because measurements in an activation window are guaranteed to be outside the normal MET ranges 80% of the time, while no more than 20% of the time in a deactivation window can be spent outside the normal MET ranges. Thus, our definition minimises the number of measurements within the normal MET ranges in an event window, and maximises it outside event windows. This becomes relevant when assessing the performance of a measurement-by-measurement classifier.

The vital-sign time-series corresponding to all the identified MET events were then shown to a panel of signal processing experts, who were asked to determine whether they were the result of artefact or if they were actual physiological data. The variability between reviewers has been analysed and is discussed in the following sections. A variety of procedures to aggregate

[5]The use of these criteria for activation and deactivation is discussed further in Chapter 7.

Algorithm 3.1 Identification of MET events.

1. Exclude measurements out of physiological range using values in Table 2.4, except that the upper BR threshold is now set to 60 rpm.
2. Apply a zero-order hold (40 seconds for HR, BR, and SpO_2. Up to 30 minutes for the blood pressure variables), to obtain one time-series sampled every second for each vital-sign channel.
3. For each channel, identify activation and deactivation windows using the modified "4-out-of-5" and "2-out-of-3" rules. Record start times and end times of event windows.
4. In cases where two or more single-channel event windows overlap in time, merge them. These merged event windows, plus all the event windows which did not overlap with any other, are the MET events.

	HR	RR	SpO_2	SBP	DBP
Low value	3	180	335	57	N/A
High value	36	135	0	26	91

Table 3.1: MET primary causes.

the experts' opinions are thereafter compared. Eventually, one aggregation method is selected which is subsequently used to summarise the experts' opinions.

3.1.4 Identification of MET events

Details of the procedure are given in Algorithm 3.1. Note that the criteria for activation and deactivation are mutually exclusive (activation and deactivation cannot happen at the same time). Using the algorithm on the UPMC Phase 1 data set yields a total of 862 MET events in 150 MET patients (47% of all patients).

The number of MET events is higher than the number of unmerged Condition C events (862 versus 564). If we consider the primary activation causes (Table 3.1, which should be compared to Table 2.3), it appears that half of this difference is due to an increased number of low SpO_2 events (from 203 to 335). This increase is explained by the new procedure used for identifying activation windows, which allows windows to be incomplete. The same holds for the other vital-sign channels, although to a lesser extent, due to lower amounts of missing data for the HR, RR, SBP, and DBP channels.

The next step is to filter out those events that were caused by artefactual measurements; i.e.,

activations that are not due to actual physiological data. In the historical labelling procedure, Condition C events were discarded if a type of artefact in the breathing rate channel was identified, or if there were fewer than three vital-sign channels available during the event. After merging, these events were presented to clinicians. In the present procedure, we do not infer a rule for identifying the breathing rate artefacts, and instead propose to rely on the expertise of the panel of experts, to identify these or other artefacts in the MET event data, so as to obtain an artefact-free set of events, corresponding to actual physiological measurements.

For the purpose of this thesis, we will not merge events which are not overlapping, even if they are relatively close to one another. The motivation for merging events in Hann (2008) was the likely dependence of two events which occurred within a few minutes of one another (which was perhaps necessary to reduce the number of events that needed to be shown to clinical experts). The merging of events which do not overlap results in the inclusion in the C' and C" events of windows comprised of data within the MET ranges. This is inconvenient when assessing the timeliness of detections, which can occur during actual PSCAs, or transition periods between C' events, and prevents us from assessing the contribution made by using the correlations between vital-sign channels. The (MET and C') single-channel events are constrained enough that we hope to detect them easily using data fusion techniques, but there is greater interest in examining detections which are not synchronous with a MET or C' event, because they will highlight periods when combinations of vital signs exhibit abnormal behaviour. Merging non-overlapping events (i.e, introducing data that are within the single-channel MET ranges into events) will hinder our ability to evaluate the performance of the novel algorithms developed in this thesis.

Before we proceed to the labelling stage, we use the newly identified set of MET events to re-examine the historical labelling and highlight some inconsistencies in the C' and C" event labels.

3.1.5 Issues with existing labelling

Further examination of the historical event labels, and the vital-sign time-series that are associated with these events, prompts a number of questions about the labelling procedure described

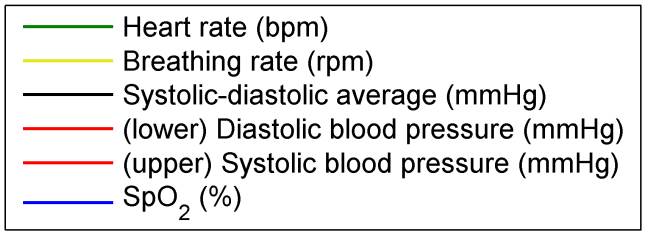


Figure 3.1: Legend for vital-sign plots.

by Hann (2008):

- Some events of interest were shown to clinicians when others were not. An example is given in Figure 3.2. The upper plot shows vital-sign data from the beginning of Patient 284's recording. Five MET events are represented by grey rectangles, and these were all caused by an elevated breathing rate. The lower plot shows the same MET periods in grey, and in pink, a binary index indicating if a point in time was included in a C" event. First, it appears that the start and end times of MET and C" events do not coincide, due to their inclusion in or exclusion from the activation and deactivation windows. The first C" event results from the union of several C events, on the basis that they occur within 30 minutes of one another. It is therefore not unexpected to find three MET events during this event. The first of these three MET events starts 5 minutes before the start of the C" event, which is solely due to the inclusion of the activation window in the MET event (which is valid, because this period contains 80% abnormal data). The third MET event ends slightly after the first C" event, because of missing data in the BR time-series (not visible at this time scale). Missing data are treated as being normal data for condition C events, whereas they are treated as neither normal or abnormal for MET events, hence prolonging the MET event in this instance. The fourth MET event represented was not labelled C" (or indeed, C'), because no SpO_2, SDA, SBP or SDA measurement was available in a 30-minute window centered on the start of the event. It was, however, labelled as being a Condition C event, but discarded before clinical review. From the vital-sign data, there is no indication that the HR or BR time-series are artefactual during this MET event, and it is likely to be as clinically significant as the other events in this figure, and should therefore be retained. There is therefore a possibility that events which are interesting

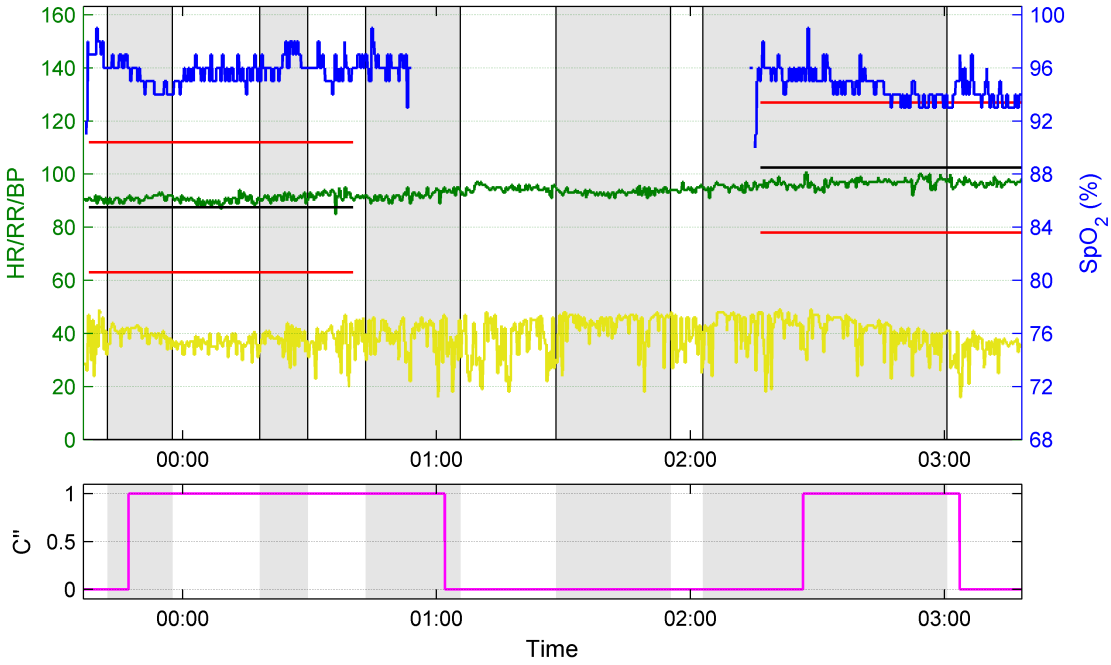


Figure 3.2: Example of labelling inconsistency: UPMC Patient 284. The upper graph shows all the vital-sign channels available. The grey area correspond to MET event periods. The lower graph shows which periods of time were labelled C" in the existing labelling. All 5 MET events are caused by an elevated breathing rate. Recordings are shown from the start of monitoring. The legend is given in Figure 3.1. In the top graph, data are presented on two different *y*-scales: the SpO_2 curve (blue) relates to the *y*-axis on the right-hand side of the graph, while all other vital signs are represented in their respective units using the scale on the *y*-axis on the left-hand side.

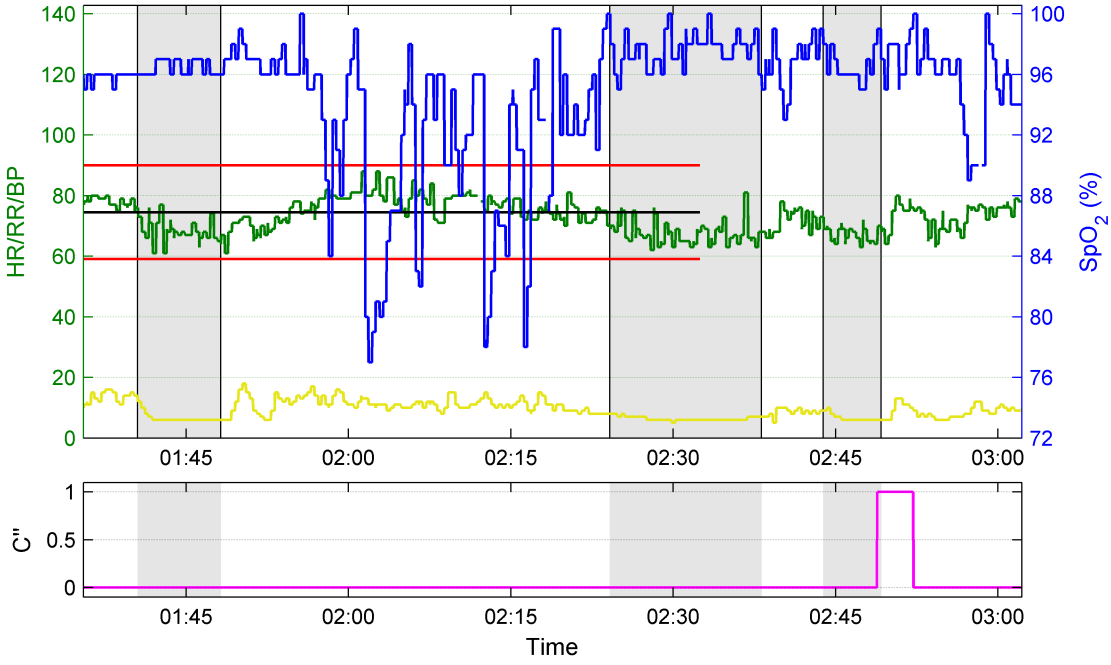


Figure 3.3: Example of labelling inconsistency: UPMC Patient 235. Plots are the same as in Figure 3.2 for this patient. All 3 MET events are caused by a low breathing rate. These recordings begin 75 hours after the start of monitoring.

from both a machine-learning and a medical viewpoints have been overlooked by the historical labelling procedure.

- C" events may have been identified due to the presence of artefactual data. Figure 3.3 illustrates this point. Patient 235's recording has 67 MET events caused by a low breathing rate, 35 of which overlap with a C' event. The remaining 32 do not overlap with a C' event either because of an insufficient number of recorded channels, or because they were deemed artefactual. Only one of those overlapping with a C' event was labelled a C" event. This is the rightmost MET event in Figure 3.3. While it cannot be claimed from the available data that the labelling was erroneous, the fact that only one out of 32 (potentially 69) events was identified as being clinically relevant, suggests that the approach of (Hann 2008) may be inappropriate, particularly when the BR data are not artefactual, as is probably the case for patient 235. On the same plot, the SpO_2 time-series appears to exhibit an erratic behaviour between $t = 1:55$ and $t = 2:30$. The excursions under 80% should perhaps not be interpreted as actual physiological values (as discussed in Section 2.5.3), but the large variations observed could be indicative of an abnormality, artefactual

or not. The SpO_2 channel does not result in a MET or C' event, because excursions below the MET threshold of $SpO_2 = 80\%$ are short and intermittent. The 5-minute activation criterion, while appropriate for detecting prolonged excursions beyond the normal physiological range, is not designed to identify abnormal dynamics of the time-series, such as that exhibited in the figure. If not artefactual, these dynamics are indicative of cardio-respiratory deterioration, yet are not identified as such by either the existing or the new labelling.

- Intra-observer variability: in Figure 3.4, we show further data from Patient 284, recorded on the fourth day of their stay on the ward. All of the 86 MET events associated with this patient overlap with a C' event. Of these, only 33 were labelled C". In the figure, the 7 earliest MET events overlap with a C" event. The remaining 6 were not classified as being C" events; i.e., they were deemed not to be clinically relevant. Visual inspection of the breathing rate time-series reveals that there is little difference in signal quality and signal values between the left and right portions of the graph (which would not have been seen simultaneously by the clinical reviewers). If they had been given longer windows of data, the clinical reviewers could plausibly have labelled all these events uniformly (as being either clinically relevant or not).

3.1.6 Visual inspection of MET events and manual labelling

A panel of biomedical engineers visually inspected data corresponding to MET events, to determine whether a particular MET event consists of valid physiological measurements, or whether it is more likely to be the result of artefacts. To this end, a MATLAB® interface was used to inspect and manually label each MET event window. For each MET event, all of the available vital-sign channels were displayed on a single graph (using the same routine as that used to generate Figures 3.2 to 3.4) for an interval of data spanning from one hour prior to the event's activation to one hour after its deactivation. The MET event window was highlighted and the vital-sign channel responsible for the activation displayed. Other MET events were displayed if they were within the display range, and the event to be labelled by the expert clearly high-

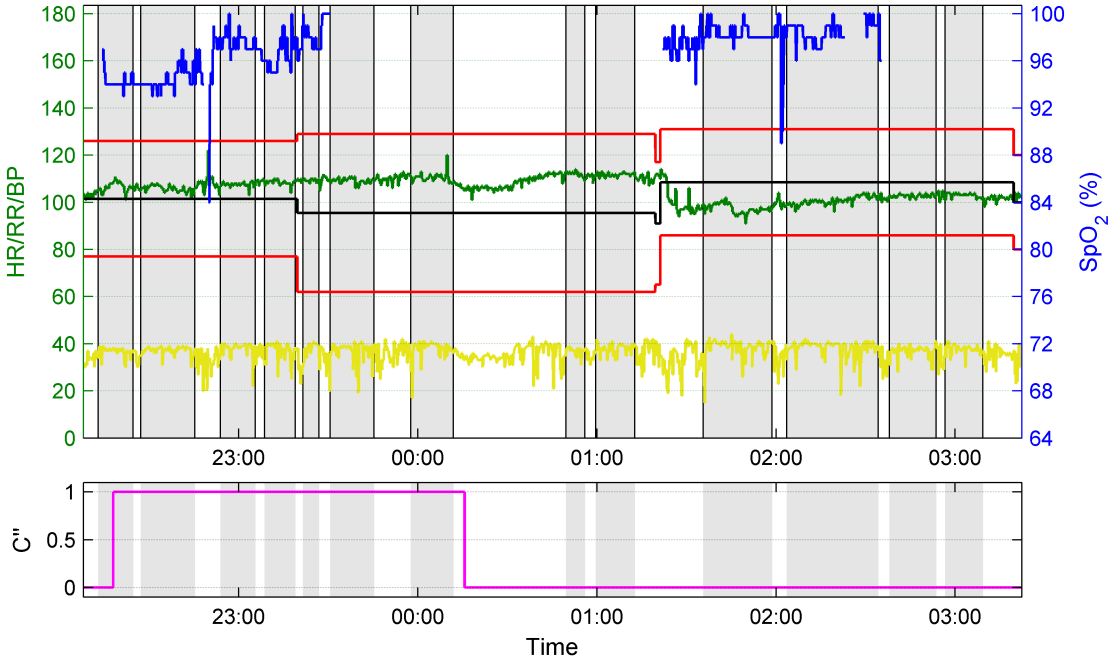


Figure 3.4: Example of labelling inconsistency: UPMC Patient 284. Plots are the same as in Figure 3.2 for this patient. All 13 MET events are caused by a high breathing rate. The patient has been monitored for 70 hours prior to the recordings shown here.

lighted. Reviewers were asked to decide whether a particular MET event activation was the result of artefactual data (aMET) or non-artefactual data (tMET).[6] Evaluation criteria (obtained from discussion with the reviewers) included the number of vital-sign channels with valid measurements, the completeness of a vital-sign channel before, during, and after the event (indicative of a probe malfunction or disconnection), the presence of erratic measurements (indicative of motion artefacts), or the presence of "jumps" in the continuous data (indicative of probe repositioning).

Three post-graduate students and a post-doctoral researcher, all experienced in analysing this type of data, comprised the panel of "experts", each of whom reviewed and classified data independently. The four reviewers are hereafter referred to as Rev1, Rev2, Rev3, Rev4.

[6]The clinical relevance of the physiological values at the time of the event was not assessed, only the quality of the time-series data.

Vital Sign	HR		BR		SpO_2	SBP		DBP	Total
Type	Low	High	Low	High	Low	Low	High	High	
Rev1	1	33	86	134	88	46	23	75	486
Rev2	1	24	30	120	76	35	18	55	359
Rev3	1	31	80	123	174	47	22	75	553
Rev4	1	34	69	134	67	53	25	85	469
Mean	1.00	30.50	66.25	127.75	101.50	45.25	22.00	72.50	466.75
% of all METs	33.33	84.72	36.80	94.62	30.29	174.03	38.59	79.67	54.08

Table 3.2: Breakdown of tMET labels by reviewer and primary activation cause.

3.1.7 Results

Table 3.2 reports the distribution of tMET labels given by each reviewer, split between MET event primary activation causes. From this table, we observe that Rev2 rejected more events than all three of the other reviewers combined. Rev3 labelled low SpO_2 events as being tMET only around 50% of the time, yet twice as often than any other reviewer. The percentages of tMET labels per primary activation cause confirm that high HR and high RR events are usually accepted, whereas low RR and low SpO_2 events seem to be the least trusted.

The total number of MET patients is 150, or 47% of the UPMC Phase 1 population. Table 3.3 shows the consensus rates for all possible combinations of two, three, and four reviewers. Three percentages are reported: the percentage of individual events on which reviewers agree, the percentage of MET patients that all reviewers agree as being either tMET patients or not tMET patients, and the percentage of all UPMC Phase 1 patients all reviewers agree are or are not tMET patients.

The consensus rate for all patients in the UPMC Phase 1 population (third row in the table) is higher than that for the population of MET patients (second row in the table), simply because no data from these patients have been shown to reviewers, and it is assumed that they are not tMET patients. These numbers are, however, relevant, because stable patients represent a large proportion of the ward population, and the implications of an early-warning system applied to these patients have to be taken into consideration, especially when evaluating false-positive rates. The consensus rates for all patients vary between 84.3% (union of all four reviewers) and 93.1% (union of Rev1 and Rev2). These numbers indicate that for 84% to 93% of patients,

Rev1	x	x	x				x	x		x	x
Rev2	x			x	x		x	x	x		x
Rev3		x		x		x	x		x	x	x
Rev4			x		x	x		x	x	x	x
Consensus (%, MET events)	77.8	72.4	86.3	67.6	77.2	72.4	58.9	70.7	58.6	65.6	54.9
Consensus (%, MET patients)	84.7	78.0	90.7	76.0	82.0	78.0	69.3	78.7	68.0	73.3	65.3
Consensus (%, all patients)	93.1	90.1	95.8	89.2	91.9	90.1	86.1	90.4	85.5	88.0	84.3

Table 3.3: Consensus rates for all groupings of reviewers on a per-event and per-patient basis. Each column corresponds to a grouping. An "x" in the first four lines means the corresponding reviewer is included in the group. The consensus rate is the proportion of events classified identically by all reviewers in the per-event case, or the proportion of patients for whom all reviewers identified at least one true MET event in the per-patient case.

reviewers agree on the presence of at least one tMET event, but they do not give any indication as to whether they labelled the same MET event as being a tMET.

Rev3 disagrees most often with the three other reviewers. Looking at the consensus rates for MET patients, any pairing including Rev3 yields a lower consensus rate ($< 69.3\%$) than the groupings of Rev1, Rev2 and Rev4 (78.7%), which suggests that Rev3 is likely to be an "outlier" in this labelling procedure. On an event basis, all four reviewers agree only 54.9% of the time, which emphasises the ambiguous nature of the labelling task, yet it is not an unusually low consensus rate for this type of data (when there are four reviewers).

Results considering the MET activation vital-sign channel that caused each MET event are presented in Table 3.4. Events related to heart rate and breathing rate are least contentious, while those related to SpO_2 are most contentious. Not only are the former labelled as being tMET events twice as often, but the resulting labellings yield, on average, a higher agreement with the majority vote (where the majority label is that label given by at least three reviewers. In case of a tie, the event is considered to be ambiguous.). The major source of disagreement between Rev3 and the other three reviewers appears to be in classifiying low SpO_2 events (hypoxaemia). The resulting differences are amplified by the fact that SpO_2-related events form the majority of all MET events.

While the consensus rate between all reviewers on an event-basis is only 54.09%, the average agreement rate with the majority vote is nearly 86% (88.86% excluding Rev3). Consequently,

	Rev1	Rev2	Rev3	Rev4	Mean
% HR events labelled tMET	87.18	64.10	82.05	89.74	80.77
% BR events labelled tMET	69.84	47.62	64.44	64.44	61.59
% SpO_2 labelled tMET	26.27	22.69	51.94	20.30	30.30
% SBP events labelled tMET	83.13	63.86	83.13	93.98	81.02
% DBP events labelled tMET	82.42	60.44	82.42	93.41	79.67
% all events labelled tMET	56.38	41.65	64.04	54.29	54.09
agreement with maj. vote (% HR events)	89.74	87.18	94.87	87.18	89.74
agreement with maj. vote (% BR events)	90.79	85.71	82.86	94.92	88.57
agreement with maj. vote (% SpO_2 events)	91.04	89.85	66.57	88.06	83.88
agreement with maj. vote (% SBP events)	96.39	81.93	79.52	83.13	85.24
agreement with maj. vote (% DBP events)	81.32	83.52	87.91	81.32	83.52
agreement with maj. vote (% all events)	90.37	86.77	77.38	89.44	85.99
agreement with maj. vote (% MET patients)	97.29	95.78	92.17	96.08	95.33
agreement with maj. vote (% all patients)	98.73	98.02	96.33	98.16	97.81

Table 3.4: Summary of reviewers' labels on an event- and patient-basis. "Maj. vote" stands for majority vote, defined as being "tMET" if at least three reviewers have labelled the event/patient true, and "aMET" if at least two reviewers have labelled the event/patient artefactual.

while consensus is difficult to obtain, a majority vote, even a majority requiring a 3-out-of-4 majority, allows us to aggregate the results in a more reliable manner, stressing the importance of using multiple reviewers.

These remarks are confirmed by the upper-left graph in Figure 3.5, where the "aggregate score" is defined as being the total number of reviewers who labelled a particular event tMET (0 if no reviewer labelled the event tMET, 4 if all reviewers agreed the MET event is a tMET event). Rev3's classifications are the most likely to be isolated (yellow part of the bar corresponding to an aggregate score of 1). The lower graph of Figure 3.5 shows the number of tMET events labelled by each reviewer and compared to the average number of tMET events. Again, Rev3 stands out as more likely to label a MET event as being true, which has significant consequences for the classification of patients as being tMET patients, because 19 patients are given no true MET events by Rev1, Rev2 and Rev4, and yet are given a single true MET event by Rev3.

Similar plots reporting results on a per-patient basis are given in Figure 3.6. Only Rev3 identified true MET patients with an aggregate score of 1. Overall the inter-observer variability appears to originate mostly from a difference in interpreting SpO_2 time-series. As discussed in Section 2.5.3, oxygen saturation measurements are the most contentious in terms of accuracy

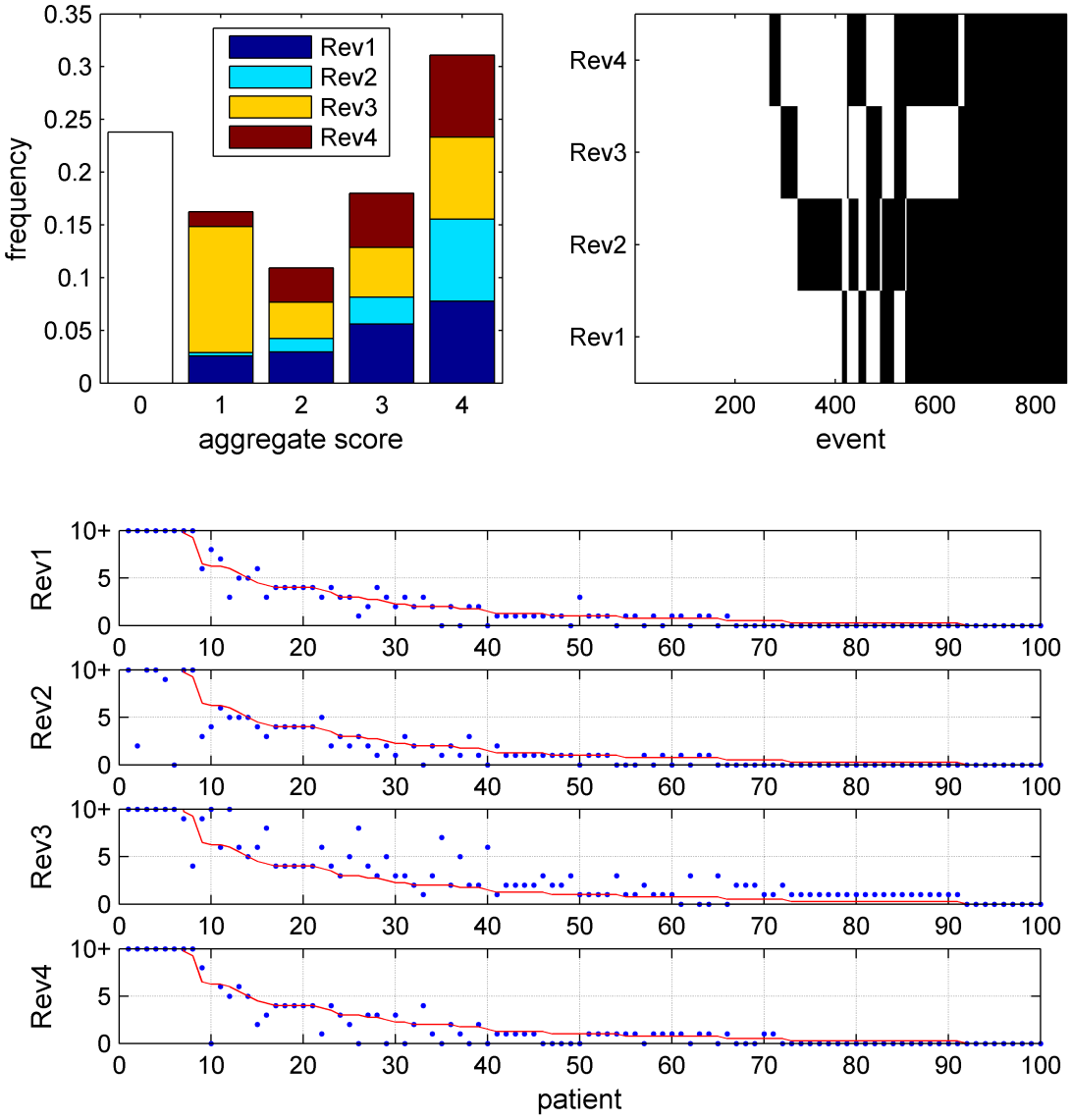


Figure 3.5: Summaries of tMET labels. Upper-left: histogram of aggregate scores. Bars are coloured according to the fraction of events labelled true by each reviewer. If no event was labelled "true", it was assigned an aggregate score of 0. All reviewers need to labelled an event "true" to obtain an aggregate score of 4. Upper-right: reviewers' labels for all events ("tMET" labels in white, "aMET" labels in black). Events are sorted according to their scores and voting patterns. Events on the far left and far right are events on which all reviewers agree. Heterogeneous columns are indicative of disagreement between reviewers. Lower: number of events per patient, according to each reviewer (blue dots). The red line in each graph is the average number of events for that patient across all 4 reviewers. Patients are sorted by decreasing average number of events. 50 patients for whom no "tMET" label was assigned by any of the reviewers are omitted.

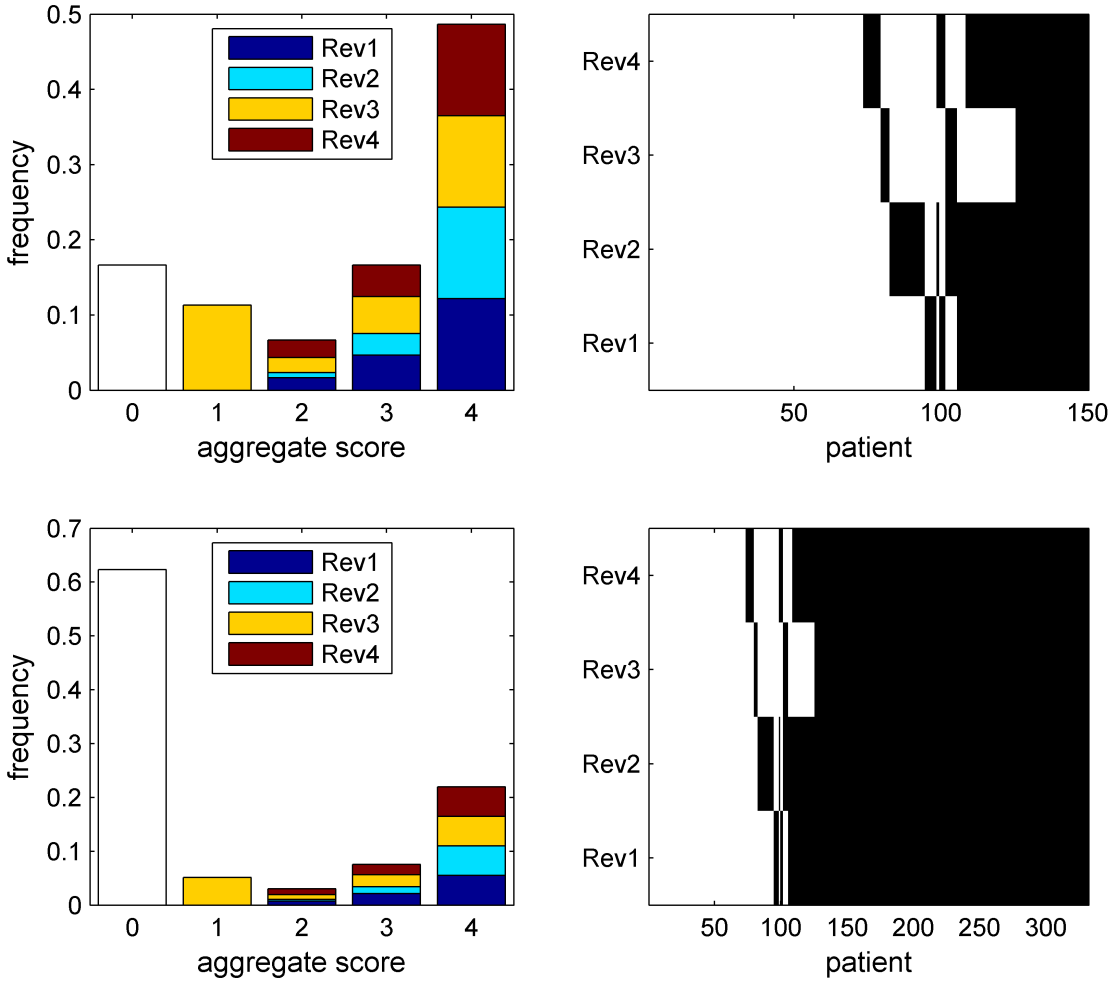


Figure 3.6: Same as the upper graphs in Figure 3.2 on a per-patient basis. For a given reviewer, a "tMET patient" is a MET patient with at least one tMET event according to this reviewer. The two top graphs only include MET patients. The two lower graphs include all 320 patients in the UPMC Phase 1 population.

and physical meaning.

Differences have been noted between the sets of labels returned by the reviewers. Excluding Rev3 from our labelling cannot be done without introducing a strong subjective bias, and would in any case leave us with three sets of labels which would still need to be aggregated to yield a single set of labels to use in the future as our reference. In the next section, we retain Rev3's labels and examine several methods of aggregating the set across all four reviewers.

3.2 Label selection

The aim of this section is to test several methods of aggregating the reviewers' set of labels and their respective robustness, in order to decide on a final set of labels.

Several methods were considered to select the final labels:

1. "majority vote using all four reviewers": a MET event is considered to be a true MET event if at least three reviewers have labelled it true, and as being an artefactual MET event otherwise.

2. "majority vote using a subset of three reviewers": considering only three reviewers, a MET event is labelled true if at least two reviewers have labelled it true. This method accounts for the inconvenience of having an even number of reviewers. Accounting for all combinations of three reviewers, this method yields four sets of labels.

3. "three-way consensus or adjudication": in turn, every reviewer is considered to be an "adjudicator". For those events for which no consensus is obtained by the other three reviewers, the label assigned by the chosen adjudicator is considered to be the final label. This method assesses differences in labelling which can be attributed to different levels of experience amongst reviewers. Events for which a three-way consensus is obtained can be said to be non-controversial. Events that are more contentious could benefit from a higher level of experience from one reviewer. Here, we do not assume prior knowledge on the respective experience of the reviewers. As with the previous method, this "consensus or adjudication" method yields four sets of labels.

The three methods of aggregation listed above result in 9 sets of labels, which yield high agreement rates. For pairs of label sets, the average agreement on an event-basis is $93.94 \pm 2.57\%$ (maximum: 97.45%, minimum: 89.10%). Across all nine sets of labels, the rate remains high at 89.09% (768 out of 862). As a consequence, fewer than 10% of the MET events are ambiguous with respect to the aggregation methods and reviewers considered.

Further examination reveals that all label sets agree if the aggregate score of a MET event is 0, 1, 3 or 4. A MET event with an aggregate score of 2 is labelled "aMET" by the four-way

majority vote. The three-way majority vote gives precedence to these reviewers who labelled the event as false. By design, the "consensus or adjudicator" method gives precedence to the adjudicator's decision.

On a per-patient basis, the consensus rate across all methods is 93.33% of MET patients (96.88% of all patients). The average agreement rate using pairs of label sets is $96.30 \pm 1.81\%$ (maximum: 98.67%, minimum: 93.33%). All methods agree on all but 10 patient labels.

We conclude that the aggregation methods considered are robust to inter-observer variability. Unfortunately, in the absence of a gold standard, the agreement or disagreement rates alone do not allow us to select any one method over the others, especially since the agreement rates are so close that all methods appear to be nearly equivalent. The favoured aggregation method will be that which on average agrees the most with all the others. This is the first method, the four-way (absolute) majority vote, which agrees on average for 95.15% of events with each of the other methods on a per-event basis.[7] Henceforth, MET events labelled "true" using the four-way majority vote are now referred to as "tMET events", and similarly for "tMET patients", "aMET events", and "aMET patients".

3.3 Summary of labels

3.3.1 Number of events

Using our final set of labels, we find that, of the 862 MET events in 150 patients, only 424 events (49.2%), occurring in 97 patients (64.7% of MET patients, 30.3% of all patients) are labelled tMET. The 438 aMET events are distributed among 110 aMET patients (34.4% of all patients), meaning that 57 patients (17.8% of all patients) had both tMET and aMET events. Primary activation causes are reported in Table 3.5, which confirms that tachycardic and tachypnoeic events are generally trusted, while bradycardic, bradypnoeic, and hypoxaemic events are generally rejected.

[7] The choice of the four-way majority vote may seem disappointing after this in-depth investigation, yet it appears to be the most sensible choice in the absence of a categorical ground truth.

	HR		BR		SpO_2	SBP		DBP	Total
	Low	High	Low	High	Low	Low	High	High	
Num. True MET events	1	29	58	134	68	44	22	68	424
% of all MET events	33.3	80.6	32.2	99.3	20.3	77.2	84.6	74.8	49.2

Table 3.5: Primary activation causes of tMET events using the absolute majority vote.

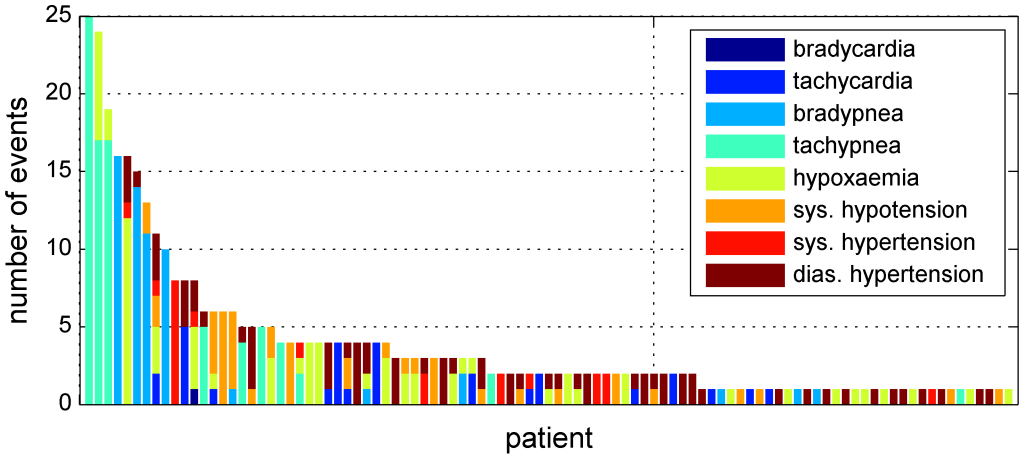


Figure 3.7: Distribution of tMET events across tMET patients. Each bar corresponds to a patient. Primary activation causes are colour-coded. The leftmost bar totals 77 tachypnoeic events and has been cropped for visualisation purposes.

3.3.2 Events per patient

Figure 3.7 shows the distribution of tMET events among the 97 tMET patients. Sixteen patients have more than 5 events, 48 have between 2 and 5 events, while 33 patients have one tMET event. In terms of primary activation causes, 33 patients have tMET events activated by two or more vital-sign channels. Overall, 49 patients (50.5% of tMET patients) had tMET events during which more than one vital sign became abnormal. In comparison, only 23.6% of aMET patients had aMET events related to multiple vital signs, from which we conclude that events during which several vital-sign channels have abnormal measurements are more likely to be tMET than aMET.

3.3.3 Event lengths

Figure 3.8 shows histograms of tMET and aMET event durations. Most MET events appear to be shorter than 10 minutes (213 tMET events or 50.2%, 294 aMET events or 67.1%). The

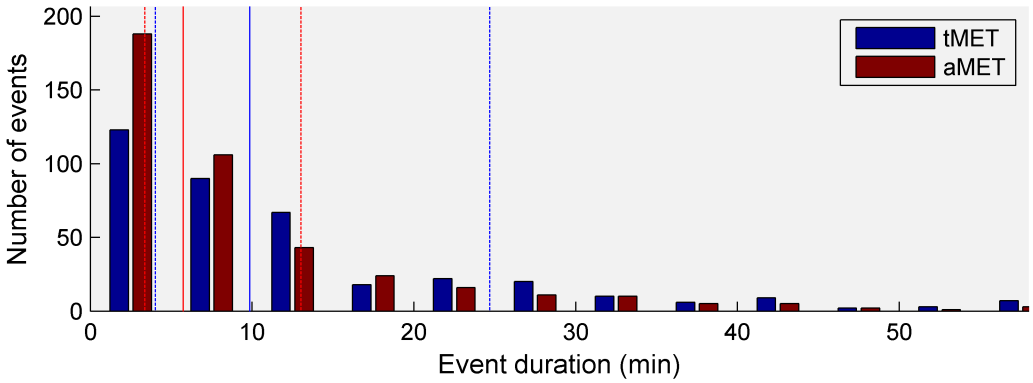


Figure 3.8: Durations of tMET and aMET events. The bin size is 5 minutes. The first bin is centred at a 2.5-minute duration. Vertical lines are (from left to right), the 25 (dashed), 50 (continuous), and 75 (dashed) percentiles for each event type. Only events shorter than 60 minutes are represented (88.9% of tMET events and 94.5% of aMET events), but all events are included in the percentile calculations.

large fraction of tMET events shorter than 10 minutes emphasises the importance of frequent observations. A significantly larger fraction of aMET events are shorter than 10 minutes, indicating that artefactual events are mostly of short duration, and conversely that short events are more likely to be artefactual. However, this difference is not sufficiently significant that it can be considered for analysis, since 42% of all MET events under 10 minutes are tMET events.

3.3.4 Channel completeness

The left-hand plot in Figure 3.9 shows the completeness of each vital-sign channel during aMET and tMET events. For every vital-sign channel, tMET events have a higher total completeness rate than aMET events. The biggest difference is observed for the HR channel (96.2% to 63.2%), which indicates that a low completeness of the HR channel correlates well with the artefactual nature of a MET event. For the SpO_2 channel, the difference in completeness rate is small (64.5% versus 62.3%), indicating that the completeness of the SpO_2 channel has little discriminative power.

The completeness rates for tMET events are comparable to those observed for C' events (left plot in Figure 2.7). They are overall slightly lower, due to the absence of constraint on the quantity of missing data and the number of available channels in the MET labelling, as occurs in Hann (2008).

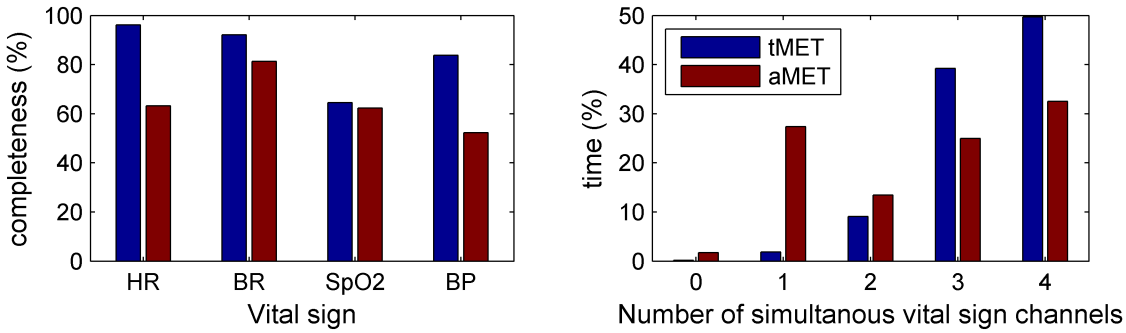


Figure 3.9: Channel completeness for tMET and aMET events. This plots should be compared to Figure 2.2 (overall channel completeness) and Figure 2.7 (equivalent plots for C' and C" events).

The impact of constraining the number of channels is made clearer in the right-hand side graph of Figure 3.9, where we show the number of simultaneously available channels during aMET and tMET events, as a percentage of event time. Reviewers appear to be more confident that an event is non-artefactual if three or more channels are available. When identifying MET events, only one channel is considered at a time, which means that the validation of the window as a MET event is not influenced by the presence or absence of other vital-sign channels. Their validation as tMET events, however, is likely to have included information from other channels, thus suggesting that the process of visual inspection would have yielded different results if the reviewers had been blinded to the additional channels available at the time of MET events. When evaluating the performance of single-channel detection algorithms, we will therefore need to remember that these labels have been obtained using information which was not made available to the single-channel algorithm in its training phase.

It is noteworthy that 4 channels are simultaneously available only 50% of the time for tMET events, which further emphasises the need for early-warning systems capable of coping with temporarily missing channels.

3.3.5 Monitoring times

Table 3.6 illustrates both the infrequency of tMET events, and the high relative prevalence of artefactual MET events. Firstly, tMET and aMET events occur in similar numbers of patients (97 tMET patients versus 110 aMET patients). Secondly, the total tMET time is only 0.89%

Patient groups	All patients	MET patients	tMET patients	aMET patients
Num. of patients	320	150	97	110
MET events duration (%)	1.56	2.48	2.97	2.63
tMET events duration (%)	0.89	1.41	1.93	1.27
aMET events duration (%)	0.67	1.07	1.04	1.36
MET event time (min/24h)	22.50	35.68	42.78	37.86
tMET event time (min/24h)	12.79	20.28	27.73	18.23
aMET event time (min/24h)	9.71	15.40	15.04	19.63
Num. MET/patient	2.69±7.35	5.75±9.90	7.57±11.80	6.85±11.27
Num. tMET/patient	1.33±5.09	2.83±7.16	4.37±8.53	2.86±8.22
Num. aMET/patient	1.37±4.13	2.92±5.65	3.20±6.77	3.98±6.27
Num. of MET event/24h	0.74±1.67	1.58±2.16	1.95±2.53	1.64±2.00
Num. of tMET event/24h	0.40±1.32	0.85±1.83	1.31±2.14	0.64±1.43
Num. of aMET event/24h	0.34±0.82	0.73±1.07	0.64±1.14	1.00±1.14

Table 3.6: MET, tMET, and aMET event durations and frequencies.

of the total monitoring time, while the total aMET time is 0.67%, or only 25% shorter than the total tMET time. Lastly, the numbers of tMET and aMET events per patient have similar means and standard deviations (1.33 ± 5.09 vs. 1.37 ± 4.13). On average, tMET patients have 1.31 tMET events and 0.64 aMET events per 24 hours of monitoring, while aMET patients have 0.64 tMET events and 1.00 tMET events over the same period.

The infrequency of tMET events means that high accuracy and specificity over time can easily be obtained by naïve classifiers, such as classifiers which always indicate that the measurements are normal. In this scenario, high accuracy and specificity are not necessarily indicators of good performance. Additionally, the high relative occurrence of artefacts and the inherent ambiguities in the time-series mean that obtaining a high sensitivity over time is extremely challenging.

3.3.6 Acceptance rates by patient

The acceptance rate is defined as the proportion of MET events which are also labelled tMET events for a given patient. Figure 3.10 shows the distribution of these acceptance rates across the 150 MET patients. Of these, 93 patients present no ambiguity: 53 (or 35%) had no MET event labelled as being tMET, while 40 (or 27%) had all of their MET events labelled as being

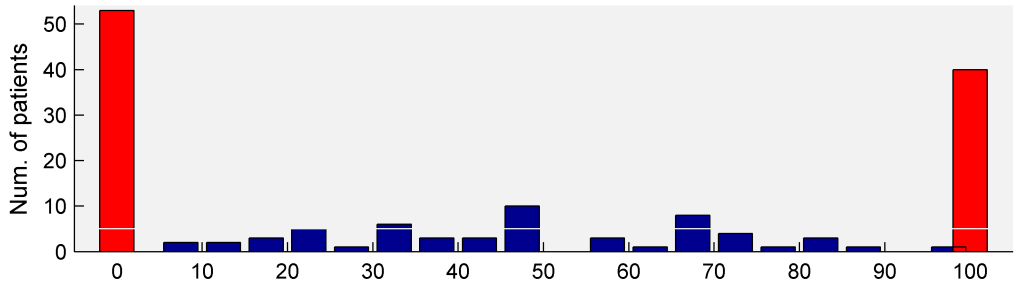


Figure 3.10: Acceptance rates of MET events by patient in the MET population. Blue bars are 5-percent wide, and the leftmost one is centred at 2.5%. The red bars correspond to 0 and 100%.

tMET events. The remaining 57 MET patients have a mixture of tMET and aMET events. Here we examine the events of the three patients at the ends of the acceptance rate spectrum: the two patients for whom the acceptance rate is lower than 15% (but not 0%) and the single patient for whom the acceptance rate is greater than 90% (but not 100%).

The only patient for whom the acceptance rate is greater than 90% is Patient 284. This patient had 81 MET events, 77 of which were labelled tMET (95.06%). A closer look at the 4 remaining aMET events was necessary to ensure that labels for this patient are consistent. The 77 tMET events are caused by a low oxygen saturation, while the 4 aMET events occur within 2 hours of one another, and are all activated by a different vital-sign channel, breathing rate, while no other vital sign is available (Figure 3.11). All four of these events have an aggregate score of 0, meaning that they were unanimously labelled as being aMET by all four reviewers. All of the events with the same activation cause were therefore labelled consistently and it was deemed unnecessary to amend any label.

At the other end of the spectrum, Patient 59 had 15 MET events, all caused by low SpO_2 measurements. Only one event (7%) was labelled as being tMET, with an aggregate score of 3, the lowest possible aggregate score for a tMET event. This event is shown in Figure 3.12, where it may be seen that the event overlaps with a condition C event, which was not deemed to be a C' event, indicating that it was discarded by clinicians being deemed artefactual. While the observed SpO_2 values below 75% are unlikely to be physiologically-accurate measurements, it is not obvious that the trend should be disregarded. There are two other indications that the trend may be physiological: first, less than five minutes after the start of the event, the heart

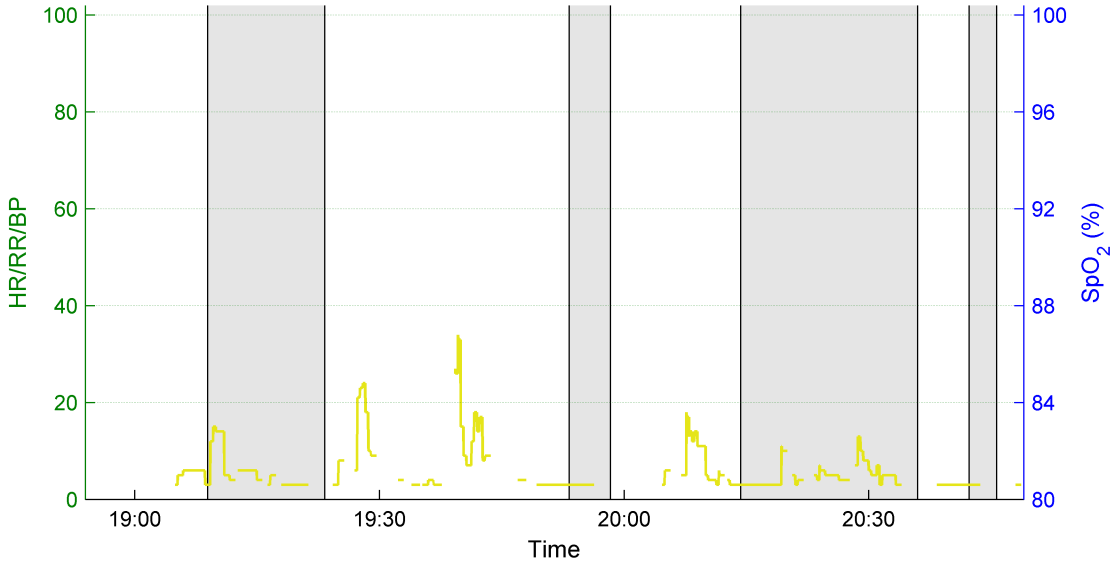


Figure 3.11: Artefactual MET events for patient 284. All four aMET events occur within 2 hours of each other and are caused by an abnormally low breathing rate (yellow line). No other vital sign is available during this period. Measurements of other vital signs restart at 20:50.

rate stabilises around HR= 70 bpm, which could have been achieved by activating an "on-demand" pacemaker (there is further evidence of pacemaker use in the patient's recording); second, shortly before the end of the event, the breathing rate decreases to BR= 15 bpm, which coincides with the first time in 20 minutes that the SpO_2 is measured above 84%.[8] Without additional information about the patient, there is no ground truth to indicate that we should overturn the reviewers' classification.

Patient 118 had 34 MET events, 32 caused by low SpO_2 measurements, one by systolic hypotensive measurements, and one by diastolic hypertensive measurements. Of these, 30 hypoxaemic events and the systolic hypotensive events were labelled aMET. The two remaining hypoxaemic events that were labelled "tMET" present similar situations to that found in the recording of Patient 59, which led to the same conclusion. The diastolic hypertensive event, despite being suspiciously short, was classified unanimously by the reviewers (aggregate score of 4). It is noteworthy that this MET event overlaps with a C event, which was not labelled C' or C".

More generally speaking, acceptance rates are a measure of how ambiguous a patient's recording may be with respect to artefacts. It appears that 93 MET patients were not ambiguous to

[8]The decrease in breathing rate and increase in SpO_2 could indicate that the patient has been given a mask with oxygen-enhanced air to breathe.

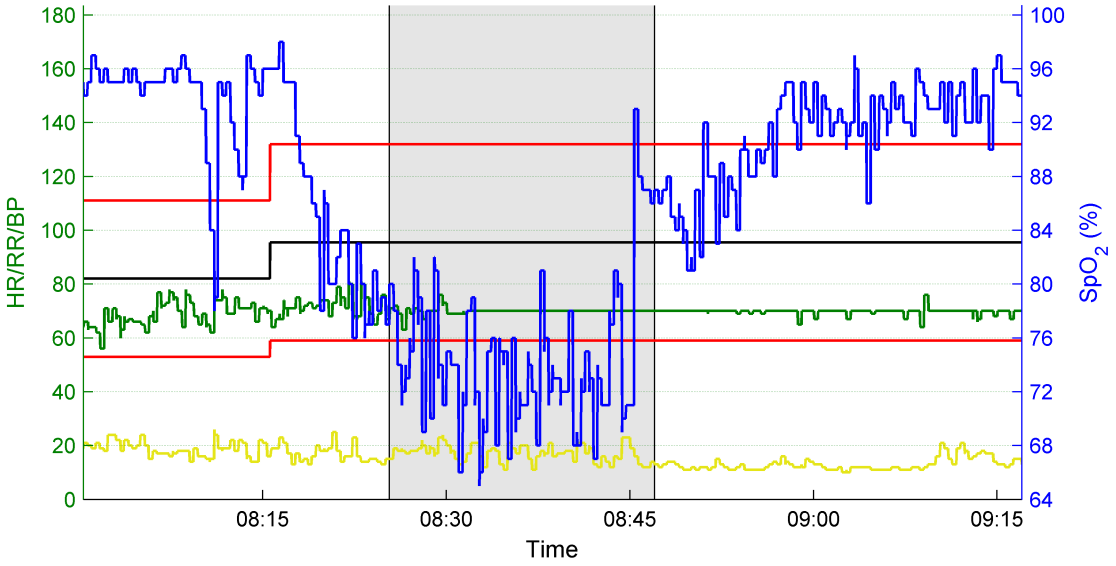


Figure 3.12: True MET event for patient 59.

reviewers, while we showed above that there was a degree of consistency in those cases for which acceptance rates are the most suspicious (i.e., very low or very high). While this analysis is by no means exhaustive, there is reasonable ground to believe that a mid-range acceptance rate is a reflection of the presence of both artefactual and physiological events in a patient's recording.

3.4 Comparison with historical labelling

The set of tMET events has been identified with the same objective as that used for the set of C' events of Hann (2008); i.e., to include all the periods containing non-artefactual abnormal measurements. Similarly, MET events are equivalent to the (unmerged) condition C events. Since there are a number of differences in the procedures used to generate both set of labels, it is not expected that they should be identical. However, comparing the events and quantifying their differences gives us insight into the validity of their respective assumptions, and into the robustness of expert labelling for multivariate vital-sign data.

The single-channel thresholds, durations of zero-order holds, and lengths of activation and de-activation windows used to identify the MET and condition C labels, were the same for both

↱ overlap with	MET	tMET	C	C'	C"
MET	*100*	*100*	100	100	100
tMET	49.07	*100*	64.13	79.41	76.11
C	63.57	86.79	*100*	*100*	*100*
C'	40.84	66.98	58.72	*100*	*100*
C"	20.07	31.84	27.60	47.68	*100*

Table 3.7: Overlap rates between types of events. This table should be read: *events* (first column) overlap with *number*% of *events* (first row) Examples: "MET events overlap with 100% of C" events" and "C" events overlap with 31.84% of tMET events". Values of 100% in italic come from the definitions of event types (e.g. tMET events, being a subset of MET events, overlap with 100% of MET events).

Hann's (2008) and our algorithms. The only difference lies in the fact that MET activations are generated even in the event of missing data (up to 20% in the activation window of MET events) resulting in potentially more MET than condition C activations (862 MET activations versus 564 condition C activations).

As a result of differences in domains of expertise, we expect that, provided with the same data for review, signal processing and clinical experts would annotate events differently. We therefore expect some of the MET events that did overlap with C events not to overlap with any C' event. Reciprocally, some C' events will overlap only with MET events and not with tMET events.

Table 3.7 shows the overlap rates for pairs of event types; i.e., the percentage of events of one type which overlap in time with events of another type. The three rightmost 100% on the table's first row indicate that at least one MET event overlaps with every C and consequently C' and C" events, which confirms the expectation that the procedure used to extract MET events is more conservative. From the first column in the table, we can see that C events overlap with only 64% of MET events, indicating that 36% of MET events were not highlighted by the historical labelling procedure, and therefore never presented for review to clinical experts.

tMET events overlap only with 79% of C' events. Since at least one MET event overlaps with each of the C' events, this means that 21% of C' events were regarded as artefactual events by the signal processing experts, while they were not regarded as such by clinical experts. Conversely, C' events only overlap with 67% of tMET events, meaning that 33% of tMET

events where not regarded as relevant by the historical labelling procedure.

tMET events overlap with 76% of C" events, which is less than the 79% of C' events which overlap with at least one tMET event. The difference between C' and C" events is the retrospective clinical belief that a MET call should have been made. Assuming that the tMET labels form a "ground truth" of periods of "non-artefactual physiological derangements meeting the MET criteria", we would expect the percentage of C" events which overlap with tMET events to be greater than it is for C' events, if only because C" events form the *clinically relevant* subset of C' events, which should include the most obvious physiological derangements. A binomial test, under the null hypothesis that the proportion of C" events which overlap with tMET events is at least that of C' events which do, yields a p-value of 0.20, indicating that the observed difference is not significant at the 0.05-level. The drop in overlap from 79 to 76% between C' events and C" events is however unexpected. Possible factors could be that the vital-sign time-series (at least in the way they are used to generate the labels) are only a part of the total quantity of information taken into account by clinicians. Another explanation could be that persistent exceedances of single-channel thresholds are less insightful than other features of the vital-sign time-series which are not included in either label-identification procedure, yet are integrated into the judgement of clinical experts (e.g., higher order statistics, "trends", etc.). Despite a high correlation between the presence of PSCAs and actual physiological deterioration, a system which aims at reliable detecting physiological deterioration would likely be required to identify and make use of these other features.

3.5 Future use of labels

In the next two chapters, we present algorithms for novelty detection in time-series based on Extreme Value Theory, first in the case of a univariate generating process, then in the more general case of a multivariate generating process. The sets of MET, tMET, and aMET labels are used in Chapters 6 and 7 to assess the characteristics and performance of our proposed novelty detection algorithms. Where possible, we shall also report results on the historical set of labels.

Chapter 4

Univariate Extreme Value Theory for Novelty Detection

After motivating the use of EVT, we list a few of the key results of univariate classical Extreme Value Theory, starting with exact distributions of minima and maxima. These are known to be degenerate when the sample size m tends to infinity, making the exact distributions computationally impractical for larger values of m. However, they are suitably modelled by a family of non-degenerate distributions, as stated by the Fisher-Tippett theorem, which forms the basis of asymptotic EVT. We reproduce some of the key results which follow from the Fisher-Tippett theorem. We then operate a shift in viewpoint, by refining our definition of extremes, to one that we argue is more suited to a novelty detection framework. By considering a "probability distribution of probability densities", under the right assumptions, we transform a potentially complicated (multimodal) distribution, for which the areas of interests may be disconnected into a simpler distribution of probability densities, whose lower tail becomes the only focus for a novelty detection application.

4.1 Motivation

Novelty detection, outlier detection, or one-class classification techniques based on the assumptions that the normal data are drawn from some data-generating distribution, also assume that observations which lie far in the tails of the data-generating distribution may be considered abnormal, or novel. Rejection of abnormal observations is performed using some metric, possibly dependent on the model. Examples of such metrics are absolute magnitude: Mahalanobis distance to the nearest mode of the distribution, or percentiles of the model. In the model used by Hann (2008), an observation is considered abnormal if the total probability mass which lies closer to the centre of the distribution that our observation exceeds a threshold, which is optimised and validated empirically. For the sake of demonstration, let us imagine that the data-generating distribution is univariate and unimodal with cdf F. Hann's method[1] is equivalent to setting a threshold $\alpha \in \left[0, \frac{1}{2}\right]$ such that an observation x is rejected if $F(x) < \frac{\alpha}{2}$ or $F(x) > 1 - \frac{\alpha}{2}$. The choice of the letter α as our threshold is deliberate, as α is indeed the significance level[2] in an instance of hypothesis testing. The null hypothesis is "observation x was generated by the model F". If $F(x)$ is not in $[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}]$ then this is evidence than the null hypothesis can be rejected at the α level of confidence (as we are performing a two-sided test in this example).

This is a theoretically sound framework, if there is only ever one available observation at a time. However, let us now consider a case where sets of observations are drawn in succession. Let us also assume, for the sake of argument, that there is at least one observation such that $F(x) < \frac{\alpha}{2}$. According to this strategy, it should be rejected as being abnormal. Now the intuition is that the more observations are drawn from F, the more likely we are to observe data of very high and low magnitude.

Figure 4.1 demonstrates how the distribution of the extremes of m samples drawn from a standard Gaussian or an exponential distribution evolve as m increases. Eventually, for sufficiently large values of m, we expect with almost certainty to draw observations such that $F(x) < \alpha$,

[1] Although Hann only describes this novelty detection method in terms of pdf-thresholding, the interpretation in terms of hypothesis testing described in this section is strictly equivalent.

[2] i.e., the test's probability of incorrectly rejecting the null hypothesis, also known as false positive rate, or Type I error.

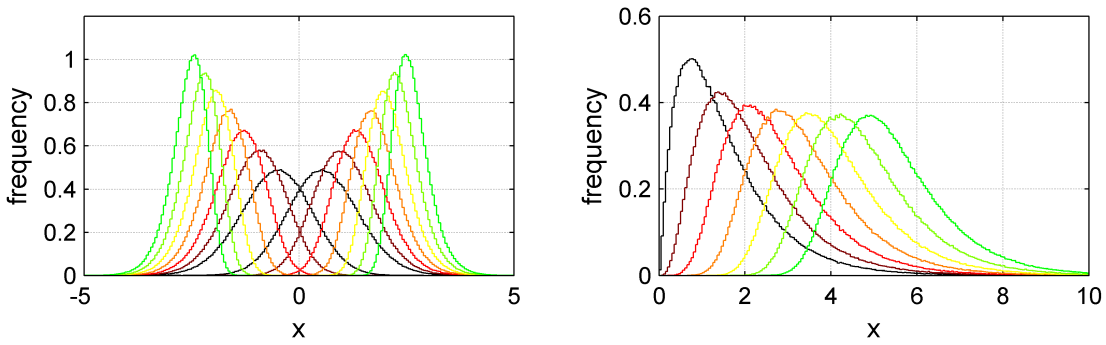


Figure 4.1: Empirical distributions of 10^6 minima and maxima for the standard Gaussian distribution (left), and of maxima for the standard exponential distribution. Values of m are 2 (black), 4, 8, 16, 28, 64 and 128 (green).

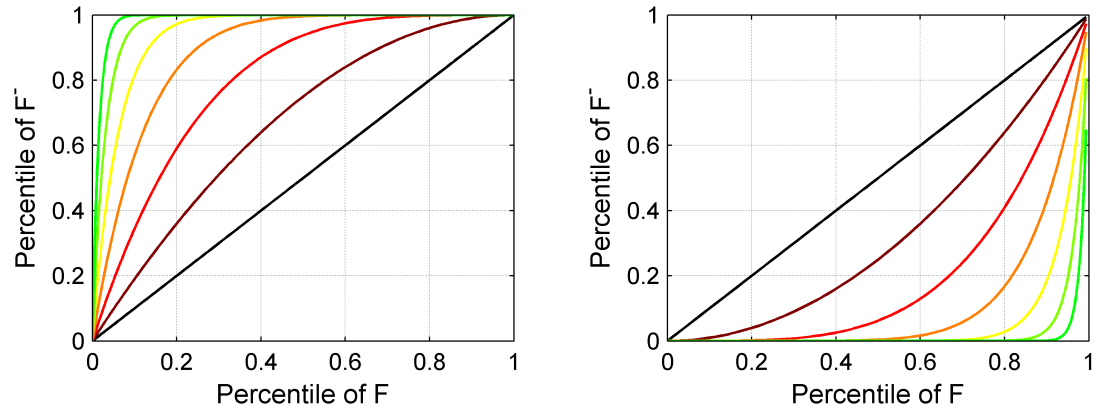


Figure 4.2: P-P plots of cdf vs. EVD percentiles for the standard Gaussian (minima, left) and exponential (maxima, right) distributions. Values of m are 1 (black) , 2, 4, 8, 16, 28, 64, and 128 (green).

which means that the null hypothesis will be falsely rejected at least once almost certainly. Figure 4.2 shows Percentile-Percentile plots[3] for the standard Gaussian and standard exponential distributions against their distributions of minima and maxima as m increases. The x-axis represents the percentiles of F. In the case of the standard Gaussian for instance, as m increases, the probability mass of the distribution of minima which lies below a given percentile of F (e.g. 10%) increases rapidly. For $m = 64$, this mass is nearly unity, which means that, out of 64 i.i.d. observations from m, we are almost guaranteed to draw at least one that lies in the lower 10^{th}-percentile of F, *despite* the i.i.d. nature of the sampling. We would therefore be wrong to reject the null hypothesis for all 64 of these observations at the 0.1 level of confidence. Applications

[3]In a Percentile-Percentile plot, or P-P plot, two distributions are compared by plotting their cdfs against each other. Perfect agreement results in the line of no discrimination. Departures from this line are indicative of a disagreement.

involving time-series such as our vital-sign monitoring problem are typically applications for which we assume samples are drawn repeatedly from the same distribution, therefore we require a theoretical framework which allows us to cope with this repeated sampling. The remainder of this chapter and the next chapter describe our developments of EVT to address this problem.

4.2 Key results of classical Extreme Value Theory

The study of extreme statistics via classical univariate Extreme Value Theory is a well-established field, and the literature on the subject is abundant. In this section we report a number of theoretical results available in the literature with three aims:

- introduce the reader to a number of notations, which will remain consistent throughout this thesis whenever possible;
- build up from intuitive observations to advanced theorems and proofs from the asymptotic theory, while motivating each step along the way;
- report a number of results, some of which may seem out of place here, but will be useful in later stages of this work.

At this point, it is understood that an *extreme* of a sample $(x_1, x_2, \ldots, x_m)$ is either the maximum or the minimum of the m observations, and we note:

Definition 1. *(Minimum and maximum of m observations)*

$$M_m^+ = \max_{i=1:m} (x_i) \tag{4.1}$$

$$M_m^- = \min_{i=1:m} (x_i) \tag{4.2}$$

4.2.1 Exact distribution of minima and maxima

Let X be a univariate random variable with pdf f, cdf F, and support[4] D_f. The joint pdf of an i.i.d. sample drawn from F is:

$$f(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} f(x_i), \tag{4.3}$$

and their cdf is:

$$F(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} F(x_i). \tag{4.4}$$

The probability that the maximum of the sample is less than a value $x \in D_f$ is given by:

$$\begin{aligned} F^+(x;m) &= \Pr\left(\max(X_1, \ldots, X_m) \leq x\right), \\ &= \Pr\left(\forall i,\ X_i \leq x\right), \\ &= \prod_{i=1}^{m} \Pr\left(X_i \leq x\right), \\ &= \prod_{i=1}^{m} F(x), \\ &= \left[F(x)\right]^m. \end{aligned} \tag{4.5}$$

The i.i.d.assumption is required to move from the second to the third line. F^+ is, by definition, the cdf of the maximum of the sample. A straightforward differentiation gives its associated pdf:

$$f^+(x;m) = mf(x)\left[F(x)\right]^{m-1}. \tag{4.6}$$

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

$$F^-(x;m) = 1 - \left[1 - F(x)\right]^m, \tag{4.7}$$

$$f^-(x;m) = mf(x)\left[1 - F(x)\right]^{m-1}. \tag{4.8}$$

Figure 4.3 shows the exact extreme value distributions (EVDs) associated with the minima of the standard Gaussian distribution ($f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$, $F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$, $\forall x \in \mathbb{R}$),

[4] The support of the definition is the set of points where the distribution's pdf is not zero.

and an excellent agreement with numerical sampling. Figure 4.4 is the equivalent figure for the maxima of the standard exponential distribution ($f(x) = \exp(-x)$, $F(x) = 1 - \exp(-x)$, $\forall x \in \mathbb{R}^+$). In all cases, the exact EVDs fit the empirical distributions perfectly.

In the limit $m \to \infty$, these distributions present two disadvantages. Firstly, they become degenerate[5] when m tends to infinity. At the limit, F^+ is zero on all of D_f, and 1 at the supremum of D_f. Secondly, they are not practical for engineering purposes at even moderate values of the parameter m. Since values of F are in general less than 1, the m^{th} power can decrease rapidly to or be below machine precision.[6] When working with small values of m, it is in our interest to use exact EVDs. When this becomes impractical for large values of m, we must resort to using the EVDs given by asymptotic EVT, which is described in the next section.

4.2.2 Asymptotic EVT

The Fisher-Tippett theorem

Classical univariate Extreme Value Theory is based on key results from Fisher & Tippett (1928): consider a set $X_m = (x_1, x_2, \ldots, x_m)$ of m independent and identically distributed random variables $x_i \in \mathbb{R}$ (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:

$$M_m^+ \to c_m.X + d_m \tag{4.9}$$

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 distributions over M_m^+:

Theorem 1. *(Fisher-Tippett theorem - Theorem 3.2.3 in (Embrechts et al. 1997), p.121)*

[5] A degenerate distribution is one whose support consists of only one value.

[6] If we are interested in values of the cdf as low as 10^a, $a < 0$, and the smallest number that can be represented by our machine is 2^b, $b < 0$, then the exact EVDs are guaranteed to take values below the precision of the machine for m greater than $\frac{b\ln(2)}{a\ln(10)}$. For the simulations in this thesis, b is of the order of -1000, which means that for $a = -3$, values of m greater than 100 are prohibited.

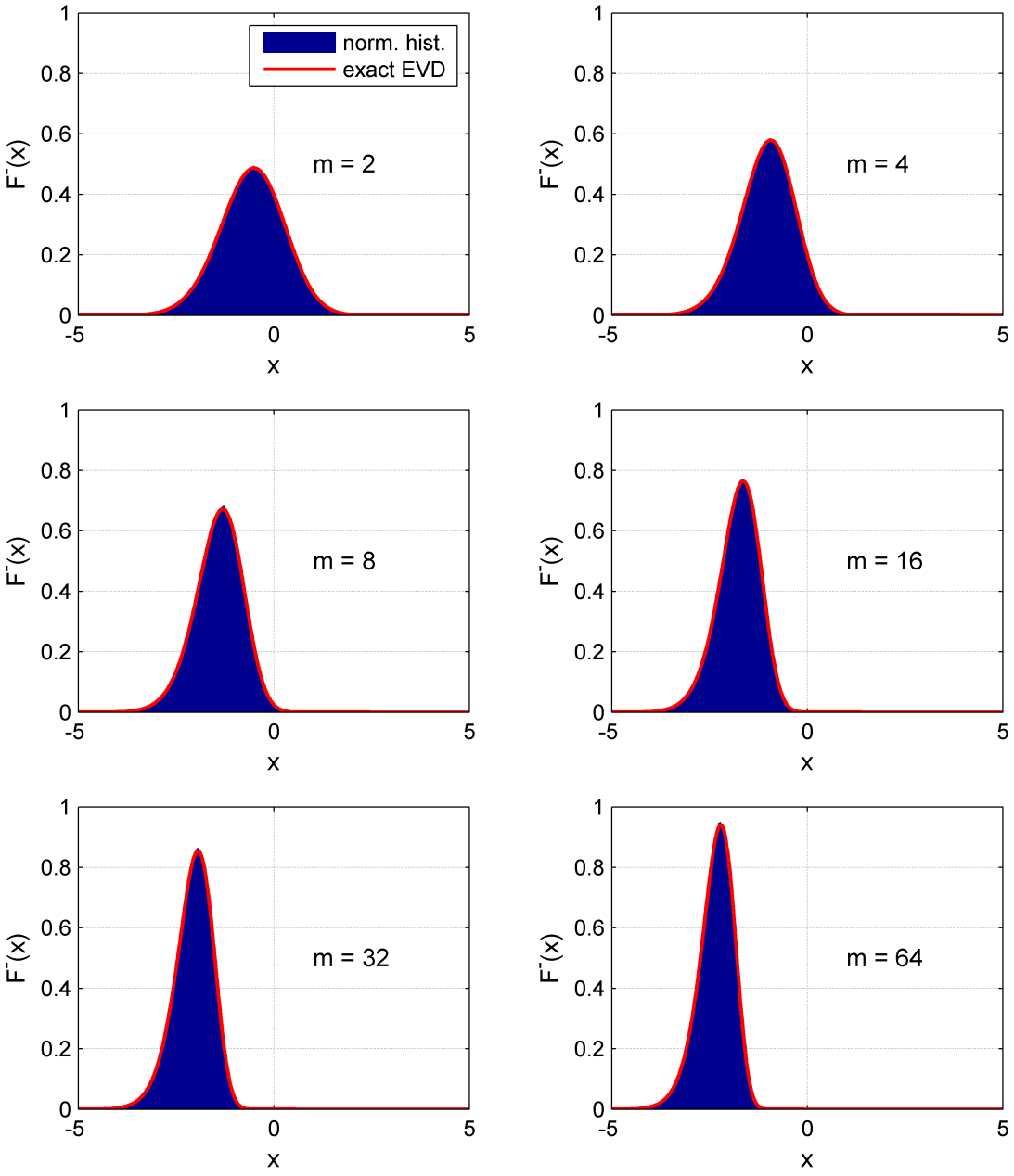


Figure 4.3: Empirical distributions of minima (10^6) and exact EVDs of minima (red) of the standard Gaussian distribution for various values of m. The exact EVDs fit the empirical distributions closely for all values of m.

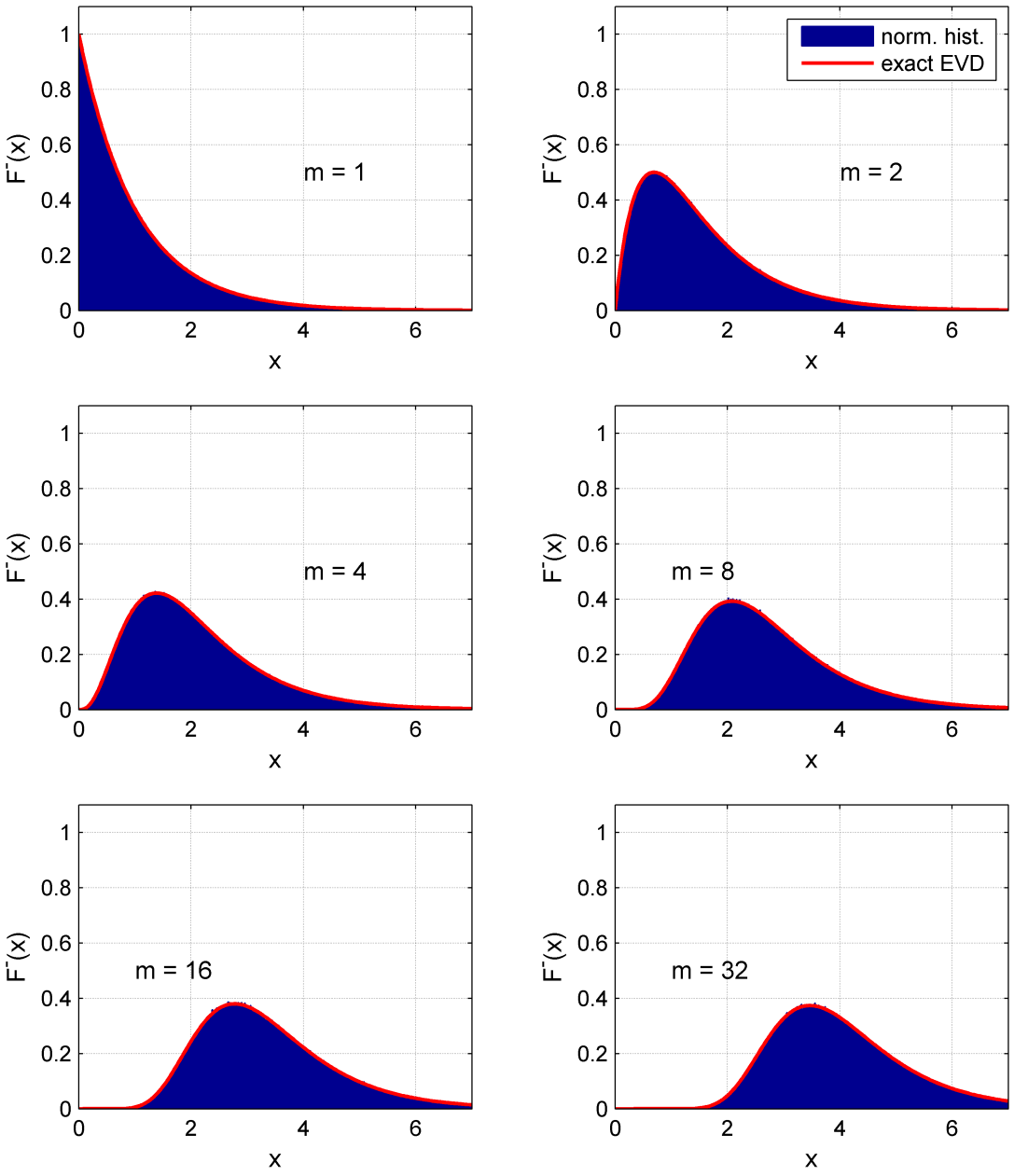


Figure 4.4: Empirical distributions of minima (10^6) and exact EVDs of maxima (red) of the standard exponential distribution for various values of m. The exact EVDs fit the empirical distributions closely for all values of m.

Let (X_m) be a sequence of i.i.d. random variables. If there exist norming constants $d_m \in \mathbb{R}$, $c_m \in \mathbb{R}^+$ and some non-degenerate distribution function H such that

$$c_m^{-1}(M_m^+ - d_m) \xrightarrow{d} H, \tag{4.10}$$

where M_m^+ is a maximal random variable, then H belongs to the type of one of the following three distribution functions:

$$\textit{Gumbel}, \quad H_1^+(x) = \exp(-\exp(-x)) \tag{4.11}$$

$$\textit{Fréchet}, \quad H_2^+(x) = \begin{cases} 0 & \textit{if } x \leq 0 \\ \exp(-x^{-\alpha_m}) & \textit{if } x > 0 \end{cases} \tag{4.12}$$

$$\textit{Weibull}, \quad H_3^+(x) = \begin{cases} \exp(-(-x)^{\alpha_m}) & \textit{if } x \leq 0 \\ 1 & \textit{if } x > 0 \end{cases} \tag{4.13}$$

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

The symbol $\xrightarrow{d}$ denotes a convergence in distribution, or weak convergence[7]. The superscript "+" refers 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. All three are special cases of the Generalised Extreme Value Distribution:

$$H^+(x) = \begin{cases} \exp\left\{-(1+\xi x)^{-1/\xi}\right\} & \text{, if } \xi \neq 0 \\ \exp\{-\exp(-x)\} & \text{, if } \xi = 0. \end{cases} \tag{4.14}$$

The H_1^+, H_2^+, and H_3^+ distributions are the cases $\xi = 0$, $\xi > 0$ and $\xi < 0$, respectively (with necessary adjustments to the norming parameters.).

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 x-axis. The minimal Weibull

[7] A distribution F convergences in distribution, convergence weakly, to a distribution H, if the limit of $F(x)$ is $H(x)$ for every x at which H is *continuous*.

distribution in particular will be of particular interest for this work:

$$H_3^-(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp\left(-(x)^{\alpha_m}\right), & x \geq 0 \end{cases}. \tag{4.15}$$

Extreme value distributions and extremal random variables are defined as follows:

Definition 2. *(Extreme Value distribution and extremal random variables - Definition 3.2.6 in (Embrechts et al. 1997), p.124)*

The distribution functions H_1^+, H_2^+, H_3^+ as presented in Theorem 1 are called standard extreme value distributions*, and the corresponding are random variables* standard extremal random variables*. Distribution functions of the types of H_1^+, H_2^+, H_3^+ are* extreme value distributions*; the corresponding random variables are the* extremal random variables*.*

Crucially, extreme value distributions are only dependent on the parameters of the generative distribution, and m, the size of the sample from which the extremum is taken. Theorem 1 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 is not one, but three limit forms. Theorem 1 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 3. *(Maximum domain of attraction - Definition 3.3.1 in (Embrechts et al. 1997), p. 128)*

We say that the random variable X (the distribution function F of X, or the distribution of X) belongs to the maximum domain of attraction of the extreme value distribution *H if there exist constants $d_m \in \mathbb{R}$ and $c_m \in \mathbb{R}^+$, such that:*

$$\frac{1}{c_m}(M_m^+ - d_m) \xrightarrow{d} H. \tag{4.16}$$

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

The following proposition offers a characterisation of the MDA:

Proposition 1. *(Characterisation of MDA(H) - Proposition 3.3.2 in (Embrechts et al. 1997), p. 129)*

The distribution function F belongs to the maximum domain of attraction of the extreme value distribution H with norming constants $d_m \in \mathbb{R}$, $c_m > 0$, if and only if

$$\lim_{m\to\infty} m\overline{F}(c_m x + d_m) = -\ln(H(x)),\ x \in \mathbb{R}. \tag{4.17}$$

When $H(x) = 0$, the limit is interpreted as ∞.

$\overline{F} = 1 - F$ is the righthand tail of F (also known as the survival function.).

Tail-equivalence

A useful concept when trying to determine the MDA of a distribution is that of tail-equivalence:

Definition 4. *(Tail-equivalence - Definition 3.3.3 in (Embrechts et al. 1997), p.129)*

Two distribution functions F and J are called tail-equivalent *if they have the same right endpoint, i.e. if $x_F = x_J$, and if there exists some constant $0 < c < \infty$, such that*

$$\lim_{x\uparrow x_F} \overline{F}(x)/\overline{J}(x) = c.$$

We note $F \overset{t}{\sim} J$

In other words, if F and J are proportional near their right endpoint, they belong to the same MDA, where the right endpoint is defined as $x_F = \sup\{x \in \mathbb{R}/F(x) < 1\}$. x_F can therefore be a finite value, or $+\infty$.

Maximum domain of attraction and norming constants

We report here the theorems given by Embrechts et al. (1997) which characterise each MDA and provide formulae for the corresponding norming constants.

Theorem 2. *(Maximum domain of attraction of H_1^+ - Theorem 3.3.26 in (Embrechts et al. 1997), p.142)*

The distribution function F with right endpoint $x_F \leq \infty$ belongs to the maximum domain of attraction of H_1^+ if and only if there exists some $z < x_F$ such that F has representation

$$\overline{F}(x) = c(x) \exp\left\{ -\int_z^x \frac{g(t)}{A(t)} dt \right\}, \qquad z < x < x_F, \tag{4.18}$$

where c and g are measurable functions satisfying $c(x) \to c > 0$, $g(x) \to 1$ as $x \uparrow x_F$, and $A(x)$ is a positive, absolutely continuous function with density $a(x)$ having $\lim_{x \uparrow x_F} a(x) = 0$. For F with representation 4.18, we can choose the norming constants as

$$d_m = F^{\leftarrow}(1 - m^{-1}), \text{ and } c_m = A(d_m).$$

A possible choice for the function a is

$$A(x) = \int_x^{x_F} \frac{\overline{F}(t)}{\overline{F}(x)} dt, \qquad x < x_F. \tag{4.19}$$

$F^{\leftarrow}(q)$ denotes the q^{th} percentile of F. A special case occurs when c is a constant and $g(t) = 1$. In this case, F is called a von Mises function, with auxiliary function a. Von Mises functions have interesting properties for the derivation of norming constants, as we demonstrate in Section 4.2.3.

Theorem 3. *(Maximum domain of attraction of H_2^+ - Theorem 3.3.7 in (Embrechts et al. 1997), p.131)*

The distribution function F belongs to the maximum domain of attraction of H_2^+, $\alpha > 0$, if and only if there exists some slowly varying function L such that $\overline{F}(x) = x^{-\alpha} L(x)$. If $F \in MDA\left(H_2^+\right)$, then

$$\frac{1}{c_m}(M_m^+ - d_m) \xrightarrow{d} H_2^+, \tag{4.20}$$

where the norming constants can be chosen as $d_m = 0$ and $c_m = (1/\overline{F})^{\leftarrow}(m)$.

L is a slowly varying function at $+\infty$; i.e., a positive function which obeys (Karamata 1933):

$$\forall t \in \mathbb{R}_+^*, \lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1. \tag{4.21}$$

Theorem 4. *(Maximum domain of attraction of H_3^+ - Theorem 3.3.12 in (Embrechts et al. 1997), p.135)*

The distribution function F belongs to the maximum domain of attraction of H_3^+, $\alpha > 0$, if and only if $x_F < \infty$ and there exists some slowly varying function L such that $\overline{F}(x_F - x^{-1}) = x^{-\alpha}L(x)$.

If $F \in MDA\left(H_3^+\right)$, then

$$\frac{1}{c_m}(M_m^+ - d_m) \xrightarrow{d} H_3^+, \tag{4.22}$$

where the norming constants can be chosen as $d_m = x_F$ and $c_m = x_F - F^{\leftarrow}(1 - m^{-1})$.

As a consequence of these three theorems, we have the following properties:

- Every $F \in \text{MDA}\left(H_2^+\right)$ has an infinite right endpoint: $x_F = \infty$.
- Every $F \in \text{MDA}\left(H_3^+\right)$ has a finite right endpoint: $x_F < \infty$.
- The Gumbel case is the ambiguous one: every $F \in \text{MDA}\left(H_1^+\right)$ can have either a finite or infinite endpoint: $x_F \leq \infty$. Perhaps a more interesting way of thinking of $\text{MDA}\left(H_1^+\right)$ is that it consists of distribution functions whose right tails decrease to zero faster than any power function.

Finally, we note that every maximum domain of attraction is closed with respect to tail-equivalence[8]. Moreover, for any two tail-equivalent distributions, one can take the same norming constants (closure properties of MDAs, also from (Embrechts et al. 1997)). Table 4.1 gives the attractor of some common distributions.

4.2.3 Derivation of the norming constants for the standard Gaussian

In this section, we propose to derive the norming constants of the H_1^+ for the distribution of maxima of the Gaussian distribution, with general mean and variance. This is performed to illustrate the process by which we identify these constants.

We first prove that the Gaussian distribution is a Von Mises function, and is therefore in the MDA of the Gumbel distribution. We proceed to find a tail-equivalent distribution, which en-

[8] i.e., if two distributions are tail-equivalent, then they belong to the same domain of attraction.

EVT Type	Distribution	Probability density function
Fréchet	Cauchy	$f(x) = \left(\pi(1+x^2)\right)^{-1}$, $x \in \mathbb{R}$
	Pareto	$f(x;k;x_m) = kx_m^k x^{-k-1}$, $x_m, k > 0$, $x > x_m$
	Loggamma	$f(x;\alpha,\beta) = \frac{\alpha^\beta}{\Gamma(\beta)}(\ln x)^{\beta-1}x^{\alpha-1}$, $x > 1$, $\alpha,\beta > 0$
Weibull	Uniform	$f(x;a,b) = (b-a)^{-1}$, $a < x < b$
	Beta	$f(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$, $0 < x < 1$, $a,b > 0$
Gumbel	Exponential	$f(x;\lambda) = \lambda\exp(-\lambda x)$, $\lambda, x > 0$
	Weibull	$f(x;\lambda) = \exp(-(-x)^\lambda)$, $x < 0$, $\alpha > 0$
	Gamma	$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$, $x > 0$, $\alpha,\beta > 0$
	Gaussian	$f(x;\mu,\sigma) = (2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/2\sigma^2}$, $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$
	Lognormal	$f(x;\mu,\sigma) = (2\pi\sigma^2x^2)^{-1/2}e^{-(\ln x-\mu)^2/2\sigma^2}$, $x > 0$, $\mu \in \mathbb{R}$, $\sigma > 0$

Table 4.1: Type of extreme value distribution for the maxima of common distributions (Embrechts *et al.* 1997).

ables us to derive a formula for the location parameter. The scale parameter is easily deduced from the previous steps using Theorem 2.

The pdf and cdf of the Gaussian distribution with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ are given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{4.23}$$

$$F(x) = \frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right),\right) \tag{4.24}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\mathrm{d}t$ is the so-called *error function*. Example 3.3.23 and proposition 3.3.25 in (Embrechts et al. 1997) provide a strategy to prove that F is in the MDA of the Gumbel distribution. We start by showing that F is a von Mises function, having noted that von Mises funciton are a special case in Theorem 2. To this end, we only need to show that $\lim_{x\uparrow x_F}\overline{F}(x)f'(x)/f(x)^2 = -1$ (Embrechts et al. 1997). Calling A its auxiliary function, the norming constants can be taken to be: $d_m = F^{\leftarrow}(1-m^{-1})$ and $c_m = a(d_m)$.

We start with Mill's ratio:

$$\begin{aligned}\frac{1-F(x)}{f(x)} &= \frac{\sqrt{2\pi\sigma^2}}{2}\left(1-\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\\ &= \frac{\sqrt{2\pi\sigma^2}}{2}\operatorname{erfc}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\end{aligned} \tag{4.25}$$

where $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty \mathrm{e}^{-t^2}dt$ is the *complementary error function*[9]. A Taylor expansion of erfc when $x \to +\infty$ yields:

$$\begin{aligned}\frac{1-F(x)}{f(x)} &= \frac{\sqrt{2\pi\sigma^2}}{2}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\left(\frac{\sigma\sqrt{2}}{(x-\mu)\sqrt{\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)(1+\mathrm{o}(1))\right),\\ &= \frac{\sigma^2}{x-\mu}+\mathrm{o}\left(\frac{1}{x}\right).\end{aligned} \tag{4.26}$$

We can therefore write:

$$1-F(x) = \frac{\sigma^2 f(x)}{x-\mu} = \frac{\sigma}{\sqrt{2\pi}(x-\mu)}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)+\mathrm{o}\left(\frac{1}{x}\right). \tag{4.27}$$

Furthermore, $f'(x) = -\frac{x-\mu}{\sigma^2}f(x)$ is negative near the right endpoint ($f(x) > 0$, $x_\mu > 0$), which means:

$$\lim_{x\to+\infty}\frac{(1-F(x))f'(x)}{f^2(x)} = -1. \tag{4.28}$$

Thus F is a von Mises function (with auxiliary function A) and, as such, is in the MDA of the Gumbel distribution.

To calculate the norming constants, we start from Equation 4.26:

$$\overline{F} \simeq \frac{\sigma^2 f(x)}{x-\mu} = \frac{\sigma}{\sqrt{2\pi}(x-\mu)}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x\to\infty \tag{4.29}$$

and interpret the right-hand side as the tail of some distribution function J. Then F and J have the same norming constants. According to Theorem 2, $d_m = J^{\leftarrow}(1-m^{-1})$. From Proposition

[9] i.e., $\operatorname{erfc}(x) = 1-\operatorname{erf}(x)$.

1, we look for a solution of $-\ln\overline{J}(d_m) = \ln m$; i.e a solution of:

$$\begin{aligned}
\ln m &= \frac{1}{2\sigma^2}(d_m-\mu)^2 + \ln(d_m-\mu) + \frac{1}{2}\ln 2\pi - \ln\sigma, \\
(d_m-\mu)^2 &= 2\sigma^2\left(\ln m - \ln(d_m-\mu) - \frac{1}{2}\ln 2\pi + \ln\sigma\right), \\
&= 2\sigma^2 \ln m\left(1 + \frac{-\ln(d_m-\mu) - \frac{1}{2}\ln 2\pi + \ln\sigma}{\ln m}\right).
\end{aligned} \tag{4.30}$$

Taking the positive root of the square, we can write the following Taylor expansion:

$$\begin{aligned}
(d_m-\mu) &= \sigma\sqrt{2\ln m}\left(1 + \frac{-\ln(d_m-\mu) - \frac{1}{2}\ln 2\pi + \ln\sigma}{\ln m}\right)^{1/2}, \\
&= \sigma\sqrt{2\ln m}\left(1 + \frac{-\ln(d_m-\mu) - \frac{1}{2}\ln 2\pi + \ln\sigma}{2\ln m} + o\left(\frac{1}{\ln m}\right)\right), \\
&= \sigma\sqrt{2\ln m} + \sigma\frac{-\ln(d_m-\mu) - \frac{1}{2}\ln 2\pi + \ln\sigma}{\sqrt{2\ln m}} + o\left(\frac{1}{\sqrt{\ln m}}\right), \\
&= \sigma\sqrt{2\ln m} + \sigma\frac{-\ln\left(\sigma\sqrt{2\ln m}\right) - \frac{1}{2}\ln 2\pi + \ln\sigma}{\sqrt{2\ln m}} + o\left(\frac{1}{\sqrt{\ln m}}\right), \\
&= \sigma\sqrt{2\ln m} - \sigma\frac{\ln(\ln m) + \ln 4\pi}{2\sqrt{2\ln m}} + o\left(\frac{1}{\sqrt{\ln m}}\right),
\end{aligned} \tag{4.31}$$

from which we deduce:

$$d_m = \sigma(2\ln m)^{1/2} + \mu - \sigma\frac{\ln\ln m + \ln 4\pi}{2(2\ln m)^{1/2}} + o\left((\ln m)^{-1/2}\right). \tag{4.32}$$

Since we can take $A(x) = \frac{1-F(x)}{f(x)}$, we have $A(x) \sim \frac{\sigma^2}{(x-\mu)}$ and therefore, first order approximations of the norming constants are given by:

$$d_m = \sigma(2\ln m)^{1/2} + \mu - \sigma\frac{\ln\ln m + \ln 4\pi}{2(2\ln m)^{1/2}} \tag{4.33}$$

$$c_m = a(d_m) \simeq \frac{\sigma}{\sqrt{2\ln m}}. \tag{4.34}$$

Figure 4.5 shows the fit between the exact EVDs and the Gumbel distributions with norming constants given by Equations 4.33 and 4.34. As we can see from the top left graph, H_1^+ is

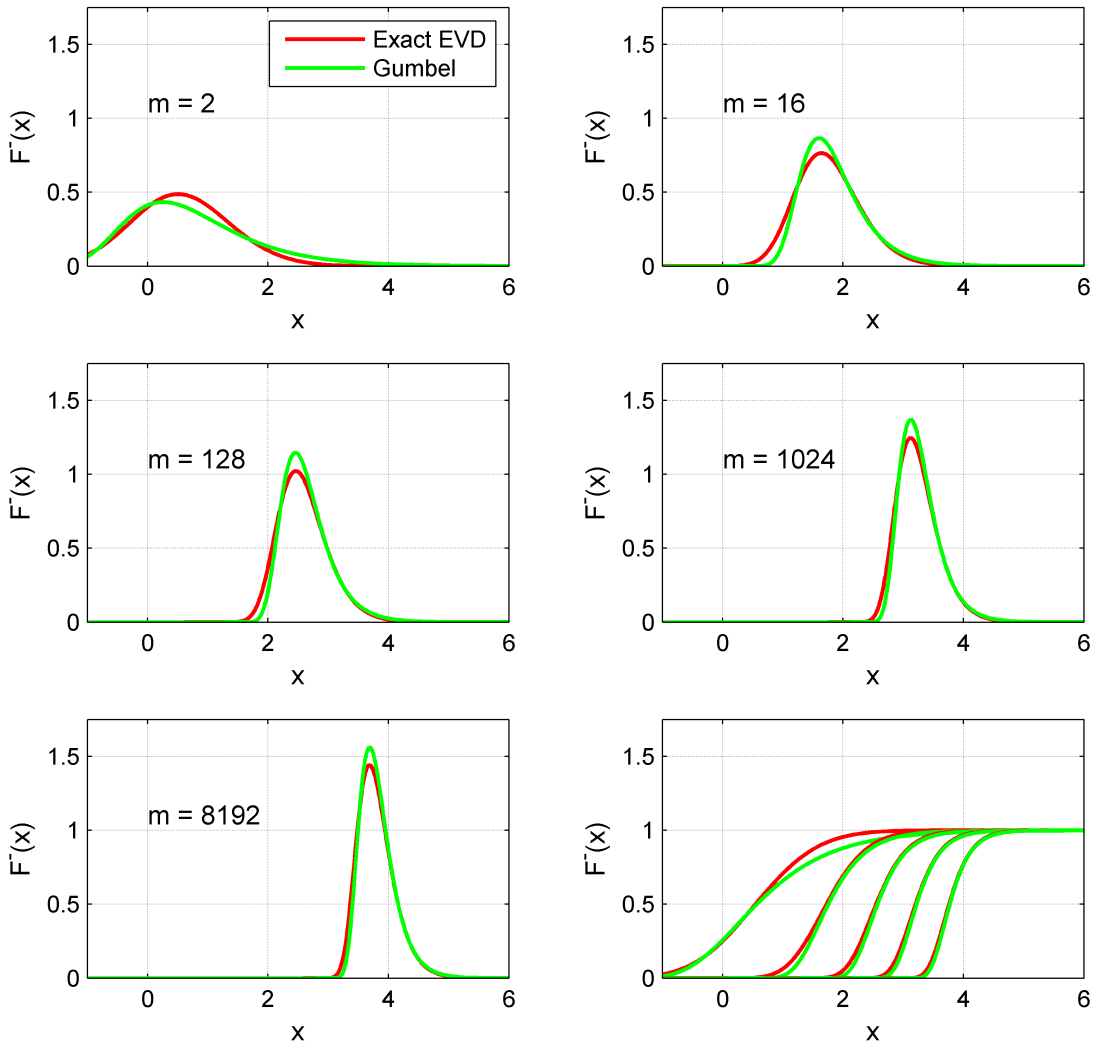


Figure 4.5: Exact EVDs (red) vs. Maximal Gumbel (green) distributions for the standard Gaussian. The bottom right graph shows the cdfs of the pdfs represented in the other five graphs. As m is increased, the Gumbel distributions become better fit to the Exact EVDs, until the cdfs become undistinguishable.

a bad fit for low values of m but becomes an increasingly better fit as m is increased. The difference becomes visually negligible in the bottom right graph, which shows the cdfs for all 5 values of m. Figure 4.6 shows the evolution of c_m and d_m as a function of m over a logarithmic scale. The location parameter d_m, grows in a sub-logarithmic fashion, while the scale parameter c_m decreases, indicating a "peakier" distribution, with lower variance as m increases. Put together, these two observations indicate that maxima move rather slowly away from the centre distribution, perhaps slower than is intuitive. As an example, there is less than a $1 : 10^3$ chance that the maximum of a 10^9 observations drawn from the standard Gaussian exceeds $x = 7.0$.

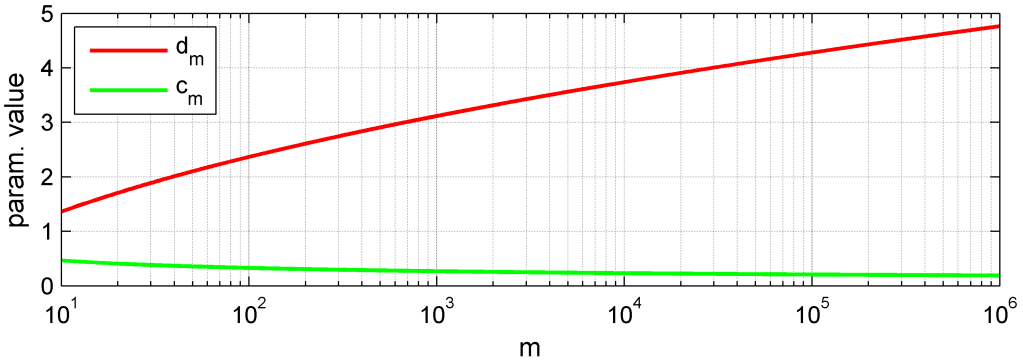


Figure 4.6: Maximal Gumbel parameters for the standard Gaussian as a function of m.

4.3 Extension to the multimodal case

4.3.1 Limitations of EVT

The theoretical results of Section 4.2.2 are insufficient to underpin a general EVT approach to novelty detection. Extremes have so far been considered to be extremes in magnitude, which means that EVT is exclusively concerned with abnormal departures from the *modes* of the data-generating distribution, near the limit of its support. While this is useful to analyse weather, tidal or insurance claim data, such abnormalities are not necessarily those in which we are interested in engineering applications, such as fault detection or network activity monitoring. Although small and large values will often be of particular interest, this is not always the case, and the focus of an EVT novelty detection scheme should be on *improbable* states, whether of large magnitudes or not; i.e, any state in which the system under consideration should *not* be in, and which may occur even when data take values that are not extreme in absolute magnitude. Figure 4.7 shows two examples of situations for which classical EVT is not suitable for novelty detection. In the case of the Beta distribution, improbable values of the support lie in its central zone and tend to accumulate around $x = 0.5$ as m increases. If we are monitoring improbable states, then clearly we have no interest in the tails, near 0 or 1, of the distribution. In the case of the mixture of two Gaussians, improbable observations will lie in both tails, but also, for smaller values of m, between the two modes of the distribution. As the pdf takes higher values there than in the tails, we will eventually lose interest in this area for larger values of m, and the distribution of extremes will look more and more like the average of the distribution of minima and the

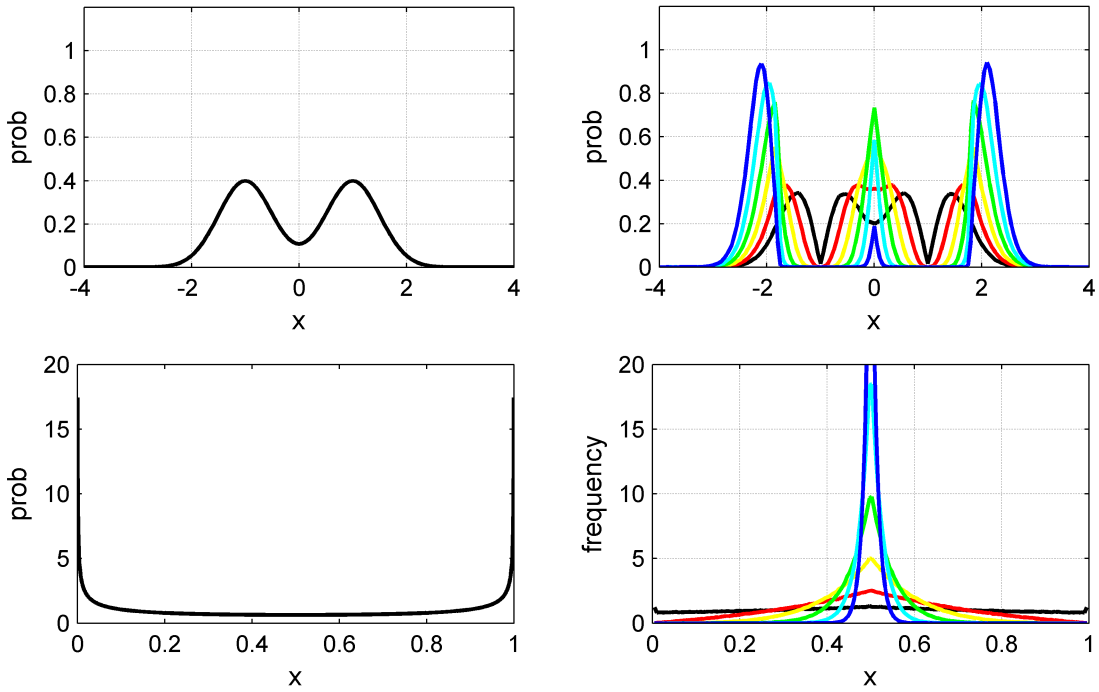


Figure 4.7: Examples of distributions for which extremes in magnitude differ from extremes in probability densities. The two generative distributions are a mixture of two Gaussians with equal variances, $\sigma^2 = \frac{1}{4}$, and centres $\mu = -1$ and 1 (top row) and a Beta distribution with parameters $(\alpha, \beta) = (0.5, 0.5)$ (bottom row). Generative distributions are shown on the left. On the right, we show the normalised histograms of 10^6 extremes in probability density, where the values of m range from 2^1 (black) to 2^6 (blue).

distribution of maxima. In addition, multivariate extensions to classical EVT theory, which we address in Chapter 5, usually involve heuristic approximations, or cumbersome definitions of multivariate extremes in the multivariate case. All of this motivates the change in viewpoint which is adopted in the next section, and which we shall retain in the rest of this thesis.

4.3.2 Redefining extremes

To address the limitations of classical univariate EVT highlighted in Section 4.3.1 and Figure 4.7, we propose a new definition of the extremes of a sample of m observations.

Definition 5. *Extremes in probability density.*

Let[10] $m \in \mathbb{N}^*$ *and* $\{X_m\}$ *be a sequence of (possibly multivariate) i.i.d. random variables, drawn from a distribution* F *with probability density function* f*. We define the extremum to be the random variable* $E_m = argmin_{i=1:m}\{f(X_i)\}$*. We define* e_m *to be an observation of* E_m*, and term* e_m *an* ***m-extreme****.*

This definition of extreme can be related to Definition 1 in the following ways:

- 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;
- Selecting the most improbable sample with respect to f is equivalent to selecting the sample of minimal magnitude with respect to the distribution over $f(E_m)$. Since f takes values in $\mathbb{R}^+$, the distribution over $f(E_m)$ is always univariate, regardless of the dimensionality of f, and the results reported in Section 4.2 remain valid, assuming this distribution has good properties.

Additionally, this definition encompasses extremes in probability densities which lie between the distribution modes, and is therefore better suited to problems where all improbable events are of interest. The next step is to characterise the distribution over $f(E_m)$, which will allow us to describe the newly defined extremes.

[10] $\mathbb{N}^*$ is the set of strictly positive integers.

4.3.3 Distribution of probability densities

Let x be a sample drawn from a (possibly multivariate) distribution with cdf F, pdf f, and support D_f.We term *probability space* the image of D_f under f; i.e, $P_f = f(D_f)$. The likelihood of obtaining a specific value y in the probability space by drawing a sample in the data space is entirely defined by the form of f. Consequently, if F is the distribution over the random variable X, there exists a distribution G over the random variable $Y = f(X)$. Drawing one sample from F implies equivalently drawing a unique sample from[11] G. G is defined as follows:

Definition 6. *Distribution of probability densities (DPD)*
Let X be a (possibly multivariate) random variable with cdf F, pdf f, support D_f, and probability space P_f. $Y = f(X)$ is a random variable, distributed according to:

$$\begin{aligned} \forall y \in P_f, \quad G(y;f) &= Pr(Y \leq y) \\ &= Pr(X \leq y) \\ &= \int_{f^{-1}(]y_{\min},y])} f(x)dx, \end{aligned} \tag{4.35}$$

where $y_{\min} = Inf(P_f)$ and $f^{-1}(]y_{\min},y])$ is the preimage[12] of $]y_{\min},y]$ under f.

The distribution $G(.;f)$ is non-degenerate, as long as f takes more than one value (i.e., f is not the uniform distribution on D_f). If f is continuous, then $G(.;f)$ is continuous[13] and differentiable. From Definition 5 and Equation 4.35, we have:

$$\begin{aligned} f(E_m) &= f(\text{argmin}_{i=1:m}\{f(X_i)\}) \\ &= \min_{i=1:m}\{f(X_i)\} \\ &= \min_{i=1:m}\{Y_i\}, \end{aligned}$$

which shows that $f(E_m)$ is a minimum of $G(.;f)$ and we have the proposition:

Proposition 2. *Distribution over $f(E_m)$.*

[11] However, drawing from G is not equivalent to drawing from F if f is not bijective over D_f and several elements of D_f have the same image under f (a bijective function, or bijection is a one-to-one correspondence between two sets.).

[12] Let f be a function from A to B. The preimage of set $C \subseteq B$ under f is defined by $f^{-1}(C) = \{x \in A | f(x) \in C\}$

[13] In fact, f only needs to be *piece-wise continuous* for $G(.;f)$ to be continuous

The distribution over $f(E_m)$ is the minimal extreme value distribution of $G(.;f)$ with sample size m, which we note $G^-(.;f,m)$.

Given a pdf f and $m \in \mathbb{N}^*$, $G^-(.;f,m)$ will be either the exact minimal distribution of $G(.;f)$, or one of the asymptotic EVDs given in Theorem 1, provided that we can identify the MDA to which $G(.;f)$ belongs, and find appropriate approximations of the corresponding norming constants. We note that according to the remarks at the end of Section 4.2.2, the left endpoint of the support of $G(.;f)$ is always finite (f cannot take negative values). $G(.;f)$ is therefore in the MDA of either the minimal Gumbel H_1^- or the minimal Weibull H_3^-.

4.4 Novelty detection with distributions of probability densities

4.4.1 Hypothesis testing with sets of observations

Let F be a (possibly multivariate) distribution function, and let f, D_f, P_f, $G(.;f)$, and $G^-(.;f,m)$ be its pdf, support, probability space, DPD, and minimal EVT of the DPD, respectively. Given a sample of m i.i.d observations $(x_1,\ldots,x_m)$:

- $f(E_m)$ is distributed according to $G^-(.;f,m)$;
- for all $y \in D_f$, $G^-(y;f,m)$, as a percentile of the distribution, gives the probability of drawing a sample of size m from G^- whose minimum is smaller than y;
- Defining e_m to be the extreme of $(x_1,\ldots,x_m)$, $G^-(f(e_m);f,m)$ can be interpreted as the probability of drawing m observations from F, whose extreme is less than e_m;
- The previous assertion can be interpreted as an instance of one-sided hypothesis testing, where the null hypothesis is now: "The extreme of m observations with respect to f is consistent with that of an i.i.d. sample of size m drawn from f". $G^-(f(e_m);f,m)$, as the probability of drawing a smaller extreme, is in this case the p-value, and the null hypothesis may be rejected at the α-level of significance if $G^-(f(e_m);f,m) < \alpha$;

Algorithm 4.1 Univariate Novelty Detection with EVT.

Let $m \in \mathbb{N}^*$ be the typical size of observed samples. Let F be the assumed data-generating model for the system of interest.

Initialisation:

1. Compute (or estimate) $G(.;f)$ and $G^-(.;f,m)$, which can be either an exact or an asymptotic EVD;
2. pick a value of the threshold α in [0,1].

Unseen samples: Given an unseen sample, $(x_1,\ldots,x_m)$

1. Compute $(f(x_1),\ldots,f(x_m))$;
2. Find the minimum e_m of $(f(x_1),\ldots,f(x_m))$;
3. Compute the novelty score $q = 1 - G^-(f(e_m);f,m)$;
4. If $1 - G^-(f(e_m);f,m)$ is greater than $1-\alpha$, label $(x_1,\ldots,x_m)$ novel.

- Rejecting the null hypothesis reflects our belief that the observed extreme is unlikely to be the extreme from a sample of f.

4.4.2 Novelty scores

The results developed so far now allow us to describe the complete theoretical framework for an original EVT-based novelty detection approach (Algorithm 4.1). Like Hann's method (2008), our method is based on the assumption that sufficiently unlikely observations with respect to a model distribution characterise "novelty"; i.e., the more improbable an observation is, the more certain we are of observing a new behaviour of the system under consideration with respect to our knowledge of its normal system behaviour.

Setting a threshold on $G^-(.;f,m)$, and subsequently classifying as novel observations that are shortfalls beneath this threshold, is similar to setting a lower threshold on the pdf f. There are, however, two major differences. Firstly, unlike the value of a threshold on density, the value of a threshold set on G^- has an actual probabilistic interpretation: it represents the probability mass below this threshold; i.e., the significance level or the *expected false-positive rate under normal behaviour*[14]. This feature is desirable, for instance, if the dimensionality of observations is variable (in the case of missing or corrupt data), and when some dimensions must thus be temporarily ignored (by marginalising out missing dimensions, for instance.). For such systems,

[14]Despite significance level and Type I error being two distinct concepts, they take equal values in the cases where the null hypothesis is systematically rejected at the α-level of significance.

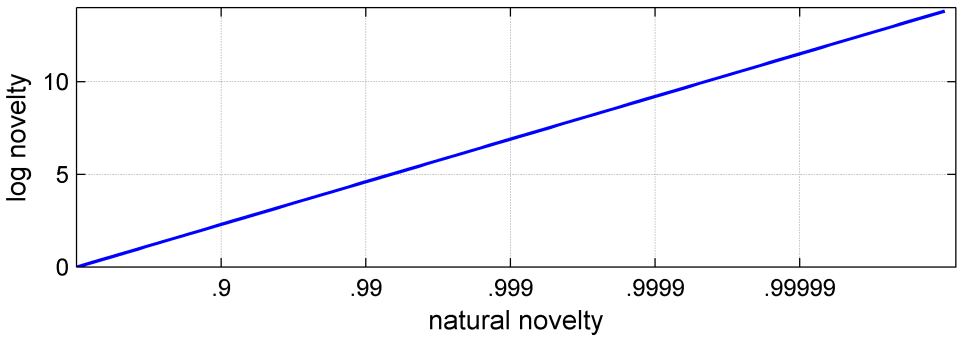


Figure 4.8: Natural scores vs. log-novelty scores. The natural novelty scale is logarithmic from 1 downwards. A natural score of 0.99 will always be equal to a log-novelty score of approximately 5, regardless of the shape of the distribution which generates them.

models of differing dimensionalities may be used, and knowing the value of an acceptable false-positive rate for the system makes defining a useful threshold after switching dimensionalities straightforward. Secondly, Algorithm 4.1 takes into account the size of the observed samples, which makes it less sensitive to those departures from the distribution modes which are expected as m increases (effects shown in Figures 4.1 and 4.2).

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

$$q(y;f,m) = \phi\left[1 - G^{-}(y;f,m)\right], \tag{4.36}$$

where ϕ is a monotonically increasing function with domain[15] $]0,1[$. In general ϕ will be the identity function when we want to relate observations to actual probabilities, and $1-G^{-}(y;f,m)$ is considered the probability that the sample is novel. We refer to these scores as the *natural novelty scores*. For classification purposes, ϕ will be a step function from $]0,1]$ to $\{0,1\}$. For visualisation purposes, we will occasionally take $\phi(.) = -\ln(1-.)$. Such scores are referred to as *log-novelty scores*. In figure 4.8, we plot natural scores versus log-novelty scores, in order to ease the interpretation of later graphs. The probabilistic interpretation of q makes it more interpretable than the log-scale used to define the PSI (Equation 1.1), as it is independent of the dimensionality of f, and indeed independent of the form of f. When using natural novelty scores, we call $G^{e}(.;f,m)$ and $F^{e}(.;f,m)$ the *novelty functions* over the probability space P_f

[15] Excluding 0 and 1 allows for the novelty of 0 and 1 to be $+\infty$.

and the data space D_f, respectively :

$$\forall y \in P_f,\ G^e(y;f,m) = 1 - G^-(y;f,m), \tag{4.37}$$

$$\forall x \in D_f,\ F^e(x;f,m) = 1 - G^-(f(x);f,m). \tag{4.38}$$

Before we can show plots of novelty scores, we need to address the missing practical step of this method: deriving the DPD $G(.,f)$.

4.5 Example: Gaussian distributions

Even if the DPDs exist, closed-form solutions for the DPDs of even a univariate distribution do not always exist, as the integral in Equation 4.35 is not always tractable. Distributions which are unimodal and symmetric about their mean are, however, easier cases. We derive here the DPD for the general Gaussian distribution, noting that we ultimately want to work with mixtures of multivariate Gaussian distributions.

4.5.1 DPD of the univariate Gaussian

Let $f = N(\mu, \sigma^2)$ be the Gaussian distribution with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}_+^*$, for which $D_f = \mathbb{R}$. The cdf and pdf are given by Equations 4.23 and 4.24. We start by noting that the DPD of any distribution is invariant under translation of the support. In the case of a Gaussian distribution, we can therefore assume $\mu = 0$ without loss of generality[16]. $P_f = f(\mathbb{R})$ is the interval $\left]0, \frac{1}{\sqrt{2\pi\sigma^2}}\right]$.

[16]This does not mean that novelty scores are the same for all Gaussian distributions with identical variance, as densities are dependent on f, and therefore dependent on μ

For all $y \in P_f$, $G(.;f)$ is defined as (Equation 4.35):

$$\begin{aligned} G(y;f) &= \Pr\left(f(x) < y\right) \\ &= 2\Pr\left(x < \sqrt{-\ln\left(2\pi\sigma^2 y^2\right)},\ x < 0\right) \\ &= 2(1 - \Pr\left(x \geq \sqrt{-\ln\left(2\pi\sigma^2 y^2\right)}\right),\ x < 0) \\ &= 2F\left(\sqrt{-\ln\left(2\pi\sigma^2 y^2\right)}\right) \\ &= 1 - \text{erf}\left(\sqrt{-\ln\left(\sqrt{2\pi\sigma^2}y\right)}\right). \end{aligned} \tag{4.39}$$

The second line is derived from the first as a result of the symmetry of f about 0. In the second line, we recognise the cdf of F. Equation 4.39 can be rewritten:

$$G(y;f) = \text{erfc}\left(\sqrt{\ln\left(\sqrt{2\pi\sigma^2}y\right)}\right). \tag{4.40}$$

A differentiation of Equation 4.39 yields:

$$g(y,f) = \left(\frac{2\sigma^2}{-\ln\left(\sqrt{2\pi\sigma^2}y\right)}\right)^{1/2}. \tag{4.41}$$

The top row of Figure 4.9 shows $f(x)$ and $G(y;f)$ for various values of the standard deviations. We note that P_f becomes smaller as the standard deviation increases, as the supremum of f decreases with increasing variance.

4.5.2 Novelty scores

Having derived a closed-form expression for $G(.;f)$, we can now plot novelty scores, using Equation 4.36. Figure 4.9 shows natural and log- novelty scores over the supports of f and g. As the tails of f become heavier, novelty scores become lower over D_f. Novelty scores agree with the symmetric nature of f. They increase as we move away from the mode of the distribution, modelling adequately our belief that minima and maxima observed further from the mode are less probable.

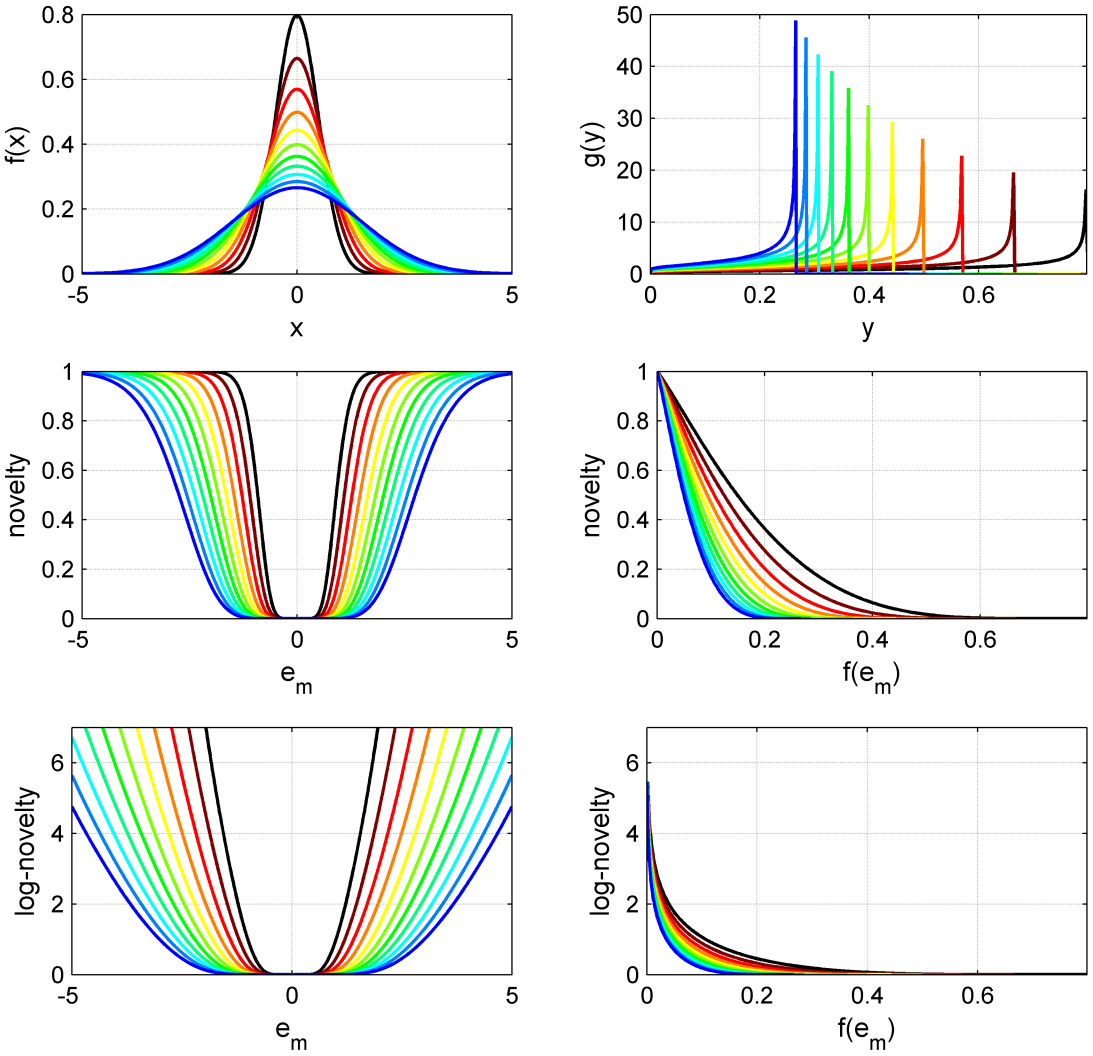


Figure 4.9: Generative distributions (top), natural novelty scores (middle), and log-novelty scores (bottom) for univariate Gaussians distributions, shown over D_f (left column) and P_f (right column). All means are 0. The standard deviation ranges from 0.5 (black) to 1.5 (dark blue) in 0.1 increment. The maximum log-novelty allowed here is 7, which corresponds to a natural novelty score of approximately 0.999. The exact minimal distribution of $G(.; f)$ is used to generate these graphs. In the top-right graph, the right endpoint decreases with increasing variance, because the maximum of f decreases.

4.5.3 Going back to the data space

Once again taking advantage of the symmetry of f, we can derive the exact distributions of extremes in the data space $\mathbb{R}$. The middle-left graph in Figure 4.9 shows that the righthand side part of the natural novelty function appears to be the cdf of extremes which lie to the right of the mean μ. The first derivative of the novelty function $F^e(x;f,m) = 1 - G^-(f(x);f,m)$ (The novelty function over D_f) for $x > \mu$ yields:

$$\begin{aligned} \forall x > 0, \frac{\mathrm{d}F^e(x;f,m)}{\mathrm{d}x} &= m \times [1 - G(f(x);f,m)]^{m-1} \times -g(f(x);f) \times \frac{\mathrm{d}f(x)}{\mathrm{d}x}, \quad (4.42) \\ &= m \times [1 - G(f(x);f,m)]^{m-1} \times -\frac{2\sigma^2}{x-\mu} \times \frac{-(x-\mu)f(x)}{\sigma^2}, \\ &= 2mf(x)\,[1 - G(f(x);f,m)]^{m-1}. \quad (4.43) \end{aligned}$$

Equation 4.42 is always valid, as long as $F^e(x)$ is differentiable. Equation 4.43 arises notably because in the case of the Gaussian distribution, $g(f(x);f) = \frac{2\sigma^2}{x}$. It can be interpreted as the pdf of the distribution over the extremes of f which lie in the righthand tail. By symmetry, the pdf of the distribution over the extremes of f which lie in the lefthand tail is also given by Equation 4.43. Concatenating the two halves of this distribution, and normalising the probability mass, we obtain an exact closed-form solution for our definition of extremes,

Proposition 3. *Data-space extremal distribution for the Gaussian distribution*
Let f be a Gaussian distribution with arbitrary mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}_+^*$ and a sample size $m \in \mathbb{N}_+^*$, then the distribution over E_m (given by definition 5), is

$$\forall x \in \mathbb{R},\ f^e(x;m) = mf(x)\,[1 - G(f(x);f,m)]^{m-1}. \quad (4.44)$$

This distribution is exact, and will be referred to as the exact data-space extremal distribution *of f.*

It is understood that finding a closed-form expression of f^e is not always possible, if only because it requires knowledge of a closed-form expression of the DPD G. Figure 4.10 shows a comparison between curves resulting from Equation 4.44 and numerically-simulated histograms of extremes for various values of m. When $m = 1$, we note that f^e is the original

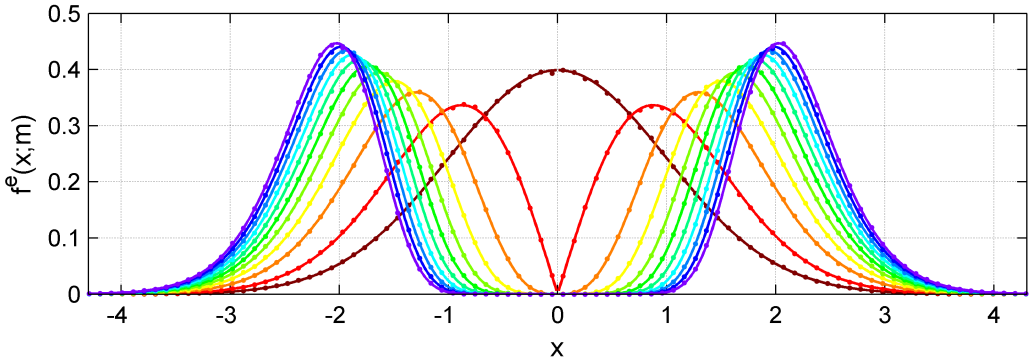


Figure 4.10: Distribution of extremes in data space for the standard Gaussian. Continuous lines are obtained using Equation 4.44. Dots represent normalised histograms of 10^6 simulated extremes. Values of m are 1 (brown), and 2 to 20 (purple) in increments of 2.

f. We argue that the observed agreement validates Proposition 3.

4.6 Novelty Detection with time-series

Section 4.3.2 laid the foundations for an EVT-based novelty detection framework that takes into account the number of data observed. In this section, we present an algorithm to perform novelty detection in univariate time-series. Let $(t_1,\ldots,t_N)$ be a series of regularly-sampled times, $X = (x(t_1),\ldots,x(t_N))$ a univariate time-series observed sequentially, and f the assumed underlying distribution. Our aim here is to determine if and when $x(t_i)$ can no longer be considered i.i.d. observations of f; i.e., if and when the underlying stochastic generative process changes[17]. Table 4.2 presents one possible algorithm, which makes use of a sliding window of fixed size m. Figure 4.11 demonstrates the use of Algorithm 4.2 on a synthetic time-series. The time-series consists of: 400 observations drawn from $N(0,1)$, 200 observations drawn from $N(0,4)$, and another 400 observations drawn from $N(0,1)$. We then test the null hypothesis that the sample is drawn from $N(0,1)$. Ideally, we want to classify only the central interval as novel. Log-novelty scores are calculated using the algorithm in Table 4.2 with exact minimal distributions of $G(.,f)$, and which are colour-coded according to whether they are above or below 3 (corresponding to natural novelty score of approximately 0.95, or a significance level of 0.05.), which we set as our log-novelty threshold. For a window size $m = 1$, the algorithm is strictly

[17] At this point, we restrict ourselves to detecting a statistically significant departure from the assumed stochastic process, and do not try to model the new random process.

Algorithm 4.2 EVT-based Novelty Detection in time-series

Let $X = (x(t_1), \ldots, x(t_N))$ be a univariate time-series observed sequentially, and f the assumed underlying distribution.

Initialisation:

1. Compute (or estimate) $G(.;f)$ and $G^{-}(.;f,m)$, which can be either an exact or an asymptotic EVD;
2. Choose a value of the threshold α in [0,1];
3. Choose a value of $m \in \mathbb{N}^*$.

Unseen data from a time-series:

1. At time t_i, $i < m$, do step 2 with $m = i$;
2. At time t_i, $i > m-1$, let $X_{t,m} = (x_{t-m+1}, ..., x_t)$ be the m most recent observations;
3. Apply Algorithm 4.1 to $X_{t,m}$. The resulting novelty score is the score at time t_i;
4. Slide the window (increment i) and iterate from step 2 (step 1 if $i+1 < m$).

equivalent to the density-thresholding method used by Hann (2008). We observe multiple false positives outside the time interval of interest, as well as a large proportion of false negatives. As m is increased, the number of transitions from a false state to a true state decreases, but novel episodes appear to last longer. Eventually, for m = 80, a large majority of the middle window is classified as novel, and no transition from a true positive state to a false positive state occurs. However, because a novel sample will remain in a window for m time steps, we observe a significant period of false positive classification at the end of the central interval. Overall, in this example, using our EVT-based scheme results in fewer misclassifications, as well as smoother novelty curves, which require little or no post-processing.

The use of a sliding window is however not the only option. One could set m to be the number of observations before time t_i (i.e., set $m = i$ at each step), or use fixed m and calculate individual scores without extracting the minimum probability density in a window. The former method is equivalent to setting a lower threshold in the probability space which becomes more conservative over time. The latter is equivalent to setting a lower fixed threshold in the probability space. The relative merits of these approaches are application-dependent and will be discussed in a later chapter (8).

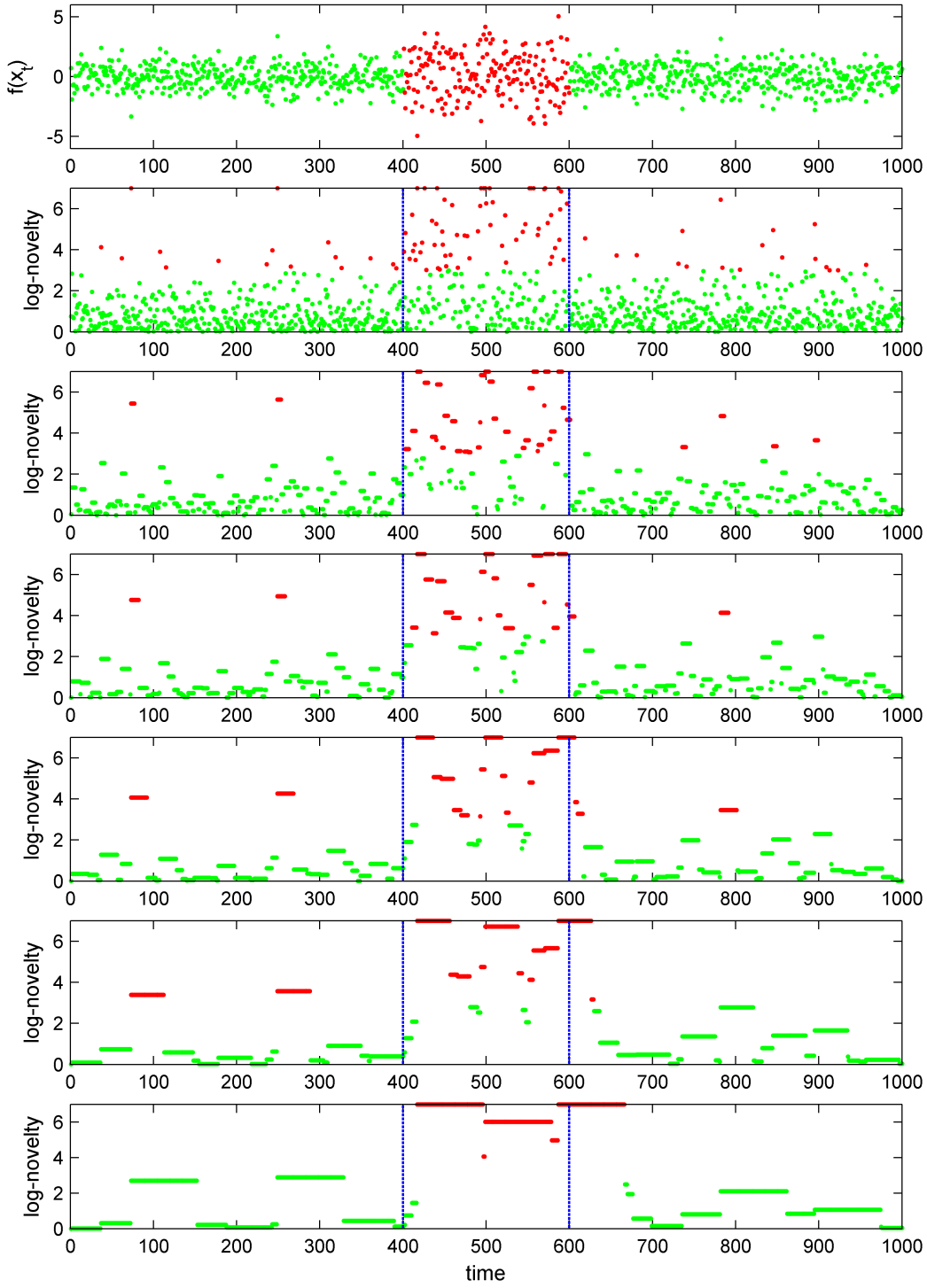


Figure 4.11: Example of novelty detection under variance shift. The top graph is a 10^3 point synthetic time-series. Green observations are drawn from $N(0,1)$, while red observations are drawn from $N(0,4)$. The remaining graphs show the log-novelty given by Algorithm 2 (saturated at 8) for a window-size m = 1, 5, 10, 20, 40, and 80, from top to bottom. Scores are plotted in red if they exceed 3 (natural score of approximately 0.95). The case $m = 1$ is strictly equivalent to the scheme used by Hann (2008).

4.7 Conclusions

Taking advantage of the availability of multiple observations, as long as it can be assumed that they come form the same underlying distribution, is an important improvement on treating (and classifying) observations fully independently. Even under the i.i.d. assumption, the sample size affects the likelihood of observing unusual data, which we argue can prove crucial when attempting to detect novelty in time-series. This assessment led us to a redefinition of extremes to include all improbable states, such that the use of classical extreme value theory remained possible. The novelty detection schemes in Algorithms 4.1 and 4.2 yield much cleaner raw novelty curves than the raw curves obtain via density thresholding. Additionally, novelty curves in the example of Figure 4.11 would require no heuristic post-processing, and have a natural probabilistic interpretation. They however rely on a strong i.i.d. assumption, which is rarely satisfied in practice, and remain sensitive to extreme outliers, while inflating the influence of single-point noise.

Another problem is that of the choice of m, which is not trivial when working with large time-series. Firstly, it could become computationally impractical to work with large values of m. Computation can be made easier by working with approximate, asymptotic EVDs, but the benefits of using a large value of m may not necessarily outweigh the disadvantage of using approximate EVDs. Secondly, changing m for $2m$ has a moderate (sub-logarithmic) effect on the novelty scores (sub-logarithmic), which pleads in favour of using reasonable values for m. Thirdly, and perhaps more importantly, m is related to how quickly we aim to detect a change in regime after it has occurred. Smaller values of m will yield earlier detections, but also more false positives, and conversely for larger values of m, and the choice is thus dependent on the dynamics of our system.

Finally, we note that we have not shown examples of EVDs (or approximations of EVDs) for the multimodal case. Section 4.3 presented a theoretical framework, but the only practical case treated in 4.4 is unimodal. The added difficulty in the multimodal case is the derivation of a closed-form expression for the DPD. Whenever this is impossible, G, or at least its lower tail, has to be estimated. A complete treatment is presented for mixtures of multivariate Gaussians

in the next chapter.

Chapter 5

Multivariate Extreme Value Theory for Novelty Detection

5.1 From univariate to multivariate data-space

If we now assume that normal data are drawn from a multivariate distribution, the aim of novelty detection remains to determine with maximal confidence when this stochastic process changes, while minimising the number of false detections (an illustration is shown in Figure 5.1). In statistical terms, we aim at determining when the null hypothesis H_0*: data are generated according to the model distribution* f does not hold to some level of significance. The case for using multivariate EVT as a novelty detection technique is the same as that made in Section 4.1: if multiple observations are available and assumed to originate from the same, known, model distribution, then simply thresholding its pdf is not theoretically appropriate, and one should take into account the size of the available sample.

To this end, we presented in Chapter 4 a novelty detection scheme based on univariate Extreme Value Theory, in which observations are ranked according to their probability densities, and the minimum observed density is evaluated against the minimal EVDs of the distribution of probability densities, which are associated with the model. The definition of extremes in terms of minimum probability density (Definition 5) is not restricted to univariate distributions, and

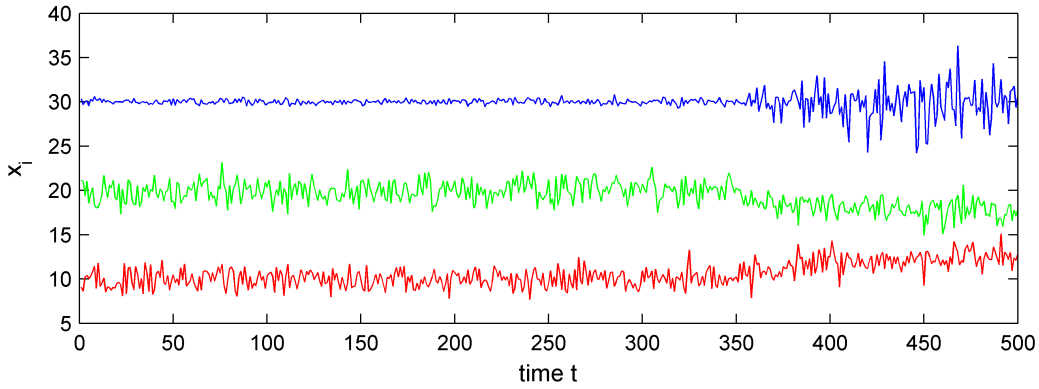


Figure 5.1: Illustration of the novelty detection problem for multivariate time-series. Each curve correspond to a dimension of the data. Units of the y-axis are arbitrary. The data-generating stochastic process remains constant until time $t = 350$. For $t > 350$, the stochastic process undergoes a change, resulting in an increased variance for the top curve and mean shifts for the bottom two. The aim of a novelty detection scheme is here to determine when this change occurs, as early as possible, while keeping the number of false detections to a minimum.

remains valid for all dimensionalities of the data space, essentially because the DPD is always a univariate distribution. Algorithm 4.1 is therefore still applicable, as long as we are able to estimate the DPD of our generative models, which will be the focus of the majority of this chapter.

Here, a generative model is assumed to be a d-dimensional, K-kernel mixture of Gaussians (MG) ($d, K \in \mathbb{N}^*$), for which the pdf is:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}) \tag{5.1}$$

$$= \sum_{k=1}^{K} \frac{\pi_k}{C_{d,k}} \exp\left(-\frac{M_k(\mathbf{x})^2}{2}\right) \tag{5.2}$$

where, for $k = 1, \ldots, K$, $M_k(\mathbf{x}) = \left((\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)^{1/2}$ is the Mahalanobis radius implied by $\boldsymbol{\Sigma}_k$, $C_{d,k} = (2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}$ is the normalisation coefficient, $\boldsymbol{\mu}_k$ is the centre, $\boldsymbol{\Sigma}_k$ is the covariance matrix and π_k is the mixing weight of the k^{th} component f_k. Additionally, we have the constraints $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and we define $y_{\max}$ to be the maximum of $f(\mathbf{x})$ over $\mathbb{R}^d$, and $P_f =]0, y_{\max}]$ to be the probability space. A bold font is used to indicate that $\mathbf{x}$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ are vectors and matrices. We write $f(\mathbf{x})$ without listing the model parameters to keep notations concise. Mixtures of Gaussians are important because they can approximate multivariate, mul-

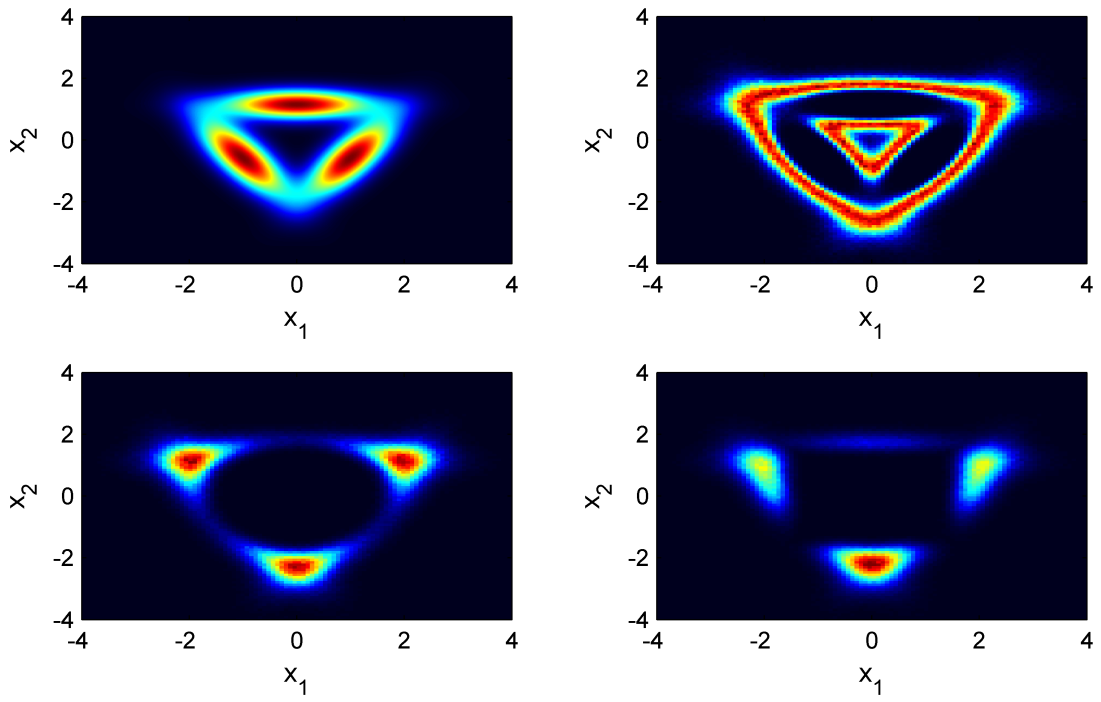


Figure 5.2: Generative bivariate distribution consisting of 3 Gaussian kernels (top-left) and empirical distributions of: extremes in probability density (top-right), maxima of the Euclidean distances to the centre $(0,0)$, of the distribution (bottom-left), and component-wise maxima (bottom-right). Component-wise maxima are taken on the absolute value of each x_i, and are therefore maxima in L_1 distance from the distribution centre. Only the extremes in probability density follow the geometry implied by the generative distribution. The sample size is $m = 10$ for all simulations.

timodal distributions, theoretically to arbitrary precision, if K is sufficiently large. In particular, Gaussians are a widespread choice in two popular kernel density estimation techniques, namely Parzen Windows (Parzen 1962) and Gaussian Mixture Models.

The main issue of the most classical approaches lies in the need to agree on a definition of extremes, as the use of different distance metrics to define magnitudes will result in different topologies of the data space, and EVDs will become dependent both on f and the chosen distance metric. This, we argue, is undesirable as the extreme character of a location in data space can be made solely dependent[1] on f, via Definition 5.

Figure 5.2 shows bivariate examples of empirical EVDs using the L_1 distance [2], the Euclidean distance, and the pdf of f to define extremes. The resulting EVDs are different enough that the

[1] Of course, this is predicated upon the assumption that f is known. If only extremes are available and f is not known, then there is no other choice than choosing a distance metric, and estimating EVDs directly.

[2] If $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, then the L_1 distance between $\mathbf{x}$ and $\mathbf{y}$ is $d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$.

choice of definition should be considered carefully. In this thesis, we will always assume that f (or an estimate of f) is known, and we will consider extremes in probability density. This approach will hereafter be referred to simply as "EVT". To our knowledge, there are two significant precedents in the use of this approach in the literature, from which this chapter is inspired. Both consider, implicitly for the earlier one (Roberts 1999, Roberts 2000) and explicitly for the later one (Clifton 2009), extremes in probability density. They are reviewed in Section 5.2. In Sections 5.3 to 5.7, we offer a number of analytical and algorithmic developments to cope with some of the weaknesses of both existing methods. The last section of this chapter offers a summary and a detailed comparison of existing and novel EVT methods.

5.2 Existing approaches to multivariate EVT

5.2.1 The "Winner-Takes-All method"

The "Winner-Takes-All" method (Roberts 1999, Roberts 2000) is a method for performing multivariate novelty detection in time-series, with examples on biomedical and financial data sets. In this approach, extremes are understood to be maxima in Mahalanobis radius with respect to all kernels in the mixture. Because the pdf of one kernel monotonically decreases as a function of the Mahalanobis radius, this definition is strictly equivalent, for a single kernel, to the definition in terms of probability density. As in the pdf thresholding method, an MG f is used to model the time-series data. At a location $\mathbf{x}$ in data space, finding the value of the EVD is reduced to a single-kernel problem: the closest kernel (determined using the Mahalanobis radii; i.e. the distances to the centre of each kernel, under the distance implied by the covariance matrix of that kernel) is assumed to dominate $f(\mathbf{x})$, and the EVD value associated with this location is considered to be the density given by the EVD of this particular kernel. The contribution of other kernels to $f(x)$ and $f^e(\mathbf{x})$ is considered to be negligible, and thus ignored. We refer to this assumption as Assumption 1 of the method.

Furthermore, Roberts notes:

- that the asymptotic EVD of the single-sided univariate Gaussian distribution is of type Gumbel, with parameter estimates resembling those of the Gumbel distribution for the two-sided Gaussian distribution (Equations 4.33 and 4.34);
- that any cross-section through the centre of a multivariate Gaussian kernel has the same shape as a single-sided univariate Gaussian distribution.

From these two observations, two assumptions are made:

- Assumption 2: that the cross-sections of the EVDs of a multivariate Gaussian kernel through its centre are asymptotically of Gumbel type;
- Assumption 3: that parameter estimators for the single-sided univariate Gaussian distribution can be used to estimate the parameters of this Gumbel distribution.

This method is hereafter referred to as the "Winner-Takes-All" (WTA) method, as was termed by Clifton (2009), who reviews it in detail, and points out problems arising due to differing kernel variances and overlaps between kernels (Section 5.3.2). Here, we only offer further insights into the limitations of the method's assumption permitted by the analytical developments in the remainder of this chapter.

We start by noting that Assumption 2 is correct. This comes from the fact that the DPD of a Gaussian kernel is in the domain of attraction of the minimal Weibull H_3^- [3] and that the mapping from data space to probability space occurs via a negative exponential. The following properties are easily verified (Embrechts et al. 1997, Sohn et al. 2005):

$$\begin{aligned} & X^{\alpha} \in \mathrm{MDA}\left(H_1^+\right) \\ \iff \quad & -\exp\left(-X\right) \in \mathrm{MDA}\left(H_3^+\right) \text{ with shape parameter } \alpha \\ \iff \quad & \exp\left(X\right) \in \mathrm{MDA}\left(H_3^-\right) \text{ with shape parameter } \alpha \end{aligned}$$

Therefore, if the DPD is in the domain of attraction of the minimal Weibull, EVDs along a Mahalanobis radius are Gumbel type.

[3] Section 5.6.1 in (Clifton 2009)

However, Clifton (2009) demonstrates that Assumption 3 is incorrect[4], as MLE estimates of the location parameter appear to move further away from 0 (in Mahalanobis radius), as the dimensionality is increased. Clifton shows empirically that this phenomenon is due to the fact that the probability mass shifts further from the distribution mode in Mahalanobis radius, as the dimensionality is increased [5]. It is one of the aims of the next section to explain this shift analytically (Section 5.33, and Figure 5.6). The presence of the Jacobian in Equation 5.12, when changing from the Cartesian coordinate system to the Mahalanobis-spherical polar coordinate system, implies that the probability mass is "diluted" along a Mahalanobis radius, and that the magnitude of this "dilution" is dependent on the dimensionality of the data space.

The results in Sections 5.4 and 5.5 allow us to use the correct EVD for a single Gaussian kernel (using, for instance, Equation 5.35 to obtain novelty scores), and therefore to examine Assumption 1. Using the notations of Equation 5.2, drawing an observation from an MG can be seen as follows:

- selecting a kernel, with each kernel having a prior probability π_k, of being selected;
- subsequently drawing an observation from this kernel.

Let us now consider a hypercube $[\mathbf{x}, \mathbf{x}+\mathrm{d}\mathbf{x}]$ in $\mathbb{R}^d$. On average, for a sample size m, we expect $m_k = m\pi_k$ of the observations in the sample to have been drawn from kernel k, as is noted by Roberts (1999). Let us call p_k the probability that at least one out of m_k observations from the k^{th} kernel lies in $[\mathbf{x}, \mathbf{x}+\mathrm{d}\mathbf{x}]$ (observations in the hypercube need not be extremes with respect to any kernel). If a single observation is generated, and t_k is the probability of that observation falling outside the hypercube, and then $p_k = (1-t_k)^{m_k}$. The probability that the j^{th} kernel generated an observation in the hypercube is then given by:

$$\frac{p_j}{\sum_{k=1}^{K} p_k} = \frac{(1-t_j)^{m_j}}{\sum_{k=1}^{K}(1-t_k)^{m_k}}. \tag{5.3}$$

This is true for all observations from the MG, and in particular, this is true if we place the hypercube around an m-extreme e_m of the MG, for some large $m \in \mathbb{N}^*$. This leads to the following

[4] Figure 5.3 in (Clifton 2009).
[5] Figure 5.6 in (Clifton 2009).

conclusions:

- An m-extreme from an MG need not be an m-extreme with respect to any of the k kernels;
- If at least two p_k have comparable magnitude (in the case of identical kernels overlapping significantly, or a large number of kernels with identical covariance matrices, as is typically the case for Parzen Windows estimators), then the choice of the nearest kernel is not adequate. Choosing the nearest kernel may result in underestimating how extreme a location should be considered, and therefore overestimating the novelty scores. In other words, even if the nearest kernel contributes the most to the density, it is that for which an observation is the least extreme (after accounting for the mixing weights), and the contribution of other kernels make the observation "more" extreme. An example of overestimation is shown in Figure 5.15;
- Alternatively, if one p_k dominates all the others, which is the case if one kernel has a large variance in all directions, then the contribution of the others kernels will be constantly ignored, even if they contribute to making the distribution asymmetric, leading to both overestimation and underestimation of novelty scores. An example is shown in Figure 5.3.

In summary, the WTA method successfully addresses the problem of multivariate EVT in some cases, those in particular for which kernels are pairwise far from each other in the Mahalanobis sense. Its major strength lies in the fact that it is fully analytical (as long as the correct EVDs of individual kernels are known analytically) and therefore requires no sampling. In the next section, we review a numerical method which does not rely on the dominant-kernel assumption.

5.2.2 The Ψ-transform method

Acknowledging the contributions of the WTA method, Clifton (2009) proposes a numerical method to estimate the EVDs and associated novelty scores for MGs. In this work, extremes are explicitly defined as extremes in probability density, and an extreme value distribution is

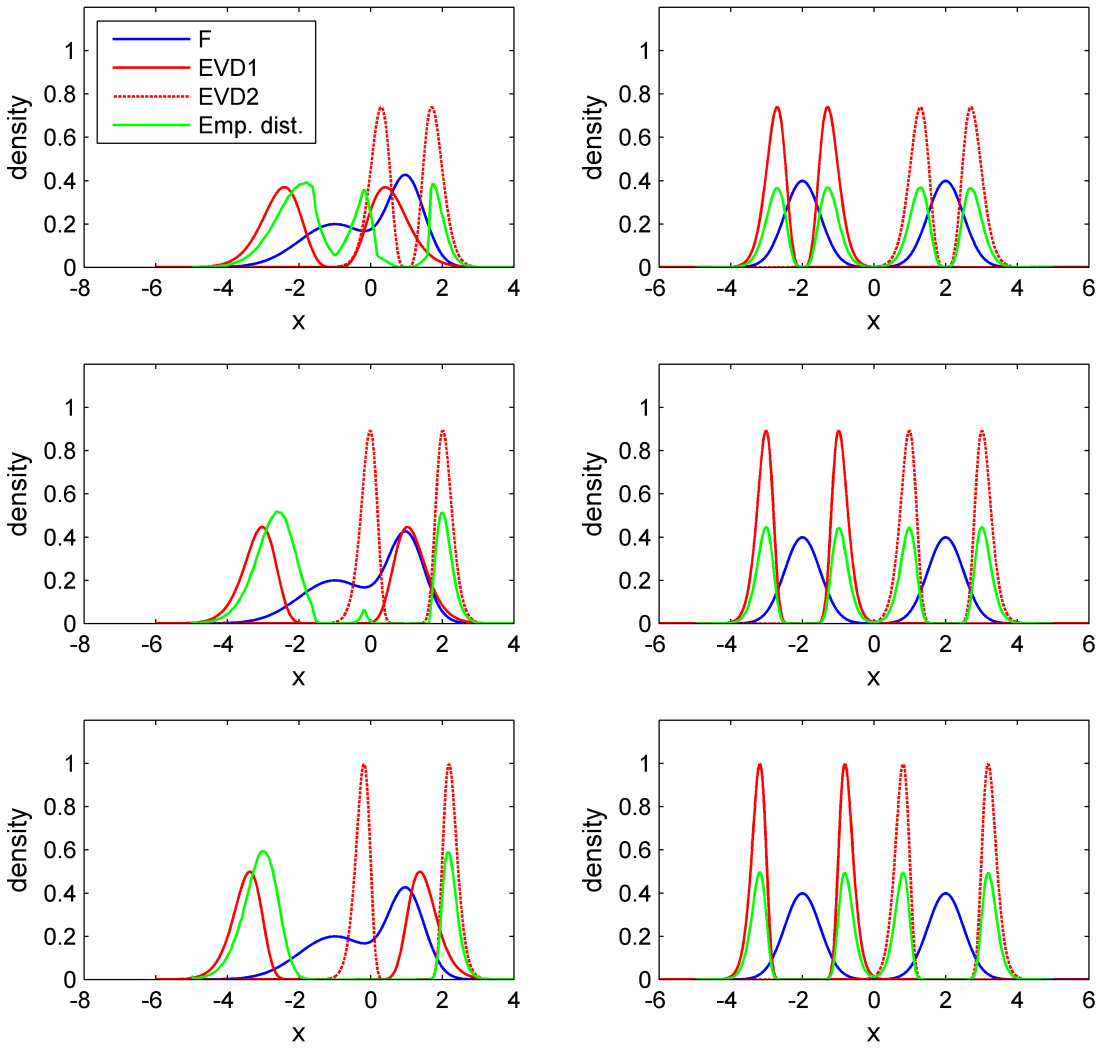


Figure 5.3: Two illustrations of the Winner-Takes-All method. Right column: a case for which the method is a good approximation. Left: for which the assumptions of the Winner-Takes-All method break down. In all graphs, f is a mixture of two Gaussians (blue), the exact EVDs of each kernel (given by Equation 4.44) are shown in red, and the empirical EVD is shown in green. From top to bottom the values of m are 5, 20, and 50. Right: the two Gaussian kernels have little overlap and equal variance, and empirical extremes are distributed according to the form of both exact EVDs; i.e., each mode of f corresponds to a clearly identified mode of one of the EVDs. Given an extreme observation, it is acceptable to use only the EVD of the nearest Gaussian kernel to assign a novelty score to it, for all values of m shown here. Left: the two Gaussian kernels overlap significantly, and have different variances. The empirical EVD does not behave according to any of the individual EVDs. On the lefthand side, the significant shift between the mode of the empirical EVD and that of the EVD of the nearest kernel, observed for all values of m shown here, indicates that the Winner-Takes-All method overestimates how far from the distribution centre extremes lie.

seen as a weighting of the MG's level sets[6]. To estimate an EVD for some sample size $m \in \mathbb{N}^*$, Clifton proposes to draw a large number N of extremes from the MG, and form an empirical distribution of their densities in the Ψ-transform space, where the Ψ-transform is defined as:

$$\Psi[f(\mathbf{x})] = \begin{cases} (-2\ln(f(\mathbf{x})) - d\ln(2\pi))^{1/2} & \text{if } f(\mathbf{x}) < (2\pi)^{d/2}, \\ 0 & \text{if } f(\mathbf{x}) \geq (2\pi)^{d/2}. \end{cases} \tag{5.4}$$

f is the pdf of the d-dimensional MG. If the original distribution were a single standard Gaussian distribution, the Ψ-transformation would map the $f(\mathbf{x})$ values back onto r, the Mahalanobis radius of $\mathbf{x}$. As such, the Ψ-transform attempts to map densities to pseudo-Mahalanobis radii[7], which we term here Ψ-radii. The case $f(\mathbf{x}) \geq (2\pi)^{d/2}$ ensures that no Ψ-radius is negative.

Observing that for a single Gaussian distribution, a Gumbel distribution can be fitted to the Mahalanobis radii, the assumption is made that a Gumbel distribution can subsequently be fitted to the distribution of Ψ-radii. Gumbel parameter values are obtained via MLE. Natural novelty scores are then straightforwardly given by the cdf values of the fitted Gumbel distribution (in Ψ-transform space, m-extreme correspond to large Ψ-radii, and the Gumbel distribution is a maximal distribution, so that there is no need to take the complement of the cdf to obtain novelty scores.). A descriptive example of the Ψ-transform method is given in Figure 5.4, which shows: an empirical distribution of m-extremes in data space, probability space, and Ψ-transform space; a Ψ-transform EVD in Ψ-transform space; and Ψ-transform natural novelty map in data space.

The Ψ-transform is shown to perform well for complex MGs in any dimension [8]. It makes a single assumption on the domain of attraction in Ψ-transform, which is proved to be correct in Section 5.7.1. Its major disadvantage is however the need to sample m-extremes. Assuming that satisfactory accuracy on the maximum likelihood estimates requires drawing N m-extremes, then the total number of required draws is $N \times m$. This is computationally impractical, especially because the asymptotic EVD obtained is only valid, in theory, for this particular value of m. This

[6] The levels sets, also known as isolines, isocontours or simply contours, of a distribution are the sets of locations over which the pdf takes a constant value.

[7] pseudo-, because it is not in general possible to define a Mahalanobis distance which would imply the geometry of an MG's level sets.

[8] Section 5.1, and Figures 5.13 to 5.16 in (Clifton 2009).

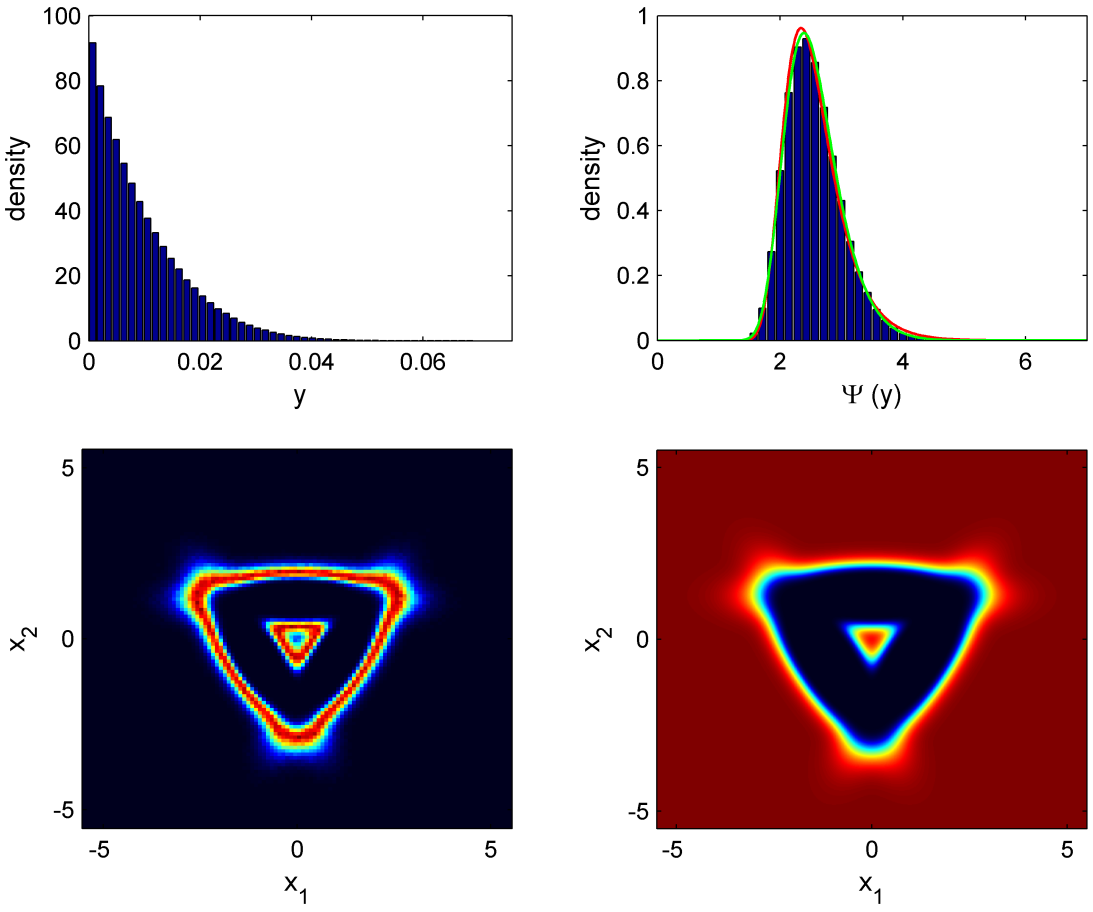


Figure 5.4: Illustration of the Ψ-transform method. Top-left: empirical EVD in the probability space. Top-right: empirical EVD in the Ψ-transform space, and MLE estimates of the Gumbel (red) and the GEV (green) distribution. Bottom-left: empirical EVD in the data space. Bottom-right: novelty map in the data space, using the estimate of the Gumbel distribution in red above.

means that if a range of sample sizes is to be investigated, m-extremes must be drawn for each value of m throughout that range.

In the remainder of this chapter, we present analytical developments which help reduce the computational cost of EVD estimation, while remaining as close to exact EVDs as possible.

5.3 Empirical cdf interpolation

In this section we propose a simple numerical algorithm to estimate EVDs and novelty scores with minimum use of sampling. To this end, we note that exact distributions, given by Equations 4.5 and 4.7 only depend on the cdf of the DPD $G(.;f)$. Therefore if $G(.;f)$ can be accurately estimated, it is easy to derive the exact EVDs, for any value of the sample size m, without the need for additional sampling. Using the notations of Equation 5.2, if f is a mixture of Gaussians, it is not in general possible to find a closed-form expression of $G(.;f)$. It is however possible to obtain an estimate $\hat{G}(.;f)$ of $G(.;f)$, simply by forming an empirical distribution $\hat{g}(.;f)$ of densities and integrating over its support.

Let us draw N observations from f and compute their densities $(f(\mathbf{x}_1),\ldots,f(\mathbf{x}_N))$. Given y in P_f, $G(y;f)$ is approximated by:

$$\hat{G}(y;f) = \frac{|\{\mathbf{x}_i / \mathbf{x}_i \leq y\}|}{N}. \tag{5.5}$$

$|.|$ denotes the cardinal number of a set[9]. $\hat{G}(y;f)$ is the empirical probability of drawing an observation less than y, which corresponds to the definition of a cumulative distribution function[10]. Since it would be impractical to draw N times from the distribution for every value of y, a set of strictly increasing values $Y_P = \{y_0, y_1, \ldots, y_P\}$, where P is an integer greater than 2, can be chosen, with $y_0 = 0$ and $Y_P = \max\{f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d\}$. For y in Y_p, $\hat{G}(y;f)$ is evaluated by drawing N observations once (or a few times and averaging the results). For values of y not in Y_P, such that $y_i < y < y_{i+1}$ for some $i \in [0, P-1]$, $\hat{G}(y;f)$ is evaluated by linearly interpolating between

[9]The number of elements in the set.

[10]This is a general scheme for estimating a cdf, which need not be that of DPD.

$(y_i, \hat{G}(y_i; f))$ and $(y_{i+1}, \hat{G}(y_{i+1}; f))$. With P and N of the orders of 10^3 and 10^6, respectively, $\hat{G}(y_i; f)$ is a good estimate of $G(y_i; f)$ for all values of y. EVDs can consequently be estimated using the formula:

$$\hat{G}^{-}(y; f, m) = 1 - (1 - \hat{G}(y; f))^m, \tag{5.6}$$

for a sample size $m \in \mathbb{N}^*$. Natural novelty scores follow directly from Equation 4.37:

$$\hat{G}^{e}(y; f, m) = 1 - \hat{G}^{-}(y; f, m). \tag{5.7}$$

$$\hat{G}^{e}(y; f, m) = \left(1 - \hat{G}(y; f, m)\right)^m. \tag{5.8}$$

An example of novelty score map obtained using Equation 5.7 for a multimodal MG is given in Figure 5.15.

The main advantage of this method is its simplicity: it does not require us to sample extremes; $\hat{G}(y; f)$ is a robust estimator, regardless of the complexity of f. Linear interpolation between two precomputed values of the cdf estimate is computationally inexpensive and can in theory be made arbitrarily accurate by increasing the value of P. Typically, for cases where the null hypothesis (Section 4.4.1) needs to be rejected at the $\alpha = 0.05$ level of significance, for moderate values of m (e.g. $m < 100$), there is no need for a more elaborate method. For large values of m however, this interpolation method breaks down for two reasons:

1. The estimator $\hat{G}(.; f)$ yields more accurate estimates for large values of $G(.; f)$ (where $G(.; f)$ is close to 1), which is unfortunate as we are interested in the lower tail of the DPD. This effect is demonstrated on a simple example in Figure 5.5. In this example, the model is a standard univariate Gaussian, for which the cdf is known analytically. Estimates of the cdfs were generated at selected locations 10^3 times, for $N = 10^5$, and their medians and inter-quartile ranges are computed. While medians appear to fit the analytical cdf closely, it is clear that the relative errors[11], and the relative inter-quartile range[12] increase by several orders of magnitude as the cdf decreases.

2. For large values of m, the probability mass of the EVD is heavily skewed towards $y = 0$. A

[11] If t is the target value and $\hat{t}$ is an estimate of t, then the relative error is defined as $\frac{|t-\hat{t}|}{t}$.
[12] The relative inter-quartile range is the quotient of the inter-quartile range and the target value.

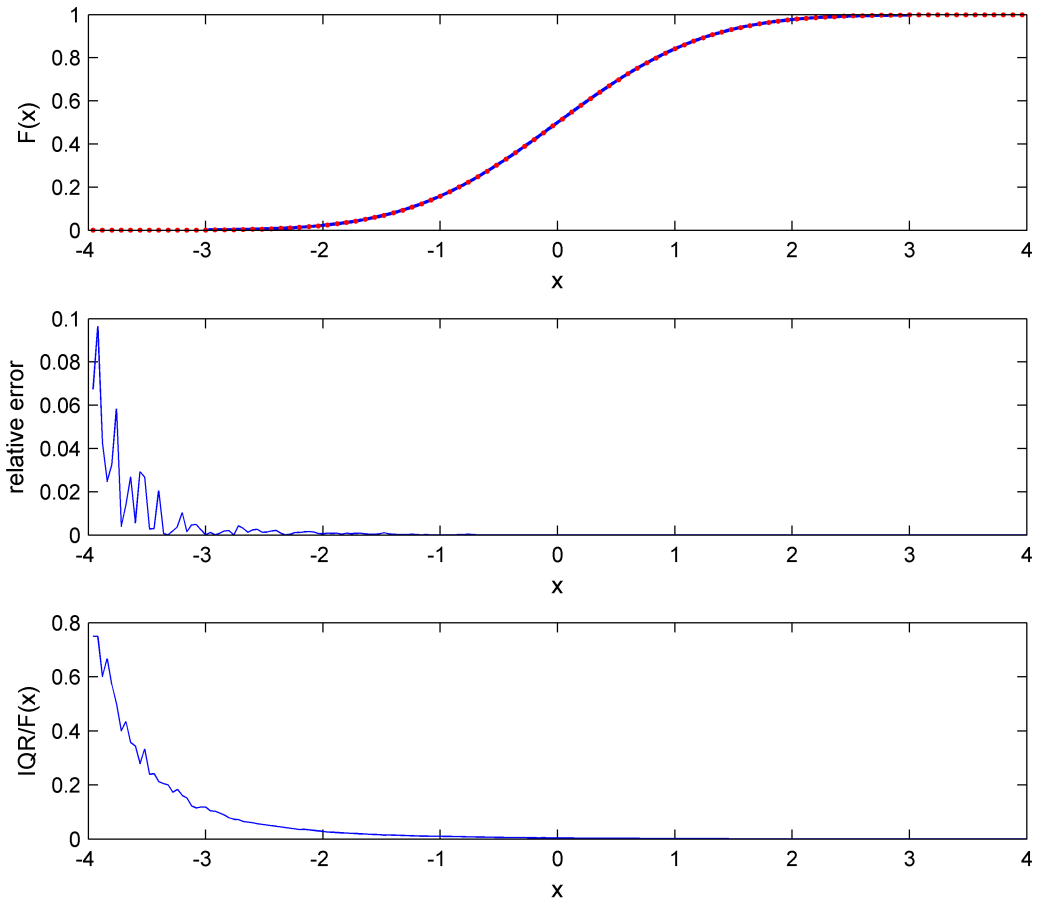


Figure 5.5: Example of cdf estimation using Equation 5.5. The model f is the standard univariate Gaussian. Its cdf is plotted in blue on the top graph. At a location x, we generate 10^3 estimates of $\hat{F}(x)$, using $N = 10^5$ observations drawn from f. Their medians are reported in red. The middle graph shows the relative error between these medians and the true values of F. The lower graph shows the relative inter-quartile ranges. In both cases, it can be seen that estimates are accurate for $x > -1$, yet exhibit significant relative errors and high relative IQR as x decreases.

simple calculation shows that if $m = 10^4$, over 99% of the EVD mass lies to the left of the 2^{nd} percentile of the DPD. To estimate the EVD accurately, we would need a significant number of y_i to lie below this percentile, which happens to be the region of space where we have the poorer estimates. Obtaining accurate estimates would require to increase N by orders of magnitude, which is undesirable.

In conclusion, it would be preferable to have at our disposal a closed-form formula for the DPD. We show in the next section that this is possible for any Gaussian distribution. For general MGs for which the DPD cannot be written in closed-form, asymptotic properties of the DPD allow us to model its tail below a certain threshold, which allows us to derive asymptotically accurate

closed-form solutions for the EVDs of an MG.

5.4 DPD of the multivariate Gaussian

In this section, we consider the multivariate Gaussian distribution F of arbitrary dimensionality $d \in \mathbb{N}^*$. The pdf is given by:

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{C_d} \exp\left(-\frac{M(\mathbf{x})^2}{2}\right), \tag{5.9}$$

where $M(\mathbf{x}) = \left((\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{1/2}$ is the Mahalanobis distance to $\boldsymbol{\mu}$, $C_d = (2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}$ is the normalisation coefficient, $\boldsymbol{\mu}$ is the centre, and $\boldsymbol{\Sigma}$ is the covariance matrix. The data space and probability space are $D_f = \mathbb{R}^d$ and $P_f = f(D_f) = \left]0, \frac{1}{C_d}\right]$, respectively.

Estimating the EVDs of multivariate Gaussians is interesting for three reasons: first, Gaussians are suitable models of normality for a number of (simple) problems; second, accurately estimating their EVDs brings an improvement to the WTA method described in Section 5.2, remembering that we have already stated that the parameter estimates previously used (Roberts 1999, Roberts 2000) do not agree with the maximum likelihood estimates as the dimensionality of the sample space increases; third, and perhaps more importantly, Gaussian kernels occur often as components in more complex models, such as Support Vector Machines or Gaussian Processes.

The ellipsoidal symmetry of f and its relatively simple form make it possible to obtain closed-form expressions of its exact and asymptotic minimal EVDs. This is done here in four steps: (i) calculating an analytical form of the distribution G associated with F; (ii) deducing the form of the exact EVDs; (iii) showing that G is, for the distribution of its minima, in the domain of attraction of the minimal Weibull distribution, (iv) estimating the Weibull parameters.

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

of the transformation is $|J_d| = |\Sigma|^{1/2} r^{d-1} \prod_{i=0}^{d-3} (\cos\theta_i)^{d-i}$ (see, for instance, (Scott 1992)). This yields:

$$
\begin{aligned}
G(y;f) &= \int_{f^{-1}(]0,y])} \frac{1}{C_d} \exp\left(-\frac{M(\mathbf{x})^2}{2}\right) \mathrm{d}\mathbf{x}, && (5.10)\\
&= \int_{f^{-1}(]0,y])} \frac{|J_d|}{C_d} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r\,\mathrm{d}\boldsymbol{\theta}, && (5.11)\\
&= \Omega_d \int_{M^{-1}(y)}^{+\infty} \frac{r^{d-1}}{(2\pi)^{d/2}} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r, && (5.12)\\
&= \int_0^y \Omega_d |\Sigma|^{1/2} \left[-2\ln(C_d u)\right]^{(d-2)/2} \mathrm{d}u. && (5.13)
\end{aligned}
$$

$M^{-1}(y) = \sqrt{-2\ln(C_d y)}$ is the unique Mahalanobis radius associated with the probability density y. Equation 5.11 is obtained by rewriting Equation 5.10 in the spherical polar coordinate system. Integrating out the angles yields Equation 5.12, where $\Omega_d = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}$ is the total solid angle subtended by the unit d-sphere, and Γ is the Gamma function[13]. Equation 5.13 is obtained after making the substitution $u = \frac{1}{C_d} \exp\left(-\frac{r^2}{2}\right)$.

The integrand in 5.13 is the pdf of G, which we aim to find,

$$
g(y;f) = \Omega_d |\Sigma|^{1/2} \left[-2\ln(C_d y)\right]^{(d-2)/2}, \tag{5.14}
$$

In the univariate case $d = 1$, noting that $\Omega_1 = 2$ and $C_1 = \sqrt{2\pi|\Sigma|}$, it follows from Equation 5.14 that:

$$
\begin{aligned}
\text{If } d = 1, \text{ then } g(y;f) &= \left(\frac{2|\Sigma|}{-\ln\left(\sqrt{2\pi|\Sigma|}y\right)}\right)^{-1/2}, && (5.15)\\
\text{and } G(y;f) &= \operatorname{erfc}\left(\sqrt{-\ln\left(\sqrt{2\pi|\Sigma|}y\right)}\right). && (5.16)
\end{aligned}
$$

Note that Equations 5.15 and 5.16 are identical to Equations 4.39 and 4.41 in Section 4.5.1 with $|\Sigma| = \sigma^2$. The derivation therefore holds for $d = 1$.

[13] $\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t)\,\mathrm{d}t$, $\forall x \in \mathbb{R}$. In particular, if n is a positive integer, $\Gamma(n) = (n-1)!$

In the bivariate case $d = 2$, the integration of Equation 5.13 is straightforward, and we have:

$$\begin{aligned} \text{If } d = 2, \text{ then } g(y;f) &= C_2, & (5.17) \\ \text{and } G(y;f) &= 2\pi|\Sigma|^{1/2}y \;=\; C_2 y. & (5.18) \end{aligned}$$

Note that if $d = 2$, $g(y,f)$ is the uniform distribution over P_f.

If $d \geq 2$, the integration of 5.13 is possible using a recursive integration by parts. Let us start by noting that the righthand side of Equation 5.12 represents the total probability mass which lies on the outside of the hyper-ellipsoid of Mahalanobis radius R; i.e., the set $\{\mathbf{x} \in \mathbb{R}^d \,/\, M(\mathbf{x}) = R\}$. Let us call $J_d(R)$ the complement of this quantity, i.e.; the total probability mass which lies inside the hyper-ellipsoid of Mahalanobis radius R:

$$\forall R \in \mathbb{R}_+,\; J_d(R) \;=\; \frac{\Omega_d}{(2\pi)^{d/2}} \int_0^R \exp\left(-\frac{r^2}{2}\right) r^{d-1} \mathrm{d}r. \tag{5.19}$$

Denoting $u_d(R)$ the integral part of Equation 5.19, we have:

$$\begin{aligned} u_1(R) &= \sqrt{\frac{\pi}{2}}\,\mathrm{erf}\left(\frac{r}{\sqrt{2}}\right), & (5.20) \\ \text{and } u_2(R) &= 1 - \exp\left(-\frac{R^2}{2}\right). & (5.21) \end{aligned}$$

Moreover, if $d > 2$, an integration by parts yields:

$$\begin{aligned} u_d(R) &= \int_0^R \exp\left(-\frac{r^2}{2}\right) r^{d-1} \mathrm{d}r \\ &= \left[-\exp\left(-\frac{r^2}{2}\right) r^{d-2}\right]_0^R + (d-2)\int_0^R \exp\left(-\frac{r^2}{2}\right) r^{d-3} \mathrm{d}r \\ &= -\exp\left(-\frac{R^2}{2}\right) R^{d-2} + (d-2)u_{d-2}(R). & (5.22) \end{aligned}$$

Equation 5.22 is a recursive relation for the sequence of functions $(u_d)_{d\in\mathbb{N}^*}$. Since u_d is a function of u_{d-2} only, we need to treat the two cases "d even" and "d odd" separately, and obtain two formulae for u_d (and therefore J_d). The initial terms $u_1(.)$ and $u_2(.)$ are given by Equations 5.20 and 5.21.

If $d = 2p$, $p \in \mathbb{N}^*$, it follows from Equations 5.21 and 5.22 that:

$$u_{2p}(R) = 2^{p-1}(p-1)!u_2(R) - \sum_{k=0}^{p-2} \frac{2^k(p-1)!}{(p-1-k)!} R^{2(p-k-1)} \exp\left(-\frac{R^2}{2}\right), \quad (5.23)$$

$$= 2^{p-1}(p-1)! - \sum_{k=0}^{p-1} \frac{2^k(p-1)!}{(p-1-k)!} R^{2(p-k-1)} \exp\left(-\frac{R^2}{2}\right). \quad (5.24)$$

If $d = 2p+1$, $p \in \mathbb{N}$, Equations 5.22 and 5.20 yield:

$$u_{2p+1}(R) = u_1(R)\frac{(2p+1)!}{2^{p-1}(p-1)!} - \sum_{k=0}^{p-1} \frac{(2p+1)!(p-k)!}{2^{k-1}(p-1)!(2p-2k)!} R^{2(p-k)-1} \exp\left(-\frac{R^2}{2}\right). \quad (5.25)$$

Writing,

$$\forall p \in \mathbb{N}*,\ A_{2p}^k = \Omega_{2p}|\Sigma|^{1/2}\frac{2^k(p-1)!}{(p-1-k)!}, \quad (5.26)$$

$$\forall p \in \mathbb{N},\ A_{2p+1}^k = \Omega_{2p+1}|\Sigma|^{1/2}\frac{(2p-1)!(p-k)!}{2^{k-1}(p-1)!(2p-2k)!}, \quad (5.27)$$

we can then write:

$$J_{2p}(R) = 1 - \exp\left(-\frac{R^2}{2}\right)\sum_{k=0}^{p-1} A_{2p}^k R^{2(p-k-1)}, \quad (5.28)$$

$$J_{2p+1}(R) = \operatorname{erf}\left(\frac{R}{\sqrt{2}}\right) - \exp\left(-\frac{R^2}{2}\right)\sum_{k=0}^{p-1} A_{2p+1}^k R^{2(p-k)-1}. \quad (5.29)$$

$$(5.30)$$

Finally, Equation 5.13 can be rewritten as:

$$G(y;f) = 1 - J_n\left(\sqrt{-2\ln\left((2\pi)^{n/2}|\Sigma|^{1/2}y\right)}\right), \quad (5.31)$$

which after substituting Equations 5.28 and 5.29 yields the two cases of DPDs for the multi-

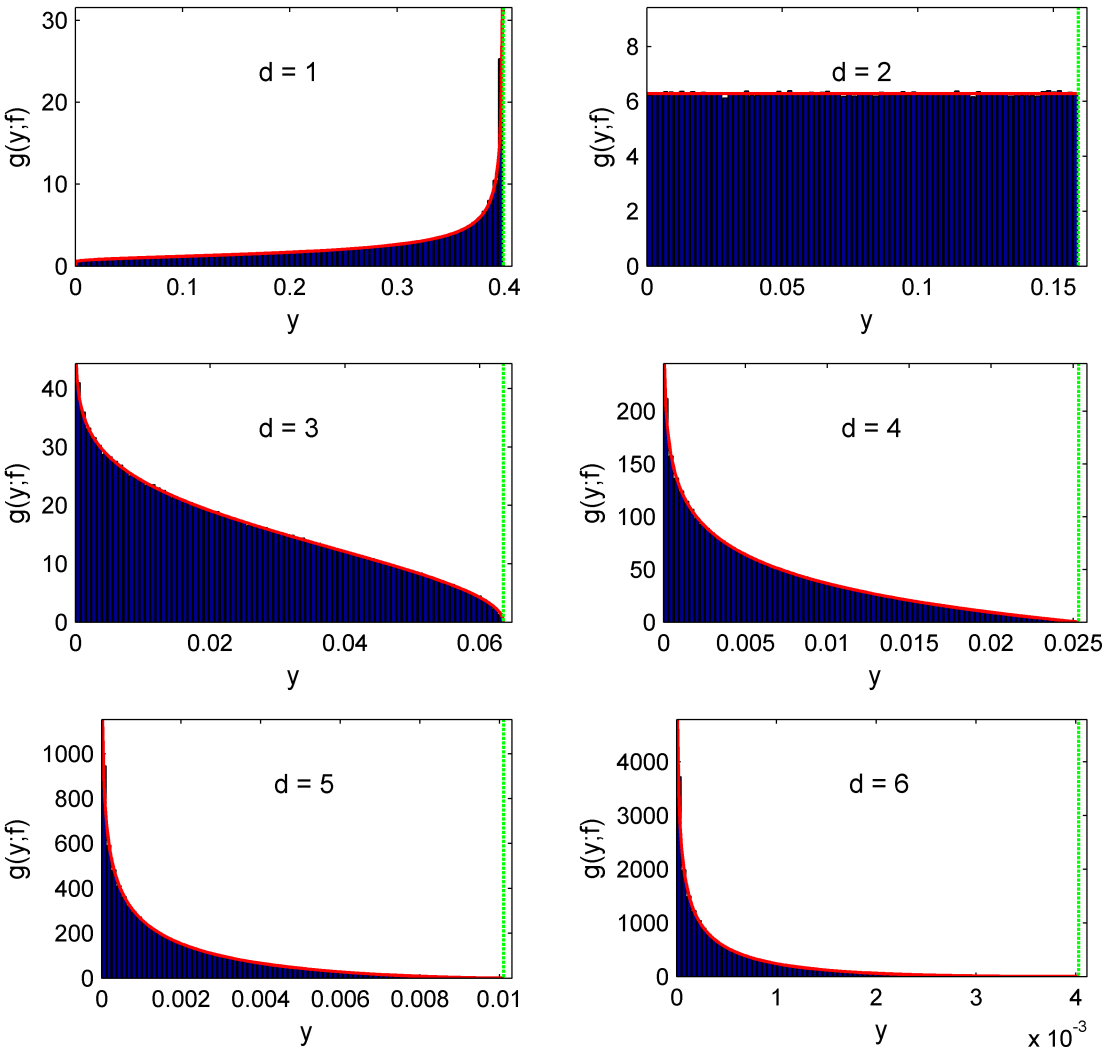


Figure 5.6: DPDs of standard multivariate Gaussian distributions for $d = 1$ to 6. In each graph, the red line is the pdf of the DPD, obtained using Equation 5.14. The empirical distribution in blue is the normalised histograms of 10^6 simulated observations. The vertical green line marks the maximum value of $f(\mathbf{x})$; i.e., C_d^{-1}.

variate Gaussian distribution:

$$G(y;f) = y\sum_{k=0}^{p-1} A_{2p}^{k}\left(-2\ln\left(C_{2p}y\right)\right)^{(p-k-1)}, \tag{5.32}$$

$$G(y;f) = \text{erfc}\left(\sqrt{-\ln\left(C_{2p+1}y\right)}\right) + y\sum_{k=0}^{p-1} A_{2p+1}^{k}\left(-2\ln\left(C_{2p+1}y\right)\right)^{\frac{2p-2k-1}{2}}. \tag{5.33}$$

Equation 5.32 is the case for which $d = 2p$, while Equation 5.33 is valid when $d = 2p+1$. Figure 5.6 shows empirical pdfs of DPDs for $d = 1$ to 6. The red lines obtained using Equation 5.14 fit the empirical pdfs adequately. The univariate Gaussian distribution is the only case for which $g(y;f)$ is an increasing function of y. In the bivariate case, $g(y;f)$ is constant, which

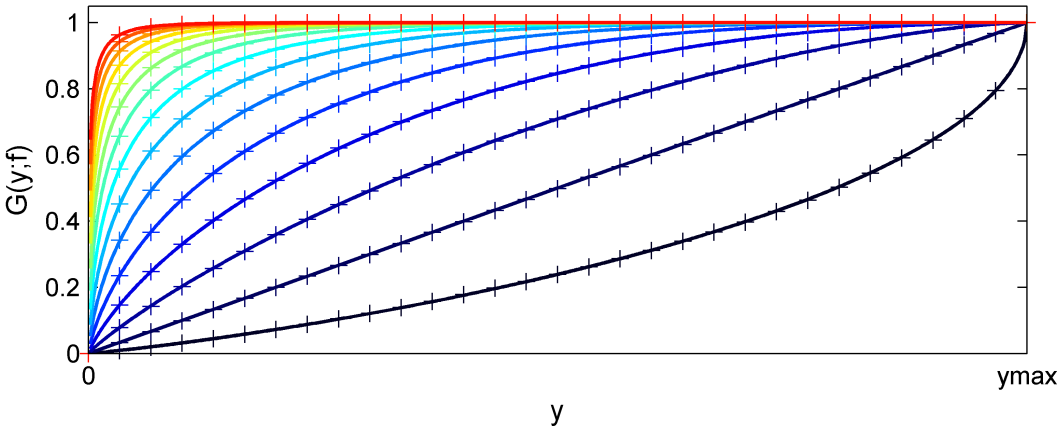


Figure 5.7: DPDs of multivariate Gaussian distributions. Continuous lines are obtained using Equations 5.32 and 5.33. Values of d range from 1 (black) to 15 (red). Crosses of each colour represent the empirical distribution obtained from 10^6 simulated observations associated with the continuous line of the same colour. For visualisation purposes, all the cdfs are represented over a scaled $[0, y_{\max}]$ range. They do not, in fact, share a right endpoint.

means that every density in $]0; \frac{1}{C_2}]$ is *equiprobable* when drawing randomly from f. For $d > 2$, $g(y; f)$ is a decreasing function of y, which indicates that, relative to the right endpoint of $g(.; f)$, the probability mass shifts towards 0 as the dimensionality d increases. In the data space, this means that the probability mass moves away from the centre of the distribution, as d increases [14]. These findings are reinforced by Figure 5.7, which shows the cdfs of DPDs for multivariate Gaussian distributions of dimensionalities 1 to 15, obtained using Equations 5.32 and 5.33. For $d = 2$, the curve of $G(.; f)$ is a straight line, under which we find only the curve of the univariate Gaussian. The progression towards the upper-left part of the graph as the dimensionality increases confirms that the probability mass moves away from the mode of the Gaussian. In other words, the probability becomes more spread out in regions of relatively low density, which themselves become dominant over regions of high density, resulting in DPDs with heavier tails.

[14] As is noted by Bishop (2007) (ex. 1.4, p.29.) and Clifton (2009) (section 5.2.2, p.66). We provided here a full analytical treatment of their observations for the multivariate Gaussian distributions.

5.5 Exact EVDs of the multivariate Gaussian

5.5.1 Exact EVDs

The exact EVD for the minima of G defined by Equations 5.32 and 5.33 is, according to the general form given by Equations 4.7 and 4.8:

$$G^{-}(y;f,m) = 1-[1-G(y;f)]^{m}, \tag{5.34}$$

$$g^{-}(y;f,m) = m\,g(y;f)[1-G(y;f)]^{m}. \tag{5.35}$$

Figure 5.8 shows examples of curves obtained using Equation 5.35 for a variety of dimensionalities and sample sizes, together with empirical EVDs resulting from the sampling of 10^6 extremes. We note that, in these graphs, the scale of the x-axis is logarithmic. Empirical EVDs are obtained using logarithmic bins. The number of samples falling in the leftmost bins is very small (a few dozens), while most of the probability mass is concentrated on the righthand side of the distributions. The bias observed at the lower end of the density axis (x-axis) are therefore expected.

5.5.2 Novelty scores

Natural novelty scores in the probability space are given by the following equation:

$$G^{e}(y;f,m) \;=\; 1-G^{-}(y;f,m) = \Big(1-G(y;f)\Big)^{m}, \tag{5.36}$$

where $G(y;f)$ is given by Equations 5.32 and 5.33. From these EVDs, we can derive the natural novelty function in data space, which we write here as a function of the Mahalanobis distance to the centre of the distribution:

$$F^{e}(r;f,m) \;=\; \Big(1-G(f(r);f)\Big)^{m}. \tag{5.37}$$

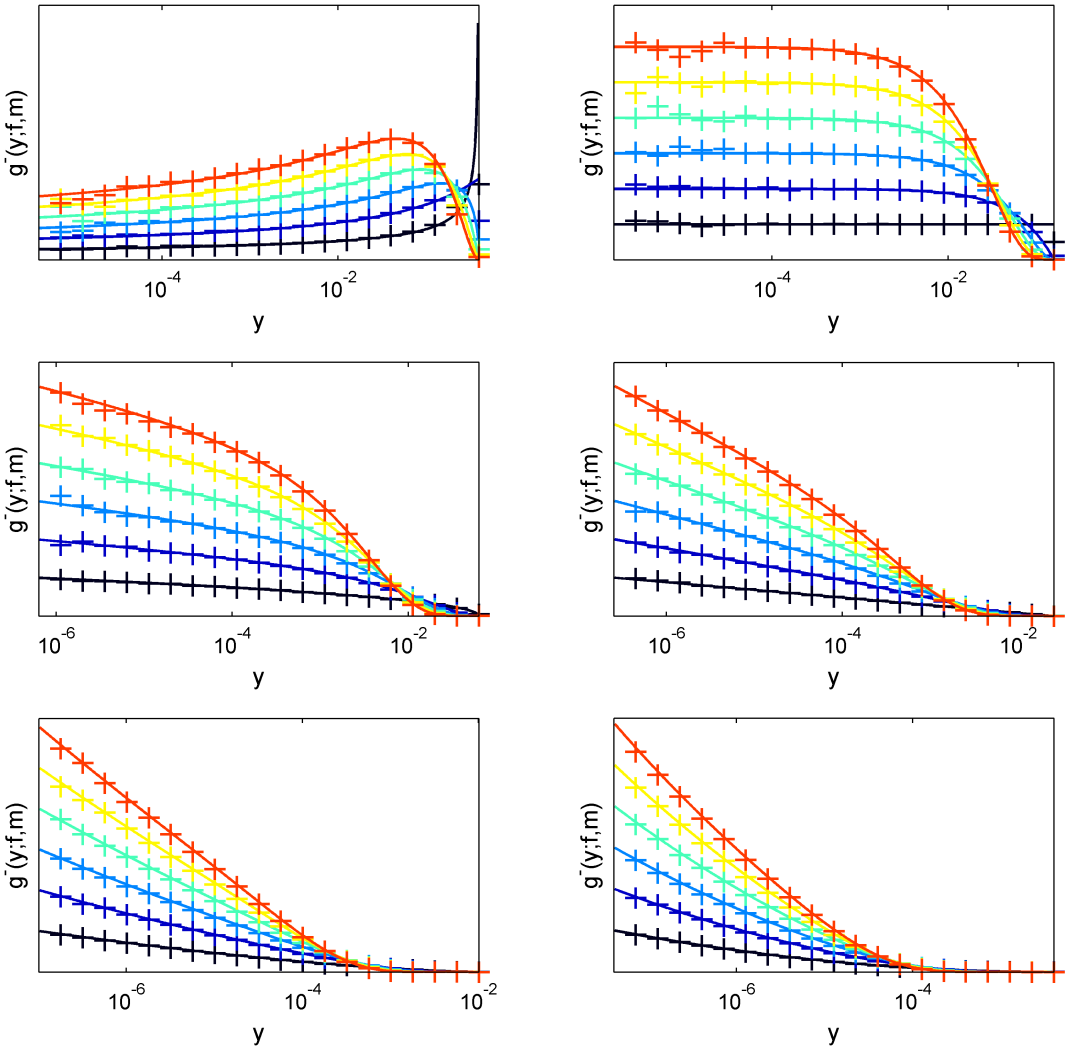


Figure 5.8: EVDs in probability space for the standard d-dimensional Gaussian. The values of m range from 1 (black) to 6 (red). The dimensionality of the kernels go from $d = 1$ (top-left) to $d = 6$ (bottom-right). The crosses are the empirical distributions (10^7 extremes), represented on logarithmic scales ranging from $\frac{10^{-5}}{C_d}$ to $\frac{1}{C_d}$. The continuous lines are obtained using Equation 5.35.

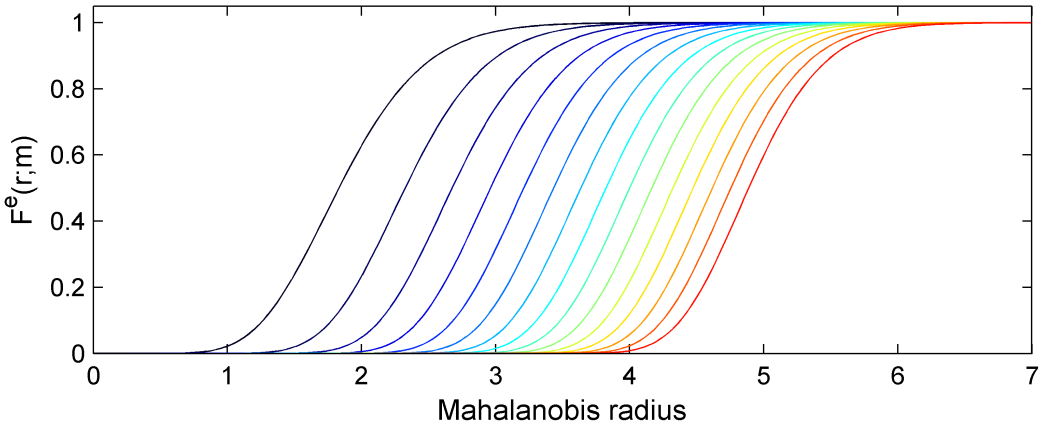


Figure 5.9: Exact EVDs as a function of the Mahalanobis distance for a sample size $m = 10$ and dimensionalities ranging from 1 (black) to 15 (red).

Given $\mathbf{x} \in \mathbb{R}^d$, the novelty score at this location can be found by calculating $r = M(\mathbf{x})$ and using Equation 5.37. Figure 5.9 shows natural novelty scores as a function of the Mahalanobis distance r for different values of d. Note that to reach the same novelty scores at a higher dimensionality, the Mahalanobis distance has to be increased, which is in line with the remarks accompanying Figure 5.7.

5.5.3 EVD in data space

The exact data-space extremal distribution, $f^e(.;m)$, can be derived in the same way it was derived in Section 4.6 to obtain Equation 4.43. We first note that, if we marginalise out all of the angles (that is, if we focus on the radial component of $f^e(.;m)$), we can write:

$$\begin{aligned} f^e(r;m) &\propto \frac{\mathrm{d}F^e(r;f,m)}{\mathrm{d}r}, \\ &\propto m[1-G(f(r);f)]^{m-1} \times -g(f(r);f) \times \frac{\mathrm{d}f(r)}{\mathrm{d}r}, \\ &\propto m[1-G(f(r);f)]^{m-1} \times -\Omega_d |\Sigma|^{1/2} r^{d-2} \times -rf(r), \\ &\propto \Omega_d |\Sigma|^{1/2} r^{d-1} m f(r) [1-G(f(r);f)]^{m-1}. \end{aligned} \tag{5.38}$$

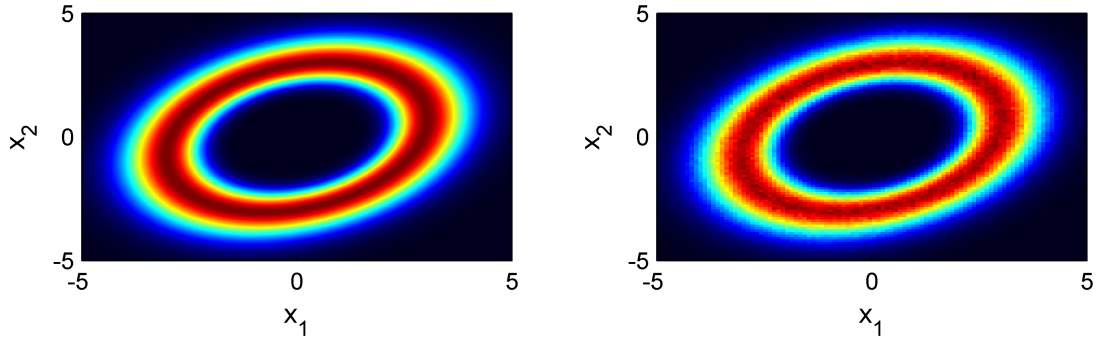


Figure 5.10: Theoretical (left) and empirical (right) EVDs of a bivariate Gaussian. The theoretical EVD is obtained using Equation 5.39. The empirical distribution is obtained after drawing 10^7 extremes. The sample size is $m = 10$.

Taking advantage of the radial symmetry of the Gaussian distribution once again, we can say that in the data space, the extremal distribution is solely dependent on r and therefore:

$$\begin{aligned} f^e(\mathbf{x};m) &\propto f^e(M(\mathbf{x});m), \\ &\propto \Omega_d |\Sigma|^{1/2} M(\mathbf{x})^{d-1} f(M(\mathbf{x})) m[1-G(f(\mathbf{x});f)]^{m-1}. \end{aligned} \tag{5.39}$$

The normalisation constant can be calculated using the same steps used to obtain Equation[15] 5.13. Figure 5.10 shows a comparison between an empirical EVD and an EVD obtained using Equation 5.38 for a bivariate Gaussian distribution.

5.6 Asymptotic EVDs of the multivariate Gaussian

To find a solution for the asymptotic EVD of the multivariate Gaussian in probability space, we first show that its DPD is, for the distribution of its minima, in the domain of attraction of the Weibull. To find the appropriate norming constants, we adapt Theorem 4 which gives a characterisation of the MDA of the maximal Weibull distribution H_3^+, to characterise the domain of attraction of the minimal Weibull distribution H_3^-:

Theorem 5. *(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

[15] This is not shown here, as the derivation is rather long, and we are predominantly interested in the form of F^e.

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 $c_m = x_F + F^{\leftarrow}(m^{-1})$ and $d_m = x_F$.

M_m^- is here the minimum of m i.i.d. random variables. x_F is the left endpoint of the df F, $F^{\leftarrow}(p)$ is the p-percentile of F, and L is a slowly varying function at $+\infty$ (Equation 4.21).

5.6.1 The DPD of the Gaussian is in the domain of attraction of H_3^-

Following from Theorem 5, we have the proposition:

Proposition 4. *Domain of attraction of the DPD for the multivariate Gaussian.*
Let f be the multivariate Gaussian distribution. F is, for the distribution of its minima in probability density, in the domain of attraction of the minimum Weibull distribution H_3^-.

For a sample size $m \in \mathbb{N}^*$, the norming constants can be chosen to be:

$$c_m = G^{\leftarrow}\left(\frac{1}{m}; f\right), \tag{5.40}$$

$$d_m = 0, \tag{5.41}$$

$$\alpha_m = 1. \tag{5.42}$$

Proof of Proposition 4.

Step 1: $yG(y^{-1}; f)$ is a sum of slowly-varying functions.

Starting with the case where $d = 2p$, $p \in \mathbb{N}^*$, it is clear from Equation 5.32 that $G(.; f)$ can be written as the product of y and a linear combination of functions of the form:

$$\forall y \in]0, a^{-1}],\ i(y; a, b) = [-\ln(ay)]^b, a > 0, b \in \mathbb{N}^* \tag{5.43}$$

Near the left endpoint 0,

$$i\left(0 + \frac{1}{y}; a, b\right) = \left[\ln\left(\frac{y}{a}\right)\right]^b, \tag{5.44}$$

and we have the following sequence:

$$\begin{aligned}
& \forall t \in \mathbb{R}_+^*, && \lim_{y\to\infty} \frac{\ln(y)+\ln(t)}{\ln(y)} = 1, \\
\Rightarrow \quad & \forall t \in \mathbb{R}_+^*, && \lim_{y\to\infty} \frac{\ln(ty)}{\ln(y)} = 1, \\
\Rightarrow \quad & \forall t, a \in \mathbb{R}_+^*, && \lim_{y\to\infty} \frac{\ln(ty/a)}{\ln(y/a)} = 1, \\
\Rightarrow \quad & \forall t, a \in \mathbb{R}_+^*, b \in \mathbb{N}^*, && \lim_{y\to\infty} \left(\frac{\ln(ty/a)}{\ln(y/a)}\right)^b = 1, \\
\Rightarrow \quad & \forall a \in \mathbb{R}_+^*, b \in \mathbb{N}^*, && i \text{ is a slowly varying function.}
\end{aligned} \tag{5.45}$$

Consequently, near the left endpoint 0, $yG(y^{-1};f)$ is a sum of slowly varying functions.

The case d odd is very similar. The only additional step is to prove that $y\operatorname{erfc}\left(\sqrt{\ln\left(\frac{y}{C_d}\right)}\right)$ is a slowly varying function of y at $+\infty$. This is straightforward, as it tends towards[16] 0 near $+\infty$, and functions which admit a limit near $+\infty$ can be shown to be slowly varying.

Step 2: A sum of slowly-varying functions is slowly varying.

If L_1 and L_2 are two slowly-varying functions, then using the definition of a limit, for all $t > 0$ and $\varepsilon > 0$, there exists $x_0 \in \mathbb{R}_+^*$ such that, for all $x > x_0$:

$$(1-\varepsilon)L_1(x) \leq L_1(tx) \leq (1+\varepsilon)L_1(x), \tag{5.46}$$

$$\text{and } (1-\varepsilon)L_2(x) \leq L_2(tx) \leq (1+\varepsilon)L_2(x) \tag{5.47}$$

using the fact that, that L_1 and L_2 are positive functions. Summing these two equations yields:

$$(1-\varepsilon)(L_1(x)+L_2(x)) \leq L_1(tx)+L_2(tx) \leq (1+\varepsilon)(L_1(x)+L_2(x)). \tag{5.48}$$

which means that $\lim_{x\to\infty} \frac{L_1(tx)+L_2(tx)}{L_1(x)+L_2(x)} = 1$; i.e, the sum of two slowly-varying functions is a slowly varying function. The proof holds for any number of terms in the sum, and therefore the sum of any number of slowly-varying functions is a slowly varying function. Similarly a positive linear combination (a linear combination with positive coefficients) of slowly-varying

[16] This is easily proven using, for instance, a first-order Taylor expansion of erfc.

functions can be shown to be a slowly-varying function. This is not necessarily true if some coefficients are negative (Jankovic & Ostrogorski 2000, Jankovic & Ostrogorski 2002).

Step 3: Norming constants

From steps 1 and 2, we conclude that $yG(y^{-1};f)$ is a slowly-varying function, and therefore that Theorem 5 applies: $G(.;f)$ is in the domain of attraction of the minimal Weibull distribution H_3^-. The choice of norming constants follows then directly from the theorem.

□.

The scale parameter c_m can be easily estimated numerically to arbitrary accuracy, as $G(.;f)$ is a strictly increasing univariate function over a finite support. Figure 5.11 shows that the formula for the scale parameter c_m is a very close approximation to the maximum likelihood estimates. However, the value of the shape parameter α_m, although theoretically guaranteed to converge to 1 in the limit $m \to +\infty$, appears to be accurate only in the case $d = 2$. It seems to decrease significantly as the dimensionality of the data space increases, and is overestimated even for large values of m. We derive a new estimator for the shape parameter in the next section.

5.6.2 New estimator for the Weibull shape parameter

The class of equivalence of H_3^- contains all the distributions with a power law behaviour at the finite left endpoint (Embrechts et al. 1997). Therefore the tail of $G(,;f)$ is, in the limit $y \to 0$, equivalent to the tail of a power law; i.e., $G(y;f) \sim Ky^s$, where K is a positive constant. It follows that $g(y;f) \sim sKy^{s-1}$; i.e., $s = y\frac{g(y;f)}{G(y;f)}$. This approximation is, of course, valid only near 0, and s should be evaluated near 0, in a manner that depends on m. We propose to evaluate s at the $\frac{1}{m}^{\text{th}}$ percentile of $G(.;f)$; i.e., calculate the power law behaviour of $G(.;f)$ at this percentile. Additionally this percentile happens to be the value of the scale parameter c_m given by Equation 5.40, so we define:

$$\alpha_m = c_m \frac{g(c_m;f)}{G(c_m;f)}. \tag{5.49}$$

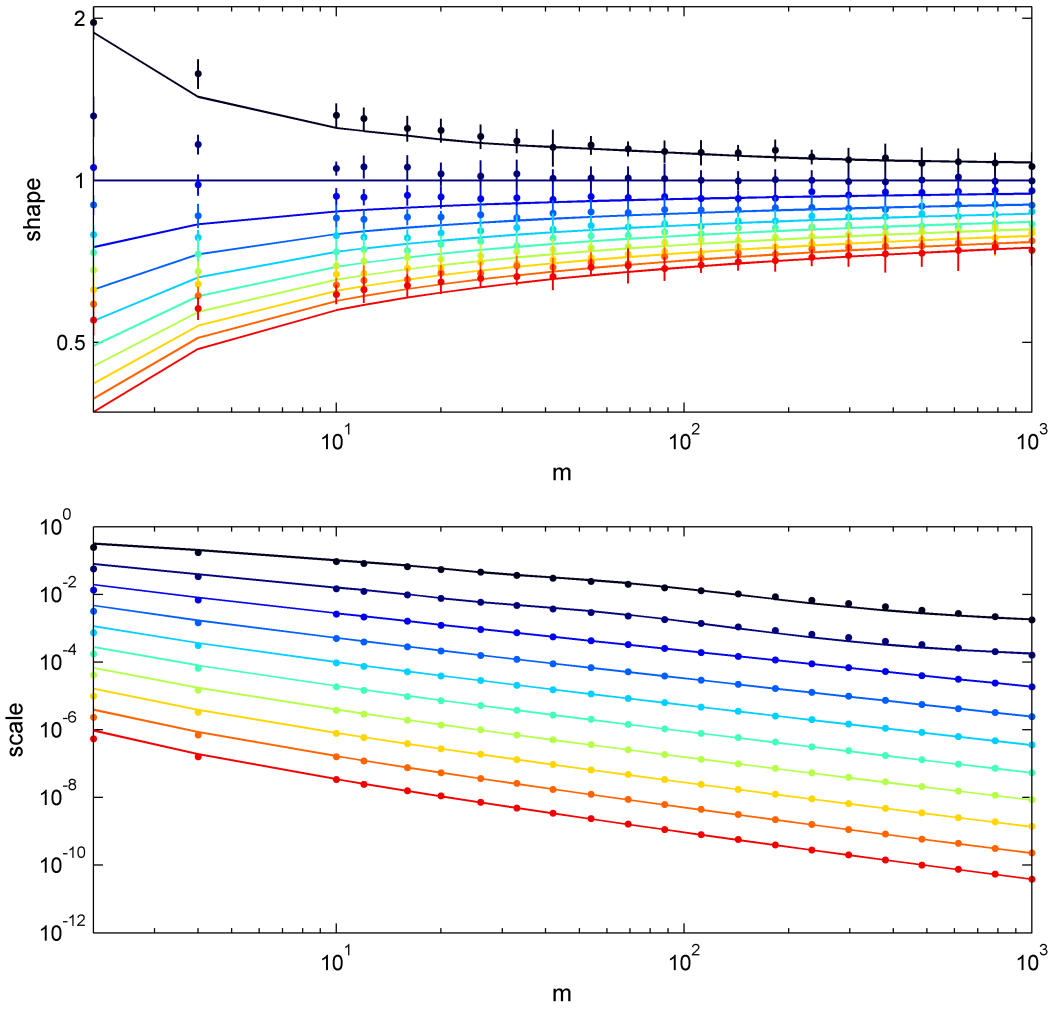


Figure 5.11: Weibull shape (top) and scale (bottom) parameters for the standard multivariate Gaussian. Dimensionalities range from 1 (black) to 10 (red). Continuous lines are the estimates obtained using Equations 5.40 for c_m and 5.49 for α_m. Dots are the mean of 10 maximum likelihood estimates (10^3 extremes per iteration), represented with error bars ($\pm$ 2 standard deviations). Error bars for the scale parameter are too small to be visible at this scale. The location parameter is fixed at 0 in all simulations.

Figure 5.11 shows that Equation 5.49 yields estimates much closer to the maximum likelihood estimates than those obtained using Equation 5.42, even for low values of m. In particular, when $d = 2$, $c_m = \frac{1}{m y_{\max}}$, $g(c_m; f) = y_{\max}$, and $G(c_m; f) = \frac{1}{m}$which means $\alpha_m = 1$, which is the value given by Proposition 4. In conclusion, we recommend the use of our original estimator over that of the estimator given by the literature.

5.6.3 Asymptotic EVDs and novelty maps

Finally, letting f be the pdf of the multivariate Gaussian in $\mathbb{R}^d$, the asymptotic EVDs of $G(.; f)$ are given in the probability space by (Equation 4.15):

$$G^e(y; f, m) = 1 - \exp\left[-(y/c_m)^{\alpha_m}\right], \tag{5.50}$$

where c_m and α_m are given by Equations 5.40 and 5.49, respectively.

The corresponding natural novelty function in data space is given by:

$$F^e(\mathbf{x}; m) = 1 - G^e(f(\mathbf{x}); m), \tag{5.51}$$

$$= \exp\left[-\left(\frac{1}{C_d c_m} \mathrm{e}^{-\frac{M(\mathbf{x})^2}{2}}\right)^{\alpha_m}\right]. \tag{5.52}$$

At a location $\mathbf{x}$, $F^e(\mathbf{x}; m)$ is the probability of drawing a sample of size m whose extremum has a density value higher than $f(\mathbf{x})$; i.e, the probability of drawing an extremum closer to the centre of the distribution (in the Mahalanobis sense). Figure 5.12 shows a comparison between the exact novelty score function (Equation 5.35) and the asymptotic one (Equation 5.52), for $d = 5$ and various sample sizes. As expected, the asymptotic novelty score provides an increasingly better fit to the exact score as m is increased (acceptable fits are obtained for values of m greater than 2^4).

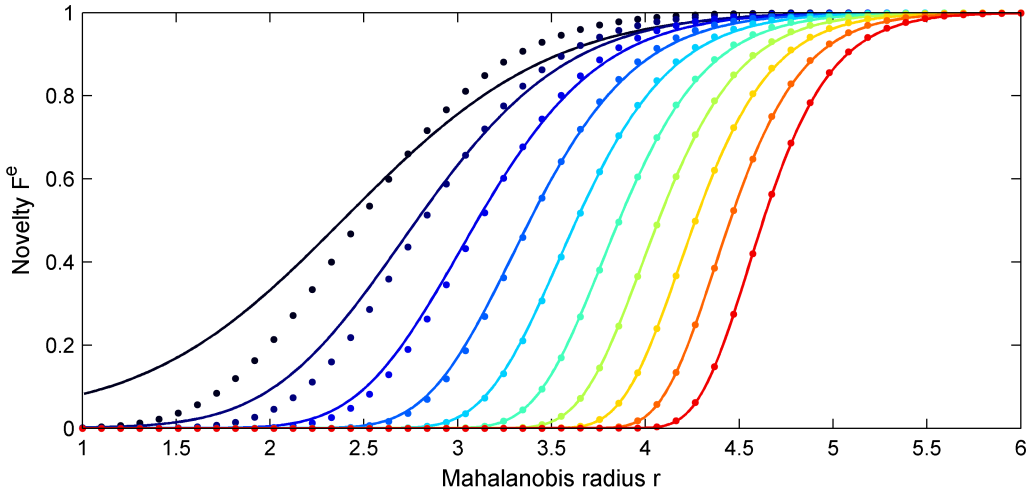


Figure 5.12: Comparison of asymptotic (continuous lines, Equation 5.35.) vs. exact (dots, Equation 5.12) EVDs, as a function of the Mahalanobis radius. The dimensionality is $d = 5$, and the sample size m ranges from 2^1 (black) to 2^{10} (red). The asymptotic EVDs fit poorly for low sample sizes, but become increasingly more accurate as the sample size increases.

5.7 EVT for mixtures of Gaussians

Having derived both exact and asymptotic extreme value distributions for the multivariate Gaussian distribution, we now focus in this section on the more general case of mixtures of Gaussians (MG). A successful example of EVD estimation for MGs was given in Section 5.2.2, using Clifton's Ψ-transform method. The main drawback of this method is the need to sample m-extremes from the generative distribution. This generative process is computationally intensive, as one needs to draw $m \times N$ to obtain N m-extremes, from which the parameters of the Gumbel distribution can be estimated via MLE. Furthermore, the process needs to be reiterated for each value of m, that one might wish to investigate.

In cases where an MG can be assumed to be responsible for generating data, we show that sampling m-extremes to estimate the EVDs is not always necessary, which considerably reduces the computational cost. In previous sections, we have seen that exact and asymptotic EVDs can be determined entirely from the knowledge of the DPD $G(.;f)$ and the sample size. We therefore have to goals: finding the MDA of $G(.;f)$ for its minima, and estimating the DPD.

In section 5.7.1, we show that $G(.;f)$ is in the domain of attraction of H_3^-. Following sections describe two original semi-numerical methods to estimate the EVDs of an MG, which follow

in the footsteps of the method adopted to estimate the EVDs of multivariate Gaussians. We show that asymptotic EVDs of MGs can be estimated by drawing N times from the MG once. These N observations are used to estimate the lower tail of the DPD. From the estimate of the DPD tail, estimates of the EVDs can be analytically derived with minimal computation, for any value[17] of m, without the need to resample.

5.7.1 The DPD of MGs is in the domain of attraction of H_3^-

Similar to Proposition 4, we have:

Proposition 5. *Domain of attraction of the DPD of an MG.*
Let f be a mixture of Gaussians. F is, for the distribution of its minima in probability density, in the domain of attraction of the minimum Weibull distribution H_3^-. For a sample size $m \in \mathbb{N}^*$, the norming constants can be chosen to be:

$$c_m = G^{\leftarrow}\left(\frac{1}{m}; f\right), \tag{5.53}$$

$$d_m = 0, \tag{5.54}$$

$$\alpha_m = c_m \frac{g(c_m; f)}{G(c_m; f)}. \tag{5.55}$$

Proof of Proposition 5.

Because the DPD of an MG has a finite left endpoint, it can only be in the domain of attraction of H_1^- or H_3^-. For $y \in P_f$:

$$G(y; f) = \int_{f(\mathbf{x})<y} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{5.56}$$

$$\geq \sum_{k=1}^{K} \pi_k \int_{f(\mathbf{x})<y/\pi_k} f_k(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{5.57}$$

$$\geq \sum_{k=1}^{K} \pi_k G(y/\pi_k; f_k). \tag{5.58}$$

The right-hand side quantity of Equation 5.57 is the sum of suprema of the individual kernel

[17] Although satisfactory accuracy cannot be achieved for the smallest values of m.

contributions to $G(y;f)$. For a given y, the contribution of f_k is less than y/π_k. Since $G(.;f_k)$ is in the domain of attraction of H_3^- (Proposition 4), $G(y/\pi_k;f_k)$ does not decrease faster than any power function (see the remarks at the end of Section 4.2.2), as $y \to 0$. Consequently, the sum of $G(y/\pi_k;f_k)$ and $G(y;f)$ cannot decrease faster than any power function, and so it cannot be in the domain of attraction of the minimal Gumbel, and is thefore in the domain of attraction of the minimal Weibull H_3^-.

The choices of norming constants come from Theorem 5, and Equation 5.49.

□.

Mixtures of Gaussians having the property of universal approximation, it could be tempting to conclude from Proposition 5 that all the true distributions that we are trying to model (and can be approximated to arbitrary accuracy by mixture of Gaussians) are in the domain of attraction of the Weibull distribution, for the distribution of their minima in probability density. We explicitly refrain of making this claim, and note that, because the domain of definition of the distributions in density are bounded by zero, they fall in the maximum domain of attraction of either the Weibull or Fréchet distribution.[18]

Proposition 5 validates the observation made for the Ψ-transform method that Ψ-radii are asymptotically distributed according to a H_1^+, since the following properties can be immediately verified (Embrechts et al. 1997, Sohn et al. 2005):

$$\begin{aligned} & X^{-1} \in \text{MDA}\left(H_3^-\right) \text{ with shape parameter } \alpha \\ \iff \quad & -X^{-1} \in \text{MDA}\left(H_3^+\right) \text{ with shape parameter } \alpha \\ \iff \quad & \ln(X)^{\alpha} \in \text{MDA}\left(H_1^+\right) \end{aligned}$$

Since the Ψ-transform consists in taking the log of the probability density, Proposition 5 implies that the distribution over Ψ-radii is in the domain of attraction of H_1^+.

[18] A study of the convergence properties of Gaussian estimator in relation to their distribution of minima in density fall outside the scope of this thesis, and constitute potential future work.

5.7.2 Tail-fitting via Least Square Estimation

Except in trivial cases[19], there is no reason why $f(\mathbf{x})$ should exhibit the sort of spherical or ellipsoidal symmetry that makes the integration of f tractable in Section 5.4. Consequently, it is not possible to find a closed-form solution for the DPD $g(.;f)$, if f is a multimodal *MG*. Because of the overlap between kernels, $g(.;f)$ may present modes anywhere in the probability space, as shown in Figure 5.13.

However, we can hypothesize and experimentally verify that, because of the Gaussian nature of the kernels forming the d-dimensional MG, the DPD of the MG will, sufficiently close to its left endpoint $y = 0$, behave like that of a d-dimensional Gaussian. An example of this is given in Figure 5.13. The top-left graph shows the empirical DPD for $N = 10^7$ observations, which cannot be approximated by the uniform distribution (the DPD of any two-dimensional Gaussian). The bottom-right graph shows the histogram of the lower 2% of these observations, which is approximated reasonably well by the horizontal red line. We term this assumption *Asymptotic Gaussianity in Density* (AGD). Under this assumption, there exists a threshold γ in probability space, below which the DPD of an MG is assumed to behave like that of a single Gaussian kernel. Consequently, below $\gamma \in P_f$, the pdf of $g(.;f)$ is approximated by a function of the form

$$l(y;\beta,d) = \Omega_d \beta \left[-2\ln\left((2\pi)^{d/2}\beta y\right)\right]^{(d-2)/2} \tag{5.59}$$

for some value of $\beta \in \mathbb{R}_+^*$. Equation 5.59 is the same as eq. 5.14, where $|\mathbf{\Sigma}|^{1/2}$ has been replaced by the parameter β, hence the requirement that β be strictly positive. It is noteworthy that l is only dependent on the dimensionality d and the scalar parameter β. In other words, the form of l does not depend on the precise form of the covariance matrix but only on its (scalar) determinant; i.e.,the DPD is invariant under mean shifts, rotations, but also any transformation which does not affect the determinant of the covariance matrix. Similarly, the expression of the cdf L associated with l is adapted from Equations 5.32 and 5.33 after replacing $|\mathbf{\Sigma}|^{1/2}$ by β. The family of parametric distributions $l(.;\beta,d)$ may be used to estimate $g(.;f)$. Estimating β cannot be performed using by straightforward maximum likelihood estimation, because the

[19] such cases are limited to MGs for which all kernels share the same mode and level sets.

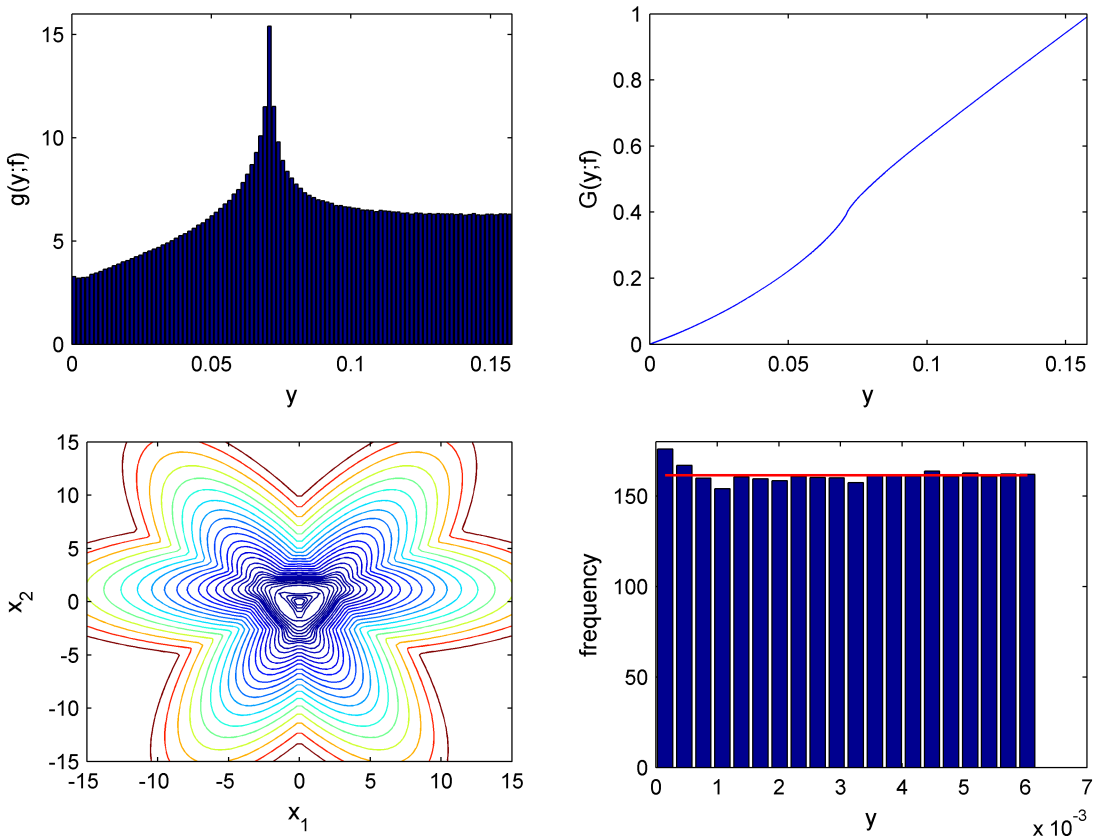


Figure 5.13: Top: empirical pdf and cdf of the MG used in Figure 5.2 ($N = 10^7$ observations). Bottom-left: level sets of the MG in data space. Away from the distribution modes, the level sets look increasingly like the contours of three intersecting ellipsoids. Bottom right: histograms of the lower 2% of observations (same data as those used to generate the top row graphs), with average of all bins in red.

Algorithm 5.1 Tail-fitting via LSE for mixture of Gaussians.

Let f be a d-dimensional, k-kernel mixture of Gaussians with probability space P_f.

1. Draw N observations from the MG and compute their densities $Y_N = (y_1, \ldots, y_N)$.
2. Plot an histogram of Y_N to assess visually where the modes (if any) of $g(,;f)$ occur in the probability space.
3. Set $\gamma \in P_f$ such that γ lies between 0 and the leftmost (non-zero) mode of the histogram, and γ corresponds to the lower percentile in which we are interested.
4. Extract the set of observed densities which are less than γ; $Y_N^- = \{y_i \in Y_N \,|y_i < \gamma\}$.
5. Form the empirical lower-tail histogram ; i.e. a j-bin histogram of Y_N, normalised so that the total area under the histogram equals the mass which lies left of γ; we define $(b_1, \ldots, b_j)$ to be the bin centres and $\hat{g}(b_i;f)$ the corresponding histogram values.
6. Minimise the quadratic error between the empirical lower-tail histogram and the function $l(y;\beta,d)$ with respect to β. ; i.e, find $\hat{\beta}$ such that:

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^*} \left\{ \sum_{i=1}^{j} \left(\hat{g}(b_i;f) - l(b_i;\beta,d)\right)^2 \right\}. \tag{5.60}$$

7. The pdf of the lower-tail is given by $l(y;\hat{\beta},d)$. The corresponding cdf, from which novelty scores are derived, is given by 5.32 or 5.33, with $|\Sigma|^{1/2} = \hat{\beta}$.

approximation is only valid on a fraction of the probability space, and the actual DPD and fitted l function are likely to have different supports. Instead, we propose to minimise the quadratic error between l and an empirical histogram of the densities less than γ, obtained after drawing N observations from the MG. This Least-Square Estimation (LSE) is naturally biased towards minimising the error near the left endpoint 0, because the probability mass shifts towards zero in the probability space as d increases. This is seen as a desirable effect, as we are interested in modelling the lower tail.

The choice of γ must be driven by the form of the MG and the range of values of m in which we are interested. Visual inspection of a histogram of densities can allow us to assess where the modes of the distribution lie, and we may then set γ between 0 and the leftmost non-zero mode. Additionally, the AGD assumption is likely to be more accurate closer to 0, therefore if we are interested in large values of m, γ can be chosen to be the $\frac{1}{m}^{\text{th}}$ percentile of the DPD. The value of N should be chosen so that enough observations lie below γ to construct a histogram. The method[20] is detailed in Algorithm 5.1. Figure 5.14 shows an example of tail-fitting for a

[20]The minimisation step (step 6) was performed using the fminbnd method in MATLAB.

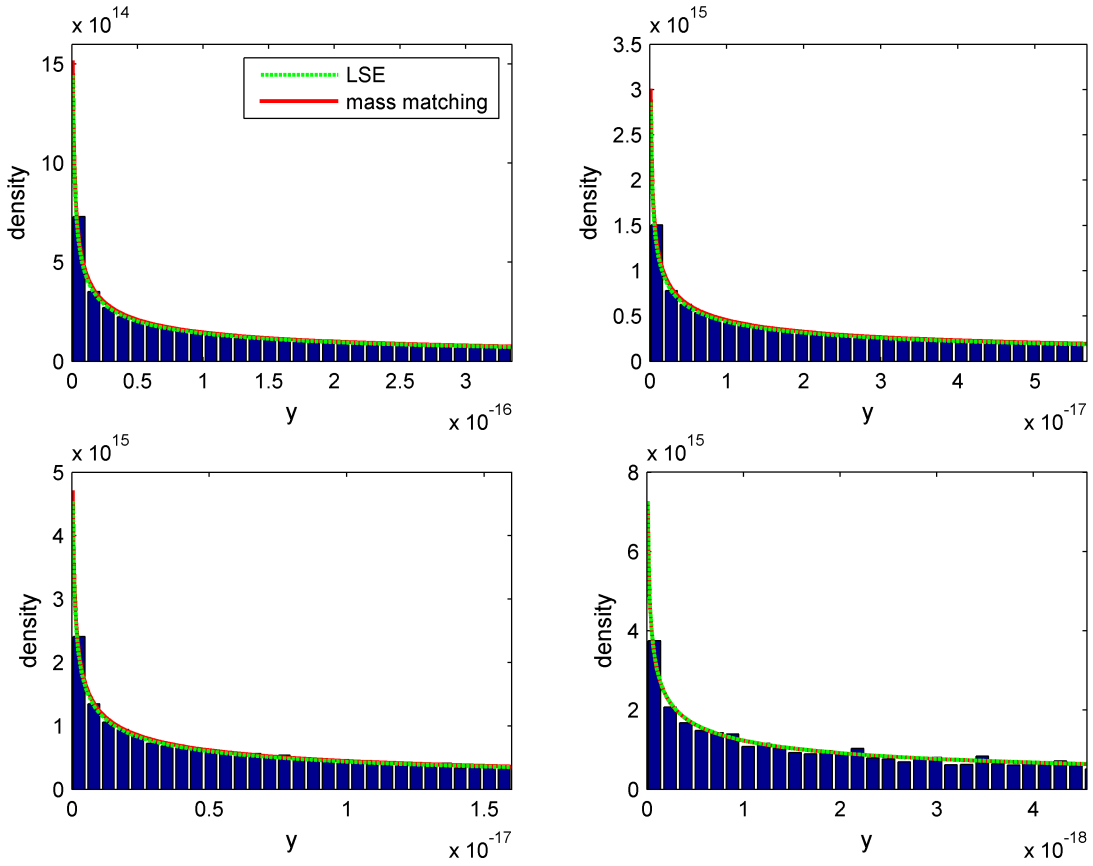


Figure 5.14: Tail-fitting under the AGD assumption. The generative distribution is a randomly generated 20-dimensional, 5-kernel MG. Histograms represent the empirical lower tail of T and all consist of 30 bins. γ is the 5-(top-left), 2-, 1- and 0.5-(bottom-right) percentile of $N = 10^6$ observations from the MG. The green lines were generated using the LSE method, and the red lines using the proposed mass-matching method.

20-dimensional MG for various values of γ, and demonstrates a good fit between the empirical tail distributions and the $l(.;\hat{\beta},20)$.

5.7.3 Tail-fitting via probability mass matching

We give here an alternative way of fitting the lower tail of the DPD under the AGD assumption, which is perhaps simpler, and does not require us to build a histogram of the tail.

Assuming that a threshold $\gamma \in P_f$ (below which the AGD assumption holds) has been set, then $G(\gamma; f)$ represents the probability mass which must lie below γ. This quantity can be estimated as was shown in Section 5.3, by considering the fraction of N observations from f whose densities are less than γ, to obtain $\hat{G}(\gamma; f)$.

Now under the AGD assumption, the estimate of the pdf $\hat{g}(.; f)$ belongs to the parametric family l defined in Equation 5.59. All the functions of this family are themselves pdfs. The probability mass which should lie below γ is known, and is exactly $G(\gamma; f)$. In other words, the probability mass modelled by the lower tail estimate should be the complement of the probability mass that does not lie in the lower tail. Therefore, an alternative estimator of the parameter of l is given by:

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^*} \{|l(\gamma; \beta, d) - \hat{G}(\gamma; f)|\}. \tag{5.61}$$

The minimisation procedure is simple, as l is a monotonically decreasing function of β (if β increases; i.e. the determinant of the covariance matrix increases, f is "flatter", and the mass of DPD shifts towards $y = 0$.). Equation 5.60 therefore has a unique solution, as long as $\hat{G}(\gamma; f) > 0$ (see Algorithm 5.2). A comparison with the LSE method is shown in Figure 5.14. The two algorithms appear to yield very similar results on this particular example. However, because it does not require us to build a histogram, the mass matching method introduces one fewer parameter (the number of bins) and does not require us to estimate values of the pdf anywhere left of γ, a region for which we have demonstrated earlier that estimates are less accurate than right of γ.

Algorithm 5.2 Tail-fitting via probability-mass matching.

Let f be a d-dimensional, k-kernel mixture of Gaussians with probability space P_f.

1. Draw N observations from the MG and compute their densities $Y_N = (y_1, \ldots, y_N)$.
2. Plot an histogram of Y_N to assess visually where the modes (if any) of $g(,;f)$ may occur in the probability space.
3. Set $\gamma \in P_f$ such that γ lies between 0 and the leftmost (non-zero) mode of the histogram, and γ corresponds to the lower percentile in which we are interested.
4. Extract the set of observed densities which are less than γ; $Y_N^- = \{y_i \in Y_N \,|\, y_i < \gamma\}$. The probability mass to match is given by $\hat{G}(\gamma; f) = \frac{|Y_N^-|}{N}$.
5. Find the value of parameter β for which $L(\gamma; \beta, d)$ best matches $\hat{G}(\gamma; f)$ (Equation 5.61).
6. The pdf of the lower-tail is given by $l(y; \hat{\beta}, d)$. The corresponding cdf, from which novelty scores are derived, is given by 5.32 or 5.33, with $|\Sigma|^{1/2} = \hat{\beta}$.

5.8 Conclusions

Table 5.1 summarises the multivariate EVT methods reviewed in this chapter, along with recommended uses. We discuss them here in order of increasing sampling requirements.

The single-kernel analytical method requires no sampling and offers closed-form solutions, but is only applicable for single kernels. The WTA method also does not require any sampling, but its assumptions break down for complex MGs. Next is the empirical cdf interpolation method, which only requires sampling to estimate the cdf of the DPD. In theory, arbitrary accuracy can be achieved, and closed-form solutions allow us to derive estimates of the EVDs and novelty scores from the cdf. However achieving high accuracy in the lower tail requires significantly more sampling than achieving the same accuracy in the upper tail, which is impractical. Next are the two tail-fitting methods under the AGD assumption, which also only require sampling of densities. Provided that γ can be estimated, there is no need for more sampling below γ to achieve accuracy. However, right of γ, the cdf of the DPD must be estimated using another method, as values given by the l function are unlikely to be appropriate. Finally, the Ψ-transform method allows us to achieve novelty scores with arbitrary accuracy, but at the cost of drawing extremes. Moreover, because it does not try to estimate the DPD, sampling must be performed for each value of m considered.

Figure 5.15 shows a comparison between four of these methods (only the LSE method is omit-

Method name (Section)	Main Assumption	Required sampling	DPDs	Novelty Scores	Recommended uses.
Single-kernel analytical (5.5,5.6)	None.	None.	Closed-form solutions.	Closed-form solutions for moderate values of m, asymptotically correct otherwise.	Only applicable to single kernels.
Winner-Takes-All (5.2.1)	Only the nearest kernel contributes to the EVD.	None.	Closed-form solutions for a single kernel.	Asymptotically correct, but can be significantly under- or overestimated.	MG with well-separated kernels, prototyping.
Ψ-transform (5.2.2)	None.	Extremes.	Not estimated.	Asymptotic only, although accuracy can be achieved even for moderate values of m.	Complex MGs, especially if the AGD assumption only holds for very small values of γ. Moderate values of m and d.
Empirical cdf interpolation (5.3)	None.	Only to estimate the DPD, no extremes.	Linearly interpolated using reference values. Less accurate in the lower tail.	Close to exact for any m. Less accurate in the lower tail of the DPD.	Any MG. Moderate values of m
DPD Tail-fitting via LSE (5.7.2)	AGD below γ	Only to estimate the DPD, no extremes.	Asymptotically correct below γ. Unknown above γ.	Asymptotically correct below γ. Unknown above γ.	Any MG for which γ can be estimated accurately. Large values of m.
DPD Tail-fitting via mass-matching (5.7.3)	AGD below γ	Only to estimate the DPD, no extremes.	Asymptotically correct below γ. Unknown above γ.	Asymptotically correct below γ. Unknown above γ.	Any MG for which γ can be estimated accurately. Large values of m.

Table 5.1: Summary of multivariate EVT methods presented in this chapter. "Asymptotically correct" refers to a guaranteed convergence to true values in the limit $m \to +\infty$. Moderate values of m are typically between 10 and 10^2. Large values of m are values greater than 10^2.

ted, as it gives a novelty map indistinguishable to that provided by the mass-matching method). The model distribution is a 15-kernel, 2-dimensional MG, whose centres are placed on a square grid (3 rows, 5 columns). All kernels have equal mixing weights and spherical covariance matrices. The latter are identical down a column. Their determinant increases across a row, yielding flatter kernels. The sample size is $m = 50$ for all methods. From the figure, it is clear that this is an MG on which the WTA method performs poorly, as it appears to overestimate scores on the entire data space. It performs especially poorly between the fourth and fifth columns of kernels, where assignment to a single kernel results in an important contribution to the density values by neighbouring kernels with large variances being discarded. The empirical cdf interpolation and the Ψ-transform (whose novelty map we use as our reference) method appear to yield very similar novelty scores. The semi-numerical tail-fitting method with mass matching appear to underestimate scores in regions of high kernel variance. We note that this is a particularly hard MG for this method, as the AGD assumption only appears to hold for $G(\gamma; f) < 10^{-3}$. Extrapolated scores for $y > \gamma$, although not guaranteed to be meaningful, are represented, and appear to be significantly overestimated between the fourth and fifth rows (lighter blue regions).

Overall, the trade-off between sampling requirements and accuracy of novelty scores is mediated by the number and magnitude of the sample sizes considered, and the desired level of significance. If m is small enough that interpolating an empirical cdf around the $\frac{1}{m}^{\text{th}}$ percentile can be performed to satisfactory accuracy (e.g., values of m less than 100), there is no need for more elaborate methods. If several values of m greater than 100 are considered, and a small level of significance required (i.e., we are interested in areas of the probability space closer to zero), then it may be preferable to make use of the AGD assumption. For $y > \gamma$, the interpolation method is more accurate and may be used to estimate DPD densities. Finally, extreme accuracy is the prerogative of the Ψ-transform method, but comes at a high computational cost.

The theoretical and numerical developments of Chapters 4 and 5 allow us to perform novelty detection in multivariate time-series generated by complex mixtures of Gaussians (Algorithms 4.1 and 4.2, replacing the DPDs and EVDs of univariate distributions by those of MGs). EVT-based methods are an improvement on the pdf thresholding method, insofar as they are designed to avoid false detections of normal variability that arises from multiple observations, and to

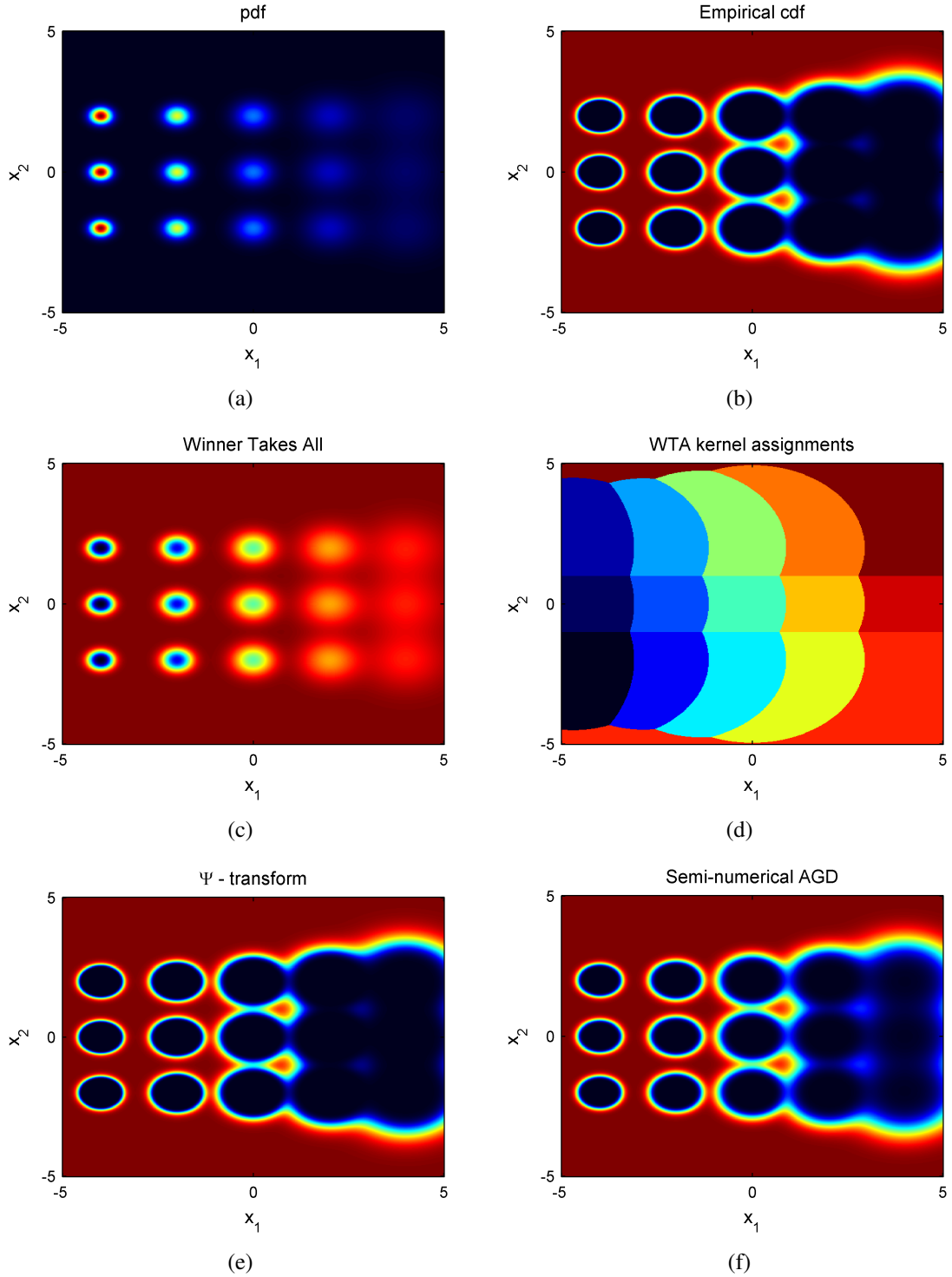


Figure 5.15: Comparison of 4 multivariate EVT methods for MGs. (a) pdf f of the MG, (b) novelty map using the empirical cdf interpolation method ($N = 10^6$), (c) novelty map using the WTA method (and exact distributions), (d) map of "nearest kernel" use for kernel assignments in the WTA method, (e) novelty map obtained using the Ψ transform method ($N = 10^4$), (f) novelty map using the AGD assumption with mass matching ($N = 10^6$, $G(\gamma; f) = 0.01$). The color scale is identical for all novelty maps and ranges from 0 to 1. The sample size is $m = 50$ for all methods.

ensure that the novelty scores they yield are meaningful with respect to hypothesis testing. However, in classification, a threshold must be set on novelty scores, which is equivalent to setting a selecting a level set of the pdf beyond which an extreme observation is considered abnormal, and it is therefore not fundamentally different to thresholding the pdf. We show in Chapter 6 on a real data set that EVT-based methods limit significantly the number of false detections, but remain sensitive to extreme transient noise. More robust novelty classifiers can be designed specifically to cope with transient noise using general order statistics, which are the focus of Chapter 7.

Chapter 6

Application of Multivariate EVT to the UPMC Data Set

6.1 Overview of this chapter

Having introduced classical Extreme Value Theory and a multivariate extension in Chapters 4 and 5, respectively, we proceed to apply our method to the UPMC data set used by Hann (2008) and described in Chapter 2. First, we consider a univariate case study in which the aim is to detect events caused by abnormal heart rates, using the HR time-series. Second, we consider a multivariate case study using the HR, BR, and SpO_2 time-series. The aim in both cases is to identify those periods of physiological data corresponding to the tMET events described in Chapter 3.

6.2 Assessing performance

6.2.1 Performance metrics

Given a set of positive and negative labels and a set of positive and negative classifications, the statistics used in this thesis are defined as follows:

$$\text{Se} = \frac{\text{True Positives}}{\text{Positives}}, \tag{6.1}$$

$$\text{Sp} = \frac{\text{True Negatives}}{\text{Negatives}}, \tag{6.2}$$

$$\text{Ac} = \frac{\text{True Positives + True Negatives}}{\text{Positives + Negatives}}, \tag{6.3}$$

$$\text{AUC} = \int_0^1 \text{Se}(\text{Sp})\, d\text{Sp}, \tag{6.4}$$

where Se is the *sensitivity* (the fraction of positives retrieved by the classifier), Sp is the *specificity* (the fraction of negatives retrieved by the classifier), Ac is the *accuracy* (the fraction of correct classifications made by the classifier), and AUC is the *area under the receiving operating characteristic* (the probability that a randomly chosen positive is given a higher classification score than a randomly chosen negative by the classifier). Throughout the remainder of this thesis, AUCs are calculated using a trapezoidal rule. When reporting results at the levels of the monitoring hour, the monitoring day, the patient, all the threshold values present in the data set are used to generate graphs and calculate AUCs. When reporting results on an observation-by-observation basis, the threshold step is set so that no fewer than 10^3 distinct threshold values are considered.

AUCs resulting from Hann (2008)'s classifiers and those introduced in this thesis are compared using the method by Hanley et al. (1983)'s. The method, suitable for comparing AUCs derived from the same test cases allows us to assess the statistical significance of the observed variations. We used the MATLAB implementation of the method referred to and used in (Mayaud et al. 2013).

6.2.2 Levels of detection and class definitions

We noted in Section 2.2.6 that Hann (2008) used a combination of event-based and patient-based statistics to form a receiving operating characteristic (ROC curve), as defined by Equations 2.1 and 2.2. The ROC curve gives information on how likely the C" events are to be identified by a classifier (Sensitivity is plotted on the y-axis) and how likely false alerts are to be generated over a patient's stay in the ward (1 - Specificity is plotted on the x-axis). The combined use of event- and patient-based statistics may be misleading, as it makes it possible to obtain ideal values (i.e., sensitivity equal to 1 and specificity equal to 1) even with a less than ideal classifier: if untimely alerts (false positives) only occur for positive patients, $\mathrm{FN}_e = 0$, and $\mathrm{FP}_p = 0$, then $\mathrm{Se}_e = 1$ and $\mathrm{Sp}_p = 1$ (with the notations in Section 2.2.6), which might lead us to the wrong conclusion, whereas it is possible for this classifier to alert regularly for positive patients, without affecting those statistics.

Conversely a patient who had events (positive patient) for which alerts are triggered only outside its event windows (false positive alert) is not accounted for in these statistics. In other words, we can add as many of these patients to the patient population without causing Sen_e or Sen_p to depart from 1, values we argue should be reserved only for a classifier which perfectly separates two classes.

The problem naturally arises from the fact that, as a result of using a combination of definitions for "positives" and "negatives", the "classes" become ambiguous. The resulting ROC-like curve loses its probabilistic interpretation, which is only retained if the positive and negative classes are unambiguously defined, and mutually exclusive.

The difficulty comes from the gap between the high-level question these definitions aim at answering and the time resolution at which they operate. Detection algorithms such as those described in Hann (2008) and in this thesis are designed to estimate the patient status every time a new vital-sign observation is made. The estimate is based on this latest observation, and perhaps a small number of observations made in the preceding few minutes (accounting for a post-processing step such as time windowing). In a classical two-class learning framework, validation of such algorithms would require a set of unseen observations, each labelled

"normal", or "abnormal", so that the validation step and the learning step are consistent. In this respect, observation-based statistics (sensitivity, specificity, AUC, etc.) are the only statistics which soundly assess the relevance of the learning step. They are not, however, the only metrics which need to be reported, for two main reasons.

Firstly, labels for individual observations are not available. The C" and MET labels were obtained using procedures which include a hysteresis component (i.e, the activation and deactivation rules). The classifications given by human reviewers in both cases are likely to rely on the identification of trends and correlations between vital-sign channels, resulting in event windows which contain both vital-sign measurements outside normal physiological ranges, as well as measurements which are within acceptable physiological ranges, but are included in the event windows due to earlier abnormal measurements.[1] This motivated the introduction of a post-processing step in Hann (2008)'s work, and our transition to classifying windows of vital-sign data.

Secondly, we state in Chapter 2 that the overall aim of this work is to detect the onset of physiological abnormality while keeping the false alert rate to a minimum. This minimum could be understood to be the minimum of false alerts per period of fixed length (e.g. per monitoring day), or the minimum of patients whose time-series data result in false alerts. Additionally, the distributions of misclassifications in time and across patients are of interest. Systems that generate a large number of false alerts over short periods of time could, for instance, be considered to be more useful than those which generate false alerts at regular intervals. Systems that generate a large number of false alerts only in a few patients should arguably not be considered equivalent to those which generate the same number of false alerts in every patient.

In order to gain the necessary insight into these higher-level aspects of system performance, while remaining within the scope of well-defined two-class classification problems, we define five levels of analysis for which statistics can be reported.

[1] To clarify, a heart rate measurement of 110 bpm can be included in an event window if the previous measurements are higher or erratic, but would not if it appeared as a single occurrence following similar or lower stable heart rate measurements.

Level (i): the observation

Novelty detectors that output a classification ("normal" or "abnormal") at the time of each observation, may be consistently analysed in terms of their performances at the scale of an individual observation. This would involve comparing each classification of an observation to its true class. True classes are defined as "positive" for observations occurring during event windows, and "negative" for observations occurring outside event windows. The observation-based true positive rate, false positive rate, sensitivity, and specificity are calculated on the basis of observations being correctly or incorrectly classified, as they would in a classical two-class classification problem.

This method of assessing the performance of an observation-by-observation classifier process is natural, and benefits from the large number of observations available. However, it is not sufficient in the context of vital-sign monitoring. Firstly, because events are rare (C" events represent 0.24% of the total monitoring time of the ward population, Table 2.6.), the accuracy of the naïve classifier on C" events is 99.76%, meaning that more sophisticated classifiers should have an accuracy around or preferably above this value. Specificity, sensitivity, and AUC would in general be more relevant statistics, yet their interpretation is made difficult by the likely presence of heterogenous observations both inside and outside the event windows; i.e, not all observations in an event window are "abnormal", and not all observations outside event windows are "normal". Secondly, an ideal alerting system would trigger a single alert per event window, as close to its start as possible. Consequently, there is no need to classify every observation in an event window as being "abnormal", as long as at least one observation at or near the start of the event window is. Lastly, reporting statistics (sensitivity, specificity, etc.) on the complete set of observations is equivalent to considering all patient recordings as one long recording (formed by appending individual patient recordings to form that single recording). This gives insufficient information on the variability in performance over time, or between patients, which is of clinical interest.

Level (ii): the event

At this level of analysis, an event is said to be correctly detected if a set number of observations, or percentage of observations in the event window are classified abnormal. The choice of this number or percentage can be made dependent on the importance of detecting events early, our confidence in the activation and deactivation times, and the relative lengths of false alerts to true alerts. In practice, we will assume that we only need one abnormal classification in the event window to trigger an alert, regardless of the event duration. This allows us to define true positive alerts (alerts triggered during an event window) and false positive alerts (alerts triggered outside any event window). "Negative events" and therefore "negative event-level alerts" are unfortunately not easily defined. At the level of the event, we will therefore only report the event-based true positive rate.

Levels (iii) and (iv): the monitoring hour and the monitoring day.

Despite not having a standard way of defining negative alerts at the level of the events, we may consider windows of vital-sign data of equal durations. If recordings are divided into consecutive non-overlapping windows of duration T, then a "positive" window can be defined as one that overlaps with an "event" window, and a "negative" window as one that does not overlap with any "event" window. For the purpose of this work, we select two durations: $T = 1$ hour (a monitoring hour), and $T = 24$ hour (a monitoring day). The former is selected because a large majority of MET events are shorter than 1 hour (as shown in Figure 3.8), and this value of T therefore minimises the number of events which spread over multiple windows. Also, events occurring within 1 hour of each other are highly likely to be correlated. The monitoring day is chosen as a more demanding alternative, used to assess the robustness of performance at the level of the monitoring hour.

Statistics on the classification of monitoring hours and days are defined using the standard definitions.

Level (v): the patient

A "positive" patient is defined as being one who had at least one event, and a "negative" patient is defined as a patient who did not have any event, regardless of the lengths of their recordings, or the number of events present in their recordings. Patient-based statistics are straightforwardly defined using the standard definitions.

6.2.3 Handling of missing data

In Hann (2008) missing data is handled as follows:

- if at a given timestep, a single vital-sign channel does not have a valid observation, it is assumed that the observation at this timestep is the channel's mean in the training data set;
- if at a given timestep more than one vital-sign do not have valid observations, the data at that timestep are ignored.

Windows which include missing data are later ignored in the novelty detection procedure. Replacing missing observations by the means of their channels in the training data set has the effect of *normalising* the multi-channel observations at these times, by assuming that they are closer to the centre of the model distribution than they are likely to be.[2]

The approach chosen here is not to make any assumption on the missing data. When observations are missing from a detection window, the number of observations (the parameter m used to calculate the novelty score) is adjusted so as to match the number of actual observations in the window, and the novelty scores are computed using this adjusted value of m. While the detection window is of fixed time duration, the number of observations is therefore allowed to vary, thus preventing us from making any assumption on the values of the missing observations.

[2] A more appropriate method of handling missing data would be to replace the missing observations by the conditional mean of the missing channels in the training data set, conditioned on the values of the observed vital-sign channels, thus introducing a normalising effect to the multivariate observation.

6.3 Univariate Novelty Detection on the UPMC Phase 1 data set

This section demonstrates the application of the multimodal, multivariate EVT technique presented in Chapter 5 to the heart rate time-series recorded in the first phase of the UPMC trial (introduced in Section 2.1). Performance is reported at the observation, event, monitoring hour, monitoring day, and patient level, using the statistics defined in Section 6.2.

First, a model of normality is constructed using "normal" training data. The unconditional probability distribution of a set of identified normal vital-sign data is estimated, using a mixture of K Gaussians. This model of normality is then assumed to be the underlying data-generating distribution; i.e., we assume that patient data are i.i.d. according to this distribution.

6.3.1 Model construction

The set of 170 patients who did not have any event (either MET or C'), which we shall term $\mathcal{P}_\varnothing$,[3] is randomly divided into two halves: one is used as a training set for model construction, the other as a "held out" control group to be used subsequently for testing.

An unconditional probability density function is estimated using the heart rate observations (subsampled every 10 seconds) for the entire recording of each patient in the training set, using a Gaussian Mixture Model (GMM) of the form:

$$f(\mathbf{x}) \quad = \quad \sum_{k=1}^{K} \frac{\pi_k}{C_{d,k}} \exp\left(-\frac{M_k(\mathbf{x})^2}{2}\right) \tag{6.5}$$

where K is the number of kernels in the GMM. The parameters of the k^{th} kernel f_k are: its centre $\boldsymbol{\mu}_k$, its covariance matrix $\boldsymbol{\Sigma}_k$, the Mahalanobis radius implied by covariance matrix

[3] $\mathcal{P}_\varnothing$ is the set of patients who were the most stable. It is our intention to use the tMET labels as event labels in the rest of this chapter. Including patients who had no MET event but one or more C' event in the set of stable patients would have been possible, but there is an ambiguity as to whether such patients were actually stable or not during their stay, since neither MET nor C events can be considered the ground truth. We therefore opted for the simpler option of excluding them for the purpose of demonstration. In clinical settings, more attention should be paid to these patients, and the definition of events.

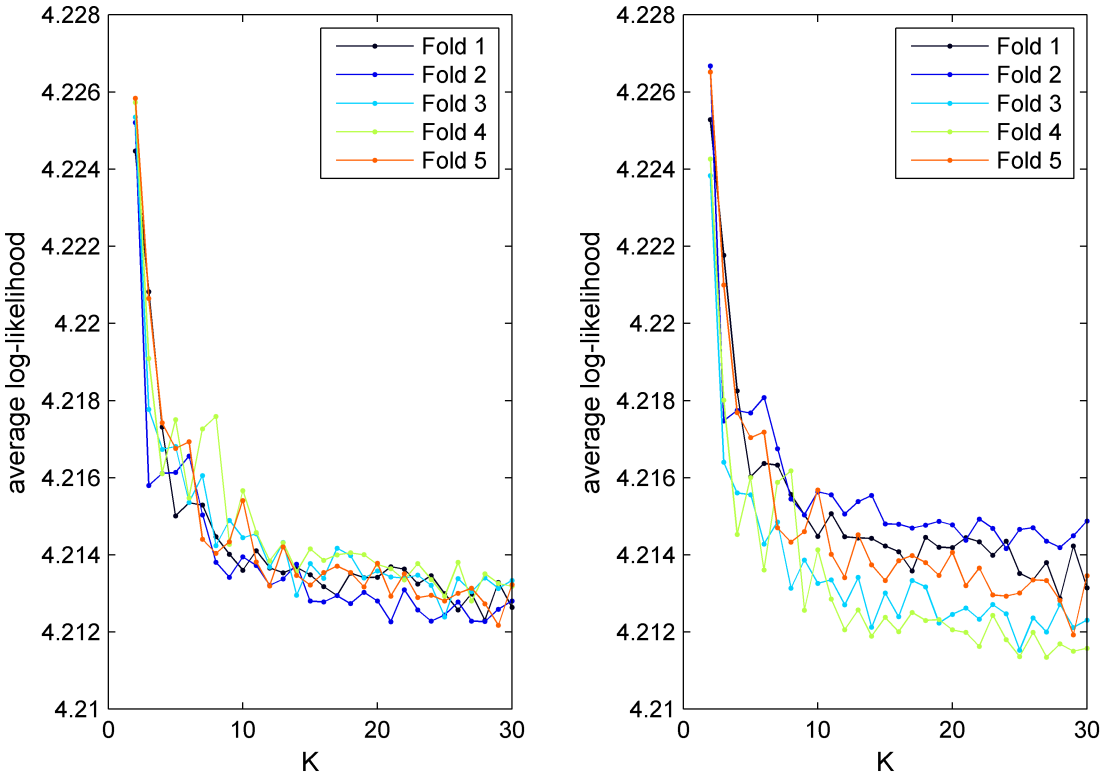
$M_k(\mathbf{x}) = \left((\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)^{1/2}$, its normalisation coefficient $C_{d,k} = (2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}$, and its mixing weight π_k.

For some value of K, a GMM is fitted to a training set of vital-sign measurements, using the Expectation-Maximisation algorithm. The number of kernels K must be chosen to be large enough such that the resulting GMM accurately describes the training data, yet small enough such that the GMM does not overfit. We used a 5-fold cross-validation procedure to select a suitable value of K: the training data, (1.4×10^6 HR observations), were randomly split into five subsets of equal sizes. Five iterations were performed, in each of which one the five sets is used as "held out" validation data, while the remaining four were used for training the GMM. We then compared the contribution to the log-likelihood of each observation ($-\ln(f(\mathbf{x}))$) (i.e., the log-likelihood of each set divided by the number of observations in this test) to determine (i) whether increasing the number of kernels led to overfitting, and (ii) the number of kernels above which increasing K did not result in a significant decrease of the training data log-likelihood.

Figure 6.1 shows the results of this process for the UPMC heart rate data. For all folds, the training log-likelihood decreases considerably as K is increased from 2 to 10, and then remains stable as K increases above 10. Hence, we conclude that a 10-kernel GMM seems to explain the distribution of the training data almost as effectively as a more complex GMM. The log-likelihood averages per sample for the test data follow the same trends, as shown in the figure. Comparing the two graphs, we note a marginal difference in log-likelihood between training and validation data, and no increase of the average log-likelihood for the validation data, even for larger values of K, which indicates that increasing K to 30 does not result in significant overfitting. The results for individual folds show that the average log-likelihood for the validation set compared to that for the corresponding training set is on average higher for Folds 1 and 2, lower for Folds 4 and 5, and similar for Fold 3.

In consequence, we conclude that K can be set to a value between 10 and 30 without overfitting, or loss of generality.

Since the log-likelihood for validation data decreases only slowly as K increases for $K \geq 10$, one could, in accordance with the principle of Occam's razor, decide to use $K = 10$ kernels.



(a) (b)

Figure 6.1: Average log-likelihood $-\ln(f(x))$ per observation on (a) the training set and (b) validation set for each of the five folds in the cross-validation process. K is the number of kernels in the GMM.

However, having demonstrated the absence of overfitting, we may choose a value of $K > 10$ for which the GMM better estimates the tails of the distribution. The need to model accurately the extreme data for stable patients means that we are less interested in correctly estimating the shape of the distribution near its modes if it impairs accurate estimation of the distribution of the tail data. In that respect, even if a GMM is a universal estimator for the distribution, this does not necessarily mean that a good approximation of the distribution as a whole provides an equally good estimation of its tails. We therefore compare the pdfs estimated by the GMMs to the empirical distribution of the training data. We retrain the GMMs for K = 2 to 30, this time using all of the available training data, without the need for cross-validation at this stage.

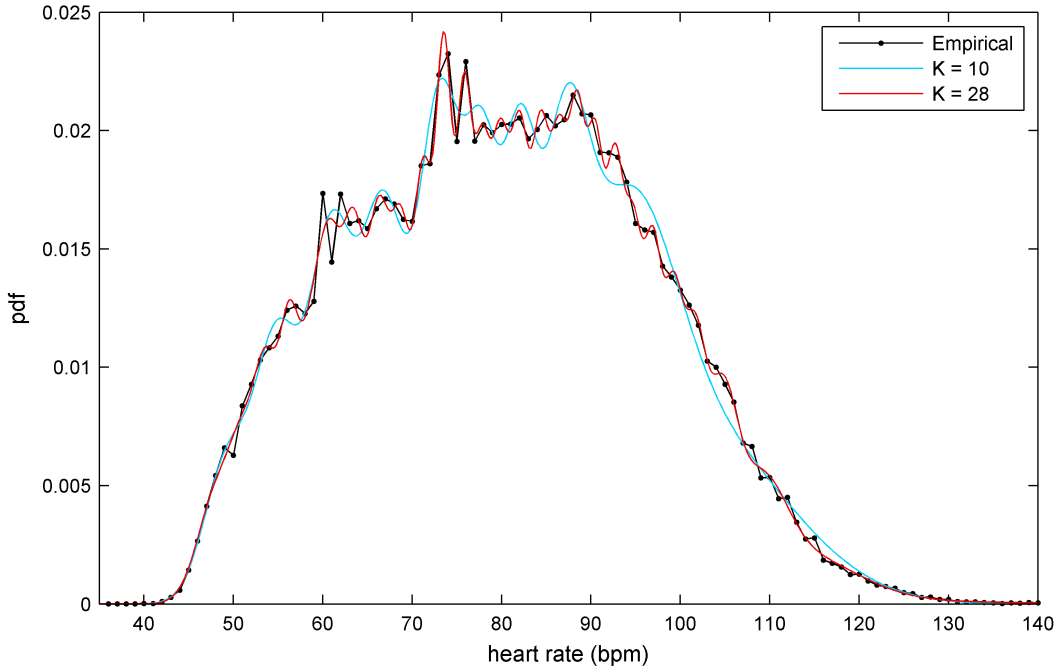
Figure 6.2 shows the empirical distribution of heart rate in the training population, as well as the pdf of two GMMs obtained using $K = 10$ and $K = 28$. Inspection of these pdfs reveals that while there is no significant gain in average log-likelihood when increasing K from 10 to 30, this is not true for the tails of the distribution. Since our interest is in modelling the tails accurately, when the log-likelihoods are the same overall, a more complex model is preferable. The choice of $K_{HR} = 28$ was made after observing no incremental improvement for $K > 28$.

The 28-kernel GMM will thereafter referred to as *the HR model*, or f_1.

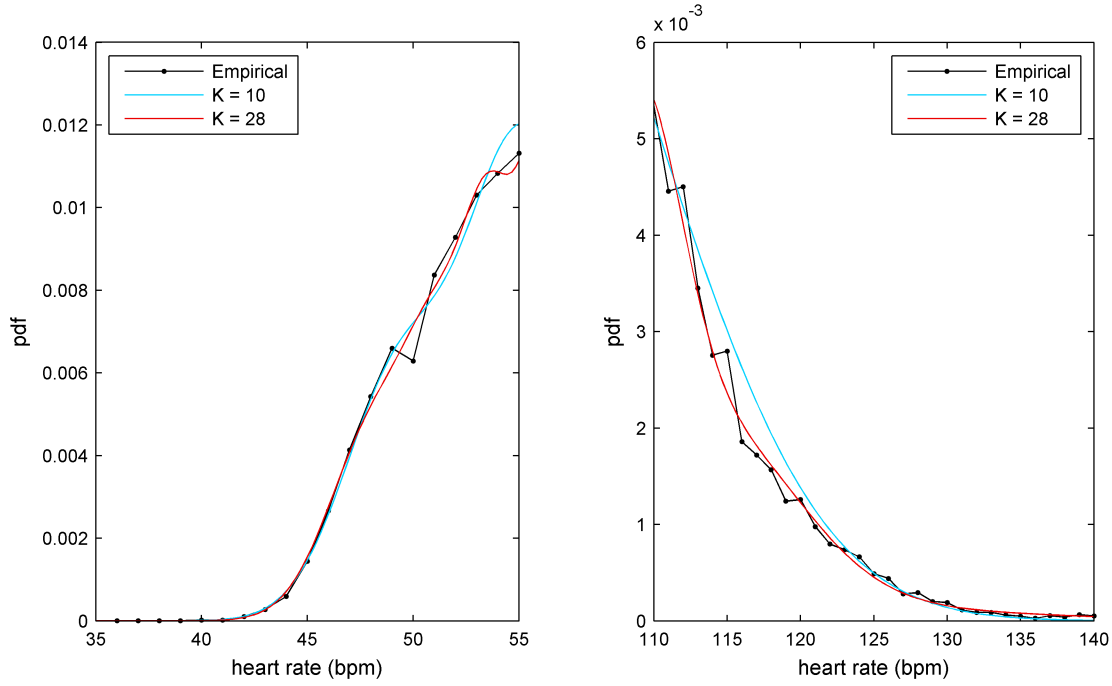
It may be noted that there exist better-founded method of choosing the model's number of kernel.Examples include shrinkage priors for Bayesian prediction (Komaki 2006), or Minimum Length description (Grünwald 2007). The cross-validation approach was deemed adequate for the purpose of demonstration in this chapter in the next, noting that the exact number of kernels (within reason) is not a dominant source of error.

6.3.2 Differences with existing work

The approach described above fundamentally differs from that of (Hann 2008), which consists of aggregating vital-sign measurements from all patients and discarding those that are the most extreme with respect to all measurements of all patients. In doing so, it is assumed that artefactual and physiologically extreme measurements are discarded before learning the model of normality. This process is "unsupervised" (in the sense that all data from all patients, whether



(a)



(b) (c)

Figure 6.2: Empirical distribution of HR from two GMMs. (a) Empirical distribution of HR in the training population (91 patients), superimposed on the pdfs from the GMMs with 10 kernels (blue) and 28 kernels (red). Magnifications of the lower (b) and upper (c) tails.

"normal" or "abnormal", are used), although the fraction of extreme data (20%) to be discarded needs to reflect the amount of either artefactual or physiologically abnormal observations that are expected to be in the "normal" data set. When testing the resulting model of normality using a different data set, it is therefore assumed that the prevalence of these abnormal measurements are comparable, ideally the same.

Classifying patients in a supervised manner allows us to identify those patients who did not have any of the previously identified events, and use as our training data all the observations in these patients' recordings. Since patients who experienced one or more events are excluded from the set of training patients, it is expected that few data corresponding to significant physiological abnormality appear in the set of training observations.

Another difference between our procedure and that described by Hann (2008) is that we use a Gaussian Mixture Model to form the unconditional pdf, rather than a Parzen Windows estimator. Given the large size of the training set, it is impractical to define a Parzen Windows estimator using every training point, and it is necessary to subsample them. This was performed using a K-means clustering procedure, where K was set to 500. The 400 most normal kernels were subsequently identified, and their centres used as the basis for a Parzen Windows estimator. It has since been shown by Wong (2011) that the K-means procedure is unsuited for subsampling, because it tends to underestimate the probability mass assigned to the tails. Additionally, the number of samples used to form the Parzen Windows estimator is heuristic, as is the number of cluster centers to be discarded. Finally, the resulting Parzen Windows estimator is a relatively complicated model comprising 400 kernels, where we have demonstrated that models comprising only a few tens of kernels may be suitable.

6.3.3 Novelty detection algorithm

Notation

The notation used in in Chapters 4 and 5 is simplified and summarised in Table 6.1. For the f_1 model, we perform novelty detection using Algorithm 4.2, using a range of values for the

Symbol	Definition	Reference
f_1	pdf of the model	-
F_1	cdf of the model	-
$D_1 = \mathbb{R}$	vital-sign (HR) space	-
$P_1 = P_{f_1} = f_1(D_1)$	probability space of the model	Section 4.3.3
y	density in P_1	Definition 5
$G_1 = G(.; f_1)$	cdf of the DPD	Definition 5
$g_1 = g(.; f_1)$	pdf of the DPD	Definition 5
$G_1^-(.;m) = G^-(.; f_1, m)$	EVD of minima in P_1 given a value of m	Proposition 2
$G_1^e(.;m) = G^e(.; f_1, m)$	natural novelty function in P_1	Equation 4.37
$F_1^e(.;m) = F^e(.; f_1, m)$	natural novelty function in vital-sign space.	Equation 4.38

Table 6.1: Summary of functions and notations for the HR model, f_1.

sample size parameter m. We shall also refer to "the distributions f_1, g_1, *etc*" to denote their estimates rather than introducing additional notation.

Distribution of Probability Density (DPD) for f_1

The first step in Algorithm 4.1 is to find an estimate of the DPD, G_1. The complexity of f_1 makes it impossible to find a closed-form solution for G_1. Instead, we have shown in Section 5.3 that an empirical distribution of densities formed using a large number of random observations from the model is a suitable approximation of the DPD, so long as the value of m is reasonably small (e.g. $m < 10^2$). In the present case, we use $N = 10^7$ (number of observations), and $P = 10^3$ (number of evenly distributed y values). The resulting cdf G_1 and its pdf g_1 are shown in Figure 6.3 The pdf g_1 exhibits a large number of modes, which is expected given its number of kernels. Most of these modes are localised in the upper part of the probability space P_1, while f_1 is approximately flat over the lower fifth of the probability space, P_1.

Extreme Value Distributions (EVDs) of f_1

The distributions of minima for G_1, termed $G_1^-(.;m)$ are obtained using Equation 5.6. Missing values of G_1 are linearly interpolated using the values estimated in the previous step. The graph on the lefthand side of Figure 6.4 shows $G_1^-(.;m)$ for sample sizes m ranging from 1 to 60. For $m = 1$, the EVD $G_1^-(.;1)$ (shown by the black line) is exactly G_1 (Figure 6.3). As m is increased,

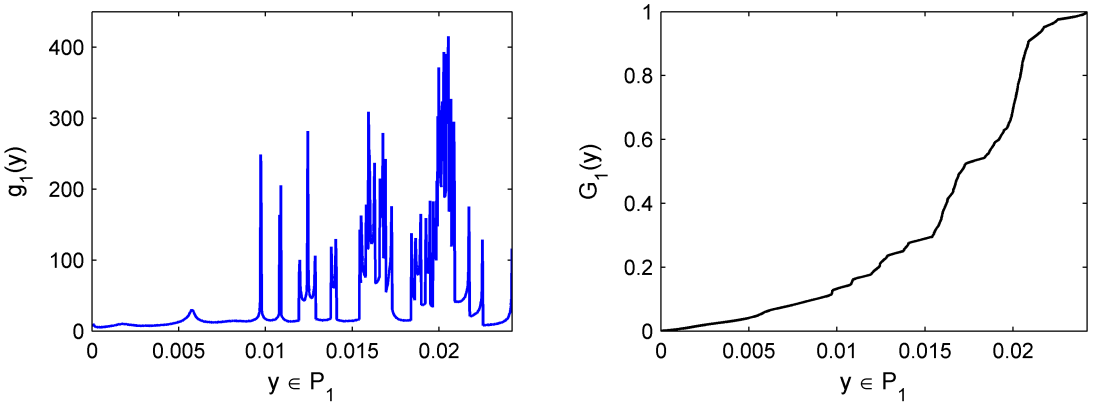


Figure 6.3: DPD of the univariate HR model: pdf, g_1 (left), and cdf, G_1 (right) over the probability space P_1. The 28-kernel GMM f_1 is that represented in Figure 6.1. Both functions are approximated using 10^7 randomly-drawn observations.

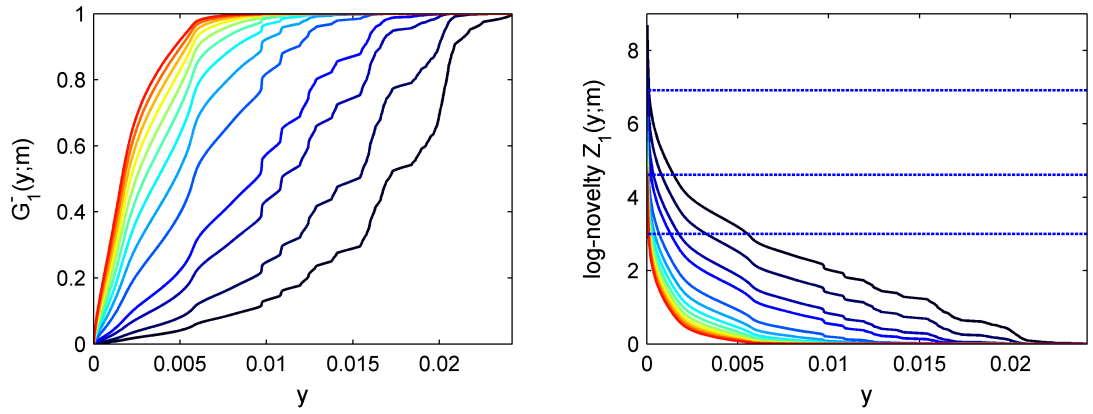


Figure 6.4: EVDs of natural (left) and log-novelty in probability space (right). Values of the sample size parameter m range from 1 (black) to 60 (red). The legend for both graphs is the same as that of Figure 6.5. In the log-novelty graph, the horizontal dotted lines correspond to natural novelty scores (or quantiles of the EVDs) of 0.95 (lowest line), 0.99, and 0.999. Conversely, they correspond to significance levels α of 0.05, 10^{-2}, and 10^{-3}.

the probability mass of the EVDs moves to the lower part of the probability space; i.e., away from the modes of f_1. As a consequence of this effect and the nature of Equation 5.6, the EVDs become increasingly smoother.

The graph on the righthand side in Figure 6.4 shows the log-novelty functions associated with these EVDs. Novelty scores are higher for lower values of m, because smaller densities are more unlikely to be observed if the sample size m is smaller. Novelty curves are also smoother as the probability mass of the associated EVDs moves away from the modes of the distribution, because the modes introduce large changes in density in G_1 and G_1^-.

Novelty scores in HR space

$G_1^e(.;m)$ is obtained from of $G_1^-(.;m)$ using Equation 4.37. Natural novelty scores in HR space (the data space) follow from Equation 4.38, which in conjunction with Equation 4.36 with $\phi(z) = -\ln(1-z)$ gives us the log-novelty scores. They are respectively denoted by:

$$z_1(x;m) = G_1^-\left(f_1(x);m\right), \tag{6.6}$$

$$\text{and} \quad Z_1(x;m) = Z(x;f_1,m) = -\ln\left[G_1^-\left(f_1(x);m\right)\right]. \tag{6.7}$$

Natural novelty and log-novelty curves are shown in Figure 6.5 for the same values of m as in Figure 6.4. The log-novelty representation appears to amplify the asymmetry in the tails of f_1. In the range 50 - 100 bpm, f_1 does not appear to be obviously skewed (red line in Figure 6.2), but its tails have greatly different shapes, due to the fact that a decrease of 1 bpm from a heart rate below 50 does not occur as often as an increase of 1 bpm from a heart rate above 100. The focus on accurately estimating the tails of the distribution rather than its core probability mass corresponds to an increased likelihood of capturing this asymmetry.

Range of m

Figure 6.5 shows that HR < 48 bpm or HR > 110 bpm are significantly abnormal at the 0.05 level of significance, if assessing individual observations ($m = 1$). These are the values of HR

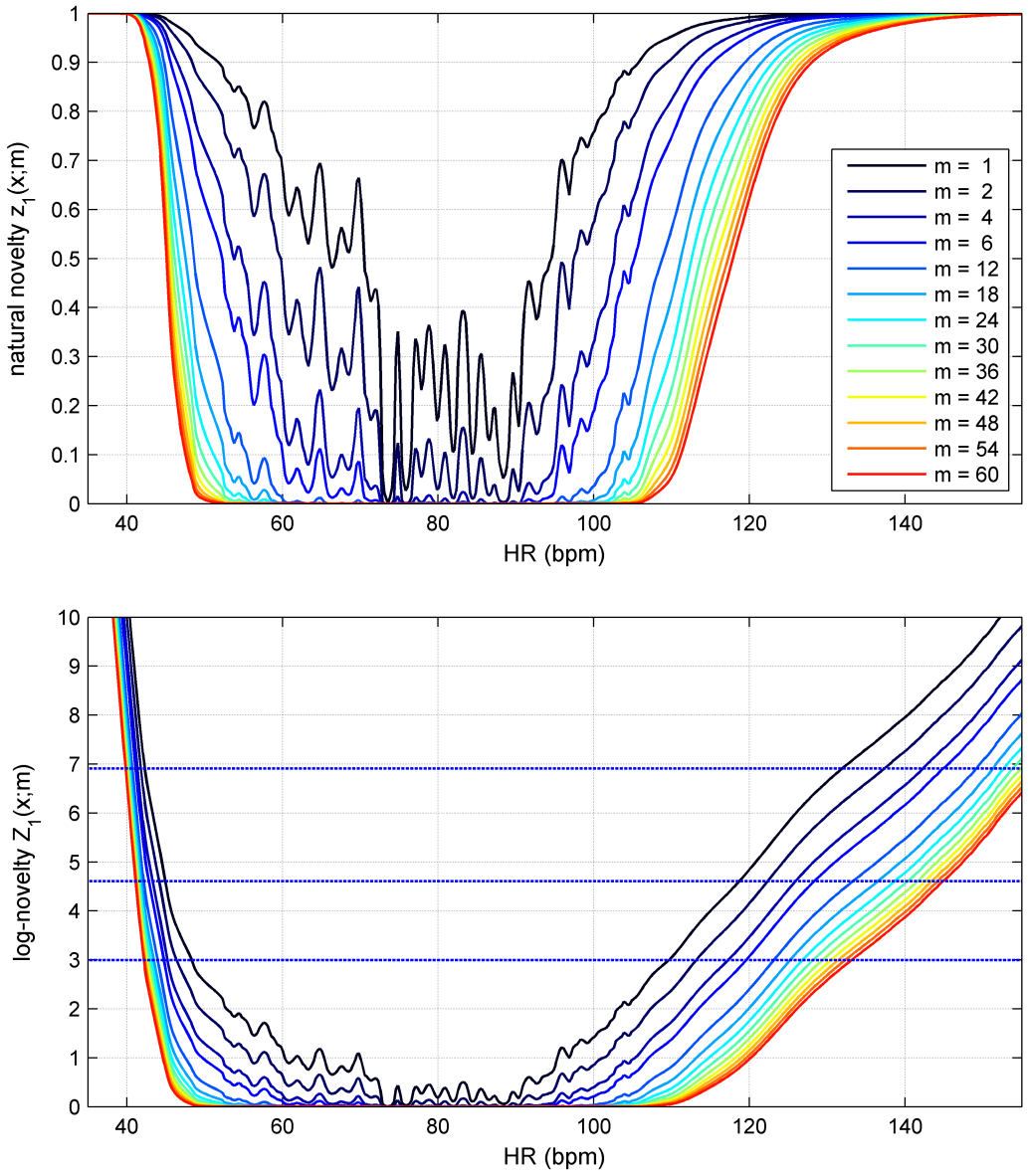


Figure 6.5: Novelty (top) and log-novelty scores (bottom) in HR space for values of m between 1 and 60. As in Figure 6.4, the horizontal dotted lines correspond to natural novelty scores of $z_1(x;m)$ =0.95 (lower line), 0.99, and 0.999.

for the line corresponding to $m = 1$ which intersects with the horizontal line corresponding to $\alpha = 0.05$. If $m = 60$, these thresholds become 42 bpm and 133 bpm. These differences are direct consequences of classifying sets of data points (as is performed by our method) rather than individual points (as has been the focus of previous work). We note that these thresholds may move indefinitely further from the centre of the distribution, as m is increased. Consequently, there exists values of m for which these thresholds are guaranteed to reach physiologically implausible values.[4]

Since the sampling period of the HR data is 10 seconds (as noted in Section 6.3.1), $m = 60$ corresponds to a window length of 10 minutes. If the window length is increased to 24 hours ($m = 8640$) then the corresponding thresholds would be HR = 39 and HR = 162 bpm for $\alpha = 0.05$, or HR = 38 and HR = 169 bpm for $\alpha = 0.01$. It is probable that a patient would be considered to be in a physiologically abnormal condition long before these thresholds were reached. It would therefore not be appropriate to use the current number of observations since the start of a patient's stay as the value of m. We therefore adopt the use of a sliding window of fixed length, which represents the assumption that physiology has limited memory.

Considering the 0.05 significance thresholds, we find that:

- the lower threshold decreases by only 1 bpm between $m = 24$ and $m = 60$ (from HR = 43 to HR = 42 bpm);
- the higher threshold increases rapidly, between $m = 1$ and $m = 30$, up to HR = 128 bpm, then increases more slowly up to 132 when $m = 60$.

Increasing m from 30 to 60 therefore has only a small effect on the physiological thresholds. For clarity of demonstration, we exclude all values of m above 30 (corresponding to a window length of 5 minutes), and only consider values between $m = 1$ and $m = 30$ as candidate values for m.

Algorithm 6.1 Novelty Detection procedure in HR time-series.

Initialisation:

1. Estimate the unconditional probability distribution of HR observations f_1.
2. Compute estimates of $G(.;f)$
3. Choose a value for the detection window size $m \in \mathbb{N}^*$ and estimate $G_1^-(.;m)$.
4. Choose a value for the level of significance $\alpha > 0$, for determining the threshold on log-novelty $Z_1(x;m)$.

For a time-series of HR observations:

1. Starting with the first window at the beginning of an HR time-series:
2. Find the m-extremum e_m of the window with respect to f_1; i.e., the HR observation of lowest density y.
3. Compute the window's log-novelty score $Z_1(x;m) = -\log[G^-(e_m;f,m)]$.
4. If $Z_1(x;m) > \alpha$, an alert is deemed to have occurred.
5. Otherwise, output 0 (normal state)
6. Advance the window by 10 seconds.
7. Repeat steps 2 to 6 until the end of the time-series is reached.

Novelty Detection procedure

Our novelty detection procedure is given in Algorithm 6.1. Windows are advanced in increments of 10 seconds. The values of m considered are 1, 2, 4, 6, 12, 18, 24, and 30 (a time window of 10 seconds, 20 seconds, 40 seconds, and 1 to 5 minutes, respectively).

Figure 6.6 shows two hours of HR data in a stable patient, as well as the corresponding log-novelty time-series, up to a maximum window length of 5 minutes. In general, peaks in log-novelty are higher for smaller values of m. However, high log-novelty values are held longer for larger values of m. The latter occurs because a high HR observation is included in more windows if windows are longer (corresponding to increasing values of m). Larger values of m lead to less variability in log-novelty time-series, which is easier to interpret.

The value HR = 111 bpm observed at $t = 54$ minutes is likely to be artefactual, considering the previous and following measurements and the general downward trend around this time. It is noteworthy that the $\alpha = 0.05$ significance threshold (log-novelty $= 2.9957$) is exceeded only for $m = 1$, which is equivalent to the pdf-thresholding method.[5]

[4] Physiologically implausible values would be reached because we use a GMM as our generative model, whose support is the entire real line $\mathbb{R}$. This would not be the case if we used a model with bounded support.

[5] The EVDs used here are estimates of the exact EVDs, rather than estimates of asymptotic EVDs, meaning that for $m = 1$, $G_1^-(.;1) = G_1$. Setting a threshold in P_1 on $G_1^-(.;1)$ is therefore equivalent to setting a threshold on G_1, which is equivalent to setting a threshold on f_1.

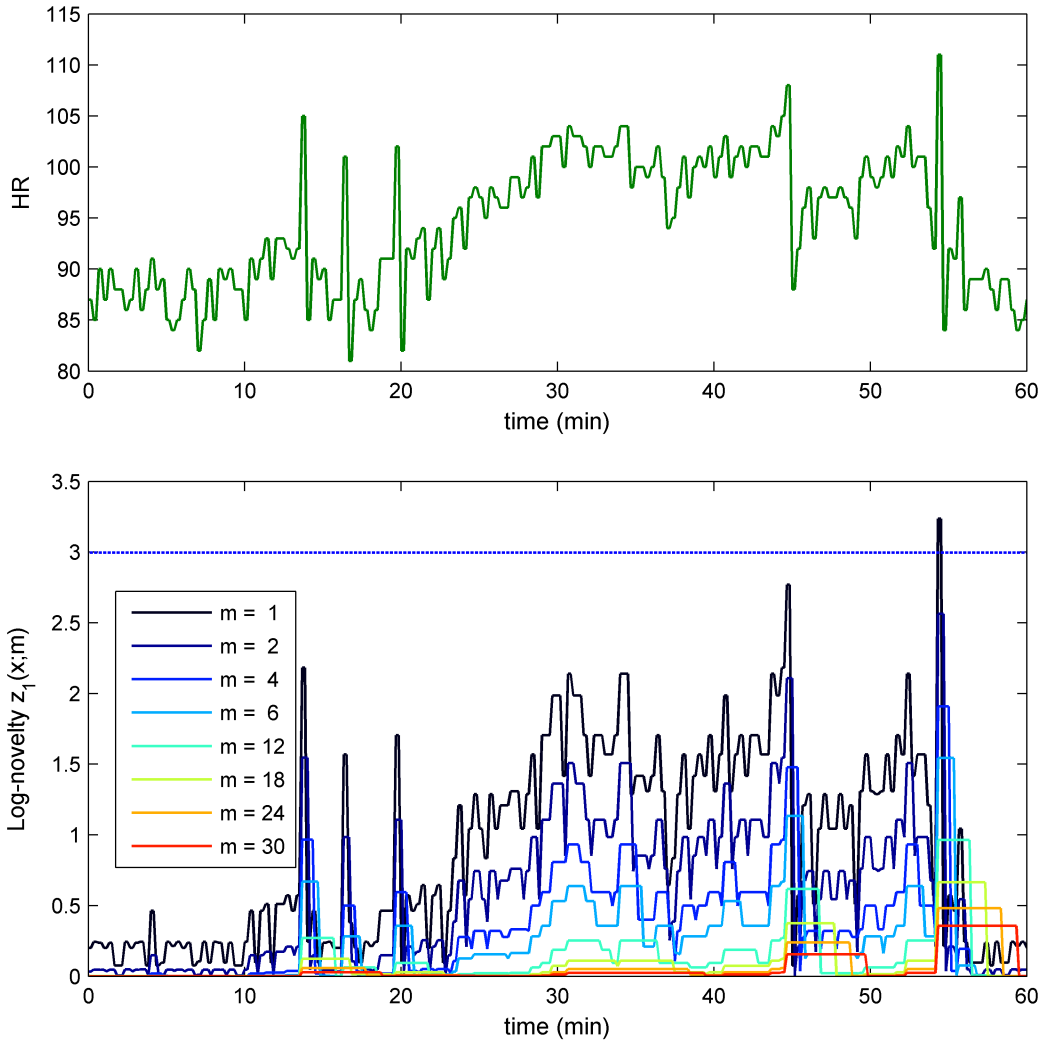


Figure 6.6: Example of univariate Novelty Detection. The upper graph shows the HR time series for two hours from UPMC Patient 1. This patient was stable throughout their stay on the ward. The lower graph shows the corresponding log-novelty time-series, with m ranging from 1 to 30. The blue dotted line represents the $\alpha = 0.05$ level of significance.

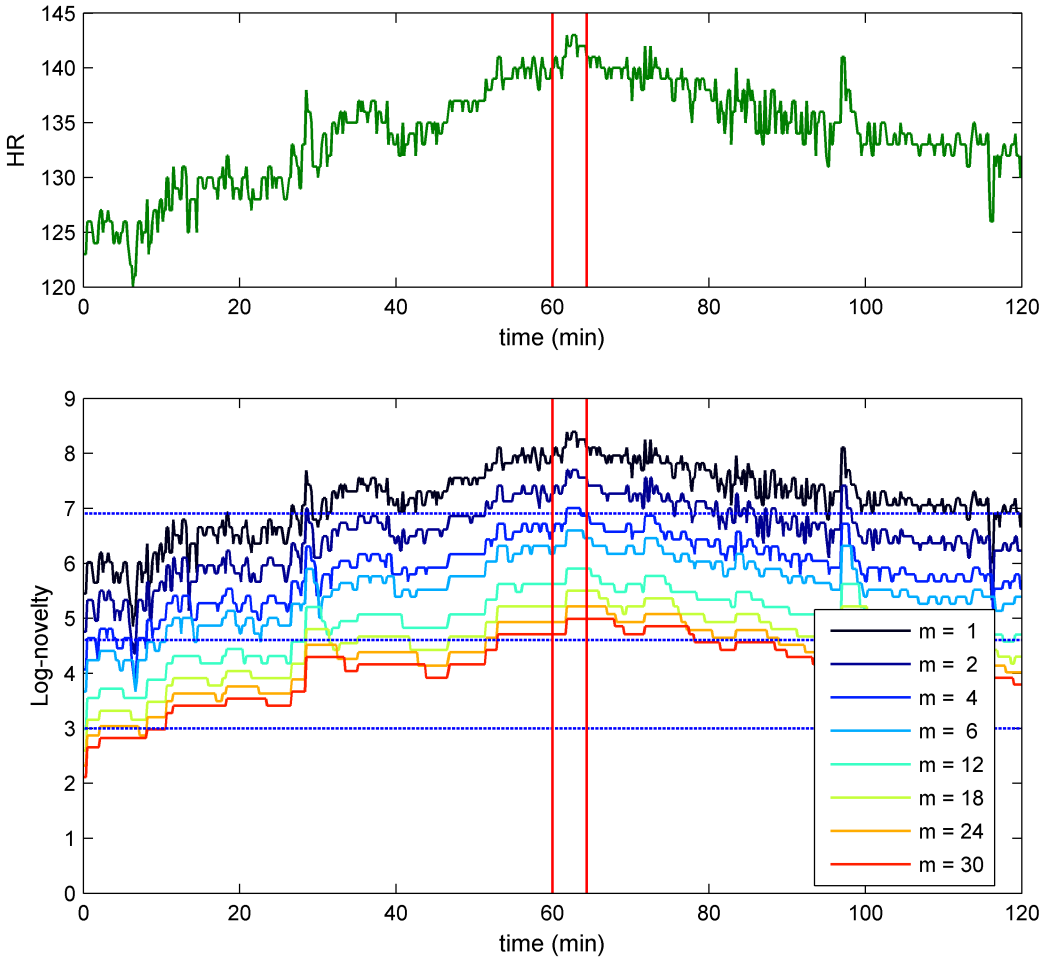


Figure 6.7: Same as Figure 6.6 for UPMC patient 32. The area between the vertical red lines indicate a tMET event. The horizontal blue lines correspond to the same significance levels and have the same meaning as those in Figure 6.4.

Figure 6.7 shows two hours of HR data around the time of a tMET event caused by an elevated heart rate for UPMC Patient 32. From the start of the tMET event HR $>$ 140 (the upper MET threshold for HR from Table 2.2) for at least 4 out of the first 5 minutes, and subsequently remains above 140 bpm for at least 2 out of 3 minutes until the end of the event. The HR gradually increases from 125 bpm one hour prior to the start of the event, reaches a maximum of HR = 143 bpm, then decreases to 130 bpm one hour after the end of the event. All the log-novelty time-series exhibit a similar trend. As in Figure 6.5, the higher the value of m, the lower the novelty scores, and the lower their variability.

6.3.4 Events and patients

The events used in this section are the tMET events caused by abnormal heart rates introduced in Chapter 3. The population of patients that we will consider for previously-unseen test data consists of the 85 patients in the "most normal" patient subset $\mathcal{P}_{\varnothing}$ whose data were not used to train the GMM f_1, and the 16 patients who had at least one tMET event caused by an abnormal heart rate (with 30 events in total). The HR time-series of these 101 patients are sampled every 10 seconds, to give 2.95×10^6 HR observations, over 8,193 hours of monitoring.

6.3.5 Results

Results of novelty detection are shown in Figure 6.8, and compared with results obtained using the algorithm described by Hann (2008). We show the following:

- False positive rate at the observation level (Figure 6.8.a): false positive rates are generally low (note the log scale used for the y-axis), due to the vast majority of the test data comprising normal observations. All methods appear to exhibit similar false positive rates for a novelty threshold less than 6. Above 6, the pdf-thresholding method appears to have lower false positive rates, indicating that the 4-out-of-5-minute hysteresis used in Hann (2008) does not improve observation-level false positive rates below high log-novelty thresholds (i.e., log-novelty less than 6). This is because an isolated artefactual

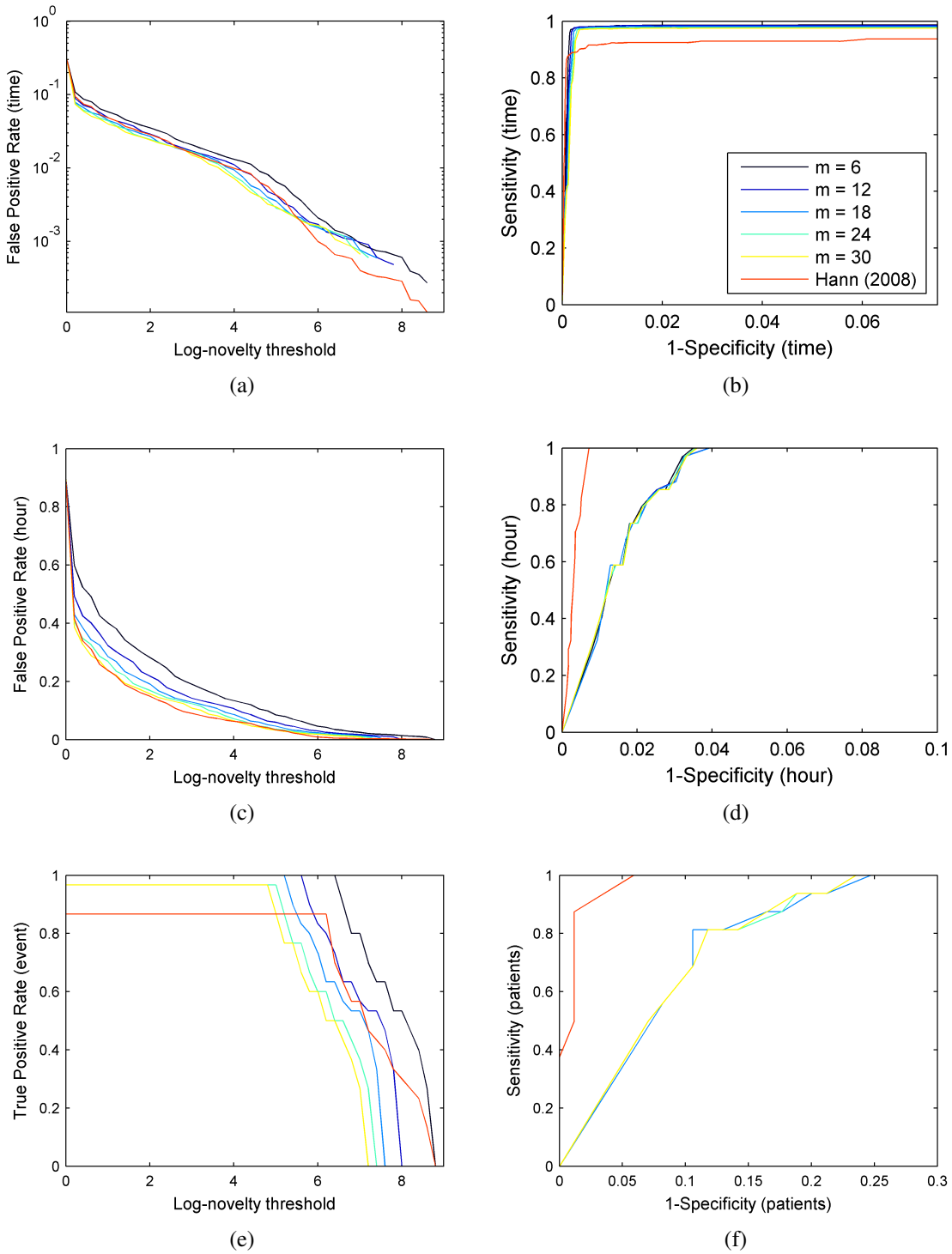


Figure 6.8: Results of the EVT-based novelty detection in HR time-series window lengths of 1 to 5 minutes ($m = 6$ to $m = 30$) compared to the results obtained using the algorithm of Hann (2008). Left, from top to bottom: false positive rates over individual observations, false positive rates over monitoring hours, and true positive rates over events, all plotted against log-novelty thresholds. Right: receiving operating characteristics at the observation (time), monitoring hour, and patient recording levels.

observation may cause multiple EVT windows of data to be misclassified (in general each observation is contained in $2m-1$ detection windows). This is not necessarily the case after post-processing in the pdf-thresholding method, where isolated artefactual observations are removed by the hysteresis.

- Observation-based ROC curve (Figure 6.8.b): all values of m yield comparable observation-based ROC curves, with AUCs greater than 0.97 (shown also in Table 6.2). The pdf-thresholding method, however, yields an AUC of 0.9361. Since it is not designed to perform classification immediately after periods of missing data (missing data do not contribute to the count of "abnormal" data, and therefore no alert can be triggered in the 4 minutes following a period of missing data), not every event observation is examined, and its sensitivity never reaches one.

  From Table 6.2, the observation level is the only detection level at which differences in AUCs are statistically significant.

- Hour-based false positive rate (Figure 6.8.c): we observe that the choice of m has an impact at the scale of the monitoring hour. The pdf-thresholding method consistently exhibits lower false positive rates than the EVT-based methods at this scale.

- Hour-based ROC curve (Figure 6.8.d): the differences in false positive rates due to the choice of m do not have a significant impact on the corresponding ROC curves, as sensitivities are all 1 for log-novelty thresholds up to 5. All AUCs are greater than 0.985 for these EVT-based classifiers, indicating that monitoring hours overlapping with HR tMET events, and monitoring hours which do not overlap with events, can be separated effectively. Choosing $m = 24$ and a log-novelty threshold of 5, for instance, allows us to classify correctly every monitoring hour which overlaps with tMET events, while maintaining a specificity of 0.965, equivalent to a false alarm rate of 3.5%. The pdf-thresholding method yields better results, with an AUC of 0.9969, due, again, to its ability to discard artefactual observations using hysteresis.

- Event-based true positive rates (Figure 6.8.e): for $m = 24$ and $m = 30$, there is one event which is never detected, as a result of the fact that the single tMET event for this patient

	AUC (time)	AUC (hour)	AUC (patient)
$m = 6$	0.986*	0.986	0.917
$m = 12$	0.984*	0.986	0.917
$m = 18$	0.981*	0.986	0.917
$m = 24$	0.977*	0.986	0.917
$m = 30$	0.974*	0.986	0.917
Hann (2008)	0.936	0.997	0.990

Table 6.2: AUCs of novelty detection in HR time-series. An asterisk next to an AUC denotes that the difference in AUC between this AUC and that obtained by Hann (2008)'s classifier at the same level of detection is statistically significant at the 0.05-level, using Hanley et al. (1983)'s method for comparing AUCs.

occurs over the first 3.5 minutes of the recording during which no classification is performed. The same event is missed by the pdf-thresholding method, along with 3 other events, which cannot be detected due to periods of missing data. On this graph, the effect of the filtering can be observed. While curves for different values of m cannot, by design, intersect that of the pdf-thresholding method despite being the lowest up to a log-novelty threshold of approximately 5, better performance than the EVT-based methods for higher values of the threshold, and the pdf-thresholding method eventually retains a higher true positive rate than all the EVT-based curves.

- Patient-based ROC curve (Figure 6.8.f): this graph shows that the pdf-thresholding method outperforms the EVT method in discriminating between patients who have events and patients who do not. This is consistent with our findings at the scale of monitoring hours, and indicate that the hours misclassified by the EVT method are spread across patients who do not have events. The observed differences in AUC are not statistically significant (Table 6.2) given the small number of patients included in this case study. They are large enough however, for us to suggest that the EVT-based classifiers would not outperform the pdf-thresholding one, even on a larger population, at this level of detection.

In summary, the pdf-thresholding is shown to outperform the EVT-based detectors at the monitoring hour and patient levels of detection on this univariate example. The next section presents a more detailed multivariate investigation involving the three continuous vital signs.

6.4 Trivariate HR/BR/SpO_2 model

6.4.1 Data

The triplets of heart rate, breathing rate, and oxygen saturation for all patients in subset $\mathcal{P}_\varnothing$ were obtained with a sampling period of 10 seconds. After exclusion of incomplete observations (at least one missing vital sign measurement), the number of training observations is 7.86×10^5. The two-dimensional marginal distributions for the three combinations of HR, BR, and SpO_2 are represented in Figure 6.9. The use of oxygen therapy results in increased frequencies of observations near and at the 100% saturation level. Ignoring periods of monitoring when patients receive oxygen therapy would result in smoother distributions, which would perhaps be easier to estimate. However, since oxygen therapy is used routinely in high-dependency units[6], it would be impractical to exclude these periods from this analysis.

The group of patients who did not have any event behaves differently from the group of patients who had at least one tMET event caused by an abnormal heart rate, breathing rate, or peripheral oxygen saturation level. In particular, the observed differences in the spreads of the tails may be detectable.

6.4.2 Model choice

The cross-validation procedure described in Section 6.3.1 is used to estimate the trivariate pdf of HR, BR, and SpO_2 for $\mathcal{P}_\varnothing$. The vector $\mathbf{x}$ is now a triplet of (HR, BR, SpO_2). The mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k$ are redefined accordingly.

Figure 6.10 shows the results of the 5-fold cross validation process in training and validation, for K ranging from 10 to 90. As was the case with the heart rate model f_1, the amount of overfitting as K is increased is limited due to the large number of training observations (each fold comprises 6.24×10^5 training observations and 1.56×10^5 validation observations). This makes the choice of the value K less clear than if the validation error were to reach a minimum.

[6] Anecdotal evidence suggests that 66% of UPMC Phase 1 patients were on oxygen therapy for more than one hour.

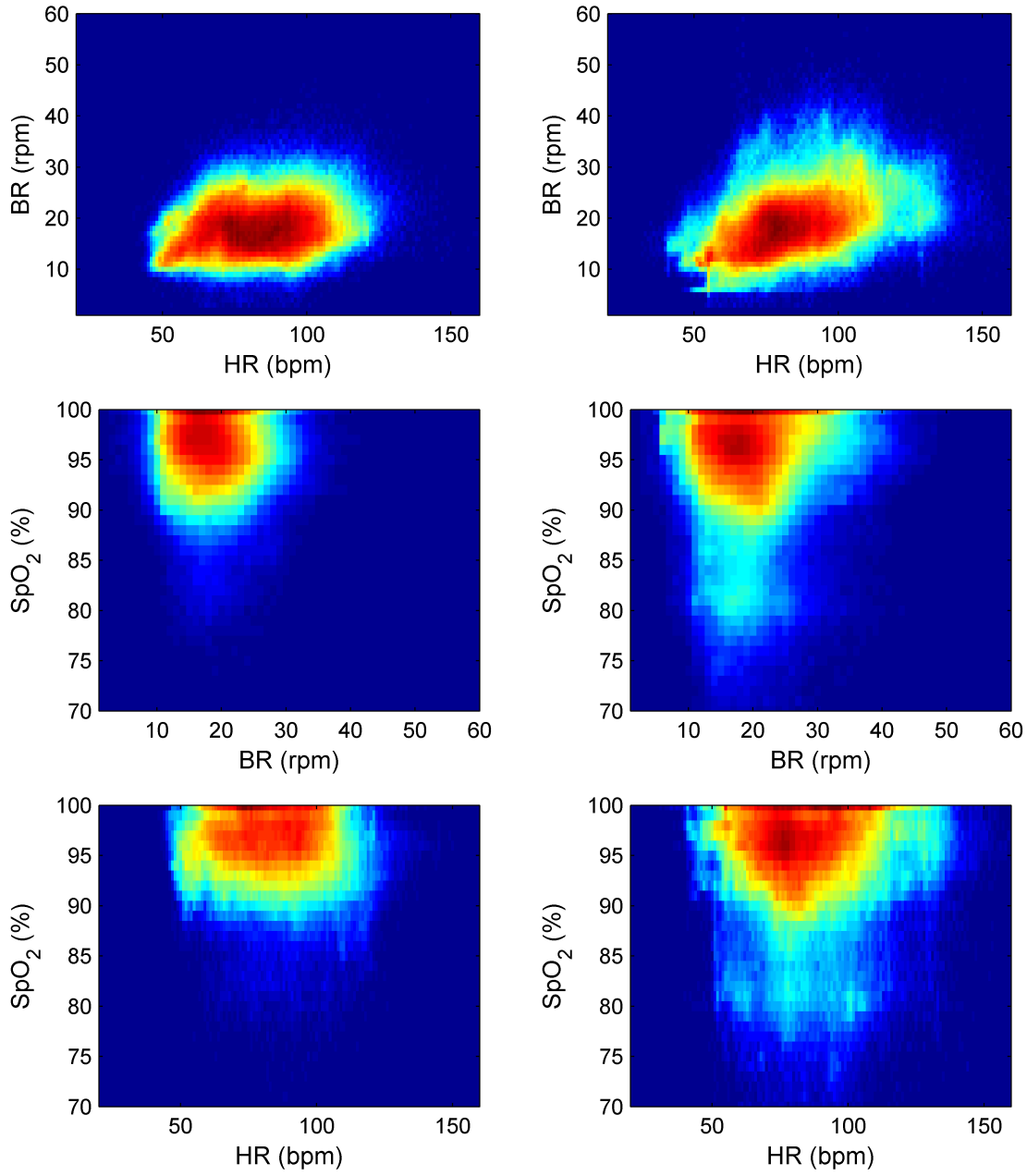


Figure 6.9: Empirical bivariate histograms of vital-sign observations for the group of patients who had no event (left column) and that of patients who had at least one tMET event related to HR, BR, or SpO_2. Values represented are logarithmic frequencies to enable an enhanced visualisation of the distribution tails. "Hotter" colours indicate higher frequencies.

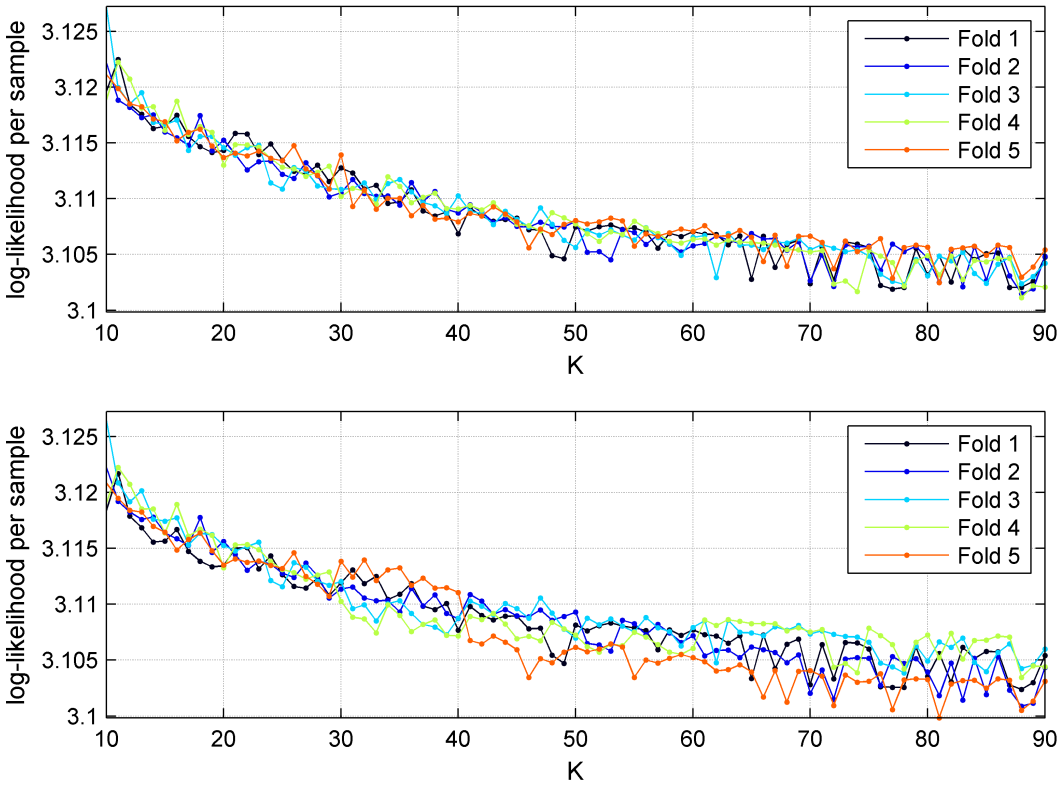


Figure 6.10: Average log-likelihood per observation for the training set (80% of data, upper graph) and validation set (20% of data) for each of the five folds in the cross-validation process. K is the number of kernels.

An indication of possible overfitting is the growing variance of the validation error across the five folds, observable for $K \geq 30$. It is noteworthy that the average data log-likelihood on the validation sets continually decreases as K increases.

The number of kernels for the trivariate model is chosen after a visual inspection of the tail estimates for the three univariate and three bivariate marginal distributions associated with each model, leading to the choice $K_2 = 60$. The bivariate marginal distributions for this value of K are shown in Figure 6.11. These can be seen to be suitable estimates of the empirical marginal distributions in Figure 6.9 (left column). From Figure 6.10, it seems that the overall averate likelihood per sample continues to decrease slightly up to 90 kernels. This improvement is marginal, and is likely to be caused by a better modelisation of the normal data. The modelisation of extreme data did not exhibit the same trend for $K > 60$. After selection of a value for K_2, a GMM is subsequently constructed using all the vital-sign training data, and is henceforth referred to as *the trivariate model*, or f_2.

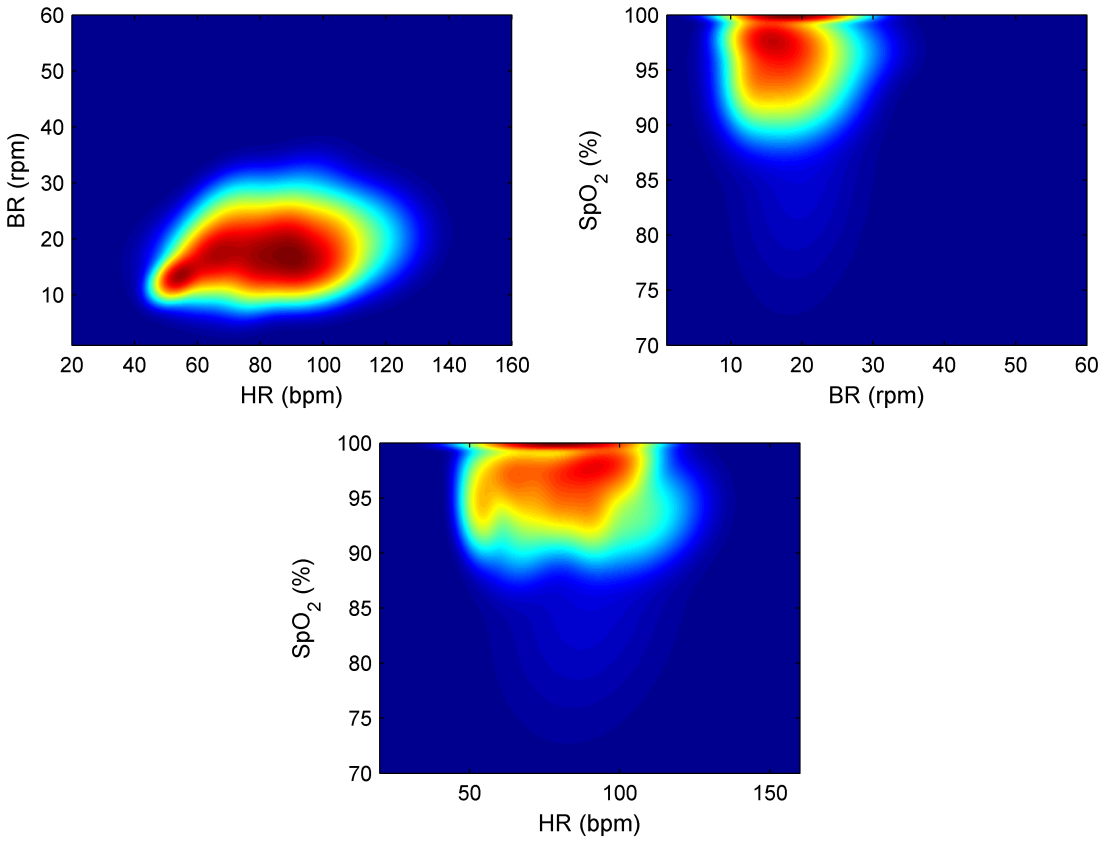


Figure 6.11: Bivariate marginal distributions of the HR/BR/SpO_2 model f_2.

It is notable that the SpO_2 variable has an upper limit (100%) near which lies a significant proportion of the probability mass. Modelling the trivariate distribution using an estimator whose support is infinite (as is the case for a GMM) means that a fraction of the model's probability mass lies outside the SpO_2 range, regardless on the number of kernels considered. This could in turn hinder our ability to derive meaningful estimator of extremal behaviour. In the present case, we note that SpO_2 near 100% is always considered normal, and that because of the presence of many patients under oxygen therapy, it would be impractical to remove those measurements from our data set. More accurate model distributions could be obtained:

- using truncated Gaussian distribution in the estimator;
- by applying transform to the SpO_2-axis (e.g., a logarithmic transform), such that the probability mass near the upper bound is spread over a larger range, allowing it, in principle to be more accurately estimated by a GMM.

The latter option was considered, and rejected for two reasons. First, there is an incentive to

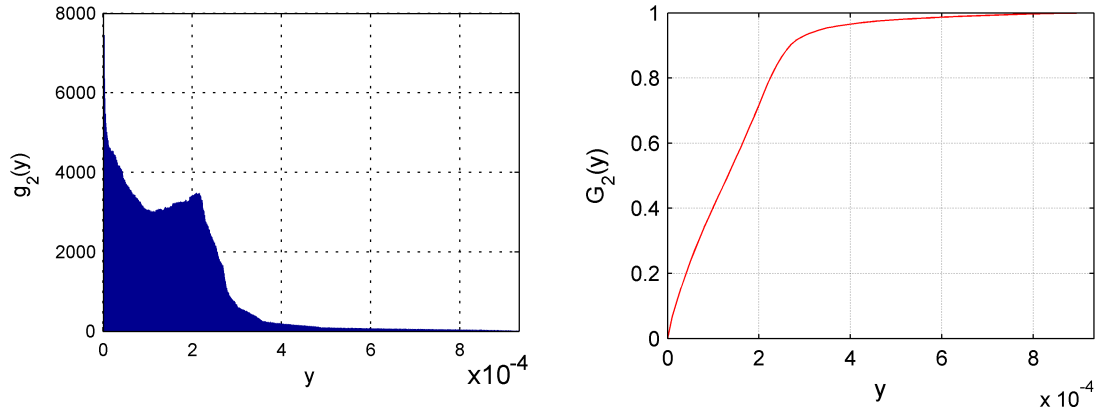


Figure 6.12: DPD of the trivariate HR/BR/SpO_2 model f_2: pdf, g_2 (left), and cdf, G_2 (right) over the probability space P_2. f_2 is the 50-kernel GMM whose marginal distributions are represented in Figure 6.11. Both graphs are obtained using 10^7 randomly-drawn observations.

model the data in the data space they are measured in, so that dimensions and units remain those that clinical practioners are familiar with.[7] Second, due to the quantization of the data (SpO_2 measurements are only available as integer values in the range 0-100%), no transform would be able to spread the high-SpO_2 data along the SpO_2-axis. As a consequence, a mixture of Gaussian distributions modelling the SpO_2 data would inevitably have probability mass which lies above the 100% threshold. For the purpose of demonstration, it was also estimated that classical, non-truncated Gaussian distributions are sufficient to model the trivariate distribution at this stage.

6.4.3 Novelty detection

Notations

The data space for this investigation is denoted by D_2, and corresponds to $\mathbb{R}^3$. We re-use the notations from Table 6.1, with the subscript 2 replacing the subscript 1, denoting this is the second case-study.

Distribution of Probability Density for f_2.

An estimate of the DPD for f_2 is generated using 10^7 randomly-generated observations from f_2, and 10^3 evenly distributed y values. The resulting pdf and cdf are shown in Figure 6.12. Unlike in Figure 6.3, the pdf g_2 exhibits few modes, indicating that f_2 is smoother than f_1. It appears that the higher-dimensional GMM f_2 is more skewed towards 0 than the one-dimensional f_1, and therefore has a heavier tail. This is anticipated by Equations 5.32 and 5.33, as well as Figure 5.6.

Events and patients

The events used in this section are the same as those in Section 6.3. The population of patients considered here consists of the 85 patients in $\mathcal{P}_\varnothing$ whose data were not used to train the GMM f_2, and the 57 patients who had at least one tMET event caused by an abnormal heart rate (30 events), respiratory rate (192 events), or oxygen saturation level (68 events). The total number of patients considered in this study is 142, while data from 85 other patients were used to construct the model f_2.

Extreme Value Distributions of f_2

The distributions of minima for G_2, termed $G_2^-(.;m)$ are obtained using Equation 5.6. Figure 6.13 shows the EVDs $G_2^-(.;m)$ for a range of sample sizes m and their associated log-novelty curves. As expected, increasing m causes the probability mass of the EVD to shift towards the lower part of the probability space. At a given density level y, novelty scores tend to decrease as m increases.

[7] It could be argued that a reverse transform may be applied to present results in the correct dimensions and units. Not transforming the data was however a deliberate choice.

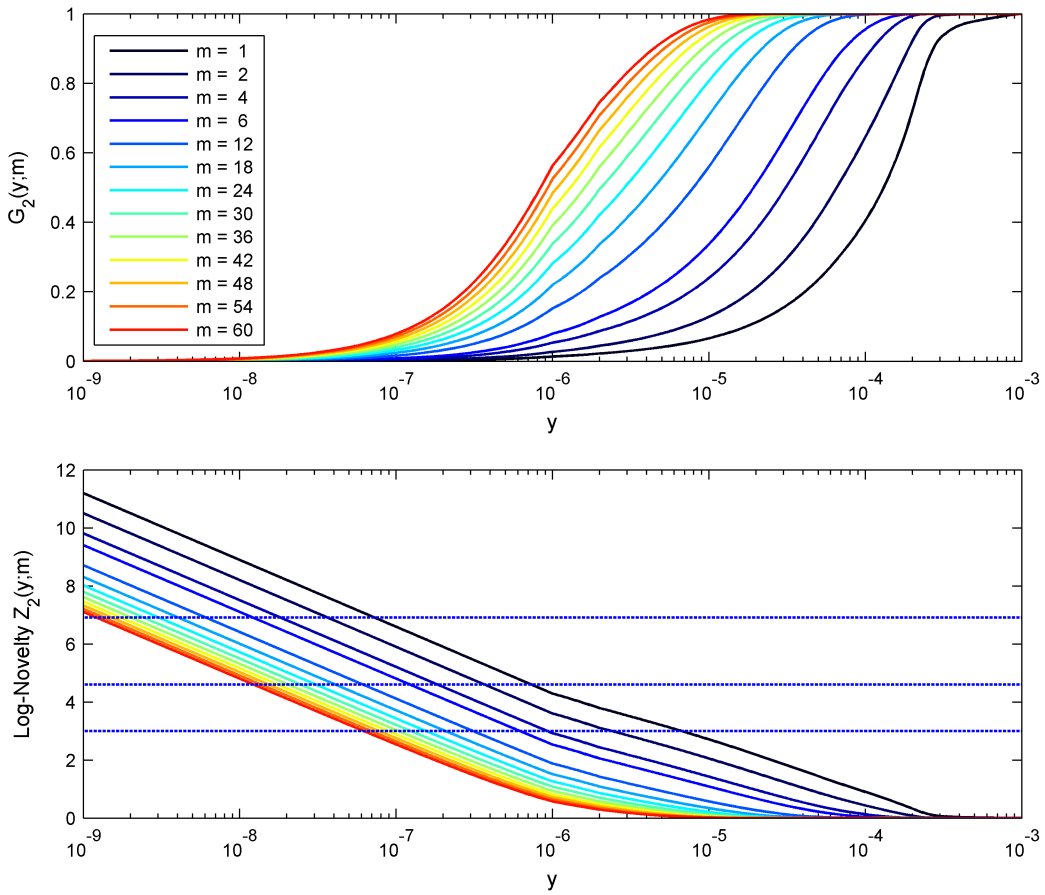


Figure 6.13: EVDs of G_2 (upper graph) and log-novelty in probability space, both represented on a semi-logarithmic scale. Values of the sample size parameter m range from 1 (black) to 60 (red). In the lower plot, the horizontal dotted lines correspond to natural novelty scores (or quantiles of the EVDs) of 0.95 (lowest line), 0.99, and 0.999. Conversely, they correspond to significance levels $\alpha = 0.05$, 10^{-2}, and 10^{-3}. This figure is the equivalent of Figure 6.4, adapted for G_2.

Range of m

Following the discussion in Section 6.3.3, we use the same values of the sample size m; i.e. $m \in \{6, 12, 18, 24, 30\}$, corresponding to window sizes of 1 to 5 minutes.

Novelty detection procedure in HR/BR/SpO_2 time-series

We apply the novelty detection procedure from Algorithm 6.1, using f_2 instead of f_1, and HR/BR/SpO_2 triplets of observations. Detection windows are again advanced in increments of 10 seconds.

The example in Figure 6.7 is considered again, this time using all three continuous vital signs in the two-hour period surrounding the tMET event caused by elevated heart rate. Figure 6.14 shows a plot of the vital-sign data and the corresponding log-novelty time-series obtained using the trivariate model f_2, rather than f_1. These log-novelty time-series differ significantly from those in Figure 6.7:

- Due to the presence of an elevated breathing rate throughout the window, novelty scores are, on average, higher in the trivariate case. For $m = 30$, the log-novelty exceeds the $\alpha = 0.01$ level of significance over 75% of the time (where it was less than 20% using f_1).
- In the univariate case, the maximum novelty score over the period is reached at the time of the labelled event, for all values of m. This is not the case in the multivariate case: assuming they correspond to accurate measurements, the simultaneous peaks of HR and BR (and possible drops in oxygen saturation), occur at $t = 30$ and 95 minutes. These increases are assigned higher novelty scores than the maximum in heart rate reached during the tMET event (at) $t = 62$ min.
- The relative decrease in novelty as m is increased appears to be less marked in the trivariate case, indicating that the length of the detection window is less critical than it is in the univariate case.

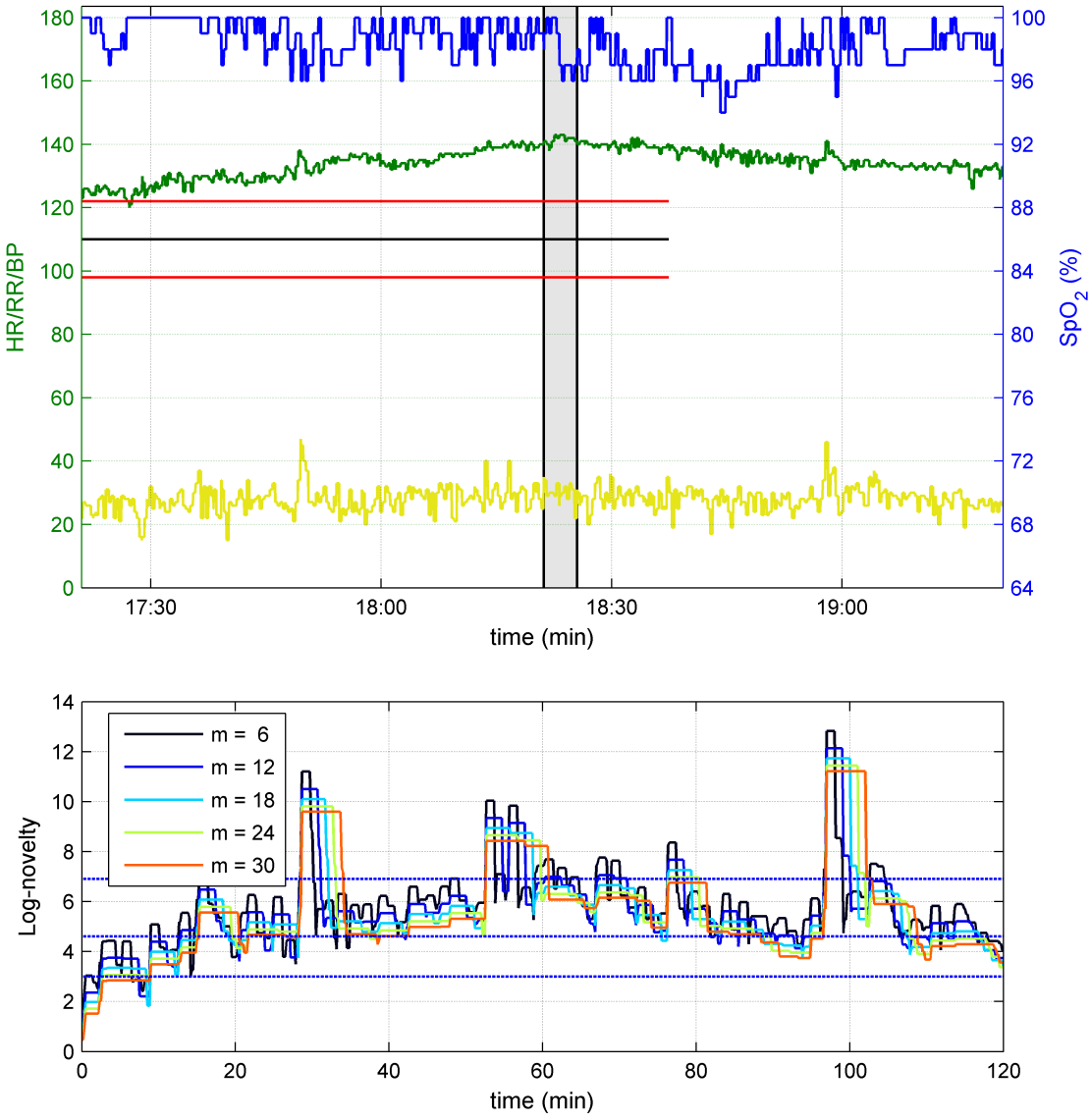


Figure 6.14: Example of multivariate novelty detection in UPMC Phase 1 Patient 32. The upper graph shows the vital-sign time-series. Blood pressure parameters (in red for systolic and diastolic blood pressure, in black for mean arterial pressure) are represented although they are not used in the novelty detection procedure. The HR data represented (in green) are the same as in the upper graph of Figure 6.7. The RR data are shown in yellow. The lower graph shows the corresponding log-novelty time-series for m ranging from 6 to 30. The horizontal dotted lines correspond to natural novelty scores of 0.95, 0.99, and 0.999.

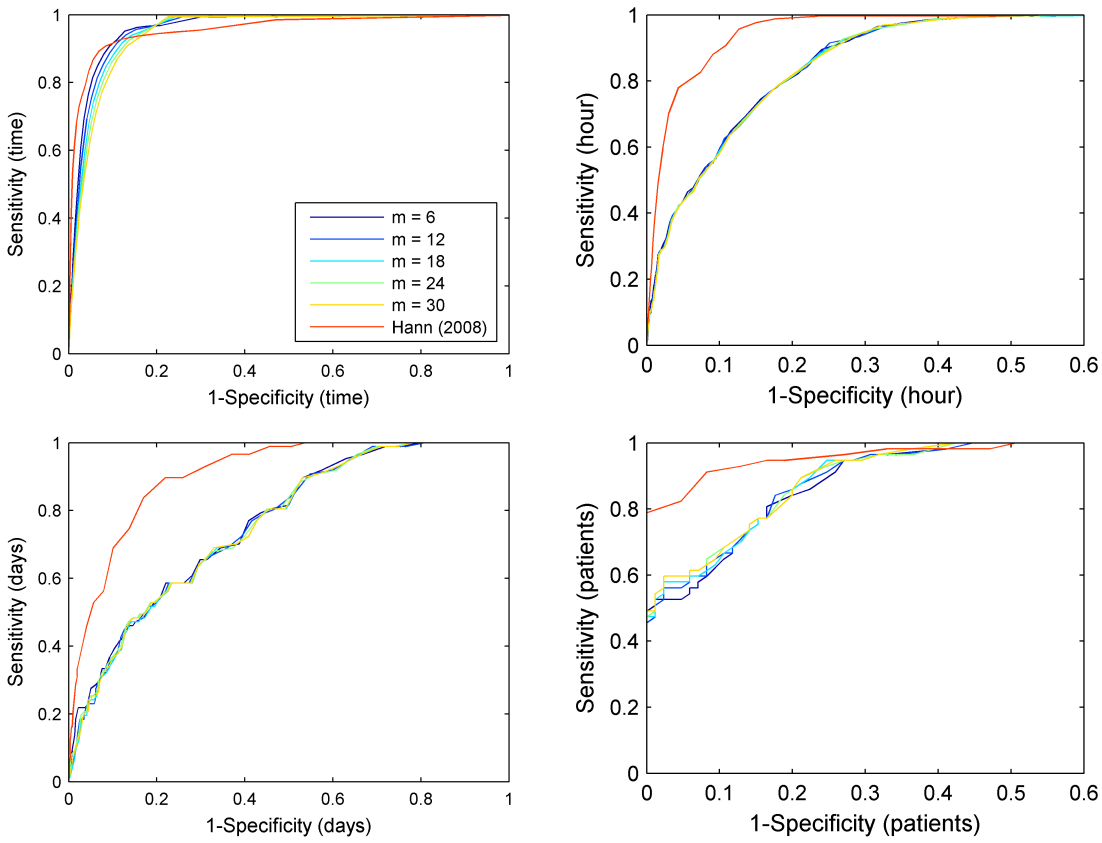


Figure 6.15: ROC curves for procedure using multivariate EVT. Levels of detection are, clockwise: an observation (time), a monitoring hour, a monitoring day, and a patient's stay in the ward.

6.4.4 Results

Algorithm 6.1 is used to perform novelty detection using the UPMC Phase 1 data set. Receiving operating characteristics at four levels of detection are shown in Figure 6.14. Results obtained using the algorithm described by Hann (2008) are also represented (red lines). The corresponding areas under the curves are reported in Table 6.3.

Performance by level of detection:

The highest AUCs are obtained while performing detection on an observation-by-observation basis (maximum AUC of 96.06% for $m = 6$). Considering the corresponding ROC curves, one could select a novelty threshold value such that both the specificity and the sensitivity of the novelty detector are high (both greater than 90%, for instance), or favour a higher sensitivity, at the expense of a moderate drop in specificity. As an example, the maximal achievable specificity

at the 97% sensitivity lies between 79% and 80% for all the values of m.

As m increases, the AUC decreases slightly, due mainly to a relative decrease in sensitivity. Increasing m means decreasing the density at which a novelty threshold is reached, and therefore classifying an increasing number of observations as being "normal". The presence of normal (or only marginally "abnormal") data in the tMET event windows means that the overall sensitivity of the detector may decrease. Additionally, abnormal observations, whether artefactual or physiological, correspond to increases in novelty scores; i.e, they are often orders of magnitude less likely than "normal" observations. Their classifications are therefore relatively unchanged as m changes, meaning that below a certain density level, specificity is not improved by small increments in the value of m.

While performance at the observation level is encouraging, it can be misleading considered on its own. High sensitivity and specificity on an observation basis allow us to know the fraction of the time the detector correctly classifies observations, but not how often it transitions from a correct classification to an incorrect classification. Results at the level of the monitoring hour allow us to quantify whether false alerts occur in the same hours as events, and how often they spread to negative (event-free) monitoring hour. Hour-based AUCs are all around 90%, a five-point drop from the observation-level ones, due to a relative increase in false positives affecting the specificity. False positives generally do not occur in the same hours as tMET events, and occur in monitoring hours which contain no event. Typically, ensuring that at least one alert occurs in each of the positive monitoring hours (i.e., those overlapping with at least one event) can only be achieved at the expense of allowing at least one false alert to occur in approximately 2 out of 5 negative monitoring hours. The same remarks are applicable to the more demanding level of the monitoring day, at which the AUCs are around 76%, a further 14-percent decrease. Guaranteeing detection of all positive monitoring days can only be achieved at the expense of at least one false alert being triggered in 4 out of 5 negative monitoring days.

This decrease in performance does not extend to the patient level, despite the average length of stay being more than one day (3.46 days, see Section 2.3.2). False positive hours and days appear to occur more frequently in patients who had tMET events; i.e, more abnormal obser-

	AUC (time)	AUC (hours)	AUC (days)	AUC (patients)
$m = 6$	0.961*	0.8986*	0.764*	0.917
$m = 12$	0.958*	0.8977*	0.760*	0.921
$m = 18$	0.954*	0.8970*	0.757*	0.922
$m = 24$	0.952*	0.8972*	0.761*	0.925
$m = 30$	0.948*	0.8968*	0.759*	0.925
Hann (2008)	0.959	0.965	0.907	0.971

Table 6.3: AUCs of novelty detection in HR/BR/SpO_2 time-series. An asterisk next to an AUC denotes that the difference in AUC between this AUC and that obtained by Hann (2008)'s classifier at the same level of detection is statistically significant at the 0.05-level, using Hanley et al. (1983)'s method.

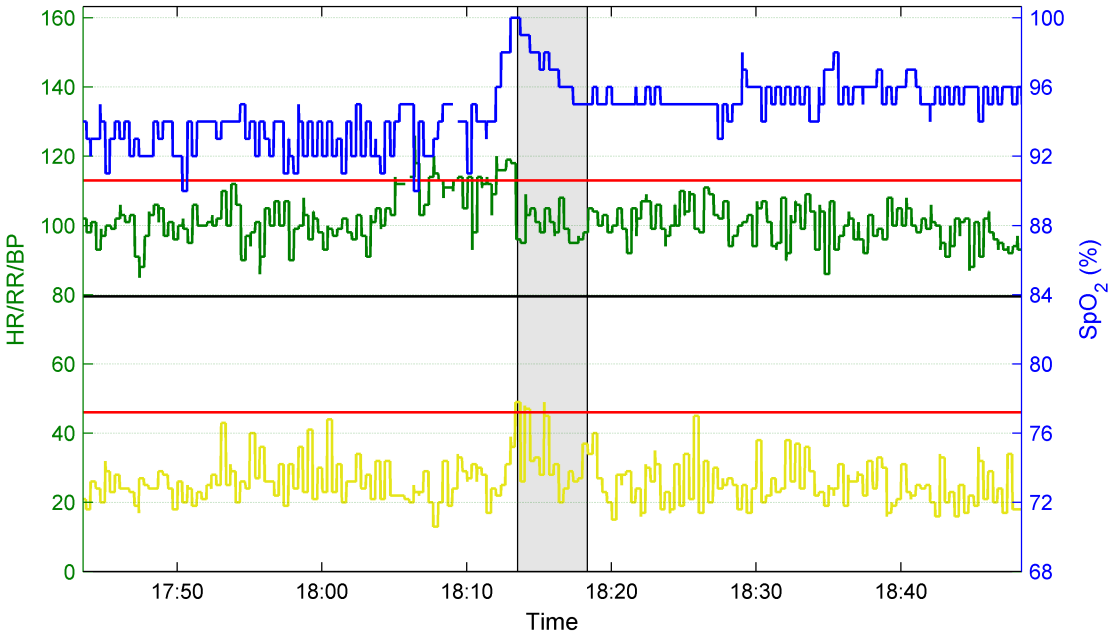
vations are identified by the EVT-based method in patients who had tMET events. While these abnormal EVT windows do not meet the criteria of being tMET events, they suggest one of two conclusions: patients who had events are more likely than others to show artefactual data, or these detections are actually physiological in nature and could correspond to unlabelled periods of abnormality.

Influence of the sample size m:

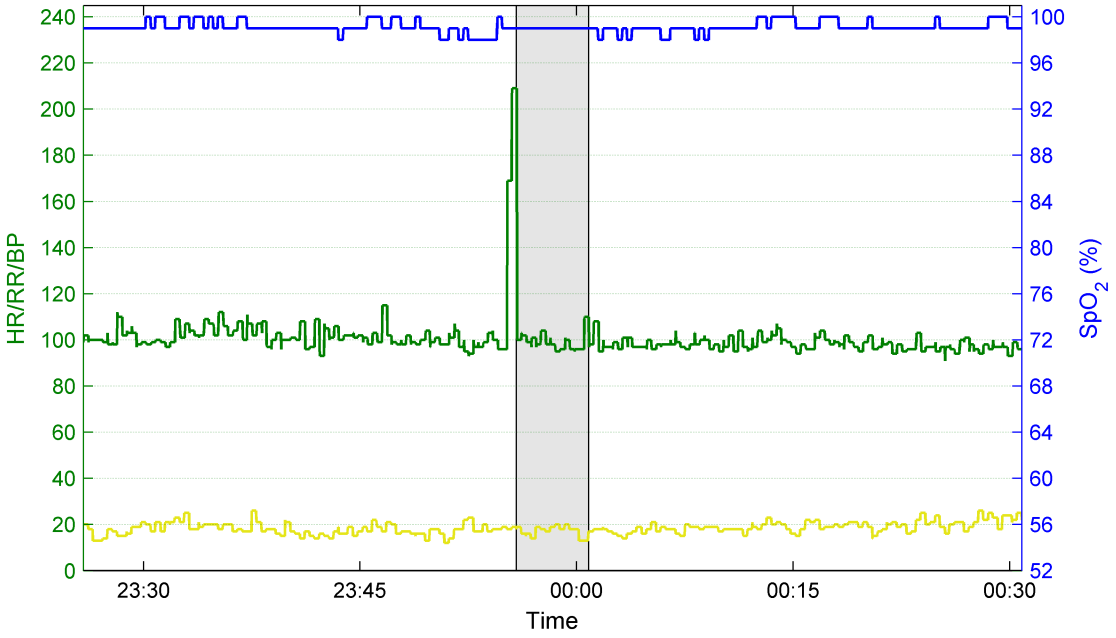
Varying the sample size m appears to have a negligible effect on the ROC curves and AUCs, perhaps contrary to theoretical predictions. This relates to comments in the previous section: abnormal observations, whether artefactual or physiological have novelty scores that are much higher than normal data with respect to the model f_2. Abnormal periods of data greatly differ from both the model f_2 and the estimate of its tails: physiological abnormality appears to occur far from the predicted edge of physiological normality, resulting in extremely high novelty scores with respect to the model. Log-novelty scores in excess of 10 are commonly observed for all values of m, correspond to breaking the $\alpha = 4.5 \times 10^{-5}$ level of significance.

Comparison with Hann's algorithm

Figure 6.14 and Table 6.3 show that Hann's algorithm significantly outperforms our EVT-based method at every detection level except one. At the observation level, our method slightly outperform Hann's, and the differences are statistically significant at the 0.05-level for every value



(a) Patient 288, day 3.



(b) Patient 271, day 2.

Figure 6.16: Example of false alerts in UPMC Phase 1 Patients 271 & 288. Both patients have no event. The area in grey correspond to the period of false detection for $m = 30$ (duration = 5 minutes).

of the sample size m. This is a positive result, recalling that the EVT-based method returns a classification at the time of each observation.

Our methods are however significantly less consistent at higher levels of detection. The explanation lies in the post-processing stage of Hann's algorithm. The 4-out-of-5 hysteresis ensures that 80% of a 5-minute window's observations are above a certain threshold. Transient peaks of less than 4 minutes are therefore discarded. On the other hand, the EVT-based methods, do not modify the ranking of observations. Peaks removed by the 4-out-of-5 hysteresis are consequently peaks in novelty. Since we do not use any post-processing step, it is therefore normal that EVT novelty detectors yield more detections in general, and more false detections in particular that Hann's method.

In Figure 6.16, we show two examples of false detections for patients who had no event. In the case of Patient 271, the two HR observations of 163 and 209 are obviously artefactual in an otherwise very stable period. In the case of patient 288, a breathing rate measured above 40 around 18:13 appears to be responsible for the false alert. There is no evidence that the false alert period is the result of artefactual data. Since it lasts under 4 minutes, it is not identified as a MET event, nor is this period detected by Hann's detector. While it is possible that it is physiologically relevant, it is an isolated occurrence in an otherwise stable patient, which, like the artefact in Patient 271, would ideally be ignored.

These false detections illustrate the lack of tolerance to noise of the EVT methods. Since they are only concerned with the most extreme observation, they cannot be robust to noise affecting this extreme. They can be made more robust by adding a filtering step in which, for instance, a median filter is applied to the probability density time-series before extracting the extremes. In Table 6.4, we show that this pre-processing stage improves the AUCs at every detection level, such that the AUCs are comparable to those obtained using the algorithm by Hann (2008). Improvements in AUCs at the observation level remain statistically significant, while nearly all classifiers result in AUCs which are no longer significantly worse than the AUCs obtained using the pdf-thresholding method. However, the choices of filter and filter parameters are heuristic, and the probabilistic interpretation of novelty scores is lost in the process.

Median filter (min)	m	AUC (time)	AUC (hours)	AUC (days)	AUC (patients)
	6	0.975*	0.944	0.838*	0.934
	12	0.974*	0.944	0.838*	0.936
2	18	0.975*	0.944	0.837*	0.937
	24	0.971*	0.943	0.838*	0.943
	30	0.970*	0.942	0.838*	0.942
	6	0.978*	0.951	0.849	0.945
	12	0.977*	0.951	0.852	0.949
3	18	0.975*	0.950	0.850	0.949
	24	0.974*	0.949	0.850	0.950
	30	0.971*	0.948	0.850	0.953
	6	0.980*	0.955	0.855	0.941
	12	0.978*	0.954	0.853	0.943
4	18	0.977*	0.953	0.855	0.944
	24	0.975*	0.951	0.855	0.944
	32	0.973*	0.951	0.855	0.945
	6	0.981*	0.958	0.866	0.9450
	12	0.980*	0.958	0.864c	0.9450
5	18	0.978*	0.956	0.865	0.954
	24	0.977*	0.954	0.863	0.952
	30	0.974*	0.954	0.863	0.952
	6	0.982*	0.960	0.875	0.953
	12	0.981*	0.960	0.870	0.951
6	18	0.979*	0.959	0.872	0.955
	24	0.978*	0.956	0.871	0.955
	30	0.976*	0.957	0.872	0.955
	6	0.983*	0.965	0.891	0.958
	12	0.982*	0.965	0.889	0.959
10	18	0.981*	0.964	0.891	0.962
	24	0.979*	0.956	0.890	0.961
	30	0.976*	0.956	0.887	0.956
	6	0.981*	0.965	0.890	0.962
	12	0.979*	0.964	0.890	0.965
20	18	0.977*	0.960	0.890	0.965
	24	0.974*	0.953	0.888	0.959
	30	0.975*	0.966	0.900	0.947
Hann (2008)		0.959	0.965	0.907	0.971

Table 6.4: AUCs of novelty detection in HR/BR/SpO_2 time-series, after median filtering of the density time-series. An asterisk next to an AUC denotes that the difference in AUC between this AUC and that obtained by Hann (2008)'s classifier at the same level of detection is statistically significant at the 0.05-level, using Hanley et al. (1983)'s method for comparing AUCs.

In order to address this robustness issue in a more principled manner, we turn our attention in the next chapter to designing a novelty detection algorithm based on multivariate order statistics.

Chapter 7

Novelty Detection with Order Statistics

7.1 Motivation

EVT for novelty detection in time-series allows us to identify extreme observations occurring in a time window as abnormal with respect to a model distribution. However, as demonstrated in the previous chapter, focusing on a window's most extreme observation does not prevent false alerts in the event of artefactual observations, or in the event of short transient departures from normality. We demonstrated that isolated outliers influence novelty scores for the length of the detection window. In the case of noisy time-series, this can result in an amplification of the effect of any noise present, because the most extreme observation in a window is always that used to determine the value of the novelty score for that window.

Aberrant observations taking values outside a physiologically-plausible range could be filtered out online by adding a pre-processing step to our novelty detection algorithm. Identifying transient excursions outside the boundaries of normality to prevent false alerting does not belong to the scope of EVT-based methods, unless a heuristic additional post-processing step, such as the 4-out-of-5-minute filtering described previously, is added. To address this issue, we turn our attention to order statistics.

Definition 7. *(Order Statistic)*

Let $(X_1, X_2, \ldots, X_m)$ *be an i.i.d. sample of size m drawn from a univariate distribution function*

Algorithm 7.1 Two-sided univariate novelty detection with order statistics, at the α level of significance.

Let $m \in \mathbb{N}^*$ be the sample size, $k \leq m$ a strictly positive integer, $(X_1, X_2, \ldots, X_m)$ m i.i.d random variables , and F their distribution function.

Initialisation:

1. Compute (or estimate) $F_{k:m}$, the distribution over $X_{k:m}$
2. Select a value of the novelty threshold $\alpha \in [0,1]$.

Unseen samples: Given an unseen sample, $(x_1, \ldots, x_m)$

1. Compute $x_{k:m}$, the k^{th} order statistic of the sample;
2. The novelty score of $x_{k:m}$ is given by $\max F_{k:m}(x_{k:m}), 1 - F_{k:m}(x_{k:m})$;
3. Label the sample novel if $F_{k:m}(x_{k:m}) \notin \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right]$.

F, with pdf f. Arrange $(X_1, X_2, \ldots, X_m)$ in 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 *of the sample.*

Definition 7 defines the k^{th} order statistic of a statistical sample as being its k^{th} smallest value. The first and last order statistics, $X_{1:m}$ and $X_{m:m}$ are the minimum and maximum of the sample, respectively, from which it follows that classical extreme value theory is simply a branch of order statistics. Given values for m and k, the distribution $F_{k:m}$ over $X_{k:m}$, assuming it is known, can be used to determine the novelty score of an observation $x_{k:m}$ in the same manner that the extreme value distribution is used in Chapter 4. Novelty detection is again considered an instance of hypothesis testing at some level of significance, as described in Algorithm 7.1; there, a sample is classified as novel if its k^{th} order statistic lies outside expected bounds of normality, defined as percentiles of the predicted distribution over the statistic.

In the case of novelty detection in i.i.d. time-series, the content of a sliding window of data is considered to be a sample, and a novelty detection algorithm similar to Algorithm 4.2 is derived in Section 7.4.

Considering the k^{th} order statistic ($1 < k < m$) rather than an extremum allows us to adjust the robustness of the novelty detection systems to outliers. The value of $x_{k:m}$ remains unchanged if every other observation x_i in the sample is translated by an arbitrary quantity, as long as $x_{k:m}$ remains the k^{th} smallest value in the sample. In particular, shifting observations greater than $x_{k:m}$ by an arbitrary positive quantity, and observations smaller than $x_{k:m}$ by an arbitrary

negative quantity has no effect on the value of the k^{th} order statistic.[1] In the context of novelty detection, short transient periods of artefactual measurements can be made to have no effect on the novelty score of a window. Order statistics are therefore potentially useful for the robust detection of persistent vital-sign abnormalities.

One important drawback of the use of order statistics is a potential deterioration of the timeliness of detection. In EVT, an abnormal observation results in a classification of abnormality in the first sliding window that contains the abnormal point, leading to an immediate alert. It is not, however, guaranteed to change the order statistics of a window (e.g., if it is greater than the previous value of the order statistic, and replaces a value that is also greater), and it might take up to $m-k$ (or $k-1$, depending on whether the abnormal observation value is greater or smaller than the order statistic) windows for an alert to be triggered. This natural trade-off between robustness and timeliness of detection drives the selection of the value of k.

7.2 Distributions of order statistics

The derivation of the exact distribution of the k^{th} order statistic is similar to that of the distribution of extrema in Section 4.2.1, with an additional combinatorial element. Let $(X_1,\ldots,X_m)$ be a series of univariate i.i.d random variables with pdf f, cdf F, and support $D_f=[a,b]$ and let $\mathrm{d}x$ be an infinitesimal element. We have:

$$\begin{aligned} P(X_i \in]x,x+\mathrm{d}x]) &= F(x+\mathrm{d}x)-F(x) \\ &\simeq f(x)\mathrm{d}x\,; \\ P(X_i \leq x) &= F(x)\,; \\ P(X_i > x+\mathrm{d}x) &= 1-F(x+\mathrm{d}x)\,. \end{aligned} \tag{7.1}$$

[1] Filtering techniques involving medians take advantage of this robustness to outliers, the median being the $\lfloor \frac{m}{2} \rfloor^{\text{th}}$ order statistic.

$X_{k:m}$ is in $]x, x+h[$ if exactly one out of the m values of X_i is in $]x, x+h[$, and exactly $k-1$ out of the $m-1$ others are less than x, and the remaining $m-k$ are greater than $x+h$. This gives:

$$\begin{aligned} P(X_{k:m} \in]x, x+\mathrm{d}x]) &= f_{k:m}(x)\mathrm{d}x \\ &= m\,(F(x) - F(x+\mathrm{d}x))\binom{m-1}{k-1}F(x)^{k-1}(1-F(x+dx))^{m-k} \\ &= m\,f(x)\mathrm{d}x\binom{m-1}{k-1}F(x)^{k-1}(1-F(x+dx))^{m-k}, \end{aligned}$$

where $\binom{m-1}{k-1}$ denotes the binomial coefficient: $\frac{(m-1)!}{(k-1)!(m-k)!}$. In the limit $\mathrm{d}x \to 0$, we therefore have the following pdf:

$$f_{k:m}(x) = m\,f(x)\binom{m-1}{k-1}F(x)^{k-1}(1-F(x))^{m-k}. \tag{7.2}$$

The expression of the cdf can be obtained by integrating Equation 7.2, or by noting that (Arnold et al. 2008):

$$\begin{aligned} F_{k:m}(x) &= \mathrm{P}(X_{k:m} \leq x) \\ &= \mathrm{P}(\text{at least } k \text{ of } X_1, X_2, \ldots, X_m \text{ are at most } x) \\ &= \sum_{j=k}^{m} \mathrm{P}(\text{exactly } j \text{ of } X_1, X_2, \ldots, X_m \text{ are at most } x) \\ &= \sum_{j=k}^{m}\binom{m}{j}F(x)^j\,[1-F(x)]^{m-j}. \end{aligned} \tag{7.3}$$

Replacing k by m in Equations 7.2 and 7.3 yields Equations 4.5 and 4.6. Example distributions are shown in Figure 7.1. The extremal distributions $f_{1:10}$ and $f_{10:10}$ are the distributions of the minimum and the maximum of 10 random observations.

Definition 7 defines a univariate order statistic based on the natural order in the univariate data space, typically a subset of $\mathbb{R}$. This definition is not adapted to novelty detection in cases where the data-generating model is multimodal and multivariate: first because novel data, including any order statistic of novel window, may lie in areas of low probability between modes, and second because there is no natural choice of ranking function in a multivariate data space.

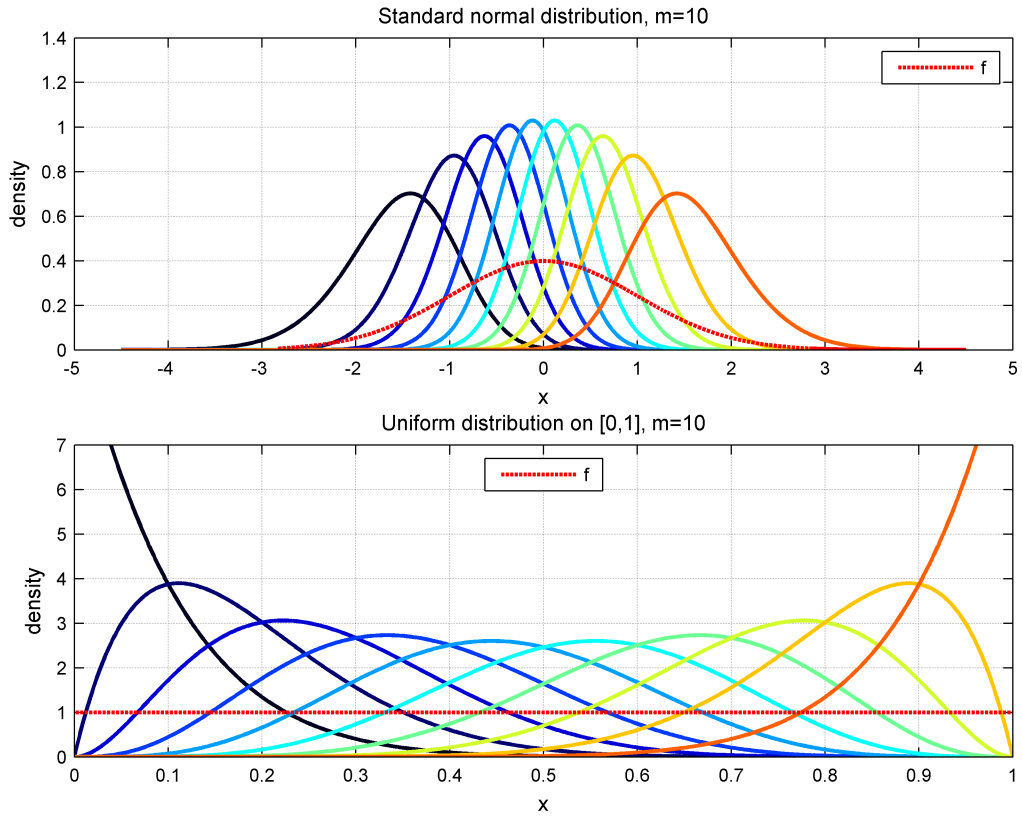


Figure 7.1: Distributions of order statistics for the standard normal and uniform distributions. In each graph, the red dotted line is the pdf of the generating distribution. The other lines are the pdfs of, from left to right, $f_{1:10}$ to $f_{10:10}$, obtained using Equation 7.2.

Similar to the multivariate extension of EVT presented in Chapter 5, we proceed to building a multivariate extension of the classical order statistic definition.

7.3 Order Statistics in multivariate space

7.3.1 Density-ordered Statistic

Inherent to the notion of order statistic is that of rank; i.e., the existence of a manner of ranking observations, such that one can be deemed greater than the other. In classical univariate order statistical theory, the ranking function is the natural order on the real line, or a subset of the real line. If the distribution support D_f is multivariate, any function $r : D_f \longrightarrow \mathbb{R}$ may be used to rank two observations $\mathbf{x}_1$ and $\mathbf{x}_2$, such as the Euclidean norm where $D_f = \mathbb{R}^d$. In accordance with Definition 5, this thesis adopts the view that novelty is characterised by an abnormally low likelihood for a sample or an observation to occur with respect to an assumed data-generating model. In the case of order statistic-based novelty detection, this translates to a low likelihood for the value of the selected order statistic to occur. We therefore need a ranking function r which respects the structure of the model; i.e. a function r such that $\forall \mathbf{x}_1, \mathbf{x}_2 \in D_f$, $f(\mathbf{x}_1) < f(\mathbf{x}_2)$, $r(\mathbf{x}_1) < r(\mathbf{x}_1)$. The natural choice is, of course, to select f itself as the ranking function, and define density-ordered statistics as:

Definition 8. *(Density-ordered statistic)*
Let $(X_1, X_2, \ldots, X_m)$ be an i.i.d. sample of size m drawn from a (possibly multivariate) distribution function F, with pdf f. Arrange $f(X_1), f(X_2), \ldots, f(X_m)$ in increasing order of magnitude and let be $Y_1, \ldots, Y_m$ be the ordered values. The k^{th} density-ordered statistic is defined as the X_i responsible for the k^{th} order statistic of the sequence $f(X_1), f(X_2), \ldots, f(X_m)$; i.e, X_i such that $f(X_i) = Y_k$. We denote $X_{k|m}$ the density-ordered statistic, and $Y_{k:m} = f(X_{k|m})$ the associated order statistic in probability space.

It follows from Definition 8 that there is a one-to-one correspondence between a density-order statistic in data space and an order statistic in the probability space $\mathcal{P}_f = f(D_f)$. Since $\mathcal{P}_f$ is univariate, the results of Section 7.2 apply, and in particular Equations 7.2 and 7.3 can be

applied to the distribution of probability densities $G(.;f)$, which allows us to compute novelty scores.

7.3.2 Novelty scores

Let $X = (X_1, X_2, \ldots, X_m)$ be a sample of observations. The novelty score of the k^{th} density-ordered statistic of X is simply that of the k^{th} order statistic of $f(X_1), f(X_2), \ldots, f(X_m)$ in the probability space. Assuming the distribution of probability density $G(.;f)$ is known, then the natural novelty score of $X_{k|m}$ in data space is that of $Y_{k:m} = f(X_{k|m})$ in probability space.

Assuming that we are predominantly concerned with unusually low values of the order statistic in probability space, the natural novelty function $z(y;f,m,k)$ over the probability space is:

$$\begin{aligned} z(y;f,m,k) &= 1 - G_{k:m}(y;f) \\ &= 1 - m\binom{m-1}{k-1}\sum_{j=0}^{m-k}\binom{m-k}{j}\frac{G(y;f)^{k+j}}{k+j}. \end{aligned} \tag{7.4}$$

Equation 7.4 is obtained by substituting $G_{k:m}(y;f)$ with its expression from Equation 7.3. In turn the novelty function over the data space is given by:

$$z(x;f,m,k) = 1 - m\binom{m-1}{k-1}\sum_{j=0}^{m-k}\binom{m-k}{j}\frac{G(f(x);f)^{k+j}}{k+j}. \tag{7.5}$$

The log-novelty functions are obtained by applying the transform $q \mapsto -\log(1-q)$ to the natural novelty functions. Figure 7.2 shows an example of novelty calculations using the standard normal distribution. The lower graph shows the pdf of $g(.;f)$, given in Equation 4.41, alongside the pdfs of $Y_{1:20}$ to $Y_{20:20}$, calculated by replacing the expression of the pdf and cdf in Equation 7.2 by the expressions in Equations 4.39 and 4.41. The upper graph shows the pdf of the model, alongside the natural novelty functions in data space (Equation 7.5), for $m = 20$ and $k = 1 \ldots m$. The critical point to observe concerning the distributions of order statistics in Figure 7.1 is that the novelty functions are symmetric: novelty scores are assigned based solely on density values, and not on the relative position of the observation to the mode of the distribution.

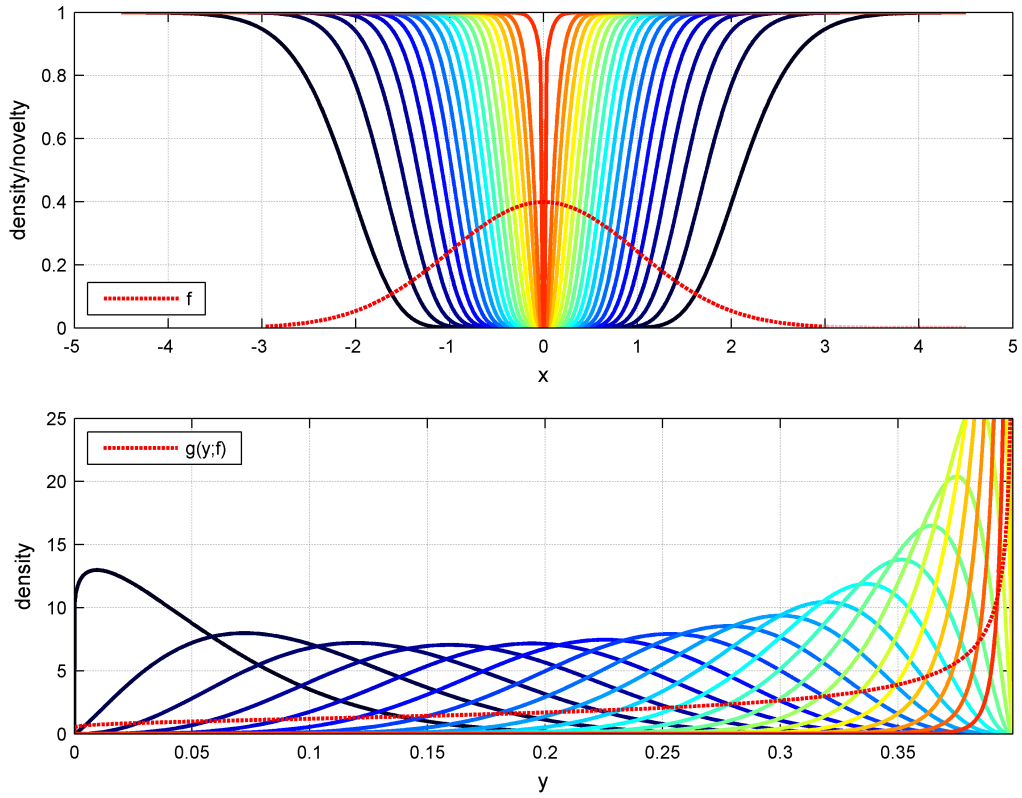


Figure 7.2: Upper graph: Novelty functions based on density-ordered statistics for the standard normal distribution. The red dotted line is the pdf of the model. The continuous lines are the novelty functions $z(x;f,20,1)$ (in black) to $z(x;f,20,20)$ (in red). Lower graph: pdf of the novelty functions in the probability space, $z(y;f,20,1)$ to $z(y;f,20,20)$ with the same colour coding. The red dotted line is the pdf of the distribution of the probability densities.

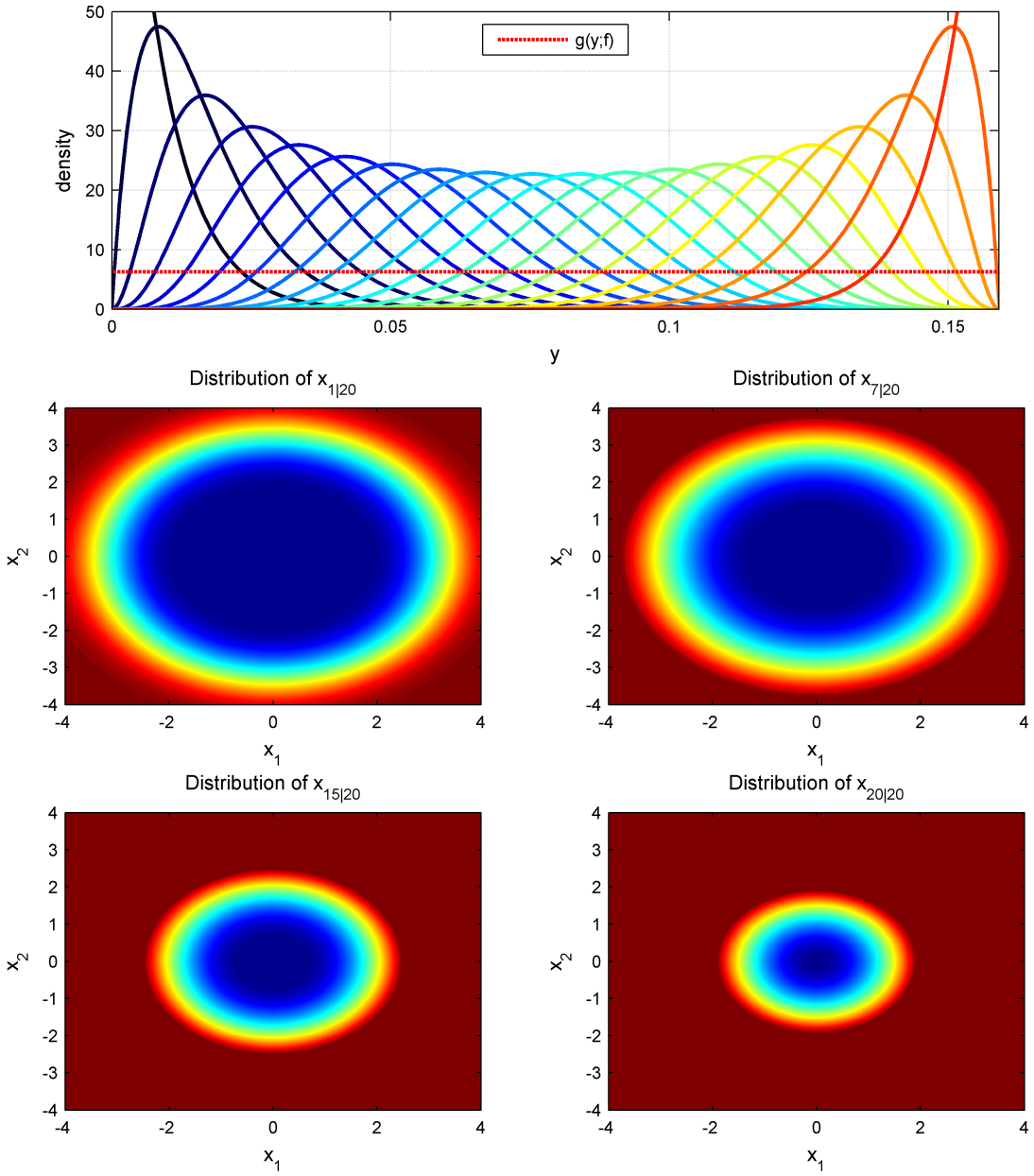


Figure 7.3: Upper graph: pdf of the novelty functions in the probability space for the standard two-dimensional Gaussian distribution, $q(y;f,20,1)$ in black, to $q(y;f,20,20)$ in red. Again the model pdf is shown by a dashed red line. Other graphs: corresponding log-novelty maps for $\mathbf{x}_{1|20}$, $\mathbf{x}_{7|20}$, $\mathbf{x}_{15|20}$, and $\mathbf{x}_{20|20}$.

The equivalent two-dimensional example is shown in Figure 7.3, where f is the standard two-dimensional Gaussian distribution. The expression of the DPD is known from 5.14 to be the uniform distribution over the probability space. The pdfs of the $Y_{k:m}$ are obtained by combining Equation 7.2 with Equations 5.14 and 5.32 and are therefore exact. The series of log-novelty maps represented behave as expected: at a location $\mathbf{x} = (x_1, x_2)$, the higher the value of k, the greater the novelty score.

7.3.3 Similarities with the "4-out-of-5" heuristic filtering

The 4-out-of-5 filtering used in Algorithm 1.1 aims at reducing the influence of transient noise on the novelty scores. The requirement that the novelty score be greater than the novelty threshold for 4 out of 5 minutes before an alert is triggered is equivalent to requiring that (assuming a fixed sample rate of the time-series) the 80% density-ordered statistic in 5 minutes be greater than some novelty threshold. Substituting an extreme statistic by a density-ordered statistic is therefore expected to increase the robustness of our classifier from Chapter 6 to such transient noise. Using Algorithm 7.1 enables us to set novelty scores which retain a probabilistic interpretation while attempting to improving robustness to transient noise.

7.4 Application to the UPMC data set

7.4.1 Novelty detection procedure

The data and model f_2 described in Section 6.4.1 and 6.4.2, respectively, are used again in this application. The novelty detection procedure is presented in Algorithm 7.2. The values of m considered are 6, 12, 18, 24, and 30 (corresponding to windows of duration 1 to 5 minutes). For each value of m, the values of k considered correspond to the 20%-, 50%-, and 80% density-ordered statistics. The model f_2 is used to rank observations.

A notable difference with earlier procedures is the addition of step 5 in the initialisation phase in Algorithm 7.2. Although $G_{k:m}(.; f_2)$ need not be estimated if an estimate of $G_2^-(.; m)$ is

Algorithm 7.2 Novelty Detection procedure in HR/BR/SpO_2 time-series with density-ordered statistics

Initialisation:

1. Estimate the probability distribution of HR/BR/SpO_2 observations f_2.
2. Compute estimates of $G(.;f)$
3. Choose a value for the detection window length $m \in \mathbb{N}^*$ and estimate $G_2^-(.;m)$.
4. Choose an order $k \in \mathbb{N}^*$, $k \leq m$.
5. Estimate the probability distribution of $y_{k:m}$, which we shall abusively term $G_{k:m}(.;f_2)$.
6. Choose a value for the level of significance $\alpha > 0$, for determining the threshold on log-novelty $Z_1(x;m)$.

For a time-series of HR/BR/SpO_2 observations:

1. Starting with the window at the beginning of the time-series:
2. Identify $y_{k:m}$ for the window with respect to f_2; i.e., the k^{th} density-ordered statistic of the window.
3. Compute the window's log-novelty score $Z(x;f_2,m,k) = -\log[G_{k:m}(y_{k:m};f_2)]$.
4. If $Z(x;m) > \alpha$, an alert is deemed to have occurred.
5. Otherwise, output $z = 0$ (normal state)
6. Advance the window by 10 seconds.
7. Repeat steps 2 to 6 until the end of the time-series is reached.

known (using Equation 7.3), in practice using the distribution of the density-ordered statistic in probability space yields better results, essentially because the i.i.d. assumption required to obtain 7.3 does not hold for vital-sign time-series. Naturally, forming the distribution $G_{k:m}(.;f_2)$ is only made possible by the large amount of data that we have at our disposal in this particular example.

7.4.2 Results

Novelty detection is performed on the UPMC data set using Algorithm 7.2. Receiver operating characteristics at the four levels of detection defined in Section 6.2.2 are shown for $k = \lfloor 0.2m \rfloor$, $\lfloor 0.5m \rfloor$, and $\lfloor 0.8m \rfloor$ in Figures 7.5, 7.4, and 7.6, respectively. AUCs for all values of m and k are reported in Table 7.1. It may be seen that, for constant m, AUCs are similar for all values of k, illustrating the transient nature of the noise present in the dataset.

Comparing these AUCs with those in Table 6.3, it is clear than the approach used in this chapter significantly outperforms the multivariate EVT approaches used in Chapter 6:

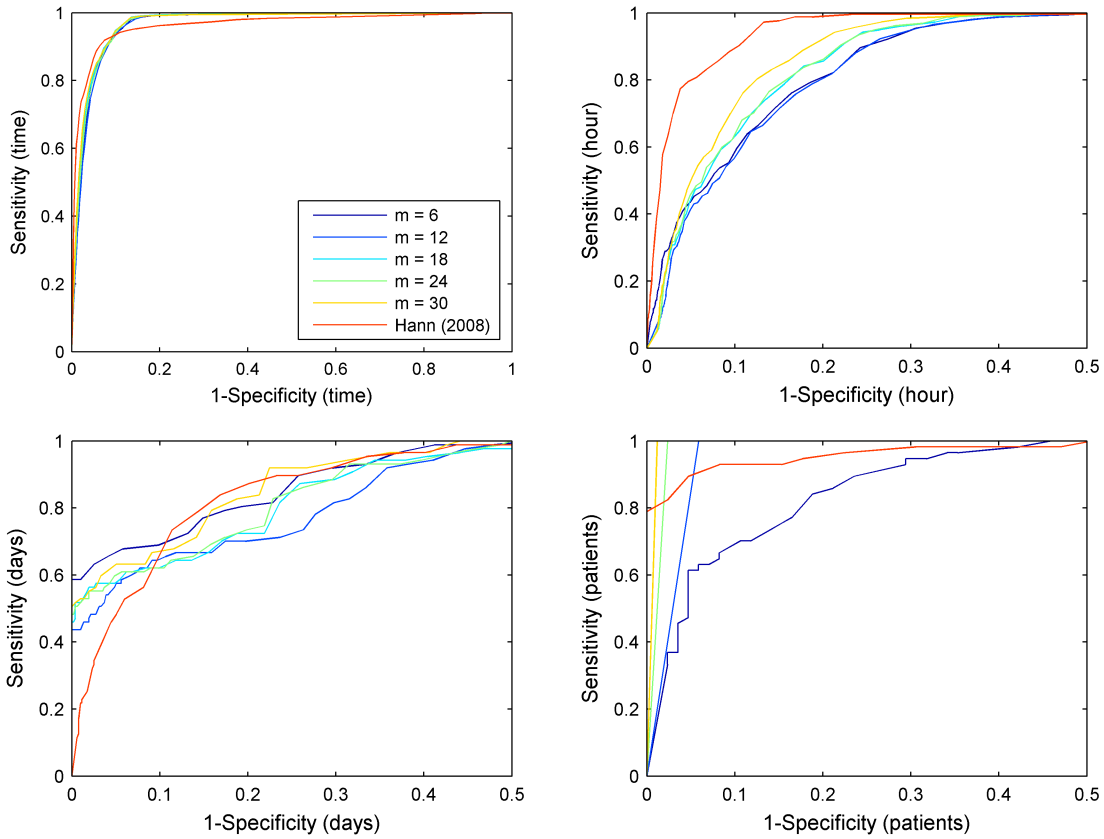


Figure 7.4: ROC curves of the novelty detection procedure using the 20% order statistic.

- At the level of the observation, which is the level of detection at which the multivariate EVT approach performs best, AUCs are improved for all values of m and k considered. This improvement can be seen to increase as m increases: for $m = 6$, the increase in AUC can be seen to be in the 0.63-0.70 percentage-point range, where it is in the 2.02-2.30 range for $m = 30$. At this level, all improvements over the pdf-thresholding method are statistically significant at the 0.05-level.

- At the level of the monitoring hour, AUCs are comparable for $m \leq 18$, and become significantly better for $m > 18$.

- The best improvements are seen at the level of the monitoring day, at which the original multivariate EVT method performed the most poorly. For $m = 6$, the difference in AUC between the two methods can be seen to be greater than 15 percentage points in favour of Algorithm 7.1, from 76.83% in Table 6.3 to 92.16%. Improvements of comparable magnitudes are observed for all values of k and m under consideration, and a number

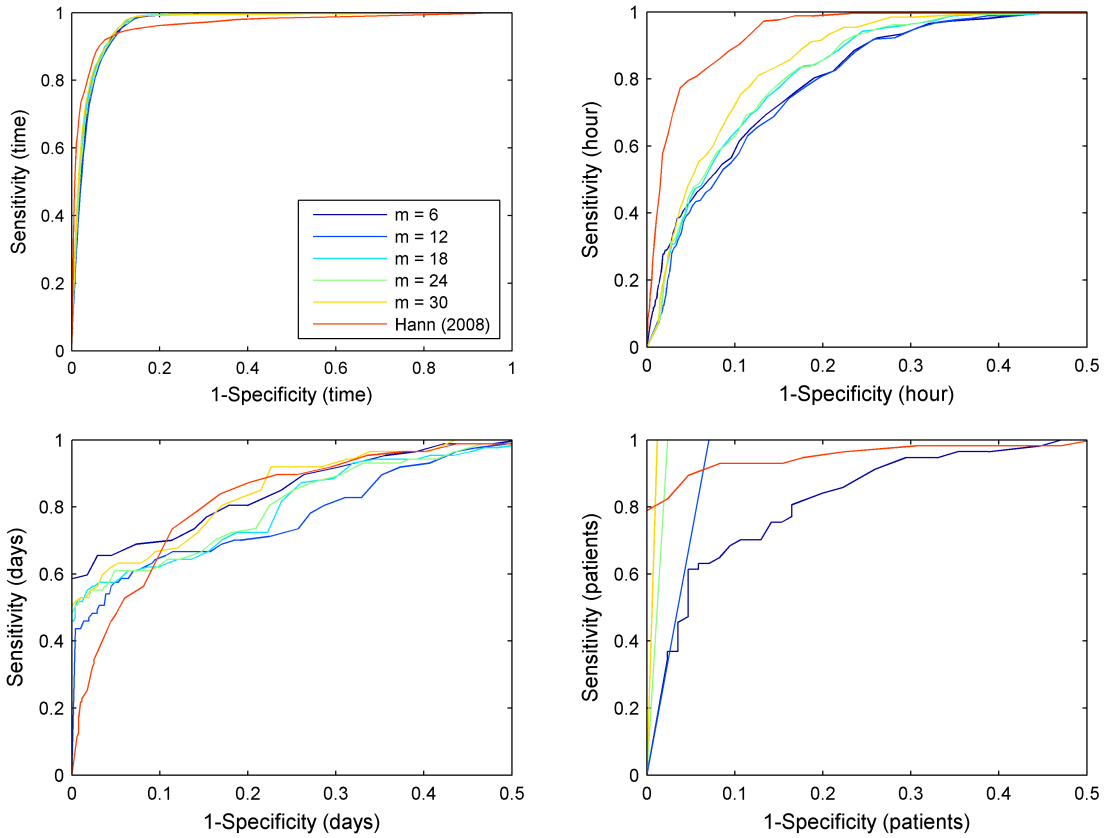


Figure 7.5: ROC curves of the novelty detection procedure with order statistics using the median. Values of k are listed in Table 7.1.

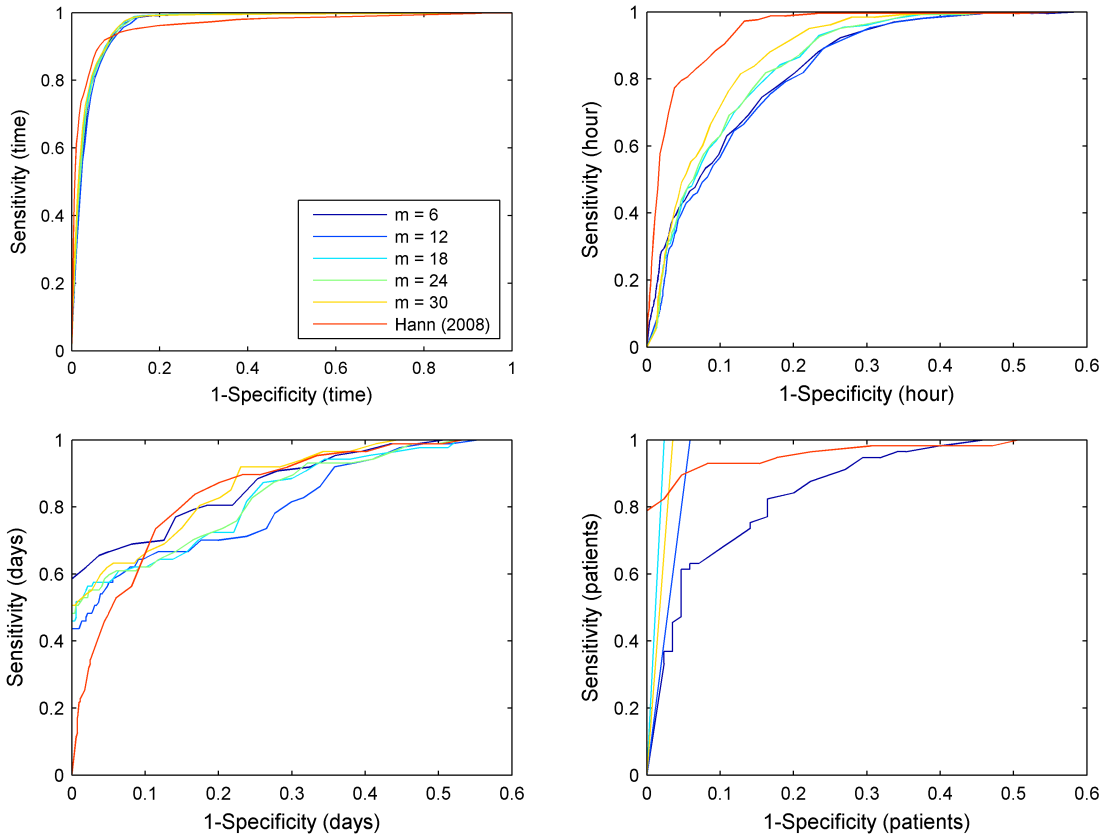


Figure 7.6: ROC curves of the novelty detection procedure using 80% order statistic. Values of k are listed in Table 7.1.

m	k	AUC (time)	AUC (hours)	AUC (days)	AUC (patients)
6	1	**96.8***	89.7*	**92.1**	90.9*
12	2	**96.6***	89.1*	88.2	97.1
18	3	**97.0***	90.7*	89.7	**98.8**
24	4	**96.9***	90.9*	89.9	**98.8**
30	6	**97.1***	92.3*	**92.1**	**99.4**
6	3	**96.7***	89.7*	**92.2**	90.8*
12	6	**96.6***	89.1*	88.1	96.5
18	9	**96.9***	90.6*	89.7	**98.9**
24	12	**96.8***	90.8*	89.9	**98.8**
30	15	**97.0***	92.3*	**92.1**	**99.4**
6	4	**96.8***	89.7*	**92.1**	91.0*
12	9	**96.6***	89.1*	88.2	97.1
18	14	**97.0***	90.6*	89.6	**98.8**
24	19	**96.9***	90.8*	89.7	**98.2**
30	24	**97.1***	92.3*	**92.0**	**98.2**
Hann (2006)		95.9	96.5	90.7	97.1

Table 7.1: AUCs of novelty detection in HR/BR/SpO_2 time-series using 20%-, 50%- and 80%-Order Statistics, as percentages. The values in bold font are those which exceeds Hann's at the same level of detection. An asterisk next to an AUC denotes that the difference in AUC between this AUC and that obtained by Hann (2008)'s classifier at the same level of detection is statistically significant at the 0.05-level, using Hanley et al. (1983)'s method for comparing AUCs.

of order-statistic based classifier outperform Hann's. The difference are however not statistically significant given the population of monitoring days considered.

- At the level of the patient stay, improvements are observed except for $m = 6$. In all other cases AUCs are improved by at least 4 percentage points, from 92.23% in Table 6.3 to 98.82% for $m = 18$ and $k = 14$. While this is a positive result, the differences are not statistically significant.

The significant improvement observed at the level of the monitoring day, compared to the lesser improvement seen at the level of the monitoring hours indicates that monitoring hours which are harder to classify tend to occur consecutively; i.e, during the same monitoring day.

Lastly, it is possible to select values of m and k for which the density-ordered approach outperforms Algorithm 1.1 at all levels of detection, except the monitoring hour. This is the case for all values of k considered when $m = 30$ (which corresponds to a detection window of 5 minutes).

We argue that we have therefore created a probabilistic approach that is robust to transient noise, and which yields fewer and more accurate alerts by monitoring day, and throughout a patient's stay.

Chapter 8

Conclusions

8.1 Summary of results

This thesis has described three approaches to novelty detection in time-series data, with applications in vital-sign monitoring. All three stem from the observation that novelty detection using complex multivariate, multimodal generative models is often an ill-defined problem when attempted directly in the data space. In situations where "novel" is equivalent to "improbable with respect to a data-generating probability distribution", we show that performing novelty detection in the (always) univariate probability space allows us to (i) use classical results of univariate statistical theory, such as those of (univariate) extreme value theory; (ii) in some cases, derive analytical forms for the distribution over probability densities (DPDs) of multivariate distributions; (iii) in cases where this is not possible, provide estimates for the tails of these DPDs.

Three methods for performing novelty detection have been derived and applied to a vital-sign data set:

- A method suitable for novelty detection in time-series generated from a univariate multimodal model (Chapter 4), in which the DPD of a multimodal univariate distribution is computed (or estimated), and the (possibly asymptotic) distribution of its lower tail is formed, to be used as the distribution of the test statistic (the extreme of newly-acquired

samples) in a hypothesis test (Algorithm 4.1). This method is extended to apply to windows of data in a time-series of i.i.d. observations (Algorithm 4.2).

- A method suitable for novelty detection in time-series generated from a multivariate, possibly multimodal model (Chapter 5). First, we extend the method to the case where the data-generating distribution is a multivariate Gaussian, for which we provide closed-form expressions of the DPDs, and both exact and asymptotic EVDs of the DPDs. Second, it is extended to the case where the data-generating distribution is a mixture of Gaussians for which we provide two methods of estimating the lower tail of the DPDs.

- A method based on order statistics, suited to time-series containing transient noise (Chapter 7), in which a non-extremal order statistic is substituted for the minimum in density, so as to make classification more robust to transient noise.

8.1.1 Vital-sign monitoring

All three methods (the univariate EVT, multivariate EVT, and model with order statistics) are applied to a vital-sign data set in which periods of physiological abnormality have previously been identified. All methods are shown to outperform the existing continuous vital-sign monitoring approach at the observation-by-observation level. The EVT methods, however, significantly underperform compared with the existing methods over longer time-scales (monitoring hours, days, and patient stays), due to their sensitivity to artefactual data. The existing approach in the literature uses a heuristic filtering step, the effect of which we attempt to reproduce with the density-ordered statistics method. This last method is shown to outperform the existing approach at all levels of detection but the monitoring hour, while retaining a probabilistic interpretation.

8.1.2 Novelty Detection

In addition to applying novelty detection to vital-sign monitoring, contributions in this thesis are also valid in the context of data-modelling problems in general. As noted in Section

4.3.1, it is easy to imagine that extreme magnitudes most often correspond to improbable states. Operating the transition from data space to probability space may prove useful in other applications. Calculations ultimately occur in a univariate space for our method, and so increasing the complexity of a model (and also increasing its dimensionality) does not necessarily imply an exponential growth in computation requirements. In cases where the model is known, obtaining a decision boundary may be done by estimating a univariate distribution (noting that each location in probability space corresponds to a hypersurface - or level set - in data space), and estimating its lower tail.

Furthermore, our methods do not include any heuristic element. They are designed to yield novelty scores that all have a probabilistic interpretation, independent of the number of modes, dimensionality of the model, or indeed of the structure of the data space. This makes them suitable for hypothesis testing, and, we argue, easier to interpret.

A limit of our methods is their reliance on the i.i.d. assumption. We show in Chapter 7 that the assumption can be relaxed by forming an empirical unconditional distribution of density-ordered statistics (or, indeed, extremes), instead of using the distributions theoretically associated with the DPD under the i.i.d. assumption, although this requires a large amount of data.

Another characteristic of our methods is that they rely on the assumption that the structure of the normal data fully inform that of the tail data, under normal circumstances. While this may be a necessary assumption in situations where few data are available, it may not hold for complex systems in which extreme, yet not abnormal behaviour, differs from normal behaviour. Potential solutions to both these limitations are presented in the next section.

8.2 Future work

8.2.1 Tail-only estimation in multivariate space

In *generative* approaches to novelty detection, the probabilistic structure of the "normal" class is modelled with pdfs $f(\mathbf{x})$ which may then be used to find a decision boundary. Examples

of such approaches include, for example, mixture models, used to estimate the density $f(\mathbf{x})$ of an observation, which can in turn be related to a novelty score. Classification may then be performed by setting a threshold on the novelty score. Generative approaches use most of the training dataset in an attempt to determine its underlying structure, assuming that the data, including the tail data is generated from that structure.

Discriminative approaches aim at modelling directly the novelty decision boundary. Examples include kernel methods, such as one-class support vector machines (SVMs). Such approaches are typically sparse, and do not attempt to discover the underlying probabilistic structure of the data, concentrating instead on estimating the decision boundary that is required to perform classification. In practice, this can mean using only that subset of the data that is required to determine the decision boundary, such as the support vectors in an SVM, which often lie in areas of low data density (Clifton, Clifton, Watkinson & Tarassenko 2011). Discriminative approaches, unlike generative approaches, do not attempt to model the distribution of the normal data, and therefore do not rely on the assumption that the structure of the tail data is necessarily informed by that of the normal data. The probability mass of the most normal data may be rearranged in such ways that the decision boundary associated with a generative approach may change, while that of a discriminative approach would not. It is however a requirement that the available data set contains a sufficient quantity of tail data to estimate accurately the decision boundary, which may not always be the case for high-integrity systems.

The peaks-over-threshold (POT) method of EVT offers an appealing compromise between using all of the data (to discover probabilistic class structure) and using only those data that lie close to the edge of the region of support of normal data. It is a generative method in which some underlying probabilistic structure is assumed (although it may not explicitly be modelled), and is conventionally used to analyse the data that lie in the tail of univariate or low-dimensional distributions.

The POT method considers exceedances over (or shortfalls under) some extremal threshold $u \in \mathbb{R}^d$, where typically $d = 1$, or a small number of dimensions (Embrechts et al. 1997). Assuming that the maximum X_m^+ of a set of m i.i.d. random variables X_i is distributed according to the

Generalised Extreme Value distribution (Equation 4.14), then it may be shown (Pickands 1975) that the distribution of exceedances $X_i > u$ converges towards a known form when $u \to +\infty$, which is the Generalised Pareto Distribution (GPD):

$$H(x) = \begin{cases} 1 - \left\{1 + \xi \left(\frac{x-c}{d}\right)^{-1/\xi}\right\} & \text{, if } \xi \neq 0 \\ 1 - \exp\left(-\frac{x-c}{d}\right) & \text{, if } \xi = 0, \end{cases} \tag{8.1}$$

where c, d, and ξ are the location, scale, and shape parameter, respectively. They can be estimated by maximum likelihood (Castillo et al. 2005).

The GPD, and therefore the POT method, are typically used to form models of extremal data directly, where such data are observed, yet relatively sparse. The selection of the value of the threshold u is important: set u to be too "extreme" (i.e, too far into the tail of f), and the tail will be fitted well by the GPD (because it is an asymptotic distribution), but there will be few training examples that exceed the threshold, resulting in estimates of the parameters which will have high variance. Conversely, setting u to be too "normal" (i.e., too close to the regions of high density) will provide a larger number of training observations that exceed the threshold, and so estimates of the parameters will have lower variance, but there will be high bias, because this non-extremal tail will not be well-fit by the GPD. Conventional practice is, therefore, to examine the variance and bias of the result of fitting a GPD to the tail data for various extremal values of u.

Using the above equation, we obtain the probability $P(X_m^+ - u | X_m^+ > u) = H(X_m^+)$, which is the probability of a maximum X_m^+ greater than u taking a value less than its own. This can therefore be used as novelty score. For example, a natural novelty score (similar to that defined in Equation 4.36) for univariate maxima would be:

$$q(x; f, m) = H(x), \tag{8.2}$$

which takes value 0 if $x < u$, and converges to 1 when $x \to +\infty$. A threshold may be set on the novelty score function for the purpose of classification.

Analysis of the tails of multivariate distributions has been limited by the conventional approach of estimating the dependence structure between the extremes of the random variables (Castillo et al. 2005). In particular, the POT method relies on our ability to identify, and therefore define, extremes in multivariate space. The same reasoning as in Sections 4.3.2 ("Redefining extremes") and 5.1 ("From univariate to multivariate data space") therefore applies, and we propose to use the method adopted in Section 4.3.2 in such cases, where evidence of novelty may be defined as the observation of improbable data, for multivariate (potentially multimodal) distributions of the normal data.

We propose to use an estimate of the pdf f for all the available training data to map the data space into the corresponding univariate probability space $P_f = f(D_f)$. According to Definition 5, the set i.i.d. of (possibly multivariate) rvs $f(X_i)$ is a i.i.d. set of univariate random variables in P_f, whose distribution function we term $G(.;f)$. Minima in density with respect to f also correspond to minima of $G(.;f)$, and the POT method may be applied to the lower tail of $G(.;f)$. Details of the method are available in (Clifton, Clifton, Hugueny & Tarassenko 2013).

The conventional POT method ideally does not rely on any knowledge of the normal data structure. Our proposed method does, however, for the purpose of mapping the data space into a one-dimensional probability space. In the probability space, only the lower-tail data are examined, and the exact structure of the normal data in probability space need not be considered. The resulting GPD therefore need not be equal to the corresponding EVD obtained using multivariate extreme value theory, thus providing a more accurate tail estimate when tail and normal data are structured differently.

8.2.2 Novelty detection with joint order statistics

The multivariate EVT approach presented in Chapter 6 makes use of a single observation, the extreme, to classify a sample comprising m observations. Similarly, the density-ordered statistic approach in Chapter 7 uses a single (ordered) observation. In both cases, the statistic used implicitly contains information about the magnitude of the remaining observations, relative to that of the order statistic. The exact locations of the unused observations does not change the

output novelty score of the whole sample.

A more comprehensive approach would be to build classifiers which consider several or all the sample's order statistics, thus making use of all the information contained in the sample.

Assuming $\{X_i\}_{i=1}^m$ random variables with univariate pdf f, the joint density function of all m order statistics is (Arnold et al. 2008):

$$f_{1,2,\ldots,m:m}(x_{1:m}, x_{2:m}, \ldots, x_{m:m}) = m! \prod_{j=1}^{m} f(x_{j:m}), \tag{8.3}$$

with, by definition, $x_{1:m} \leq x_{2:m} \leq \ldots \leq x_{m:m}$. Now, let f be a multivariate pdf, and $y_{i:m}$ be the i^{th} density order statistic of $\{\mathbf{x}_i\}_{i=1}^m$. In probability space, we have:

$$g_{1,2,\ldots,m:m}(y_{1:m}, y_{2:m}, \ldots, y_{m:m}; f) = m! \prod_{j=1}^{m} g(x_{j:m}; f). \tag{8.4}$$

The above equation defines the m-variate pdf of the densities of all m order statistics. All variables of this pdf have the same support and dimension. To perform novelty detection, a threshold may be applied to this pdf, which we term $gg(.;f)$[1]. If $gg(.;f)$ is not degenerate [2], gg may be used to perform hypothesis testing, in a similar fashion to Algorithm 4.1. Such a test would be an exact test of whether the density-ordered statistics of the sample are collectively distributed as expected if the sample were indeed drawn from f.

In practice, analytical expressions of gg are unlikely to exist for even simple models, and would have to be numerically estimated, making such a test impractical for use with large sample sizes. Additionally, testing using all of the information in the sample means that the test is extremely sensitive to even small deviations from the model, as well as cases where the i.i.d. hypothesis is not valid. In particular, it could not be adapted to novelty detection in vital-sign time-series, where the population pdf is used as the model, but where the i.i.d. hypothesis does not hold for any particular patient or window of data.

[1] or the distribution of densities of $g_{1,2,\ldots,m:m}$ may be calculated (or estimated), resulting in a univariate distribution (over the probability space of $g_{1,2,\ldots,m:m}$).

[2] which happens, if for instance, g is constant, as is the case if f is a univariate uniform distribution, or a bivariate Gaussian distribution (see Figure 5.6).

By selecting a subset of density-ordered statistics, such as the minimum and the median for instance, it appears possible to construct classifiers which have some tolerance to single-point noise, while retaining sensitivity to deviation of the most improbable observation.

8.2.3 Towards an Extreme Function Theory

A different approach to novelty detection with time-series which makes use of all the information in a window, is one that we term "extreme function theory" (EFT). In previous chapters, we have adapted point-wise methods to be applicable to arbitrary sample sizes with the essential requirement that samples must be i.i.d. Furthermore, we have limited ourselves to using a feature of the samples (i.e., an order statistic), leaving the dynamics of sample unconsidered. We noted in the previous section that increasing the amount of the information in use under the i.i.d. assumption makes the resulting classifiers too sensitive to be used with real-life time-series.

Instead, we may assume that the data represent a time-series, where the i.i.d. assumption does not hold. Adopting a point-wise, sample-by-sample approach to classification can result in large numbers of assumedly-independent decisions (perhaps at the sampling rate of the data, as we have described in earlier chapters). Instead, we suggest that a single decision can be taken, testing an entire time-series, which represents a test function. We thus adopt a functional approach to novelty detection, where functions are represented by time-series of discrete observations.

Following the novelty detection approach, in which a model of normality $\mathcal{M}$ is constructed from "normal" training examples, we assume a set of training time-series X_T, from which a model $\mathcal{M}$ is constructed. We let $\mathcal{M}$ be a Gaussian Process (GP), which will allow us to use some of the calculations of Chapter 5 (specifically DPD calculations in Section 5.4). The hyperparameters of the GP used to represent $\mathcal{M}$ could be determined using a number of methods (maximisation of the joint likelihood of all training time-series, for instance). We however note that the structure of the GP's covariance matrix allows us to model time-series for which subsequent observations are neither identically distributed, nor independent. Assuming the presence of a GP model $\mathcal{M}$, we propose to formulate a hypothesis test H_0 to classify a test time-series as being "normal" (i.e, generated from $\mathcal{M}$ to some significance level α), or otherwise "abnormal".

Assigning probabilities to functions

To formulate our hypothesis test, we require a mapping from functions to probabilities. The conventional approach consisting of considering the marginal likelihood given the test of input data X yields probability densities, not probabilities, which therefore scale with the dimensionality n of the GP.

The GP offers the convenience that the joint distribution over the test data is multivariate Gaussian, and so we could formulate H_0 by determining the probability of observing an n dimensional vector from the n-dimensional multivariate Gaussian distribution, which can be performed following the results in Section 5.4 by forming the distribution over the probability densities allowed by the GP $\mathcal{M}$.

For some n, the GP distribution over functions is multivariate Gaussian by definition, and a function s may be evaluated to give a density y using the GP, so that $y = f_n(s)$, where f_n is a multivariate Gaussian pdf defined by the GP. The DPD $G_n(.; f_n)$ is such that it can be used to assign probabilities to each function that we wish to examine. An initial hypothesis test to classify a single function H_0 could be used to compare $G_n(x, f_n)$ to some threshold significance level $\alpha \in [0,1]$, and classify those functions as "abnormal" for which $G_n(x; f_n) < \alpha$. This corresponds to rejecting "abnormal" functions that could have been generated from the GP model $\mathcal{M}$ with a probability below significance level α. This probabilistic interpretation is not affected by the dimensionality of $\mathcal{M}$.

Extreme function theory

Let $\{X_i\}_{i=1}^m$ be a set of m functions, and let us assume that they were independently drawn from the GP model $\mathcal{M}$ (though observations forming a given function may not be independent, nor identically distributed). Each of the X_i may be considered as being a multivariate vector of the GP model $\mathcal{M}$, and the set of test functions may be considered as being an i.i.d. sample of the generative Gaussian distribution. All the methods and results in Sections 5.5 and 5.6 apply, and can be used to determined if, of a sample of i.i.d. functions, the most improbable breaks a level

of significance.

We have therefore sketched an extreme function theory for novelty detection, which allow us to classify with a probabilistic interpretation how "improbable" a sample of i.i.d. observed functions is with respect to a GP model. While the set of observed functions needs to be i.i.d., the discrete functions themselves may exhibit any behaviour that can be modelled by a Gaussian process. A full technical treatment, as well as applications to a vital-sign data set is given in (Clifton, Clifton, Hugueny, Wong & Tarassenko 2013).

Bibliography

Abildstrom, S. Z., Jensen, B. T., Agner, E., Torp-Pedersen, C., Nyvad, O., Wachtell, K., Ottesen, M. M. & Kanters, J. (2003), 'Heart rate versus heart rate variability in risk prediction after myocardial infarction.', *Journal of cardiovascular electrophysiology* **14**(2), 168–73.

Aboukhalil, A., Nielsen, L., Saeed, M., Mark, R. G. & Clifford, G. D. (2008), 'Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform.', *Journal of biomedical informatics* **41**(3), 442–51.

Abraham, B. & Chuang, A. (1989), 'Outlier detection and time series modeling', *Technometrics* **31**(2), 241–248.

Anderson, J. (2008), Biomedical Signal Processing : Application of Support Vector Machines for Detecting Deterioration in Acutely Ill Patients, Technical Report September.

Arnold, B. C., Balakrishnan, N. & Nagaraja, H. N. (2008), *A First Course in Order Statistics (Classics in Applied Mathematics)*, SIAM.

Arroliga, A. C., Sanchez, J. & Sikka, P. (2011), 'Monitoring the ill: Is this another measurement or will it change outcomes?', *Crit Care Med* **39**(1), 202–203.

Basmann, R. (2003), 'Statistical outlier analysis in litigation support: the case of Paul F. Engler and Cactus Feeders, Inc., v. Oprah Winfrey et al.', *Journal of Econometrics* **113**(1), 159–200.

Basu, S., Bilenko, M. & Mooney, R. J. (2004), 'A probabilistic framework for semi-supervised clustering', *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04* pp. 59–68.

Bellomo, R., Goldsmith, D., Uchino, S., Buckmaster, J., Hart, G. K., Opdam, H., Silvester, W., Doolan, L. & Gutteridge, G. (2003), 'A prospective before-and-after trial of a medical emergency team', *Medical Journal of Australia* **179**(6), 283–288.

Bishop, C. M. (1994), 'Novelty detection and neural network validation', *IEE Proceedings-Vision, Image and Signal processing* **141**(4), 217–222.

Bishop, C. M. (2007), *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. 2006. corr. 2nd printing edn, Springer.

Brennan, P. & Silman, A. (1992), 'Statistical methods for assessing observer variability in clinical measures.', *British Medical Journal* **304**(6840), 1491.

Broomhead, D. S. & King, G. P. (1986), 'Extracting qualitative dynamics from experimental data', *Physica D: Nonlinear Phenomena* **20**(2-3), 217–236.

Buchman, T. G. (2010), 'Novel representation of physiologic states during critical illness and recovery.', *Critical care (London, England)* **14**(2), 127.

Budzyński, R. J., Kondracki, W. & Królak, A. (2008), 'Applications of distance between probability distributions to gravitational wave data analysis', *Classical and Quantum Gravity* **25**(1), 015005.

Burbea, J. & Rao, C. R. (1982), 'Entropy differential metric, distance and divergence measures in probability spaces: A unified approach', *Journal of Multivariate Analysis* **12**(4), 575–596.

Burridge, P. & Robert Taylor, A. M. (2006), 'Additive outlier detection via extreme-value theory', *Journal of Time Series Analysis* **27**(5), 685–702.

Castillo, E., Hadi, A. S., Balakrishnan, N. & Sarabia, J. M. (2005), *Extreme value and related models with applications in engineering and science*, Wiley-Interscience.

Chavez-Demoulin, V. & Roehrl, A. (2004), 'Extreme Value Theory can save your neck', *Approximity whitepaper, Bk* pp. 1–5.

Choe, J. & Shroff, N. B. (1998), 'A Central-Limit-Theorem-Based Approach for Analyzing Queue Behavior in High-Speed Networks', *IEEE/ACM Transactions on networking* **6**(5), 659–671.

Classen, D. C., Resar, R., Griffin, F., Federico, F., Frankel, T., Kimmel, N., Whittington, J. C., Frankel, Seger & James, B. C. (2011), 'Global Trigger Tool Shows That Adverse Events In Hospitals May Be Ten Times Greater Than Previously Measured', *Health Affairs* **30**(4), 581–589.

Clifton, D. A. (2009), Novelty detection with extreme value theory in jet engine vibration data, PhD thesis, University of Oxford.

Clifton, D. A., Hugueny, S. & Tarassenko, L. (2009*a*), A comparison of approaches to multivariate extreme value theory for novelty detection, *in* 'IEEE Workshop on Statistical Signal Processing Proceedings', pp. 13–16.

Clifton, D. A., McGrogan, N., Tarassenko, L., King, D., King, S. & Anuzis, P. (2008), Bayesian extreme value statistics for novelty detection in gas-turbine engines, *in* 'Proceedings of IEEE Aerospace', Citeseer.

Clifton, D., Clifton, L., Hugueny, S. & Tarassenko, L. (2013), 'Extending the Generalised Pareto Distribution for Novelty Detection in high-dimensional spaces', *Journal of Signal Processing Systems* pp. 1–17.

Clifton, D., Clifton, L., Hugueny, S., Wong, D. & Tarassenko, L. (2013), 'An extreme function theory for novelty detection', *Selected Topics in Signal Processing, IEEE Journal of* **7**(1), 28–37.

Clifton, D., Hugueny, S. & Tarassenko, L. (2009*b*), Novelty detection with multivariate extreme value theory, part i: A numerical approach to multimodal estimation, *in* 'Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on', pp. 1 –6.

Clifton, D., Hugueny, S. & Tarassenko, L. (2009*c*), 'Novelty detection with multivariate extreme value theory, part I: Numerical approach for multimodal estimation', *Proceedings of IEEE Workshop on Machine Learning in Signal Processing* .

Clifton, D., Hugueny, S. & Tarassenko, L. (2010), 'Novelty detection with multivariate extreme value statistics', *Journal of Signal Processing Systems* pp. 1–19–19.

Clifton, D., Hugueny, S. & Tarassenko, L. (2011), Pinning the tail on the distribution: A multivariate extension to the generalised pareto distribution, *in* 'Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on', pp. 1–6.

Clifton, L., Clifton, D., Watkinson, P. & Tarassenko, L. (2011), Identification of patient deterioration in vital-sign data using one-class support vector machines, *in* 'Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on', pp. 125–131.

Cohen, M. J., Grossman, A. D., Morabito, D., Knudson, M. M., Butte, A. J. & Manley, G. T. (2010), 'Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis.', *Critical care (London, England)* **14**(1).

Coles, S. G. & Tawn, J. A. (1996), 'A Bayesian analysis of extreme rainfall data', *Applied Statistics* **45**(4), 463–478.

Cooley, D., Nychka, D. & Naveau, P. (2007), 'Bayesian spatial modeling of extreme precipitation return levels', *Journal of the American Statistical Association* **102**(479), 824–840.

Culverhouse, P., Williams, R., Reguera, B., Herry, V. & González-Gil, S. (2003), 'Do experts make mistakes? A comparison of human and machine identification of dinoflagellates', *Marine Ecology Progress Series* **247**, 17–25.

De Haan, L. & de Ronde, J. (1998), 'Sea and wind: multivariate extremes at work', *Extremes* **1**(1), 7–45.

De Haan, L. & Ferreira, A. (2006), *Extreme value theory: an introduction*, Springer Verlag.

De Haan, L. & Resnick, S. I. (1993), 'Estimating the limit distribution of multivariate extremes', *Stochastic Models* **9**(2), 275–309.

De Jong, P. & Penzer, J. (1998), 'Diagnosing shocks in time series', *Journal of the American Statistical Association* **93**(442), 796–806.

De Oliveira, J. T. (1965), 'Statistical decision for bivariate extremes', *Portugal. Math* .

DeMeulenaere, S. (2007), 'Pulse oximetry: uses and limitations', *The Journal for Nurse Practitioners* **3**(5), 312 – 317.

Devita, M. A., Braithwaite, R. S., Mahidhara, R., Stuart, S., Foraida, M. & Simmons, R. L. (2004), 'Use of medical emergency team responses to reduce hospital cardiopulmonary arrests', *Quality and Safety in Health Care* **13**(4), 251.

Diehl, C. P. & Hampshire, J. B. I. (2002), 'Real-time object classification and novelty detection for collaborative video surveillance', *Proceedings of the 2002 International Joint Conference on Neural Networks IJCNN02 Cat No02CH37290* **3**, 2620–2625.

Duda, R. O., Hart, P. E. & Stork, D. G. (2012), *Pattern classification*, John Wiley & Sons.

Dupuis, D. J. & Morgenthaler, S. (2002), 'Robust weighted likelihood estimators with an application to bivariate extreme value problems', *Canadian Journal of Statistics* **30**(1), 17–36.

Dutta, H., Giannella, C., Borne, K. & Kargupta, H. (2007), Distributed top-k outlier detection from astronomy catalogs using the demac system.

Einmahl, J. H. J. & Magnus, J. R. (2008), 'Records in athletics through extreme-value theory', *Journal of the American Statistical Association* **103**(484), 1382–1391.

Einmahl, J. H. J. & Smeets, S. (2009), 'Ultimate 100m world records through extreme-value theory', *papers.ssrn.com* .

Einmahl, J. H. J. & Smeets, S. G. W. R. (2011), 'Ultimate 100-m world records through extreme-value theory', *Statistica Neerlandica* **65**(1), 32–42.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Springer.

Embrechts, P., Lambrigger, D. D. & Wüthrich, M. V. (2009), 'Multivariate extremes and the aggregation of dependent risks: examples and couter-examples', *Extremes* **24**(12), 107–127.

Endres, D. M. & Schindelin, J. E. (2003), 'A new metric for probability distributions', *IEEE Transactions on Information theory* **49**(7), 1858–1860.

Escalante, H. J. (2005), 'A comparison of outlier detection algorithms for machine learning'.

Falk, M. (2007), 'It was 30 years ago today when Laurens de Haan went the multivariate way', *Extremes* **11**(1), 55–80.

Fasen, V., Klüppelberg, C. & Schlather, M. (2008), 'High-level dependence in time series models', *Extremes* pp. 1–33.

Field, C. & Smith, B. (2011), 'Robust estimation: a weighted maximum likelihood approach', *International Statistical Review* **62**(3), 405–424.

Filippone, M. & Sanguinetti, G. (2010), 'Information theoretic novelty detection', *Pattern Recognition* .

Findlay, G., Shotton, H. ad Kelly, K. & Mason, M. (2012), 'Cardiac arrest procedures: Time to Intervene?', *National Confidential Inquiry into Patient Outcome and Death* .

Fisher, R. & Tippett, L. (1928), 'Limiting forms of the frequency distributions of the largest or smallest members of a sample', *Proc. Camb. Philos. Soc.* **24**, 180–190.

Folke, M., Cernerud, L., Ekstrm, M. & Hk, B. (2003), 'Critical review of non-invasive respiratory monitoring in medical care', *Medical and Biological Engineering and Computing* **41**, 377–383.

Galbraith, J. & Zernov, S. (2009), 'Extreme dependence in the NASDAQ and S&P 500 composite indexes', *Applied Financial Economics* **19**(13), 1019–1028.

Gardner, A. B., Krieger, A. M., Vachtsevanos, G. & Litt, B. (2006), 'One-class novelty detection for seizure analysis from intracranial eeg', *Journal of Machine Learning Research* **7**, 1025–1044.

Geddes, L., Voelz, M., Combs, C., Reiner, D. & Babbs, C. (1982), 'Characterization of the oscillometric method for measuring indirect blood pressure', *Annals of Biomedical Engineering* **10**, 271–280. 10.1007/BF02367308.

Gray, R. M. & Davisson, L. D. (2004), *An introduction to statistical signal processing*, Vol. 1, Cambridge Univ Pr.

Grünwald, P. D. (2007), *The minimum description length principle*, MIT press.

Gumbel, E. J. (1935), Les valeurs extrêmes des distributions statistiques, *in* 'Annales Henri Poincare', Vol. 4, pp. 115–158.

Hanley, J. A., McNeil, B. J. et al. (1983), 'A method of comparing the areas under receiver operating characteristic curves derived from the same cases', *Radiology* **148**(3), 839–843.

Hann, A. (2008), Multi-parameter Monitoring for Early Warning of Patient Detoriation, DPhil, University of Oxford.

Harlow, D. G., Smith, R. L. & Taylor, H. M. (1983), 'Lower tail analysis of the distribution of the strength of load-sharing systems', *Journal of Applied Probability* **20**(2), 358–367.

Harrison, G. A., Jacques, T., McLaws, M.-L. & Kilborn, G. (2006), 'Combinations of early signs of critical illness predict in-hospital death–the soccer study (signs of critical conditions and emergency responses)', *Resuscitation* **71**(3), 327 – 334.

Hart, D. K., Rutherford, S. E. & Wickham, A. H. S. (1985), 'Structural reliability analysis of stiffened panels', *RINA Transactions* **128**, 293–310.

Hautamaki, V., Karkkainen, I. & Franti, P. (2004), 'Outlier detection using k-nearest neighbour graph', *Proceedings of the 17th International Conference on Pattern Recognition 2004 ICPR 2004* **3**, 4–7.

Heffernan, J. E. & Tawn, J. A. (2004), 'A conditional approach for multivariate extreme values', *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **66**(3), 497–546.

Heldt, T., Long, B., Verghese, G. C., Szolovits, P. & Mark, R. G. (2006), 'Integrating data, models, and reasoning in critical care.', *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* **1**, 350–3.

Hero III, A. O. (n.d.), 'Geometric entropy minimization (GEM) for anomaly detection and localization', *Ann Arbor* **1001**, 48109–2122.

Hill, E. & Stoneham, M. (2000), 'Pratical applications of pulse oximetry', *Practical Procedures* (11).

Hillman, K. M., Chen, J., Cretikos, M., Bellomo, R., Brown, D., Doig, G., Finfer, S. & Flabouris, A. (2005), 'Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial', *The Lancet* **365**(9477), 2091–2097.

Hodge, V. & Austin, J. (2004), 'A survey of outlier detection methodologies', *Artificial Intelligence Review* **22**(2), 85–126.

Hodgetts, T. J., Kenward, G., Vlackonikolis, I., Payne, S., Castle, N., Crouch, R., Ineson, N. & Shaikh, L. (2002), 'Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital', *Resuscitation* **54**(2), 115–123.

Hogue, C. W., Stearns, J. D., Colantuoni, E., Robinson, K., Stierer, T., Mitter, N., Pronovost, P. J. & Needham, D. M. (2009), 'The impact of obesity on outcomes after critical illness: a meta-analysis.', *Intensive care medicine* **35**(7), 1152–70.

Hourihan, F., Bishop, G., Hillman, K. M. & Daffurn, K. (1995), 'The medical emergency team: a new strategy to identify and intervene in high-risk patients', *Clinical Intensive Care* **6**(6), 269–272.

Hravnak, M. (n.d.), Ability of Electronic Integrated Monitoring System to Impact Duration of Patient Instability on a Step Down Unit.

Hravnak, M., Devita, M. A., Clontz, A., Edwards, L., Valenta, C. & Pinsky, M. R. (2011), 'Cardiorespiratory instability before and after implementing an integrated monitoring system.', *Crit Care Med* **39**(1), 65–72.

Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M. A. & Pinsky, M. R. (2008), 'Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system.', *Archives of internal medicine* **168**(12), 1300–8.

Hugueny, S., Clifton, D. A. & Tarassenko, L. (2009*a*), 'Extreme Value Theory for Novelty Detection in Vital Sign Monitoring. Technical report'.

Hugueny, S., Clifton, D. & Tarassenko, L. (2009*b*), Novelty detection with multivariate extreme value theory, part ii: An analytical approach to unimodal estimation, *in* 'Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on', pp. 1–6.

Hyndman, R. J. (1996), 'Computing and Graphing Highest Density Regions.', *The American Statistician* **50**(2).

Iyengar, A., Baxter, A. & Forster, A. (2009), 'Using Medical Emergency Teams to detect preventable adverse events', *Critical Care* **13**(4).

Jankovic, S. & Ostrogorski, T. (2000), 'Differences of slowly varying functions', *Journal of Mathematical Analysis and Applications* **247**(2), 478 – 488.

Jankovic, S. & Ostrogorski, T. (2002), 'Convex additively slowly varying functions', *Journal of Mathematical Analysis and Applications* **274**(1), 228 – 238.

Jones, D., Egi, M., Bellomo, R. & Goldsmith, D. (2007), 'Effect of the medical emergency team on long-term mortality following major surgery', *Critical Care* **11**(1).

Jyothsna, V., V Rama Prasad, V. & Munivara Prasad, K. (2011), 'A Review of Anomaly based Intrusion Detection Systems', *International Journal of Computer Applications* **28**, 26–35.

Kallstrom, T. J. (2002), 'AARC Guideline: oxygen therapy for adults in the acute care facility.', *American Association for Respiratory Care Clinical Practice Guidelines* .

Karamata, J. (1933), 'Sur un mode de croissance régulière. théorèmes fondamentaux.', *Bulletin de la Société Mathématique de France* **61**, 55–62.

Katz, R. (1999), 'Extreme value theory for precipitation: sensitivity analysis for climate change', *Advances in Water Resources* **23**(2), 133–139.

Kbar, F. A. & Campbell, I. A. (2006), 'Oxygen therapy in hospitalized patients: the impact of local guidelines', *Journal of Evaluation in Clinical Practice* **12**(1), 31–36.

Kim, H. (1996), 'On a Constrained Optimal Rule for Classification with Unknown Prior Individual Group Membership', *Journal of Multivariate Analysis* **59**(2), 166–186.

Knorr, E. M. & Ng, R. T. (1998), Algorithms for Mining Distance-Based Outliers in Large Datasets, *in* 'Proceedings of the 24th VLDB Conference', Vol. 98, pp. 392–403.

Komaki, F. (2006), 'Shrinkage priors for bayesian prediction', *the Annals of Statistics* pp. 808–819.

Kong, M. H., Shaw, L. K., O'Connor, C., Califf, R. M., Blazing, M. A. & Al-Khatib, S. M. (2010), 'Is rhythm-control superior to rate-control in patients with atrial fibrillation and diastolic heart failure?', *Annals of noninvasive electrocardiology : the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc* **15**(3), 209–17.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Lee, H.-j. & Roberts, S. J. (2008), On-line novelty detection using the Kalman filter and extreme value theory, *in* 'Pattern Recognition, 2008. ICPR 2008. 19th International Conference on', number 1, pp. 1–4.

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P. & Zhang, H. (2001), 'An information-based sequence distance and its application to whole mitochondrial genome phylogeny.', *Bioinformatics* **17**(2), 149–154.

Li, M., Chen, X., Li, X., Ma, B. & Vitanyi, P. M. B. (2004), 'The Similarity Metric', *IEEE Transactions on Information Theory* **50**(12), 3250–3264.

Li, Y. (2008), A surface representation approach for novelty detection, *in* 'Information and Automation, 2008. ICIA 2008. International Conference on', pp. 1464–1468.

Lin, J. (1991), 'Divergence measures based on the Shannon entropy', *IEEE Transactions on Information Theory* **37**(1), 145–151.

Liu, H., Shah, S. & Jiang, W. (2004), 'On-line outlier detection and data cleaning', *Computers and Chemical Engineering* **28**(9), 1635–1647.

Longin, F. & Solnik, B. (2001), 'Extreme Correlation of International Equity Markets', *The Journal of Finance* **56**(2), 649–676.

Lowe, D. & Tipping, M. E. (1997), 'Neuroscale: novel topographic feature extraction using RBF networks', *Advances in Neural Information Processing Systems* pp. 543–549.

Magnussen, H., Kirsten, A.-M., Khler, D., Morr, H., Sitter, H. & Worth, H. (2008), 'Guidelines for long-term oxygen therapy: German society for pneumology and respiratory medicine [leitlinien zur langzeit-sauerstofftherapie: Deutsche gesellschaft fr pneumologie und beatmungsmedizin e. v.]', *Pneumologie* **62**(12), 748–756.

Mahalanobis, P. C. (1936), On the generalized distance in statistics, *in* 'Proceedings of the National Institute of Science, Calcutta', Vol. 12, p. 49.

Manson, G., Worden, K. & Allman, D. (2003), 'Experimental validation of a structural health monitoring methodology: part III. Damage location on an aircraft wing', *Journal of Sound and Vibration* **259**(2), 365–385.

Markou, M. & Singh, S. (2003*a*), 'Novelty detection: a review, part 1: statistical approaches', *Signal Processing* **83**(12), 2481–2497.

Markou, M. & Singh, S. (2003*b*), 'Novelty detection: a review, part 2: neural network based approaches', *Signal Processing* **83**(12), 2499–2521.

Markou, M. & Singh, S. (2006), 'A Neural Network-Based Novelty Detector for Image Sequence Analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**, 1664–1677.

Marshall, A. W. & Olkin, I. (1983), 'Domains of attraction of multivariate extreme value distributions', *The Annals of Probability* **11**(1), 168–177.

Mayaud, L., Lai, P. S., Clifford, G. D., Tarassenko, L., Celi, L. A. & Annane, D. (2013), 'Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension*', *Critical care medicine* **41**(4), 954–962.

McGloin, H., Adam, S. K. & Singer, M. (1999), 'Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable?', *J R Coll Physicians Lond* **33**(3), 255–259.

McQuillan, P., Pilkington, S., Allan, A., Taylor, B., Short, A., Morgan, G., Nielsen, M., Barrett, D., Smith, G. & Collins, C. H. (1998), 'Confidential inquiry into quality of care before admission to intensive care', *BMJ British Medical Journal* **316**, 1853–1858.

Meerschaert, M. M. & Stoev, S. A. (2009), 'Extremal limit theorems for observations separated by random power law waiting times', *Journal of Statistical Planning and Inference* **139**(7), 2175–2188.

Miljkovic, D. (2010), Review of novelty detection methods, *in* 'MIPRO, 2010 Proceedings of the 33rd International Convention', IEEE, pp. 593–598.

Nabney, I. T. (2002), *NETLAB: algorithms for pattern recognition*, Springer Verlag.

Neamatullah, I., Douglass, M. M., Lehman, L.-w. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G. & Clifford, G. D. (2008), 'Automated de-identification of free-text medical records.', *BMC medical informatics and decision making* **8**, 32.

NICE Short Clinical Guidelines Technical Team (2006), 'Acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital. Technical report'.

O'Driscoll, B. R., Howard, L. S. & Davison, A. G. (2008), 'Bts guideline for emergency oxygen use in adult patients', *Thorax* **63**(Suppl 6), vi1–vi68.

Österreicher, F. & Vajda, I. (2003), 'A new class of metric divergences on probability spaces and its applicability in statistics', *Annals of the Institute of Statistical Mathematics* **55**(3), 639–653.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The annals of mathematical statistics* **33**(3), 1065–1076.

Patcha, A. & Park, J. (2007), 'An overview of anomaly detection techniques: Existing solutions and latest technological trends', *Computer Networks* **51**, 3448–3470.

Pickands, J. (1975), 'Statistical Inference Using Extreme Order Statistics', *Annals of Statistics* **3**, 119–131.

Poon, S. H., Rockinger, M. & Tawn, J. A. (2004), 'Extreme value dependence in financial markets: Diagnostics, models, and financial implications', *Review of Financial Studies* **17**(2), 581.

Pursell, L. E. (1968), 'Introduction to Real Analysis.', *The American Mathematical Monthly* **75**(8), 922.

Quinn, J. A. & Williams, C. K. I. (2007), 'Known unknowns: Novelty detection in condition monitoring', *Pattern Recognition and Image Analysis* pp. 1–6.

Quinn, J. A., Williams, C. K. I. & Mcintosh, N. (2009), 'Factorial Switching Linear Dynamical Systems Applied to Physiological Condition', *IEEE Transactions on Pattern Analysis* **31**(9), 1537–1551.

Riedi, R. H. (2006), 'An Introduction to Statistical Signal Processing ', *Journal of the American Statistical Association* **101**(475), 1317.

Roberts, S. J. (1999), 'Novelty detection using extreme value statistics', *IEE Proceedings-Vision, Image and Signal Processing* **146**(3), 124–129.

Roberts, S. J. (2000), 'Extreme value statistics for novelty detection in biomedical data processing', *Science, Measurement and Technology, IEE Proceedings* - **147**(6), 363 –367.

Roberts, S. & Tarassenko, L. (1994), 'A Probabilistic Resource Allocating Network for Novelty Detection', *Neural Computation* **6**(2), 270–284.

Sainfort, F. (1991), 'Evaluation of Medical Technologies: A Generalized ROC Analysis', *Medical Decision Making* **11**(3), 208.

Sammon Jr, J. W. (1969), 'A nonlinear mapping for data structure analysis', *IEEE Transactions on computers* **100**(18), 401–409.

Savran, Y. & Gunsel, B. (2010), 'Novelty Detection on Metallic Surfaces by GMM Learning in Gabor Space', pp. 325–334.

Schein, R. M., Hazday, N., Pena, M., Ruben, B. H. & Sprung, C. L. (1990), 'Clinical antecedents to in-hospital cardiopulmonary arrest', *Chest* **98**(6), 1388.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J. & Platt, J. (2000), 'Support vector method for novelty detection', *Proceedings of NIPS* pp. 582–588.

Scott, D. W. (1992), *Multivariate density estimation: theory, practice, and visualization*, Wiley-Interscience.

Smith, A. & Wood, J. (1998), 'Can some hospital cardio-repiratory arrests be prevented? a prospective study', *Resuscitation* **37**, 133–137.

Sofman, B., Neuman, B., Stentz, A. & Bagnell, J. A. (2011), 'Anytime online novelty and change detection for mobile robots', *Journal of Field Robotics* **28**(4), 589–618.

Sohn, H., Allen, D. W., Worden, K. & Farrar, C. R. (2005), 'Structural damage classification using extreme value statistics', *Journal of Dynamic Systems, Measurement, and Control* **127**(1), 125–132.

Spearpoint, K. G., Gruber, P. C. & Brett, S. J. (2009), 'Impact of the Immediate Life Support course on the incidence and outcome of in-hospital cardiac arrest calls: an observational study over 6 years.', *Resuscitation* **80**(6), 638–43.

Subbe, C. P., Kruger, M., Rutherford, P. & Gemmel, L. (2001), 'Validation of a modified Early Warning Score in medical admissions', *QJM* **94**(10), 521.

Surace, C. & Worden, K. (2010), 'Novelty detection in a changing environment: A negative selection approach', *Mechanical Systems and Signal Processing* **24**(4), 1114 – 1128.

Takano, C. & Ohta, Y. (2007), 'Heart rate measurement based on a time-lapse image.', *Medical engineering & physics* **29**(8), 853–7.

Takeuchi, J. . & Yamanishi, K. (2006), 'A unifying framework for detecting outliers and change points from time series', *IEEE Transactions on Knowledge and Data Engineering* **18**(4), 482–492.

Tarassenko, L., Clifton, D. A., Bannister, P. R., King, S. & King, D. (2009), 'Novelty Detection', *Encyclopedia of Structural Health Monitoring* .

Tarassenko, L., Hann, A., Patterson, A., Braithwaite, E., Davidson, K., Barber, V. & Young, D. (2005), Biosign: Multi-parameter monitoring for early warning of patient deterioration, *in* 'Medical Applications of Signal Processing, 2005. The 3rd IEE International Seminar on (Ref. No. 2005-1119)', pp. 71–76.

Tarassenko, L., Hann, A. & Young, D. (2006), 'Integrated monitoring and analysis for early warning of patient deterioration', *British Journal of Anaesthesia* **97**(1), 64.

Tarassenko, L., Hayton, P., Cerneaz, N. & Brady, M. (1995), Novelty detection for the identification of masses in mammograms, *in* 'Artificial Neural Networks, 1995., Fourth International Conference on', pp. 442–447.

Tarassenko, L., Nairac, Townsend, N., Buxton, I. & Cowley, P. (2000), 'Novelty detection for the identification of abnormalities', *International Journal of Systems Science* **31**(11), 1427–1439.

Tawn, J. A. (1988), 'Bivariate extreme value theory: models and estimation', *Biometrika* **75**(3), 397.

Tawn, J. A. (1990), 'Modelling multivariate extreme value distributions', *Biometrika* **77**(2), 245.

Tax, D. M. J. & Duin, R. P. W. (1999), 'Data domain description using support vectors', *Neural Networks* **256**, 251–256.

Tremblay (2003), 'Impact of Body Mass Index on Outcomes Following Critical Care', *Chest* **123**(4), 1202–1207.

Tremblay, A. & Bandi, V. (2003), 'Impact of body mass index on outcomes following critical care.', *Chest* **123**(4), 1202–7.

Tsay, R. S., Pena, D. & Pankratz, A. E. (2000), 'Outliers in multivariate time series', *Biometrika* **87**(4), pp. 789–804.

Vieira Neto, H. & Nehmzow, U. (2007), 'Real-time Automated Visual Inspection using Mobile Robots', *Journal of Intelligent and Robotic Systems* **49**, 293–307.

Webster, J. G. (1998), *Medical Instrumentation : application and design*, J. Wiley & Sons.

Williams, C. K. I., Quinn, J. A. & Mcintosh, N. (2006), 'Factorial switching Kalman filters for condition monitoring in neonatal intensive care', *Advances in Neural Information Processing Systems* **18**, 1513.

Wong, D. (2011), Identifying Vital Sign Abnormality in Acutely-Ill Patients, DPhil, University of Oxford.

Worden, K., Manson, G. & Allman, D. (2003), 'Experimental validation of a structural health monitoring methodology: Part I. Novelty detection on a laboratory structure', *Journal of Sound and Vibration* **259**(2), 323–343.

Zhang, Y. & Szolovits, P. (2008), 'Patient-specific learning in real time for adaptive monitoring in critical care.', *Journal of biomedical informatics* **41**(3), 452–60.

Zhang, Y. Z. Y., Meratnia, N. & Havinga, P. (2010), 'Outlier Detection Techniques for Wireless Sensor Networks: A Survey', *IEEE Communications Surveys Tutorials* **12**, 159–170.