Extreme Value-based Novelty Detection

Matthys Lucas Steyn

Report presented in partial fulfilment
of the requirements for the degree of
MComm (Mathematical Statistics)
at the University of Stellenbosch

Supervisor: Professor T. de Wet

Degree of confidentiality: C

December 2017

PLAGIARISM DECLARATION

1. Plagiarism is the use of ideas, material and/or intellectual property of another's work and presenting it as my own.

2. I agree that plagiarism is a punishable offence because it constitutes theft.

3. I also understand that direct translations are plagiarism.

4. Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.

5. I declare that the work contained in this assignment, except otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

Student number	**Signature**
Initials and surname	**Date**

Acknowledgements

I hereby wish to acknowledge prof. T. de Wet for supervising my research. Furthermore, I acknowledge the Department of Actuarial Science and Statistics and the Stellenbosch University Library for providing me with the necessary books and journals. Finally, I wish to thank Me. A. Matthee for assisting me with the editing.

Abstract

This dissertation investigates extreme value-based novelty detection. An in-depth review of the theoretical proofs and an analytical investigation of current novelty detection methods are given. It is concluded that the use of extreme value theory for novelty detection leads to superior results.

The first part of this dissertation provides an overview of novelty detection and the various methods available to construct a novelty detection algorithm. Four broad approaches are discussed, with this dissertation focusing on probabilistic novelty detection. A summary of the applications of novelty detection and the properties of an efficient novelty detection algorithm are also provided.

The theory of extremes plays a vital role in this work. Therefore, a comprehensive description of the main theorems and modelling approaches of extreme value theory is given. These results are used to construct various novelty detection algorithms based on extreme value theory.

The first extreme value-based novelty detection algorithm is termed the Winner-Takes-All method. The model's strong theoretical underpinning as well as its disadvantages are discussed. The second method reformulates extreme value theory in terms of extreme probability density. This definition is utilised to derive a closed-form expression of the probability distribution of a Gaussian probability density. It is shown that this distribution is in the minimum domain of attraction of the extremal Weibull distribution.

Two other methods to perform novelty detection with extreme value theory are explored, namely the numerical approach and the approach based on modern extreme value theory. Both these methods approximate the distribution of the extreme probability density values under the assumption of a Gaussian mixture model. In turn, novelty detection can be performed in complex settings using extreme value theory.

To demonstrate an application of the discussed methods a banknote authentication dataset is analysed. It is clearly shown that extreme value-based novelty detection methods are extremely efficient in detecting forged banknotes. This demonstrates the practicality of the different approaches.

The concluding chapter compares the theoretical justification, predictive power and efficiency of the different approaches. Proposals for future research are also discussed.

Opsomming

Hierdie verhandeling ondersoek anomalie-opsporing wat op ekstreemwaardeteorie gegrond is. Die teoretiese bewyse word breedvoerig beskryf en huidige metodes word ontleed. Daar word bevind dat die gebruik van ekstreemwaardeteorie vir anomalie-opsporing tot uitsonderlike resultate lei.

Die eerste deel van die verhandeling bied 'n oorsig van anomalie-opsporing en verskillende metodes wat gebruik kan word om 'n anomalie-opsporingsalgoritme te formuleer. Vier benaderings tot anomalie-opsporing word bespreek. Die verhandeling lê klem op een daarvan, naamlik probabilistiese anomalie-opsporing. Die gedeelte sluit af met 'n opsomming van die praktiese toepassings van anomalie-opsporing en die eienskappe van 'n doeltreffende anomalie-opsporingsalgoritme.

Ekstreemwaardeteorie speel 'n uiters belangrike rol in hierdie werk. Daarom word 'n omvattende beskrywing van die vernaamste grondbeginsels en modelleringsbenaderings tot ekstreemwaardeteorie gegee. Dié resultate word benut om verskeie anomalie-opsporingsalgoritmes te formuleer wat op ekstreemwaardeteorie gegrond is.

Daar word eerstens gekyk na die anomalie-opsporingsalgoritme wat op ekstreemwaardeteorie gegrond is en wat die Wenner-Vat-Alles-metode genoem word. Daar word bewys dat die model teoreties korrek is. In die tweede metode word ekstreemwaardeteorie ten opsigte van ekstreme waarskynlikheidsdigtheid geherdefinieer. Hierdie definisie word gebruik om 'n geslote-vorm uitdrukking van die waarskynlikheidsverdeling van 'n Gaussiese waarskynlikheidsdigtheid af te lei. Gevolglik word daar aangetoon dat hierdie verdeling in die minimum aantrekkingsgebied van die ekstreme Weibull-verdeling val.

Daarna volg 'n oorsig van twee ander metodes wat vir anomalie-opsporing met ekstreemwaardeteorie gebruik kan word, naamlik die numeriese metode en die metode gebaseer op moderne ekstreemwaardeteorie. In albei hierdie metodes word die verdeling van die ekstreme waarskynlikheidsdigtheidwaardes op die veronderstelling van 'n Gaussiese mengselmodel gegrond. Anomalie-opsporing kan dus in komplekse omgewings uitgevoer word deur ekstreemwaardeteorie te gebruik.

Om te demonstreer hoe hierdie metodes prakties toegepas kan word, word 'n datastel vir banknoot-verifikasie ontleed. Daar word duidelik aangetoon dat anomalie-opsporing wat op ekstreemwaardeteorie gegrond is uiters doeltreffend is om vervalste banknote uit te ken. Dit beklemtoon die praktiese toepassing van die verskillende benaderings.

Die laaste hoofstuk vergelyk die teoretiese regverdiging, voorspellingskrag en doeltreffendheid van die verskillende benaderings. Voorstelle vir toekomstige navorsing word ook bespreek.

Table of contents

CHAPTER 1: INTRODUCTION

CHAPTER 2: REVIEW OF NOVELTY DETECTION

CHAPTER 3: A REVIEW OF EXTREME VALUE THEORY

CHAPTER 6: PRACTICAL APPLICATION OF EXTREME VALUE-BASED NOVELTY DETECTION

CHAPTER 7: CONCLUSIONS AND FUTURE RESEARCH

List of tables

List of figures

List of abbreviations and/or acronyms

AIC	Akaike information criterion
AGD	Asymptotic Gaussianity in Density (assumption)
BIC	Bayesian information criterion
EM	expectation-maximisation (algorithm)
EVI	extreme value index
GEV	generalised extreme value (distribution)
GMM	Gaussian mixture model
GP	generalised Pareto (distribution)
iid	independent and identically distributed (random variables)
KNN	K-nearest neighbour
PCA	principal component analysis
POT	peaks-over-threshold (method)
PWMs	probability-weighted moments
QQ	quantile-quantile (plots)
r.v.	regularly varying (function)
s.v.	slowly varying (function)
SVDD	support vector domain description (method)
SVM	support vector machine (SVM-1) (algorithm)
WTA	winner-takes-all (method)

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Novelty detection is a method used to detect when new data differs to some extent from what is expected to be normal. Conventionally, classification is performed via a supervised approach. This approach assumes that all the classes under investigation are well-sampled. A classifier is constructed to classify a new observation to the class that has the maximum posterior probability, given the data and prior beliefs. However, if one or more of the classes are severely under-sampled, it is not possible to accurately estimate the probability distribution of those classes. For this reason, a novelty detection approach must be considered. One-class classification ultimately finds an accurate estimate of the probability distribution of the class that is sampled sufficiently. This class is termed the normal class. New data is classified as belonging to the normal class or as being novel in terms of the class of normality.

In general, novelty detection is the only solution in high-integrity systems. Such systems refer to scenarios where deviations from the normal class may have catastrophic impacts. For example, one major concern for banks is credit card fraud. However, it is difficult – if not impossible – for a bank to obtain a good sample of fraudulent credit card transactions. A supervised approach will fail to discriminate between legitimate and fraudulent transactions. Alternatively, a model based on legitimate credit card transactions and the personal or demographic information of the account holder can be constructed to represent normal transactions. Thereafter, new transactions can be tested against this model to determine whether they are legitimate or fraudulent. Other examples of high-integrity systems include jet-engine monitoring, banknote authentication and cybersecurity.

Once a model representing the normal class has been constructed, a threshold must be selected to define the decision boundary. Recently, extreme value theory has been proposed as an efficient and theoretically grounded approach to threshold the model of normality.

Extreme value theory is a field of statistics used to model rare or extreme events. Intuitively, extreme value theory is well-suited for novelty detection because it is believed that novel events are extreme in terms of the system under normal observation. This dissertation investigates different methods of constructing a novelty detection algorithm based on extreme value theory.

1.2 RESEARCH OBJECTIVES AND BENEFITS OF THE STUDY

The literature on novelty detection is usually found in the fields of computer science and engineering. One of the objectives of this dissertation is to introduce statisticians to literature from other broad fields of research. Statisticians can benefit from this by being introduced to innovative ways of thinking about a problem. Furthermore, researchers in the computer science or engineering fields can benefit from statisticians improving the theoretical understanding of these methods.

The main research objective of this dissertation is to give an in-depth account of the use of extreme value theory for novelty detection. This class of models has not been described from a mathematical statistical point of view. It will be motivated why extreme value theory is well-suited for novelty detection. Moreover, the theoretical justification and practicality of this class of models will be investigated. The results found in this dissertation should then indicate whether extreme value theory is a powerful tool for novelty detection.

Using extreme value theory to perform novelty detection has only recently been proposed. Additionally, not much research on high-dimensional or multimodal novelty detection has been done. Hence, there is a need to discuss these methods in a principled manner, allowing for future research to be undertaken on this class of models. Therefore, the advantages and disadvantages of current methods are investigated and viable solutions are discussed. New research to improve the shortcomings of current methods can then be conducted.

Numerous algorithms have been proposed to perform classification. These algorithms are extremely powerful if the main assumptions of the model are satisfied. However, in modern times it is likely to encounter datasets with class imbalance. In such cases, supervised algorithms cannot accurately model the probability distribution of the under-sampled class. Novelty detection will then prove to be a valuable alternative. It is never the case that one method is superior to all other approaches. Therefore, it is important to be comfortable with several ways of building a model for discrimination. The reason for this is generally the complexity or form of the data that governs the optimal approach.

1.3 LITERATURE REVIEW

Two broad research areas are covered in this dissertation, namely extreme value theory and novelty detection. The problem considered throughout pertains to novelty detection. Once the probability density of the normal class has been estimated, extreme value theory can be utilised to threshold this estimated probability density function.

The model used in this dissertation to estimate the probability density function of the normal class is the Gaussian mixture model. The book *Multivariate Density Estimation* by Scott (2015) gives an in-depth analysis of density estimation. Specifically, this book contains results on the transformations of multivariate Gaussian distributions. These results are used in Chapter 5. Another important reference on statistical modelling is *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman (2009). In this book, supervised and unsupervised learning methods are discussed. Specifically, Chapter 8 of this book describes the expectation-maximisation (EM) algorithm which is traditionally used to fit a Gaussian mixture model.

Extreme value theory provides the required methods to model the tails of distributions – extreme observations. The book *An Introduction to Statistical Modelling of Extreme Values* by Coles (2000) serves as a good introductory text for extreme value theory. This book provides the theorems and approaches generally used in extreme value theory. Extreme value theory is also explored in the book *Statistics of Extremes: Theory and Applications* by Beirlant, Goegebeur, Segers and Teugels (2004). Both the classical and modern approaches of extreme value theory are discussed. Furthermore, sketches of the proofs of the two main theorems used, namely the Fisher-Tippett and Pickands-Balkema-de Haan theorems, are presented. The book also covers the results on regular variation, univariate and multivariate extreme value theory, and extreme value theory for time series data.

Increased attention has been given to novelty detection in recent years. The book *Outliers in Statistical Data* by Barnett and Lewis (1994) provides the concepts associated with outliers. This theory is closely related to that of novelty detection. Outlier detection for univariate and multivariate data and regression models is considered in this book. *Learning with Kernels: Support Vector Machines, Regularisation, Optimisation and Beyond* by Schölkopf and Smola (2002) discusses various kernel-based methods for statistical modelling. In this book, the one-class support vector machine is defined. This is a powerful domain-based novelty detection algorithm. Chapter 8 of this book describes single-class problems and novelty detection. A helpful review of anomaly detection is given in *Anomaly detection: A Survey* by Chandola, Banerjee and Kumar (2009). This article covers all aspects of anomaly detection thoroughly. The main aspects and types of anomalies are highlighted and the approaches used in different application areas are discussed. Furthermore, the different techniques to perform anomaly detection are reviewed. A comprehensive review of novelty detection is given in *A review of novelty detection* by Pimentel, Clifton, Clifton and Tarassenko (2014). This article describes novelty detection as four broad approaches. It also provides the advantages and disadvantages of each approach, and references to the most recent methods of novelty detection. Additionally, a section on the practical uses of novelty detection is presented. Many of the topics discussed in Chapter 2 of this dissertation have been extracted from this article.

The two researchers who have contributed proficiently to the literature on extreme value-based novelty detection are Stephen Roberts and David Clifton. Roberts (1999) defined the first novelty detection algorithm relying on extreme value theory. Many of the concepts proposed in his article were used to build more efficient extreme value-based novelty detection algorithms. In the DPhil thesis of Clifton (2009) the method of Roberts (1999) was explored in terms of its usefulness and limitations. Clifton (2009) then redefined the meaning of an extreme observation such that extreme value theory is more suitable for novelty detection. These results were restated in the article *Novelty Detection with Multivariate Extreme Value Statistics* by Clifton, Hugueny and Tarassenko (2011). This article proposed an analytical method and a numerical method to perform novelty detection with extreme value theory. Most of the results discussed in Chapters 4 and 5 of this dissertation are found in these articles on extreme value-based novelty detection.

1.4 CHAPTER OUTLINE

Chapter 2 explores novelty detection. The basic terminology and definitions in the field of novelty detection are given. It also explains why novelty detection is approached as a one-class classification problem. This chapter covers the four general approaches to perform novelty detection, namely the probabilistic, distance-based, reconstruction-based and domain-based approaches. The method and the advantages and disadvantages of each of these approaches are discussed. Next, an outline is given of the properties of an efficient novelty detection algorithm. Chapter 2 is concluded with an overview of the practical applications of novelty detection.

An overview of extreme value theory is given in Chapter 3. The two main approaches of extreme value theory – the classical approach and the modern approach – are explored. The problem statements of both these methods are discussed. It is also demonstrated how the resulting limiting distributions are estimated and validated from a finite sample. This chapter is concluded with a section that relates novelty detection to extreme value theory. Chapter 3 serves as a review of extreme value theory so that these results can be used in Chapters 4 and 5.

In Chapter 4 extreme value-based novelty detection is considered. The chapter starts by highlighting when conventional threshold methods fail to accurately threshold the distribution of normality. Consequently, extreme value theory is proposed as an alternative method to threshold the distribution of the normal class. It is argued that extreme value theory overcomes the disadvantages of conditional methods used to threshold the distribution of the system under normal behaviour. However, traditional extreme value theory has some limitations which, as a standalone approach, makes it unsuitable for novelty detection. Next, the limitations of traditional extreme value theory for novelty detection are discussed. Reflecting on these shortcomings, a first extreme value-based novelty detection algorithm is proposed. This model is shown to hold analytically under the appropriate assumptions. The chapter is concluded with the limitations of this extreme value-based novelty detection algorithm.

Chapter 5 considers recent advances in novelty detection based on extreme value theory. Extreme value theory is redefined in terms of minimum probability density. This definition of extreme value theory reduces multivariate problems to an equivalent univariate case. Hence, this definition can be utilised to perform novelty detection in complex scenarios. The first case considered is the multivariate Gaussian case. It is shown that a closed-form expression exists for the distribution of the probability density of a multivariate Gaussian distribution. This expression is then used to prove that the distribution of the density function is in the minimal domain of attraction of the Weibull class of generalised extreme value distributions. However, this approach is constrained by the assumption that the distribution describing the normal class is multivariate Gaussian. Therefore, a numerical approach for Gaussian mixture models is also discussed. The theoretical underpinning of this method and its application in complex settings are explained. It is concluded that the method is applicable for multivariate and multimodal distributions. However, the computational efficiency of the model is weak. Consequently, advances to speed up the computational time of the method are discussed. The concluding section of this chapter considers the modern approach of extreme value theory for novelty detection. Very little research regarding this method has been done. Thus, a possible method to construct a novelty detection algorithm with modern extreme value theory is discussed. Both the multivariate Gaussian distribution and Gaussian mixture model are considered for this approach.

A practical application of the methods discussed in Chapter 5 is given in Chapter 6. A banknote authentication dataset is used for this purpose. Both the classical and modern approaches of extreme value theory are used to detect forged banknotes. The advantage of this methodology is that only real banknotes are needed during the training phase of the model. It is shown that, for this dataset, the application of extreme value theory to perform novelty detection produces highly competitive results.

This dissertation is concluded in Chapter 7. A comparison of all the extreme value-based novelty detection approaches is given. The chapter highlights the disadvantages of each approach and the preference of certain approaches over others. In the conclusion, it is argued that extreme value theory, when used appropriately, leads to superior results when probabilistic novelty detection is performed. Finally, future research areas to improve this class of models are proposed.

1.5 REMARK ON TERMINOLOGY AND NOTATION

The research underpinning this dissertation is mostly extracted from the fields of engineering and computer science. To be consistent, the terminology and notations of these disciplines have therefore been used. However, the definitions and derivations are given strictly from a statistical point of view.

CHAPTER 2

REVIEW OF NOVELTY DETECTION

2.1 INTRODUCTION

Novelty detection is an approach used to detect whether new observations differ significantly from the estimated probability generating mechanism. Generally, a model is fitted to training data. This model represents some normal class. Thereafter, new data containing examples from both the normal class and the novel classes are classified as normal or novel using the estimated model.

There are slight differences between novelty detection, anomaly detection and outlier detection. Barnett and Lewis (1994) defined outliers as observations that are not consistent with the other observations in the sample. These unwanted observations may result from a different probability distribution or just be the extreme observations of the underlying class. Consequently, outliers can be detected and better coped with when a model is built to describe the normal class. Similarly, anomaly detection can be defined as detecting irregularities in the sample. It is believed that the anomalous observations that do not conform to the expected, normal behaviour distort the results. Generally, these observations are removed from the sample during training (Chandola, Banerjee & Kumar, 2009). Novelty detection also tries to identify observations that do not resemble the normal class. However, instead of removing these observations, the novel events are added to a test set. The model is then used to discriminate between normal and novel data. Although the ultimate goals of these three problems might differ, they are used interchangeably in the literature. This is because the same methods are generally used for all domains.

This chapter describes the fundamental concepts of novelty detection. A definition and the main problem of novelty detection are given. It is explained how and why novelty detection can be viewed as a one-class classification problem. Next, an overview is given of the most general approaches to perform novelty detection and the properties that an efficient novelty detection algorithm should have. The chapter is concluded with practical applications of novelty detection.

2.2 DEFINITION AND BASIC CONCEPTS

This dissertation defines novelty detection as the procedure to detect events that differ in some manner from the expected behaviour. In order to detect novel events, some measure of similarity between the training sample and new observations is required. Therefore, a general approach is to build a model that describes the expected behaviour. This class of observations is referred to as the normal or positive class. New observations are tested against this model to produce a novelty score. Finally, each new observation is classified as normal or novel based on the novelty scores (Pimentel, Clifton, Clifton & Tarassenko, 2014).

Consider the random variable Y termed the response or dependent variable. Furthermore, let $\underline{X}^T = \left(X_1, X_2, \ldots, X_d\right)$ be the d-dimensional vector of predictor or independent variables. The response variable is coded as, for example, a binary variable such $Y = 1$ if an observation is from the normal class and $Y = 0$ if the observation is from some other class. Novelty detection attempts to train a model on the normal class – predictor variables for which the response is 1. This produces a novelty score $z\left(\underline{X}\right)$. These novelty scores are then compared to some threshold t. High novelty scores generally indicate that the observation is abnormal (Pimentel *et al*., 2014). Consequently, if the novelty score produced by the predictor variable is below the specified threshold, the response is labelled as belonging to the normal class. Hence, the d-dimensional surface $z\left(\underline{X}\right) = t$ represents the decision boundary between the normal class and novel observations.

Notice that only the data that represents the normal or positive class is used to ultimately discriminate between normal and novel observations. Furthermore, different types of novelties arise from a variety of problems. For example, novelty detection has been used to detect network intrusion. A model is built for the normal network features. Any anomalous activity, relative to this model, can therefore be flagged as an intrusion. On the other hand, a novel observation could be a new class not seen at training. The formation or disappearing of classes seen at training is known as concept drift in the computer science literature. It refers to the fact that, over lengthy periods of time, new classes may appear or some may disappear (Chen & Liu, 2016).

2.3 NOVELTY DETECTION AND ONE-CLASS CLASSIFICATION

As mentioned in Section 2.2, a general approach for novelty detection is to build a model based on the positive class and test new observations against this model. However, from a supervised learning perspective one would consider the normal data as well as the novel data. Consequently, it is a binary classification problem where a model is built using examples from both the normal and the novel class. Many algorithms have been proposed for this problem – see Hastie, Tibshirani and Friedman (2009). Unfortunately, these methods rely heavily on the assumption that all the classes are well sampled. In cases where all the classes are well sampled, supervised classification algorithms are extremely powerful.

However, this assumption breaks down in some situations. If any of the classes are significantly under-sampled this assumption breaks down. More worryingly, it is often the under-sampled – if observed at all – class that has vital consequences in the real world. For example, the goal might be to detect fraud at an insurance company. There may be very few or no observations which are labelled as fraudulent claims. Supervised models also break down if the number of possible classes is specified incorrectly (Hugueny, 2013). There are distinct reasons why the number of classes may be estimated wrongly. It might be that the training data is incomplete. Hence, the analyst has too little information to know that there is another class. New classes may also form over time; supervised models are not built to handle such changes. Observations belonging to an unseen class are mistakenly classified into one of the classes used to train the model.

As a result of the shortcomings of supervised classification algorithms an alternative approach must be used. In general, problems involving novelty detection usually have a very well-sampled positive class. However, observations from the novel classes might be difficult to obtain. These anomalies might be difficult to obtain due to high measurement cost or the infrequent appearance of novel classes (He & Garcia, 2009). This problem is worsened by the fact that observations from the novel class generally have significant variability. Consequently, the credibility of the novel samples limits the use of these observations for discrimination (Lee & Cho, 2006). Therefore, novelty detection is tackled as a one-class classification problem (Moya, Koch & Hostetler, 1993). This means a model is built on the positive data to represent the normal class. In turn, new observations are tested against this model. Hence, there is a positive class of interest and other novel classes which are only classified as not being in the same class of the normal model.

Lee and Cho (2006) mentioned that abnormal observations can also be used for one-class classification. Thus, a one-class classifier is still constructed but the model considers normal as well as novel observations. Although this approach might distort the results if one of the classes is highly underrepresented, it may improve the predictive power significantly if a class is only moderately underrepresented. Furthermore, as the class imbalance reduces, the predictive power of a novelty detection algorithm using samples from the abnormal class as well as the normal class improves.

In terms of model validation, it must be mentioned that the misclassification error does not give an accurate estimate of model performance if the novel class is highly under-sampled. Consider a test set where 99% of the data belongs to the normal class. If a model were to predict that all observations belong to the normal class, the test error (using the misclassification error) would be 1%. However, not a single novelty would have been detected by the model. Due to the class imbalance inherent to novelty detection, the performance of the model must be measured by also considering errors regarding normal observations and errors regarding novel observations separately. Visualisations such as ROC curves are also useful.

As a result of the obstacles encountered by a supervised classification model, various approaches for one-class classification, and specifically novelty detection, have been proposed. The next section introduces the most common methods to construct a novelty detection algorithm.

2.4 APPROACHES TO NOVELTY DETECTION

This section presents an overview of different methods to discriminate between normal and novel observations. As mentioned in Pimentel *et al*. (2014), novelty detection models can be divided into four main categories, namely the probabilistic, distance-based, reconstruction-based and domain-based approaches. These methods are now discussed.

2.4.1 Probabilistic approach to novelty detection

Probabilistic novelty detection assumes that the normal class is generated by some probability distribution F. This approach starts by estimating the probability distribution of the normal data. The estimated distribution is denoted as $\hat{F}$ and represents a model for the positive class. Hence, this distribution should have high density for positive examples and low density for novel observations. A novelty score is obtained by setting a novelty threshold on the density function of the estimated distribution. In turn, a new observation $\underline{x}$ is classified as novel if $\hat{f}\left(\underline{x}\right)<t$, where $\hat{f}$ is the estimated probability density function of the normal class and t is the novelty threshold. The novelty threshold must be set such that most of the positive samples are within the boundary. Hence, t is chosen such that the probability of a normal observation lying interior to $\left\{\underline{x}:f\left(\underline{x}\right)\geq t\right\}$ is large. However, the boundary must not be too wide so that novel events are within the boundary (Hugueny, 2013; Pimentel *et al*., 2014). Hence, a novelty threshold is set using a probabilistic approach to define the normal class.

One of the simplest probabilistic approaches to novelty detection is the Grubbs' test (Grubbs, 1969). This test assumes that the observations are univariate and normally distributed. The distances from the sample mean to each observation are computed and standardised in terms of the sample standard deviation. Usually, if any one of the computed standardised distances is greater than 3, it is considered an outlier. The Grubbs' test has some disadvantages. It assumes a normal distribution which might be restrictive and it only tests one observation at a time. Nevertheless, it is a very simple test to understand and might be useful to identify possible outliers. These possible outliers can then be checked with more efficient detection algorithms.

In the light of outlier detection another simple test is Tukey's rule or variants thereof. For example, Solberg and Lahti (2005) used the Box-Cox transformation to transform the data to an approximate normal distribution. Thereafter, Tukey's rule is used to detect outliers. Tukey's rule classifies observations as outliers if they are outside the range

$$\left[Q_1-1.5\times IQR,Q_3+1.5\times IQR\right]. \tag{2.1}$$

In equation (2.1), Q_1 and Q_3 are the first and third quartile, respectively, and IQR is the interquartile range. It has been shown that this test has the ability to detect outliers. However, the algorithm breaks down due to the Box-Cox transformation (Pimentel *et al*., 2014).

Datasets for novelty detection are generally highly complex and require state-of-the-art procedures. Therefore, statistical modelling techniques must be used to perform probabilistic novelty detection. Statistical modelling can broadly be divided into a parametric or non-parametric approach. The parametric approach assumes that the underlying probability distribution can be modelled by a parametric function. In turn, the problem is reduced to estimating the parameters of the assumed model. Conversely, the non-parametric approach assumes no form, but finds a function that is close to the data while being adequately smooth.

Semi-parametric approaches fall in between these two methods, thereby improving the interpretability of the model and using the data directly to improve the model fit.

The simplest parametric approach to novelty detection is to assume that the normal class is generated by a parametric distribution. As a first step, the data should be reduced by removing all the known novel events. Hence, the reduced dataset contains only observations that are believed to be normal. Thereafter, a parametric distribution can be assumed for the normal data – for example, a Gaussian distribution. In turn, only the parameters of the distribution must be estimated. This is a very simple approach. However, it might be that a single distribution is too restrictive for the normal class. Therefore, mixture models have been used widely.

Mixture models are highly suitable for novelty detection. The most popular mixture model is the Gaussian mixture model. Again, the normal class is modelled by some distribution. As an example, consider the Gaussian mixture model. Hence, the probability density function of the normal or positive class is assumed to be a mixture of normal densities. Consequently, for multivariate data, the probability density function of the normal class is given by

$$f\left(\underline{x}\right)=\sum_{m=1}^{M}\alpha_m f_m\left(\underline{x},\underline{\mu}_m,\Sigma_m\right),\ \sum_{m=1}^{M}\alpha_m=1. \tag{2.2}$$

In equation (2.2), M is the number of distributions used, α_m , $m=1,2,\ldots,M$ are the mixing proportions, $f_m\left(\underline{x},\underline{\mu}_m,\Sigma_m\right)$ is the Gaussian probability density and $\underline{\mu}_m$ and Σ_m are the mean vector and covariance matrix of the m^{th} Gaussian distribution, respectively. To estimate the parameters in the model the EM algorithm is generally used (Hastie *et al*., 2009).

The output of the EM algorithm returns estimates of the mean vector, covariance matrix and mixing proportion of each component in the model. There is only one tuning parameter, namely the number of mixture components to use. This parameter plays a cardinal role in the ultimate goodness-of-fit. If there are too many mixture components in the model (large M), the model will overfit the data and have high variance. Conversely, if there are too few mixture components (small M), the model will be too rigid, thereby missing important structures in the data which leads to a high bias. Hence, there is a bias-variance trade-off. Conventionally, selection of the number of mixture components is based on the likelihood of the model or information theoretic criteria. The latter includes the Akaike information criterion (AIC) and Bayesian information criterion (BIC) (Huang, Peng & Zhang, 2013).

Extreme value theory is a parametric approach that has gained popularity in novelty detection literature. The theory of extremes is generally used to set a novelty threshold. Intuitively, it is believed that novel events are extreme in some sense. This might mean that they are close to the decision boundary or have a low probability of being in the normal class. Extreme value theory provides a theoretical framework that could be used to detect anomalous events, as will be explored in this dissertation.

Instead of using the parametric framework, non-parametric estimation can be used to model the normal class. Two of the most common approaches are kernel density estimation and negative selection. These two approaches are now briefly discussed.

Kernel density estimation is an unsupervised learning approach to model the normal class. Again, only the observations that are believed to be normal are considered. An estimate for the probability density is then found for the normal class. Given a new observation x_0, an estimate of the density at this point is obtained by using the Parzen estimate, namely

$$\hat{f}(x_0) = \frac{1}{N\lambda}\sum_{i=1}^{N} K_\lambda(x_0, x_i). \tag{2.3}$$

In equation (2.3) N is the sample size, λ is the width of the kernel, x_i, $i = 1,\dots,N$ are the sample observations and $K_\lambda(x_0, x_i)$ is the kernel. Notice that this non-parametric technique considers all the observations and weighs them based on their distance from the target point. The weights (kernel) decrease smoothly with the distance from the target point such that the density estimates are smooth. A commonly used kernel is the Gaussian kernel,

$$K_\lambda(x_0, x) = \phi\left({|x - x_0|}/{\lambda} \right). \tag{2.4}$$

Here, $\phi(\cdot)$ is the standard Gaussian kernel. Notice that $\phi\left({|x - x_0|}/{\lambda} \right) = \lambda\phi_\lambda(|x - x_0|)$, where $\phi_\lambda(|x - x_0|)$ is the Gaussian density with standard deviation λ. In turn, the estimated density is given by

$$\hat{f}(x_0) = \frac{1}{N\lambda}\sum_{i=1}^{N}\phi\left({|x - x_0|}/{\lambda} \right) = \frac{1}{N}\sum_{i=1}^{N}\phi_\lambda(|x - x_0|) = (\hat{F} * \phi_\lambda)(x). \tag{2.5}$$

Hence, the density estimate is the convolution of the empirical distribution and the Gaussian distribution with standard deviation λ. This means that the discontinuous empirical distribution function is smoothed by adding Gaussian noise to each observation in the sample (Hastie *et al.*, 2009). The obtained kernel density estimate provides a model for the normal class. Consequently, new observations are compared to this distribution. A novelty threshold must be selected such that if $\hat{f}\left(x_0\right) < t$, observation x_0 is classified as novel.

Although kernel density estimation is a powerful technique to model the positive class, it has some disadvantages. The width of the kernel determines the quality of the fit. Therefore, this parameter must be selected with care. Additionally, the entire sample must be considered for each new observation. Thus, for large datasets this approach is inefficient.

Negative selection is a non-parametric approach that was inspired by the human immune system. This approach was originally introduced by Forrest, Perelson, Allen and Cherukuri (1994). The human immune system discriminates between what is part of the body and anything anomalous. This process is known as self-nonself discrimination. T-cell receptors are generated by random processes of genetic rearrangements. These receptors identify anomalous cells, viruses or bacteria in the body. Any cells that do not successfully bind with the self-cells are considered anomalous by the immune system, and consequently destroyed. Negative selection is an idea based on how the immune system identifies viruses and/or bacteria in the body. This approach has been widely used for novelty and change-point detection (Pimentel *et al*., 2014).

Various other probabilistic approaches can be used to perform novelty detection, as explained by, among others, Pimentel *et al*. (2014). Probabilistic novelty detection has the ability to perform novelty detection accurately if a good estimate of the distribution of the normal class can be obtained. Furthermore, these approaches are generally represented in a mathematical framework. Consequently, inference on the results can be performed. However, the predictive power of these methods relies on the availability of a large sample (Pimentel *et al*., 2014).

2.4.2 Distance-based approach to novelty detection

This section describes distance-based methods for novelty detection. Distance-based methods rely on the use of a distance metric or similarity measure to determine the correspondence between two observations. These methods assume that observations close to a target point are similar (Hautamäki, Kärkkäinen & Fränti, 2004; Pimentel *et al*., 2009).

The K-nearest neighbour (KNN) algorithm is a simple non-parametric method for classification and regression. The algorithm is initialised by finding the K closest observations to a target point x_0. These observations form the neighbourhood around x_0. If regression is the ultimate task the average of the response of the observations in the neighbourhood is computed, or, if classification is the ultimate task a majority vote is used – the modal class is used as a prediction (Hastie *et al*., 2009). Two factors play a vital role in the performance of the KNN algorithm. That is, the value of K and the distance metric used. The former is usually chosen by cross-validation whereas the latter should be pre-specified.

The KNN algorithm can also be used to perform novelty detection and/or outlier detection. One approach is to find the K closest observations to the target point and compute the distances from the target point to each neighbour. If the target point is more than a distance $d_{\min}$ from each observation in the neighbourhood, it is considered an outlier (Pimentel *et al*., 2014).

A wide range of distance metrics has been used in the KNN algorithm (Duda, Hart & Stork, 2001). The most popular distance metrics are the Euclidean and Mahalanobis distances. Hence, the distances from the target observation to each other observation in the neighbourhood represent novelty scores of the similarity between the target observation and the samples in the neighbourhood. Instead of calculating the distance from a target observation to each of the K samples in the neighbourhood, some methods find the distance from the target point to the mean of the K observations in the sample. Techniques that use this approach are termed density-based methods (Hautamäki *et al*., 2004).

Since distances must be calculated, a natural question is how to handle categorical variables. One approach is the simple matching coefficient method. This method counts the number of attributes that match (have the same categorical response) and divides it by the total number of attributes. More sophisticated methods for dealing with categorical variables have been proposed by, among others, Boriah, Chandola and Kumar (2008).

Clustering algorithms are useful for novelty detection. The most popular clustering algorithm is the K-means clustering algorithm. This algorithm is initialised by specifying an initial set of K centres. As a second step, the observations closest to each centre are found. Hence, the data is divided into clusters where each cluster contains the observations closest to that cluster’s centre. The average of the observations in each cluster is computed and the cluster centres are updated as the mean of that cluster. Next, observations are again divided into clusters based on the new centres. This is repeated until convergence, *i.e.* the centres do not change (Hastie *et al*., 2009). The final centres can be used for novelty detection. Similar to the KNN algorithm, if a target point is too far from all the clusters it is considered an outlier (Pimentel *et al*., 2014). An approach followed by Clifton, Bannister and Tarassenko (2006) is to define novelty scores based on how many standard deviations a target point is away from its cluster centre, relative to the distribution of clusters.

One of the major problems with distance-based methods is the difficulty to handle high-dimensional data. This is a result of the curse of dimensionality. As the dimension increases, the hypervolume in which the observations are distributed increases exponentially. This means that naturally all points are a greater distance from one another. Another manifestation of the curse of dimensionality is the fact that observations move to the boundaries of the sample as the dimension increases. To see how this happens consider a hypersphere with radius 1 inscribed in a hypercube with edges of length 2 such that the sphere touches the cube at each side. Figure 2.1 shows such a setup in two dimensions.

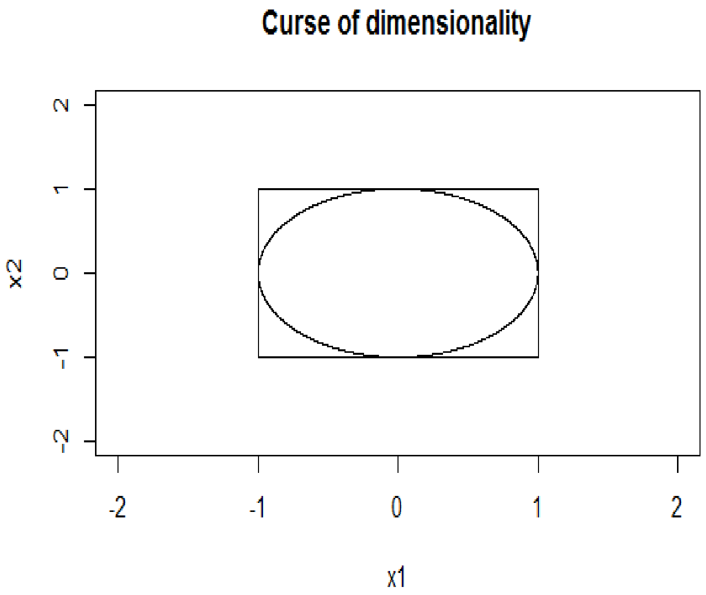


Figure 2.1: Curse of dimensionality in two dimensions

It can be shown that the volume of the hypersphere relative to that of the hypercube tends to zero as the dimension tends to infinity. On top of this, the convergence is remarkably fast. The convergence is shown in Figure 2.2. For each variable in the p-dimensional space an independent sequence of uniformly distributed random numbers between -1 and 1 is generated. The distance from each of the p-dimensional vectors to the origin is computed. If this distance is greater than 1 (the radius of the hypersphere) this vector falls outside the sphere. Figure 2.2 shows the proportion of vectors falling outside the sphere and, hence, in the edges of the sample. This was done for a sample size of 1 000.

Figure 2.2 shows how remarkably fast the proportion converges to 1. Therefore, it makes one believe that, in high dimensions, distance-based methods might lead to normal observations being classified as novel observations. Furthermore, due to the exponentially increasing size of the volume, the distance threshold strongly depends on the dimension. Other manifestations of the curse of dimensionality are mentioned in Hastie *et al*. (2009).

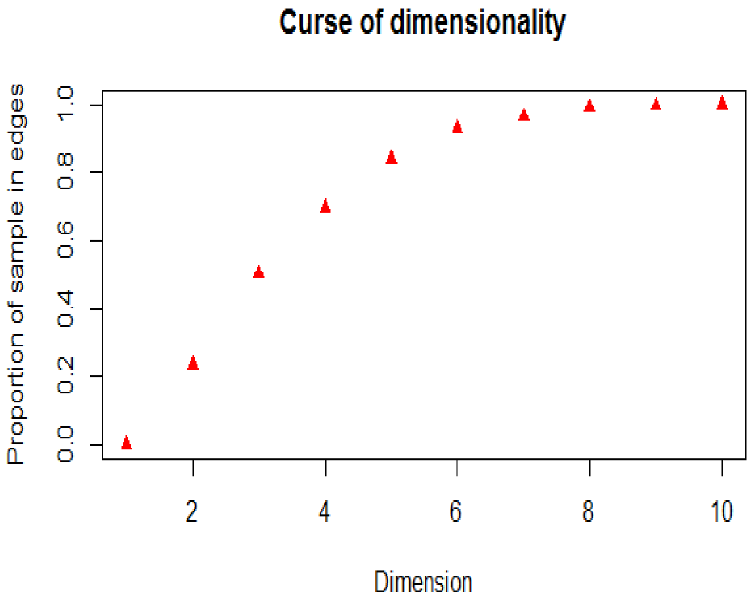


Figure 2.2: Proportion of points at edges of sample

Nevertheless, modifications to distance-based methods have been proposed to deal with the curse of dimensionality. One method would be to perform variable selection. This can be achieved by splitting the data into a training set and a validation set. The combination of variables that produces an optimal model performance on the validation set is selected. However, in extremely high-dimensional cases manually selecting variables is not feasible. Angiully and Pizzuti (2002) considered a weighted sum from the target point to each observation in the neighbourhood. The observations that produce the largest weighted sums are considered outliers or novel observations. Other methods to deal with high-dimensional data are mentioned in Pimentel *et al*. (2014).

Distance-based methods have the advantage that they are not based on an assumption regarding the distribution of the data. Therefore, in lower-dimensional settings these methods perform relatively well. Additionally, the methods proposed to deal with high dimensions can be used in complex settings (such as high dimensions) where assumptions cannot be validated easily.

2.4.3 Reconstruction-based approach to novelty detection

Two reconstruction-based approaches are neural network-based and subspace-based approaches. Many neural network-based approaches have been proposed for novelty detection. A review of these methods is given in Markou and Singh (2003). These methods will not be discussed in this dissertation.

Subspace-based approaches rely on the assumption that the data can be mapped onto a lower-dimensional manifold where the normal and novel observations are separated better. Thus, the data is transformed to a lower-dimensional space in such a way that class separation is maintained or improved.

Principal component analysis (PCA) is an unsupervised dimension reduction technique. The singular value decomposition of the data matrix decomposes this matrix into its principal component directions of the variables and the singular values. Each successive principal component direction explains less of the variability in the data. Therefore, only the first few principal components are selected. These orthogonal components are then used to transform the data matrix to a low-dimensional space. Notice that this technique does not provide a method to discriminate between normal and novel observations. Instead, it is a pre-processing step to reduce the dimension and/or improve class separation in an efficient manner.

Some extensions have been proposed to deal with novelty detection if the data is not linearly separable. One approach is kernel PCA. Kernel PCA first transforms the data to a higher-dimensional space where the data is better separable linearly. Principal component analysis is then performed in the transformed space. Hoffmann (2007) applied kernel PCA for novelty detection to the handwritten digits dataset and breast cancer cytology which demonstrated the competitiveness of this method. The data was transformed to an infinite dimensional feature space in which PCA was performed. Novel events were classified based on the squared distance to the corresponding principal subspace.

Subspace-based methods are useful when the data is not separated well or has a high dimension. However, these methods generally do not give a classifier to discriminate between normal and novel events. Instead, the data is mapped onto a lower-dimensional subspace such that novel events can be detected more easily.

2.4.4 Domain-based approach to novelty detection

The final method discussed in this chapter is a domain-based method for novelty detection. Domain-based methods describe the boundary of the normal class as opposed to its density. This means only the observations at the boundary are used to determine a novelty detection classifier. Therefore, this class of methods is generally robust against the distribution of the normal class (Pimentel *et al*., 2014).

A popular domain-based approach to novelty detection is the one-class support vector machine (SVM-1) algorithm. The one-class support vector classifier was defined by Schölkopf, Williamson, Platt, Shawe-Taylor and Smola (2000) as the solution to the quadratic program

$$\min_{\underline{\beta},\, \underline{\xi}\in\mathbb{R}^N,\, \rho\in\mathbb{R}} \left\{ \frac{1}{2}\left\|\underline{\beta}\right\|^2 + \frac{1}{\nu N}\sum\nolimits_i \xi_i - \rho \right\} \text{ s.t. } \left\langle \underline{\beta}, \Phi\left(\underline{x}_i\right)\right\rangle \geq \rho - \xi_i \text{ , } \xi_i \geq 0 \,. \qquad (2.6)$$

In equation (2.6), the vector $\underline{\beta}$ and parameter ρ are the regression coefficients and intercept defining a hyperplane in feature space, respectively. The sample size is denoted by N, the function $\Phi(\cdot)$ is a feature map that maps each observation to some feature space and the vector $\underline{\xi}$ is a vector containing the slack variables, ξ_i , $i = 1,\ldots,N$. Finally, the parameter ν is a tuning parameter. This tuning parameter controls the complexity of the one-class support vector machine.

The one-class support vector machine classifier considers only the training data of normal instances. Let $\underline{x} \in \mathrm{X}$ represent a positive observation and consider the mapping

$$\Phi : \mathrm{X} \rightarrow \mathrm{H} \,. \qquad (2.7)$$

Thus, the function $\Phi\left(\underline{x}\right)$ maps the vector $\underline{x} \in \mathrm{X}$ onto a feature space of possibly higher dimension. A similarity measure is defined as the inner product between samples in the feature space. This function describing the inner product is termed a kernel function and is denoted by

$$k\left(\underline{x},\underline{x}'\right)=\left\langle \Phi\left(\underline{x}\right),\Phi\left(\underline{x}'\right)\right\rangle_{\mathrm{H}}. \tag{2.8}$$

Kernel functions play a vital role in the theory of reproducing kernel Hilbert spaces. Interestingly, the solution to the optimisation problem of the SVM-1 algorithm only depends on the original data through the kernel. In turn, the feature map $\Phi\left(\underline{x}\right)$ need not be known. Furthermore, different kernels (which represent different feature mappings) can be used.

The SVM-1 algorithm maps the predictor space onto a feature space such that the positive data is separable from the origin. In the spirit of conventional support vector machines, the SVM-1 algorithm seeks the hyperplane $\left\langle \underline{\beta},\Phi\left(\underline{x}\right)\right\rangle=\rho$ such that the margin between the data and the origin is a maximum. New data falling above the hyperplane is considered normal and data falling below the hyperplane is considered novel. Ultimately, the hyperplane in feature space defines a non-linear decision boundary such that a function returns a 1 for a small region capturing most of the data and -1 elsewhere (Schölkopf *et al.*, 2000). Hence, the function to be estimated is

$$f\left(\underline{x}\right)=sign\left\{\left\langle \underline{\beta},\Phi\left(\underline{x}\right)\right\rangle-\rho\right\}. \tag{2.9}$$

It can be shown that, for the Lagrange multipliers $\alpha_i \geq 0$, $i=1,\ldots,N$, the coefficient vector is given by

$$\underline{\beta}=\sum_i \alpha_i \Phi\left(\underline{x}_i\right). \tag{2.10}$$

In turn, the decision function becomes

$$f\left(\underline{x}\right)=sign\left\{\sum_i \alpha_i k\left(\underline{x}_i,\underline{x}\right)-\rho\right\}. \tag{2.11}$$

Finally, the parameter ρ is recovered from the fact that for non-zero α_i the corresponding observation $\underline{x}_i$ satisfies

$$\rho=\left\langle \underline{\beta},\Phi\left(\underline{x}_i\right)\right\rangle=\sum_j \alpha_j k\left(\underline{x}_j,\underline{x}_i\right). \tag{2.12}$$

For the full derivation of the SVM-1 algorithm refer to Schölkopf *et al*. (2000). The resulting hyperplane separates the data a maximum margin of $\rho/\|\underline{\beta}\|$ from the origin. Furthermore, for each $\underline{x}_j$ that is misclassified the observation is a distance of $\xi_j/\|\underline{\beta}\|$ from the optimal hyperplane in feature space. Given that the SVM-1 algorithm is a constrained quadratic program, efficient optimisation strategies exist. Furthermore, from the derivation of the classifier (using the Lagrange multiplier) it is seen that only the observations at or within the margin determine the optimal solution. These observations are known as support vectors. Therefore, this method is a domain-based method as only the observations near the boundary of the normal class are used.

One aspect that needs careful consideration is the type of kernel function used to map the data to some feature space. Specifically, it is assumed that observations with high density in the normal class are mapped far from the origin whereas low-density observations are closer to the origin. Furthermore, possible novel observations should be the closest to the origin. A kernel that achieves this is the Gaussian kernel given by

$$k_G\left(\underline{x},\underline{x}'\right)=\exp\left\{-\frac{\|\underline{x}-\underline{x}'\|^2}{2\sigma^2}\right\}. \qquad (2.13)$$

Notice that this kernel is maximal at $k_G\left(\underline{x},\underline{x}\right)=1\ \forall\ \underline{x}\in\mathrm{X}$. Furthermore, as observations move away from each other the kernel moves towards zero. Therefore, observations far from the density of the normal class will be closer to the origin.

A method closely related to one-class support vector machines is the support vector domain description (SVDD) method. If a Gaussian kernel is used (or any kernel that only depends on $\underline{x}-\underline{x}'$) the SVDD method is equivalent to the one-class support vector machine (Schölkopf & Smola, 2002). The SVDD algorithm, proposed by Tax and Duin (1999), finds the hypersphere with minimum volume that surrounds the positive data. Let R and $\underline{a}$ be the radius and centre of the hypersphere, respectively. Furthermore, to allow small errors let $\underline{\xi}$ be a vector of slack variables with elements ξ_i , $i=1,\ldots,N$ describing how far a corresponding observation $\underline{x}_i$ lies outside the hypersphere. The SVDD optimisation is formulated as

$$\min\left\{R^2+C\sum_i\xi_i\right\}\text{ s.t. }\|\underline{x}_i-\underline{a}\|^2\leq R^2+\xi_i\ ,\ \xi_i\geq 0\ \forall\ i\,. \qquad (2.14)$$

In equation (2.14) C is a tuning parameter controlling the flexibility of the model. Again, through using Lagrange multipliers, it is seen that the optimisation only depends on the data through inner products – the solutions to all the parameters only depend on $\langle \underline{x}, \underline{x} \rangle$. In turn, any basis expansion (in the predictor space) could be used to improve the classifier. Moreover, it is known that the inner product in some feature space is represented by a reproducing kernel which means that $k(\underline{x}, \underline{x}') = \langle \Phi(\underline{x}), \Phi(\underline{x}') \rangle_{\mathrm{H}}$, as seen previously. Hence, the kernel trick can be used in the SVDD algorithm (all calculations can be done by only using the kernel). Furthermore, it is again the case that only the observations that lie at the boundary or outside the hypersphere are used to determine the solution. These observations are known as support vectors (Tax & Duin, 1999).

Domain-based methods, specifically the SVM-1 and SVDD methods, have the ability to handle high-dimensional data. Although overfitting must still be controlled, almost no assumptions other than that the data describes the normal class are made. Therefore, the algorithm can be seamlessly applied to high-dimensional data with the use of appropriate regularisation. A disadvantage of these approaches is that they do not produce probabilities of the certainty of the classified observation. The classifier only returns a 1 if the observation is predicted to be normal and a -1 if the observation is predicted novel.

2.5 PROPERTIES OF AN EFFICIENT NOVELTY DETECTION ALGORITHM

It is now clear that there are many approaches to novelty detection. The question is which of these methods perform the best in general. There is not a universal method that produces superior results on all datasets. However, there are some properties that a good novelty detection algorithm should possess. These properties are now discussed.

2.5.1 Predictive power

The algorithm should be able to detect novel events and correctly classify normal observations as normal. Generally, there is a trade-off between these two requirements. Algorithms that detect novel observations with high sensitivity might misclassify normal observations as novel. Conversely, algorithms that are too robust to novel observations might misclassify novel observations as normal whereas most of the normal observations are classified correctly. Therefore, it is important to investigate the misclassification rate of the model as well as to examine where the model makes errors. Additionally, the model should generalise well to new data. Hence, a validation set is required to validate the model. Models that are too flexible generally have a low error on the training data and a high error on the validation or test data.

2.5.2 Interpretability

Model interpretability is a cardinal property of efficient algorithms. It is extremely helpful to be able to interpret which predictors have the biggest influence on the model. A good interpretation of how a model performs helps one to understand how the model detects novelties. Consequently, informed decisions can be made based on the model. In general, there is a trade-off between model interpretability and model flexibility. Probabilistic methods have the advantage of producing class probabilities. Hence, the confidence of the classification of the model can be estimated.

2.5.3 Computational time

Computational time is another factor that can govern the selection of a novelty detection algorithm. The computational time depends on the size and dimension of the dataset. For example, in a low-dimensional setting mixture models are useful to determine a density estimate of the normal class. In turn, the density estimates are used to perform novelty detection. However, in a high-dimensional setting it is difficult to train a mixture model on the data. Furthermore, multivariate mixture densities have a long computational time. Therefore, methods such as these are not suitable for data streaming or high-dimensional settings. On the other hand, one-class support vector machines can be regularised and trained on high-dimensional data. Other incremental learning algorithms have been proposed for novelty detection. These algorithms are designed to update as a new observation is added to the dataset. In some cases, data arrives continuously, which means that incremental learning algorithms must be used. In general, as a new observation enters the data, the model classifies this observation as normal or novel. Using this information, the algorithm is updated.

2.5.4 Ability to handle high-dimensional data

The ability to handle high-dimensional data is becoming increasingly important as the availability of data increases. It is not always the case that data is high-dimensional. In low-dimensional settings strict assumptions (such as normality) can be validated easily. However, in high-dimensional settings it is difficult to validate assumptions or to explore the structure in the data. Therefore, in these settings it is useful to consider a model that can be trained without much information on the structure of the data. Alternatively, assumptions to simplify the model can be made. For example, in high-dimensional settings it might be useful to assume only main effects are present (additive assumption) if a density estimate must be obtained. Furthermore, it is recommended to perform variable selection or dimension reduction to simplify the model.

2.5.5 Ability to handle unbalanced datasets

A specific characteristic of novelty detection datasets is the imbalance of classes. It is generally the case that there are many observations from the normal class and only a few, if any, novel observations. This is because novelty detection is usually performed on high-integrity systems – detecting system failure, detecting fraud or monitoring vital signs. Therefore, in many cases, it is too expensive to sample novel observations. The effect of the class imbalance on the classifier must be controlled. For example, supervised algorithms cannot be used if no novel observations have been observed. Even if novel observations are present in the training data, the class imbalance causes supervised algorithms to break down. Novelty detection handles class imbalance by only using the normal class and then testing new observations against this class. This is because the normal class is generally well sampled such that accurate density estimates can be obtained. Some novelty detection algorithms incorporate novel observations into the algorithm to improve the prediction accuracy. Other methods rely on resampling techniques to reduce class imbalance. Usually, the resampling is performed by only considering the normal data.

As mentioned previously, the class imbalance also limits the available model validation techniques. It would be wrong to only consider a misclassification error as it would be at least as low as the proportion of the under-sampled class (or number of novelties). Therefore, techniques that better describe the trade-off between the specificity and sensitivity of the model must be used.

2.6 APPLICATIONS OF NOVELTY DETECTION

In this section, a summary of the practical applications of novelty detection is given. Novelty detection is applied in many practical fields. As discussed in this section, novelty detection plays a vital role in high-integrity systems where the cost of a misclassification is high. Therefore, these techniques are of extreme importance.

2.6.1 Fraud detection

Novelty detection data is used to build algorithms that help to detect fraud. Fraudulent observations do not occur frequently. Hence, in general, this class will be highly under-sampled. It might also be the case that there are no observations that have been explicitly marked as fraudulent. For this reason, novelty detection is a useful approach. Different types of fraud detection include credit card fraud, insurance fraud, telecommunication fraud and insider trading. In cases like these the algorithm must be able to classify accurately and fast. This is because data from these scenarios are generated continuously and the cost of a misclassification is high. Fawcett and Provost (1999) proposed a general approach for fraud detection termed activity monitoring. In general, a model is built for the usage profile of each customer or record. New observations are compared to the usage profile of the corresponding customer to detect anomalous activity.

2.6.2 Image detection

Image detection has also been performed using novelty detection. A search engine consists of two parts – image retrieval and image detection. During the image retrieval phase a shortlist of images that correspond with a query is constructed. These images are possible images that match the query. Image detection is the process of detecting which images in the shortlist do not match the query. Furon and Jégou (2013) proposed an approach based on extreme value theory to detect possible outliers from the shortlist of images. Similarity scores were constructed for each image in the shortlist. The generalised Pareto distribution together with the limiting distribution of order statistics were used to formulate a hypothesis test based on the similarity scores. The algorithm was then applied in a sequential manner such that images that most likely do not match the query were removed first.

2.6.3 Network security

As technology progresses, the risk of network intrusion increases. Cyber security has received a significant amount of attention over the past decade. It is becoming increasingly important for companies and individuals to protect their classified or personal information as a result of the online environment in which we operate. The difficulty of network intrusion is that data arises in a data streaming manner. Therefore, the algorithm must be able to perform online analysis relatively fast. Another challenge in terms of network intrusion is the fact that hackers are aware of the current methods used to prevent network intrusion. Consequently, the techniques used to commit network intrusion are constantly changing. This makes detecting network intrusion specifically difficult because a feature, which serves as a major indicator of an intrusion, may have no significance if a different intrusion method is used.

2.6.4 Medical safety

Novelty detection has many applications in the medical field. One application is the detection of rare diseases. It is usually too expensive to obtain a representative sample of patients with a rare disease. Therefore, this class would be substantially under-sampled. In turn, supervised approaches cannot be used efficiently without taking the class imbalance into account. Novelty detection is therefore a general approach. Hence, only a good sample of patients without the disease is required. A model is built based on these patients and new patients are tested against the normal model. Another interesting problem in medical safety is the modelling of disease outbreak. Again, it is the case that possibly no observations of the disease have been made. Agarwal (2007) used a Bayesian approach on the logs of emergency visits to the hospital. A model based on patients without the disease is constructed. Thereafter, patients were compared to this model to detect a possible disease outbreak. Another example of novelty detection in the medical field is vital-sign monitoring. Clifton, Hugueny and Tarassenko (2011) used the heart rate and breathing rate of patients as the two predictors of vital-sign monitoring. The goal of the experiment was to detect when patients require immediate attention from the medical staff. This problem is complicated by the fact that data arrived every two seconds. Furthermore, the variability between different patients is large. Hence, the model describing the normal heart rate and breathing rate must be able to incorporate these variances.

2.7 Conclusion

Novelty detection has significant applications in a variety of fields. It is generally the case that the cost of a misclassification is high. Novelty detection algorithms must therefore be able to identify novel events accurately while maintaining a low false positive rate. Furthermore, the data used for novelty detection is complicated and might arrive continuously. It is therefore important to investigate the nature of the novel events. This will aid the analyst in constructing an efficient novelty detection algorithm.

Various methods can be used for novelty detection. It is not the case that one method dominates the others. Instead, the type of novelty to be detected and the complexity of the data govern the choice of the algorithm. Finally, as a result of the class imbalance present in novelty detection problems, conventional supervised learning algorithms break down. Consequently, an approach based on novelty detection must be used.

CHAPTER 3

A REVIEW OF EXTREME VALUE THEORY

3.1 INTRODUCTION

This chapter focuses on the theory of extreme value statistics. Extreme value theory is a branch of statistics that provides tools to analyse the tails of distributions. Conventional applications of extreme value theory include estimating maximum flood levels, predicting the minimum or maximum yield of crops, assessing the probability of large claims for reinsurance companies and estimating the Value-at-Risk of a portfolio (Beirlant *et al*., 2004). Recently, novelty detection with the use of extreme value theory has been proposed.

The aim of this chapter is to give a review of extreme value theory. Two approaches are given to estimate the tails of a distribution. The basic idea underlying these approaches is discussed. It is shown how the models are fitted to data and how to perform visual goodness-of-fit tests. Finally, a short section explains the challenge of multivariate extremes. This chapter is concluded with a discussion on the use of extreme value theory for novelty detection. Most of the theory discussions in this chapter, as well as advanced discussions of extreme value theory, can be found in Beirlant *et al*. (2004).

3.2 CLASSICAL EXTREME VALUE THEORY

In this section, the classical approach to extreme value theory is discussed.

3.2.1 Problem statement

Consider the sequence $X_1, X_2, X_3, \ldots$ of independent and identically distributed (iid) random variables with distribution function F. The central limit theorem provides a limiting distribution for the sum of the independent and identically distributed random variables. It is well known that under the assumption of a finite variance the sum of iid random variables converges in distribution to a normal distribution. In turn, a limiting distribution for the sample mean is recovered.

The classical approach to extreme value theory seeks to find a limiting distribution for the maximum (or minimum) of a sequence of iid random variables. Hence, under some mild assumptions on the distribution of the random variables, a parametric (limiting) distribution is obtained for the maximum or minimum of a sequence of iid random variables. Inferential statistics can therefore be derived based on this limiting distribution. This approach is now discussed.

3.2.2 The Fisher-Tippett theorem

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with cumulative distribution function F and let $X_{n,n} = \max\{X_1, X_2, \ldots, X_n\}$. Given that a non-degenerate distribution G and sequences of constants $\{a_n\} > 0$ and $\{b_n\}$ exist such that

$$a_n^{-1}\left(X_{n,n} - b_n\right) \to G(x) \text{ , } n \to \infty \text{ ,} \tag{3.1}$$

then $G(x)$ is a generalised extreme value (GEV) distribution with cumulative distribution function given by

$$G_\gamma(x) = \begin{cases} \exp\left\{-(1+\gamma x)^{-1/\gamma}\right\} & \gamma \neq 0 \text{ , } \quad (1+\gamma x) > 0 \\ \exp\left\{-\exp(-x)\right\} & \gamma = 0 \qquad x \in \mathbb{R} \text{ .} \end{cases} \tag{3.2}$$

For proof of this theorem refer to Beirlant *et al*. (2004). In the classical approach, two parts to the problem have been identified. The first is termed the extremal limit problem. This part of the problem, solved by Fisher and Tippett (1928), seeks to find all possible non-degenerate distributions that can appear in the limit in equation (3.1) for a given set of sequences $\{a_n\} > 0$ and $\{b_n\}$. They showed that the possible limits are the class of GEV distributions (Beirlant *et al.*, 2004). The second part of the problem is termed the domain of attraction condition. Some definitions are required to state this condition. Therefore, these definitions are given below. The quantile function of a random variable X is defined as

$$Q(u) = \inf\left\{x : F(x) \geq u\right\} \text{ , } u \in (0,1) \text{.} \tag{3.3}$$

The tail-quantile function, $U(x)$, is then defined as

$$U(x) = Q\left(1 - \frac{1}{x}\right) \text{ , } x \in (1,\infty) \text{.} \tag{3.4}$$

The domain of attraction condition states the condition on F, or equivalently on U, required such that for given sequences of constants $\{a_n\} > 0$ and $\{b_n\}$, the standardised maximum

$\dfrac{X_{n,n}-b_n}{a_n}$ converges in distribution to the GEV distribution. The condition on the distribution function, or equivalently in terms of the tail quantile function, states that for some positive function $a(.)$ and for any $u>0$, $\lim_{x\to\infty}\left\{\dfrac{U(xu)-U(x)}{a(x)}\right\}=:h(u)$ exists and is not identically zero. Additionally, it can be shown that the only possible limits satisfying this condition is of the form $c\cdot h_\gamma(u)=\dfrac{c\left(u^\gamma-1\right)}{\gamma}$ where $c>0$. Given the fact that $c>0$ and $a(x)$ is an ultimately positive function, the constant c can be absorbed into $a(x)$ such that the condition becomes

$$\left(C_\gamma\right):\ \lim_{x\to\infty}\left\{\frac{U(xu)-U(x)}{a(x)}\right\}=:h_\gamma(u)=\frac{u^\gamma-1}{\gamma}\ ,\ \forall\ u>0\,. \tag{3.5}$$

It is assumed that $h_0(u)=\ln(u)$. Additionally, by choosing $a_n=a(n)$ and $b_n=U(n)$, it follows that under this condition on F, $\dfrac{X_{n:n}-b_n}{a_n}\xrightarrow{D}G_\gamma$ as $n\to\infty$ where G_γ is the generalised extreme value distribution. Consequently, under the domain of attraction condition $\left(C_\gamma\right)$ and for argument's sake assuming $\gamma\neq 0$,

$$F_{X_{n,n}}\left(a_nx+b_n\right)\to\exp\left\{-\left(1+\gamma x\right)^{-1/\gamma}\right\}\ ,\ n\to\infty\,. \tag{3.6}$$

Furthermore, the distribution of the maximum in terms of the marginal distribution of the sequence of iid random variables is easily obtained. Notice that

$$F_{X_{n,n}}(x)=P\left(X_{n,n}\leq x\right)=P\left(X_1\leq x,\ldots,X_n\leq x\right)=\prod_{i=1}^{n}P\left(X_i\leq x\right)=F^n(x)\,. \tag{3.7}$$

Hence, $F_{X_{n,n}}\left(a_nx+b_n\right)\overset{d}{=}F^n\left(a_nx+b_n\right)\to G_\gamma(x)$, $n\to\infty$ such that

$F\left(a_nx+b_n\right)\overset{d}{\approx}G_\gamma^{1/n}(x)$ for large n.

Thus, it is possible to determine extreme probabilities or quantiles of the underlying distribution by utilising the information in the tails of this distribution. Consider equation (3.6). Equivalently, this equation can be written as

$$F_{X_{n,n}}(x) \approx \exp\left\{-\left(1+\gamma\frac{x-b_n}{a_n}\right)^{-1/\gamma}\right\} \text{ for large } n . \tag{3.8}$$

Let $b_n = \eta$ and $a_n = \beta$ represent location and scale parameters, respectively. Then, from equation (3.8), the parameterised GEV distribution has the form

$$G_{\gamma,\eta,\beta}(x) = \exp\left\{-\left(1+\gamma\frac{x-\eta}{\beta}\right)^{-1/\gamma}\right\} , \gamma,\eta \in \mathbb{R} , \beta > 0 , \left(1+\gamma\frac{x-\eta}{\beta}\right) > 0 . \tag{3.9.1}$$

If $\gamma = 0$, the parameterised GEV distribution is given by

$$G_{0,\eta,\beta}(x) = \exp\left\{-\exp\left\{-\frac{x-\eta}{\beta}\right\}\right\} ,\eta \in \mathbb{R} , \beta > 0 , x \in \mathbb{R} . \tag{3.9.2}$$

Hence, the generalised extreme value distribution has three parameters, namely a location parameter, η, a scale parameter, β, and a shape parameter, γ. The shape parameter is termed the extreme value index (EVI). This parameter is crucial to the theory of extremes as it specifies the type of GEV distribution. There are three types of GEV distributions, namely the Gumbel type, Fréchet-Pareto type and extremal Weibull type. These types of GEV distributions are also known as type I, type II and type III extreme value distributions, respectively. The different types of GEV distributions result from the type of tail the GEV distribution describes (in the limit).

The Pareto type is categorised by a positive EVI – $\gamma > 0$. Distributions in the domain of attraction of the Pareto type distributions (denoted as $F \in D\left(\Phi_{1/\gamma}\right)$) are known to have heavier than exponential tails. The tail of such a distribution decays at approximately a polynomial rate. Furthermore, distributions in this class are identified by having a regularly varying (r.v.) tail-quantile function with an index of regular variation γ (Beirlant *et al*., 2004). Hence,

$$U(x) \sim x^{\gamma}\ell_U(x) , x \to \infty . \tag{3.10}$$

The function $\ell_U(x)$ is a slowly varying (s.v.) function which means

$$\lim_{x\to\infty}\frac{\ell_U(xu)}{\ell_U(x)} = 1 \; \forall \; u > 0 . \tag{3.11}$$

Given that the tail-quantile function of a random variable X is of the form in (3.10), the sequences of constants can be chosen as

$$a_n = a(n) = \gamma n^{\gamma}\ell_U(n) = \gamma \cdot U(n)\text{, and} \tag{3.12.1}$$

$$b_n = n^{\gamma}\ell_U(n) = U(n)\,, \tag{3.12.2}$$

since then,

$$\frac{U(xu)-U(x)}{a(x)} = \frac{(xu)^{\lambda}\ell_U(xu)-x^{\gamma}\ell_U(x)}{\gamma x^{\gamma}\ell_U(x)} = \frac{1}{\gamma}\left(u^{\gamma}\frac{\ell_U(xu)}{\ell_U(x)}-1\right) \sim h_{\gamma}(u)\ ,\ x\to\infty\,.$$

Hence, condition (C_{γ}) is satisfied for any $u>0$. In general, it is observed that condition (C_{γ}) holds with $\gamma>0$ if $a(x)/U(x)\sim\gamma$ (Beirlant *et al*., 2004).

The Gumbel type is characterised by an EVI of zero – $\gamma=0$. Distributions in this domain of attraction share the property of tails that decrease at approximately an exponential rate. In the thesis of De Haan (1970) it was shown that a distribution F is in the domain of attraction of the Gumbel type of GEV distributions (denoted by $F\in D(\Lambda)$) if and only if

$$\frac{1-F(y+b(y)v)}{1-F(y)} \to \exp\{-v\}\ \ \forall\ \ v>0 \text{ as } y\to x_u\,. \tag{3.13}$$

Notice that x_u is the maximum possible value that the random variable can take on – the upper bound. Furthermore, if (3.13) holds it follows further that

$$\frac{b(y+b(y)v)}{b(y)} \to 1. \tag{3.14}$$

In equations (3.13) and (3.14) the function $b(\cdot)$ is an auxiliary function. This function can be chosen as $b(y)=a(U^{-1}(y))$ (Beirlant *et al*., 2004). This alternative condition is rather different to condition (C_{γ}). However, it is the building block of the modern approach to extreme value theory.

The extremal Weibull type is characterised by a negative EVI – $\gamma < 0$. Distributions in the domain of attraction of the extremal Weibull type of GEV distributions (denoted by $F \in D\left(\Psi_{-1/\gamma}\right)$) are identified as having a finite upper bound for the maximum. The results for this type of GEV distribution are very similar to the Pareto type. Distributions in the domain of attraction of the extremal Weibull type have a tail-quantile function of the form

$$U(y) = x_u - y^{\gamma}\ell_U(y) \text{ , } y \to \infty . \tag{3.15}$$

Using condition $\left(C_\gamma\right)$ it is seen that

$$\frac{U(xu)-U(x)}{a(x)} = \frac{x^{\gamma}\ell_U(x)-(xu)^{\gamma}\ell_U(xu)}{a(x)} = \frac{x^{\gamma}\ell_U(x)}{a(x)}\left[1-u^{\gamma}\frac{\ell_U(xu)}{\ell_U(x)}\right].$$

Hence, condition $\left(C_\gamma\right)$ is satisfied for any $u > 0$ if $a(x) \sim -\gamma \cdot x^{\gamma}\ell_U(x) = -\gamma\left(x_u - U(x)\right)$. Thus, all that is required is that

$$\frac{a(x)}{x_u - U(x)} \to -\gamma \text{ , } x \to x_u . \tag{3.16}$$

It is now clear that, as a result of the Fisher-Tippett theorem, the family of GEV distributions is the only possible limit that can appear for a normalised sequence of maxima. Furthermore, the domain of attraction condition and the EVI specify which type of GEV distribution appears in the limit. It must be mentioned that all these results are equivalent if the limiting distribution of the minimum is investigated. Notice that $\min\left\{X_1,\ldots,X_n\right\} = -\max\left\{-X_1,\ldots,-X_n\right\}$. Once this transformation is performed the results for the minimum follow directly.

3.2.3 The block-maxima method

One difficulty in estimating the limiting distribution is the fact that a sample only contains one maximum. Therefore, some method must be used to obtain an approximate sample of maxima that represents the upper values of the sample. This is what the block-maxima method entails.

Consider a sample denoted by $\left\{X_1, X_2, \ldots, X_N\right\}$. The sample is divided into M blocks of approximately equal size. For each block, the maximum in that block is computed which is termed the block-maximum of block $m = 1,\ldots,M$. This leads to a sample of block-maxima denoted by $\{Y_1,\ldots,Y_M\}$. Consequently, a sample of maxima is obtained. This sample is used for parameter estimation.

A bias-variance trade-off is present which is controlled by the number of blocks used to estimate maxima. If M is too small, there will only be a few block-maxima. Notice that each block would then contain many observations so that the maxima are close to the theoretical maxima with high probability. Hence, these maxima have low bias. However, due to the small sample size of the block-maxima, the variance of the estimates would be large. In turn, the variability in the estimates distorts the results. Conversely, if M is too large, there are too many block-maxima. Their estimates will have lower variability since more maxima are used to estimate the GEV distribution. However, the blocks do not contain enough sample points to accurately approximate the GEV distribution. In turn, these block-maxima cause high bias in the estimates.

Choosing the number of blocks is a challenge of the block-maxima method. No fixed method is available to choose the optimal number of blocks. It is important to keep in mind that the number of blocks influences the bias and the variance in the estimates and some optimal value must be determined. In some cases, novelty detection has an advantage here. If a validation set is available (or through cross-validation), the algorithm can be tested for different numbers of blocks. Consequently, a numerical estimate of the optimal number of blocks can be obtained.

Finally, it must be mentioned that the disadvantage of the block-maxima method is the reduction in the sample size. A significant amount of information is lost by assuming that only the block-maxima carry information regarding the tail of a distribution. It might be that one block contains more information than other blocks. Still, only the maximum of these blocks is used. Fortunately, the modern approach (discussed in Section 3.3) overcomes this shortcoming of the block-maxima approach.

3.2.4 Parameter estimation

Recall that the parameterised GEV distribution has the form

$$G_{\gamma,\eta,\beta}(x)=\begin{cases}\exp\left\{-\left(1+\gamma\dfrac{x-\eta}{\beta}\right)^{-1/\gamma}\right\} & \gamma\neq 0\ ,\ \eta\in\mathbb{R},\ \beta>0,\quad \left(1+\gamma\dfrac{x-\eta}{\beta}\right)>0\\[2ex] \exp\left\{-\exp\left(-\dfrac{x-\eta}{\beta}\right)\right\} & \gamma=0\ ,\ \eta\in\mathbb{R},\ \beta>0,\quad x\in\mathbb{R}\ .\end{cases} \tag{3.17}$$

This distribution, together with the sample of block-maxima, is used to estimate the parameters of the GEV distribution. Given that the original sample is assumed to be an iid sample, the block-maxima also form an iid sample of maxima.

The method of maximum likelihood is a popular approach to estimate the parameters of the GEV distribution. Notice that the likelihood is over $\{Y_1,\ldots,Y_M\}$ as it is assumed that each Y_m is approximately GEV and the M block-maxima form an iid sample. Hence, for $\gamma \neq 0$, the log-likelihood is given by

$$\log L(\gamma,\beta,\mu) = -M\ln(\beta) - \left(\frac{1}{\gamma}+1\right)\sum_{m=1}^{M}\ln\left(1+\gamma\frac{Y_m-\eta}{\beta}\right) - \sum_{m=1}^{M}\left(1+\gamma\frac{Y_m-\eta}{\beta}\right)^{-1/\gamma}. \tag{3.18}$$

Furthermore, if $\gamma = 0$, the log-likelihood is given by

$$\log L(0,\beta,\mu) = -M\ln(\beta) - \sum_{m=1}^{M}\exp\left\{-\frac{Y_m-\eta}{\beta}\right\} - \sum_{m=1}^{M}\frac{Y_m-\eta}{\beta}. \tag{3.19}$$

Notice that (3.18) is only valid for Y_m if $1+\gamma\frac{Y_m-\eta}{\beta}>0$. The vector of parameter values that maximises (3.18) or (3.19) is the maximum likelihood estimator, denoted by $(\hat{\gamma},\hat{\beta},\hat{\eta})$. In general, there is no closed-form expression for the maximum likelihood estimator. Nevertheless, efficient procedures exist to solve this optimisation. For example, Hosking (1985) used the Newton-Raphson iteration with some modifications to speed up the convergence. Furthermore, open source code is available on all major computing platforms.

Finally, as stated in Beirlant *et al*. (2004), if $\gamma > -0.5$, the maximum likelihood estimators possess the usual properties of consistency, asymptotic efficiency and asymptotic normality.

Another approach that can be used to estimate the parameters of the GEV distribution is the method of probability-weighted moments (PWMs). Defined by Greenwood, Landwehr, Matalas and Wallis (1979), the probability-weighted moments for $p,r,s \in \mathbb{R}$ of the random variable X with distribution F are

$$M_{p,r,s} = E\left[X^p\{F(X)\}^r\{1-F(X)\}^s\right]. \tag{3.20}$$

Specifically, for the GEV distribution the PWMs with $p=1$, $r=0,1,2$ and $s=0$ are used. A general expression for these PWMs is

$$M_{1,r,0}=\frac{1}{r+1}\left\{\eta-\frac{\beta}{\gamma}\left[1-\left(r+1\right)^{\gamma}\Gamma\left(1-\gamma\right)\right]\right\}\text{ , }\gamma<1. \tag{3.21}$$

These PWMs are estimated from the sample of block-maxima. An unbiased estimator for the PWMs is

$$\hat{M}_{1,r,0}=\frac{1}{M}\sum\nolimits_{m=1}^{M}\left(\prod_{j=1}^{r}\frac{m-j}{M-j}\right)Y_{m,M}\ . \tag{3.22}$$

In equation (3.22), the variables $Y_{m,M}$, $m=1,\ldots,M$ are the ranked sample of maxima. Finally, $\hat{\gamma}_{PWM}$ is obtained by solving equation (3.23.1) numerically, given by

$$\frac{3\hat{M}_{1,2,0}-\hat{M}_{1,0,0}}{2\hat{M}_{1,1,0}-\hat{M}_{1,0,0}}=\frac{3^{\hat{\gamma}_{PWM}}-1}{2^{\hat{\gamma}_{PWM}}-1}. \tag{3.23.1}$$

Using this estimate for $\hat{\gamma}_{PWM}$, the parameter estimate for β is found from

$$\hat{\beta}_{PWM}=\frac{\hat{\gamma}_{PWM}\left(2\hat{M}_{1,1,0}-\hat{M}_{1,0,0}\right)}{\Gamma\left(1-\hat{\gamma}_{PWM}\right)\left(2^{\hat{\gamma}_{PWM}}-1\right)}. \tag{3.23.2}$$

The parameter estimate for η is then obtained from

$$\hat{\eta}_{PWM}=\hat{M}_{1,0,0}+\frac{\hat{\beta}_{PWM}}{\hat{\gamma}_{PWM}}\left(1-\Gamma\left(1-\hat{\gamma}_{PWM}\right)\right). \tag{3.23.3}$$

These results were formally derived by Hosking, Wallis and Wood (1985) and summarised in Beirlant *et al*. (2004: 133-135). In some cases, the sample size is too small for the maximum likelihood estimator to be efficient. The PWM estimator is more robust to sample variability. Furthermore, the equations to be solved are simple and, therefore, solutions can be obtained quickly.

3.2.5 Goodness-of-fit evaluation

In this section, some approaches to evaluate the goodness-of-fit of the estimated GEV distribution are discussed. Assessing the goodness-of-fit for the GEV distribution can be undertook in many ways. In this section, only some visual diagnostic checks are discussed. However, more sophisticated tests – such as the Anderson-Darling, Kolmogorov-Smirnov or correlation coefficient goodness-of-fit tests – are also recommended.

The challenge when constructing quantile-quantile plots (QQ-plots) for the GEV distribution is the incorporation of the EVI parameter. Therefore, consider the case where $\gamma = 0$ – hence the Gumbel distribution. The Gumbel distribution is characterised by the quantile function

$$Q(u) = F^{-1}(u) = \eta - \beta \ln(-\ln(u)) \text{ , } \eta, \beta \in \mathbb{R},\ \beta > 0 \text{ ; } u \in (0,1) . \tag{3.24}$$

If $Q_s(u) = -\ln(-\ln u)$ is the standard Gumbel quantile function, it means that for $u \in (0,1)$,

$$Q(u) = \eta + \beta \cdot Q_s(u) . \tag{3.25}$$

Notice that the Gumbel distribution only has a location and scale parameter. Therefore, a QQ-plot can be constructed using (3.25). Consider a ranked sample $\{x_{1,N} < x_{2,N} < \ldots < x_{N,N}\}$ of block-maxima that is believed to be Gumbel distributed. This sample is compared to the theoretical quantiles of the standard Gumbel distribution. To find the respective theoretical quantiles corresponding with the sample, plotting-position estimates are used. These estimates are

$$p_{i,N} = {}^{i}\!\big/\!{}_{(N+1)} \text{ , } i = 1, \ldots, N . \tag{3.26}$$

Notice that other definitions of plotting-positions exist. It is then clear that

$$X_{i,N} \approx Q\left({}^{i}\!\big/\!{}_{(N+1)} \right) = \eta + \beta \cdot Q_s\left({}^{i}\!\big/\!{}_{(N+1)} \right) . \tag{3.27}$$

Therefore, the coordinates $\left(-\ln\left(-\ln\left({}^{i}\!\big/\!{}_{(N+1)} \right) \right), x_{i,N} \right)$, $i = 1, \ldots, N$ are plotted. The plotted points should fall on a straight line if the data corresponds with a Gumbel distribution. Furthermore, as mentioned by Beirlant *et al*. (2004), the QQ-plot can then be used to estimate the location and scale parameters. This is done by fitting the least squares line to the points. The estimates of the intercept and slope are estimates for the location and scale parameters, respectively. This remark is also seen in equation (3.27). Additionally, the coefficient of correlation quantifies how well the Gumbel distribution fits the data.

As an example, consider generating 100 observations from a Gumbel distribution with location parameter 1 and scale parameter 2. To each generated observation an independent noise component from an $N(0,0.05)$ distribution is added. Comparing this sample to the standard Gumbel quantiles produces the QQ-plot, as can be seen in Figure 3.1:

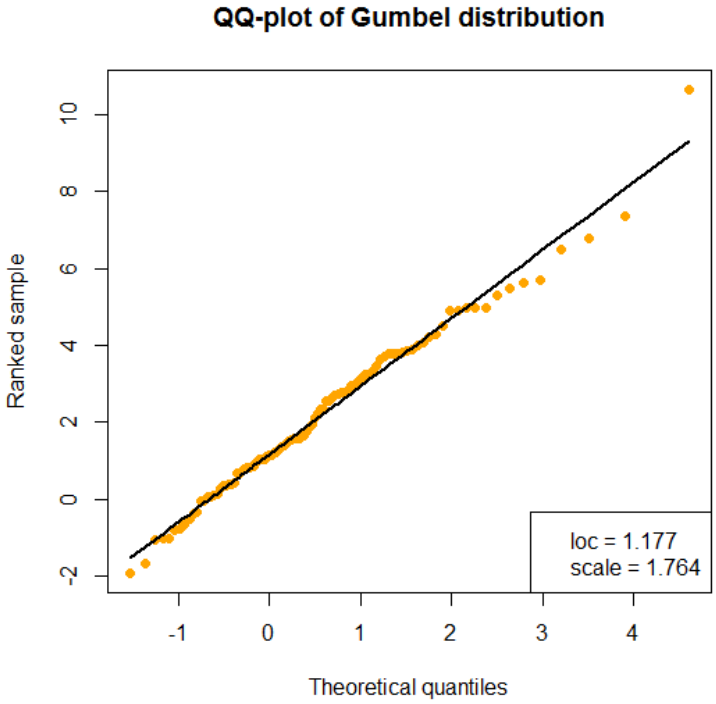


Figure 3.1: Gumbel QQ-plot

It is clear that the points fall approximately on a straight line. Furthermore, the least squares estimates are 1.177 for the intercept and 1.764 for the slope. This is close to the true parameters. The R^2 obtained for this linear model is 0.987.

Consider the case where $\gamma \neq 0$. The fact that the EVI is non-zero causes some problems. The expression is not neat as in equation (3.27) when the distribution only has a location and scale parameter. To obtain a goodness-of-fit QQ-plot, the shape parameter γ (the EVI) must first be estimated. Once an estimate for the EVI has been obtained a QQ-plot based on the GEV distribution with this value for γ is constructed. This QQ-plot should be approximately a straight line. Any deviances from a straight line indicate that the proposed model does not fit the GEV distribution well (Coles, 2000). Notice that the method of maximum likelihood requires all parameters to be estimated numerically. Therefore, the method of probability weighted moments should be used to estimate the EVI. Alternatively, as mentioned in Beirlant *et al*. (2004), a grid of values for the EVI can be considered. For each grid value, the correlation between the theoretical quantiles and the sample quantiles (with this EVI) is computed. The EVI is then set to the value that maximises this correlation. Call this estimate $\hat{\gamma}_{opt}$. It then follows that

$$x_{i,N} \approx G^{-1}_{\hat{\gamma}_{opt},\eta,\beta}\left({i}/{(N+1)} \right) = \eta + \beta \cdot G^{-1}_{\hat{\gamma}_{opt}}\left({i}/{(N+1)} \right) \text{ where} \tag{3.28.1}$$

$$G^{-1}_{\hat{\gamma}_{opt}}(u) = \frac{1}{\hat{\gamma}_{opt}}\left\{ (-\ln u)^{-\hat{\gamma}_{opt}} - 1 \right\}. \tag{3.28.2}$$

Therefore, once an optimal EVI has been estimated the same approach as for the Gumbel distribution can be followed.

As an example, consider generating 100 observations from a GEV distribution with $\eta = 1,\ \beta = 2,\ \gamma = 0.3$. For each generated value, an independent noise component from an $N(0,0.05)$ is added. A grid of values ranging from 0 to 1 (excluding zero) was constructed for the EVI. The value that maximised the correlation between the theoretical and sample quantiles was $\hat{\gamma}_{opt} = 0.34$. This is close to the true value of 0.3. Using this estimate for the EVI, the QQ-plot based on the GEV distribution is constructed. This QQ-plot is given in Figure 3.2.

It is clear that the points approximately fall on a straight line. Furthermore, the linear regression estimates of the location and scale parameters are 1.179 and 1.608, respectively, and the obtained R^2 value is 0.967. The methods described in this section are simple to use and give a visual indication of whether the GEV distribution serves as a good model. However, the QQ-plot gives no indication of the accuracy of the estimated parameters.

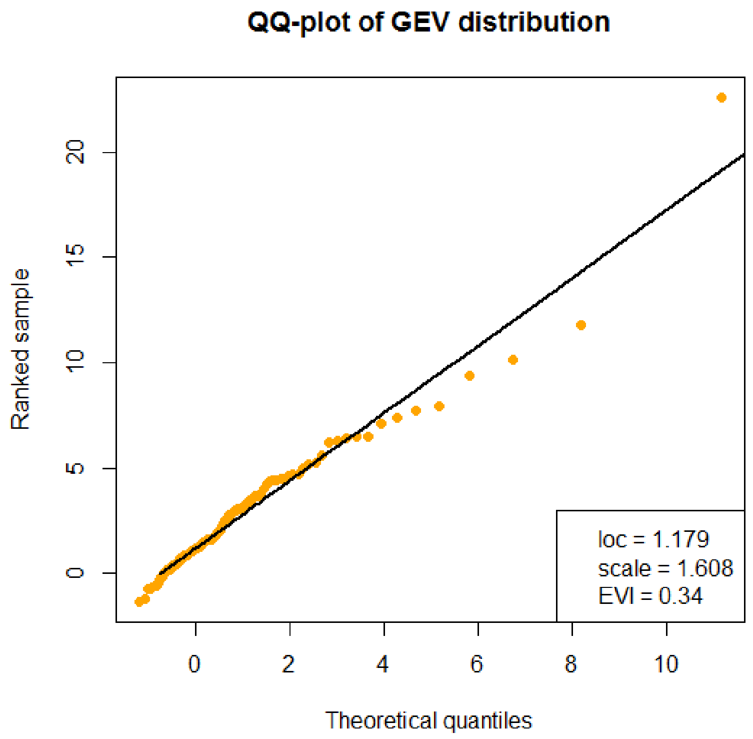


Figure 3.2: GEV QQ-plot

3.3 MODERN EXTREME VALUE THEORY

In this section, the modern approach to extreme value theory is discussed.

3.3.1 Problem statement

Recall that a disadvantage of the classical approach was the loss of a significant amount of information when only the maximum in each block is used to estimate the GEV distribution. It is believed that using more of the upper-order statistics may improve the results as more information about the structure of the tail is available. It was shown in De Haan (1970) that an alternative condition equivalently demonstrates that a distribution is in the domain of attraction of the Gumbel type (see equation (3.13)). This result was generalised to the case where $\gamma \in \mathbb{R}$ and led to the modern approach of extreme value theory. The advantage of this approach is that more of the data in the tails is used to estimate a limiting distribution. In turn, the additional information improves the accuracy of the results. Furthermore, there is full equivalence between the classical and modern approaches – both estimate a limiting distribution for the tail of some distribution.

3.3.2 The Pickands-Balkema-de Haan theorem

The distribution F is in the domain of attraction of the GEV distribution if and only if for some auxiliary function $b(\cdot)$ and for all $1+\gamma v>0$,

$$\frac{1-F\left(y+b(y)v\right)}{1-F(y)} \to (1+\gamma v)^{-1/\gamma} \text{ as } y\to x^{*}. \tag{3.29}$$

Furthermore, under (3.29),

$$\frac{b\left(y+b(y)v\right)}{b(y)} \to u^{\gamma}=1+\gamma v .$$

This result was extracted from Beirlant *et al*. (2004) and was formulated by De Haan (1970). The condition in (3.29) is denoted by $\left(C_{\gamma}^{*}\right)$. Additionally, the auxiliary function is $b(y)=a\left(U^{-1}(y)\right)$.

Consider the sequence of iid random variables $\{X_1,X_2,\ldots\}$ and define the excesses above some high threshold t as $Z=X-t$. Then,

$$P\left(Z>z\cdot b(t)\middle|X>t\right)=P\left(\frac{X-t}{b(t)}>z\middle|X>t\right)=\bar{F}_t\left(z\cdot b(t)\right)=\frac{1-F\left(t+b(t)\cdot z\right)}{1-F(t)}. \tag{3.30}$$

Hence, under $\left(C_{\gamma}^{*}\right)$, it follows that

$$\bar{F}_t(z)\sim\left(1+\gamma\frac{z}{b(t)}\right)^{-1/\gamma} \text{ for large enough } t. \tag{3.31}$$

Finally, the auxiliary function at the threshold t is interpreted as a scale parameter. Hence, the parametric form of the generalised Pareto (GP) distribution is given by

$$H(z)=\begin{cases} 1-\left(1+\gamma\dfrac{z}{\beta}\right)^{-1/\gamma} & z\in(0,\infty) \quad \gamma>0 \\ 1-\exp\left\{-\dfrac{z}{\beta}\right\} & z\in(0,\infty) \quad \gamma=0 \\ 1-\left(1+\gamma\dfrac{z}{\beta}\right)^{-1/\gamma} & z\in\left(0,-\dfrac{\beta}{\gamma}\right) \quad \gamma<0. \end{cases} \tag{3.32}$$

These results can all be found in Beirlant *et al*. (2004). Notice that again there are three types of GP distributions. Again, these types of distributions are characterised by their EVI. Although this alternative approach seems somewhat different from the classical approach, it has been shown, for example in De Haan and Ferreira (2006), that there is full equivalence between these two approaches. More specifically, as stated in De Haan and Ferreira (2006), for $\gamma\in\mathbb{R}$ statements 1 to 4 are equivalent.

1. There exist real constants $a_n>0$ and b_n such that for all x where $1+\gamma x>0$,

$$\lim_{n\to\infty} F^n\left(a_n x+b_n\right)=G_\gamma(x)=\exp\left\{-(1+\gamma x)^{-1/\gamma}\right\}.$$

2. There exists a positive function $a(\cdot)$ such that for any $u>0$,

$$\lim_{x\to\infty}\frac{U(xu)-U(x)}{a(x)}=h_\gamma(u)=\frac{u^\gamma-1}{\gamma}.$$

Furthermore, by definition $h_0(u)=\ln(u)$.

3. There exists a positive function $a(\cdot)$ such that for all x where $1+\gamma x>0$,

$$\lim_{t\to\infty} t\left\{1-F\left[a(t)x+U(t)\right]\right\}=(1+\gamma x)^{-1/\gamma}.$$

4. There exists a positive function $b(\cdot)$ such that for all x where $1+\gamma x>0$ and where $x_u=\sup\{x:F(x)<1\}$,

$$\lim_{t\to x_u}\frac{1-F\left(t+b(t)x\right)}{1-F(t)}=(1+\gamma x)^{-1/\gamma}.$$

Hence, if any one of these four statements is shown to be true, it implies the other three also hold. Notice that Statement 1 is the Fisher-Tippett theorem, Statement 2 is the domain of attraction condition, Statement 3 is another condition that can be used to show $F \in D\left(G_\gamma\right)$ and Statement 4 is part of the Pickands-Balkema-de Haan theorem.

The utilisation of the Pickands-Balkema-de Haan theorem to model the tails of a distribution will now be discussed.

3.3.3 The peaks-over-threshold method

Consider a sample $\left\{X_1, X_2, \ldots, X_N\right\}$ of the random variable X. In the previous section, it was shown that the peaks of a distribution above some high enough threshold converge in distribution to a generalised Pareto distribution. To set up the data for this method is rather simple. Let $Z_m = X_m - t \middle| X_m > t$, $m = 1, \ldots, M$ be the exceedances above some threshold. Notice that there are $M << N$ exceedances above this threshold. Furthermore, notice that the index of the original sample is adjusted such that the value Z_j might correspond with the exceedance for the sample value X_i. The sample $\left\{Z_1, Z_2, \ldots, Z_M\right\}$ is used to fit the limiting distribution. This method is known as the peaks-over-threshold (POT) method.

Up to now nothing has been said about choosing the threshold t. There is no optimal method to choose the optimal threshold. In general, the threshold is set equal to some high-order statistic in the sample. Hence, let $t = X_{N-k,N}$, $k << N$. Expressing the threshold as an upper-order statistic has the advantage that the number of exceedances above this threshold is known. Hence, there are k exceedances above the order statistic $X_{N-k,N}$ and the ordered exceedances are given by $Z_{j,k} = X_{N-k+j,N} - X_{N-k,N}$, $j = 1, 2, \ldots, k$ (Beirlant *et al.*, 2004). However, the problem of selecting a threshold has now only been expressed as a problem to select the number of order statistics to use.

The role that the threshold plays is related to the bias and variance in the estimated model. If the threshold is specified to be too large, only a few upper-order statistics will exceed this threshold. These exceedances are expected to be high in the tail which means the bias in these exceedances is low. However, estimating a model with only few observations leads to high variability. The model would be highly dependent on the sample and results obtained from this model could suffer from high variability. In contrast, if the threshold is set too low, many of the sample values will exceed it. This model would have less variability as more observations are used to estimate it. However, some of the exceedances will correspond with observations that are near the centre of the distribution. These observations are not extreme and should not be used to model the GP distribution. Hence, such a model would have large bias.

It is seen that the POT method overcomes the disadvantage associated with the block-maxima approach. Instead of only using the maximum in each block, the POT method uses all the sample values that are believed to be extreme. Furthermore, the similarity between these two approaches is the number of sample values to be used. For the block-maxima approach the number of blocks controls the bias-variance trade-off. The threshold or number of upper-order statistics to be used controls the bias-variance trade-off in the POT method. It is now discussed how the parameters of the GP distribution are estimated.

3.3.4 Parameter estimation

Consider the sample of exceedances $\{Z_1,\ldots,Z_M\}$ and assume $\gamma \neq 0$. The log-likelihood of the GP distribution for this sample is

$$\log L(\gamma,\beta) = -M\ln(\beta) - \left(\frac{1}{\gamma}+1\right)\sum_{m=1}^{M}\ln\left(1+\gamma\frac{Z_m}{\beta}\right). \tag{3.33}$$

Notice that only the sample values Z_m for which $1+\gamma\frac{Z_m}{\beta} > 0$ are valid. If $\gamma = 0$, the log-likelihood reduces to

$$\log L(\gamma,\beta) = -M\ln(\beta) - \frac{1}{\beta}\sum_{m=1}^{M}Z_m \,. \tag{3.34}$$

The method of maximum likelihood is most easily applied by first defining the vector $(\gamma,\tau)\rightarrow\left(\gamma,{}^{\gamma}\!/_{\beta}\right)$ (Beirlant *et al.*, 2004). The maximum likelihood estimator is then found by solving the equation (for $\gamma \neq 0$)

$$\frac{1}{\hat{\tau}} - \left(\frac{1}{\hat{\gamma}}+1\right)\frac{1}{M}\sum_{m=1}^{M}\frac{Z_m}{1+\hat{\tau}Z_m} = 0\text{ , where } \hat{\gamma} = \frac{1}{M}\sum_{m=1}^{M}\ln\left(1+\hat{\tau}Z_m\right). \tag{3.35}$$

Equation (3.35) must be solved numerically as no closed-form expression exists for the case where $\gamma \neq 0$. Although the parameters cannot be solved analytically, numerical techniques exist that can be used to solve this optimisation efficiently. For the case where $\gamma = 0$, equation (3.34) is a maximum where

$$\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} Z_m \; . \tag{3.36}$$

Other parameter estimation approaches for the GP distribution are the method of moments and probability-weighted moments. As mentioned in Beirlant *et al*. (2004), the EVI determines how many of the moments exist. For the r^{th} moment to exist requires that $\gamma < 1/r$. Thus, if $\gamma \geq 0.5$ the method of moments cannot be used. The mean and variance of the GP distribution are given in equation (3.37) below.

$$E[Z] = \frac{\beta}{1-\gamma} \; ; Var(Z) = \frac{\beta^2}{(1-\gamma)^2 (1-2\gamma)} \tag{3.37}$$

The mean and the variance of the exceedances are estimated with their unbiased sample statistics. These values of the statistics are then used in equation (3.37) to estimate the parameters of the GP distribution by utilising the sample moments.

Recall from Section 3.2.4 that for real p, r, s the probability-weighted moments of the random variable X with distribution function F are given by

$$M_{p,r,s} = E\left[X^p \{F(X)\}^r \{1-F(X)\}^s \right].$$

To estimate the parameters of the GP distribution the PWMs $M_{1,0,s}$, $s = 0,1$ are used. An equation for these PWMs for the GP distribution is given as

$$M_{1,0,s} = \frac{\beta}{(s+1)(s+1-\gamma)} \;, \gamma < 1. \tag{3.38}$$

Furthermore, these moments must be estimated. An unbiased estimator for this type of PWM is

$$\hat{M}_{1,0,s} = \frac{1}{M}\sum_{m=1}^{M}\left(\prod_{j=1}^{s}\frac{M-m-j+1}{M-j}\right)Z_{m,M} \ . \tag{3.39}$$

Using this estimator, the parameters of the GP distribution are estimated as

$$\hat{\gamma}_{PWM} = 2 - \frac{\hat{M}_{1,0,0}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}} \text{ and } \hat{\beta}_{PWM} = \frac{2\hat{M}_{1,0,0}\cdot\hat{M}_{1,0,1}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}} \ . \tag{3.40}$$

These results can be found in Beirlant *et al*. (2004). Although the moment and probability-weighted moment estimators are easy to compute, they have the disadvantage that they might not exist. This depends on the tail-heaviness of the underlying distribution. If $\gamma \geq 1$, none of these moment estimators can be used. Furthermore, Beirlant *et al*. (2004) mentioned that if $\gamma < 0$ some observations in the sample might exceed the estimated endpoint.

3.3.5 Goodness-of-fit evaluation

The goodness-of-fit of the generalised Pareto distribution is now discussed. Traditional goodness-of-fit hypothesis tests can be used to test whether the data follows a GP distribution. However, in this section it is discussed how a visual test based on the QQ-plot of the GP distribution can be performed.

If $\gamma = 0$, the GP distribution has a quantile function given by

$$Q(u) = -\beta\ln(1-u) \ , \beta \in \mathbb{R}, \beta > 0 \ ; u \in (0,1) \ . \tag{3.41}$$

Notice that this is an exponential distribution with parameter $\lambda = {}^{1}\!/_{\beta}$. Therefore, a QQ-plot can be constructed using the coordinates $\left(-\ln\left(1-{}^{i}\!/_{(N+1)}\right), z_{i,N}\right)$, $i = 1,\ldots,N$. If the data fits a GP distribution with $\gamma = 0$ well, the points should fall on a straight line. Furthermore, the least squares estimate of the slope serves as an estimate of β. Notice that the sample quantiles used are based on the exceedances above some high threshold.

As an example, consider simulating 100 observations from a GP distribution with $\gamma = 0$ and $\beta = 2$. To each sampled value an error from an $N(0,0.05)$ distribution is added. This ranked sample can now be compared to the theoretical quantiles of the GP distribution with $\gamma = 0$. A QQ-plot of the simulated data is given in Figure 3.3 below.

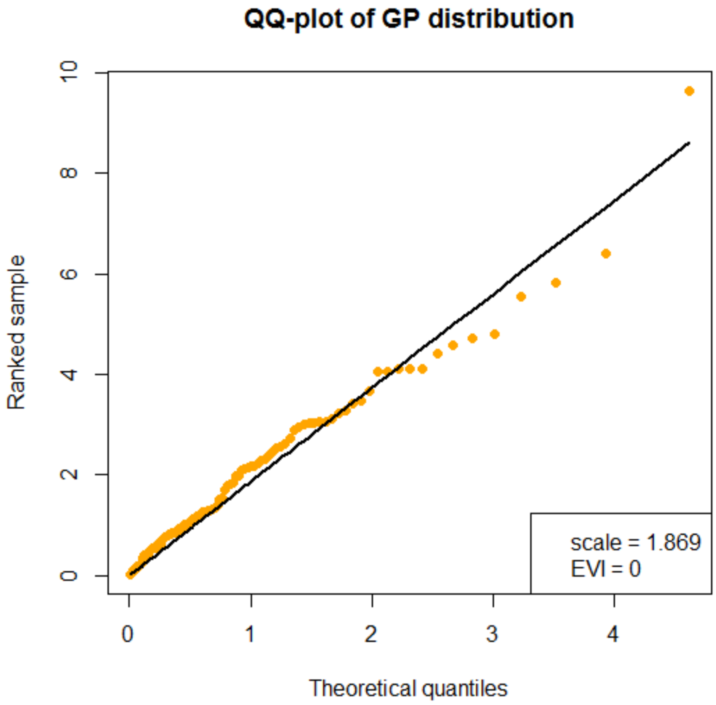


Figure 3.3: QQ-plot of GP distribution with zero EVI

Clearly, the points fall approximately on a straight line. The estimate of the slope of a linear line of best fit was estimated as 1.869. This is very close to the true scale parameter of 2. The R^2 value of the regression line is 0.989.

If $\gamma \neq 0$, the GP distribution is characterised by a quantile function of

$$Q(u) = \beta/\gamma \left\{ (1-u)^{-\gamma} - 1 \right\} \text{ , } \beta, \gamma \in \mathbb{R} \text{ , } \beta > 0 \text{ , } u \in (0,1) \text{ .} \qquad (3.42)$$

Again, the EVI must be estimated before a QQ-plot can be constructed. Therefore, the method of probability-weighted moments or the grid of values method must be used to estimate the EVI. Notice that again the maximum likelihood estimates must be found simultaneously and, therefore, this method cannot be used to only estimate the EVI. Given the estimated value of the EVI (denoted by $\hat{\gamma}_{opt}$), the QQ-plot can be constructed using the coordinates $\left(\frac{1}{\hat{\gamma}_{opt}} \left\{ \left(1 - \frac{i}{(N+1)} \right)^{-\hat{\gamma}_{opt}} - 1 \right\}, \; z_{i,N} \right)$, $i = 1, \ldots, N$. If the estimated model fits the GP distribution well, the points should fall on a straight line. Notice that once the EVI has been estimated it follows that

$$z_{i,N} \approx \beta \cdot Q_{\hat{\gamma}_{opt}}\left(\frac{i}{(N+1)} \right) \text{ where } Q_{\hat{\gamma}_{opt}}(u) = \frac{1}{\hat{\gamma}_{opt}} \left\{ \left[\log(1-u) \right]^{-\hat{\gamma}_{opt}} - 1 \right\}. \tag{3.43}$$

Thus, after the EVI (and not β) has been estimated the QQ-plot of the theoretical GP distribution and the ranked sample can be constructed. A linear regression line can then be fitted to this set of coordinates. The estimated slope of the straight line serves as an estimate of β.

Consider generating 100 observations from a GP distribution with $\beta = 2$ and $\gamma = 0.3$. A random error from an $N(0, 0.05)$ distribution is added to each observation in the sample. As a first step, the EVI must be estimated. Thus, a grid of values for the EVI ranging from 0 to 1 (excluding zero) was constructed. The value in the grid of values that maximises the correlation between the theoretical and sample quantiles is chosen as the EVI. For this example, it was found that $\hat{\gamma}_{opt} = 0.4$. This is relatively close to the true value of 0.3. Using this value for the EVI, the QQ-plot based on the theoretical quantiles of a GP distribution with $\beta = 1$ and $\gamma = \hat{\gamma}_{opt}$ is constructed. This QQ-plot is represented in Figure 3.4.

As seen in Figure 3.4, the points fall approximately on a straight line. Again, the scale parameter can be estimated by performing a linear least squares regression on the set of theoretical and sample quantile coordinates. This method produced an estimate for the scale parameter of 1.483. This is fairly close to the true value of 2. The obtained R^2 for the linear fit is 0.970.

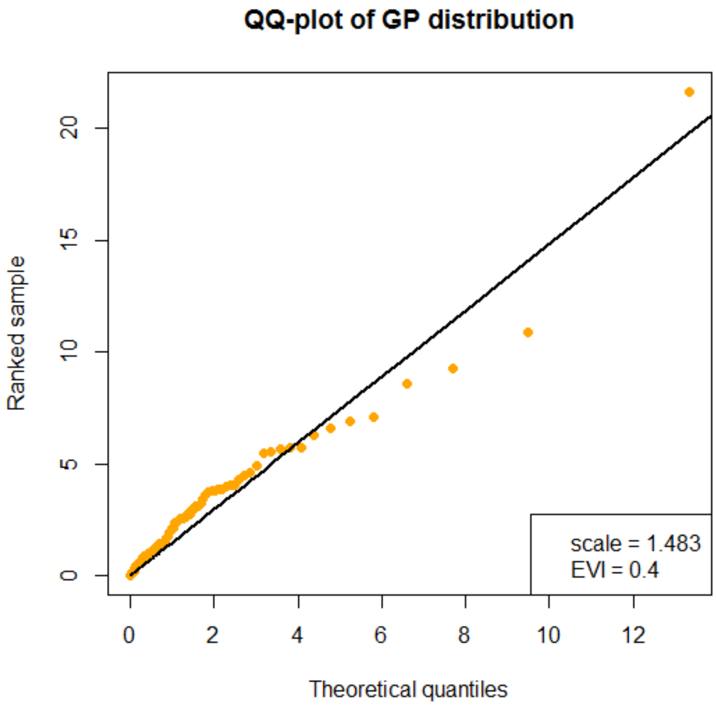


Figure 3.4: QQ-plot of GP distribution with positive EVI

This is a simple approach to validate that the GP distribution fits the peaks over a high threshold. Any deviances from a straight line indicate that the GP distribution does not fit the data well. Again, it must be mentioned that the QQ-plot does not indicate the accuracy of the estimated parameters.

3.4 THE CHALLENGE OF MULTIVARIATE EXTREME VALUE THEORY

The challenge of multivariate extreme value statistics is discussed in this section. Some problems already arise in the bivariate case. Firstly, the bivariate case is discussed. This is followed by a discussion of general considerations and problems with multivariate extremes. This section only considers the GEV distribution but similar issues arise for the GP distribution.

Consider a sample of independent pairs of random variables $\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_N,Y_N)\}$ with cumulative distribution function $F(x,y)$. A first issue is defining the block-maxima. One approach is to use component-wise block-maxima. Thus, divide the data into blocks as in the univariate case. For each block, find the maximum of X and the maximum of Y. If there are B blocks, the component-wise block-maxima for X and Y respectively are given by

$$\underline{M}_b = \left(Z_{1b}, Z_{2b}\right) \text{ , } b = 1,\ldots,B . \tag{3.44}$$

Notice that Z_{1b} and Z_{2b} are the maxima of X and Y in block b, respectively. Furthermore, the coordinate $\left(Z_{1b}, Z_{2b}\right)$ is not necessarily in the original sample.

As a first step, the univariate block-maxima of each variable is used to fit a univariate GEV distribution. Hence, the parameter vectors $\left(\hat{\gamma}_1, \hat{\beta}_1, \hat{\eta}_1\right)$ corresponding with the maximum of X and $\left(\hat{\gamma}_2, \hat{\beta}_2, \hat{\eta}_2\right)$ corresponding with the maximum of Y are obtained. Let $Z_1 = \max_{j=1,\ldots,N}\left\{X_j\right\}$ and $Z_2 = \max_{j=1,\ldots,N}\left\{Y_j\right\}$ such that

$$Z_i \sim GEV\left(\hat{\gamma}_i, \hat{\beta}_i, \hat{\eta}_i\right) \text{ , } i = 1,2 . \tag{3.45}$$

Consider the transformation

$$\breve{Z}_i = \left[1 + \hat{\gamma}_i\left(\frac{Z_i - \hat{\eta}_i}{\hat{\beta}_i}\right)\right]^{1/\hat{\gamma}_i} \text{ , } i = 1,2 . \tag{3.46}$$

Clearly, the distribution of $\breve{Z}_i$ is approximately standard Fréchet. Hence, obtain the transformed component-wise block-maxima given by $\left(\breve{z}_{1b}, \breve{z}_{2b}\right)$, $b = 1,\ldots,B$. It can be shown that if $\breve{Z}_i$, $i = 1,2$ are independent standard Fréchet distributed random variables, then

$$P\left(\frac{\breve{Z}_1}{N} \leq x, \frac{\breve{Z}_2}{N} \leq y\right) \xrightarrow{D} \exp\left\{-V\left(x,y\right)\right\} \text{ , } x > 0,\ y > 0 \text{ , } N \to \infty . \tag{3.47}$$

Furthermore, the function $V\left(x,y\right)$ is given by

$$V\left(x,y\right) = 2\int_0^1 \max\left\{\frac{\omega}{x}, \frac{1-\omega}{y}\right\} dH\left(\omega\right), \tag{3.48}$$

where $H\left(\omega\right)$ is any distribution on $\left[0,1\right]$ satisfying the mean constraint $\int_0^1 \omega dH\left(\omega\right) = 0.5$. These results can be found in Coles (2000). Hence, the limiting probability density of the component-wise block-maxima can be derived from (3.47). From this the maximum likelihood estimates can be found.

Notice that the convenient parametric distribution obtained as the limit in the univariate case is no longer valid. Instead, a semi-parametric model for the component-wise maxima arises. Specifying the distribution $H(\omega)$ is challenging. This distribution governs the quality of the obtained model. Some choices for this distribution are discussed in Coles (2000).

The analysis of the extremes of a multidimensional distribution is challenging. Of more concern is the fact that continuously improving data collecting technologies lead to a norm that datasets are naturally of higher dimension. Therefore, the theory of extreme value analysis for the multivariate case is becoming increasingly important. Fortunately, efficient methods have been proposed to estimate the dependence structure of multivariate random variables. A key piece of information in high dimensions is the dependence structure of the tail of the distribution. If this structure can be estimated accurately, important observations can be made about how extremes occur. For example, the dependence structure might indicate which variables have extremes that occur simultaneously and which variables have lower values (at these extremes). Consequently, a more comprehensive picture of the dependence in the tails is sketched (Goix, Sabourin & Clémençon, 2016).

Focusing on novelty detection, an increase in the dimension of the problem also makes the estimation procedure more difficult. As will be discussed in Chapters 4 and 5, a variety of methods have been proposed to reformulate a multidimensional novelty detection problem as an equivalent univariate problem. Nevertheless, these approaches are not without challenges. For example, say the multidimensional predictor space is mapped onto a univariate space by calculating the Euclidean distance for each observation from the mean in the normal class. Although the problem reduces to a univariate problem, the curse of dimensionality has not been dealt with properly. It might still be the case that some predictors carry no information regarding the detection of anomalies. The inclusion of these predictors (in the Euclidean distance) only increases variability in the model.

As discussed in this section, high-dimensional data poses challenges that are not apparent in a low-dimensional setting. Developing models to handle high dimensions is challenging, but efficient methods have been proposed. The most popular are algorithms that automatically perform variable selection or regularisation.

3.5 RELATING EXTREME VALUE THEORY TO NOVELTY DETECTION

Novelty detection is the process of finding observations that do not conform to the normal class or expected behaviour. It is believed that these events occur only infrequently. Furthermore, the novel events should be far from the normal class in the sense that there is a significant difference between the novel observation and the normal class.

As discussed in this chapter, extreme value theory provides the basis for analysing the extreme observations of a random event. It has been shown that two equivalent methods can be used to model the tails of a random variable. Intuitively, it seems that extreme value theory will be useful for novelty detection. As mentioned by Coles (2000), the extreme values of a variable are sparse. Hence, extreme events occur infrequently such that a sample only has a few, if any, extreme events. This is in line with the problem of detecting novelties. Since these novel observations are also expected to occur infrequently, there is an imbalance between the classes. Additionally, it is believed that novel events should be near the boundary of some predictor variable. Observations in the boundary of the variables imply observations with low density in terms of the trained model. Therefore, the theory of extremes provides a theoretical basis for novelty detection.

In Chapters 4 and 5 it will be shown that the theory of extremes can be reformulated to perform novelty detection in cases where the probability of a novel event does not necessarily correspond with an extreme value in the sample. For example, if the underlying normal class is bimodal, there are low density observations between the two modes. Extreme value theory can still be used in such settings. These methods will be discussed in Chapters 4 and 5.

On the other hand, extreme value theory can be used to threshold novelty scores. Say any algorithm is used to produce a novelty score associated with each observation in the sample. Thresholding these scores is a difficult task and generally requires an additional dataset. To our advantage, extreme value theory can be applied to these scores to threshold them in a probabilistic manner. In turn, the scores are translated to estimates of the probability that an observation is in the normal class. Consequently, using the complement of this probability, the probability that an observation is novel is recovered. This approach has the advantages that a probabilistic estimate is obtained and that no additional dataset is required to determine the threshold.

As mentioned by Clifton *et al*. (2011), extreme value theory provides a threshold that does not require updating when the sample size changes. If one were to set a threshold on the original estimated distribution of the normal class, the threshold must be adjusted if more than one observation is checked. As more observations are investigated to exceed a threshold, the probability of exceeding the threshold increases. However, extreme value theory incorporates this adaptive threshold in its solution. Therefore, the threshold is assumed constant irrespective of the number of observations checked. This is a major advantage of using extreme value theory to determine a threshold for detecting novel events.

Although only mentioned here, extreme value theory provides an intuitive method for novelty detection. Furthermore, the estimates have probabilistic meaning. In turn, the confidence of the resulting classifications can be quantified. Furthermore, the probabilistic nature of the classifier makes it possible to probabilistically infer results. For example, using one of the limiting distributions described in this chapter, confidence intervals can be constructed for the probabilities. The next chapter introduces these results.

3.6 CONCLUSION

The aim of this chapter was to explore extreme value theory. Conventionally, extreme value theory has been used to model the tails of distributions for other reasons than novelty detection. A cardinal application of extreme value theory is to predict an n-year return level. For example, the annual maximum flood level of a dam or river is taken as the block-maximum of that year. The GEV distribution is fitted to these block-maxima. In turn, the tail-quantile function of the maxima at n, $U(n)$ then represents the n-year return level of the expected maximum flood level over an n-year period. Other applications of extreme value theory are discussed in Beirlant *et al*. (2004).

It was seen that, broadly, extreme value theory is divided into the classical and modern approaches. The classical approach follows the same reasoning as the central limit theorem but the focus is on the maximum (or minimum) of a sample. Decades later, De Haan (1970) proposed an approach now known as the POT method. In general, the POT method has the advantage that more data points are used such that the obtained estimates are more accurate. However, in some cases the data naturally calls for the block-maxima approach. This would be the case if, say, only the annual maximum flood level is known. The data is therefore already in the form of block-maxima with each block representing one year.

Recently, novelty detection underpinned by extreme value theory has been proposed. This approach has some advantages over other probabilistic approaches. One advantage is the fact that this approach does not require an additional dataset to determine a novelty threshold. Furthermore, the results have probabilistic meaning. However, no approach is without some disadvantages. One of the main obstacles of using the theory of extremes for novelty detection is the difficulty of multivariate extremes. It is difficult to model the tails of high-dimensional distributions accurately. If the challenges associated with multivariate extreme value theory are not confronted, the resulting estimates will be inaccurate. Consequently, the novelty detection algorithm will produce results that cannot be trusted. Fortunately, methods to deal with high-dimensional extremes for conventional extreme value theory and novelty detection have been proposed.

The next chapter introduces a first approach to novelty detection through the theory of extremes. Chapter 5 will then use the findings from Chapter 4 to discuss more sophisticated approaches to novelty detection underpinned by extreme value theory. All the novelty detection approaches discussed in Chapters 4 and 5 rely on a transformation that maps the potentially multidimensional data onto a univariate space. In turn, the results of this chapter can directly be applied to the transformed data.

CHAPTER 4

NOVELTY DETECTION WITH UNIVARIATE EXTREME VALUE THEORY I

4.1 INTRODUCTION

Novelty detection is an approach used to detect if new observations differ significantly from the estimated probability generating mechanism. It is assumed that the process under investigation is in a normal state most of the time. Hence, a model is built to describe this state of the system. Next, new observations are tested against the normal model and, consequently, labelled as belonging to the normal class or being novel.

It is generally the case that only a few or none of the observations in the sample represent novelties. Therefore, a multiclass (supervised) approach cannot be considered. Even if some novel observations were present in the sample, they would probably have high variability since novelties arise for varied reasons. Hence, a one-class classification approach must be followed. Furthermore, novelty detection is generally the only option in high-integrity systems because the detection of novel observations in such systems does not occur frequently and when they do, they have major consequences (Clifton *et al*., 2011). Novelty detection in high-integrity systems can lead to the timely discovery of jet-engine failure, power plant failure and the deterioration of vital signs which could indicate a critical condition in a patient.

This chapter introduces the concept of performing novelty detection with univariate extreme value theory. Recall from Chapter 2 that there are many ways to construct a model representing the normal class. One such method is the probabilistic approach. Extreme value theory for novelty detection falls under this class of methods for novelty detection. A probabilistic approach therefore has many advantages as it gives a probabilistic interpretation of the confidence in classifying a new observation.

The next section of this chapter gives a simple example of the conventional method of setting a novelty threshold for the normal class. It is motivated why the extreme value distributions are best to use when a threshold is set on the distribution of the normal state. In Section 4.3 some limitations of the use of traditional extreme value theory for novelty detection are discussed. These disadvantages point to the need to adjust conventional techniques of extreme value theory. Gaussian mixture models play an important role in current research involving extreme value-based novelty detection. Therefore, an overview of this class of models is given in Section 4.4. Section 4.5 introduces the first extreme value-based novelty detection approach. The theoretical justification and the pitfalls of this model are discussed. Furthermore, a small simulation study is conducted to investigate the rate of convergence of the proposed model. It is argued that although this model is not suitable for any dataset, the concepts from it are essential for a more advanced model, as discussed in Chapter 5. Chapter 4 is concluded with other concerns that have caused this model to break down.

4.2 CONVENTIONAL THRESHOLD METHODS

In this section, a simple univariate problem is considered. The goal is to show intuitively what type of problems arises when conventional methods are used to threshold the distribution describing the normal class. An extreme value-based approach is then given as an alternative to overcome the shortcomings faced by conventional threshold methods.

Assume that a novelty detection problem can be described by a univariate random variable. Furthermore, let this random variable be unimodal and continuous. Hence, it is assumed that the random variable X describing the normal class is univariate, unimodal and continuous with distribution function denoted by $F(x)$. This distribution is used to classify some new observation x_{new} as normal or novel. One approach is to construct the hypothesis

H_0 : x_{new} is an observation from F.

This hypothesis must be tested at some significance level – say α. Hann (2008) used this method which leads to the rejection of the null hypothesis if $F(x_{new}) \notin \left[{}^{\alpha}\!/_{2}, 1 - {}^{\alpha}\!/_{2} \right]$. Notice that, for this example, it is implicitly assumed that novelties are extreme in terms of the distribution of the normal data. Considering this, novelty detection boils down to selecting a threshold α such that if $P(X < x_{new}) < \alpha/2$ or $P(X < x_{new}) > 1 - \alpha/2$, the test observation is classified as novel (Clifton *et al*., 2011; Hugueny, 2013).

In the past, it has been suggested to set a threshold on the density $f(x) = \tau$ such that the test observation is classified as novel if $f(x) < \tau$. However, setting a threshold in this manner has no probabilistic meaning and only represents a novelty score. It has been proposed by researchers such as Hyndman (1996) that this threshold can be treated probabilistically by considering the cumulative probability associated with the data exceeding the threshold – the data representing the normal state. Hence, if $S = \{x : f(x) \geq \tau\}$ he finds $F_S(\tau) = \int_S f(x)dx$. As mentioned by Clifton *et al*. (2011), a new observation sampled from F is then falsely classified as novel with probability $1 - F_S(\tau)$. A novelty threshold must then be set at $f(x) = \tau$ so that $F_S(\tau)$ is some high probability.

Although this method is theoretically sound for testing one observation, a problem arises when more than one sample point is considered. It makes intuitive sense that as more observations are sampled from the normal class, their extremes increase in magnitude. Thus, when a larger sample is considered it is expected that this sample will have more extreme observations than when only a few observations are sampled. Based on this, it would be wrong to set a constant threshold on the univariate distribution of the independent and identically distributed (iid) random variables in the validation set. The conventional method of using a validation set to fix a threshold is therefore dependent on the size of the validation set (Clifton *et al*., 2011). To see this, consider the example below.

Let the random variable X describing the normal class be a standard Gaussian random variable. Thus, the distribution describing the normal class is given by $F(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp\left\{-t^2/2\right\} dt$. For a sample of size N, the false positive error or α is the probability that an observation from the normal class is wrongly classified as novel. To investigate this error the distribution of the maximum and minimum of the normal distribution must be considered. In this example only the maximum is considered for illustration purposes. The distribution of the maximum is used to investigate how the upper extremes of a sample from the normal distribution change for an increasing sample size. As more data points are considered, the probability of observing extremes above the specified threshold increases quickly. Therefore, using a constant threshold method is not adequate.

Figure 4.1 shows the densities of the maximum of a Gaussian distribution for different sample sizes. It is clear that if only one observation is considered, the densities of X and the maximum of this variable are equal. However, as more observations are generated, the probability of an observation exceeding a specified threshold increases. The threshold was chosen to be $F(\tau) = 0.95$ such that an observation from the normal class is classified novel with probability $1 - F(\tau)$. Hence, $\tau = 1.645$. The threshold is the dotted, vertical line. Clearly, if $N = 1$ the density of the maximum is the same as the Gaussian distribution. However, as the sample size increases the modal density shifts to the right. Hence, the probability that a sample from the normal class exceeds the threshold increases. For example, the distribution of the maximum of 10 normal random variables is

$$F_{10}(x) = P\left(\max\{X_1,\ldots,X_{10}\} \le x\right) = F^{10}(x).$$

Thus, the probability that at least one of these 10 sample points exceeds the threshold of 0.95 is $1 - F_{10}(1.645) = 1 - (0.95)^{10} \approx 0.4$.

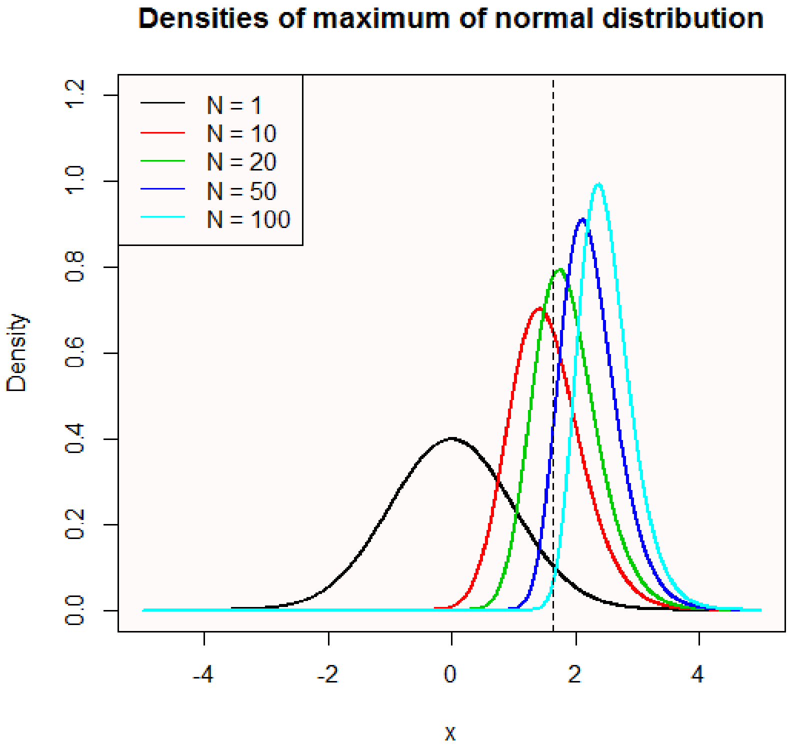


Figure 4.1: Densities of maximum for increasing sample size

Hence, there is approximately a 40% chance that at least one of the normal observations exceeds the threshold and is, consequently, wrongly classified as novel. Figure 4.2 shows the probability that the threshold is exceeded for increasing sample sizes. This graph is for the same example where the novelty threshold is set at 0.95 on the Gaussian distribution. Clearly, this error increases rapidly with the sample size.

Thus, setting a threshold based on the underlying distribution describing the normal class is only correct if one observation is considered. For example, this happens when a photo of one patient's lungs is compared to a normal model. However, as soon as more than one observation is considered the false positive rate is not constant (Clifton *et al*., 2011).

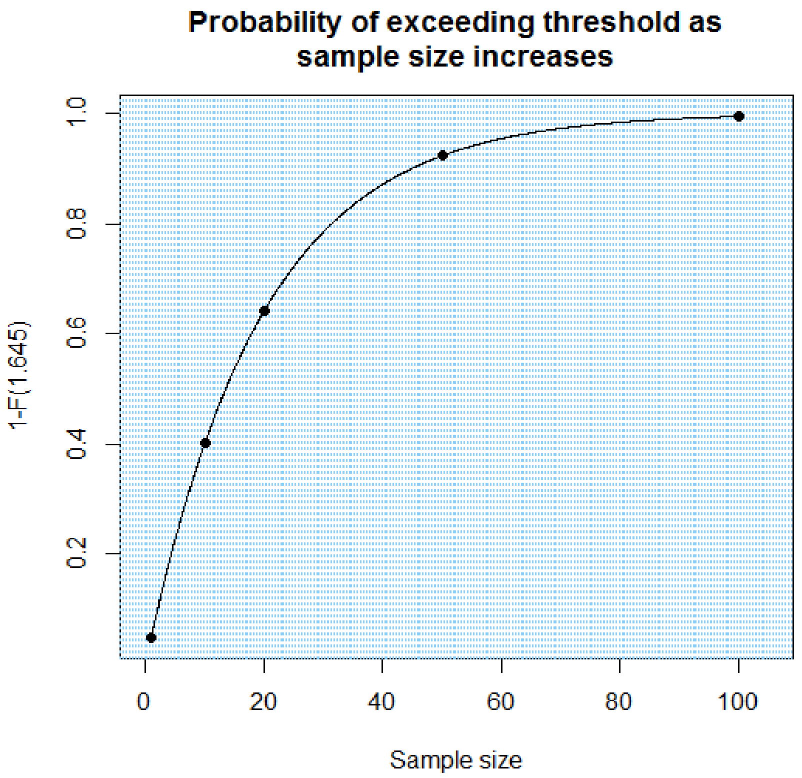


Figure 4.2: Probability of exceeding a threshold as a function of the sample size

Thus, a threshold must rather be set on the distribution of the maximum (and minimum) of the N iid random variables. As Clifton *et al*. (2011) stated, if more than one observation is considered only extreme value theory provides the correct probability distribution on which a boundary for the normal class must be set. This means that we can use the result from Chapter 3 to set a novelty threshold – that is

$$F_{X_{N,N}}(x) = P\left(\max\left\{X_1,\ldots,X_N\right\} \leq x\right) = F^N(x). \qquad (4.1)$$

The distribution in equation (4.1) has some disadvantages. Firstly, it is degenerate in the limit. For the normal example, if $N \to \infty$, the distribution for the maximum will place all its mass at $x = \infty$. Furthermore, since any distribution function takes on probabilities between $[0,1]$, the values of the distribution decrease rapidly for increasing N and become very small for large samples. These probabilities soon decrease below machine precision, making them useless for machine learning purposes (Hugueny, 2013). However, in Chapter 3 it was shown that for the correct normalising constants, the distribution of $a_N^{-1}\left(X_{N,N}-b_N\right)$ tends in distribution to the class of GEV distributions. Consequently, the class of GEV distributions can be used to set the novelty threshold.

Therefore, two distributions are useful to set a novelty threshold. If the sample size is moderately small, a threshold can be set on the exact distributions of the maximum and minimum. Conversely, for large sample sizes the limiting distribution of the maximum and minimum, which is the GEV distribution, is used to set a novelty threshold.

This example demonstrates that extreme value theory provides the correct probability distribution to threshold the normal class. In turn, new observations are tested against this threshold (determined by the extreme value distributions) and, consequently, classified as belonging to the normal class or being novel. Using this method, it is expected that the probability of misclassifying a normal instance remains constant. An important remark is that the only assumption made on how novelties occur is that anomalous events are more extreme with respect to the extremes of the data describing a normal state. If this assumption is violated, extreme value theory on its own will not succeed in discriminating between normal and novel data.

Furthermore, since only data that represents the normal state is used to train the model, the threshold must be set at a high value. Ideally, the threshold should be high enough such that none of the normal data exceeds it, but low enough such that novelties do exceed it. Extreme value theory provides the necessary tools to set a threshold such that the rate at which new observations from the normal class are classified as novelties remains constant.

4.3 LIMITATIONS OF CONVENTIONAL EXTREME VALUE THEORY

In Section 4.2 it was explained that conventional thresholding methods do not provide an accurate probabilistic interpretation. It was also argued that the theory of extremes is a viable solution to setting a boundary of normality. However, this theory on its own is not adequate. The conventional methods of extreme value theory break down when the training data becomes more complex in terms of dimensionality and modality. Therefore, this section points out the main limitations of traditional extreme value theory for novelty detection.

4.3.1 Distribution of the normal class is multimodal

Using the extreme value distribution proposed in Section 4.2 to threshold the distribution of normality makes an implicit assumption. It is assumed that novelties occur in the tails of the distribution describing the normal class. If the distribution describing the normal class (hereafter referred to as F) is unimodal, it makes sense that extremes in magnitude are improbable observations (low probability density values). However, if F has more than one mode, extremes in probability do not necessarily correspond with extremes in magnitude.

This is a major concern in novelty detection with traditional extreme value theory. It is usually the case that the data describing the normal class is complex. For example, monitoring the blood pressure of patients of different ages might result in a multimodal distribution describing the blood pressure of patients who are labelled as “normal”. This is because it is expected that the mean blood pressure of older and younger patients differs to some extent. Extreme events between these two modes therefore have no influence on the tails of the underlying distribution of normality.

Figure 4.3 shows a bimodal distribution. This distribution is a mixture of two Gaussian distributions with equal priors. The means and variances of the two normal components are $\mu_1 = 1, \mu_2 = 7$ and $\sigma_1^2 = 1, \sigma_2^2 = 2$, respectively. Hence, the density of this normal mixture is given by

$$f(x) = \frac{1}{2}\sum_{k=1}^{2} f_k\left(x, \mu_k, \sigma_k^2\right), \; x \in \mathbb{R}. \qquad (4.2)$$

Furthermore, a threshold on the normal data was set at $f(x) = 0.05$ indicated by the horizontal dashed line. The blue region in Figure 4.3 indicates the region where the density is greater than 0.05.

It is clear that the extreme value distribution of the bimodal normal distribution will detect novelties in the tails. However, at approximately 3.5 there is a local minimum of the density. Observations that fall in a range close to this value will have low probability of being sampled from the bimodal Gaussian. Unfortunately, extreme value theory does not provide the tools necessary to detect novelties falling between these two modes.

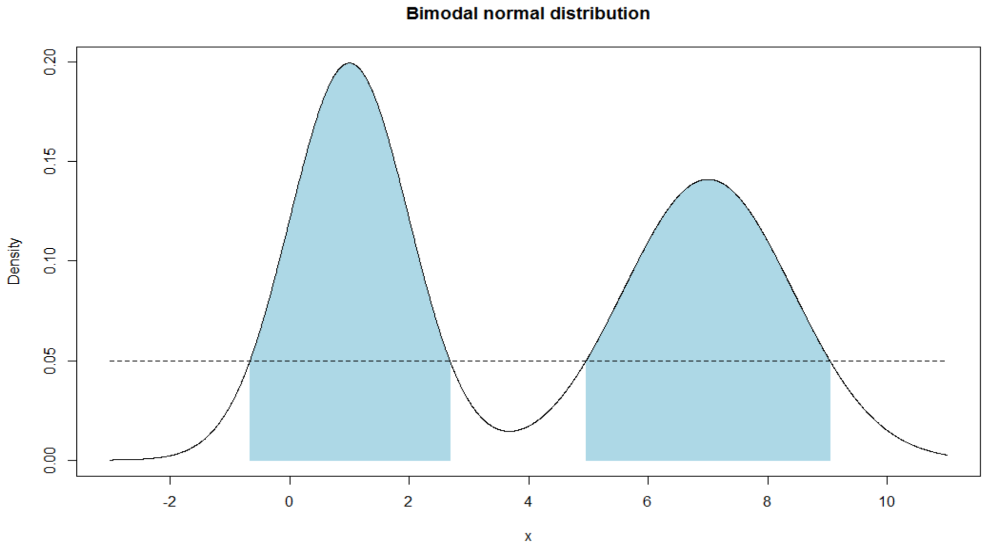


Figure 4.3: Density of bimodal Gaussian distribution

The problem of detecting novelties between modes becomes increasingly apparent as the number of modes increases. For this reason, extreme value theory must be formulated in a manner that detects extremes of low density as opposed to extremes of high or low magnitude.

4.3.2 Data of the normal class is multivariate

So far, only univariate data has been considered. It is the norm that datasets are generally of higher dimension than unity. Performing traditional extreme value theory in high dimensions is challenging. Furthermore, it is not recommended to use the results of univariate extreme value theory to build a multivariate extreme value model.

In conventional multivariate extreme value theory, the idea of component-wise maxima (minima) has been used to model the tails of multivariate distributions. This idea was mentioned in Section 3.4 for the bivariate case. It was shown that the parametric limiting distribution of the univariate case does not appear for multidimensional distributions. Rather, a semi-parametric limiting distribution was recovered. As the dimension increases above two, it becomes even more difficult to find a limiting distribution for the extremes of a sample.

It seems tempting to assume that the components are independent. The univariate extreme value approach is therefore applied to each variable and then fused. However, ignoring the covariance structure of the random variables might lead to serious errors. Such models have no method of determining which variables have extremes that occur together or which variables take on values near their centre when other variables are extreme. In novelty detection, a model of normality is constructed in multivariate space. Therefore, the extreme value distributions of each univariate variable are not adequate to identify extremes with respect to the normal model.

It must be mentioned that the theory of multivariate extremes can be used to threshold the multivariate distribution describing the normal class. As with multivariate or multimodal novelty detection, not much research has been done on using multivariate extreme value theory for novelty detection. Goix *et al*. (2016) proposed a multivariate extreme value distribution using the angular measure. As an application, anomaly detection is mentioned. A significant advantage of this method is the ability to find a sparse representation of the multivariate extreme value distribution. Hence, the problem is reduced to a lower-dimensional case which eases the interpretation of the model and adds robustness to the estimates. This method is not discussed further; interested readers are referred to Goix *et al*. (2016).

Recently, the problem of extreme value theory for novelty detection in multivariate space has been transformed to the univariate case. The univariate theory of extremes can then be utilised to threshold the distribution of normality. The first approach was proposed by Roberts (1999). Improving on the results of Roberts (1999), Clifton (2009) and Clifton *et al*. (2011) proposed an extreme value model for multivariate Gaussian distributions and a numerical method to handle multivariate, multimodal distributions of high complexity. Most of these methods rely on the use of a Gaussian mixture model to fit the data of the normal class. The next section therefore provides an overview of the Gaussian mixture model (GMM). This is followed by a discussion of the extreme value-based method of Roberts (1999). The methods of Clifton *et al*. (2011) are discussed in Chapter 5.

4.4 THE GAUSSIAN MIXTURE MODEL

The Gaussian mixture model (GMM) is a popular choice to estimate the density of complex distributions. Consider the random variable $\underline{X} \in \mathbb{R}^d$ and assume the distribution of this variable is a mixture of Gaussian distributions. The GMM consisting of K components (Gaussian distributions) representing the probability density function associated with $\underline{X}$ is then

$$f\left(\underline{x}\right) = \sum_{k=1}^{K} \alpha_k f_k\left(\underline{x}, \underline{\mu}_k, \Sigma_k\right), \ \sum_k \alpha_k = 1. \tag{4.3}$$

In equation (4.3), the vector $\underline{\mu}_k$ and matrix Σ_k are the mean and covariance structure of the k^{th} Gaussian component, respectively. Furthermore, $f_k(\cdot)$ refers to the k^{th} Gaussian component and α_k is the k^{th} mixing proportion. From a Bayesian point of view, the mixing proportions are the prior probabilities that an observation belongs to the different Gaussian densities (components). Therefore, it is expected that for a sample size N the number of observations from the k^{th} component is $N_k = \alpha_k N$.

The GMM is generally fitted by the expectation-maximisation (EM) algorithm. The steps to fit a GMM with the EM algorithm are given in Bishop (1995) and can also be found in Hastie *et al.* (2009: 277). Readers not familiar with fitting a GMM are referred to these texts. Additionally, most open source computer programs have packages that can fit GMMs in complex settings.

Usually, a penalised likelihood-based approach is used to fit the model. A variety of penalties can be imposed on the (log-) likelihood. In general, the likelihood is penalised by the number of components or free parameters in the model. This ensures that the GMM does not overfit the data and, consequently, generalises poorly.

Various valuable properties make GMMs a popular choice for density estimation. Gaussian mixture models generally approximate complex densities relatively well. Roberts (1999) stated that GMMs have the ability to approximate a density with arbitrary precision – remarking that complexity must still be controlled to cope with overfitting. Additionally, these mixture models have a closed-form analytical expression. The analytical expression of the GMM is advantageous as it eases deriving properties of the model.

It must be mentioned that GMMs are computationally slow to implement in high dimensions. If the optimal number of components in the model is unknown and needs to be estimated, the training time of the GMM becomes significantly long. Fortunately, methods have been proposed to speed up the computational time of GMMs. Verbeek, Vlassis and Kröse (2003) proposed a greedy approach where mixture components are added sequentially. At each step, the algorithm finds the optimal Gaussian component to insert. The complexity of the model can again be controlled by one of the traditional methods (Akaike information criterion, Bayesian information criterion, Minimum description length). Other similar methods to reduce the function space of the model can be found in the literature on Gaussian mixture models.

4.5 WINNER-TAKES-ALL APPROACH

This approach was formulated by Roberts (1999) and termed the winner-takes-all (WTA) method by Clifton (2009). As a first step, a GMM is fitted to the (multivariate) data. The output of the GMM is then used to construct an extreme value-based distribution that is used to threshold the normal class.

4.5.1 Description of the WTA method

Let $\underline{X}^T = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ be a random vector with distribution and density functions given by $F(\underline{x})$ and $f(\underline{x})$, respectively. It is assumed that a representative sample (data matrix) of the normal class is given by X. This matrix has dimensions $N \times d$ such that each row represents an observation (strictly from the normal state) and each column a variable. As a first step, the distribution describing the normal state must be estimated.

Roberts (1999) used a GMM to estimate the distribution of normality. Assuming there are K Gaussian components in the model, the estimated density is given by

$$\hat{f}(\underline{x}) = \sum_{k=1}^{K} \frac{\hat{\alpha}_k}{(2\pi)^{d/2}} \left|\hat{\Sigma}_k\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underline{x} - \hat{\underline{\mu}}_k)^T \hat{\Sigma}_k^{-1}(\underline{x} - \hat{\underline{\mu}}_k)\right\}, \ \sum_k \hat{\alpha}_k = 1. \tag{4.4}$$

Notice that the parameters in equation (4.4) are estimated with the EM algorithm. The main assumption made by Roberts (1999) is that the probability that a sample is a novelty is dominated by the Gaussian component closest in the Mahalanobis distance to this observation. The Mahalanobis distance metric of $\underline{x}$ to the k^{th} Gaussian component is given by

$$M(\underline{x})_k = \left[(\underline{x} - \underline{\mu}_k)^T \Sigma_k^{-1}(\underline{x} - \underline{\mu}_k)\right]^{1/2}. \tag{4.5}$$

Roberts (1999) stated that the density function of the Mahalanobis distance in terms of a Gaussian distribution is

$$f_M(M(\underline{x})) = \frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} M(\underline{x})^2\right\}, \ M(\underline{x}) \geq 0. \tag{4.6}$$

However, notice that this result is only true for the univariate case. Let the random vector $\underline{X} \in \mathbb{R}^d$ be multivariate normal with mean vector and covariance matrix $\underline{\mu}$ and Σ, respectively. It is well known that $Y = \left(\underline{X} - \underline{\mu}\right)^T \Sigma^{-1} \left(\underline{X} - \underline{\mu}\right)$ is chi-squared distributed with d degrees of freedom. Hence, for $\underline{X} \in \mathbb{R}^d$ and multivariate normal, the quadratic form Y has a density function given by

$$f_Y\left(y\right) = \frac{1}{2^{d/2}\Gamma\left(d/2\right)} y^{d/2-1} \exp\left\{-y/2\right\} \text{ , } y > 0 . \tag{4.7}$$

Clearly, for any dimension d, the density function of $M\left(\underline{X}\right)$ is recovered from the transformation $M\left(\underline{X}\right) = \sqrt{Y}$. Thus, using the Jacobian,

$$f_M\left(M\left(\underline{x}\right)\right) = \frac{2^{1-d/2}}{\Gamma\left(d/2\right)} M\left(\underline{x}\right)^{d-1} \exp\left\{-M\left(\underline{x}\right)^2 / 2\right\} \text{ , } M\left(\underline{x}\right) > 0 . \tag{4.8}$$

It is now clear that if the data is univariate – meaning $d = 1$ – the density in (4.6) is obtained. However, if $d = 2$ this density becomes

$$f_M\left(M\left(\underline{x}\right)\right) = M\left(\underline{x}\right) \cdot \exp\left\{-M\left(\underline{x}\right)^2 / 2\right\} \text{ , } M\left(\underline{x}\right) > 0,\ \underline{x} \in \mathbb{R}^2,\ M\left(\underline{x}\right) \in \mathbb{R}^+ . \tag{4.9}$$

The density in (4.9) is clearly a Rayleigh density function with $\sigma^2 = 1$. As the dimension increases above 2, the density function has no well-known tractable form. It is thus seen that the statement of Roberts (1999), namely that the Mahalanobis distances are absolutely normally distributed, is only true for the univariate case. However, this statement was only used to determine the limiting distribution of the maximum Mahalanobis distance of a multivariate normal distribution. Roberts (1999) stated that the distribution of the Mahalanobis distance of the multivariate Gaussian vector is one-sided normal. Therefore, the distribution of the Mahalanobis distance is in the maximum domain of attraction of the Gumbel class. This follows from the fact that the one-sided normal distribution belongs to this class of GEV distributions.

Under the assumption that the distribution of the maximum Mahalanobis distance (normalised) is Gumbel in the limit, Roberts (1999) further assumed that the parameters of the Gumbel distribution of the maximum Mahalanobis distance associated with the k^{th} Gaussian component can be estimated from the one-sided normal distribution. Denote the location and scale parameters of the Gumbel distribution by η and β, respectively. Roberts (1999) stated that these parameters can be estimated as

$$\hat{\eta} = \left(2\ln\hat{N}_k\right)^{1/2} - \frac{\ln\left(\ln\hat{N}_k\right) + \ln 2\pi}{2\left(2\ln\hat{N}_k\right)^{1/2}} \text{ and } \hat{\beta} = \left(2\ln\hat{N}_k\right)^{-1/2}. \tag{4.10}$$

In (4.10), the parameter $\hat{N}_k = N \cdot \hat{\alpha}_k$ is the expected number of observations generated by the k^{th} Gaussian component in the GMM. The parameter estimators in (4.10) were derived by assuming the Mahalanobis distances are one-sided normally distributed (Roberts, 1999). An analytical expression for the location and scale parameter of the Gumbel distribution can therefore be derived.

For a new observation, Roberts (1999) suggested finding the Gaussian component closest to this observation in the Mahalanobis distance. Let the closest component to a new observation $\underline{x}_{new}$ be k^*. Thus, $k^* = \underset{k}{\arg\min} M\left(x_{new}\right)_k$. The Mahalanobis distance of $\underline{x}_{new}$ with respect to the Gaussian component k^* is then $\hat{M}\left(\underline{x}_{new}\right)_{k^*}$ as in (4.5). Using the parameter estimates (dependent on the number of observations seen by the k^{th} component) in (4.10), the new observation is classified as novel if

$$1 - \hat{G}\left(\hat{M}\left(\underline{x}_{new}\right)_{k^*}, \hat{\eta}_{k^*}, \hat{\beta}_{k^*}\right) < \alpha. \tag{4.11}$$

In equation (4.11) the distribution $\hat{G}(\cdot)$ is the estimated Gumbel distribution of the maximum Mahalanobis distance corresponding with the Gaussian component k^* and α is a significance level. Notice that the quantity on the left-hand side of (4.11) is the probability that the Mahalanobis distance $\hat{M}\left(\underline{x}_{new}\right)_{k^*}$ is greater than the distribution of the Mahalanobis distances implied by the normal class. Therefore, this probability will be very close to 1 for observations deemed normal. Hence, any value for α slightly lower than, say, 0.99 is appropriate.

4.5.2 Maximum domain of attraction of the Mahalanobis distance of Gaussian vectors

Although the probability density function in (4.6) is not recovered for Gaussian distributions of any dimension, the distribution of the Mahalanobis distance of a multivariate Gaussian random vector is in the domain of attraction of the Gumbel class of GEV distributions. To prove this remark, the Von Mises' theorem for the Gumbel class is used. This result states that $F \in D(\Lambda)$ if, and only if,

$$\lim_{x\to\infty} \frac{\bar{F}(x)F^{\prime\prime}(x)}{f^2(x)} = -1. \qquad (4.12)$$

Notice that $\bar{F} = 1 - F$. Consider the probability density function in (4.8) with $K_d = \frac{2^{1-d/2}}{\Gamma(d/2)}$ and $M(\underline{x}) = u$. The derivative of this function is

$$f_M^{\prime}(u) = K_d\left[(d-1)u^{d-2} - u^d\right]\exp\left\{-u^2/2\right\} = f_M(u)u^{-1}\left((d-1)-u^2\right) = u\cdot f_M(u)\left(\frac{d-1}{u^2}-1\right).$$

However, as $u \to \infty$ the term $(d-1)/u^2 \to 0$. Hence,

$$f_M^{\prime}(u) \approx -u\cdot f_M(u) \text{ and } u\cdot f_M^{\prime}(u) = f_M(u)\left((d-1)-u^2\right). \qquad (4.13)$$

Returning to (4.12) it follows that

$$R(u) \equiv \frac{\bar{F}_M(u)F_M^{\prime\prime}(u)}{f_M^2(u)} = \frac{\bar{F}_M(u)f_M^{\prime}(u)}{f_M^2(u)} = \frac{\bar{F}_M(u)\cdot u\cdot\left((d-1)/u^2 - 1\right)}{f_M(u)}. \qquad (4.14)$$

The numerator and denominator in (4.14) both converge to zero as $u \to \infty$. Therefore, L'Hospital's rule can be applied. This gives (in the limit)

$$R(u) \approx -\frac{u\cdot\bar{F}_M(u)}{f_M(u)} \overset{L'Hospital}{\approx} -\frac{\bar{F}_M(u) - u\cdot f_M(u)}{f_M^{\prime}(u)} = -\frac{\bar{F}_M(u) - u\cdot f_M(u)}{u\cdot f_M(u)\left((d-1)/u^2 - 1\right)} \approx \frac{\bar{F}_M(u)}{u\cdot f_M(u)} - 1.$$

Again, the numerator and denominator of $\bar{F}_M(u)/(u\cdot f_M(u))$ converge to zero so that L'Hospital's rule can be applied. Hence, in the limit,

$$\frac{\bar{F}_M(u)}{u\cdot f_M(u)} \approx -\frac{f_M(u)}{f_M(u) + u\cdot f_M^{\prime}(u)} = -\frac{f_M(u)}{f_M(u) + f_M(u)\left((d-1)-u^2\right)} = -\frac{1}{d-u^2} \to 0. \qquad (4.15)$$

Hence,

$$R(u) \to 0 - 1 = -1 \text{ as } u \to \infty . \tag{4.16}$$

This proves that the distribution of the maximum Mahalanobis distance of a d -dimensional Gaussian distribution converges in distribution to the Gumbel class of GEV distributions.

4.5.3 Rate of convergence of the maximum Mahalanobis distance GEV distribution

It is now investigated how fast the distribution of the maximum Mahalanobis distance converges to the Gumbel distribution. Furthermore, the parameter estimators proposed by Roberts (1999) are checked. Consider generating B samples of size N from a multivariate Gaussian distribution of dimension d . For each sample, the N Gaussian vectors are mapped onto the Mahalanobis radius and the maximum of these distances is stored. Let this maximum distance be $M_j = \max_{i=1,\ldots,N}\left\{M\left(\underline{x}_{ij}\right)\right\}$, $j = 1,\ldots,B$ where $M\left(\underline{x}_{ij}\right)$ is the i^{th} Mahalanobis distance of the j^{th} sample. The sample $\left\{M_j\right\}_{j=1}^{B}$ should be approximately Gumbel distributed if the sample size is large enough. Furthermore, it must be determined whether the parameter estimates in (4.10) are adequate for dimensions higher than 1.

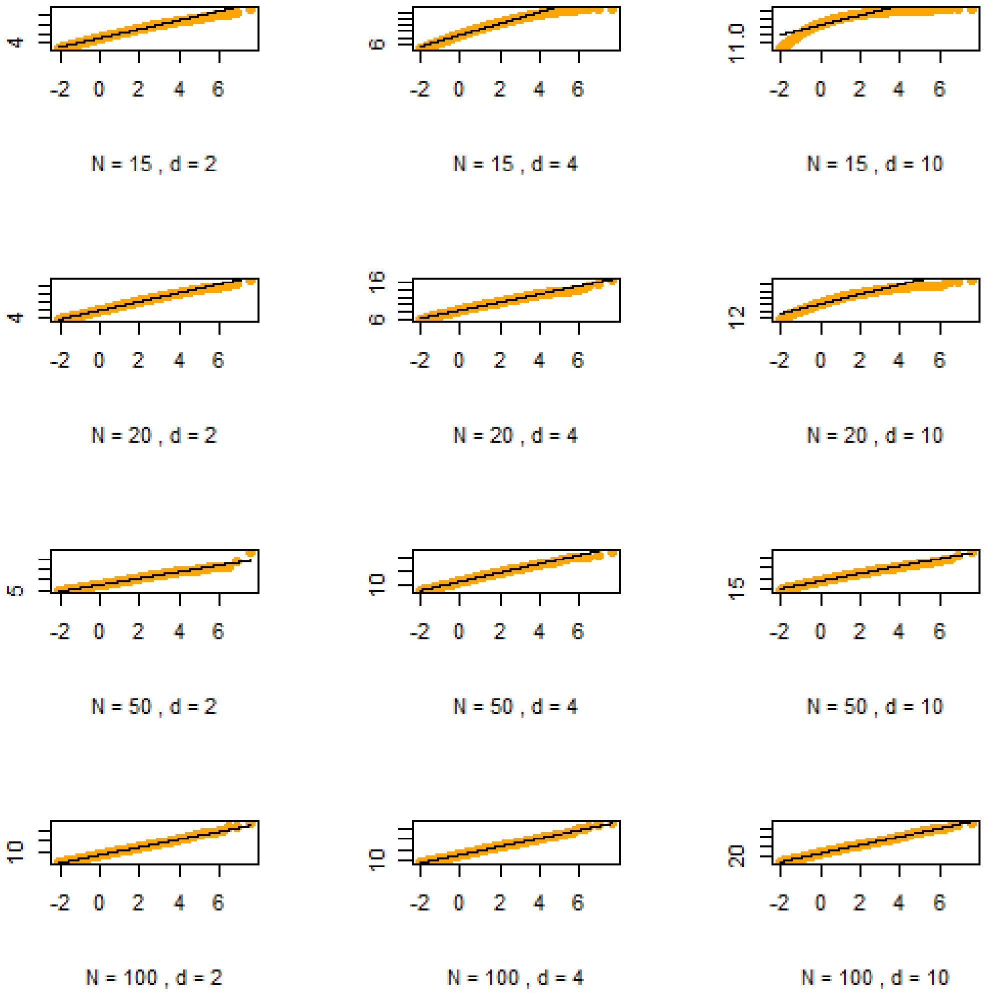
After performing a simulation for different mean vectors, covariance structures and dimensions it was seen that the distribution of the maximum Mahalanobis distances of a multivariate Gaussian distribution converges to a Gumbel distribution relatively fast. If the dimension is low, a small sample of maxima is adequate to model the distribution of the maximum Mahalanobis distance with a Gumbel distribution. However, for high-dimensional data a much larger sample is required. Nevertheless, for a sample size of $N \geq 50$ the Gumbel distribution fits the block-maxima of Mahalanobis distances well for higher-dimensional data. It is noted that the parameter estimates obtained from the univariate one-sided Gaussian distribution are not accurate estimates of the location and scale of the Gumbel distribution.

Figures 4.4 and 4.5 show the QQ-plots and histograms with the fitted density function superimposed, respectively. Three multivariate normal densities were generated, and their properties are given by

$\underline{\mu}_1^T = (0,0)$, $\underline{\mu}_2^T = (-3,2,0,-1)$, $\underline{\mu}_3^T = (0,0,0,0,1,0,0,0,0,0)$ and

$$\Sigma_1 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 2 & 0.5 & 1 & 0.2 \\ 0.5 & 1 & 0.4 & 0.1 \\ 1 & 0.4 & 2 & 0.6 \\ 0.2 & 0.1 & 0.6 & 1 \end{bmatrix},\ \Sigma_3 = \mathrm{I}_{10}\,.$$

The sample size and the dimension are given below each graph. Furthermore, the density superimposed on the histograms was constructed using the maximum likelihood estimates of the Gumbel distribution. The linear function sketched on the QQ-plots was obtained by fitting a linear regression model on the order statistics of the maximum Mahalanobis distances using the theoretical quantiles of the Gumbel distribution.



N = 15 , d = 2

N = 15 , d = 4

N = 15 , d = 10

N = 20 , d = 2

N = 20 , d = 4

N = 20 , d = 10

N = 50 , d = 2

N = 50 , d = 4

N = 50 , d = 10

N = 100 , d = 2

N = 100 , d = 4

N = 100 , d = 10

Figure 4.4: QQ-plots of Mahalanobis maxima

Figure 4.4 shows that for $N = 15$ none of the QQ-plots are straight lines. Therefore, such a sample will be too small to fit a Gumbel distribution to the block-maxima of Mahalanobis distances. For $N = 20$ only the QQ-plot with $d = 2$ is roughly a straight line. Fortunately, when $N \geq 50$ the QQ-plots are all straight lines. Similar conclusions can be drawn from Figure 4.5. In this figure, the density superimposed on the histograms is a Gumbel density with the location and scale parameter estimated by maximum likelihood. Hence, it is concluded that the distribution of the maximum Mahalanobis distance of a multivariate Gaussian converges relatively fast to a Gumbel distribution. However, the rate of convergence is strongly dependent on the dimension.

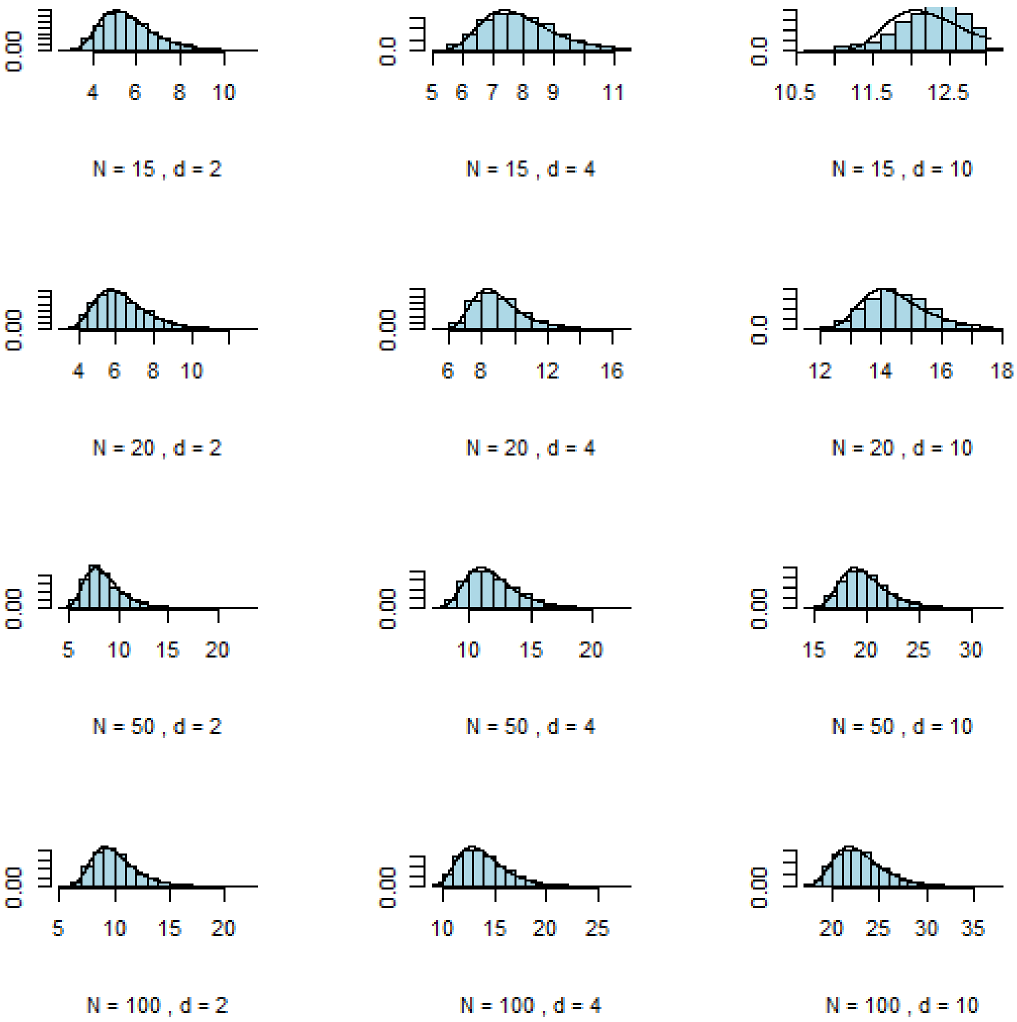


Figure 4.5: Histograms of Mahalanobis maxima

Table 4.1 provides the statistics obtained from the simulation. These are the estimates of the location and scale parameters of the Gumbel distribution denoted by η and β, respectively. The subscript indicates whether the parameters were estimated using the approach of Roberts (1999), maximum likelihood estimation or the linear regression approach. Clearly, the estimators proposed by Roberts (1999) underestimated the true parameters. However, the estimates obtained through maximum likelihood estimation and the regression technique are very close to each other. Therefore, one of these methods should be used as opposed to the estimators in (4.10).

It is concluded that the assumption that the parameters of the Gumbel distribution in the WTA approach can be estimated with equation (4.10) is not correct. Clifton (2009) and Clifton *et al*. (2011) also pointed out that these estimates underestimate the true parameters as the dimension increases. As a result, the method of Roberts (1999) cannot be used appropriately without estimating the parameters by means of some other method.

Although the parameter estimators of Roberts (1999) do not fit the data well for high dimensions, this approach is a first approach that transforms multivariate data to the univariate case and, consequently, models the (transformed) data with univariate extreme value theory. Should the sample size allow this, the parameters of the Gumbel distribution should be estimated via block-maxima using the maximum likelihood or probability-weighted moment estimators. The Gumbel distribution will then be an accurate approximation of the maximum Mahalanobis distance of a multivariate Gaussian distribution. However, the multivariate Gaussian of the closest component approximates the global GMM fitted to the normal model. Unfortunately, the approach of Roberts (1999) has more concerning issues, which makes it inappropriate to use this model for complex GMMs.

Table 4.1: Statistics of simulation

N	d	$\hat{\eta}_{ROB}$	$\hat{\eta}_{MLE}$	$\hat{\eta}_{REG}$	$\hat{\beta}_{ROB}$	$\hat{\beta}_{MLE}$	$\hat{\beta}_{REG}$
15	2	4.081	5.120	5.139	0.220	1.042	1.001
15	4	4.081	7.376	7.413	0.220	1.003	0.897
15	10	4.081	12.060	12.105	0.220	0.459	0.300
20	2	4.148	5.704	5.716	0.217	1.146	1.119
20	4	4.148	8.302	8.338	0.217	1.218	1.111
20	10	4.148	14.040	14.088	0.217	0.898	0.729
50	2	4.352	7.600	7.590	0.208	1.513	1.553
50	4	4.352	10.962	10.971	0.208	1.671	1.648
50	10	4.352	18.769	18.806	0.208	1.843	1.702
100	2	4.502	9.178	9.162	0.202	1.740	1.805
100	4	4.502	12.856	12.866	0.202	1.903	1.877
100	10	4.502	21.796	21.828	0.202	2.197	2.090

4.5.4 Other concerns with the WTA method

Given that the GMM fits the normal data well, it is expected that this method would be sufficient to detect novelties. However, Clifton *et al*. (2011) pointed out a much bigger problem. If the covariances of the components differ significantly or if the components overlap severely, the main assumption made by Roberts (1999) is violated. That is, extreme observations cannot solely be described by the Gaussian component closest in the Mahalanobis distance. The other components also influence the obtained statistics.

Thus, this method (using any parameter estimates) fails to correctly discriminate between novelties and normal data in complex settings. Other components also carry information regarding the extremeness of a point. Therefore, ignoring these contributions distorts the results.

It is for this reason that Clifton *et al.* (2011) proposed an approach based on modelling the minimum density. This approach overcomes the issue of the influence of different components as it uses the contributions of each Gaussian component in the GMM. In the light of modelling the minimum density, they proposed a theoretical framework for the multivariate Gaussian distributions and a numerical scheme for a GMM. These two methods are discussed in the next chapter.

Although the method of Roberts (1999) has some shortcomings, it is a first model relying on extreme value theory for novelty detection. Therefore, the ideas used to construct this model have led to more sophisticated novelty detection algorithms utilising extreme value theory. Furthermore, in cases where the distribution of the data is in fact approximated well by a multivariate Gaussian distribution this approach should be efficient to detect novelties.

4.6 CONCLUSION

This chapter introduced the concept of using extreme value theory to perform novelty detection. It was explained why the conventional threshold methods fail by setting a constant threshold on the data describing the normal class. Using this approach, the threshold must be updated when the sample size changes. It is therefore difficult to apply the final model to datasets of varied sizes describing the same normal state.

To overcome the disadvantage of the conventional threshold method, it was argued that the distribution of the minimum and maximum provides the correct model to threshold the normal data. The location and scale of this distribution are automatically adjusted for different sample sizes. However, using this model solely is not adequate. Traditional extreme value theory is difficult to apply in high-dimensional cases. Furthermore, in multimodal scenarios traditional extreme value theory fails to detect novelties that occur between two or more modes.

Roberts (1999) was first to propose a model based on extreme value theory to detect novelties. This model is strongly dependent on the assumption that the probability of an observation being novel is dominated by the Gaussian component closest in the Mahalanobis distance to this observation. The Mahalanobis distances of each observation were therefore used to construct a novelty detection model. The attractiveness of this approach is the way in which the data was seamlessly mapped onto a univariate space. Consequently, the theory of univariate extremes can be applied to the data.

Unfortunately, Clifton *et al*. (2011) remarked that the model of Roberts (1999) breaks down in two general scenarios. The first problem arises when the covariances of the Gaussian components differ significantly. Secondly, if the components overlap significantly the proposed Gumbel distribution also fits the Mahalanobis distance block-maxima poorly. Hence, the assumption that only the Gaussian component closest in the Mahalanobis distance governs the probability of a sample being novel is not necessarily satisfied. For this reason, Clifton (2009) and Clifton *et al*. (2011) proposed an alternative method for thresholding the normal data using extreme value theory. These approaches are discussed in the next chapter.

CHAPTER 5

NOVELTY DETECTION WITH UNIVARIATE EXTREME VALUE THEORY II

5.1 INTRODUCTION

The previous chapter introduced the use of extreme value theory to perform novelty detection. Although the theoretical justification of the model is strong, the main assumption made by Roberts (1999) breaks down in complex settings. That is, if a more complex distribution than a multivariate Gaussian distribution is required to describe the extremes of the normal data, this approach fails to accurately detect novel observations.

Clifton (2009) and Clifton *et al*. (2011) extended the research of Roberts (1999) and proposed two methods to perform extreme value-based novelty detection. Both these methods consider extreme value theory in terms of modelling the areas where the probability density function of the data describing the normal class is low. Therefore, extreme value theory is first redefined in terms of low probability density. This definition of extreme value theory is then exploited to derive an analytical and numerical approach for novelty detection with extreme value theory.

Recall from Chapter 3 the two general approaches to modelling extreme values, namely the classical approach and modern approach. The models proposed by Clifton (2009) and Clifton *et al*. (2011) consider the classical approach. Hence, the goal is to derive a limiting distribution for the minimum or maximum of some underlying distribution. More recently, the modern approach of extreme value theory has been used to build the same models. As will be discussed, the theory of both approaches is similar.

In the next section, extreme value theory is redefined in terms of the data points where the probability density function of the underlying distribution is a minimum. In order to find the distribution of the minimum probability density, the distribution of the probability density values must first be found. Therefore, the distribution of the probability density function is defined. It is then shown how these definitions are used to find analytical expressions for the distribution of the probability density function and minimum probability density function. Consequently, an analytical method for performing novelty detection with extreme value theory is constructed. Section 5.3 formulates the numerical scheme proposed by Clifton (2009) and Clifton *et al*. (2011). This approach is suitable for datasets of high dimensions and multimodality. It is shown that the distribution of the minimum probability density values is a transformation of the equiprobable contours of the underlying distribution. This result is used to derive a numerical scheme to perform novelty detection. The disadvantage of this method is that it requires the generation of data from a GMM. Therefore, Section 5.4 discusses proposals to improve the computational time as well as the theoretical justification of the model. The final section of this chapter considers extreme value-based novelty detection with the modern approach.

5.2 DISTRIBUTION OF THE MINIMUM DENSITY

The previous chapter unpacked the need to redefine the extremes of a sample for novelty detection. Traditional extreme value theory cannot be applied in multimodal cases and is difficult to apply in high-dimensional situations. For this reason, Clifton *et al*. (2011) considered extremes in terms of minimum density as opposed to extremes in magnitude.

Let $\left\{\underline{X}_i\right\}_{i=1}^{N}$, $\underline{X}_i \in \mathbb{R}^d$ be a sequence of iid random vectors with distribution and density functions F and f , respectively. The extremum of this sequence is defined as

$$E_N = \underset{\underline{X}_i\ ,\ i=1,\ldots,N}{\arg\min}\left\{f\left(\underline{X}_i\right)\right\}. \tag{5.1}$$

This definition of extremes is a generalisation of the traditional definition. Given that the distribution is unimodal, extremes in magnitude correspond with extremes in density. As observations move away from their mode, the values of the density function decrease monotonically. Furthermore, any density function has the property that $f\left(\cdot\right) \in \mathbb{R}^+$. Hence, this definition reduces a multivariate problem to the univariate case. Therefore, if a non-degenerate distribution over the density values $f\left(\underline{x}\right)$ can be found, the results of univariate extreme value theory can be utilised (Clifton *et al*., 2011).

5.2.1 Distribution of the density function

Before the distribution of E_N can be obtained, the distribution of the density function must be defined. Let $\underline{X} \in \mathrm{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$ be a random vector with a corresponding density function $f\left(\underline{x}\right) \in P_f$. Furthermore, consider the transformation $Y = f\left(\underline{X}\right)$. Then, for all $y \in P_f$, the distribution function of $Y = f\left(\underline{X}\right)$ is defined as

$$G_d\left(y, f\right) \equiv P\left(f\left(\underline{X}\right) \leq y\right) = \int_{f^{-1}\left(]0, y]\right)} f\left(\underline{x}\right) d\underline{x}\,. \tag{5.2}$$

In equation (5.2) the integral is over $f^{-1}(]0,y]) = \{\underline{x} : f(\underline{x}) \leq y\}$. Under this definition, $f(E_N)$ is the minimum of a sequence of N random variables of $Y = f(\underline{X})$ (Hugueny, 2013). This follows from the fact that

$$f(E_N) = f\left(\underset{\underline{X}_i\ ,\ i=1,\ldots,N}{\operatorname{argmin}}\{f(\underline{X}_i)\}\right) = \min_{i=1,\ldots,N}\{f(\underline{X}_i)\} = \min_{i=1,\ldots,N}\{Y_i\}.$$

Hence, the distribution of $f(E_N)$ is the extreme value distribution of the minimum probability density function of a random vector $\underline{X} \in \mathrm{X}$ (Hugueny, 2013). Given the fact that a probability density function is always positive, it has a minimum lower bound of zero. Therefore, the distribution of the probability density values is assumed to be bounded. Thus, it is expected that the minimal extreme value distribution of $G_d(y,f)$ can only be in the domain of attraction of the extremal Weibull class.

This new definition of extremes can be utilised to perform novelty detection. As a first step, the distribution over the probability density values of the normal class is obtained. This distribution is denoted by $G_d(\cdot,f)$. Once this distribution is obtained, the distribution of the minimum probability density values can be derived. Let the probability distribution function of $f(E_N)$ be denoted by $P(f(E_N) \leq y) \equiv G_d^{\min}(y,f)$. For a new (transformed) observation $y^* = f(\underline{x}^*)$, the probability that the minimum density of $\underline{X}$ is less than y^* is given by $G_d^{\min}(y^*,f)$. This distribution function therefore estimates how far into the tails new observations lie in terms of the normal class.

One vital step that has not been mentioned is how to derive the form of the distribution $G_d^{\min}(\cdot,f)$. This is now discussed for the multivariate Gaussian distribution.

5.2.2 Distribution of the density function for the multivariate Gaussian

Let $\underline{X}$ be a random vector distributed as a multivariate Gaussian distribution with mean vector $\underline{\mu}$ and covariance matrix Σ. The density function of $\underline{X}$ is thus

$$f(\underline{x}) = \frac{1}{(2\pi)^{d/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})\right\},\ \underline{x} \in \mathbb{R}^d. \tag{5.3}$$

The goal of this section is to find the distribution of $Y = f\left(\underline{X}\right)$ denoted by $G_d\left(y, f\right)$. Once this distribution is obtained, the extreme value distribution describing the minimum of $Y = f\left(\underline{X}\right)$ can be recovered. The results of this section are based on the work of Clifton *et al*. (2011). Using equation (5.3), the distribution function $G_d\left(y, f\right)$ is

$$G_d\left(y, f\right) = \int_{f^{-1}(]0,y])} f\left(\underline{x}\right) d\underline{x} = \int_{f^{-1}(]0,y])} \frac{1}{\left(2\pi\right)^{d/2}} \left|\Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{x} - \underline{\mu}\right)\right\} d\underline{x}\,.$$

Consider the transformation to polar coordinates using the Mahalanobis radius. Let $\underline{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_{d-1}\right)$ be angles such that $\theta_j \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $j = 1, 2, \ldots, d-2$ and $\theta_{d-1} \in \left[0, 2\pi\right]$. The transformation to polar coordinates is given by

$$r^2 = M\left(\underline{x}\right)^2 = \left(\underline{x} - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{x} - \underline{\mu}\right) \text{ and} \qquad (5.4.1)$$

$$x_j = r \cdot \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{d-j} \sin\theta_{d-j+1}\ , j = 1, \ldots, d\,. \qquad (5.4.2)$$

The Jacobian of this well-known transformation (see, for example, Scott, 2015) is

$$\left|J\right| = \left\|\begin{bmatrix} \frac{\partial \underline{x}}{\partial r} & \frac{\partial \underline{x}}{\partial \underline{\theta}} \end{bmatrix}\right\| = \left|\Sigma\right|^{1/2} r^{d-1} \prod_{j=1}^{d-2}\left(\cos\theta_j\right)^{d-j-1}. \qquad (5.5)$$

Let $M^*\left(y\right)$ represent the Mahalanobis distance of $\underline{x}$ with corresponding probability density value $y = f\left(\underline{x}\right)$. Notice that

$$y = f\left(\underline{x}\right) = \frac{1}{\left(2\pi\right)^{d/2}} \left|\Sigma\right|^{-1/2} \exp\left\{-\frac{M\left(\underline{x}\right)^2}{2}\right\} \Rightarrow M\left(\underline{x}\right)^2 = -2\ln\left[\left(2\pi\right)^{d/2} \left|\Sigma\right|^{1/2} y\right], \text{ and therefore,}$$

$$M^*\left(y\right) = \sqrt{-2\ln\left[\left(2\pi\right)^{d/2} \left|\Sigma\right|^{1/2} y\right]}\,.$$

Under transformation (5.4), the distribution over the density function of the multivariate Gaussian random vector is

$$G_d\left(y,f\right)=\int_{M^*(y)}^{\infty}\int_{0}^{2\pi}\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2}\frac{|J|}{\left(2\pi\right)^{d/2}}\left|\Sigma\right|^{-1/2}\exp\left\{-r^2/2\right\}d\theta_1\ldots d\theta_{d-1}dr\ . \tag{5.6}$$

Substituting in the Jacobian and simplifying the integral leads to

$$G_d\left(y,f\right)=\frac{1}{\left(2\pi\right)^{d/2}}\int_{M^*(y)}^{\infty}r^{d-1}\exp\left\{-r^2/2\right\}\int_{0}^{2\pi}\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2}\prod_{j=1}^{d-2}\left(\cos\theta_j\right)^{d-j-1}d\theta_1\ldots d\theta_{d-1}dr\ . \tag{5.7}$$

Consider the integral

$$\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2}\prod_{j=1}^{d-2}\left(\cos\theta_j\right)^{d-j-1}d\theta_1\ldots d\theta_{d-2}\ . \tag{5.8}$$

Scott (2015: 29) showed that

$$\int_{-\pi/2}^{\pi/2}\left(\cos\theta\right)^k d\theta=\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k+2}{2}\right)}=B\left(\frac{1}{2},\frac{k+1}{2}\right). \tag{5.9}$$

Hence,

$$\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2}\prod_{j=1}^{d-2}\left(\cos\theta_j\right)^{d-j-1}d\theta_1\ldots d\theta_{d-2}=\prod_{j=1}^{d-2}B\left(\frac{1}{2},\frac{d-j}{2}\right)=\frac{\Gamma\left(\frac{1}{2}\right)^{d-2}}{\Gamma\left(\frac{d}{2}\right)}\ . \tag{5.10}$$

All the angles except the base angle, θ_{d-1}, have been integrated out. Integrating over the base angle gives

$$\int_{0}^{2\pi}\frac{\Gamma\left(\frac{1}{2}\right)^{d-2}}{\Gamma\left(\frac{d}{2}\right)}d\theta_{d-1}=\frac{2\pi\cdot\Gamma\left(\frac{1}{2}\right)^{d-2}}{\Gamma\left(\frac{d}{2}\right)}=\frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}\equiv\Omega_d\ . \tag{5.11}$$

Finally, the integral in (5.6) simplifies to

$$G_d(y,f)=\Omega_d\int_{M^*(y)}^{\infty}\frac{r^{d-1}}{(2\pi)^{d/2}}\exp\left\{-r^2/2\right\}dr\,. \tag{5.12}$$

Consider the substitution

$$u=\frac{1}{(2\pi)^{d/2}}|\Sigma|^{-1/2}\exp\left\{-r^2/2\right\}\ ;\ du=-\frac{r}{(2\pi)^{d/2}}|\Sigma|^{-1/2}\exp\left\{-r^2/2\right\}dr\,. \tag{5.13}$$

Let $C_d=(2\pi)^{d/2}|\Sigma|^{1/2}$ so that

$$G_d(y,f)=\Omega_d|\Sigma|^{1/2}\int_0^y\left[-2\cdot\ln(C_d u)\right]^{(d-2)/2}du\,. \tag{5.14}$$

Differentiating the probability distribution function in (5.14) gives the corresponding probability density function as

$$g_d(y,f)=\Omega_d|\Sigma|^{1/2}\left[-2\cdot\ln(C_d y)\right]^{(d-2)/2}\ ,\ y\in\left(0,\max_{\underline{x}}\left\{f(\underline{x})\right\}\right). \tag{5.15}$$

Using the property of the spherical symmetry of a multivariate Gaussian density, one can take the upper bound of Y as $\max\left\{f(\underline{x})\right\}=f(\underline{\mu})$. The distribution function, $G_d(y,f)$, can now be derived for three cases – the univariate case, and the two cases where the dimension of the data space is odd or even.

Consider first the case where $d=1$. Equation (5.12) then becomes

$$G_1(y,f)=\Omega_1\int_{\sqrt{-2\ln(C_1 y)}}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-r^2/2\right\}dr=\frac{2}{\sqrt{2\pi}}\int_{\sqrt{-2\ln(C_1 y)}}^{\infty}\exp\left\{-r^2/2\right\}dr\,. \tag{5.16}$$

Let $u=r/\sqrt{2}\Rightarrow dr=\sqrt{2}du$. Then, the integral in (5.16) becomes

$$G_1(y,f)=\frac{2}{\sqrt{\pi}}\int_{\sqrt{-\ln(C_1 y)}}^{\infty}\exp\left\{-u^2\right\}du=erfc\left(\sqrt{-\ln(C_1 y)}\right). \tag{5.17}$$

The function $erfc(\cdot)$ is the complementary error function given by

$$erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} \exp\{-u^2\} du \, . \qquad (5.18)$$

The function in equation (5.17) is the cumulative probability distribution of the probability density function of a univariate normal distribution. For the multivariate case, the complement of $G_d(y,f)$ in (5.12) is considered. Hence, if $M^*(y) = R$ is the Mahalanobis distance associated with $y = f(\underline{x})$, the complement of $G_d(y,f)$ is the integral over the region of the density interior to the hyper-ellipsoid of radius R. Let this region be $B = \{\underline{x} : M(\underline{x}) \leq R\}$, $R \in \mathbb{R}^+$ and the complementary integral be defined as $\bar{G}_d(y,f)$. The following results are derived in Hugueny (2013).

The probability $\bar{G}_d(y,f)$ is given by

$$\bar{G}_d(y,f) = \Omega_d \int_{0}^{R} \frac{r^{d-1}}{(2\pi)^{d/2}} \exp\left\{-r^2/2\right\} dr \equiv \frac{\Omega_d}{(2\pi)^{d/2}} u_d(R) \, . \qquad (5.19)$$

Consider the integral $u_d(R)$. This integral can be written as

$$u_d(R) = \int_{0}^{R} r^{d-1} \exp\left\{-r^2/2\right\} dr = \int_{0}^{R} r^{d-2} \left[r \cdot \exp\left\{-r^2/2\right\} \right] dr \, . \qquad (5.20)$$

Using integration by parts, this equation becomes

$$u_d(R) = \int_{0}^{R} r^{d-2} \frac{d}{dr} \left[-\exp\left\{-r^2/2\right\} \right] dr = -\exp\left\{-r^2/2\right\} r^{d-2} \Big]_{0}^{R} + (d-2) \int_{0}^{R} r^{d-3} \exp\left\{-r^2/2\right\} dr \, .$$

Hence,

$$u_d(R) = -\exp\left\{-R^2/2\right\} R^{d-2} + (d-2) \cdot u_{d-2}(R) \, . \qquad (5.21)$$

Clearly, this function depends on $u_{d-2}(R)$. Therefore, the even and odd cases must be considered separately. One more ingredient needed is the form of $u_1(R)$ and $u_2(R)$. Using equation (5.20), it follows directly that

$$u_1(R) = \sqrt{\frac{\pi}{2}}\, erf\left\{ R \big/ \sqrt{2} \right\} \text{ and } u_2(R) = 1 - \exp\left\{ -R^2 \big/ 2 \right\}. \tag{5.22}$$

Taking $d = 2p$, $p \in \mathbb{N}$ and using equation (5.21) and $u_2(R)$, give

$$u_{2p}(R) = 2^{p-1}(p-1)! - \sum_{k=0}^{p-1} \frac{2^k (p-1)!}{(p-1-k)!} R^{2(p-1-k)} \exp\left\{ -R^2 \big/ 2 \right\}. \tag{5.23.1}$$

Similarly, if $d = 2p+1$, $p \in \mathbb{N}$, then $u_d(R)$ is recovered from equation (5.21) and $u_1(R)$ as

$$u_{2p+1}(R) = u_1(R) \frac{(2p+1)!}{2^{p-1}(p-1)!} - \sum_{k=0}^{p-1} \frac{(2p+1)!(p-k)!}{2^{k-1}(p-1)!(2p-2k)!} R^{2(p-k)-1} \exp\left\{ -R^2 \big/ 2 \right\}. \tag{5.23.2}$$

Returning to equation (5.19) the survival function $\bar{G}_d(y,f)$ is then

$$\bar{G}_{2p}(y,f) = 1 - \exp\left\{ -R^2 \big/ 2 \right\} \sum_{k=0}^{p-1} A_{2p}^k R^{2(p-1-k)} \text{ , and} \tag{5.24.1}$$

$$\bar{G}_{2p+1}(y,f) = erf\left(R \big/ \sqrt{2} \right) - \exp\left\{ -R^2 \big/ 2 \right\} \sum_{k=0}^{p-1} A_{2p+1}^k R^{2(p-k)-1}. \tag{5.24.2}$$

In these two equations, the constants A_{2p}^k and A_{2p+1}^k are given by

$$A_{2p}^k = \Omega_{2p} \left| \Sigma \right|^{1/2} \frac{2^k (p-1)!}{(p-k-1)!} \text{ , and} \tag{5.25.1}$$

$$A_{2p+1}^k = \Omega_{2p+1} \left| \Sigma \right|^{1/2} \frac{(2p-1)!(p-k)!}{2^{k-1}(p-1)!(2p-2k)!}. \tag{5.25.2}$$

Hence, from the facts that $G_d(y,f) = 1 - \bar{G}_d(y,f)$ and $R = \sqrt{-2\ln(C_d y)}$, the two forms of the distribution function are

$$G_{2p}(y,f) = y \sum_{k=0}^{p-1} A_{2p}^k \left[-2\ln(C_{2p} y) \right]^{p-1-k} \text{ , and} \tag{5.26.1}$$

$$G_{2p+1}(y,f) = erfc\left(\sqrt{-\ln(C_{2p+1} y)} \right) + y \sum_{k=0}^{p-1} A_{2p+1}^k \left[-2\ln(C_{2p+1} y) \right]^{\frac{2p-2k-1}{2}}. \tag{5.26.2}$$

This completes the derivation. Thus, a closed-form expression exists for the distribution function of the probability density of a multivariate Gaussian distribution. In turn, the distribution of $f\left(E_N\right)$ can be utilised to perform novelty detection probabilistically. To find this distribution the Fisher-Tippett theorem from Chapter 3 is utilised for the minimum of a sample.

5.2.3 Distribution of the minimum density of a multivariate Gaussian random vector

In this section, the asymptotic distribution of the minimum probability density function of a multivariate Gaussian vector is derived. It was seen in the previous section that a closed-form expression for the distribution function of the probability density function of a multivariate Gaussian vector exists. Remarkably, the multivariate problem has been reduced to an equivalent univariate case. Therefore, the Fisher-Tippett theorem can be applied to that distribution.

Chapter 3 discussed the condition on the tail quantile function for a distribution to be in the maximal domain of attraction of the extremal Weibull class of GEV distributions. However, of interest is rather the limiting distribution of the minimum density. Recall that a random variable X with distribution and tail-quantile function F and U, respectively, is in the maximum domain of attraction of the extremal Weibull class of GEV distributions if, and only if,

$$U\left(x\right)=x_u-x^{\gamma}\ell_U\left(x\right)\ ,x\rightarrow\infty\,.$$

As mentioned in Beirlant *et al*. (2004: 67), an equivalent condition on F is

$$1-F\left(x_u-\tfrac{1}{x}\right)=x^{\frac{1}{\gamma}}\ell_F\left(x\right)\ ,x\rightarrow\infty,\ \gamma<0\,. \qquad (5.27)$$

In equation (5.27) the quantity $x_u<\infty$ is the upper bound of X and $\ell_F\left(x\right)$ is a slowly varying function linked to $\ell_U\left(x\right)$ via the De Bruyn conjugate. An equivalent statement for the minimum is obtained by remarking that $X_{1,N}=\min\left\{X_i\right\}_{i=1}^{N}=-\max\left\{-X_i\right\}_{i=1}^{N}$. Furthermore,

$$1-P\left(-X>x\right)=P\left(X\leq -x\right)=F\left(-x\right). \qquad (5.28)$$

The statement in equation (5.27) can be expressed as

$$1-F\left(x\right)=\left(x_u-x\right)^{-\frac{1}{\gamma}}\ell_F\left(\left(x_u-x\right)^{-1}\right)\ ,x\rightarrow x_u,\ \gamma<0\,. \qquad (5.29)$$

In turn, the condition on F such that $X_{1,N} = \min\limits_{i=1,\ldots,N}\{X_i\}$ (normalised) converges in distribution to the minimal Weibull class of GEV distributions is

$$F_X(x) = 1 - F_{-X}(-x) = (x_{-u} + x)^{-1/\gamma}\,\ell_F\left((x_{-u} + x)^{-1}\right)\text{ , } x \to x_{-u},\ \gamma < 0\,. \tag{5.30}$$

This follows from the fact that for the minimum of a sequence of iid random variables the maximum of $\{-X\}_{i=1}^{N}$ is considered. Furthermore, x_{-u} is the upper bound of the random variable $-X$. Again, equation (5.30) is better expressed as

$$F\left(x_L + {}^{1}\!/_{x}\right) = x^{-\alpha}\ell_F(x)\text{ , } x \to \infty,\ \alpha > 0\,. \tag{5.31}$$

In (5.31) the quantity $x_L = -x_{-u}$ is the lower bound of X and $\alpha = -{}^{1}\!/_{\gamma}$. If this condition is satisfied, sequences of constants $a_N > 0$ and b_N exist such that $a_N^{-1}\left(X_{1,N} - b_N\right)$ converges in distribution to the minimal Weibull distribution denoted by $G^{-}(x)$. Furthermore, as mentioned in Clifton *et al*. (2011), the norming constants (for a fixed N) can be taken as

$$a_N = x_L + F^{-1}\left({}^{1}\!/_{N}\right)\text{ and } b_N = x_L\,. \tag{5.32}$$

Finally, using equation (5.28) and the form of the extremal Weibull distribution, it follows that the distribution of $X_{1,N} = \min\limits_{i=1,\ldots,N}\{X_i\}$, for large N, is approximated by

$$P\left(X_{1,N} \le x\right) = P\left(a_N^{-1}\left(X_{1,N} - b_N\right) \le a_N^{-1}(x - b_N)\right) \approx G^{-}\left(a_N^{-1}(x - b_N)\right) \equiv 1 - \exp\left\{-\left(\frac{x - b_N}{a_N}\right)^{\alpha}\right\}. \tag{5.33}$$

Note that the EVI is $\gamma = -1/\alpha$, $\alpha > 0$. It will now be shown that the distribution of $Y = f\left(\underline{X}\right)$ is in the domain of attraction of the minimal Weibull distribution if $f\left(\underline{x}\right)$ is the probability density function of a multivariate Gaussian distribution. Notice that any Gaussian probability density function has the property that $f\left(\underline{x}\right) \ge 0$. Thus, the lower bound of Y is $x_L = 0$. Therefore, Y is in the domain of attraction of the minimal Weibull class of GEV distributions if, and only if,

$$G_d\left({}^{1}\!/_{x}, f\right) = x^{-\alpha}\ell_F(x)\text{ , } x \to \infty\,. \tag{5.34}$$

Consider the case where $d = 2p$, $p \in \mathbb{N}$. The left-hand side of (5.31) is then

$$G_{2p}\left(x^{-1},f\right) = x^{-1}\sum_{k=0}^{p-1} A_{2p}^{k}\left[-2\ln\left(\frac{C_{2p}}{x}\right)\right]^{p-1-k} .$$

If the function $\ell(x) = \sum_{k=0}^{p-1} A_{2p}^{k}\left[-2\ln\left(\frac{C_{2p}}{x}\right)\right]^{p-1-k}$ is slowly varying, it follows directly that condition (5.31) is satisfied. It is well known that for all $\beta \in \mathbb{R}$ and for $x > 1$ the function $\left[-\ln\left(\frac{1}{x}\right)\right]^{\beta} = \left[\ln(x)\right]^{\beta}$ is slowly varying – see Beirlant *et al*. (2004: 78). Hence, each term in the sum of $\ell(x)$ is slowly varying. Additionally, the sum of slowly varying functions is also slowly varying. Hence, $\ell(x)$ is slowly varying. Therefore, condition (5.31) is satisfied. It was thus argued that

$$G_{2p}\left(x^{-1},f\right) = x^{-1}\ell(x) \ , x \to \infty . \tag{5.35}$$

Furthermore, the sequences of constants are given by

$$a_N = G_{2p}^{-1}\left(\frac{1}{N}\right) \text{ and } b_N = 0 . \tag{5.36}$$

Notice that a_N acts as a scale parameter in the normalisation. This parameter must be estimated numerically.

For the case where $d = 2p+1$, $p \in \mathbb{N}$ the same arguments are followed. Recall that the distribution of the probability density values for the odd case is

$$G_{2p+1}(y,f) = erfc\left(\sqrt{-\ln\left(C_{2p+1}y\right)}\right) + y\sum_{k=0}^{p-1} A_{2p+1}^{k}\left[-2\ln\left(C_{2p+1}y\right)\right]^{\frac{2p-2k-1}{2}} .$$

Hence,

$$G_{2p+1}\left(x^{-1},f\right) = x^{-1}\left[\sum_{k=0}^{p-1} A_{2p+1}^{k}\left[-2\ln\left(C_{2p+1}x^{-1}\right)\right]^{\frac{2p-2k-1}{2}} + x\cdot erfc\left(\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}\right)\right] . \tag{5.37}$$

For the distribution function in (5.37) to be in the domain of attraction of the minimal Weibull distribution, it must be that $\ell(x)$ is slowly varying, where

$$\ell(x)=\ell_1(x)+\ell_2(x)=\sum_{k=0}^{p-1}A_{2p+1}^{k}\left[-2\ln\left(C_{2p+1}x^{-1}\right)\right]^{\frac{2p-2k-1}{2}}+x\cdot erfc\left(\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}\right). \qquad (5.38)$$

Notice that $\ell_1(x)$ is slowly varying for similar reasons than for the even case. Hence, if $\ell_2(x)$ is slowly varying, the distribution of the probability density values of a multivariate Gaussian is in the domain of attraction of the minimal Weibull distribution. Consider the complementary error function in (5.38). Let this function be denoted by $\ell_{Err}(x)$. Then,

$$\lim_{x\to\infty}\frac{\ell_{Err}\left(\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}\right)}{\ell_{Err}\left(\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}\right)}=\lim_{x\to\infty}\frac{\int_{\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}}^{\infty}\exp\left\{-s^2\right\}ds}{\int_{\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}}^{\infty}\exp\left\{-s^2\right\}ds}. \qquad (5.39)$$

Clearly, both the numerator and denominator converge to zero as $x\to\infty$. Therefore, L'Hospital's rule can be applied. This gives

$$\lim_{x\to\infty}\frac{\ell_{Err}\left(\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}\right)}{\ell_{Err}\left(\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}\right)}\approx\lim_{x\to\infty}\frac{\exp\left\{\ln\left(C_{2p+1}(tx)^{-1}\right)\right\}}{\exp\left\{\ln\left(C_{2p+1}x^{-1}\right)\right\}}\cdot\frac{\frac{d}{dx}\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}}{\frac{d}{dx}\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}}. \qquad (5.40)$$

But, $\exp\left\{\ln\left(C_{2p+1}(tx)^{-1}\right)\right\}=C_{2p+1}(tx)^{-1}$ and $\exp\left\{\ln\left(C_{2p+1}x^{-1}\right)\right\}=C_{2p+1}x^{-1}$. Thus, their ratio is

$$\frac{C_{2p+1}(tx)^{-1}}{C_{2p+1}x^{-1}}=t^{-1}. \qquad (5.41)$$

Furthermore,

$$\frac{d}{dx}\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}=\frac{C_{2p+1}t^{-1}x^{-2}}{2C_{2p+1}(tx)^{-1}\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}}=\frac{1}{2\cdot x\cdot\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}}, \text{ and}$$

$$\frac{d}{dx}\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)} = \frac{C_{2p+1}x^{-2}}{2C_{2p+1}x^{-1}\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}} = \frac{1}{2\cdot x\cdot\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}}.$$

Hence, their ratio is

$$\frac{\frac{d}{dx}\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}}{\frac{d}{dx}\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}} = \left[\frac{\ln\left(C_{2p+1}(tx)^{-1}\right)}{\ln\left(C_{2p+1}x^{-1}\right)}\right]^{-1/2}. \qquad (5.42)$$

Finally, it follows that

$$\lim_{x\to\infty}\frac{\ell_{Err}\left(\sqrt{-\ln\left(C_{2p+1}(tx)^{-1}\right)}\right)}{\ell_{Err}\left(\sqrt{-\ln\left(C_{2p+1}x^{-1}\right)}\right)} = t^{-1}\cdot\lim_{x\to\infty}\left[\frac{\ln\left(C_{2p+1}(tx)^{-1}\right)}{\ln\left(C_{2p+1}x^{-1}\right)}\right]^{-1/2} = t^{-1}\cdot\lim_{x\to\infty}\left[\frac{\frac{\ln C_{2p+1}-\ln t}{\ln x}-1}{\frac{\ln C_{2p+1}}{\ln x}-1}\right]^{-1/2} = t^{-1}.$$

Thus,

$$\lim_{x\to\infty}\frac{\ell_2(tx)}{\ell_2(x)} = \lim_{x\to\infty}\frac{tx}{x}\cdot t^{-1} = 1. \qquad (5.43)$$

Consequently, both $\ell_1(x)$ and $\ell_2(x)$ are slowly varying. Their sum is therefore slowly varying and it follows that $\ell(x)$ is slowly varying. Hence, $G_{2p+1}\left(x^{-1},f\right) = x^{-1}\ell(x)$ and, therefore, $G_{2p+1}(\cdot,f)$ is in the domain of attraction of the minimal Weibull class of GEV distributions.

In conclusion, it has been shown that, for large N, the distribution of $f(E_N)$ can be approximated by

$$G_d^{\min}(y,f) = P\left[f(E_N)\leq y\right] = P\left[a_N^{-1}f(E_N)\leq a_N^{-1}y\right] \approx 1-\exp\left\{-\left[a_N^{-1}y\right]\right\}, \; y\in P_f. \qquad (5.44)$$

Furthermore, the scale parameter (from the normalisation) is denoted by $a_N = G_{2p}^{-1}\left(1/N\right)$ and the shape parameter is $\alpha = 1$, as shown by Clifton *et al.* (2011). The authors mentioned further that although the limiting distribution has $\alpha = 1$, some issues arise when estimating this quantity from a finite sequence. This parameter, when estimated by maximum likelihood, is not close to unity for finite samples. The error in the estimate becomes more apparent as the dimension increases. Therefore, the authors proposed estimating the shape parameter with

$$\alpha_N = a_N \frac{g_d(a_N)}{G_d(a_N)}. \tag{5.45}$$

This estimator follows directly from the fact that distributions in the domain of attraction of the minimal Weibull class of GEV distributions have the property that $G_d(y,f) \sim Ky^s$, $y \to 0$. Thus, based on Karamata's Tauberian theorem (Bingham, Goldie & Teugels, 1987), $g_d(y,f) \sim sKy^{s-1}$, $y \to 0$ such that $s \sim y\frac{g_d(y)}{G_d(y)}$. The equation in (5.45) is then obtained by taking $y = a_N$. Note that $a_N \to 0$ as $N \to \infty$ so that Karamata's Tauberian theorem holds for $y = a_N$. The limiting distribution of $a_N^{-1} \cdot f(E_N)$ is approximated from a finite sample by $1-\exp\{-y^{\alpha_N}\}$, with α_N given in (5.45).

5.2.4 Novelty scores and final classification

The previous two sections demonstrated how a limiting distribution of the minimum density of a multivariate Gaussian sequence of random vectors is derived. It was seen that the distribution of $Y = f(\underline{X})$ is in the domain of attraction of the extremal Weibull class of GEV distributions. Therefore, the limiting distribution of $f(E_N)$ can be approximated from a finite sample by

$$P(f(E_N) \le y) \approx 1 - \exp\left\{-\left(y/a_N\right)^{\alpha_N}\right\}. \tag{5.46}$$

For a new observation $y^* = f(\underline{x}^*)$, the quantity $P(f(E_N) \le y^*)$ is approximated as in (5.46). This approximation is the approximate probability of drawing an extremum $f(E_N)$ less than the observed quantity y^* – the probability that y^* is within the boundary. Hence, the measure $P(f(E_N) > y^*)$ represents the probability that y^*, and consequently $\underline{x}^*$, is more extreme (in the probability space) than what is expected as normal behaviour (Clifton *et al*., 2011). Again, this survival function is approximated similarly as in equation (5.46).

Furthermore, Clifton *et al*. (2011) stated that the novelty score in data space is the probability that the observation is (in terms of the Mahalanobis distance) further from the centre of the distribution than the minimum density $f\left(E_N\right)$ of the normal class. Therefore, the novelty score at the location $y = f\left(\underline{x}\right)$ in the data space is given by

$$\bar{G}_d^{\min}\left(y,f\right) = 1 - G_d^{\min}\left(f\left(\underline{x}\right),f\right) \approx \exp\left\{-\left(\frac{1}{a_N \cdot C_d}\exp\left\{-{M\left(\underline{x}\right)^2}\big/{2}\right\}\right)^{\alpha_N}\right\}. \qquad (5.47)$$

As before, $C_d = \left(2\pi\right)^{d/2}\left|\Sigma\right|^{1/2}$. A high value indicates that an observation is novel. Notice that for normal observations, (5.47) will be close to zero. Hence, a probabilistic novelty detection algorithm has been constructed. The main assumption of this model is that data is generated from a multivariate Gaussian distribution. In turn, the definition of extreme value theory in a probability space and the Fisher-Tippett theorem are utilised to find a limiting distribution for the minimum probability density of the multivariate Gaussian vector. Thus, the challenge of multivariate extreme value theory has been circumvented by moving from the data space to the probability space of the density function. Remarkably, a fully parametric distribution is recovered in the limit for any dimension. Moreover, a novelty threshold of normality can now be set with a complete probabilistic interpretation. Thus, a new observation is classified as normal or novel at a specified probability.

One disadvantage of this model is the assumption that the random vector is multivariate Gaussian. This assumption is too restrictive in multimodal cases. As discussed previously, the extremes of a multivariate Gaussian mixture model do not necessarily only depend on the closest Gaussian component. Rather, multiple components may influence the extremeness of an observed vector. Clifton *et al*. (2011) proposed a numerical scheme to handle complex scenarios such as multimodality. This scheme is now discussed.

5.3 A NUMERICAL SCHEME FOR GAUSSIAN MIXTURE MODELS

Consider the case where data is multivariate and multimodal. Such complex problems for novelty detection have only been considered recently. The multimodality in the data implies that a multivariate normal distribution cannot adequately describe the data-generating mechanism. Therefore, a more complex model that does not rely on too strong assumptions (such as unimodality) must be formulated. One approach would be to assume the underlying density function is a mixture of multivariate Gaussian distributions. Then, the probability density function is of the form

$$f\left(\underline{x}\right)=\sum_{k=1}^{K}\alpha_k f_k\left(\underline{x},\underline{\mu}_k,\Sigma_k\right)\ ,\ \sum_k \alpha_k=1.$$

Let $\underline{X}_1,\underline{X}_2,\ldots,\underline{X}_N$ be an iid sequence from $f\left(\underline{X}\right)$. The goal is to approximate the distribution of $f\left(E_N\right)=\min\limits_{i=1,\ldots,N} f\left(\underline{X}_i\right)$. Clifton *et al.* (2011) remarked that the probability density (and distribution) function of the extreme random variable $f\left(E_N\right)$ are equal for level sets in data space of the underlying distribution function, $F\left(\underline{x}\right)$. This point is verified in the next section.

As a result, the extreme value distribution of the minimum of the sample of probability density values can be regarded as a transformation of the contours of $f\left(\underline{x}\right)$. In turn, Clifton *et al*. (2011) used this generalisation to formulate a numerical scheme for approximating the distribution of the minimum density of a GMM.

This section is concluded with a discussion of the numerical scheme proposed by Clifton *et al.* (2011).

5.3.1 Equiprobable contours of the density of the minimum probability density

Recall that an extreme vector is defined as $E_N=\arg\min\limits_{\underline{X}_i\ ,\ i=1,\ldots,N}\left[f\left(\underline{X}_i\right)\right]$. The distribution function $G_d\left(y,f\right)$ and probability density function $g_d\left(y,f\right)$ of $f\left(\underline{X}\right)$ were therefore derived for the Gaussian case. Our goal is to find the density and distribution functions of $f\left(E_N\right)$ (normalised), denoted by $g_{Y_{1,N}}\left(y\right)$ and $G_{Y_{1,N}}\left(y\right)$ in this section, respectively, for cases other than the multivariate Gaussian case. It is now argued that, in general, the distribution and density functions of $f\left(E_N\right)$ are equal, respectively, at the contours of $f\left(\underline{x}\right)$.

Let $\underline{X}\in\mathbb{R}^d$ be any random vector with probability distribution and density functions denoted by $F\left(\underline{x}\right)$ and $f\left(\underline{x}\right)$, respectively. Furthermore, assume that the distribution and density functions of $Y=f\left(\underline{X}\right)$ exist and denote them by $G\left(y\right)$ and $g\left(y\right)$, respectively. Consider the set of contours $S=\left\{\underline{x}:f\left(\underline{x}\right)=p\right\}$. Notice that for all $\underline{x}_i,\underline{x}_j\in S$ it must be, by definition, that $f\left(\underline{x}_i\right)=f\left(\underline{x}_j\right)$. Therefore, if $y_k=f\left(\underline{x}_k\right)$, $k=i,j$, it must be that $G\left(y_i\right)=G\left(y_j\right)$ and

$g(y_i) = g(y_j)$ for all $\underline{x}_i, \underline{x}_j \in S$. Clearly, this is the case since G and g are functions of $y = f(\underline{x})$. However, the probability distribution and density functions of $Y_{1,N} = \min_{i=1,\ldots,N}\{Y_i\}$ as a function of G and g are given by

$$G_{Y_{1,N}}(y) = P\left(\min_{i=1,\ldots,N}\{Y_i\} \leq y\right) \equiv 1 - \left[1 - G(y)\right]^N \text{, and} \qquad (5.48.1)$$

$$g_{Y_{1,N}}(y) \equiv N \cdot g(y)\left[1 - G(y)\right]^{N-1} \text{, respectively.} \qquad (5.48.2)$$

Hence, given that G and g exist for all $\underline{x}_i, \underline{x}_j \in S$, it must be that

$$G_{Y_{1,N}}\left(f(\underline{x}_i)\right) = G_{Y_{1,N}}\left(f(\underline{x}_j)\right) \text{ and } g_{Y_{1,N}}\left(f(\underline{x}_i)\right) = g_{Y_{1,N}}\left(f(\underline{x}_j)\right). \qquad (5.49)$$

Notice that $G_{Y_{1,N}}$ is the distribution function of $f(E_N)$. For the multivariate Gaussian case, this distribution function was denoted by $G_d^{\min}(\cdot, f)$. Therefore, in general, the probability density function of $f(E_N)$ can be regarded as a weighted function of the equiprobable contours of $f(\underline{x})$ (Clifton *et al*., 2011).

Given this reasoning, assume that the density of $f(E_N)$ is a weighted function Λ of the probability density function $f(\underline{x})$. Thus,

$$g_{Y_{1,N}}(y) = \Lambda(y) = \Lambda\left(f(\underline{x})\right). \qquad (5.50)$$

If the function Λ can be estimated accurately, the probability density function of $f(E_N)$ is recovered. One method, termed the Ψ-transform, has been proposed by Clifton (2009) and Clifton *et al*. (2011). This method is now discussed.

5.3.2 The Ψ-transform method

This section describes the numerical method for approximating the limiting distribution of the minimum density of a Gaussian mixture model. Clifton (2009) first proposed this method which has been applied to monitor the vital signs of patients (Clifton *et al*., 2011). In Section 5.3.1 it was argued that $g_{Y_{1,N}}(y)$ can be viewed as a weighted function of the contours of $f(\underline{x})$.

Consequently, if this function is estimated accurately, the resulting extreme value model would be an accurate estimate of the distribution of the minimum probability density of a GMM. Let $\underline{X} \in \mathbb{R}^d$, $d \in \mathbb{N}$ and consider the standard Gaussian density

$$f\left(\underline{x}\right) = \left(2\pi\right)^{-d/2} \exp\left\{-\underline{x}^T \underline{x}/2\right\}. \tag{5.51}$$

Notice that for the standard Gaussian case, the squared Mahalanobis distance is $M\left(\underline{x}\right)^2 = \underline{x}^T \underline{x}$. Therefore, this squared distance in terms of the density contours is

$$\underline{x}^T \underline{x} = -2\ln\left(f\left(\underline{x}\right)\right) - d\ln\left(2\pi\right). \tag{5.52}$$

Thus, since the Mahalanobis distance and inner product of a standard Gaussian vector are equal, the distance of $\underline{x}$ to zero (mean) with respect to the identity covariance is

$$M\left(\underline{x}\right) = \left\|\underline{x}\right\| = \left[-2\ln\left(f\left(\underline{x}\right)\right) - d\ln\left(2\pi\right)\right]^{1/2}. \tag{5.53}$$

Roberts (1999) and Clifton (2009) remarked that the maximum Mahalanobis radii of a multivariate normal distribution is in the domain of attraction of the Gumbel class of GEV distributions. This was also proved analytically in Section 4.5.2. Therefore, Clifton (2009) and Clifton *et al*. (2011) proposed to transform the extrema of $\underline{x}$ to

$$\Psi\left[f\left(\underline{x}\right)\right] = \begin{cases} \left[-2\ln\left(f\left(\underline{x}\right)\right) - d\ln\left(2\pi\right)\right]^{1/2} & f\left(\underline{x}\right) < K \\ 0 & f\left(\underline{x}\right) \geq K \end{cases}. \tag{5.54}$$

In equation (5.54), the boundary is at $K = \left(2\pi\right)^{-d/2}$. Notice that the transformation is in terms of the $\underline{x}$ which is extreme in probability space. Given the fact that for the standard multivariate Gaussian case the maximum of the transformation $\Psi\left[f\left(\underline{X}\right)\right]$ is Gumbel distributed, it is argued that for any GMM the function $\Psi\left[f\left(\underline{x}\right)\right]$ transforms the density contours $f\left(\underline{x}\right)$ such that $\Psi\left[f\left(\underline{X}\right)\right]$ is expected to be in the domain of attraction of the Gumbel distribution. This assumption was validated empirically by Clifton *et al*. (2011).

In equation (5.54), the upper bound is $K=(2\pi)^{-d/2}$. This bound is set such that all values resulting from (5.54) are positive. Furthermore, as $f(\underline{x})$ tends to zero, $\Psi\left[f(\underline{x})\right]$ increases to infinity. Therefore, it is of interest to fit a Gumbel distribution representing the maxima of a sample. Note that the transformation $\Psi\left[f(\underline{X})\right]$ is only applied to the $\underline{X}$ which is regarded extreme in probability space – the transformation is rather over $f(E_N)=\min_{i=1,\dots,N} f(\underline{X}_i)$. Therefore, extrema in probability space must be generated to apply this method. If the sample size permits it, the block-maxima method could be used to generate a sample of $f(E_N)$ and, consequently, estimate the parameters of the Gumbel distribution. Clifton (2009) and Clifton *et al*. (2011) proposed using the parametric bootstrap as an alternative method to estimate the parameters of the Gumbel distribution.

Consider a sample of data describing the normal class. Assume the dimension is $d \in \mathbb{N}$ and each predictor consists of N observations. The first step is to fit a GMM to this data. Therefore, the normal class is used to estimate the mixing proportion, mean vector and covariance matrix of each component. These estimates are denoted by $\hat{\alpha}_k$, $\underline{\hat{\mu}}_k$ and $\hat{\Sigma}_k$ for the k^{th} component, respectively. In turn, the estimated probability density of the data is given by

$$\hat{f}(\underline{x})=\sum_{k=1}^{K}\hat{\alpha}_k(2\pi)^{-d/2}\left|\hat{\Sigma}_k\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\underline{x}-\underline{\hat{\mu}}_k\right)^T\hat{\Sigma}_k^{-1}\left(\underline{x}-\underline{\hat{\mu}}_k\right)\right\}. \tag{5.55}$$

Assume that model (5.55) is an accurate estimate of the (multivariate) density of the normal class. Under this assumption, B samples of size N can be generated from $\hat{f}(\underline{x})$ in (5.55). Let the b^{th} bootstrap sample be denoted by $\left\{\underline{X}_{1b}^*,\underline{X}_{2b}^*,\dots,\underline{X}_{Nb}^*\right\}$ for $b=1,\dots,B$. From each bootstrap sample, a corresponding sample of probability density estimates is obtained. Denote the j^{th} generated density estimate of the b^{th} bootstrap sample by $\hat{f}\left(\underline{x}_{jb}^*\right)$. Consequently, the sample of minimum density estimates used in the Ψ-transform are given by

$$\hat{f}\left(E_{Nb}^*\right)=\min_{j=1,\dots,N}\left\{\hat{f}\left(\underline{X}_{jb}^*\right)\right\}\text{ , } b=1,\dots,B. \tag{5.56}$$

Finally, the bootstrap sample of $\Psi\left[f(E_N)\right]$ is then given by $\Psi\left[f\left(E_{Nb}^*\right)\right]$, $b=1,\dots,B$. Clifton (2009) and Clifton *et al*. (2011) argued that this sample (normalised appropriately) is approximately Gumbel distributed. The validity of this assumption was motivated in the above discussion. Furthermore, they suggest estimating the parameters using maximum likelihood estimation. Ultimately, for large N, the distribution of $\Psi\left[f\left(E_N\right)\right]$ is approximated by

$$\breve{G}_d^{\min}\left(\Psi\left(y\right),f\right) \equiv P\left(\frac{\Psi\left[f\left(E_N\right)\right]-b_N}{a_N} \le \frac{\Psi\left(y\right)-b_N}{a_N}\right) \approx \exp\left\{-\exp\left[-\frac{\Psi\left(y\right)-b_N}{a_N}\right]\right\}. \qquad (5.57)$$

Hence, a limiting distribution for the (transformed) minimum probability density of a GMM has been approximated. Note that the distribution of the transformed minimum density is denoted by $\breve{G}_d^{\min}\left(\cdot,f\right)$ to discriminate between this model and the distribution of the minimum density of only one Gaussian random vector. Consequently, due to the monotonicity of the transformation $\Psi\left(\cdot\right)$, the novelty scores are obtained from

$$P\left(f\left(E_N\right) > y\right) = \breve{G}_d^{\min}\left(\Psi\left(y\right),f\right) \approx \exp\left\{-\exp\left[-\frac{\Psi\left(y\right)-b_N}{a_N}\right]\right\}. \qquad (5.58)$$

Therefore, a vector $\underline{x}$ is regarded as novel with respect to the normal data if $\breve{G}_d^{\min}\left(\Psi\left(y\right),f\right)$, where $y = f\left(\underline{x}\right)$, is above some threshold since this is the probability that a new observation is novel. In conclusion, using a numerical method such as the bootstrap, a limiting distribution for the minimum density of a GMM is approximated.

The significant advantage of the numerical model over the analytical model described in Section 5.2 is that the underlying density function is assumed to be accurately estimated by a GMM and not just one Gaussian distribution. Consequently, datasets of high complexity can be considered. Novelty detection can now be performed in scenarios where the data is multivariate and multimodal. Clifton (2009) stated that a GMM can accurately approximate the probability density of almost any (numerical) data. Given that the GMM approximates the density of the data accurately, the novelty detection algorithm produces a novelty score that probabilistically discriminates between normal and novel data.

However, this model does not come without disadvantages. Firstly, the model requires one to generate a large number of extrema in probability. Clifton *et al.* (2011) generated $B = 10^5$ extrema before fitting a Gumbel distribution. For complex GMMs, this can become inefficient. To sample from a multivariate Gaussian mixture model is slow. Furthermore, fitting a complex GMM also has a long training time. The Ψ-transform method relies on a GMM that accurately estimates the underlying density. If this is not the case, the model will break down. On top of this, in complex situations accurate initial estimates for the mixing proportions, mean vectors and covariance matrices of each Gaussian component are difficult to specify. In turn, the EM algorithm takes longer to train the GMM.

This method is numerical because a closed-form expression cannot necessarily be derived for the distribution of the minimum density of a GMM. The problem is that the domain over which each Gaussian component must be integrated is unknown because there is no reason for the GMM to exhibit the spherical form of a single multivariate Gaussian distribution. However, Hugueny (2013) has made some proposals to improve the computational time of this method. These proposals are discussed in the next section.

It is now demonstrated that the Ψ-transform method is applicable to datasets of high dimension and multimodality. The first example considers a bivariate GMM with 5 Gaussian components. This example is used to demonstrate that the Ψ-transform can be used in multimodal scenarios. Thereafter, an example of a 6-dimensional GMM with two components is considered. In turn, it is argued that the dimensionality of the data does not impact the goodness-of-fit of this model. Notice that this example is for illustration purposes only and that the parameters of the GMM were not estimated but rather assumed to be known.

Consider the GMM consisting of 5 components. The properties of this GMM are

$$\underline{\mu}_1=(0,0),\ \underline{\mu}_2=(3,3),\ \underline{\mu}_3=(5,-2),\ \underline{\mu}_4=(-3,4),\ \underline{\mu}_5=(1,-3),\text{ and}$$

$$\Sigma_1=\begin{bmatrix}1&0\\0&1\end{bmatrix},\ \Sigma_2=\begin{bmatrix}3&1\\1&2\end{bmatrix},\ \Sigma_3=\begin{bmatrix}4&-2\\-2&3\end{bmatrix},\ \Sigma_4=\begin{bmatrix}2&-0.9\\-0.9&2\end{bmatrix},\ \Sigma_5=\begin{bmatrix}5&-3\\-3&4\end{bmatrix}.$$

The two-dimensional case was used so that the density contours can be plotted. These contours are given in Figure 5.1. For this graph, mixing proportions $\underline{\alpha}=(0.05,0.25,0.3,0.2,0.2)$ were used. Additionally, 100 randomly generated vectors are also plotted on a contour plot. The generated data is dense near the modes of the 5-component Gaussian mixture.

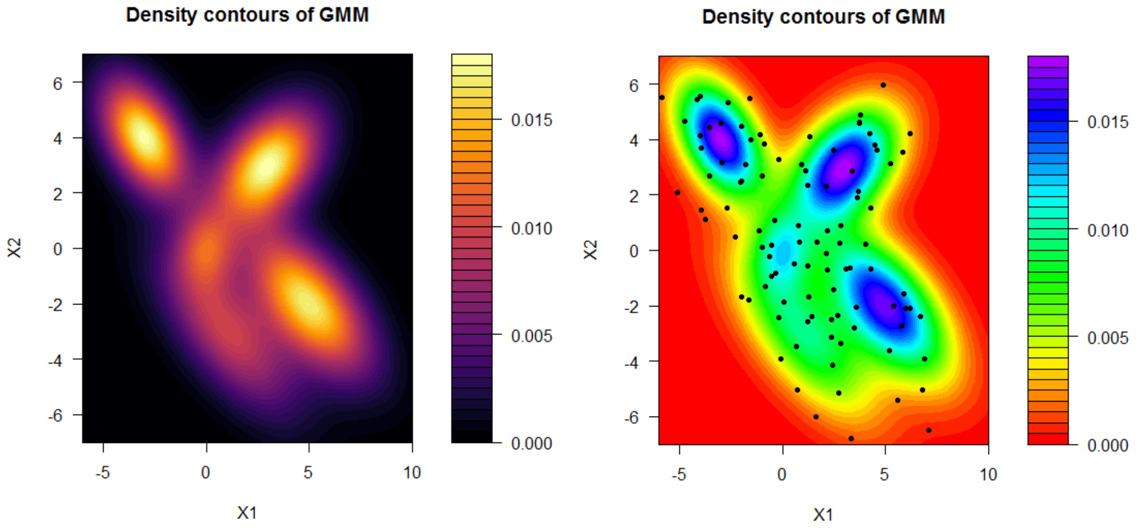


Figure 5.1: Density contours of bivariate Gaussian mixture model

Given this density, the Ψ-transform method relies on simulating many observations from the GMM. Hence, $B = 3000$ minimum probability density values were simulated from this GMM. Notice that this requires one to simulate 100 observations from a bivariate, 5-component GMM 3 000 times. The untransformed minima (in probability space) is highly skewed to the right. Most of the observations lead to low-density values close to zero. Only a few observations have (minimum) density estimates greater than 0.001.

Figure 5.2 displays the untransformed and the Ψ-transformed minimum probability density values. Furthermore, the probability density function of the Gumbel distribution is superimposed on the histogram of the transformed minima. The parameters for the Gumbel distribution were estimated using maximum likelihood estimation. This example shows that for a 5-component GMM, the Ψ-transform method accurately transforms the simulated minima to a space on which a Gumbel distribution can be fitted. Hence, the Ψ-transform method is applicable to situations where more than one multivariate Gaussian distribution is required to model the data. Finally, notice that the Gumbel probability density closely traces the histogram of the transformed minima.

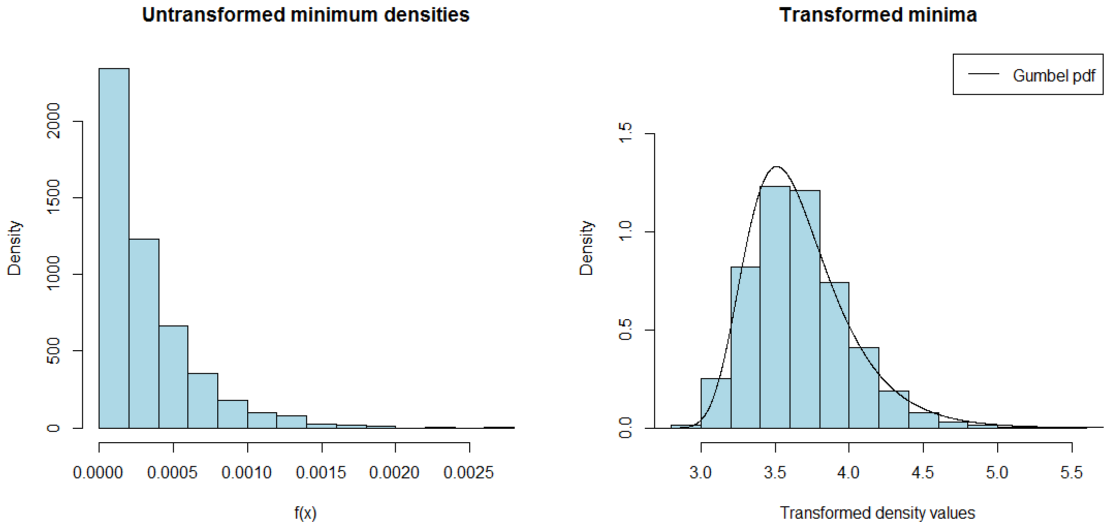


Figure 5.2: Transformed and untransformed minimum density values of bivariate GMM

It must now be determined if the Ψ-transform method can be used in higher-dimensional scenarios. To test this, the same simulation is done for a 6-dimensional, 2-component GMM. The properties for this model are

$\underline{\mu}_1 = \left(-4,6,2,0,1,6,\right)$, $\underline{\mu}_2 = \left(3,3,0,1,5,-2\right)$, and

$$\Sigma_1 = \begin{bmatrix} 4 & 1 & 0 & 0 & 2 & 0 \\ 1 & 7 & 1 & 2 & 3 & 1 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 3 & 2 & 0 \\ 2 & 3 & 1 & 2 & 5 & 1 \\ 0 & 1 & 0 & 0 & 1 & 2 \end{bmatrix}, \ \Sigma_2 = \mathrm{I}_6 .$$

Unfortunately, the same contour plots cannot be drawn for the 6-dimensional GMM. Therefore, only the histograms of the untransformed and transformed probability density values are given. For this example, mixing proportions of $\underline{\alpha} = \left(0.65, 0.35\right)$ and $B = 5000$ bootstrap repetitions were used. Figure 5.3 displays the histograms of the untransformed and transformed minima of the probability density values of the GMM. Clearly, the histogram of the transformed minima closely resembles the Gumbel distribution. Notice that the Gumbel probability density function is again superimposed on the histogram of the transformed probability density values by estimating the parameters via maximum likelihood. This example shows that the Ψ-transform method can be applied to high-dimensional GMMs. Thus, this extreme value-based novelty detection algorithm is appropriate for multimodal and multidimensional data.

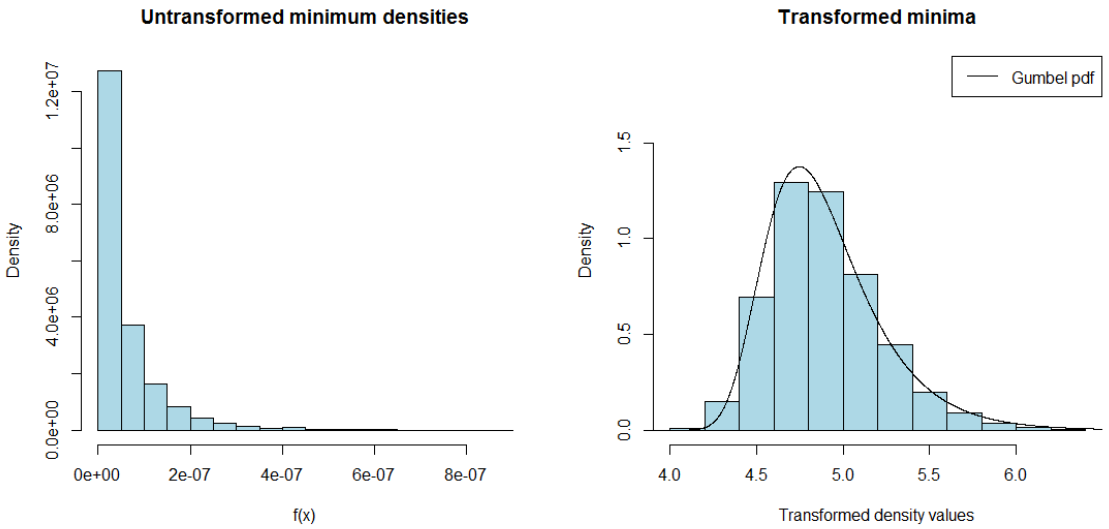


Figure 5.3: Transformed and untransformed minimum density values of a 6-dimensional GMM

Hence, a numerical extreme value-based novelty detection algorithm that is suitable for multimodal and multivariate data has been derived. Furthermore, this model has a fully probabilistic interpretation. Given that a new observation is classified as novel, it means that the probability – of the minimum of the probability density function of the normal class being below the probability density value implied by the new observation – is less than the specified confidence level. One disadvantage of this method is that numerous observations must be generated from a multidimensional GMM. Fortunately, Hugueny (2013) proposed some ideas to reduce the number of observations to be sampled from the GMM. These proposals are now discussed.

5.4 ADVANCES FOR THE Ψ-TRANSFORM METHOD

As mentioned towards the end of Section 5.3, the Ψ-transform method lacks the training speed required to handle large datasets. Today's state-of-the-art algorithms have both the properties of high prediction accuracy and a (relatively) quick training time. However, Hugueny (2013) made a few proposals to improve the efficiency of the Ψ-transform method. Hugueny (2013) first discussed a possible approximation for the distribution of the probability density of a GMM. Consequently, the distribution of $f(E_N)$ scaled by its normalising constants can be approximated with a simpler function. Under this approximation, Hugueny (2013) proposed a method termed tail-fitting via least squares to approximate the limiting distribution of the minimum probability density of a GMM. This method does not require one to generate bootstrap samples. Hence, it is faster. Furthermore, this proposal sheds some light on the justification of the method discussed towards the end of Section 5.5.

5.4.1 The Asymptotic Gaussianity in Density assumption

Clifton (2009) and Clifton *et al*. (2011) argued that the closest individual Gaussian component is not sufficient to describe the extrema in probability space. Intuitively, if some observation is far away from all the modes of a GMM, the density estimate of this sample will be low. Furthermore, this observation is with high probability in data space in a region where little overlap between the Gaussian components occur. Hence, the distribution of densities in this region of low densities should be well approximated by a single Gaussian component. The assumption is termed the Asymptotic Gaussianity in Density (AGD) assumption by Hugueny (2013).

This assumption essentially says that if the resulting probability density estimate is sufficiently low, it is valid to assume that the overlap at this region in the data space is very small. Therefore, the distribution describing the probability that this low probability density estimate is novel would be well approximated by that of a single Gaussian distribution. The distribution of the latter has a closed-form expression. In turn, the problem is simplified to the scenario in Section 5.2.3 where only one multivariate Gaussian distribution is considered.

Hugueny (2013) validated this assumption empirically. He generated data from a complex two-dimensional GMM and determined the corresponding probability density values. It is then observed that the histogram of the probability density values is approximately uniform if the densities are below a certain threshold. Alternatively, it can be thought of that the density estimates adhering to the AGD assumption are sufficiently far away from each mode. Notice that if $\underline{X} \in \mathbb{R}^2$ is a bivariate Gaussian random vector, the distribution and density functions in (5.14) and (5.15), respectively, are that of a uniform distribution.

Formally, the AGD assumption states that if $f\left(\underline{x}\right) \in P_f$ is the probability density function of a GMM, a sufficiently low constant $\rho \in P_f$ exists such that the distribution of $f\left(\underline{X}\right) \middle| f\left(\underline{X}\right) \leq \rho$ is approximately $G_d\left(\cdot, f_{Gaus}\right)$. Notice that P_f is the set of values the density of the GMM can take on while f_{Gaus} refers to the probability density function of a multivariate Gaussian distribution.

It is now discussed how this assumption is utilised to improve the Ψ-transform method.

5.4.2 Tail-fitting via least squares

Recall that the Ψ-transform method requires one to simulate data from a GMM. This section describes the proposal of Hugueny (2013) to improve the speed of this algorithm. Under the AGD assumption, the distribution in the Ψ-transform can be approximated without generating B bootstrap samples.

Consider a sample $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ and let the density of this data be described by a multivariate GMM denoted by $f(\underline{x})$. Assume that the AGD assumption holds. Therefore, assume there exists a $\rho \in P_f$ such that $f(\underline{X}_j) | f(\underline{X}_j) \le \rho$, $j = 1, \ldots, N_\rho$ are similarly distributed as in the case if $f(\underline{x})$ was a multivariate Gaussian density function. Notice that N_ρ is the number of density estimates below $\rho \in P_f$. This means that the distribution and density functions of $f(\underline{X}_j) | f(\underline{X}_j) \le \rho$, $j = 1, \ldots, N_\rho$, respectively, can be approximated with

$$G_d(y, f) = \Omega_d |\Sigma|^{1/2} \int_0^y \left[-2 \cdot \ln(C_d u)\right]^{(d-2)/2} du \text{, and,}$$

$$g_d(y, f) = \Omega_d |\Sigma|^{1/2} \left[-2 \cdot \ln(C_d y)\right]^{(d-2)/2} \quad , y \in \left(0, \max_{\underline{x}} \{f(\underline{x})\}\right).$$

These two equations were equations (5.14) and (5.15). Additionally, Hugueny (2013) argued that this density depends on the underlying covariance of the Gaussian distribution only through the determinant of this matrix. Therefore, this determinant can be approximated by a constant. Hence, for some low value $\rho \in P_f$, the density of $f(\underline{X}_j) | f(\underline{X}_j) \le \rho$, $j = 1, \ldots, N_\rho$ is approximated by

$$\tilde{g}_d(y, \lambda) = \Omega_d \lambda \left[-2 \ln\left((2\pi)^{d/2} \lambda y\right)\right]^{\frac{d-2}{2}} . \tag{5.59}$$

Notice that the square root of the determinant of the covariance matrix has been approximated by $|\Sigma|^{1/2} = \lambda$. Hugueny (2013) stated that maximum likelihood estimation is not feasible to estimate the parameter in (5.59) as this approximation is only valid for a small fraction of density values $f(\underline{x}_i) \le \rho$. Therefore, it is proposed to use a regression approach. This algorithm is given in Algorithm 5.1.

5.4.2.1 Algorithm 5.1

Consider the multivariate dataset $\left\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\right\}$. Assume that a GMM denoted by $\hat{f}\left(\underline{x}\right)$ describes the density associated with each iid $\underline{X}_i$, $i = 1,\ldots,N$ sufficiently well. The following steps are proposed by Hugueny (2013):

1. Generate N observations from the multivariate GMM. Denote these vectors by $\left\{\underline{x}_1^*, \underline{x}_2^*, \ldots, \underline{x}_N^*\right\}$. Each vector is a d dimensional observation from $\underline{X}$. Compute the densities of this sample associated with the GMM and denote them by $y_i^* = \hat{f}\left(\underline{x}_i^*\right)$, $i = 1,\ldots,N$.
2. Plot a histogram of the y_i^* , $i = 1,\ldots,N$ and inspect where the modes of the empirical density of Y^* occur.
3. Select a $\rho \in P_f$ which is the lower percentile of the density values that are assumed to adhere to the AGD assumption. Note that ρ is non-zero and smaller than the smallest y^* at which there is a mode.
4. Subset the densities which are less than ρ. Assume there are N_ρ densities for which $y_i^* \leq \rho$ and that this set is given by $\left\{\tilde{y}_1^*, \tilde{y}_2^*, \ldots, \tilde{y}_{N_\rho}^*\right\}$.
5. Plot a histogram of $\tilde{y}_i^*$, $i = 1,\ldots,N_\rho$ (normalised to have an area equal to the mass of $y_i^* \leq \rho$) consisting of B bins with centres denoted by b_i , $i = 1,2,\ldots,B$. Let the corresponding histogram values at the centre of each bin be $g^*\left(b_i, \hat{f}\right)$, $i = 1,\ldots,B$.
6. Estimate λ by performing least squares regression on the $g^*\left(b_i, \hat{f}\right)$, $i = 1,\ldots,B$ using equation (5.59). Thus, this parameter estimate is
$$\hat{\lambda} = \underset{\lambda>0}{\arg\min}\left\{\sum_{i=1}^{B}\left(g^*\left(b_i, f\right) - \Omega_d \lambda\left[-2\ln\left(\left(2\pi\right)^{d/2}\lambda b_i\right)\right]^{\frac{d-2}{2}}\right)^2\right\}. \qquad (*)$$
7. The estimated probability density function of the GMM densities below ρ is given by $\tilde{g}_d\left(\cdot, \hat{\lambda}\right)$. Furthermore, the distribution function is given in equation (5.26.1) and (5.26.2) with $\left|\Sigma\right|^{1/2} = \hat{\lambda}$.

Hence, a simple least squares regression approach has been formulated to approximate the distribution of the densities in areas where the densities of the GMM are sufficiently low to satisfy the AGD assumption. Most importantly, this method does not require the generation of B multivariate samples of size N. Therefore, this method is applicable to datasets that have many observations and are of high dimension.

5.5 DISTRIBUTION OF EXTREME DENSITIES USING THE MODERN APPROACH

The final method of this chapter considers the modern approach of extreme value theory to model the probability distribution of the lower-density estimates of a random variable. Using the peaks-over-threshold (POT) method of extreme value theory for novelty detection has only recently been defined. This method has been validated empirically and practically in the literature. Only one academic article has been published on this topic. Therefore, there is much room for further theoretical justification and exploration of this method. The article referred to here is that of Clifton, Clifton, Hugueny and Tarassenko (2014).

This section aims to highlight the advantages of the modern approach of extreme value theory when novelty detection is the goal. Although there is no general proof for this method yet, some remarks about the theoretical justification are given. It is shown how this method could be applied if the underlying distribution is a multivariate Gaussian distribution. Finally, it is argued that a similar approach can be taken in the case of a multivariate GMM.

Recall from Chapter 3 that if a distribution F is in the maximum domain of attraction of the GEV distribution, the exceedances of this distribution above a sufficiently high threshold is well approximated by the generalised Pareto (GP) distribution. This assumption is strongly supported by the Pickands-Balkema-de Haan theorem stating that, as in the Chapter 3 equation (3.29), the distribution F is in the domain of attraction of the GEV distribution if, and only if, for some auxiliary function $b(\cdot)$ and for all $1+\gamma v>0$,

$$\frac{1-F\left(z+b(z)v\right)}{1-F(z)} \rightarrow \left(1+\gamma v\right)^{-1/\gamma} \text{ as } z \rightarrow x_u . \tag{3.29}$$

Furthermore, under (3.29),

$$\frac{b\left(z+b(z)v\right)}{b(z)} \rightarrow u^{\gamma} = 1+\gamma v .$$

Notice that $Z = X - t \mid X > t$ for some high threshold t to avoid confusion with the notation of the definition $Y = f\left(\underline{X}\right)$. This theorem ultimately finds a limiting distribution for the peaks above a sufficiently high threshold. Therefore,

$$\bar{F}_t(z) = P(Z > z) \sim \left(1 + \gamma \frac{z}{b(t)}\right)^{-1/\gamma}, \text{ for large enough } t.$$

Three parametric classes of GP distributions were recovered which depended only on tail-heaviness – the EVI. For our purposes, the Gumbel type GP distribution will be used. For the Gumbel class of GP distributions and a sufficiently high threshold, the distribution of $Z = X - t | X > t$ is approximated by

$$P(Z > z) \approx 1 - \exp\left\{-z/\beta\right\}, \; z > 0, \; \beta > 0. \qquad (5.60)$$

In equation (5.60), the parameter $\beta \in \mathbb{R}^+$ is a scale parameter.

It is now argued why the modern approach of extreme value theory is suitable and advantageous compared to the classical approach of extreme value theory for novelty detection. The Pickands-Balkema-de Haan theorem is then utilised to construct a probabilistic novelty detection algorithm.

5.5.1 Why use the modern approach of extreme value theory?

In Chapter 3, it was pointed out that the block-maxima approach has the disadvantage that only the maximum of each block is used to estimate the parameters of the GEV distribution. Consequently, information regarding the tail data is ignored. One way to obtain better parameter estimates is to use the upper-order statistics to estimate the GEV parameters. Alternatively, the peaks-over-threshold method uses more data by approximating the distribution of the exceedances above a high enough threshold.

Consider a random variable X for which a large value indicates a novelty in terms of the normal state. A threshold t can be selected such that if $P\left[X - t > x - t | X > t\right] \le \alpha$, the observation x is considered novel. Thus, a novelty detection threshold is set probabilistically in terms of the exceedances above a high enough threshold. The threshold is selected such that all data below it is considered normal. Furthermore, exceedances above the threshold are not necessarily novel observations. These exceedances are regarded as tail data. Hence, only these data points are used to discriminate between novel and normal observations (Clifton *et al*., 2014). Note that a novelty is detected as an observation in the far tail of the exceedance distribution.

The advantage of this approach is that all observations that are regarded to be exceedances are used to fit the GP distribution. Consequently, more information is used than in the case of the GEV for the minimum density. Therefore, it is expected that novel observations can be detected more precisely because the extremes of the normal class are modelled more efficiently.

Clifton *et al*. (2014) stated that this approach is a mixture of a discriminative and generative approach. It is generative in the sense that a generative distribution is assumed for the normal class. From this distribution, a distribution for the exceedances is approximated. The algorithm is discriminative from the point of view that only the data exceeding the threshold is used to build the model. Hence, only data close to the decision boundary is used to discriminate between normal and novel observations.

The next section describes how the POT method is used to threshold the density function of the normal class if the underlying distribution is a multivariate Gaussian distribution. The approach followed in Sections 5.5.2 and 5.5.3 differs from that of Clifton *et al*. (2014). However, the main results are found in that article.

5.5.2 Modern extreme value theory for multivariate Gaussian distributions

Consider a multivariate Gaussian random vector $\underline{X} \in \mathbb{R}^d$ and let the probability density function be denoted by $f\left(\underline{x}\right)$. It was shown in Section 5.2.3 that the distribution of $f\left(\underline{X}\right)$, denoted by $G_d\left(\cdot,f\right)$, is in the domain of attraction of the Weibull distribution (considering the minimum of f). Clifton *et al*. (2014) used this fact to argue that the distribution of the exceedances of $f\left(\underline{X}\right)$ below a sufficiently small threshold is distributed as the Weibull class of GP distributions. Although this is theoretically sound, the density estimates of $f\left(\underline{X}\right)$ below a low threshold are extremely small. Hence, it becomes more difficult to, firstly, select the threshold and, secondly, work with such small exceedances. Therefore, a transformation is applied to the density estimates before applying the POT method.

It is known that $G_d\left(\cdot,f\right)$ is in the domain of attraction of the extremal Weibull distribution with an EVI of -1. It then follows that the transformation $T\left(\underline{x}\right) = -\ln\left[f\left(\underline{x}\right)\right]$ would have a distribution in the domain of attraction of the Gumbel distribution. This is stated in Hugueny (2013: 148). Therefore, define the transformation

$$T\left(\underline{x}\right) = -\ln\left[f\left(\underline{x}\right)\right]. \tag{5.61}$$

Given that the distribution of the maximum of (5.61), centred and scaled appropriately, is approximately a Gumbel distribution, the exceedances of (5.61) above a high enough threshold is distributed as the Gumbel class of GP distributions. This follows directly from the equivalence conditions given in De Haan and Ferreira (2006). Additionally, notice that the transformation implies that as $f(\underline{x})$ decreases, the function in (5.61) increases. Hence, the exceedances above a high enough threshold are of concern.

Consider a sample of random vectors $\{\underline{X}_i\}_{i=1}^{N}$ assumed to be sampled from a multivariate Gaussian distribution with probability density function $f(\underline{x})$. Find the density values $f(\underline{X}_i)$, $i=1,\ldots,N$ and compute their corresponding transformations $T(\underline{X}_i)$, $i=1,\ldots,N$. Select a high enough threshold u and define the exceedances as

$$Z_j = T(\underline{X}_j) - u \,|\, T(\underline{X}_j) > u \text{ , } j=1,\ldots,N_u . \tag{5.62}$$

Notice that N_u is the number of observations above the threshold. These exceedances are expected to be GP distributed. Therefore, a limiting distribution for the exceedances of a multivariate Gaussian probability density below a certain threshold has been recovered.

To validate this reasoning, consider generating 10 000 vectors from a d-dimensional Gaussian distribution. To each generated random vector, a noise component from a standard Gaussian is added. Thereafter, the parameters of the Gaussian distribution are estimated and the transformed density values are determined as $\hat{T}(\underline{x}_i)$, $i=1,\ldots,10\,000$. The threshold is selected such that only 10% of the data exceeds it. Consequently, the exceedances are given by

$$\hat{z}_j = \hat{T}(\underline{x}_j) - u \,|\, \hat{T}(\underline{x}_j) > u \text{ , } j=1,\ldots,N_u . \tag{5.63}$$

These exceedances are expected to be Gumbel distributed. Figure 5.4 displays the histograms of the exceedances of 9 multivariate Gaussian distributions. The dimension of each simulation is given below the histograms. Furthermore, the probability density function of the Gumbel class of GP distributions is superimposed on each histogram. Maximum likelihood was used to estimate the scale parameter of the GP distribution. Finally, it must be mentioned that the covariance of each of the Gaussian distributions had non-zero off-diagonal entries – the variables were assumed to be correlated.

It is clear that the Gumbel class of GP distributions fits the transformed exceedances very well. The probability density function superimposed on the histograms closely traces the centres of the histograms where the dimensions ranged from 1 to 9. This example demonstrates that the synthetic data has been efficiently transformed to a univariate space onto which a GP distribution can be fitted.

It has thus been validated through simulation that over this range of Gaussian distributions the GP distribution is a good approximation of the distribution of the (transformed) exceedances. This distribution has been constructed by equivalently using the exceedances of the probability density values below a given threshold. For a new observation $\underline{x}^*$, the first step is to obtain the corresponding exceedance. This exceedance is

$$z^* = \max\left\{-\ln\left[\hat{f}\left(\underline{x}^*\right)\right] - u, 0\right\}. \tag{5.64}$$

If the exceedance is zero, the observation is assumed to belong to the normal class. Conversely, if the exceedance is positive, the observation is regarded as tail data and must be further investigated. Assuming the exceedance is positive, the new observation is regarded as novel if

$$P\left(Z > z^*\right) < \kappa. \tag{5.65}$$

Notice that κ is some high probability.

It is expected that thresholding the distribution describing the normal class using this approach would be more efficient in detecting novelties than in the case of the GEV distribution. This is because more data is used to find the decision boundary. However, the disadvantage is that the distribution of the normal class is assumed to be Gaussian. It might be that this assumption is too restrictive to determine novelties effectively. Therefore, it is now discussed how this approach can be used if the density function of the normal class is assumed to be a GMM.

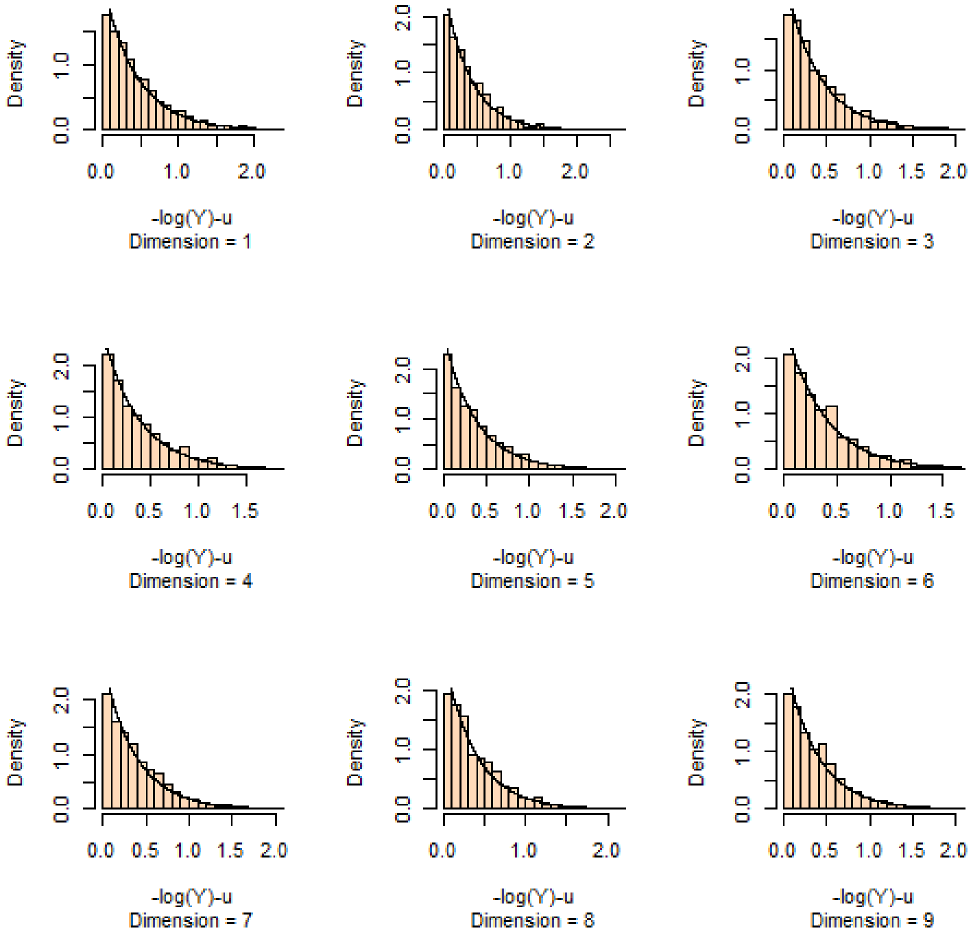


Figure 5.4: Probability density function of exceedances of multivariate Gaussian density values

5.5.3 Modern extreme value theory for mixtures of multivariate Gaussian distributions

This section discusses a possible approach to apply the POT method to the probability density of a GMM. It will be shown that under an appropriate assumption the same approach as in Section 5.5.2 can be followed.

Assume that $\underline{X} \in \mathbb{R}^d$ is a random vector generated from a GMM with density function $f(\underline{x})$. In Section 5.4.1 the AGD assumption was used to approximate the distribution of $f(E_N)$ (scaled). Similarly, this assumption can be used to approximate the distribution of the exceedances of $f(\underline{X})$ below a sufficiently small threshold. Assume the AGD assumption holds for $f(\underline{X}) \leq \rho$. Based on this, the distribution of $f(\underline{X}) \leq \rho$ is approximately $G_d(\cdot, f)$ as in the multivariate Gaussian case. Therefore, if $f\left(\underline{X}\right)\leq\rho$ it must be that $-\ln\left[f\left(\underline{X}\right)\right]\geq-\ln\left(\rho\right)$ such that the results from Section 5.5.2 can be used. That is, it is assumed that the threshold selected in the transformed space is high enough such that only density estimates adhering to the AGD assumption result in exceedances above the threshold.

Consider a sample $\left\{\underline{X}_i\right\}_{i=1}^{N}$ and assume a GMM with probability density function $f\left(\underline{x}\right)\in P_f$ is a good approximation of the true probability density. Assume a $\rho\in P_f$ exists such that the probability density below this threshold satisfies the AGD assumption. Choose a threshold $u>-\ln\left(\rho\right)$ and define the transformed exceedances as

$$Z_j=T\left(\underline{X}_j\right)-u\middle|T\left(\underline{X}_j\right)>u\ \ ,j=1,\ldots,N_u\text{ , with} \tag{5.66}$$

$$T\left(\underline{X}\right)=-\ln\left[f\left(\underline{X}\right)\right].$$

Notice that it is assumed that there are $N_u<<N$ exceedances. Under the AGD assumption, the exceedances in (5.66) are approximately distributed as a Gumbel type GP distribution. This fact follows from the above discussion.

To validate whether this approach can be used to efficiently detect novelties a small simulation study was carried out. Nine simulations were performed where the dimension ranged from 2 to 15 and the number of components in the GMM ranged from 5 to 15. As a first step data was generated from a multivariate GMM. To each observation a noise component from a Gaussian distribution with a mean vector of zero and diagonal covariance matrix was added. Furthermore, each covariance matrix had non-zero entries for all the elements. Hence, the variables were correlated. The covariances were quite large and therefore the variance of each variable in the noise component was 20. Once the observations were generated the parameters of the GMM were determined. These parameters were used to estimate the probability density function of the GMM as

$$\hat{f}\left(\underline{x}\right)=\sum_{k=1}^{K}\hat{\alpha}_k\hat{f}_k\left(\underline{x},\underline{\hat{\mu}}_k,\hat{\Sigma}_k\right). \tag{5.67}$$

Notice that the k^{th} estimated Gaussian component with its corresponding estimated mean vector and covariance matrix is denoted by $\hat{f}_k\left(\cdot,\underline{\hat{\mu}}_k,\hat{\Sigma}_k\right)$. The next step was to select an appropriately high threshold u and define the transformed exceedances as

$$\hat{Z}_j = \hat{T}\left(\underline{X}_j\right) - u \left| \hat{T}\left(\underline{X}_j\right) > u \right., \ j = 1, \ldots, N_u \text{, with} \qquad (5.68)$$

$$\hat{T}\left(\underline{X}_j\right) = -\ln\left[\hat{f}\left(\underline{X}_j\right)\right].$$

Given that the threshold is high enough such that the (-log) densities that do exceed the threshold satisfy the AGD assumption, the exceedances in (5.68) are approximately distributed as a Gumbel type GP distribution.

Figure 5.5 displays the histograms of the transformed exceedances with the Gumbel type GP probability density function superimposed. The dimension and number of components are given below each graph. Even with the artificial noise added to the data the density function of the Gumbel type GP distribution closely traces the histograms for all the cases. The number of observations generated was again 10 000 and the threshold selected such that 10% of the data exceeds it. This simulation provides some justification that the modern approach of extreme value theory is a plausible method to threshold the distribution describing the normal class if the underlying density function is modelled with a GMM.

As in Section 5.5.2, for a new observation $\underline{x}^*$ the first step is to find the corresponding exceedance given by

$$z^* = \max\left\{-\ln\left[\hat{f}\left(\underline{x}^*\right)\right] - u, \ 0\right\}. \qquad (5.69)$$

If this exceedance is zero, the observation is classified as belonging to the normal class. Conversely, if the exceedance is positive it is investigated whether such an exceedance is too large to represent an exceedance from the normal class. Hence, the observation is classified as novel if

$$P\left(Z > z^*\right) < \kappa . \qquad (5.70)$$

Again, the quantity κ is some high probability.

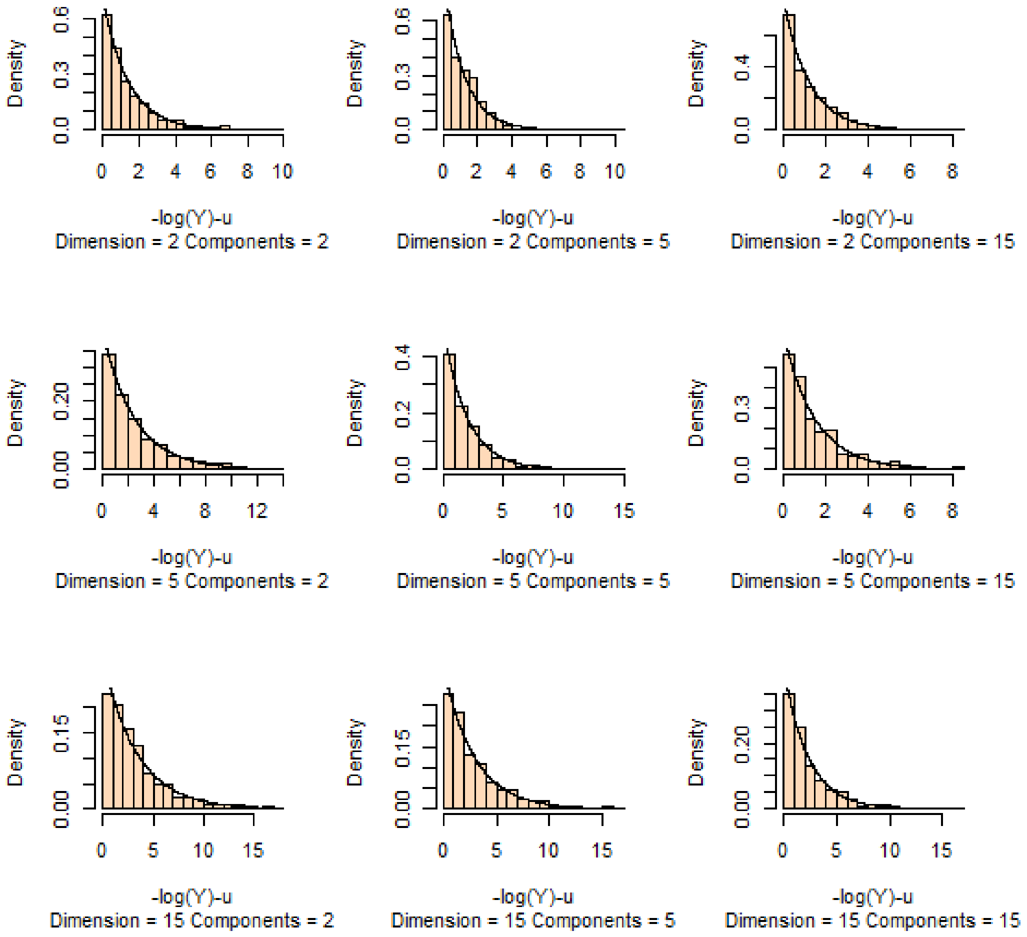


Figure 5.5: Probability density function of exceedances of multivariate GMM

This section demonstrated one approach that can be used to perform probabilistic novelty detection with the modern approach of extreme value theory. Although very little research has been done on this topic, it is believed that research in this area will lead to powerful methods for novelty detection relying on the modern approach of extreme value theory.

5.6 CONCLUSION

This chapter and Chapter 4 investigated the most recent methods to perform novelty detection with extreme value theory. In Chapter 4 it was shown that univariate extreme value theory can be applied to the Mahalanobis distances resulting from a multivariate Gaussian distribution. Although this method lacked important qualities, this idea was carried forward in Chapter 5. Clifton (2009) and Clifton *et al*. (2011) showed how extreme value theory can be applied to the distribution of the probability density values. This interesting way of thinking about probabilistic novelty detection has led to powerful algorithms that can perform novelty detection in multimodal and, probably more importantly, multivariate cases.

The common theme in these approaches is to map multivariate data to some univariate space where extreme value theory can be applied. If the data is Gaussian distributed, the maximum Mahalanobis distance and minimum probability density value approaches are equivalent. This remark was highlighted in Clifton (2009) and Clifton *et al*. (2011). However, what makes the minimum probability density value approach so attractive is that it incorporates the global contribution of the distribution when discriminating between normal and novel data. That is, in the case of a GMM, the contributions of each Gaussian component are used to discriminate between normal and novel observations.

In Section 5.2 an extreme value-based novelty detection algorithm was derived for multivariate Gaussian distributions. The main assumption was that the distribution generating the normal class is well defined by a multivariate Gaussian distribution. Given that this assumption is satisfied, this model would be powerful in discriminating between novel and normal observations. Furthermore, the fully analytical justification for the model improves the interpretability of the model. For example, expressions for the parameters of the limiting Weibull (GEV) distribution can be obtained. This is because the exact form of the distribution of the probability density values is known. However, it is seldom the case that the distribution of data can be accurately modelled by a single, multivariate Gaussian distribution. Therefore, this approach would break down if the Gaussian distribution is too rigid for the data. Consequently, more sophisticated approaches are required.

Section 5.3 considered the numerical scheme to perform extreme value-based novelty detection in multimodal and multivariate cases. Remarkably, this approach seems unaffected by the number of modes or the dimension of the data. However, the approach relies strongly on two steps. The first step is to fit a multivariate GMM to the data. If this model does not describe the density values of the data well, the Ψ-transform would do poorly in discriminating between novelties and normal observations. Therefore, this step should be performed carefully with adequate testing of the goodness-of-fit of the GMM. Once this step has been completed, many observations must be sampled from the estimated GMM. Note that if the data is high dimensional and there are many components, many bootstrap samples are required to estimate the location and scale of the Gumbel distribution. Clifton *et al*. (2011) used 10 000 bootstrap samples throughout their work. Both these steps can become slow if the data is highly complex.

Fortunately, Hugueny (2013) improved the computational time of the model with a method he termed tail-fitting via least squares. This approach was based on the AGD assumption. Under this assumption the problem is simplified to that of the multivariate Gaussian case. Closed-form expressions for the distribution and density functions of $f\left(\underline{x}\right)$ exist. Hence, it is argued that this method is a suitable candidate for novelty detection if the sample size or dimension is very large.

Finally, the modern approach of extreme value theory was used to derive similar novelty detection algorithms. The full equivalence between the Fisher-Tippett and Pickands-Balkema-de Haan theorems implies that since the GEV can be used a similar result should hold when exceedances as opposed to block-maxima are considered. This method is still very new and requires deeper research to improve the theoretical justification and understanding of the method. Nevertheless, the simulation study clearly showed that the proposed method works well on synthetic data.

Novelty detection can be elegantly performed using extreme value theory. Looking at these approaches as one, the advantage is that the threshold has a probabilistic interpretation and does not depend on the sample size. Future research could look at extending these models using second-order theory. This would shed light on the rate of convergence of the distribution of the minimum density values to the GEV distribution. Furthermore, Clifton (2009) and Clifton *et al*. (2011) only considered Gaussian distributions or mixtures thereof. Using the definition of an extremum in probability space, the expressions of the distributions of the probability density values of non-Gaussian distributions can be derived. For example, the student-t distribution could be considered to formulate a more robust model.

Robustness is an important property that improves the generalisation power of a model. Based on the tail-fitting via least squares approach of Hugueny (2013), robust regression approaches could be used to improve this model. For example, other loss functions than the squared error loss could be used. In turn, the errors can be penalised more severely or more lightly. Moreover, it might prove useful to penalise the complexity of this assumed function by regularising the parameter in the model.

Finally, one broad research area in terms of these approaches to novelty detection is to generalise the models to such an extent that the GMM is no longer needed. For example, cluster analysis and variations thereof are powerful unsupervised learning techniques. Perhaps this class of models can be used to model the data describing the normal class. Extreme value-based novelty detection approaches could then be derived to test when observations do not belong to any of the clusters. Alternatively, probabilistic graphical models could be used to approximate the underlying probability density of the normal class.

The next chapter investigates real-world datasets as a practical application of the techniques discussed in this dissertation. The practical application of these techniques will also be demonstrated and the usefulness of the models will be explored.

CHAPTER 6

PRACTICAL APPLICATION OF EXTREME VALUE-BASED NOVELTY DETECTION

6.1 INTRODUCTION

The methods developed in Chapter 5 will now be applied to a dataset. Before the data is analysed, some definitions that will be used to build and test the model are given. In general, extreme value-based novelty detection consists of two steps. Firstly, the distribution describing the normal class must be estimated. Thereafter, a threshold of normality is set on the distribution describing the normal class. These steps will be followed explicitly in this chapter.

The dataset analysed in this chapter is a banknote authentication dataset. This dataset has been modified to include none of the forged banknotes during training. It is argued that the approaches discussed in this dissertation adequately discriminate between real and forged banknotes. The major advantage of a novelty detection algorithm as opposed to a supervised, two-class approach is that no samples of forged banknotes are required to build the model. Consequently, sampling data to train the model can be done in an inexpensive manner.

The next section of this chapter summarises the definitions and procedures that will be followed to build and validate the model. This section therefore serves as a foundation on which the models will be trained and tested. In Section 6.3 the banknote authentication dataset is described and the model representing the real banknotes (normal class) is constructed. Section 6.4 considers the different methods explained in Chapter 5 to threshold the model of normality. The first method considered is the Ψ-transform approach. It is argued that this method is effective in detecting forged banknotes. However, generating bootstrap samples from the GMM is inefficient. Therefore, the modern approach of extreme value theory is also used to threshold the estimated probability densities. The chapter is concluded with a comparison of the two broad approaches.

6.2 PRELIMINARIES

This section summarises the definitions and procedures that will be used to construct a novelty detection algorithm. The steps to train and to test the model will be considered separately.

6.2.1 Model training

As a first step, a GMM must be fitted to the data describing the normal class. To mitigate overfitting, the goal is to find the simplest model describing the distribution of the normal class accurately. Therefore, a penalised likelihood-based approach is used to train the GMM.

In the literature, various methods have been proposed to penalise the likelihood of the model. In this chapter, the Bayesian information criterion (BIC) is used to penalise the likelihood of the GMM. Let $f\left(\underline{x}\right)$ be the GMM with K Gaussian components. The k^{th} Gaussian component, with mean vector $\underline{\mu}_k$ and covariance matrix Σ_k, is denoted by $f_k\left(\underline{x},\underline{\mu}_k,\Sigma_k\right)$. Furthermore, the K mixing proportions are denoted by α_k , $k=1,2,\ldots,K$. Then, the BIC penalised (log-) likelihood of the model is

$$L_p\left(\underline{x}|\Theta\right)=L\left(\underline{x}|\Theta\right)-\frac{1}{2}N_d\left(K\right)\cdot\log\left(N\right). \qquad (6.1)$$

In equation (6.1), $L\left(\underline{x}|\Theta\right)$ is the unpenalised likelihood of the GMM with parameter space Θ consisting of the mixing proportions, mean vectors and covariance matrices, N is the sample size and $N_d\left(K\right)$ is the number of free parameters in a d-dimensional, K-component GMM. Notice that for a d-dimensional GMM with K components the number of free parameters is

$$N_d\left(K\right)=K\left[1+2\cdot d+\frac{d\left(d-1\right)}{2}\right]-1. \qquad (6.2)$$

Thus, there is a tradeoff between model complexity and goodness-of-fit. The unpenalised likelihood increases as more Gaussian components are considered. Therefore, not penalising the likelihood leads to models that are too complex and, consequently, result in overfitting. The penalisation subtracted from the likelihood penalises the number of free parameters and, hence, the number of components in the model. The optimal parameters in the model are those that maximise (6.1).

Once the GMM has been estimated, this approximation is used to implement the extreme value-based novelty detection algorithms. Given that these approaches are probabilistic approaches, there is no need to minimise some misclassification error on a validation set. Rather, a novelty detection threshold is set at some significance level. However, the posterior probability that an observation from the training data is normal can be used to make sure that the normal observations are not wrongly classified as being novel.

6.2.2 Model testing

After a model has been built, it must be determined whether the model generalises sufficiently well to be applied to new data. Therefore, a new test dataset is required to test the model. Let the observations in the data considered to be from the normal class be termed positives and the observations considered to be novel be termed negatives. The total accuracy of the model is then

$$Ac = \frac{\text{True positives} + \text{True negatives}}{N}. \tag{6.3}$$

The true positives are the normal data that has been classified as normal and the true negatives are the novel observations that were correctly classified as anomalous. Furthermore, N is the number of observations tested. Notice that the accuracy of the model summarises the proportion of test observations correctly classified. To determine where the model makes errors, the sensitivity and specificity of the model should also be checked. The sensitivity of the model is

$$Se = \frac{\text{True positives}}{\text{Positives}}. \tag{6.4}$$

This is the proportion of normal data classified as normal. Finally, the specificity is

$$Sp = \frac{\text{True negatives}}{\text{Negatives}}. \tag{6.5}$$

The quantity in (6.5) is the proportion of novel observations correctly detected. These three measures of model performance will ultimately quantify how well the trained model performs on test data. It is desirable to have all three measures as close to 1 as possible.

It can be argued that if the sensitivity is low and the specificity is high the model overfits. This means that the model is too dependent on the data on which it was trained. Hence, slight changes in the data are causing the model to classify an observation as novel. Alternatively, if the sensitivity is high and the specificity is low the model is too rigid. Consequently, the model is too simple to detect deviations from the class of normality.

The definitions, notations and procedures described in this section will be used throughout this chapter. In the next section, the extreme value-based novelty detection algorithms of Chapter 5 are applied to the banknote authentication data. This dataset considered is publicly available and appropriately referenced in the appendix.

6.3 BANKNOTE AUTHENTICATION

In this section, the results from Chapter 5 are used to construct different extreme value-based novelty detection algorithms that can be used to test whether a banknote is real or forged. This is an important problem in countries with high crime rates. Therefore, an effective banknote authentication algorithm can mitigate the risk of retailers accepting forged banknotes. Furthermore, only real banknotes are used to build the model. If this model effectively discriminates between real and forged banknotes it is concluded that none of the forged banknotes are required during training. Hence, data can be sampled inexpensively.

6.3.1 Description of the data

The banknote authentication dataset consists of 762 authentic banknotes and 610 forged banknotes. This dataset is divided into a training set that contains 500 authentic banknotes. These observations represent the normal class. The test data consists of the remaining 262 authentic and 610 forged banknotes. Although both classes are relatively well sampled, the goal here is to demonstrate how a banknote authentication algorithm can be constructed if only observations of real banknotes are available. Therefore, only real banknotes are included in the training data.

This dataset is extracted from the UCI Machine Learning Repository. To locate this dataset online refer to Doerksen (2012) in the reference list. Each image in the dataset was obtained with industrial cameras usually used for print inspection. Thereafter, the images were transformed with a discrete wavelet transform. The features for each image are extracted as the variance, skewness, kurtosis and entropy of the wavelet transformed image. All four predictor variables are continuous.

A general approach to extract features from images is the wavelet transform. This class of filtering techniques has the property of time and frequency localisation. Time and frequency localisation refers to the fact that the wavelets' bases efficiently represent both smooth and locally bumpy functions (Hastie *et al*., 2009). In turn, features are extracted from the wavelet transformations.

Discrete wavelet transforms are popular tools in image processing. It is assumed that an image consists of a signal component and a noise component. Using the definitions from computer science literature, discrete wavelet transforms decompose the signal of an image into a detail level and an approximation level. This is the first level of the decomposition. The first approximation level can then be further decomposed into the second detail and approximation level, and so forth. Next, the detail levels of the multilevel decomposition are investigated and a threshold is selected to filter the noise from the levels. Notice that the detail levels catch the high frequencies of the signal and the approximation levels capture the low frequencies of the signal. In turn, the frequencies of the detail levels below the threshold are set to zero, whereas the frequencies above this threshold are left unchanged (hard thresholding), or, shrunk towards the threshold (soft thresholding). Finally, the image is reconstructed using the inverse wavelet transform. The reconstructed image (or matrix of pixels) should only contain the signal describing the significant structure of the image.

The banknote authentication dataset gives the variance, skewness, kurtosis and entropy of the wavelet transformed images. For the normal class, it is assumed that the signal of each image is similar. Therefore, the features extracted from the wavelet transform should be similar. As opposed to this, the signal of forged banknotes should differ from the normal class. Thus, their features should differ significantly from that of the normal class. More importantly, the statistics of the forged banknotes are expected to differ from each other significantly. This is expected because these banknotes are fraudulent and have nothing in common other than being fake. Therefore, constructing a model to represent the forged class is not advised. The advantage of the wavelet transform is that the dimension has been reduced significantly. Each grayscale image is 400 x 400 pixels. Hence, there are 160 000 pixels or dimensions. However, using the features extracted from the wavelet transform, the dimensionality has been reduced to 4. Although this reduction seems too large, the wavelet transform is highly efficient in capturing the true structure in the data.

As a first step, a model must be constructed to describe the distribution (density) of the features of the real banknotes. It is assumed that the GMM is an appropriate model to estimate the density of the underlying normal class. This model is constructed in the next subsection.

6.3.2 Training the multivariate Gaussian mixture model

The first step is to estimate the probability density function of the data. Hence, assume that a Gaussian mixture model is an accurate model to approximate the probability density function of the transformed images of real banknotes. It is assumed that the Gaussian components in the GMM have full covariance matrices. Furthermore, initial estimates are obtained by specifying the number of components and assuming a uniform Dirichlet prior distribution for the mixing proportions. The training data is then randomly binned into distinct sets. Each bin is assumed to be a multivariate Gaussian from which the initial estimates of the mean vector and covariance matrix are obtained by using the empirical, unbiased estimate.

Recall that the EM algorithm is used to estimate the parameters. This algorithm produces a set of parameter estimates and the (log-) likelihood of the model. The BIC penalisation is then subtracted from the estimated likelihood.

To find the GMM that best describes the probability density of the data, the optimal number of Gaussian components must be determined. Although the validation set approach and k-fold cross-validation are plausible methods, the banknote authentication dataset is relatively small. It is not recommended to train a potentially complex GMM on an even smaller dataset. Notice that a GMM has many free parameters. Consequently, this model can overfit the training data if the optimisation is unpenalised. Therefore, the BIC penalised likelihood approach is followed.

A well-known concept in statistics is the principle of parsimony. That is, the best model is the simplest model that accurately describes the data. The first step is to specify a sequence of the possible number of components. For this dataset, it is assumed that the optimal number of components is one of $k=\left(2,3,4,5,6,7,8,9\right)$. Using this set of candidate number of components, the probability density function of the data is approximated with a GMM for each number of components in the set. Hence, for each specified number of components, the parameter estimates and the maximum log-likelihood are returned. The BIC criterion, which is a function of the dimension and k, is then subtracted from the maximum log-likelihood.

Table 6.1 lists the BIC penalised log-likelihood of the GMM for a differing number of components. Additionally, Figure 6.1 displays these results graphically. One challenge with training a GMM is that the final parameter estimates and the maximum likelihood are dependent on the starting values supplied to the EM algorithm. Therefore, the training procedure was repeated three times. Thereafter, the average of the penalised log-likelihoods of the three repetitions was computed for each specified number of components. It is believed that the relative differences between these averaged BIC penalised log-likelihoods are better suited to choose an optimal model.

Table 6.1: BIC penalised log-likelihood for Gaussian mixture models

k	2	3	4	5	6	7	8	9
Run 1	-5216.6	-4809.4	-4743.5	-4715.1	-4677.6	-4812	-4604	-4717.1
Run 2	-5117.5	-4966.6	-4918.1	-4765.3	-4740.2	-4773.5	-4745.6	-4842.4
Run 3	-5117.5	-5145.5	-4844.6	-4780.9	-4899.4	-4805.5	-4856.5	-4649.9
Average	-5150.5	-4973.8	-4835.4	-4753.8	-4772.4	-4797	-4735.4	-4736.5

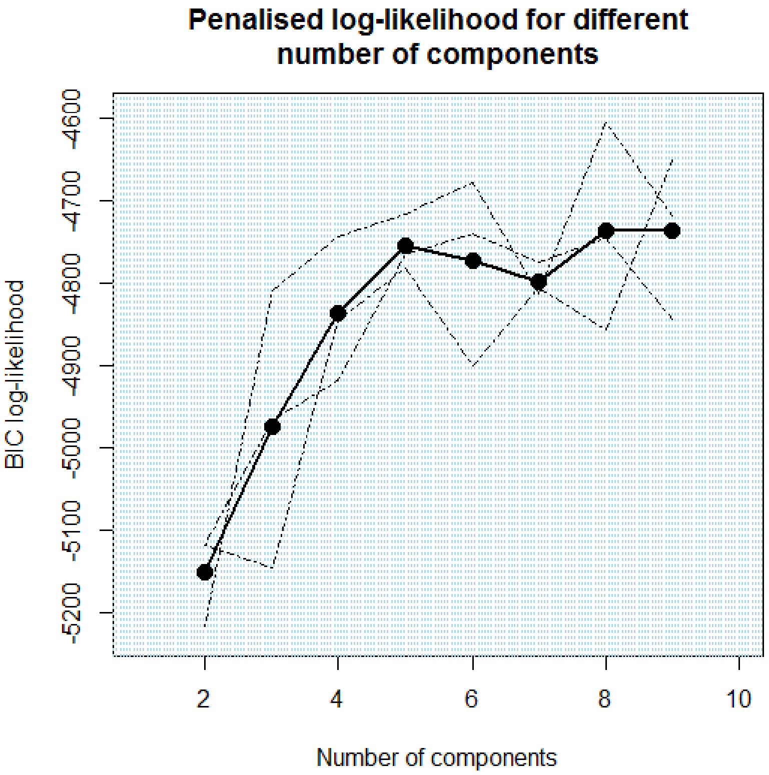


Figure 6.1: BIC penalised log-likelihood for Gaussian mixture models

The penalised log-likelihood increases abruptly as the number of components in the GMM increases from 2 to 5. After 5 components, the penalised log-likelihood seems to have levelled out. It is expected that after this point the increase in accuracy that results from increasing the number of components is insignificant. Using the principle of parsimony, the optimal number of components chosen is 5. Notice that by choosing more components one runs the risk of overfitting. Although the log-likelihood increases beyond 5 components, the increase is deemed insignificant.

Now that the optimal number of components has been estimated, the GMM is approximated using five components. The EM algorithm therefore approximates the mixing proportions, mean vectors and covariance matrices of the five Gaussian components. The estimated mean vectors obtained were

$$\begin{aligned}
\hat{\underline{\mu}}_1 &= \left(3.245 \quad \text{-}2.001 \quad 2.423 \quad 0.515\right) \\
\hat{\underline{\mu}}_2 &= \left(3.195 \quad 9.693 \quad \text{-}3.577 \quad \text{-}3.172\right) \\
\hat{\underline{\mu}}_3 &= \left(2.415 \quad 7.333 \quad \text{-}0.806 \quad \text{-}1.206\right) \\
\hat{\underline{\mu}}_4 &= \left(2.232 \quad 2.441 \quad 2.412 \quad 0.506\right) \\
\hat{\underline{\mu}}_5 &= \left(\text{-}0.011 \quad 4.896 \quad 2.878 \quad \text{-}3.262\right).
\end{aligned}$$

Furthermore, the estimated mixing proportions were

$$\hat{\underline{\alpha}} = \left(0.184 \quad 0.166 \quad 0.248 \quad 0.216 \quad 0.186\right).$$

Recall that only the determinants of the covariance matrices of each Gaussian component are needed. Therefore, only these quantities are given. The determinants of each estimated covariance matrix are

$$\begin{aligned}
\left|\hat{\Sigma}_1\right| &= 0.312 \\
\left|\hat{\Sigma}_2\right| &= 0.05 \\
\left|\hat{\Sigma}_3\right| &= 1.566 \\
\left|\hat{\Sigma}_4\right| &= 1.181 \\
\left|\hat{\Sigma}_5\right| &= 172.192.
\end{aligned}$$

These estimates are now used to build the extreme value-based novelty detection models. The different methods, as explained in Chapter 5, are now used to threshold the estimated GMM representing the probability density of the real banknotes.

6.4 EXTREME VALUE-BASED NOVELTY DETECTION OF BANKNOTES

Various approaches are now used to threshold the probability density function of the normal class probabilistically. As a result, discrimination between real and forged banknotes is done in a probabilistic manner.

6.4.1 The classical extreme value-based novelty detection algorithm

Now that the GMM has been trained, a threshold of normality must be set on the model describing the real banknotes. To threshold the model of normality the results from classical extreme value theory are used. That is, the first novelty detection algorithm constructed is the Ψ-transform model. Recall that this model requires parametric bootstrap samples to approximate the distribution of the minimum probability density. This model will be compared to the model using the modern approach of extreme value theory.

Let the approximated GMM be denoted by $\hat{f}\left(\underline{x}\right)$. Sampling from a GMM consists of two steps. Firstly, an indicator variable is sampled to indicate from which Gaussian component to sample. The probability that an observation is sampled from a specific component equals its estimated mixing proportion. Then, an observation is sampled from the selected Gaussian component. Consider generating $B = 10\ 000$ bootstrap samples of size $N_{test} = 872$ from $\hat{f}\left(\underline{x}\right)$. For each bootstrap sample the minimum probability density of that sample is stored and transformed with the Ψ-transform. Hence, the resulting bootstrap sample is

$$\Psi\left[\hat{f}\left(E_{N_{test}b}\right)\right]\ ,\ b = 1,\ldots,10\ 000\ . \tag{6.6}$$

It is expected that the sample in (6.6) is approximately Gumbel distributed. Under this assumption, the parameter estimates of the Gumbel distribution are

$$\hat{\eta} = 4.8052 \text{ and } \hat{\beta} = 0.2311. \tag{6.7}$$

Recall that, by definition, the Gumbel class of GEV distributions has an EVI of zero. These parameter estimates fully parameterise the extreme value model. Therefore, the probability that an observation is normal, or more specifically, the probability that a banknote is real, is given by

$$\hat{\breve{G}}_d^{\min}\left(y,f\right) = 1 - \exp\left\{-\exp\left[-\frac{\Psi\left(y\right) - \hat{\eta}}{\hat{\beta}}\right]\right\}. \tag{6.8}$$

A confidence level of 0.99 was chosen. That is, a real banknote is classified as forged with probability 0.01. For a new observation $y^* = \hat{f}\left(\underline{x}^*\right)$ the probability that this banknote is real is $\hat{\breve{G}}_d^{\min}\left(y^*,f\right)$. Therefore, if $\hat{\breve{G}}_d^{\min}\left(y^*,f\right) < 0.99$, the observation is classified as a forged banknote.

Based on this approach, the model classified 3 out of the 500 real banknotes in the training set as forged notes. Hence, the training error is approximately 0.6%. Using the test data, the probability density of each observation in this set is determined. Thereafter, these density estimates are used as input in (6.8). Finally, an observation is classified as novel if this probability is below 0.99. This approach led to a confusion matrix given in Table 6.2.

Table 6.2: Confusion matrix of banknote authentication test data

TRUE LABEL	PREDICTED	
	Real	Forged
Real	262	0
Forged	0	610

Clearly, the Ψ-transform performs very well on this data. The total accuracy of the model is

$$Ac = \frac{262 + 610}{872} = 1. \qquad (6.9)$$

Furthermore, the sensitivity and specificity are, respectively,

$$Se = \frac{262}{262} = 1 \text{ and } Sp = \frac{610}{610} = 1. \qquad (6.10)$$

Incredibly, the model correctly classifies all the test observations. Hence, the accuracy on the test set is 100%. Furthermore, the threshold set on the GMM implies that approximately 1% of the real banknotes is expected to be wrongly classified as novel. This is almost the case given that approximately 0.6% of the training data (which consists of only real banknotes) was classified as forged banknotes. Fortunately, none of the real banknotes were wrongly classified during testing and, more importantly, all the fake banknotes were detected.

It was shown that the numerical method of Clifton (2009) and Clifton *et al*. (2011) is highly effective in discriminating between real and forged banknotes. One disadvantage of this model is the fact that observations must be sampled from a complex GMM. Therefore, the modern approach of extreme value theory is utilised to speed up the training time of the model.

6.4.2 The modern extreme value-based novelty detection algorithm

In this section, the same steps are followed using the modern approach of extreme value theory. The advantage of this approach is that no resampling is done to fit the GP distribution. Using the same estimated GMM as before, each density estimate of the training set is transformed to

$$\hat{T}\left(\underline{x}_i\right) = -\log\left[\hat{f}\left(\underline{x}_i\right)\right] \ , i = 1,\ldots,600 . \tag{6.11}$$

The exceedances of the transformed density estimates above a high threshold u is then

$$\hat{z}_j = \hat{T}\left(\underline{x}_j\right) - u \left| \hat{T}\left(\underline{x}_j\right) > u \right. \ , j = 1,\ldots,10 . \tag{6.12}$$

The threshold was selected such that only 10 of the 500 real banknotes were regarded as tail data. Fitting a Gumbel type GP distribution to the exceedances in (6.12) resulted in an estimate of the scale parameter of $\hat{\beta} = 1.4397$. Hence, a new observation with a corresponding exceedance z^* is regarded as novel if

$$\hat{P}\left(Z > z^*\right) = \exp\left\{ -z^* \middle/ \hat{\beta} \right\} < \kappa . \tag{6.13}$$

Under this model, 10 out of the 500 training samples are classified as forged banknotes. Although this is undesirable, thresholding the model of normality in this manner implies that some of the training observations are classified as novel. However, testing the model on the test set of 262 real banknotes and 610 forged notes leads to acceptable results. Table 6.3 lists the confusion matrix of the test dataset using the modern approach of extreme value theory. It is shown that all the forged banknotes in the test dataset have been detected. However, 6 out of the 262 real banknotes in the test dataset were classified as forged banknotes. Using Table 6.3, the accuracy of the model is

$$Ac = \frac{256 + 610}{872} = 0.9931. \tag{6.14}$$

Furthermore, the sensitivity and specificity, respectively, are

$$Se = \frac{256}{262} = 0.9771 \text{ and } Sp = \frac{610}{610} = 1. \tag{6.15}$$

Table 6.3: Confusion matrix of banknote authentication test data II

TRUE LABEL	PREDICTED	
	Real	Forged
Real	256	6
Forged	0	610

Although the sensitivity is lower when the modern approach is used, it is still expected that only about 2% of the real banknotes will be falsely classified as novel and must, consequently, be expected manually. Additionally, the algorithm successfully detects all the forged banknotes. It is argued that the modern approach of extreme value theory is significantly faster than the numerical approach of Clifton (2009) and Clifton *et al*. (2011) and only slightly lacks the efficiency of the latter. Thus, the modern approach shows promising results.

6.5 CONCLUSION

This chapter served as a practical application of novelty detection with extreme value theory. The dataset was rather clean which increases the precision of the model. Novelty detection was performed probabilistically by thresholding the distribution (density) of normality with extreme value theory. The results indicated that extreme value theory is a valued method to perform novelty detection.

The first considered was the Ψ-transform. This method performed very well in detecting novel observations. However, the model requires many computations to train. The long training time is a consequence of the substantial number of bootstrap samples required to fit the GEV distribution. This problem, however, is negligible for lower-dimensional datasets or cases where the approximated density is a relatively simple function – for example, if the number of components in the GMM is low. A rather bigger concern is that the number of bootstrap repetitions required increases with model complexity. Therefore, in more complex scenarios than in this dataset, training efficiency may become troublesome. Nevertheless, this model performed remarkably well on the banknote authentication dataset.

A second approach using the modern approach of extreme value theory was also considered. This approach does not require bootstrap samples and, hence, has a significantly shorter training time. Additionally, competitive results were obtained. The modern approach of extreme value theory detected forged banknotes just as efficiently as the Ψ-transform method. Excluding the training time of the Gaussian mixture model, this approach took 2.22 seconds to execute all the required procedures. These same procedures took several minutes using the Ψ-transform method. Unfortunately, the false positive rate of the modern approach is slightly higher than with the numerical method – keeping in mind that the former model only misclassified approximately 2% of the real banknotes in the test dataset. It is noted that the dataset analysed in this chapter had only 500 observations. Hence, the number of samples regarded as tail data might be too low. This could be the reason why the numerical approach is slightly more accurate.

If all the classes of a population are well sampled, supervised algorithms are very powerful. However, in scenarios where rare events in a population need to be detected, it is seldom the case that the class of interest is well sampled. The approaches discussed in this dissertation require only observations from the well-sampled class to build a model for discrimination. Therefore, sampling can be done in an inexpensive and efficient manner. This is a significant advantage in cases where it is expensive or impossible to obtain a representative sample of a class.

CHAPTER 7

CONCLUSIONS AND FUTURE RESEARCH

This chapter summarises the findings from the previous chapters. The approaches are compared in terms of their assumptions, theoretical justification, predictive power and efficiency. In the concluding section, proposals for further research for novelty detection with extreme value theory are given.

7.1 A COMPARISON OF THE DISCUSSED APPROACHES

The main research objective of this dissertation was to discuss extreme value-based novelty detection from a statistical point of view. Recall that four methods were discussed, namely the Winner-Takes-All (WTA) method of Roberts (1999), the analytical and numerical approaches of Clifton *et al*. (2011), and the approach based on the modern extreme value theory of Clifton *et al*. (2014). This section gives a comparison of these methods.

7.1.1 Assumptions and theoretical justification

The main assumption of the WTA approach of Roberts (2009) is that, assuming a GMM sufficiently describes the density of the normal class, the probability that an observation is novel is dominated by the Gaussian component closest in the Mahalanobis distance. Clifton (2009) and Clifton *et al*. (2011) argued that this is not always the case. If there are regions where severe overlap between the components occur or the components have differing covariances, this assumption is not satisfied. However, the Asymptotic Gaussianity in Density (AGD) assumption was only proposed at a later stage. In essence, this assumption is equivalent to the assumption of Roberts (1999). That is, if the observation is far away from each Gaussian component’s centre, it is viable to assume only the closest component dominates the probability that an observation is novel.

Unfortunately, it is the parameter estimates proposed by Roberts (1999) that cause the model to break down. It was shown that the distribution of the Mahalanobis distance is in the maximum domain of attraction of the Gumbel class of GEV distributions. However, the parameters do not correspond with that from the one-sided normal distribution. Nevertheless, the theoretical justification of this model is strong. Therefore, if the appropriate parameter estimators are used, this model can detect novelties in complex settings.

The second approach discussed was the analytical approach of Clifton *et al*. (2011). This model assumed that the distribution of the normal class is a multivariate Gaussian distribution. Under this assumption, a closed-form expression for the distribution of the probability density of the Gaussian distribution was derived. In turn, the limiting distribution of the minimum probability density was shown to be the extremal Weibull class of GEV distributions.

An advantage of this method is its very strong theoretical justification. Closed-form expressions exist for the distribution and density function used to perform novelty detection. It is, however, frequently the case that a single Gaussian distribution is not adequate to model the distribution describing the normal class. Note again that the validation of the AGD assumption implies that this model would be adequate if the underlying density is modelled with a GMM. Therefore, this assumption might not be too restrictive.

To improve on the analytical approach, the numerical approach of Clifton *et al*. (2011) was discussed. This approach only assumed that the underlying probability density of the normal class is a mixture of Gaussian distributions. Unfortunately, not much theoretical justification for this method has been given explicitly. However, again the AGD assumption sheds light on its theoretical justification. As mentioned in Hugueny (2013), if the distribution of a random variable X is in the minimum domain of attraction of the Weibull class of GEV distributions with EVI -1, then the distribution of $-\ln X$ is in the maximum domain of attraction of the Gumbel class of GEV distributions. Using the AGD assumption, it can then be argued why the Ψ-transform maps the data onto a space where a Gumbel distribution can be fitted.

Finally, the modern approach of extreme value theory was used to construct novelty detection algorithms. Not much research has been done on this class of models. Therefore, concrete theoretical justifications have not been derived. However, given the equivalence between the GEV and GP distributions, it is argued that these models are theoretically supported in an analogous way as the previous methods.

A key insight into these models is the AGD assumption. This assumption connects these models such that the interpretation of the models becomes clearer. One way to think of this class of models is that the underlying assumption is that the tails of the normal class can be approximated by that of a Gaussian distribution. Hence, irrespective of which density was assumed for the normal class, all the results of this dissertation can be applied. Thus, an interesting question is which one of these methods is the most robust against violations of this assumption. Note that if this assumption is satisfied, it is the algorithm that best approximates the density of the normal class that will be optimal in detecting novelties. It is believed that the numerical approach is optimal if the AGD assumption is violated. This is because this model is built under the assumption that the probability density of the normal class is well approximated by a GMM.

7.1.2 Predictive power

Chapter 6 considered a banknote authentication dataset. Only the Ψ-transform and modern approach were considered. The chapter demonstrated that the proposed extreme value-based novelty detection methods have high predictive power. Of course, it is not the case that the predictive power will be this high on all datasets. However, this application certainly demonstrates the power of these methods. The sensitivity and specificity of both approaches were very close to unity.

In the case where the lower density estimates are approximately distributed as that of a Gaussian distribution, it is believed that the analytical approach or modern approach will be optimal to threshold the probability density describing the normal class. Notice that the underlying density may still be approximated with a GMM. It is only the lower density estimates that are assumed to be distributed as that of a single Gaussian distribution.

The predictive power of all these methods strongly depends on the accuracy of the density estimates. This can become a problem in highly complex scenarios. It is not a simple task to estimate a probability density function parametrically in high dimensions. Therefore, it is argued that this limitation of this class of methods must still be improved.

7.1.3 Model efficiency

A first remark on model efficiency relates to the use of the GMM. The assumption of the GMM has its advantages and disadvantages. Given the analytical method of a single Gaussian distribution and the AGD assumption, the GMM generalises the analytical approach to be suitable for much more complex datasets. However, the training time of a GMM can become very slow if the dimension and/or sample size is very large. Furthermore, validating the model also becomes troublesome.

If the data permits a GMM to be fitted, only the numerical approach can become inefficient. Generating bootstrap samples from a GMM is relatively slow. One observation never mentioned by Clifton *et al*. (2011) is the use of block-maxima as an alternative method to estimate the parameters of the Gumbel distribution. Bootstrap samples are generated only to estimate the location and scale parameter of the Gumbel distribution. If the sample size is large it might be possible to estimate these parameters accurately with the block-maxima method. This would speed up the computational time of this method significantly.

Notice that the other approaches do not require bootstrap samples. Therefore, these models are approximated with far less computations. As a result, computational time is not an issue for these models.

7.1.4 Final remarks

Novelty detection with extreme value theory is a new and exciting field of research. Although not much literature exists on this subject, current methods have demonstrated the advantages of using extreme value theory to perform novelty detection. There is still much room for improvements and advances in this area of research. It is the collaboration of statistics, computer science and creative thinking that will lead to powerful advances in this new research area.

7.2 FURTHER RESEARCH AREAS

Going forward, research on this class of models can focus on two specific aspects. The first aspect is improving the precision of the estimated probabilities that observations are novel. Another aspect is the method used to estimate the probability density of the normal class – keeping in mind that novelty detection is the ultimate goal. Some proposals on these two topics are now mentioned.

7.2.1 Methods to improve the precision of the classifier

Assume that the probability density function of the normal class is approximated by a GMM. Thus, the sample of estimated densities is found from the estimated GMM and denoted by $\left\{f\left(\underline{x}_i\right)\right\}_{i=1}^{N}$. The methods discussed in this dissertation (excluding the modern approach) approximates the limiting distribution of $f\left(E_N\right)$, shifted and scaled. All the methods discussed in this dissertation consider the theory of first order regular variation to approximate the parameters and, ultimately, the limiting distribution from a finite sample.

A first extension can be to consider the theory of second order regular variation to estimate the limiting distribution from a finite sample. If the AGD assumption holds, it is possible to investigate the rate of convergence of $f\left(E_N\right)$ to the minimal Weibull class of GEV distributions by using the derived closed-form expression for the distribution of $f\left(\underline{X}\right)$. Recall that this distribution was denoted by $G_d(\cdot, f)$. Furthermore, the second order theory allows one to use more (lower) order statistics to approximate the limiting distribution of the minimum probability density. For example, it is possible to derive expressions for the asymptotic bias as a function of the number of order statistics. In turn, the limiting distribution can be refined to incorporate the estimation error from a finite sample more appropriately. Note that if more order statistics are used to estimate the GEV distribution the variance of the model decreases. Therefore, of concern is the increase in the bias. However, by refining the limiting distribution via second order theory allows one to consider the asymptotic bias and, hopefully, reduce this unwanted quantity.

Another interesting extension could be a new way to utilise copulas. Copulas are used to model the covariance structure between random variables. In terms of this research, copulas can be used to model the tail-dependence between a set of random variables. Consider estimating the underlying density of the normal class with a GMM. Assume the optimal number of components is k^* and that this quantity is not very large (say less than 10). Still under the AGD assumption, it is then possible to extract a sample of densities for each cluster or component. One approach is to group the observations into clusters which consist of the observations closest to the centre of a cluster. This leads to k^* variables for which we each have a sample of densities. Denote these variables and samples, respectively, by $\left\{f_1(\underline{X}), f_2(\underline{X}), \ldots, f_{k^*}(\underline{X})\right\}$ and $\left\{f_j(\underline{x}_i)\right\}_{i=1}^{N_j}$, $j=1,\ldots,k^*$ where N_j is the number of observations belonging to the j^{th} cluster. Instead of only using univariate extreme value theory, multivariate extreme value theory via copulas can be applied to the vector of probability density functions $\left\{f_1(\underline{X}), f_2(\underline{X}), \ldots, f_{k^*}(\underline{X})\right\}$. Therefore, this approach sheds light on how these clusters depend on each other when a novelty is detected. It might be that this approach improves the selection of the threshold by incorporating the tail-dependence between the clusters of densities.

Finally, it must be mentioned that extreme value theory is a well-studied research area where a variety of powerful improvements have been proposed. All these methods can be considered in the context of novelty detection. It is now explained how the density estimators can be adjusted for novelty detection.

7.2.2 Methods to generalise the density estimators for novelty detection

A major theme in all these methods is estimating the underlying probability density of the normal class with a Gaussian distribution or mixtures thereof. Although the class of GMMs has properties that make it an attractive choice as a density estimator, it might be too restrictive in some cases to assume the Gaussian kernel. Hence, it might prove worthwhile to consider other density estimators and, in turn, derive expressions for the distribution of the minimum probability density. Some approaches are now discussed.

Sticking with the GMM, other methods can be formulated to estimate the unknown parameters in the GMM. For example, Wang, Zhang and Lu (2005) proposed boosting a GMM. The authors proposed to expand the negative log-likelihood with a Taylor expansion at the current density estimate. Functional gradient descent is then used to update the model by strictly increasing the log-likelihood until a stopping criterion is met. This research demonstrates the possibilities to use other ensemble learning algorithms to approximate the probability density with a GMM. Alternative ensembles include bagging, stacking or regression forests. The contribution of these new estimators is the ability to train Gaussian mixture models in high dimensions with little tuning required. Furthermore, alternative estimation methods such as neural networks could be considered.

Research can also be conducted to consider other density estimators than GMMs. For example, robust mixture models can be examined. A good example of this class of distributions is the student-t distribution. Mixtures of student-t distributions can be considered as density estimators. Consequently, the same results can be derived to perform novelty detection for this class of mixture models. Another class of robust density estimators is the class of quantile mixtures. Sillitto (1969) proposed to estimate a quantile function (inverse of the probability distribution function) with polynomials. This idea was later extended by Karvanen (2006) when he formulated quantile mixture models. Although this research focused on univariate data, it is possible to extend it to the multivariate case. Karvanen (2006) proposed, among other mixtures, the Normal-polynomial quantile mixture model. This mixture model is the weighted sum of a quantile function of the Gaussian distribution and a specified number of power functions. He also showed that efficient and simple parameter estimators for this model can be derived with linear moments. This will lead to a class of density estimators for which the limiting distribution of the minimum probability density can be derived.

Another potential area for future research is the use of non-parametric or semi-parametric density estimators. It will prove very useful if the need of the parametric density estimator can be bypassed. If this can be achieved, the class of extreme value-based novelty detection algorithms can be generalised to datasets where a (parametric) density estimate is difficult to specify. One approach is to use conventional non-parametric density estimators. Hence, these density estimators are used to derive the density estimates and, consequently, a limiting distribution for the minimum of the probability density of the normal class is derived. A breakthrough would be the derivation of an extreme value-based novelty detection model without the need to derive a density estimate. This can be achieved by using distance-based or kernel-based approaches. Distance-based approaches could be used to quantify the distance, and hence the similarity, of a new observation to the normal class. In turn, extreme value theory can be used to derive probabilities that an observation is novel based on these distances. Another open question is whether a kernel function can be formulated to model the similarity between an observation and the rest of the sample. Similarly, such a kernel could be used together with extreme value theory to formulate a novelty detection algorithm. One such method (which does not use extreme value theory) is the support vector domain description (SVDD) mentioned in Chapter 2. This algorithm considers the hypersphere with minimum radius surrounding the normal class. A similar method might incorporate extreme value theory based on the maximum radii from a kernelised hypersphere.

7.3 CONCLUSION

This dissertation explored extreme value-based novelty detection. Although research into this field is new, some interesting and powerful methods have been proposed. It is believed that this project motivated the need to incorporate literature from the computer science and engineering fields into traditional statistical research. Furthermore, it has been shown that extreme value theory has a new application domain other than modelling probabilities of rare events as in traditional applications such as hydrology, financial risk analysis and insurance.

It was shown that with extreme value theory and GMMs, novelty detection can be performed in multidimensional and multimodal cases. Future research into these models would lead to very strong models to detect novelties and would, in effect, improve the world in which we live.

REFERENCE LIST

Agarwal, D. 2007. Detecting Anomalies in Cross-classified Streams: A Bayesian Approach. *Knowledge and Information Systems,* 11(1), 29-44.

Angiully, F. & Pizzuti, C. 2002. Fast Outlier Detection in High Dimensional Spaces. *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*. London: Springer-Verlag,15-26.

Barnett, V. & Lewis, T. 1994. *Outliers in Statistical Data*. Third edition. New York: John Wiley & Sons.

Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. 2004. *Statistics of Extremes: Theory and Applications.* Chichester: John Wiley & Sons.

Bingham, N.H., Goldie, C.M. & Teugels, J.L. 1987. Regular variation. *Encyclopedia of Mathematics and its Applications.* Cambridge: Cambridge University Press.

Bishop, C.M. 1995. *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press.

Boriah, S., Chandola, V. & Kumar, V. 2008. Similarity Measures for Categorical Data: A Comparative Evaluation. *Proceedings of the 8th SIAM International Conference on Data Mining*, 243-254.

Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly Detection: A Survey. *ACM Computing Surveys,* 41(3), 1-58.

Chen, Z. & Liu, B. 2016. Lifelong Machine Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning,* 33. San Rafael, CA: Morgan & Claypool Publishers.

Clifton, D. 2009. *Novelty Detection with Extreme Value Theory in Jet Engine Vibration Data*. DPhil. Oxford: University of Oxford.

Clifton, D., Bannister, P. & Tarassenko, L. 2006. Learning Shape for Jet Engine Novelty Detection. *Advances in Neural Networks,* ISNN, 3973, 828-835.

Clifton, D.A., Clifton, L., Hugueny, S. & Tarassenko, L. 2014. Extending the Generalised Pareto Distribution for Novelty Detection in High-Dimensional Spaces. *Journal of Signal Processing Systems*, 74, 323-339.

Clifton, D.A., Hugueny, S. & Tarassenko, L. 2011. Novelty Detection with Multivariate Extreme Value Statistics. *Journal of Signal Processing Systems,* 65, 371-389.

Coles, S. 2000. *An Introduction to Statistical Modelling of Extreme Values.* Springer Series in Statistics. London: Springer-Verlag.

De Haan, L. & Ferreira, A. 2006. *Extreme Value Theory: An Introduction.* Springer Series in Operations Research and Financial Engineering. New York: Springer-Verlag.

De Haan, L. 1970. *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*. Amsterdam: Mathematical Centre.

Doerksen, H. 2012. *Banknote Authentication Data Set*. UCI Machine Learning Repository. [Online] Available: https://archive.ics.uci.edu/ml/datasets/banknote+authentication. Accessed: 7 February 2017.

Duda, R.O., Hart, P.E. & Stork, D.G. 2001. *Pattern Classification*. New York: Wiley.

Fawcett, T. & Provost F. 1999. Activity Monitoring: Noticing Interesting Changes in Behaviour. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM Press, 53-62.

Fisher, R.A. & Tippett, L.H.C. 1928. On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24(2), 180-190.

Forrest, A., Perelson, A., Allen, L. & Cherukuri, R. 1994. Self-Nonself Discrimination in a Computer. *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy,* IEEE, 202-212.

Furon, T. & Jégou, H. 2013. *Using Extreme Value Theory for Image Detection*. Technical Report RR-8244, INRIA.

Goix, N., Sabourin, A. & Clémençon, S. 2016. Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. [Online] Available: https://hal.archives-ouvertes.fr/hal-01295301/document. Accessed: 8 February 2017.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C. & Wallis, J.R. 1979. Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressible in Inverse Form. *Water Resources Research*, 15, 1049-1054.

Grubbs, F.E. 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1), 1-21.

Hann, A. 2008. *Multi-parameter Monitoring for Early Warning of Patient Deterioration.* DPhil. Oxford: University of Oxford.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Second edition. New York: Springer.

Hautamäki, V., Kärkkäinen, I. & Fränti, P. 2004. Outlier Detection Using K-nearest Neighbour Graph. *Proceedings of the 17th International Conference on Pattern Recognition, ICPR, IEEE*, 3, 430-433.

He, H. & Garcia, E.A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* 21(9), 1263-1284.

Hoffmann, H. 2007. Kernel PCA for Novelty Detection. *Pattern Recognition*, 40(3), 863-874.

Hosking, J.R.M. 1985. Algorithm AS 215: Maximum-Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 34(3), 301-310.

Hosking, J.R.M., Wallis, J.R. & Wood, E.F. 1985. Estimation of the Generalized Extreme Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, 27(3), 251-261.

Huang, T., Peng, H. & Zhang, K. 2013. Model Selection for Gaussian Mixture Models. [Online] Available: https://arxiv.org/abs/1301.3558. Accessed: 3 December 2016.

Hugueny, S. 2013. *Novelty Detection with Extreme Value Theory in Vital-Sign Monitoring.* DPhil. Oxford: University of Oxford.

Hyndman, R. 1996. Computing and Graphing High Density Regions. *The American Statistician*, 50(2), 120-126.

Karvanen, J. 2006. Estimation of quantile mixtures via L-moments and trimmed L-moments. *Computational Statistics & Data Analysis,* 51(2), 947-959.

Lee, H.J. & Cho, S. 2006. The Novelty Detection Approach for Different Degrees of Class Imbalance. *Neural Information Processing. Lecture Notes in Computer Science,* 4233, 21-30.

Markou, M. & Singh, S. 2003. Novelty Detection: A Review – Part 2: Neural Network Based Approaches. *Signal Processing,* 83(12), 2499-2521.

Moya, M., Koch, M. & Hostetler, L. 1993. One-class classifier networks for target recognition applications. *Proceedings of the World Congress on Neural Networks. International Neural Network Society*, 797-801.

Pimentel, M.A.F., Clifton, D.A., Clifton, L. & Tarassenko, L. 2014. A Review of Novelty Detection. *Signal Processing*, 99, 215-249.

Roberts, S.J. 1999. Novelty detection using extreme value statistics. *IEE Proceedings on Vision, Image and Signal Processing*, 146(3), 124-129.

Schölkopf, B. & Smola, A.J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimisation and Beyond.* Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press.

Schölkopf, B., Williamson, R., Platt, J.C., Shawe-Taylor, J. & Smola, A.J. 2000. Support Vector Method for Novelty Detection. *Advances in Neural Information Processing Systems,* 12(3), 582-588.

Scott, D.W. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualisation.* Wiley Series in Probability and Statistics. Second edition. New York: John Wiley & Sons.

Sillitto, G.P. 1969. Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika,* 56(3), 641-650.

Solberg, H. & Lahti, A. 2005. Detection of Outliers in Reference Distributions: Performance of Horn's Algorithm. *Clinical Chemistry,* 51(12), 2326-2332.

Tax, D. & Duin, R. 1999. Support Vector Domain Description. *Pattern Recognition Letters*, 20(11), 1191-1199.

Verbeek, J.J., Vlassis, N. & Kröse, B. 2003. Efficient Greedy Learning of Gaussian Mixture Models. *Neural Computation,* 15(2), 469-485.

Wang, F., Zhang, C. & Lu, N. 2005. Boosting GMM and Its Two Applications. In: Oza N.C., Polikar R., Kittler J. & Roli F. (eds.), *Multiple Classifier Systems. Lecture Notes in Computer Science*, 3541. Berlin, Heidelberg: Springer.