Extremal Discriminant Analysis

DISSERTATION
zur Erlangung des Grades eines Doktors
der Naturwissenschaften

vorgelegt von
M.Sc. Manjunath, B G
geb. am 24.07.1981 in Bangalore, India

eingereicht beim Fachbereich Mathematik
der Universität Siegen
Siegen 2010

Acknowledgment

This work was created in the course of my Doctoral degree in Statistics at the Department of Mathematics, University of Siegen, Siegen. At the onset, I am heartily thankful to my supervisor Prof. Dr. R.–D. Reiss, whose encouragement, supervision and support from the preliminary to the concluding level enabled me to develop this subject. Above all and the most needed, he provided me unflinching encouragement and support in various ways. The lively and productive discussions with him will always remain a very good memory.

In addition, I am very grateful to Prof. Dr. H.J. Vaman, Prof. Dr. S. Manjunath and other faculty members at the Department of Statistics, Bangalore University, Bangalore, whose support, encouragement and appreciations are memorable.

Further more, I would like to thank Prof. Dr. E. Kaufmann and other faculty members at the Department of Mathematics, University of Siegen, Siegen, for providing pleasant work group, which makes the last three years unforgettable.

I would like to express the deepest appreciation to my colleagues, Dr. Melanie Frick, Dr. Ulf Cormann and others, whose persistent help and useful discussions would help me to complete the dissertation.

Indeed, I would like to thank all our previous and present department secretaries for helping me in all non-technical issues, which made me have a pleasant stay at the department.

Finally, I would like to thank my parents, friends and also everybody who was important to the successful realization of the thesis.

Manjunath, B G, University of Siegen.

Kurzzusammenfassung

Das Hauptziel der vorliegenden Dissertation ist die Einführung von Extremwertmodellen in der Diskriminanzanalyse. Die klassische Diskriminanzanalyse konzentriert sich auf Normalverteilungs und nichtparametrische Modelle, bei denen im zweiten Fall die unbekannten Dichten durch Kerndichten ersetzt werden, die auf der Lernstichprobe basieren. Im Folgenden nimmt man an, dass es genügt die Klassifizierung auf Basis von Überschreitungen über einer Schranke vorzunehmen. Diese Überschreitungen können als Beobachtungen im bedingten Rahmen interpretiert werden. Daher ist lediglich die statistische Modellierung von abgeschnitten Verteilungen erforderlich. In diesem Zusammenhang ist eine nichtparametrische Modellierung nicht adäquat, da die Methode bezüglich der Kerndichte im Bereich der oberen Flanke nicht exakt ist. Dennoch kann eine abgeschnittene Verteilung wie die Normalverteilung verwendet werden. Es ist das primäre Ziel, abgeschnittene Normalverteilungen durch geeignete verallgemeinerte Pareto-Verteilungen zu ersetzen und Eigenschaften und die Beziehung der Diskriminanzfunktionen in beiden Modellen zu untersuchen. Anders als beim klassischen Vorgehen in der Diskriminanzanalyse wird auch die Konvergenz der klassischen Diskriminanzfunktionen untersucht.

Abstract

The main goal of this dissertation is to introduce an extreme value model to discriminant analysis. A classical discriminant analysis focuses on Gaussian and nonparametric models where in the second case, the unknown densities are replaced by kernel densities based on the training sample. In the present text we assume that it suffices to base the classification on exceedances above higher thresholds, which can be interpreted as observations in a conditional framework. Therefore, the statistical modeling of truncated distributions is merely required. In this context, a nonparametric modeling is not adequate because the kernel method is inaccurate in the upper tail region. Yet one may deal with truncated parametric distributions like the Gaussian ones. The primary aim is to replace truncated Gaussian distributions by appropriate generalized Pareto distributions and to explore properties and the relationship of discriminant functions in both models. Different to the classical work on discriminant analysis, we are also interested in the convergence of the classical discriminant function.

Contents

List of Figures

List of special symbols

Ω	population of objects
K	number of classes
d	dimension of a vector
$\mathbf{x}$	a realization of random variable
$\mathbb{R}^d$	d–dimensional real space
$p(k)$	prior probability of k^{th} class
$C(j\|i)$	cost of classifying an object i^{th} to j^{th} class
Σ	$d \times d$ covariance matrix
$\boldsymbol{\mu}$	$d \times 1$ location parameter
$\boldsymbol{\sigma}$	$d \times 1$ scale parameter
$\varphi_{\mu,\Sigma}(.)$	Gaussian density with location and scale parameter
$\Phi_{\mu,\Sigma}(.)$	Gaussian distribution function with location and scale parameter
$\rho_{1,2}$	correlation coefficient between X_1 and X_2
$g(.)$	density generator
$\mathbb{R}$	set of real numbers
x_{disc}	discriminant point

Chapter 0

Introduction

The specification of extreme value models and exploring properties of their applications is an active research area. The theory was pioneered by Leonard Tippett with the help of R. A. Fisher [12]. The field of extreme value theory has been a very large in expanse in recent years. This is due to the fact that it is of relevance in many practical problems such as in analysis of sea levels, rainfall, financial and insurance data, risk analysis, etc. For the naturally broad variety of literature concerning extreme value theory we refer to Reiss and Thomas [32] and Beirlant et al. [5] for general introductions. The probabilistic background is discussed in detail, in Falk et al. [9], Resnick [33], Galambos [13] and Reiss [31].

The goal of the present manuscript is to introduce multivariate extreme value models to discriminant analysis. As we know, the theory of classical discriminant analysis merely focuses on Gaussian model in such case we have an explicit expression to classify. In the second place, the nonparametric decision rule is such that the unknown densities are replaced by kernel densities based on the training sample. We assume that it suffices to base the classification on exceedances above higher thresholds. In such cases a nonparametric modeling is not adequate due to the small number of exceedances and also the kernel density is inaccurate in the upper tail region. Nevertheless, one can use truncated parametric distributions like the Gaussian ones. Our primary aim is to replace truncated distributions by appropriate generalized Pareto distributions and to explore properties and the relationship of discriminant functions in both models. Subsequently, we also unfold the problem to elliptical family by using the results in Harshorva [14],[16]. In relevance we are also interested in the convergence of discriminant procedure. We remark that discriminant analysis within univariate extreme value models was inves-

tigated by Abdalla [1] and Nguimbi [29].

It is well known that the asymptotic distribution of exceedances over high thresholds is that of a generalized Pareto (GP) random vector if and only if the corresponding maxima are asymptotically distributed according to an extreme value distribution (EVD). In the univariate case, where GPDs have turned out to be crucial models for the peaks-over-threshold (POT), which is presented in Section 1.3 of Reiss et al. [9], page 21. In the framework of the construction of multivariate GPDs there are different approaches by different authors and all these definitions are closely related to each other. For broad variety of literature on GPD construction we refer to Kaufmann and Reiss [21], Tajvidi [35], Reiss et al. [9], Beirlant et al. [5] and Rootzén and Tajvidi [34]. Multivariate GPDs in the framework of extreme value theory are still under scrutiny. So, due to the limits in definition, in the present manuscript we confine to the procedure given by Tajvidi [35] and Rootzén and Tajvidi [34].

Though the content of the present manuscript is small but it is comprehensive. The mentioned problems are just a small sample of those which have to be worked on for a better understanding of extremal discriminant procedure. In a nutshell, the investigation of the discriminant analysis in the framework of extreme value models, i.e., extremal discriminant analysis, is still in the beginning and will surely be an active research area in the coming years.

In Chapter 1 we begin with some basic definitions and construction of classical discriminant rule. We will also discuss the truncated Gaussian model and present the pertaining discriminant function.

Chapter 2 will be dedicated to the derivation of multivariate Hüsler–Reiss GP density from the Hüsler–Reiss EVD. Subsequently, simulation algorithm and parameter estimation of the Hüsler–Reiss GP density have been presented. Finally, we deduce the discriminant function for the multivariate Hüsler–Reiss GP density and explore some properties of it.

Having established the densities and discriminant functions in previous Chapter 2 for the rectangularly truncated Gaussian density and the Hüsler–Reiss GP density. In Chapter 3 we present the convergence theorem that relates both models to each other.

In Chapter 4 we will generalize our main motivation to the family of elliptical distributions. Consequently, we construct rectangularly truncated elliptical density and present the convergence theorem that relates to the

Hüsler–Reiss GP density.

An additional to the present research, author also affiliated in an algorithm evolution and implementation for ***tmvtnorm***: Truncated Multivariate Normal and Student-t Distribution, `R` on–line package, cf. Wilhelm and Manjunath [42], [41]. Some of the contributed algorithms are presented in Appendix B.

Chapter 1

Classical discriminant analysis

1.0 Introduction

This chapter begins with some basic definitions and construction of discriminant procedure. The definition used here is originated from Falk et al. [10] and Lachenbruch [25]. In the present text we confine ourselves to the case of two classes; the modifications required for dealing with more than two classes are straightforward.

Section 1.3 and 1.4 will begin with deducing discriminant function for Gaussian and truncated Gaussian distribution, which are of benchmark to our current text. An analogous results for discriminant functions when the covariance matrices are not identical, is also been presented. An addition, in Section 1.4 we are introducing discriminant function for the elliptically truncated Gaussian model, see Tallis [36]. At the end of each section simulated results are presented.

1.1 Theory of discriminant analysis

The discriminant procedure, which was originally introduced by Fisher [11] has suggested using a linear combination of the observations, and the coefficients are chosen in such way that the ratio of the difference of the means of the linear combination in the two population to its variance is maximized. As a contradictory Fisher's convention doesn't depend on the population density. Later, in the framework of the construction of discriminant rule there are different approaches by different authors. Some of them are noted here: Welch [40] suggested the idea of minimizing the total probability of misclassification; Von Mises [28] suggested minimiz-

ing the maximum probability of misclassification; Anderson [3] proposed minimizing the total cost of misclassification.

A contemporary procedure in discriminant analysis which is based on Bayes rule is, classify an observation to the group with the largest posterior probability. This is equivalent to the rule that minimizes the total probability of misclassification.

In the following section we will construct Bayes discriminant procedure and throughout our manuscript we adapt the same structure to construct classification rule.

1.2 Discriminant analysis

The basic idea of discriminant analysis is to classify an object of unknown origin to one of several given classes based on the measurement vector (also called discriminator) within a d–dimensional space. The available data sets to do this are samples of objects of which both their class memberships and their measurements are known.

Now, we will formulate mathematical definition of discriminant analysis, a population Ω of objects which is divided into $K \geq 2$ disjoint subsets $\Omega_1, ..., \Omega_K$ (called classes). Let $\omega \in \Omega = \bigcup_{k=1}^{K} \Omega_k$, be an object, whose actual class is unknown. We know that the object ω carries d observable characteristics, which is a d-dimensional measurement vector $\mathbf{x} = \mathbf{x}(\omega) \in S \subset \mathbb{R}^d$, a function from $\boldsymbol{X} : \Omega \to S$ (also called as discriminator). The set S is determined by all possible measurement vectors and the number K is known.

For the classification of an object we divide the sample space S into K disjoint and nonempty subsets $G_1, ..., G_K$ and estimate the class index of ω by the decision rule, ω is classified to $\Omega_k \Leftrightarrow \mathbf{x} \in G_k$. The subsets will be chosen in some optimal way through probabilistic reasoning. The true class index $k = k(\omega)$ is interpreted to be realizations of random variables $\kappa : \Omega \to \{1, ..., K\}$, where $\kappa(\omega) := k$ for $\omega \in \Omega_k$, $k = 1, ..., K$. The random variables $\boldsymbol{X}$ and κ will generally be dependent variables, as the vector $\boldsymbol{X}(\omega)$ contains the observable information about the class index $\kappa(\omega)$.

1.2.1 The model for discriminant analysis

The sampling of an object from the population Ω is modeled by a probability distribution P, is called prior distribution. Denote by $p(k) := P(\Omega_k)$

is prior probability of the event $\Omega_k = \{\kappa = k\}$, i.e., of the event that κ attains the value k. We assume that there is a positive probability of an object be in a sample from any of the subclass. As a class specific distribution on Ω, let $\boldsymbol{X}$ be a d-dimensional random variable with density $w(\mathbf{x})$, then the conditional distribution of $\boldsymbol{X}$, given that $\kappa = k$, is

$$P(\boldsymbol{X} \in \cdot | \kappa = k) = P\{\boldsymbol{X} \in \cdot, \kappa = k\} / p(k) \tag{1.1}$$

has a d-dimensional density $w(\mathbf{x}|k)$, $k = 1, ..., K$.

The unconditional distribution of $\boldsymbol{X}$ on $S \subset \mathbb{R}^d$ has the density

$$w(\mathbf{x}) := \sum_{k=1}^{K} p(k) w(\mathbf{x}|k), \mathbf{x} \in S,$$

and the conditional distribution of κ, given that $\boldsymbol{X} = \mathbf{x}$ is given by

$$P(\kappa \in \cdot | \boldsymbol{X} = \mathbf{x}).$$

1.2.2 Expected loss under misclassification

Let $\{G_1, ..., G_K\}$ be an arbitrary partition of S. Consider an event $\{\boldsymbol{X} \in G_j$, $\kappa = i\}$ with $j \neq i$ which leads to an event of misclassification and the loss of this event is $C(j|i) > 0$,
where $C(\cdot|\cdot) : \{1, ..., K\} \times \{1, ..., K\} \to [0, \infty)$, with $C(k|k) = 0, k = 1, ..., K$, is a loss function.

The expected loss or Bayesian risk R is given by

$$\begin{aligned} R := R(G_1, ..., G_K) &:= E\left(C\left(\sum_{j=1}^{K} j 1_{G_j}(\mathbf{X}) | \kappa\right)\right) \\ &= \sum_{j=1}^{K} \sum_{i=1}^{K} p(i) C(j|i) \int_{G_j} w(\mathbf{x}|i) d\mathbf{x}. \end{aligned}$$

A partition $\{G_1, ..., G_K\}$ which minimizes the risk is an optimal space for classification rule.

1.2.3 Optimal partition

Let $\mathbf{x} \in S$ and $k = 1, ..., K$, define a function

$$d_k(\mathbf{x}) := \sum_{i=1}^{K} p(i) C(k|i) w(\mathbf{x}|i). \tag{1.2}$$

The risk R is minimized by the partition

$$\begin{aligned} G_1^* &= \left\{ \mathbf{z} \in S : d_1(\mathbf{z}) = \min_{1 \le k \le K} d_k(\mathbf{z}) \right\}, \\ G_j^* &= \left\{ \mathbf{z} \in S : d_j(\mathbf{z}) = \min_{1 \le k \le K} d_k(\mathbf{z}) \right\} \setminus \bigcup_{i=1}^{j-1} G_i^*, j = 2, ..., K. \end{aligned}$$

The funtions $d_k : S \to [0, \infty)$ are called discriminant functions.

Remark 1.1 *For a symmetric loss function $C(j|i) = C > 0$ for all i, j and $p(k) = 1/K$, then Bayes' rule in (1.2.3) becomes the maximum-likelihood rule, i.e., the smallest class index $\hat{k}$ which maximizes the group-specific density.*

1.2.4 Optimal partition when $K = 2$

Discriminant function for classifying an unknown observation between two classes is given by, $d_1(\mathbf{x}) = p(2)C(1|2)w(\mathbf{x}|2)$ and $d_2(\mathbf{x}) = p(1)C(2|1)w(\mathbf{x}|1)$. The optimal discriminant partition is,

$$\begin{aligned} G_1 &= \{ \mathbf{z} \in S : d_1(\mathbf{z}) \le d_2(\mathbf{z}) \} \\ G_2 &= \{ \mathbf{z} \in S : d_2(\mathbf{z}) < d_1(\mathbf{z}) \}. \end{aligned}$$

The above rule also give rise to,

$$\frac{w(\mathbf{x}|1)}{w(\mathbf{x}|2)} \ge \frac{p(2)C(1|2)}{p(1)C(2|1)}. \tag{1.3}$$

For notational convince we denote, let p_1, p_2 and c_1, c_2 be the corresponding prior probabilities and costs of misclassification to the first and the second population, respectively. The optimal discriminant decision is determined by the following rule: an observation vector $\mathbf{x}$ is classified to class 1 if the inequality

$$\frac{w(\mathbf{x}|1)}{w(\mathbf{x}|2)} \ge \frac{c_2 p_2}{c_1 p_1} \tag{1.4}$$

is fulfilled. The optimal common border or discriminant function is obtained by formulating (1.4) as an equation and solving it as a function in the discriminator $\mathbf{x}$. Throughout our manuscript we use above formulation to construct the discriminant function and also we confine ourselves to the case of two classes; the modifications required for dealing with more than two classes are straightforward.

1.3 Discriminant function for Gaussian model

Classical discriminant analysis focuses on the Gaussian model. In that case one gets an explicit representation of the discriminant function. Denote by

$$\varphi_{\mu,\Sigma}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}, \quad \mathbf{x}\in\mathbb{R}^d \tag{1.5}$$

the d-dimensional Gaussian density with location parameter vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and non-singular covariance matrix Σ. The pertaining distribution function is denoted by $\Phi_{\mu,\Sigma}(\mathbf{x})$.

The corresponding discriminant function for classifying an observation (using equation (1.4)) $\mathbf{x}$ between $\varphi_{\boldsymbol{\mu}^{(1)},\Sigma^{(1)}}(\mathbf{x}|1)$ and $\varphi_{\boldsymbol{\mu}^{(2)},\Sigma^{(2)}}(\mathbf{x}|2)$ is

$$\begin{aligned} D_Q(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}^{(1)})^T{\Sigma^{(1)}}^{-1}(\mathbf{x}-\boldsymbol{\mu}^{(1)}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}^{(2)})^T \\ &\quad {\Sigma^{(2)}}^{-1}(\mathbf{x}-\boldsymbol{\mu}^{(2)}) - \log\frac{c_2 p_2\left|\Sigma^{(1)}\right|^{1/2}}{c_1 p_1\left|\Sigma^{(2)}\right|^{1/2}}. \end{aligned}$$

Therefore, the decision rule entails that an observation vector $\mathbf{x}$ is classified to $\varphi_{\boldsymbol{\mu}^{(1)},\Sigma^{(1)}}(\mathbf{x}|1)$ if $D_Q(\mathbf{x}) \geq 0$, cf. Lachenbruch [25], page 11, or Falk et al. [10], page 231. This function is quadratic in $\mathbf{x}$. In addition, in the case of identical covariance matrices $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$ the discriminant function

$$D_L(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}^{(1)}+\boldsymbol{\mu}^{(2)})\right]^T \Sigma^{-1}(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}) - \log\frac{c_2 p_2}{c_1 p_1} \tag{1.6}$$

is linear, and the common border constitutes a hyperplane. This result can be regarded as a benchmark in discriminant analysis.

1.4 Discriminant function for truncated Gaussian density

In this section we discuss the truncated Gaussian density and present the pertaining discriminant function. We use both rectangular (see Horrace [17]) and elliptical truncations (see Tallis [36]). Truncation of distributions outside of the upper tail region is a crucial idea in extreme value theory.

1.4.1 Rectangular truncation

Now, let $\mathbf{X} = (X_1, ..., X_d)^T$ be a d-dimensional random vector from the rectangularly truncated Gaussian model. Then $\boldsymbol{X}$ has the density

$$f_{RT}(\mathbf{x}) = \begin{cases} \frac{\varphi_{\boldsymbol{\mu},\Sigma}(\mathbf{x})}{P\{\mathbf{X}>\boldsymbol{c}\}}, & \text{for } \mathbf{x} > \boldsymbol{c} \\ 0, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{c} = (c_1, ..., c_d)^T \in \mathbb{R}^d$, Σ is a non-singular covariance matrix and $\boldsymbol{\mu} \in \mathbb{R}^d$ is a location parameter.

The discriminant function for classifying an observation between two classes with the densities $f_{RT}(\mathbf{x}|1)$ and $f_{RT}(\mathbf{x}|2)$ which have different location parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ and truncation vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, respectively, can be determined by using equation (1.4). We have

$$D_{RT}(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})\right]^T \Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) - \log\frac{c_2 p_2}{c_1 p_1} + T_r(\boldsymbol{c}_1, \boldsymbol{c}_2),$$

where $T_r(\boldsymbol{c}_1, \boldsymbol{c}_2)$ is given by

$$T_r(\boldsymbol{c}_1, \boldsymbol{c}_2) = \log P\{\mathbf{X} > \boldsymbol{c}_2\} - \log P\{\mathbf{X} > \boldsymbol{c}_1\}.$$

Note that $D_{RT}(\mathbf{x})$ is linear in $\mathbf{x}$. We refer to Kocherlakota et al. [22] for discriminant analysis concerning to the truncated univariate Gaussian distributions.

Remark 1.2 *When the support of the two truncated densities are different and they are known, then the rule $D_{RT}(\mathbf{x}) \geq 0$ is admissible within the intersected support region. If an observation $\mathbf{x}$ falls outside of the intersected support region then the optimal rule is predetermined by using the probability of existence of an observation $\mathbf{x}$ between first and second density.*

Example 1.1 *We display simulated samples from the rectangularly truncated bivariate Gaussian density with truncation vector $\boldsymbol{c} = \boldsymbol{0}$, cf. Figure 1.1.*

1.4.2 Elliptical truncation

Now, let $\mathbf{X} = (X_1, ..., X_d)^T$ be a d-dimensional random vector from the elliptically truncated Gaussian model. Then $\mathbf{X}$ has the density

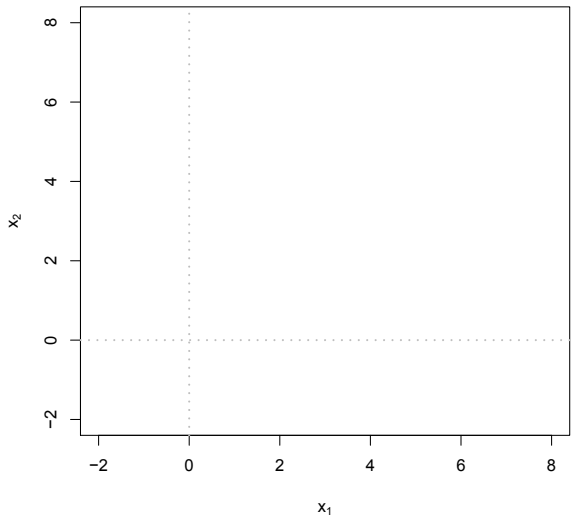


Figure 1.1: Rectangular truncation of the bivariate Gaussian density

$$f_{ET}(\mathbf{x}) = \begin{cases} \frac{\varphi_{\mu,\Sigma}(\mathbf{x})}{P\{\mathbf{X}\in E\}}, & \text{for } \mathbf{x} \in E \\ 0, & \text{otherwise}, \end{cases}$$

where $E = \{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \geq u\}$, u is non-negative real value, and Σ is a non-singular covariance matrix and $\boldsymbol{\mu} \in \mathbb{R}^d$ is a location parameter.

The discriminant function for classifying an observation between two classes which the densities $f_{ET}(\mathbf{x}|1)$ and $f_{ET}(\mathbf{x}|2)$ which have different location parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ and truncation regions E_1 and E_2, respectively, can again be determined by using equation (1.4). We have

$$D_{ET}(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})\right]^T \Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) - \log\frac{c_2 p_2}{c_1 p_1} + T_e(E_1, E_2)$$

where $T_e(E_1, E_2)$ is given by

$$T_e(E_1, E_2) = \log P\{\mathbf{X} \in E_2\} - \log P\{\mathbf{X} \in E_1\}.$$

Obviously, the two truncation borders have the same shape but differ in the shift which depends on the type of truncation.

Example 1.2 *For an illustration simulated samples from the elliptically truncated bivariate Gaussian density has been plotted with the threshold $u = 1.5$, cf. Figure 1.2.*

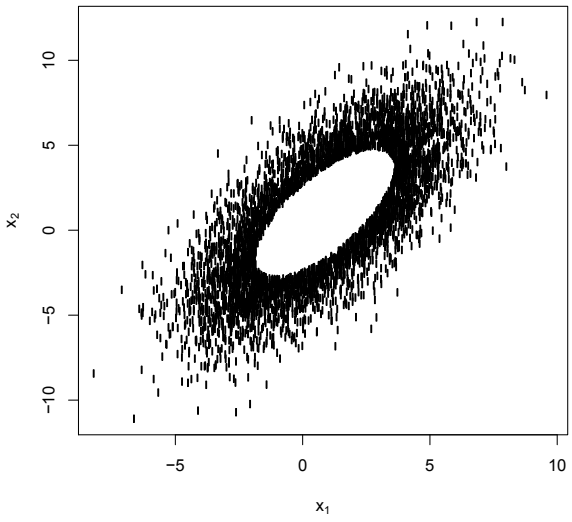


Figure 1.2: Elliptical truncation of the bivariate Gaussian density

Remark 1.3 *Note that an estimated discriminant function can be obtained by plugging in corresponding parameter estimates. In general, maximum likelihood estimator (MLE) of discriminant function is obtained by replacing all parameters with their corresponding MLE estimates.*

1.5 Non-parametric discriminant analysis

In the above sections we assumed that the group-specific distributions $w(\mathbf{x}|i)$ within the classes were known. The particular case of Gaussian and truncated Gaussian was investigated in section 1.3 and 1.4. However, if the densities $w(\mathbf{x}|i)$ are unknown then the natural idea is to replace them by kernel densities based on the training sample, see page 246 of Falk et al. [10] .

The kernel density estimator approach: let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ be a d–dimensional random vector with density $w(\mathbf{x})$. Then the kernel density approximation

of $w(\mathbf{x})$ is

$$\widehat{w}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} g\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \mathbf{x} \in \mathbb{R}^d, \tag{1.7}$$

with kernel $g : \mathbb{R}^d \to \mathbb{R}$ and bandwidth $h > 0$. With the property that

$$g(\mathbf{x}) \geq 0 \text{ and } \int g(\mathbf{x}) d\mathbf{x} = 1.$$

In addition, if the prior probabilities $p(k)$ are unknown. Then we can estimate them by the empirical frequencies, i.e., $\widehat{p}(k) = n_k/n$ of associated classes, where $n = n_1 + ... + n_K$ and n_k is sample size of k^{th} class. By substituting $w(\mathbf{x}|i)$ and $p(k)$ by $\widehat{w}(\mathbf{x}|i)$ and $\widehat{p}(k)$ in equation (1.4) we obtain non-parametric discriminant function.

In the above sections we assumed that the prior informations (or prior probabilities) and sample sizes are fixed. In general, this scenario do not leads to practical situation. Because as we know from the theory of exceedances, the number of exceedances is a Poisson distributed random variable. In that context, one might use random sample size for classification.

Chapter 2

Hüsler-Reiss GPD

2.0 Introduction

In this chapter we are primarily interested in modeling the upper tail of Gaussian density, which can be done by replacing appropriate generalized Pareto distribution. According to Theorem 1 of Hüsler and Reiss [19] the asymptotic distribution of the maxima of a triangular scheme of Gaussian random vectors is the Hüsler–Reiss extreme value distribution. Along with the approach of Rootzén and Tajvidi [34] we will deduce the pertaining GPD. The derived density will be the tail approximation of corresponding Gaussian one.

Subsequently, in the chapter we present the discriminant function and some properties of the Hüsler–Reiss GP density. Finally, we finish the chapter with some simulation results.

In the following section our results are primarily presented for the bivariate random vector and subsequently it has been extended to arbitrary dimension d.

2.1 Extreme value and generalized Pareto models

Now, let (X_1, X_2) be a bivariate random vector with associated distribution function $F_{\rho_{1,2}}$, where X_1 and X_2 are standard Gaussian random variables and $\rho_{1,2}$ is the correlation coefficient. Subsequently, let the correlation coefficient depend on the sample size n. Then, according to Theorem

1 by Hüsler and Reiss [19], the following result holds. If

$$(1-\rho_{1,2}(n))\log n \to \lambda_{1,2}^2 \in [0,\infty], \quad n \to \infty, \tag{2.1}$$

then

$$\lim_{n\to\infty} F^n_{\rho_{1,2}(n)}(b_n + x_1/b_n, b_n + x_2/b_n) = H_{\lambda_{1,2}}(x_1,x_2)$$

for every x_1, $x_2 \in \mathbb{R}$, where $b_n = n\varphi(b_n)$, φ is the standard Gaussian density, and the limiting function is given by

$$\begin{aligned} H_{\lambda_{1,2}}(x_1,x_2) = \exp\Bigg[&-\Phi\left(\lambda_{1,2} + \frac{x_1-x_2}{2\lambda_{1,2}}\right) e^{-x_2} \\ &-\Phi\left(\lambda_{1,2} + \frac{x_2-x_1}{2\lambda_{1,2}}\right) e^{-x_1}\Bigg], \end{aligned} \tag{2.2}$$

Φ being the standard Gaussian distribution function. For an explicit, approximate solution to the equation $b_n = n\varphi(b_n)$ we refer to Reiss [30], page 161. Moreover, independence and complete dependence are achieved at $\lambda_{1,2} = \infty$ and $\lambda_{1,2} = 0$, respectively, i.e.,

$$\begin{aligned} H_\infty(x_1,x_2) &= \exp(-e^{-x_1})\exp(-e^{-x_2}) \quad \text{and} \\ H_0(x_1,x_2) &= \exp\left(-e^{-\min(x_1,x_2)}\right). \end{aligned}$$

Now, following Section 3 by Hüsler and Reiss [19] and Section 12.1 by Reiss and Thomas [32], page 297, let $\mathbf{X} = (X_1,...,X_d)^T$ be a d-dimensional standard Gaussian random vector with df F_Σ, where $\Sigma = (\rho_{i,j})_{i,j\le d}$ is the correlation matrix. Apparently, by imposing a certain rate of convergence on $\rho_{i,j}(n)$, i.e., for $1 \le i,j \le d$,

$$(1-\rho_{i,j}(n))\log n \to \lambda_{i,j}^2 \in [0,\infty], \quad n \to \infty,$$

the limit of the standardized distribution function $F^n_{\Sigma(n)}$, as $n \to \infty$, is the d-dimensional Hüsler-Reiss extreme value distribution

$$H_\Lambda(\mathbf{x}) = \exp\left(-\sum_{k=1}^{d}\int_{x_k}^{\infty} \Phi_{\Sigma(k)}\left(\left(\lambda_{i,k} + \frac{x_i - z}{2\lambda_{i,k}}\right)_{i=1}^{k-1}\right) e^{-z}dz\right) \tag{2.3}$$

(in the representation given by Joe [20]) where Λ is a symmetric $d \times d$-matrix $\Lambda = (\lambda_{i,j})$ with $\lambda_{i,j} > 0$ if $i \ne j$ and $\lambda_{i,i} = 0$, and $\Phi_{\Sigma(k)}$ is a $(k-1)$–variate Gaussian distribution function (with the convention $\Phi_{\Sigma(1)} = 1$).

The mean vector of $\Phi_{\Sigma(k)}$ is zero and $\Sigma(k) = (\sigma_{i,j}(k))$ is the correlation matrix given by

$$\sigma_{i,j}(k) = \begin{cases} \frac{1}{2\lambda_{i,k}\lambda_{j,k}}\left(\lambda_{i,k}^2 + \lambda_{j,k}^2 - \lambda_{i,j}^2\right), & 1 \le i < j \le k-1, \\ 1, & i = j. \end{cases} \tag{2.4}$$

Now we discuss the construction of a GPD belonging to an EVD. The derivation of univariate GPDs, which is presented in Section 1.3 of Reiss et al. [9], page 21, has to be modified in the multivariate case. In the framework of the construction of multivariate GPDs there are different approaches by different authors. The first one can be found in the dissertation of Tajvidi [35], another one in Kaufmann and Reiss [21] and in Section 5.1 by Reiss et al. [9], and still another one in Section 8.3 by Beirlant et al. [5]. In the present work we use the definition given by Tajvidi [35], which is investigated in detail in Rootzén and Tajvidi [34].

Definition 2.1 *Let H(***x***) be a d–variate EVD. Then the corresponding GPD has the representation*

$$W(\mathbf{x}) := \begin{cases} 0, & \text{otherwise} \\ 1 - \frac{\log H(\mathbf{x})}{\log H(\mathbf{0})}, & \text{if } \mathbf{x} \ge \mathbf{0}. \end{cases} \tag{2.5}$$

The above definition has independently also been noted in Section 8.3.1 of Beirlant et al. [5], page 278. Similar definition with close to the origin on the entire negative quadrant is in Lemma 5.1.3 of Reiss et al. [9].

Remark 2.1 *Note that the Hüsler–Reiss EVD is the limiting distribution of triangular scheme of Gaussian random vectors. i.e., for each positive n we are sampling from the Gaussian distribution with increasing in parameter value (depend on n). Nevertheless, the above GPD definition is still valid for triangular scheme of random arrays because of independence at each positive n, see Hüsler et. al. [2].*

Hence, the multivariate Hüsler–Reiss GPD has the form

$$\begin{aligned} W_\Lambda(\mathbf{x}) &= 1 - (\log H_\Lambda(\mathbf{x}) / \log H_\Lambda(\mathbf{0})) \\ &= 1 - \frac{1}{C(\lambda)} \left(\sum_{k=1}^{d} \int_{x_k}^{\infty} \Phi_{\Sigma(k)} \left(\left(\lambda_{i,k} + \frac{x_i - z}{2\lambda_{i,k}} \right)_{i=1}^{k-1} \right) e^{-z} dz \right) \end{aligned} \tag{2.6}$$

where

$$C(\lambda) = \sum_{k=1}^{d} \int_{0}^{\infty} \Phi_{\Sigma(k)} \left(\left(\lambda_{i,k} - \frac{z}{2\lambda_{i,k}} \right)_{i=1}^{k-1} \right) e^{-z} dz.$$

Remark 2.2 *For $d = 2$ the constant $C(\lambda)$ reduces to*

$$\begin{aligned} C(\lambda) &= \int_0^\infty e^{-z} dz + \int_0^\infty \Phi_{\Sigma(2)} \left(\lambda_{1,2} - \frac{z}{2\lambda_{1,2}} \right) e^{-z} dz \\ &= 2\Phi_{\Sigma(2)}(\lambda_{1,2}). \end{aligned}$$

Multivariate GPDs in the framework of extreme value theory are still under scrutiny. So, due to the limits in defining a multivariate GPD we use the above definition. One shortfall of it, as discussed by Tajvidi [35], is that there is some probability mass on the each axis. i.e., the threshold line which consists of null sets with respect to the Lebesgue measure has a positive probability. This leads to one d-dimensional measure on $\mathbb{R}_+^d$ and d univariate measures on each axis. This point is also been noted in Section 2 by Michel [27].

In the following theorem we present the density of the multivariate Hüsler–Reiss GPD.

Theorem 2.1 *Let $W_\Lambda(\mathbf{x})$ be the Hüsler–Reiss GPD as defined in equation (2.6). Then for each $0 < \lambda_{i,j} < \infty$, $i < j \leq d-1$, the multivariate Hüsler-Reiss GP density is of the form*

$$w(\mathbf{x}) = \frac{e^{-x_d}}{2^{d-1} \left(\prod_{i=1}^{d-1} \lambda_{i,d} \right) C^*(\lambda)} \varphi_{\Sigma(d)} \left(\left(\lambda_{i,d} + \frac{x_i - x_d}{2\lambda_{i,d}} \right)_{i=1}^{d-1} \right) \tag{2.7}$$

for $\mathbf{x} > \mathbf{0}$, where $\varphi_{\Sigma(d)}$ is the $(d-1)$–variate Gaussian density and $C^*(\lambda) = C(\lambda)(1 - K(\lambda))$, $K(\lambda)$ denote total mass on the d axes and $C(\lambda)$ defined in (2.7) . The mean vector of $\varphi_{\Sigma(d)}$ is zero and $\Sigma(d) = (\sigma_{i,j}(d))$ is the correlation matrix in (2.4) for $k = d$.

Proof. We first prove the assertion for the bivariate case. Plugging equation (2.2) into (2.5) we obtain

$$W_{\lambda_{1,2}}(x_1, x_2) = 1 - \frac{\Phi(\lambda_{1,2} + \frac{x_1 - x_2}{2\lambda_{1,2}}) e^{-x_2} + \Phi(\lambda_{1,2} + \frac{x_2 - x_1}{2\lambda_{1,2}}) e^{-x_1}}{2\Phi(\lambda_{1,2})}.$$

If the continuous partial derivate of $W_{\lambda_{1,2}}$ exist on the open support then according to Theorem A.2.2 in Bhattacharya and Rao [7], page 264, the density is given by

$$
\begin{aligned}
w^*_{\lambda_{1,2}}(x_1, x_2) &= \frac{\partial^2 W_{\lambda_{1,2}}(x_1, x_2)}{\partial x_2 \partial x_1} \\
&= \frac{1}{2\Phi(\lambda_{1,2})} \left[\frac{e^{-x_2}}{4\lambda_{1,2}^2} \varphi' \left(\lambda_{1,2} + \frac{x_1 - x_2}{2\lambda_{1,2}} \right) \right. \\
&+ \frac{e^{-x_1}}{4\lambda_{1,2}^2} \varphi' \left(\lambda_{1,2} + \frac{x_2 - x_1}{2\lambda_{1,2}} \right) + \frac{e^{-x_2}}{2\lambda_{1,2}} \varphi \left(\lambda_{1,2} + \frac{x_1 - x_2}{2\lambda_{1,2}} \right) \\
&+ \left. \frac{e^{-x_1}}{2\lambda_{1,2}} \varphi \left(\lambda_{1,2} + \frac{x_2 - x_1}{2\lambda_{1,2}} \right) \right]
\end{aligned}
$$

where $\varphi'(a) = (-a)\varphi(a)$. Note that

$$
e^{-x_2} \varphi \left(\lambda_{1,2} + \frac{x_1 - x_2}{2\lambda_{1,2}} \right) = e^{-x_1} \varphi \left(\lambda_{1,2} + \frac{x_2 - x_1}{2\lambda_{1,2}} \right)
$$

according to Reiss and Thomas [32], page 296. With this identity the function reduces to

$$
w^*_{\lambda_{1,2}}(x_1, x_2) = \frac{e^{-x_2} \varphi(\lambda_{1,2} + \frac{x_1 - x_2}{2\lambda_{1,2}})}{4\lambda_{1,2}\Phi(\lambda_{1,2})}, \quad x_1,\, x_2 \geq 0. \tag{2.8}
$$

As discussed, if we integrate $w^*_{\lambda_{1,2}}(x_1, x_2)$ on the entire support, the total mass is less then one, namely, $(1 - \Phi(\lambda_{1,2}))/\Phi(\lambda_{1,2})$. Of course, the total mass sums up to one if we additionally consider the mass on the axes. Now, the mass on the x_2–axis is equal to $W(0, \infty) = (2\Phi(\lambda_{1,2}) - 1)/2\Phi(\lambda_{1,2})$. Since the bivariate Hüsler–Reiss distribution function is symmetric in x_1 and x_2, the same mass is obtained on the x_1–axis. One can easily see that the mass on the axes increases as $\lambda_{1,2}$ increases, i.e., the degree of independence between the two variables increases. In case of independence the entire mass lies on the axes. If $\lambda_{1,2}$ tends to zero, i.e., we move towards complete dependence, the mass on the axes converges to zero. Therefore it has been investigated that the mass on the axes is directly related to the strength of the tail dependence. The conditional bivariate density on $\mathbb{R}^2_+$ is obtained by truncating the mass on each axis. This implies that we are truncating the observations on each axis. Further, it means that we are modeling in an open rectangle of $\mathbb{R}^2_+$. So, by dividing the function in (2.8) by $(1 - \Phi(\lambda_{1,2}))/\Phi(\lambda_{1,2})$ (which is calculated within the truncated model), we obtain the bivariate Hüsler–Reiss GP conditional density.

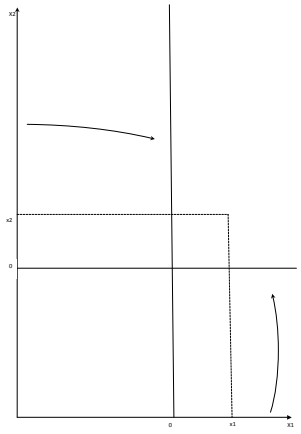


Figure 2.1: Bivariate GPD with positive support. The arrows illustrate that the mass in the each region is moved to the axes.

$$w_{\lambda_{1,2}}(x_1, x_2) = \frac{e^{-x_2}\varphi(\lambda_{1,2} + \frac{x_1 - x_2}{2\lambda_{1,2}})}{4\lambda_{1,2}(1 - \Phi(\lambda_{1,2}))}, \quad x_1,\, x_2 > 0. \tag{2.9}$$

Now we generalize our proof to arbitrary dimensions. According to Theorem A.2.2 in Bhattacharya and Rao [7], page 264, the partial derivate of W_Λ with respect to $(x_1, ..., x_d)$ exists on an open interval and the density is given by

$$w^*(\mathbf{x}) = \frac{e^{-x_d}}{2^{d-1}\left(\prod_{i=1}^{d-1}\lambda_{i,d}\right)C(\lambda)}\varphi_{\Sigma(d)}\left(\left(\lambda_{i,d} + \frac{x_i - x_d}{2\lambda_{i,d}}\right)_{i=1}^{d-1}\right). \tag{2.10}$$

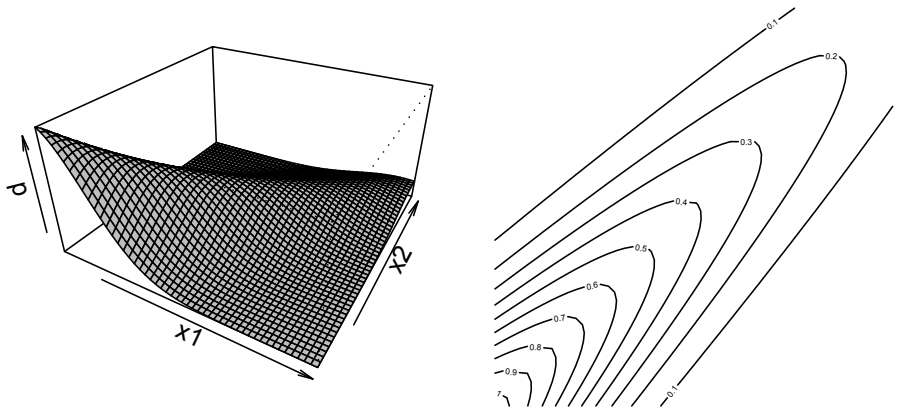
Similarly as in the bivariate case the above function leads to positive mass on each axis. The mass on the i^{th} axis can easily be determined by calculating $W_\Lambda(0, ..., 0, \infty, 0, ..., 0)$. The total mass on the d axes is denote by $K(\lambda)$. We know that the sum of the mass on the axes and the mass on $\mathbb{R}_+^d$ will add up to one. Now, similar to the bivariate case we are interested in the density upon the open rectangle of $\mathbb{R}_+^d$. Therefore, the new scaling factor is given by

$$C^*(\lambda) = C(\lambda)(1 - K(\lambda)).$$

Replacing $C(\lambda)$ by $C^*(\lambda)$ in (2.10) completes the proof. □

Remark 2.3 *For $d = 2$ the constant $C^*(\lambda)$ is given by $C^*(\lambda) = 1 - \Phi(\lambda)$. Generally, if all $\lambda_{i,j}$ are close to 0, $C^*(\lambda)$ is approximately equal to $C(\lambda)$.*

Example 2.1 *For an illustration of the bivariate Hüsler–Reiss GP density. The contour and density plots are demonstrated for $\lambda_{1,2} = 0.2$ (close to complete dependence) and $\lambda_{1,2} = 2.0$ (close to complete independence), see Figure 2.2 and 2.3,respectively.*

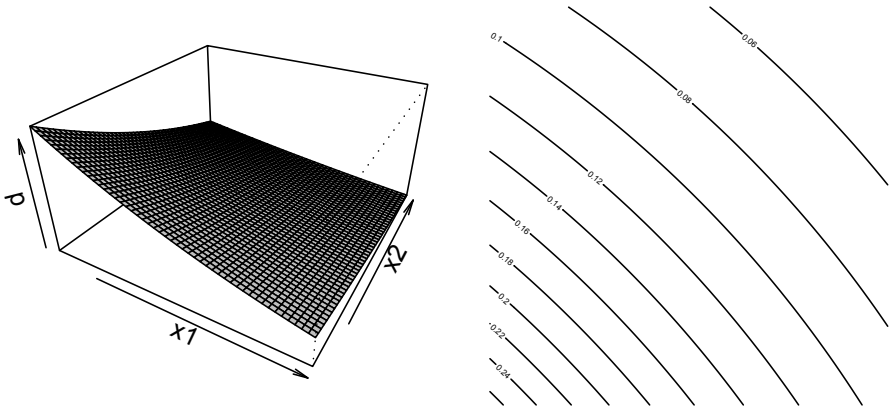


Density plot contour plot

Figure 2.2: Density and contour plot of the bivariate Hüsler–Reiss GP density for $\lambda_{1,2} = 0.2$, i.e., close to complete dependence.

Example 2.2 *As we noted from the above section that there exist a positive mass on the threshold line. In the following we will illustrate another example of the similar type.*

Let (X_1, X_2) be a bivariate random vector with the following distribution func-



Density plot contour plot

Figure 2.3: Density and contour plot of the bivariate Hüsler–Reiss GP density for $\lambda_{1,2} = 2.0$, i.e., close to complete independence

tion

$$W(x_1, x_2) = 1 - \left((-x_1)^\lambda + (-x_2)^\lambda\right)^{\frac{1}{\lambda}}, \ \left((-x_1)^\lambda + (-x_2)^\lambda\right)^{\frac{1}{\lambda}} \leq 1, \tag{2.11}$$

where $\lambda \geq 1$. Which is the derived GPD from the Gumbel EVD with negative uniform margins. And, a GPD definition follows from the Lemma given in Section 5.1.3 of Reiss et al. [9]. For the above model complete independence and dependence are achieved at $\lambda = 1$ and $\lambda = \infty$, respectively, see Figure 2.4.

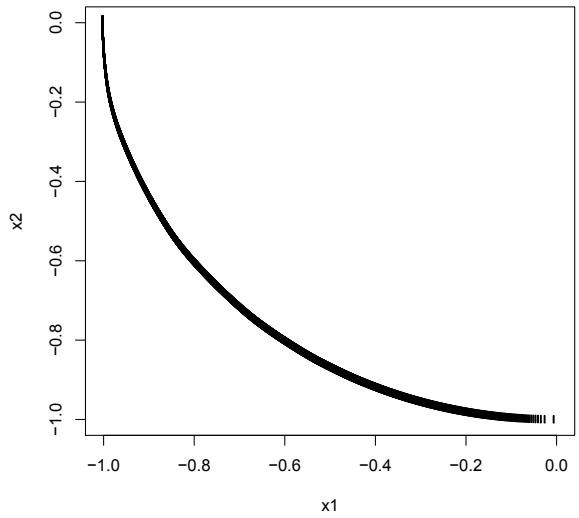


Figure 2.4: Threshold line for the Gumbel GPD by using the definition in Section 5.1.3 of Reiss et al. [9].

Now, conditional distribution function of X_1 given $X_2 = x_2$ is given by

$$\begin{aligned} P(X_1 \leq x_1 | X_2 = x_2) &= \lim_{\epsilon \downarrow 0} \frac{P\{X_1 \leq x_1, X_2 \in [x_2, x_2 + \epsilon]\}}{P\{X_2 \in [x_2, x_2 + \epsilon]\}} \\ &= \lim_{\epsilon \downarrow 0} \frac{W(x_1, x_2 + \epsilon) - W(x_1, x_2)}{x_2 + \epsilon - x_2} \\ &= \frac{d}{dx_2} W(x_1, x_2). \end{aligned} \tag{2.12}$$

So, the conditional distribution function of (2.11) reduces to,

$$P(X_1 \leq x_1 | X_2 = x_2) = \frac{-(x_2)^{\lambda - 1}}{\left((-x_1)^{\lambda} + (-x_2)^{\lambda}\right)^{1 - \frac{1}{\lambda}}}.$$

On similar line we can show that,

$$P(X_1 \leq -(1 - (-x_2)^{\lambda})^{\frac{1}{\lambda}} | X_2 = x_2) = (-x_2)^{\lambda - 1}.$$

The above probability can be rewritten as,

$$P(X_1 \leq -(1 - (-x_2)^{\lambda})^{\frac{1}{\lambda}} | X_2 = x_2) = P((-X_1)^{\lambda} + (-X_2)^{\lambda} = 1 | X_2 = x_2).$$

The total probability on the threshold line can be calculated from,

$$P((-X_1)^\lambda + (-X_2)^\lambda = 1) = \int_{-1}^{0} P((-X_1)^\lambda + (-X_2)^\lambda = 1 \mid X_2 = x_2) dx_2 = \frac{1}{\lambda}. \quad (2.13)$$

It as been shown that the total mass on the threshold line of the distribution (2.11) is $1/\lambda$*. This mass converges to zero if we move towards complete dependence, i.e, when* $\lambda \to \infty$*. Consequently, at* $\lambda = 1$ *the complete probability mass lies on the threshold line, so at* $\lambda = 1$ *we do not have density in an open space.*

2.2 Properties of the Hüsler–Reiss GPD

In the following section we will establish some properties of multivariate Hüsler–Reiss GP density.

2.2.1 Peaks-over-threshold

It is well known that the univariate GPDs are characterized by their peaks-over-threshold (POT) stability. By POT stability we mean that the excess distributions are invariant to the choice of threshold. In the frame work of multivariate POT stability, it is not uniquely determined as stated in Section 5.3. of Reiss et. al. [9] and Falk and Guillou [8]. In an article by Rootzén and Tajvidi [34] establishes the characterization of a GPD that is defined in terms of a POT stability.

According to Theorem 2.2 of Rootzén and Tajvidi [34] a GPD definition given in (2.1) persist POT stability. Which implies that the Hüsler-Reiss GPD induces POT stability.

2.2.2 Conditional density

In the following section primarily, we present the property to the bivariate case. However, it can be generalized to arbitrary dimension in a straightforward manner. We also note here that the above derived Hüsler-Reiss GP density has similar structure to Gauss-exponential density. Expect that in the Gauss-exponential density the variable in exponential model and the variables in Gaussian model are independent but the Hüsler-Reiss GP density has dependent structure. The application of the Gauss-exponential density to high risk data and properties of them are presented in Balkema and Embrechts [4].

From the above model in equation (2.9) the marginal density of X_2 is given by,

$$w_2(x_2) = \frac{e^{-x_2}\left(1-\Phi\left(\lambda_{1,2}-\frac{x_2}{2\lambda_{1,2}}\right)\right)}{2(1-\Phi(\lambda_{1,2}))}, \quad x_2 > 0 \tag{2.14}$$

Now, conditional density of X_1 given $X_2 = x_2$ is,

$$w_1(x_1|x_2) = \frac{\varphi\left(\lambda_{1,2}+\frac{x_1-x_2}{2\lambda_{1,2}}\right)}{2\lambda_{1,2}\left(1-\Phi\left(\lambda_{1,2}-\frac{x_2}{2\lambda_{1,2}}\right)\right)}, \quad x_1 > 0 \tag{2.15}$$

In general, the conditional distribution of $X_1, ..., X_{d-1}$ given $X_d = x_d$ is given by,

$$w_3(x_1, x_2, ..., x_{d-1}|x_d) = \frac{\varphi_{\Sigma(d)}\left(\left(\lambda_{i,d}+\frac{x_i-x_d}{2\lambda_{i,d}}\right)_{i=1}^{d-1}\right)}{\int_0^\infty ... \int_0^\infty \varphi_{\Sigma(d)}\left(\left(\lambda_{i,d}+\frac{x_i-x_d}{2\lambda_{i,d}}\right)_{i=1}^{d-1}\right) dx_1....dx_{d-1}}. \tag{2.16}$$

Note that the above conditional model is independent of normalizing constant. From the above model it implies that while conditioning on the last component of random vector we have truncated Gaussian model and truncated at **0**. So, by conditioning, the derived density has the tail property as such as to the original density.

In the following section we present an algorithm to simulate from the Hüsler–Reiss GP density.

2.3 Simulating from the Hüsler-Reiss GP density

The conditional distributions are useful for generating random variables in two steps: firstly, generate X_2 from marginal distribution function and secondly, generate X_1 under that condition $X_2 = x_2$. An algorithm for simulating from the bivariate Hüsler-Reiss GP density is given below,

1. distribution function of X_2 is given by,

$$\begin{aligned} W_2(x_2) &= \frac{1}{2(1-\Phi(\lambda_{1,2}))}\left\{1-e^{-x_2}+e^{-x_2}\Phi\left(\lambda_{1,2}-\frac{x_2}{2\lambda_{1,2}}\right)\right. \\ &\quad \left.+\Phi\left(\lambda_{1,2}+\frac{x_2}{2\lambda_{1,2}}\right)-2\Phi(\lambda_{1,2})\right\}. \end{aligned} \tag{2.17}$$

Now, find X_2 such that $W_2(x_2) = U$ is satisfied, where U is a random observation from $U(0, 1)$.

2. Draw X_1 from the truncated Gaussian with mean $\mu = x_2 - 2\lambda_{1,2}^2$ and $\sigma = 2\lambda_{1,2}$, truncated at 0.

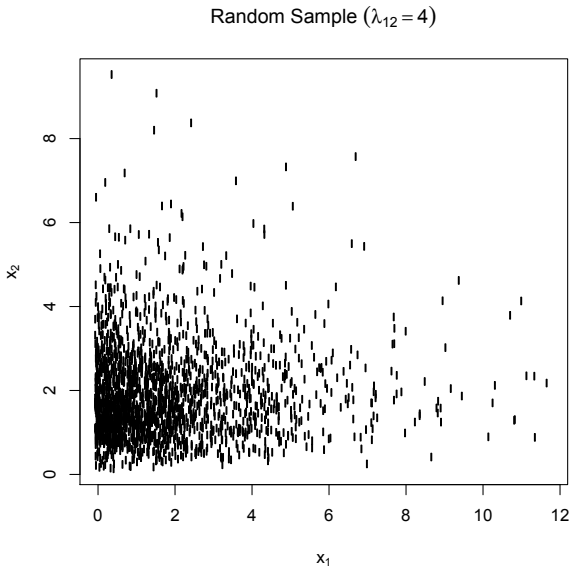


Figure 2.5: Simulation from the bivariate Hüsler–Reiss GP density for $\lambda_{1,2} = 4.0$

On the similar line, a random sample from the multivariate Hüsler–Reiss GP density can be drawn in two steps: firstly, draw X_d from the marginal distribution and secondly draw $(d-1)$–dimensional random variable from the truncated Gaussian. For an illustration 2–dimension and 3–dimension simulated observations are plotted, cf. Figure 2.5, 2.6, 2.7 and 2.8. Simulating from the multivariate truncated Gaussian using Gibbs sample approach has been presented in Appendix B.

2.4 Estimation of parameters

In the following section we present maximum-likelihood estimator for the location vector and dependence parameter of the bivariate Hüsler–Reiss GP density. We also discuss the complication in obtaining MLE estimator if we allow scale parameter in the model. The results can be generalized to arbitrary dimension in a similar way.

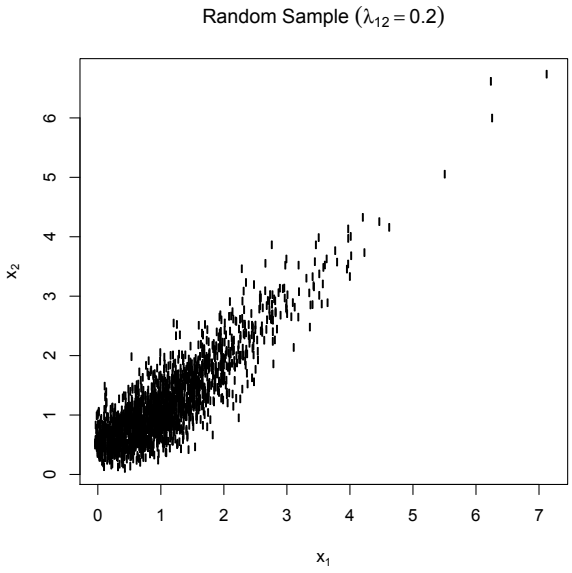


Figure 2.6: Simulation from the bivariate Hüsler-Reiss GP density for $\lambda_{1,2} = 0.2$

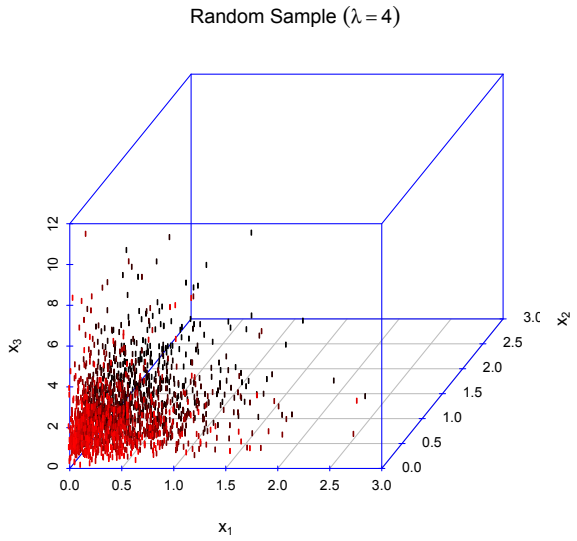


Figure 2.7: Simulation from the 3–dimensional Hüsler–Reiss GP density for $\lambda_{1,2} = \lambda_{1,3} = \lambda_{2,3} = \lambda = 4.0$

To the model in (2.9) we introduce location vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$, then the

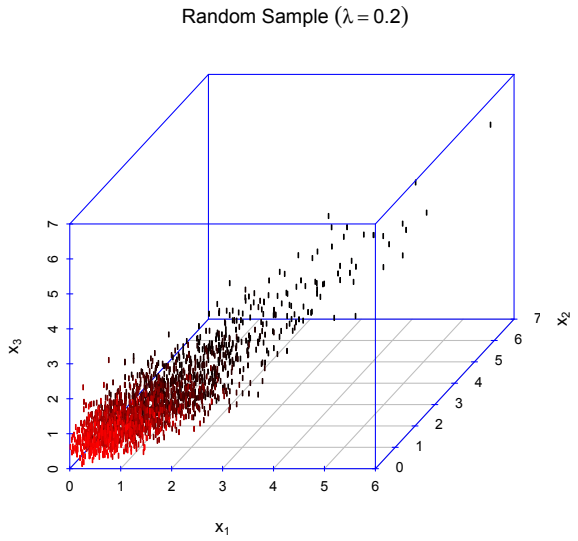


Figure 2.8: Simulation from the 3–dimensional Hüsler–Reiss GP density at $\lambda_{1,2} = \lambda_{1,3} = \lambda_{2,3} = \lambda = 0.2$

density is given by,

$$w_{\lambda_{1,2},\boldsymbol{\mu}}(x_1, x_2) = \frac{e^{-(x_2-\mu_2)}\,\varphi\left(\lambda_{1,2}+\frac{(x_1-\mu_1)-(x_2-\mu_2)}{2\lambda_{1,2}}\right)}{4\lambda_{1,2}(1-\Phi(\lambda_{1,2}))}, \quad x_1 > \mu_1,\, x_2 > \mu_2. \tag{2.18}$$

Let $(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2)$ be a n random sample vector from the above density. Then, the likelihood function is given by,

$$L(\lambda_{1,2}, \boldsymbol{\mu}; \widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2) = \frac{e^{-\sum_{i=1}^{n}(x_{2i}-\mu_2)}\prod_{i=1}^{n}\varphi\left(\lambda_{1,2}+\frac{(x_{1i}-\mu_1)-(x_{2i}-\mu_2)}{2\lambda_{1,2}}\right)}{(4\lambda_{1,2}(1-\Phi(\lambda_{1,2})))^n} \tag{2.19}$$

As we know that the support depends on the location parameter $\boldsymbol{\mu}$. So, with usual modifications in likelihood function, we define a function,

$$H(a,b) := \begin{cases} 1, & \text{if } a < b \\ 0, & \text{otherwise .} \end{cases}$$

Using the above function in (2.21) we have,

$$L(\lambda_{1,2},\boldsymbol{\mu};\widetilde{\mathbf{X}}_1,\widetilde{\mathbf{X}}_2) = \frac{e^{-\sum_{i=1}^{n}(x_{2i}-\mu_2)}\prod_{i=1}^{n}\left\{\varphi\left(\lambda_{1,2}+\frac{(x_{1i}-\mu_1)-(x_{2i}-\mu_2)}{2\lambda_{1,2}}\right)H(\mu_1,x_{1i})H(\mu_2,x_{2i})\right\}}{\left(4\lambda_{1,2}(1-\Phi(\lambda_{1,2}))\right)^n}. \tag{2.20}$$

The above likelihood function is maximized at replacing μ_1 by $min(x_{11},...,x_{1n})$ and μ_2 by $min(x_{21},...,x_{2n})$. Which implies that MLE estimator of $\boldsymbol{\mu}$ is

$$\begin{aligned}\widehat{\mu}_1 &= \min(X_{11},..,X_{1n})\\ \widehat{\mu}_2 &= \min(X_{21},..,X_{2n}).\end{aligned}$$

For estimating $\lambda_{1,2}$ we take log to (2.21),

$$\begin{aligned}\log L(\lambda_{1,2},\boldsymbol{\mu}) &= -\sum_{i=1}^{n}(x_{2i}-\mu_2)-\frac{1}{2}\sum_{i=1}^{n}\left(\lambda_{1,2}+\frac{(x_{1i}-\mu_1)-(x_{2i}-\mu_2)}{2\lambda_{1,2}}\right)^2\\ &\quad -n\log\left(4\sqrt{2\pi}\lambda_{1,2}(1-\Phi(\lambda_{1,2}))\right)\end{aligned} \tag{2.21}$$

Now, by taking partial derivative with respect to $\lambda_{1,2}$ and equating them to 0, we obtain

$$-\tfrac{n}{\lambda_{1,2}}+\tfrac{n\varphi(\lambda_{1,2})}{1-\Phi(\lambda_{1,2})}-n\lambda_{1,2}+\tfrac{1}{4\lambda_{1,2}^3}\textstyle\sum_{i=1}^{n}\left((x_{1i}-\mu_1)-(x_{2i}-\mu_2)\right)^2=0. \tag{2.22}$$

From the above equality find a value of $\lambda_{1,2}$ such that the condition is satisfied and which will be a MLE estimate of $\lambda_{1,2}$.

When $\lambda_{1,2}$ is very large then using Mills ratio,

$$(1-\Phi(\lambda_{1,2}))\approx\frac{\varphi(\lambda_{1,2})}{\lambda_{1,2}},\ \lambda_{1,2}\to\infty.$$

MLE estimator of $\lambda_{1,2}$ is given by

$$\widehat{\lambda}_{1,2}=\frac{1}{2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((x_{1i}-\mu_1)-(x_{2i}-\mu_2)\right)^2}. \tag{2.23}$$

A complete MLE estimator of (2.22) and (2.23) are obtained by plugging in MLE estimator of $\boldsymbol{\mu}$. Note that for numerical solution in (2.22) we can use (2.23) as our initial guess.

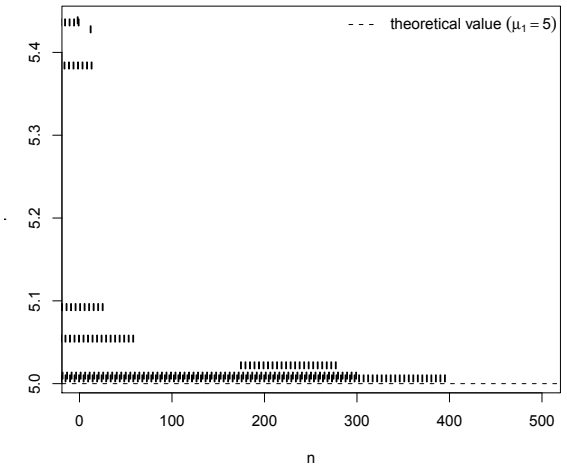


Figure 2.9: Estimate values of location parameter μ and theoretical value is set at $\mu_1 = \mu_2 = 5.0$

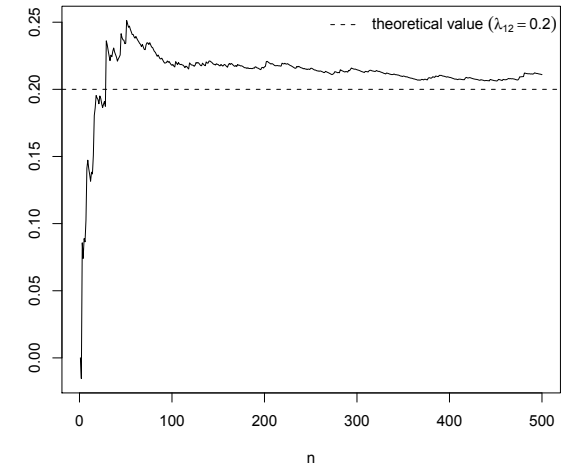


Figure 2.10: Estimate values of dependence parameter and theoretical value is set at $\lambda_{1,2} = 0.2$

Example 2.3 *For an illustration we simulated a random samples from the bivariate Hüsler-Reiss GP density with location parameter* $\mu_1 = \mu_2 = 5.0$ *and dependence parameter* $\lambda_{1,2} = 0.2$*. A convergence of MLE estimator to the true value for each increasing n up to* $n = 500$ *are presented in a graphs, see Figure 2.9 and 2.10.*

Now, introducing scale parameter $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ as well to the model (2.18) and the corresponding likelihood function is given by,

$$L(\lambda_{1,2}, \boldsymbol{\mu}, \boldsymbol{\sigma}; \widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2) = \frac{e^{-\sum_{i=1}^{n}(\frac{x_{2i}-\mu_2}{\sigma_2})} \prod_{i=1}^{n} \varphi\left(\lambda_{1,2} + \frac{\frac{(x_{1i}-\mu_1)}{\sigma_1} - \frac{(x_{2i}-\mu_2)}{\sigma_2}}{2\lambda_{1,2}}\right)}{(4\sigma_1\sigma_2\lambda_{1,2}(1-\Phi(\lambda_{1,2})))^n}. \tag{2.24}$$

From the above equation finding an exact MLE estimator of σ_1 and σ_2 is not possible. So, we can use numerical solution or other estimation technique to obtain an estimate values. One could use `nlm()` function in `R` software for numerical solution. As we note here, in general for arbitrary dimension, marginal parameters, like location and scale, can be estimated from the marginal density and the dependence parameter from the joint density.

2.5 Discriminant analysis for the Hüsler-Reiss GP model

In this section we construct the discriminant function within the multivariate Hüsler-Reiss GP model based on rectangular truncation. Therefore, we extend the multivariate Hüsler–Reiss GPD in (2.6) by a location parameter $\boldsymbol{\mu} \in \mathbb{R}^d$ and scale parameter $\boldsymbol{\sigma} > \mathbf{0}$, i.e. W_Λ becomes

$$\begin{aligned} W_{\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{x}) &= 1 - \left(\log H_\Lambda\left(\frac{\mathbf{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right) / \log H_\Lambda\left(\frac{\mathbf{0}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)\right) \\ &= 1 - \frac{1}{C(\Lambda, \boldsymbol{\mu}, \boldsymbol{\sigma})} \\ &\quad \times \left(\sum_{k=1}^{d} \int_{\left(\frac{x_k-\mu_k}{\sigma_k}\right)}^{\infty} \Phi_{\Sigma(k)}\left(\left(\lambda_{i,k} + \frac{\left(\frac{x_i-\mu_i}{\sigma_k}\right) - z}{2\lambda_{i,k}}\right)_{i=1}^{k-1}\right) e^{-z} dz\right) \end{aligned} \tag{2.25}$$

where

$$C(\Lambda, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{k=1}^{d} \int_{\left(-\frac{\mu_k}{\sigma_k}\right)}^{\infty} \Phi_{\Sigma(k)}\left(\left(\lambda_{i,k} - \frac{z}{2\lambda_{i,k}}\right)_{i=1}^{k-1}\right) e^{-z} dz.$$

Theorem 2.2 *Let $w_{\Lambda,\boldsymbol{\mu}^{(1)},\boldsymbol{\sigma}}(\mathbf{x}|1)$ and $w_{\Lambda,\boldsymbol{\mu}^{(2)},\boldsymbol{\sigma}}(\mathbf{x}|2)$ be two multivariate Hüsler-Reiss GP densities which differ in the location parameter $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$. Then,*

using equation (1.4) we obtain the optimal common border

$$\begin{aligned} D_{HR}(\boldsymbol{x}) &= (\Delta^{-1}\mathbf{x})^T\Sigma(d)^{-1}(\Delta^{-1}(\Gamma^{(2)}-\Gamma^{(1)})) \\ &\quad +\frac{1}{2}(2\boldsymbol{L}+\Delta^{-1}(\Gamma^{(2)}+\Gamma^{(1)}))^T\Sigma(d)^{-1}(\Delta^{-1}(\Gamma^{(2)}-\Gamma^{(1)}))-C, \end{aligned}$$

where $\boldsymbol{L}=(\lambda_{1,d},...,\lambda_{d-1,d})^T$, $\mathbf{x}=((\sigma_d x_1-\sigma_1 x_d),...,(\sigma_d x_{d-1}-\sigma_{d-1}x_d))^T$ *and* $\Gamma^{(i)}=((\sigma_1\mu_d^{(i)}-\sigma_d\mu_1^{(i)}),...,(\sigma_{d-1}\mu_d^{(i)}-\sigma_d\mu_{d-1}^{(i)}))^T$, $i=1,2$, *are* $(d-1)$*–dimensional vectors, and* $\Delta=diag(2\sigma_1\sigma_d\lambda_{1,d},...,2\sigma_{d-1}\sigma_d\lambda_{d-1,d})$ *is a* $(d-1)\times(d-1)$ *diagonal matrix. Obviously,* D_{HR} *is linear in* $\boldsymbol{x}$.

Proof. We rewrite the density $w_{\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{x})$ in (2.7) using d–dimensional vectors with location parameter vector, i.e.

$$w_{\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{x}) = \frac{e^{-\left(\frac{x_d-\mu_d}{\sigma_d}\right)}\exp\left(-\frac{1}{2}\boldsymbol{z}^T\Sigma(d)^{-1}\boldsymbol{z}\right)}{2^{d-1}\left(\prod_{i=1}^{d}\sigma_i\right)\left(\prod_{i=1}^{d-1}\lambda_{i,d}\right)C^*(\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma})(2\pi)^{(d-1)/2}|\Sigma(d)|^{1/2}}, \tag{2.26}$$

where $\boldsymbol{z}=\boldsymbol{L}+\Delta^{-1}(\mathbf{x}+\Gamma)$ and $\boldsymbol{L}$, Γ and Δ are defined as above.

The scaling factor $C^*(\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma})$ is now given by

$$C^*(\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma})=C(\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma})(1-K(\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma})),$$

where $K(\Lambda,\boldsymbol{\mu},\boldsymbol{\sigma})$ is the total mass on the d axes in the extended model.

Now, using (1.4), we obtain

$$-\frac{1}{2}\left(\boldsymbol{z}^{(1)}\right)^T\Sigma(d)^{-1}\boldsymbol{z}^{(1)}+\frac{1}{2}\left(\boldsymbol{z}^{(2)}\right)^T\Sigma(d)^{-1}\boldsymbol{z}^{(2)}=C, \tag{2.27}$$

where

$$\begin{aligned} C &= \log((c_2p_2)/(c_1p_1))+\frac{1}{\sigma_d}\left(\mu_d{}^{(2)}-\mu_d{}^{(1)}\right) \\ &\quad +\log\left(C^*\left(\Lambda,\boldsymbol{\mu}^{(1)},\boldsymbol{\sigma}\right)/C^*\left(\Lambda,\boldsymbol{\mu}^{(2)},\boldsymbol{\sigma}\right)\right) \end{aligned}$$

is a constant and $\boldsymbol{z}^{(1)}=\boldsymbol{L}+\Delta^{-1}\left(\mathbf{x}+\Gamma^{(1)}\right)$ and $\boldsymbol{z}^{(2)}=\boldsymbol{L}+\Delta^{-1}\left(\mathbf{x}+\Gamma^{(2)}\right)$. By substituting $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ in equation (2.27) we obtain

$$\begin{aligned} &-\frac{1}{2}\left(\boldsymbol{L}+\Delta^{-1}\left(\mathbf{x}+\Gamma^{(1)}\right)\right)^T\Sigma(d)^{-1}\left(\boldsymbol{L}+\Delta^{-1}\left(\mathbf{x}+\Gamma^{(1)}\right)\right)+ \\ &\quad\frac{1}{2}\left(\boldsymbol{L}+\Delta^{-1}\left(\mathbf{x}+\Gamma^{(2)}\right)\right)^T\Sigma(d)^{-1}\left(\boldsymbol{L}+\Delta^{-1}\left(\mathbf{x}+\Gamma^{(2)}\right)\right) = C, \end{aligned}$$

which is equivalent to

$$\left(\Delta^{-1}\mathbf{x}\right)^T \Sigma(d)^{-1}\left(\boldsymbol{C}^{(2)}-\boldsymbol{C}^{(1)}\right) + \frac{1}{2}\left(\boldsymbol{C}^{(2)}+\boldsymbol{C}^{(1)}\right)^T \Sigma(d)^{-1}\left(\boldsymbol{C}^{(2)}-\boldsymbol{C}^{(1)}\right) = C$$

where $\boldsymbol{C}^{(1)} = \boldsymbol{L} + \Delta^{-1}\Gamma^{(1)}$ and $\boldsymbol{C}^{(2)} = \boldsymbol{L} + \Delta^{-1}\Gamma^{(2)}$. Further simplification leads to the discriminant function

$$\begin{aligned} D_{HR}(\mathbf{x}) &= (\Delta^{-1}\mathbf{x})^T\Sigma(d)^{-1}(\Delta^{-1}(\Gamma^{(2)}-\Gamma^{(1)})) \\ &\quad +\frac{1}{2}(2\boldsymbol{L}+\Delta^{-1}(\Gamma^{(2)}+\Gamma^{(1)}))^T\Sigma(d)^{-1}(\Delta^{-1}(\Gamma^{(2)}-\Gamma^{(1)}))-C, \end{aligned}$$

which is a linear function in $\mathbf{x}$. Hence the proof is complete. In the following we proved that the discriminant function property is still retained in the upper tail of the Gaussian density. □

Example 2.4 *For an illustration for linear discriminant function. Discriminant function for classifying between the Hüsler-Reiss GP density with different location parameter* $\mu_1^{(1)} = 1$, $\mu_2^{(1)} = 1$ *and* $\mu_1^{(2)} = 3$, $\mu_2^{(2)} = 1$ *and common scale parameter* $\sigma_1 = 4$, $\sigma_2 = 5$ *and dependence parameter* $\lambda_{1,2} = 0.2$ *has been plotted in Figure 2.11*

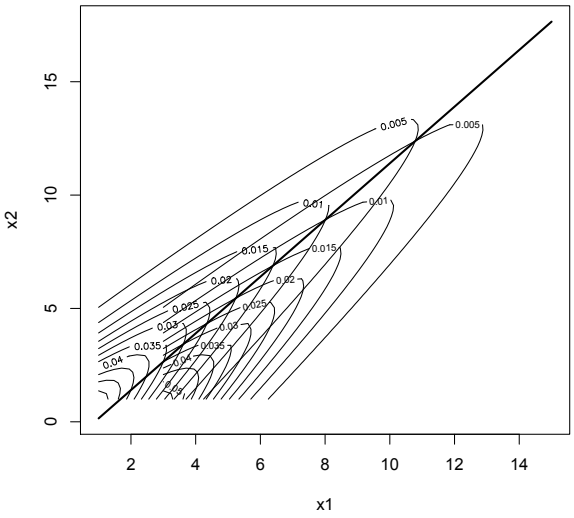


Figure 2.11: Linear discriminant function of the Hüsler-Reiss GP density.

2.6 Quadratic discriminant function

When the correlation matrices $\Sigma(d)$ and scale parameter σ are not identical, then the common border will be a quadratic function in $\mathbf{x}$ which can be solved numerically. Quadratic discriminant function is given by,

$$-\frac{1}{2}\left(\boldsymbol{L}^{(1)}+(\Delta^{-1})^{(1)}\left(\mathbf{x}^{(1)}+\Gamma^{(1)}\right)\right)^{T}\Sigma^{(1)}(d)^{-1}\left(\boldsymbol{L}^{(1)}+(\Delta^{-1})^{(1)}\left(\mathbf{x}^{(1)}+\Gamma^{(1)}\right)\right)+\frac{1}{2}\left(\boldsymbol{L}^{(2)}+(\Delta^{-1})^{(2)}\left(\mathbf{x}^{(2)}+\Gamma^{(2)}\right)\right)^{T}\Sigma^{(2)}(d)^{-1}\left(\boldsymbol{L}^{(2)}+(\Delta^{-1})^{(2)}\left(\mathbf{x}^{(2)}+\Gamma^{(2)}\right)\right)=C, \tag{2.28}$$

where

$$C=\log((c_2p_2)/(c_1p_1))+\frac{1}{\sigma_d}\left({\mu_d}^{(2)}-{\mu_d}^{(1)}\right)+\log\left(C^{*}\left(\Lambda^{(1)},\boldsymbol{\mu}^{(1)},\boldsymbol{\sigma}^{(1)}\right)/C^{*}\left(\Lambda^{(2)},\boldsymbol{\mu}^{(2)},\boldsymbol{\sigma}^{(2)}\right)\right)+\log\frac{\left(\prod_{i=1}^{d}\sigma_i^{(1)}\right)\left(\prod_{i=1}^{d-1}\lambda_{i,d}^{(1)}\right)\left|\Sigma^{(1)}(d)\right|^{1/2}}{\left(\prod_{i=1}^{d}\sigma_i^{(2)}\right)\left(\prod_{i=1}^{d-1}\lambda_{i,d}^{(2)}\right)\left|\Sigma^{(2)}(d)\right|^{1/2}}$$

is a constant.

Example 2.5 *Discriminant procedure for classifying between the Hüsler-Reiss GP density with parameter* $\mu_1^{(1)}=1$, $\mu_2^{(1)}=1$, $\lambda_{1,2}^{(1)}=0.2$ *and same density with* $\mu_1^{(2)}=3$, $\mu_2^{(2)}=1$, $\lambda_{1,2}^{(2)}=0.4$ *with common scale parameter* $\sigma_1=4$, $\sigma_2=5$. *In the Figure 2.12, the shaded region gives the optimal space for classifying an observation to* 1^{st} *density.*

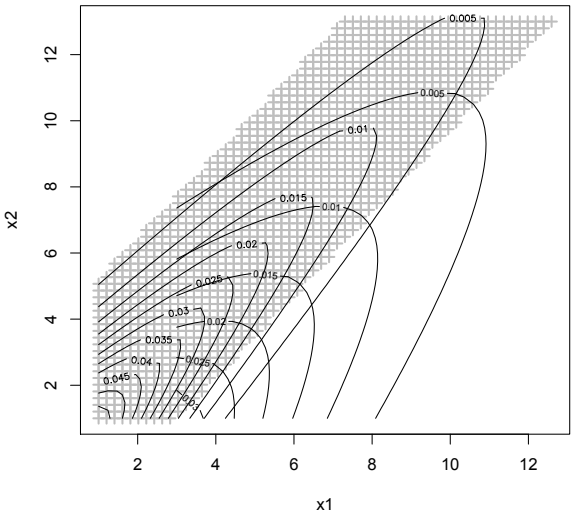


Figure 2.12: Quadratic discriminant function of the Hüsler-Reiss GP density.

Chapter 3

Convergence of discriminant procedure for the truncated Gaussian density

3.0 Introduction

Having established the densities and discriminant functions within the truncated model and the Hüsler–Reiss GP model, in the following chapter we will present a convergence theorem that relates both models to each other.

According to Theorem 1 by Hüsler and Reiss [19], we have

$$\lim_{n\to\infty} F^n_{\rho_{1,2}(n)}(b_n + x_1/b_n, b_n + x_2/b_n) = H_{\lambda_{1,2}}(x_1, x_2).$$

If

$$(1-\rho_{1,2}(n))\log n \to \lambda^2_{1,2} \in [0,\infty], \quad n \to \infty, \tag{3.1}$$

where X_1 and X_2 are standard Gaussian random variables, $\rho_{1,2}$ is the correlation coefficient and $b_n = n\varphi(b_n)$, φ is the standard Gaussian density. The limiting distribution function $H_{\lambda_{1,2}}(x_1, x_2)$ will be the Hüsler-Reiss EVD, cf. Section 2.1.

Subsequently, with suitable normalizing constants, we can show that

$$\lim_{n\to\infty} F^n_{\rho_{1,2}(n)}(d_{1,n} + x_1/a_{1,n}, d_{2,n} + x_2/a_{2,n}) = H_{\lambda_{1,2}}\left(\frac{x_1-\mu_1}{\sigma_1}, \frac{x_2-\mu_2}{\sigma_2}\right), \tag{3.2}$$

here again the condition (3.1) is satisfied, where $a_{i,n} = b_n\sigma_i$, $d_{i,n} = b_n - \mu_i/(b_n\sigma_i)$ for $i = 1, 2$; $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ are location and scale

parameters, respectively.

Now, by using the definition (1.7) for the distribution $F_{\rho_{1,2}(n)}(d_{1,n} + x_1/a_{1,n}, d_{2,n} + x_2/a_{2,n})$, we have

$$f^*_{RT,\rho_{1,2}(n)}(x_1, x_2) = \begin{cases} \frac{\varphi_{\boldsymbol{\mu}^*,\Sigma^*}(x_1,x_2)}{N_{\boldsymbol{\mu}^*,\Sigma^*}(\boldsymbol{c},\boldsymbol{\infty})}, & \text{for } x_1 > c_1,\ x_2 > c_2, \\ 0, & \text{otherwise}, \end{cases} \tag{3.3}$$

where $\boldsymbol{c} = (c_1, c_2)^T \in \mathbb{R}^2$, $\boldsymbol{\mu}^* = (-d_{1,n}a_{1,n}, -d_{2,n}a_{2,n})$ is a location parameter and non-singular covariance matrix

$$\Sigma^* = \begin{pmatrix} a^2_{1,n} & a_{1,n}a_{2,n}\rho_{1,2}(n) \\ a_{1,n}a_{2,n}\rho_{1,2}(n) & a^2_{2,n} \end{pmatrix},$$

and $N_{\boldsymbol{\mu}^*,\Sigma^*}(\boldsymbol{c}, \boldsymbol{\infty})$ is survival function of $\varphi_{\boldsymbol{\mu}^*,\Sigma^*}(x_1, x_2)$.

Note that the above density in (3.3) can also be represented as

$$f_{RT,\rho_{1,2}(n)}(x_1, x_2) = \begin{cases} \frac{\varphi_{\boldsymbol{\mu}^*,\Sigma^*}(x_1,x_2)}{N_{\boldsymbol{0},\Sigma}(\boldsymbol{c}^*,\boldsymbol{\infty})}, & \text{for } x_1 > c_1,\ x_2 > c_2, \\ 0, & \text{otherwise}, \end{cases} \tag{3.4}$$

where $\boldsymbol{c}^* = (d_{1,n} + c_1/a_{1,n}, d_{2,n} + c_2/a_{2,n})^T$ and

$$\Sigma = \begin{pmatrix} 1 & \rho_{1,2}(n) \\ \rho_{1,2}(n) & 1 \end{pmatrix}.$$

However,

$$N_{\boldsymbol{0},\Sigma}(\boldsymbol{c}^*, \boldsymbol{\infty}) = P\left\{X_1 > d_{1,n} + c_1/a_{1,n}, X_2 > d_{2,n} + c_2/a_{2,n}\right\},$$

here (X_1, X_2) be a standard bivariate Gaussian vector with correlation coefficient $\rho_{1,2}(n)$.

3.1 Convergence of truncated Gaussian to the Hüsler–Reiss GP model

In the following theorem we restrict ourselves to the bivariate case. Nevertheless the proof be generalized to arbitrary dimensions in a straightforward manner. Concerning density convergences in the univariate case we refer to Hüsler and Li [18].

Theorem 3.1 *Let $f_{RT,\rho_{1,2}(n)}$ be the density of the bivariate rectangularly truncated Gaussian model, given by*

$$f_{RT,\rho_{1,2}(n)}(x_1,x_2)=\begin{cases}\frac{\varphi_{\boldsymbol{\mu}^*,\Sigma^*}(x_1,x_2)}{N_{\mathbf{0},\rho_{1,2}(n)}(\boldsymbol{c}^*,\boldsymbol{\infty})}, & \text{for } x_1>c_1,\ x_2>c_2,\\ 0, & \text{otherwise},\end{cases} \tag{3.5}$$

where $N_{\mathbf{0},\rho_{1,2}(n)}(\boldsymbol{c}^*,\boldsymbol{\infty})$ is survival function of the standard bivariate Gaussian vector with correlation coefficient $\rho_{1,2}(n)$, $\boldsymbol{c}^*=(d_{1,n}+c_1/a_{1,n},d_{2,n}+c_2/a_{2,n})^T$, $\boldsymbol{\mu}^*=(-d_{1,n}a_{1,n},-d_{2,n}a_{2,n})$ and

$$\Sigma^*=\begin{pmatrix}a_{1,n}^2 & a_{1,n}a_{2,n}\rho_{1,2}(n)\\ a_{1,n}a_{2,n}\rho_{1,2}(n) & a_{2,n}^2\end{pmatrix}.$$

Put $a_{i,n}=b_n\sigma_i$, $d_{i,n}=b_n-\mu_i/(b_n\sigma_i)$ for $i=1,2$; with $b_n=n\varphi(b_n)$, φ is the standard Gaussian density, where the correlation coefficient $\rho_{1,2}(n)$ satisfies again

$$(1-\rho_{1,2}(n))\log n\to\lambda_{1,2}^2\in[0,\infty],\quad n\to\infty,$$

cf. (2.1). Then we have

$$\lim_{n\to\infty}f_{RT,\rho_{1,2}(n)}(x_1,x_2)=w_{\lambda_{1,2},\boldsymbol{\mu},\boldsymbol{\sigma}}(x_1,x_2),$$

for every $\mathbf{x}>\mathbf{c}$, and the limiting function is given by

$$\begin{aligned}&w_{\lambda_{1,2}}(x_1,x_2)\\ &=e^{-(x_2-\mu_2)/\sigma_2}\varphi\left(\lambda_{1,2}+\frac{(x_1-\mu_1)/\sigma_1-(x_2-\mu_2)/\sigma_2}{2\lambda_{1,2}}\right)/\\ &\quad\left\{2\lambda_{1,2}\sigma_1\sigma_2\left[\left(1-\Phi\left(\lambda_{1,2}+\frac{(c_2-\mu_2)/\sigma_2-(c_1-\mu_1)/\sigma_1}{2\lambda_{1,2}}\right)\right)e^{-(c_1-\mu_1)/\sigma_1}\right.\right.\\ &\qquad\left.\left.+\left(1-\Phi\left(\lambda_{1,2}+\frac{(c_1-\mu_1)/\sigma_1-(c_2-\mu_2)/\sigma_2}{2\lambda_{1,2}}\right)\right)e^{-(c_2-\mu_2)/\sigma_2}\right]\right\},\end{aligned}$$

where $x_1>c_1$, $x_2>c_2$.

Proof. Using definition (3.4) for $\mathbf{x}>\mathbf{c}$ we obtain

$$\begin{aligned}&f_{RT,\rho_{1,2}(n)}(x_1,x_2)\\ &=\frac{\exp\left(-\frac{1}{2}\frac{\left(\frac{x_1-\mu_1}{\sigma_1b_n}+b_n\right)^2-2\rho_{1,2}(n)\left(\frac{x_1-\mu_1}{\sigma_1b_n}+b_n\right)\left(\frac{x_2-\mu_2}{\sigma_2b_n}+b_n\right)+\left(\frac{x_2-\mu_2}{\sigma_2b_n}+b_n\right)^2}{(1-\rho_{1,2}(n)^2)}\right)}{2\pi\sigma_1\sigma_2b_n^2\sqrt{1-\rho_{1,2}(n)^2}\,P\left\{X_1>\frac{c_1-\mu_1}{\sigma_1b_n}+b_n,X_2>\frac{c_2-\mu_2}{\sigma_2b_n}+b_n\right\}}.\end{aligned} \tag{3.6}$$

Corresponding to the proof of Theorem 1 in Hüsler and Reiss [19] one gets

$$\begin{aligned} & nP\left\{X_1 > \frac{c_1-\mu_1}{\sigma_1 b_n} + b_n, X_2 > \frac{c_2-\mu_2}{\sigma_2 b_n} + b_n\right\} \\ & \to \left(1-\Phi\left(\lambda_{1,2} + \frac{(c_2-\mu_2)/\sigma_2 - (c_1-\mu_1)/\sigma_1}{2\lambda_{1,2}}\right)\right) e^{-(c_1-\mu_1)/\sigma_1} \\ & \qquad + \left(1-\Phi\left(\lambda_{1,2} + \frac{(c_1-\mu_1)/\sigma_1 - (c_2-\mu_2)/\sigma_2}{2\lambda_{1,2}}\right)\right) e^{-(c_2-\mu_2)/\sigma_2}, \end{aligned} \tag{3.7}$$

as $n \to \infty$, and

$$b_n\sqrt{1-\rho_{1,2}(n)^2} \to 2\lambda_{1,2}, \tag{3.8}$$

as $n \to \infty$, which proves the convergence of the denominator.

From the proof in Hüsler and Reiss [19] we also deduce that

$$\frac{\frac{x_1-\mu_1}{\sigma_1 b_n} + b_n - \rho_{1,2}(n)\left(\frac{x_2-\mu_2}{\sigma_2 b_n} + b_n\right)}{\sqrt{1-\rho_{1,2}(n)^2}} \to \lambda_{1,2} + \frac{(x_1-\mu_1)/\sigma_1 - (x_2-\mu_2)/\sigma_2}{2\lambda_{1,2}}, \tag{3.9}$$

as $n \to \infty$, which we use to show the convergence of the numerator. We can write

$$\begin{aligned} & \frac{n \exp\left(-\frac{1}{2}\frac{\left(\frac{x_1-\mu_1}{\sigma_1 b_n}+b_n\right)^2 - 2\rho_{1,2}(n)\left(\frac{x_1-\mu_1}{\sigma_1 b_n}+b_n\right)\left(\frac{x_2-\mu_2}{\sigma_2 b_n}+b_n\right) + \left(\frac{x_2-\mu_2}{\sigma_2 b_n}+b_n\right)^2}{(1-\rho_{1,2}(n)^2)}\right)}{2\pi b_n} \\ & = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\frac{x_1-\mu_1}{\sigma_1 b_n} + b_n - \rho_{1,2}(n)\left(\frac{x_2-\mu_2}{\sigma_2 b_n} + b_n\right)}{\sqrt{1-\rho_{1,2}(n)^2}}\right)^2\right) \\ & \quad \times \frac{n}{b_n}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2 b_n} + b_n\right)^2\right), \end{aligned} \tag{3.10}$$

where the first factor converges to

$$\varphi\left(\lambda_{1,2} + \frac{(x_1-\mu_1)/\sigma_1 - (x_2-\mu_2)/\sigma_2}{2\lambda_{1,2}}\right),$$

as $n \to \infty$, because of (3.10) and the second factor satisfies

$$\begin{aligned}
&\frac{n}{b_n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2 b_n}+b_n\right)^2\right)\\
&=\frac{n}{b_n}\varphi(b_n)\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2 b_n}\right)^2\right)e^{-(x_2-\mu_2)/\sigma_2}\\
&=\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2 b_n}\right)^2\right)e^{-(x_2-\mu_2)/\sigma_2}\\
&\to e^{-(x_2-\mu_2)/\sigma_2}, \quad n\to\infty. \qquad (3.11)
\end{aligned}$$

Combining the above convergences completes the proof. □

Because the discriminant functions are obtained by using the inequality (1.4) which contains a ratio of densities, Theorem 4.1 directly implies the convergence of the discriminant functions. More precisely, the discriminant function D_{RT} (appropriately normalized) of the rectangularly truncated Gaussian model converges to the discriminant function D_{HR} of the Hüsler-Reiss GP model.

An analogous result still holds if different covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are chosen. In this case the quadratic discriminant function of the truncated Gaussian model converges to the quadratic one in the Hüsler–Reiss GP model, cf. Section 2.6.

Chapter 4

Elliptical family

4.0 Introduction

In the following chapter we will generalize our main goal to the family of elliptical distributions. The most important subclass of elliptically symmetric (or elliptical) distributions is Gaussian one. In some recent papers by Hashorva [14], [16] shows that the multivariate Hüsler–Reiss distribution is as well the limiting distribution of multivariate maxima of elliptical triangular arrays if the random radius of the elliptical random vectors belongs to the max–domain of attraction of a Gumbel distribution. In the present text we assume that the results in Hashorva [14], [16] are suffices to extend our main motivation to elliptical family.

We refer to Koutras [24] and Wakaki [39] for discriminant analysis concerning spherical and elliptical distributions. For general discussion on extreme value theory for elliptical and spherical distributions, see Kotz and Nadarajah[23] and Hashorva [15].

4.1 Rectangularly truncated elliptical density

A d–dimensional random vector $\boldsymbol{X}$ is said to have an elliptically contoured distribution if its joint density takes the form (see Kotz and Nadarajah [23]),

$$f_{\boldsymbol{\mu},\Sigma}(\mathbf{x}) = |\Sigma|^{-\frac{1}{2}} g\left((\mathbf{x}-\boldsymbol{\mu})^T |\Sigma|^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) \tag{4.1}$$

where $g(.)$ is a scale function (also called as density generator), Σ is $d \times d$ constant matrix and $\boldsymbol{\mu}$ is a $d \times 1$ vector.

Example 4.1 *For*

$$g(x) = \frac{\exp(-x/2)}{(2\pi)^{d/2}}$$

then $f(\mathbf{x})$ *will be the d–dimensional Gaussian density.*

Now, let $\mathbf{X} = (X_1, ..., X_d)^T$ be a d-dimensional random vector from the rectangularly truncated elliptical model (4.1). Then $\boldsymbol{X}$ has the density

$$f_{ERT}(\mathbf{x}) = \begin{cases} \frac{f_{\boldsymbol{\mu},\Sigma}(\mathbf{x})}{P\{\mathbf{X}>\boldsymbol{c}\}}, & \text{for } \mathbf{x} > \boldsymbol{c} \\ 0, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{c} = (c_1, ..., c_d)^T \in \mathbb{R}^d$.

4.2 Extreme value and generalized Pareto for elliptical distributions

Let (X_1, X_2) be a triangular array of the bivariate elliptically distributed random vector. Then, (X_1, X_2) satisfy below equality in distribution

$$(X_1, X_2) =^d (X_1^* \rho_{1,2} + \sqrt{1 - \rho_{1,2}^2} X_2^*) \tag{4.2}$$

where $\rho_{1,2}$ is the correlation coefficient and X_1^*, X_2^* are spherically distributed random variables.

Now, according to Theorem 2.1 of Hashorva [14] if the random radius $R := \sqrt{X_1^{*2} + X_2^{*2}}$ has the distribution function F which is in the max-domain of attraction of the standard Gumbel distribution function with auxiliary function w, is defined as

$$w(t) = \frac{(1 + o(1))[1 - F(t)]}{\int_t^w [1 - F(s)]ds}, \; t \to w \tag{4.3}$$

and additionally, if

$$(1 - \rho_{1,2}(n))b(n)w(b(n)) \to 2\lambda_{1,2}^2 \in [0, \infty], \quad n \to \infty, \tag{4.4}$$

then

$$\lim_{n\to\infty} \sup_{(x_1,x_2)\in\mathbb{R}} \left| P\left\{ \max_{1\le i\le n} X_{1i} \le x_1 a(x) + b(n), \max_{1\le i\le n} X_{2i} \le x_2 a(x) + b(n) \right\} - H_{\lambda_{1,2}}(x_1, x_2) \right| = 0,$$

with $b(n) = G^{-1}(1 - 1/n)$, $a(n) = 1/w(b(n))$, $n \in \mathbb{N}$ and G is the distribution function of X_1^*.

An analogous result for the multivariate elliptical distribution, see Theorem 2.1 of Hashorva [16].

Remark 4.1 *For the Gaussian case we have* $b(n)w(b(n)) = b(n)/a(n) = (1 + o(1))2\log n$.

Remark 4.2 *From Berman [6], if* X_1^*, X_2^* *are identically distributed with distribution function* G *and additionally* G *is symmetric at 0, then we have below equality in distribution*

$$(X_1^*, X_2^*) =^d (R\cos\theta, R\sin\theta),$$

where θ *is uniformly distributed in* $(-\pi, \pi)$. *Consequently, if* X_1^*, X_2^* *are independent standard Gaussian random variables. Then, the distribution function of* R^2 *(Chi-Square distribution) is in the max-domain of attraction of the Gumbel distribution.*

4.3 Convergence of truncated elliptical density to the Hüsler–Reiss GP density

In the following theorem we restrict ourselves to the bivariate case. Nevertheless the proof be generalized to arbitrary dimensions in a straightforward manner.

Theorem 4.1 *Let* $f_{ERT,\rho_{1,2}}$ *be the density of the bivariate rectangularly truncated standard elliptical model, given by*

$$f_{ERT,\rho_{1,2}(n)}(x_1, x_2) = \begin{cases} \frac{f(t_n(x_1), t_n(x_2))}{(a^2(n))^{-1} P\{X_1 > t_n(x_1), X_2 > t_n(x_2)\}}, & \text{for } x_1 > c_1, x_2 > c_2 \\ 0, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{c} = (c_1, c_2)^T \in \mathbb{R}^2$ *and* $t_n(s) := a(n)s + b(n)$. *Put* $b(n) = G^{-1}(1 - 1/n)$, $a(n) = 1/w(b(n))$, $n \in \mathbb{N}$ *with* G *is the distribution function of* X_1^* *and* w *is defined as*

$$w(t) = \frac{(1 + o(1))[1 - F(t)]}{\int_t^w [1 - F(s)]ds},\ t \to w$$

with F *be the distribution function of random radius* $R := \sqrt{X_1^{*2} + X_2^{*2}}$ *where random vector* (X_1^*, X_2^*) *is spherically distributed and additionally, if*

$$(1 - \rho_{1,2}(n))b(n)w(b(n)) \to 2\lambda_{1,2}^2 \in [0, \infty], \quad n \to \infty.$$

Then we have

$$\lim_{n\to\infty} f_{ERT,\rho_{1,2}(n)}(x_1,x_2) = w_{\lambda_{1,2}}(x_1,x_2),$$

for every $\mathbf{x} > \mathbf{c}$, *and the limiting function is given by*

$$w_{\lambda_{1,2}}(x_1,x_2) = e^{-x_2}\varphi\left(\lambda_{1,2}+\frac{x_1-x_2}{2\lambda_{1,2}}\right) / \left\{2\lambda_{1,2}\left[\left(1-\Phi\left(\lambda_{1,2}+\frac{c_2-c_1}{2\lambda_{1,2}}\right)\right)e^{-c_1} + \left(1-\Phi\left(\lambda_{1,2}+\frac{c_1-c_2}{2\lambda_{1,2}}\right)\right)e^{-c_2}\right]\right\},$$

where $x_1 > c_1$, $x_2 > c_2$.

Proof. Using the definition of the truncated elliptical model for $x_1 > c_1$, $x_2 > c_2$ we have

$$f_{ERT,\rho_{1,2}(n)}(x_1,x_2) = \frac{g(t_n^2(x_1)+t_n^2(x_2)-2\rho_{1,2}(n)t_n^2(x_1)t_n^2(x_1))}{(a^2(n))^{-1}\sqrt{1-\rho_{1,2}(n)^2}\,P\{X_1 > (a(n)c_1+b(n)), X_2 > (a(n)c_2+b(n))\}}$$

where g(.) is density generator, cf. (4.1).

According to Lemma 3.3 of Hashorva [14] we have following equality in distribution,

$$(X_1,X_2) =^d (R\cos(\theta), R\cos(\theta-\psi)), \tag{4.5}$$

cf. Remark 3.2, where θ is uniformly distributed in $(-\pi,\pi)$ and $\psi(n) := \arccos(\rho_{1,2}(n)) \in [0,\pi/2]$. Now, setting $\beta(n) := -arctan((\rho_{1,2}(n)-y/x)/\sin(\psi(n)))$ and $v(n) := \sqrt{(b(n)w(b(n)))}$ then corresponding to the Proof of Lemma 3.3 of Hashorva [14] one gets

$$\begin{aligned} &nP\{X_1 > (a(n)c_1+b(n)), X_2 > (a(n)c_2+b(n))\} \\ &\to \left(1-\Phi\left(\lambda_{1,2}+\frac{c_2-c_1}{2\lambda_{1,2}}\right)\right)e^{-c_1} \\ &\quad + \left(1-\Phi\left(\lambda_{1,2}+\frac{c_1-c_2}{2\lambda_{1,2}}\right)\right)e^{-c_2}, \end{aligned} \tag{4.6}$$

as $n\to\infty$, and

$$\sqrt{2(1-\rho_{1,2}(n)^2)}v(n) \to 2\lambda_{1,2}, \tag{4.7}$$

as $n\to\infty$, which proves the convergence of the denominator.

Similarly using Lemma 3.3 of Hashorva [14] and Theorem 12.3.1 of Berman [6] with the identity according to Reiss and Thomas [32], page 296,

$$e^{-x_2}\varphi\left(\lambda_{1,2}+\frac{x_1-x_2}{2\lambda_{1,2}}\right)=e^{-x_1}\varphi\left(\lambda_{1,2}+\frac{x_2-x_1}{2\lambda_{1,2}}\right),$$

the convergence of numerator follows. □

An analogous result still holds if we introduce location and scale parameter in the above proof. As we discussed in Section 3.1 the above Theorem directly implies the convergence of the discriminant functions.

Remark 4.3 *In general when the population densities are non-normal, several difficulties arises, the main one being that the usual statistical optimality criteria do not lead to simple decision rules. As we already mentioned that the model for the upper tail of the elliptically distributed random variable can be approximated by the Hüsler – Reiss GP distribution provided that the random radius belongs to the Gumbel max-domain of attraction. Apparently, though the initial discriminant functions concerning elliptical family need not lead to an explicit expression or simple decision rule to classify but in the limiting discriminant function we have an explicit expression. Which is consider as remarkable result in limiting discriminant function.*

Appendix $\mathcal{A}$

Extremal discriminant function against to the classical function in R

In the following chapter we demonstrate simulation experiment which entitles the significance of extremal discriminant function against to the classical function in R software.

In the following section we illustrates how classification can be performed in the R software.

A.0.1 Linear discriminant analysis – R function lda

The multivariate normal based linear and quadratic classification rules are implemented in the R functions `lda` and `qda`, respectively. These functions are part of the `MASS` package, cf. Venables and Ripley [37] [38], which needs to be activated to make the functions available. This can be done by simply entering,

```
library(MASS)
```

To perform a linear discriminant analysis, call the function `lda`, for instance

```
Z = lda(X,factor,prior=c(1,1)/2)
```

The first argument is a matrix or data frame or matrix containing the explanatory variables. The second argument is the `factor` and the function is used to create a factor. Factors in R are stored as a vector of integer values with a corresponding set of character values to use when the factor

is displayed. The only required argument to `factor` is a vector of values which will be returned as a vector of factor values. Both numeric and character variables can be made into factors. The argument `prior` allows to specify the prior probabilities. If this command is omitted, the function uses the proportions of π_1 and π_2 objects in the sample. Check the content by entering result in the command line. Predictions/classifications can be made by the function `predict`, for instance

```
p = predict(Z,test)$class
```

The first argument to be supplied to the function `predict` is an `R` object resulting from an `lda` call; the second argument is a data frame that needs to be classified.

Now, consider the following simulation experiment, firstly, we will simulate $n_1 = 1000$ observations from the bivariate Hüsler-Reiss GP density with location and scale parameters $\mu_1^{(1)} = 1$, $\mu_2^{(1)} = 1$ and $\sigma_1 = 1$, $\sigma_2 = 1$ and dependence parameter $\lambda_{1,2} = 0.2$. In the second step, we will simulate $n_2 = 1000$ observations from the bivariate Hüsler-Reiss GP density with location and scale parameters $\mu_1^{(2)} = 3$, $\mu_2^{(2)} = 2$ and $\sigma_1 = 1$, $\sigma_2 = 1$ and dependence parameter $\lambda_{1,2} = 0.2$. For an algorithm to simulate from the bivariate Hüsler-Reiss GP density cf. Section 2.3.

Note that we have equality in scale and dependence parameter so obtained discriminant function will have linearity property. In the next step we use a function `lda()` in `R` software and the training data set to it will be simulated samples from the above models.

Again, simulate $n = 500$ observations from the second population and use it has test data. Now, we input this data sets to both `lda()` and the Hüsler-Reiss discriminant function to classify between two densities. In Figure A.1, first graph represent number of observations classified to density one and two by using `lda()` function. As we know our test data is simulated data set from the second population. The number of observations in first density represents the total number of misclassification. Now, according to the Hüsler-Reiss discriminant rule we have 19 misclassification. We can conclude that both the discriminant rule almost agree each other.

In the second experiment, we repeat above procedure with increasing scale parameter $\sigma_1 = 4$, $\sigma_2 = 5$. In Figure A.2 we can clearly observe that the number of misclassification is almost 200 by using `lda()` function. And, according to the Hüsler-Reiss discriminant rule we have 116

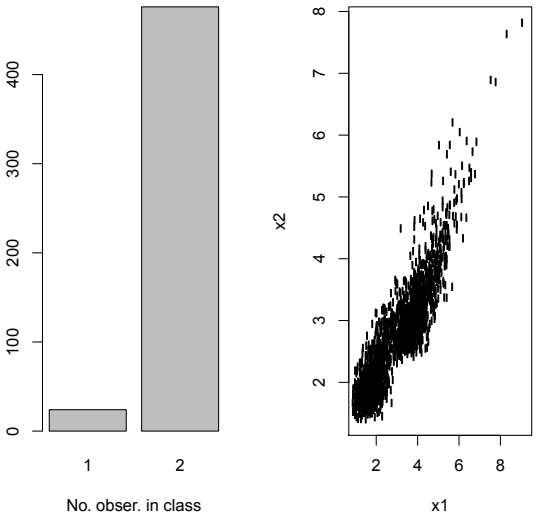


Figure A.1: Number of observations classified to class 1 & 2 using R function.

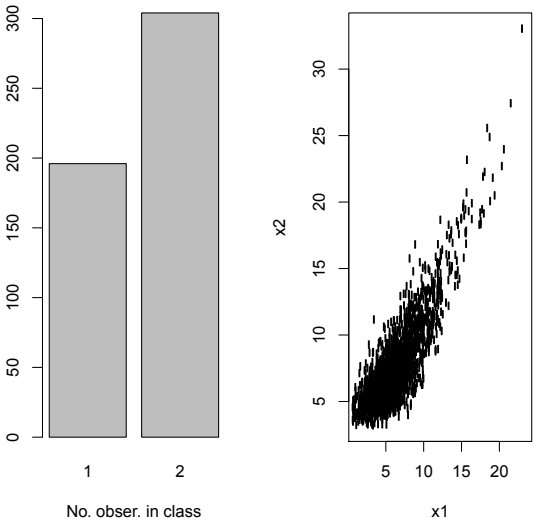


Figure A.2: Number of observations classified to class 1 & 2 using R function.

misclassification. As we know from the theory of discriminant analysis that the misclassification rate increases with increasing in variance of the population. With the following experiment we can conclude that still the

Hüsler-Reiss discriminant rule leads to a minimum number of misclassification.

In the following we can observe that even though the classical discriminant function and the Hüsler-Reiss discriminant function share similar property, i.e, linearity, but they differ in their coefficients. Like, the Hüsler-Reiss discriminant function coefficients are function of tail dependence parameter $\lambda_{1,2}$, which is a crucial parameter for heavy tail models. Here we can conclude that classical discriminant function do not leads to a minimum misclassification in all the scenarios. So, we insist extremal discriminant function when the data sets are from heavy tail models.

Appendix $\mathcal{B}$

R Contribution

B.0 Introduction

In the following we present some of the contributed algorithm to `tmvtnorm`: Truncated Multivariate Normal and Student-t Distribution, `R` on–line package.

B.1 Gibbs sampling from truncated multivariate Gaussian model

Gibbs algorithm is an approach which uses univariate conditional density for simulating from multivariate model. In the following we present an algorithm for simulating from the double truncated multivariate Gaussian model. The enhancing nature of the multivariate truncated Gaussian model is, marginal densities are not necessarily truncated Gaussian but conditional densities are truncated Gaussian model. This property simplifies our simulation algorithm.

The probability density function of the multivariate double truncated Gaussian random variable can be expressed as:

$$f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_g, \mathbf{a}, \mathbf{b}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}}{\int_{\mathbf{a}}^{\mathbf{b}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} d\mathbf{x}}$$

for $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ and 0 otherwise.

The conditional density of X_i given $\boldsymbol{X}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$ follows

$N_T(\mu_i^*, \sigma_i^*)$ (univariate truncated normal) and has the support $a_i \leq x_i \leq b_i$. The μ_i^* and σ_i^* are give by,

$$\Sigma_g = \begin{bmatrix} \sigma_{ii} & \Sigma_i \\ \Sigma_i^T & \Sigma \end{bmatrix}$$
$$\mu_i^* = \mu_{ii} + \Sigma_i^T \Sigma^{-1} (\mathbf{x}_{-i} - \boldsymbol{\mu}_{-i})$$
$$\sigma_i^* = \sigma_{ii} - \Sigma_i^T \Sigma^{-1} \Sigma_i.$$

The matrix Σ is a sub matrix of Σ_g for order $(d-1) \times (d-1)$, which is obtained by deleting i^{th} row and i^{th} column from Σ_g. And, the vector Σ_i, is $(d-1)$-dimensional vector, which is i^{th} row vector expect by deleting i^{th} element. Since we have univariate conditional density so the procedure for drawing a random observation from univariate conditional density of the above model is same as that of procedure for drawing a random observation from univariate truncated normal density.

$$x_i = \mu_i^* + \sqrt{\sigma_i^*} \Phi^{-1} \left[\Phi\left(\frac{a_i - \mu_i^*}{\sqrt{\sigma_i^*}} \right) + u \left(\Phi\left(\frac{b_i - \mu_i^*}{\sqrt{\sigma_i^*}} \right) - \Phi\left(\frac{a_i - \mu_i^*}{\sqrt{\sigma_i^*}} \right) \right) \right]$$

where u is a random observation from $U(0,1)$.

Algorithm:

1. choose an observation vector $\mathbf{x}^0$ from $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$, may be $x^0 = \mathbf{a}$ or $x^0 = \mathbf{b}$
2. generate an observation x_1 from the conditional density of X_1 given $\boldsymbol{X}_{-1} = (x_2^0, ..., x_d^0)$
3. generate an observation x_2 from the conditional density of X_2 given $\boldsymbol{X}_{-2} = (x_1, x_3^0, ..., x_d^0)$
4. x_i from the conditional density of X_i given $\boldsymbol{X}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}^0,$ $..., x_d^0)$
5. x_d from the conditional density of X_d given $\boldsymbol{X}_{-d} = (x_1, ..., x_{d-1})$.

The above procedure will give an one set of d-dimensional random sample $\mathbf{x}^{(1)}$ and for later draw use $\mathbf{x}^{(1)}$ as initial choice. Repeat the above procedure n times to have n sample observation.

Like all MCMC methods, the first iterates of the chain will not have the exact target distribution and are strongly dependent on the start value. To reduce this kind of problem, the first iterations are considered as a burn-in period which a user might want to discard.

As a major drawback, random samples produced by Gibbs sampling are not independent, but correlated. The degree of correlation depends on the covariance matrix $\mathbf{\Sigma}_g$ as well on the dimensionality. Taking only a nonsuccessive subsequence of the Markov chain, say every k^{th} sample, can greatly reduce the auto–correlation among random points and is called as "thinning". For detailed discussion, cf. Wilhelm and Manjunath [41].

Example B.1
```
mu = c(0.5, 0.5)
sigma = matrix(c(1, 0.8, 0.8, 2), 2, 2)
a = c(-1, -Inf)
b = c(0.5, 4)
X = rtmvnorm(n=10000, mean=mu, sigma=sigma, lower=a, upper=b,
 algorithm="gibbs", burn.in.samples=100,thinning = 5).
```

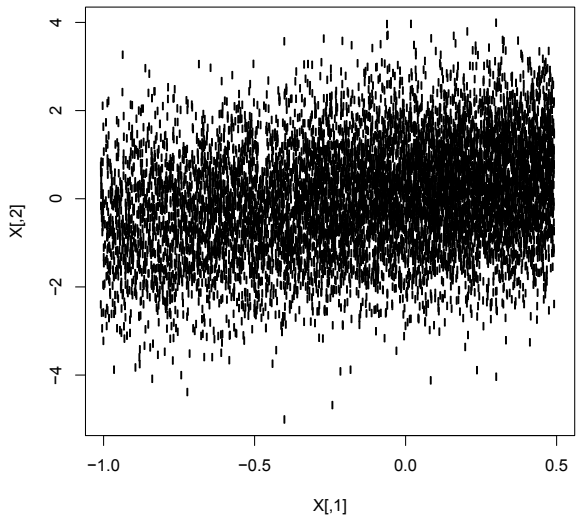


Figure B.1: Gibbs sample from multivariate the truncated Gaussian

Remark B.1 *Note that when the truncation region is too small then the rejection sampling approach will completely fails or utilizes more system time to generate samples. However, in this context Gibbs sampling approach performs well than rejection sampling approach.*

B.2 Gibbs sampling from truncated Student-t density

Let $\boldsymbol{X}$ be an d-dimensional random variable has density $T(\boldsymbol{\mu}, \Sigma; m)$ (multivariate truncated Student-t distribution), where $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$.

We know that the multivariate Student-t as the ratio of a multidimensional Gaussian to an independent $[\chi^2(m)/m]^{(1/2)}$ ($\chi^2(m)$ is Chi-Square distribution with m degrees of freedom) leads to a Gibbs sampling algorithm for $(w, z_1, ..., z_n)$ followed by the construction $x = \boldsymbol{\mu} + \boldsymbol{z}w^{-1}$.

Algorithm:

1. At stage i we have $w^{(i-1)}$ and $\boldsymbol{z}^{(i-1)}$ from the previous pass. Now, draw $w^{(i)}$ from $[\chi^2(m)/m]^{\frac{1}{2}}$ such that $a_i w^{(i)} \leq Z_i^{(i-1)} \leq b_i w^{(i)}$ is satisfied for all $i = 1, ..., d$. Note that at $i = 1$, $z^{(0)}$ will be a random observation from truncated multivariate Gaussian distribution.

2. Draw $\boldsymbol{z}$ from the multivariate Gaussian conditional on $w^{(i)}$ with support $a_i w^{(i)} \leq Z_i^{(i)} \leq b_i w^{(i)}$. Use above Section 7.1 procedure to generate an observation.

3. take a transformation $\mathbf{x}^{(i)} = \boldsymbol{\mu} + \boldsymbol{z}(w^{(i)})^{-1}$.

Example B.2
```
df = 2
mu = c(1,1,1)
sigma = matrix(c(  1, 0.5, 0.5, 0.5, 1, 0.5, 0.5, 0.5,
   1), 3, 3)
lower = c(-2,-2,-2)
upper = c(2, 2, 2)
X = rtmvt(n=10000, mu, sigma, df, lower, upper,
algorithm="gibbs").
```

B.3 Moments calculation for multivariate truncated Gaussian model

The computation of the first and second moments (mean vector $\boldsymbol{\mu}^* = E[\mathbf{x}]$ and covariance matrix $\boldsymbol{\Sigma}^*$ respectively) is not trivial for the truncated case, since they are obviously not the same as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the parametrization of $TN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{a}, \mathbf{b})$. We presented the computation of moments for the general double-truncated case and implemented the algorithm in the method `mtmvnorm()`, see Manjunath and Stefan [26].

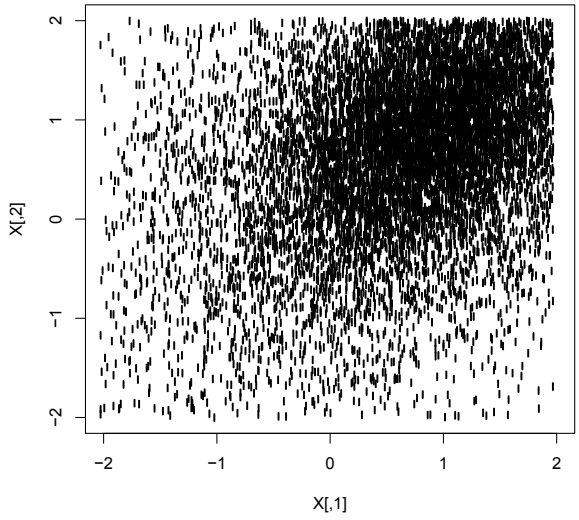


Figure B.2: Gibbs sample from the multivariate truncated Student-t distribution.

Example B.3

```
mu  = c(0.5, 0.5, 0.5)
sigma = matrix(c(  1,  0.6, 0.3, 0.6, 1, 0.2, 0.3,
  0.2,   2), 3, 3)
a  <- c(-Inf, -Inf, -Inf)
b  <- c(1, 1, 1)
mtmvnorm(mu, sigma, lower=a, upper=b).

 Output:
 $tmean
[1] -0.1524505 -0.1260711 -0.4050841

$tvar
           [,1]       [,2]       [,3]
[1,] 0.52235239 0.22046649 0.07733722
[2,] 0.22048774 0.51593373 0.03943862
[3,] 0.07734545 0.03941757 0.94595143.
```

Appendix C

Univariate extremal discriminant analysis

In the following we present some of the work on univariate extremal discriminant analysis, cf. Abdalla [1] and Nguimbi [29].

In Section 5.3 of Abdalla [1] introduces the univariate extremal discriminant analysis by considering exponential mixtures for classification. The deduced discriminant point has an explicit expression given by

$$x_{disc} := \frac{\mu^{(2)}\sigma^{(1)} - \mu^{(1)}\sigma^{(2)}}{\sigma^{(1)} - \sigma^{(2)}} + \log\left(\frac{p_2 c_1 \sigma^{(1)}}{p_1 c_2 \sigma^{(2)}}\right)^{\left(\frac{\sigma^{(1)}\sigma^{(2)}}{\sigma^{(1)}-\sigma^{(2)}}\right)}, \qquad \text{(C.1)}$$

where $\mu^{(i)}$ and $\sigma^{(i)}$ are location and scale parameters of corresponding exponential model.

Example C.1 *For classification between* $\mu^{(1)} = 10$, $\sigma^{(1)} = 2.5$ *and* $\mu^{(2)} = 8$, $\sigma^{(2)} = 4$ *with prior information* $p_1 = 0.35$, $p_2 = 0.65$ *and* $c_1 = c_2 = 1$. *Then the two exponential densities are discriminated at a point* $x_{disc} = 12.34$.

According to the derivation in Section 5.3.2 of Abdalla [1] with assumption of equality of location parameter, the optimal discriminant point for classifying mixtures of Beta and Pareto distributions can be given as

$$x_{disc} := \begin{cases} \mu + \left(\frac{\alpha^{(1)}(\sigma^{(1)})^{\alpha^{(1)}} p_1 c_1}{\alpha^{(2)}(\sigma^{(2)})^{\alpha^{(2)}} p_2 c_2}\right)^{\frac{1}{\alpha^{(1)}-\alpha^{(2)}}}, & \alpha^{(1)} < 0, \alpha^{(2)} < 0 \\ \mu - \left(\frac{|\alpha^{(1)}|(\sigma^{(1)})^{\alpha^{(1)}} p_1 c_1}{|\alpha^{(2)}|(\sigma^{(2)})^{\alpha^{(2)}} p_2 c_2}\right)^{\frac{1}{\alpha^{(1)}-\alpha^{(2)}}}, & \alpha^{(1)} > 0, \alpha^{(2)} > 0, \end{cases} \qquad \text{(C.2)}$$

where $\alpha^{(i)}$ and $\sigma^{(i)}$ are shape and scale parameters of respective densities.

Later, in the frame work of the constructing univariate extremal discriminant analysis Nguimbi [29] introduces discriminant analysis to Gompertz density.

In Section 5.2 of Nguimbi [29] the discriminant function for classifying between the two Gompertz density is given by,

$$\begin{aligned} x_{disc} &= \left(\frac{x-\mu^{(2)}}{\sigma^{(2)}}\right) - \left(\frac{x-\mu^{(1)}}{\sigma^{(1)}}\right) - \exp\left(\frac{x-\mu^{(2)}}{\sigma^{(2)}}\right) \\ &\quad + \exp\left(\frac{x-\mu^{(1)}}{\sigma^{(1)}}\right) + \log\frac{\sigma^{(1)}}{\sigma^{(2)}} \end{aligned} \tag{C.3}$$

where $p_1 = p_2 = 0.5$ and $c_1 = c_2 = 1$. Note that the above discriminant function do not leads to an explicit expression. So, one has to use numerical method for obtaining optimal discriminant point, cf. page 24 Nguimbi [29].

Now, if we assume equality of scale parameter, i.e., $\sigma^{(1)} = \sigma^{(2)} = \sigma = 1$, then the optimal discriminant point is given by,

$$x_{disc} = \log\left\{\frac{\mu^{(1)} - \mu^{(2)}}{\exp\left(-\mu^{(2)}\right) - \exp\left(-\mu^{(1)}\right)}\right\} \tag{C.4}$$

is an explicit expression for classification.

We now construct discriminant analysis for the truncated Gompertz density. The standard truncated Gompertz density, truncated at $X = u$ is given by,

$$f(x) = \exp(x)\exp\left(e^u - e^x\right),\ x > u. \tag{C.5}$$

Now, optimal discriminant point for classify between the two truncated Gompertz densities and they truncated at $u^{(1)}$ and $u^{(2)}$, respectively is given by

$$\begin{aligned} x_{disc} &= \left(\frac{x-\mu^{(2)}}{\sigma^{(2)}}\right) - \left(\frac{x-\mu^{(1)}}{\sigma^{(1)}}\right) - \exp\left(\frac{x-\mu^{(2)}}{\sigma^{(2)}}\right) \\ &\quad + \exp\left(\frac{x-\mu^{(1)}}{\sigma^{(1)}}\right) + \log\frac{\sigma^{(1)}}{\sigma^{(2)}} + e^{u^{(2)}} - e^{u^{(1)}}. \end{aligned} \tag{C.6}$$

If we have $u^{(1)} = u^{(2)} = u$ then above equation reduces to (C.3) and in addition if we assume equality of scale parameter $\sigma^{(1)} = \sigma^{(2)} = \sigma = 1$ it further reduces to (C.4).

In Chapter 6 of Nguimbi [29] discusses Ultimate and Penultimate convergence of the truncated Gompertz density to the exponential and Beta density, respectively. Further which implies that the convergence of (C.6) discriminant point to exponential and Beta discriminant point.

C.1 Predicting life span

In this section we present an analysis on predicting life span between males and females.

The given data are the life spans over 90 years old of person born in the Netherlands in the year 1877–1881. The data is found in the Xtremes software under the file name um-lspdu.dat. Beta sub-models with its finite upper tail limit is fitted to male and female age at death data. A calculated discriminant point between male and female age at death is 94, i.e., the life span mixture is discriminated at the age 94. In general, if do we have an observation or recorded life span value less than 94 we classify an observation to female population. Similarly an observation greater than 94 classified to male population. This is an example where one can find optimal class, i.e, male or female using recorded age at death value. For detailed discussion on this experiment cf. Section 5.3.5 Abdalla [1].

Classifying between the two subpopulation; given by the different genes, should be of more importance.

Author Index

Subject Index

Bibliography

[1] E.M.O. Abdalla. Structural modeling of aggreated financial ratios. *Ph.D. dissertation*, University of Siegen, 2007.

[2] C.W. Anderson, S.G. Coles, and J. Hüsler. Maxima of Poisson like variables and related triangular arrays. *Anna. of Appl. Probab.*, 7:953–971, 1997.

[3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Son, Inc., New York, 1958.

[4] G. Balkema and P. Embrechts. Multivariate excess distributions. *http://www.math.ethz.ch/%7Ebaltes/ftp/guuspe08Jun04.pdf2004*, 2004.

[5] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes, Theory and Applications*. Wiley, Chichester, 2005.

[6] M. S. Berman. *Sojurns and Extremes of Sochastic Processes*. Wadsworth, Belmont, CA, 1992.

[7] R. Bhattacharya and R. Rao. *Normal Approximation and Asymptotic Expansions*. Wiley, New York, 1976.

[8] M. Falk and A. Guillou. Peaks-over-threshold stability of multivariate generalized Pareto distributions. *J. Mult. Analysis*, 99:715–734, 2005.

[9] M. Falk, J. Hüsler, and R.-D. Reiss. *Laws of Small Numbers: Extremes and Rare Events*. Birkhäuser, Basel, 2nd ed (1st ed., birkhäuser, basel, 1994) edition, 2004.

[10] M. Falk, F. Marohn, and B. Tewes. *Foundations of Statistical Analyses and Applications with SAS*. Birkhäuser, Basel, 2002.

[11] R.A. Fisher. The use of multiple measurement in taxonomic problems. In *Ann. Eugen.*, volume 7, pages 376–386, 1936.

[12] R.A. Fisher and L.H.C. Tippett. On the estimation of the frequency distributions of the largest or smallest member of a sample. In *Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190, 1928.

[13] J. Galambos. *The Asymptotic Theory of Extreme Order Statistics*. Wiley, United States, 1978.

[14] E. Hashorva. Elliptical triangular arrays in the max-domain of attraction of Hüsler-Reiss distribution. *Statist. Probab. Letters*, 72:125–135, 2005.

[15] E. Hashorva. Extremes of asymptotically spherical and elliptical random vectors. *Insur.: Mathem. Econom.*, 36:285–302, 2005.

[16] E. Hashorva. On the multivariate Hüsler-Reiss distribution attracting the maxima of elliptical triangular arrays. *Statist. Probab. Letters*, 76:2027–2035, 2006.

[17] W. C. Horrace. Some results on the multivariate truncated normal distribution. *J. Mult. Analysis*, 94:209–221, 2005.

[18] J. Hüsler and D. Li. Tail approximations to the density function in EVT. *Extremes*, 9:131–149, 2006.

[19] J. Hüsler and R.-D. Reiss. Maxima of normal random vectors: between independence and complete dependence. *Statist. Probab. Letters 7*, pages 283–286, 1989.

[20] H. Joe. Multivariate extreme-value distributions with applications to environmental data. *Canad. J. Statist.*, 22:47–64, 1994.

[21] E. Kaufmann and R.-D. Reiss. Approximation rates for multivariate exceedances. *J. Statist. Plann. Inference*, 45:235–245, 1995.

[22] S. Kocherlakota, N. Balakrishnan, and K. Kocherlakota. The linear discriminant function: sampling from the truncated normal distribution. *Biomet. J.*, 29:131–139, 1987.

[23] S. Kotz and S. Nadarajah. Some extremal type elliptical distributions. *Statist. Probab. Letters*, 54:171–182, 2001.

[24] M. Koutras. On the performance of the linear discriminant function for spherical distributions. *J. Mult. Anaylsis*, 22:1–12, 1987.

[25] P.A. Lachenbruch. *Discriminant Analysis*. Hafner Press, New York, 1975.

[26] B.G. Manjunath and S. Wilhelm. Moments calculation for the double truncated multivariate normal density. *http://ssrn.com/abstract=1472153*, September, 2009.

[27] R. Michel. Simulation of certain multivariate generalized Pareto. *Extremes*, 10:83–107, 2007.

[28] R. Von Mises. On the classification of observation data into distinct groups. *Ann. Math. Stat.*, 16:68–73, 1945.

[29] J.A. Nguimbi. Extremale Diskriminanzanalyse, eindimensionaler Fall. *Bachelor dissertation*, University of Siegen, 2009.

[30] R.-D. Reiss. *Approximate Distributions of Order Statistics*. Springer, New York, 1989.

[31] R.-D. Reiss. *A Course on Point Processes*. Springer, New York, 1993.

[32] R.-D. Reiss and M. Thomas. *Statistical Analysis of Extreme Values*. Birkhäuser, Basel, 3rd ed. (1st ed. 1997, 2nd ed. 2001, both in birkhäuser, basel) edition, 2007.

[33] S.I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer, New York, 1987.

[34] H. Rootzén and N. Tajvidi. Multivariate generalized Pareto distributions. *Bernoulli*, 12:917–93, 2006.

[35] N. Tajvidi. Characterization and some statistical aspects of univariate and multivariate generalised Pareto distributions. *Ph.D. dissertation*, University of Göteborg, 1996.

[36] G. M. Tallis. Elliptical and radial truncation in normal populations. *Ann. Math. Statist.*, 34:940–944, 1963.

[37] Venables and Ripley. Linear discriminant analysis. *http://stat.ethz.ch/R-manual/R-devel/library/MASS/html/lda.html*.

[38] Venables and Ripley. Package MASS. *http://cran.r-project.org/web/packages/MASS/MASS.pdf*.

[39] H. Wakaki. Discriminant analysis under elliptical populations. *Biometrika*, 24:257–298, 1994.

[40] B. L. Welch. Note on discriminant functions. *Biometrika*, 31:218–220, 1939.

[41] S. Wilhelm and B.G. Manjunath. tmvtnorm: A package for the truncated multivariate normal distribution. *The R Journal*, 2(1):25–29, June 2010.

[42] S. Wilhelm and B.G. Manjunath. tmvtnorm: Truncated multivariate normal and Student-t distribution. *http://cran.r-project.org/web/packages/tmvtnorm/index.html*, 2010.