DEPENDENCE STRUCTURES BEYOND COPULAS: A NEW MODEL OF A MULTIVARIATE REGULAR VARYING DISTRIBUTION BASED ON A FINITE VON MISES-FISHER MIXTURE MODEL

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Joerg Rothenbuehler

January 2005

DEPENDENCE STRUCTURES BEYOND COPULAS: A NEW MODEL OF A MULTIVARIATE REGULAR VARYING DISTRIBUTION BASED ON A FINITE VON MISES-FISHER MIXTURE MODEL

Joerg Rothenbuehler, Ph.D.

Cornell University 2005

A multivariate regular varying distribution can be characterized by its marginals and a finite measure on the unit sphere. That measure is referred to as the spectral measure of the distribution. The spectral measure describes the structure of the dependence between the marginal distributions. An important class of multivariate regular varying distributions are multivariate extreme value distributions. Existing models for multivariate regular varying distributions in general and multivariate extreme value distributions in particular do not utilize the spectral measure. They focus on closed form equations of the cumulative distribution function. The resulting models are not flexible enough to give a realistic and adequate description of the dependence structure of real life data.

We propose a new model for multivariate regular varying distributions, based on a very flexible parametric model of the spectral measure. We use a finite mixture model to obtain a model with as much flexibility as needed to accurately describe the spectral measure of real life data.

Since the spectral measure is a measure on the unit sphere, we chose directional distributions as the distributions of the components of the mixture model. Directional distributions provide models for the distribution of random variables on unit spheres. In particular, we use the von Mises-Fisher distribution. Its properties allow it to be interpreted as an directional analogue of the well known normal distribution on a Euclidian space.

We describe how to estimate the parameters of this new model from datasets. We introduce a modified version of the likelihood ratio test to decide on how many components are needed for an accurate model of the spectral measure.

We show how our model explains the structure of the spectral measure of several financial time series. We develop a comprehensive model for a multivariate regular varying distribution that is based on our model of the spectral measure. As one particular application of this new model we describe how it can be used for portfolio optimization. We found that our model gives much more accurate results than two other well established models. It significantly improves on the deficiencies of the two existing models.

BIOGRAPHICAL SKETCH

Joerg Rothenbuehler was born in Muensterlingen, Switzerland, in July 1973. He attended high school at the “Kantonschule Kreuzlingen” between April 1988 and January 1993. Following his graduation he served in the Swiss Army for 4 month before starting his undergraduate studies in Mathematics at the Swiss Federal Institutes of Technology in Zurich, Switzerland in November 1993. He graduated with Honors in April 1998. After serving another 4 month in the army he was promoted to the rank of a corporal. While working as a teaching assistant at the Swiss Federal Institutes of Technology in Zurich he got admitted to the doctoral program at the School of OR&IE, Cornell University. He started his studies on the hill in August 1999. He was awarded the special Masters degree in Fall 2002.

This thesis is dedicated to my parents and their love and support

ACKNOWLEDGEMENTS

I would fist and foremost like to thank my advisor, Gennady Samorodnitsky, for his constant and valuable guidance and support. He always managed to help me find a way out when the harsh realities of the data and computational feasibility were closing in on me and threatened my ideas. His thorough knowledge and understanding of a wide range of topics in probability were an invaluable help that made this thesis possible.

I am also thankful for having Bob Jarrow and Bruce Turnbull on my special committee. Bob Jarrow introduced me to the world of mathematical finance. Bruce Turnbull's hints and ideas influenced my work considerably.

I have many fond memories for working as a TA for Philip Protter. I will not forget his humor and many words of wisdom and advice that he offered me throughout our quest to teach the students probability and finance. I would like to thank Sid Resnick for teaching me the rigors and foundations of heavy tailed analysis including many fruitful discussion that benefitted my work and Shane Henderson for volunteering as faculty advisor to the Swiss Club of Cornell.

During my five years on the hill have met many wonderful people and as I leave I take memories with me that shall not be forgotten. I will fondly remember Ipsita Mukherjee and our nights out dancing, Tuncay Alparslan for putting up with me as a roommate, Sam Steckley for the good company on many a Monday night at Benchwarmer's, Stefan Wild, Marc Berthoud and Martin Roth for joining forces in getting the Swiss Club started, Rita Bakalian for being my patient landlady in my first two years in town, Davina Kunvipusilkul for putting up with my two left feet at our first swing dance classes and finally Yuliya, Devashish, Pascal, Millie, Kay, Pascal, Patrick, Yuriy, Milan and everybody who let me win a game of tennis every now and then for their friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In the recent years significant attention has been paid to the development of models for multivariate distributions. The main motivation was the lack of good models for the joint distribution of returns of financial assets. The difficulty in creating reasonable models is the often complex structure of the dependence between such assets. Increasingly, the concept of copulas has been advertised as a versatile tool to create such distributions. A copula is a multivariate distribution whose marginals have a uniform distribution on [0,1]. It is used as a model of the dependence structure. Together with appropriate models for the marginal distributions they can be used as a model for the joint distribution of the assets of interest. We refer to Joe (1997) and Embrechts et al. (2003) as excellent references on copulas and multivariate distributions.

The most popular copulas are the ones based on elliptical distributions. Prominent members of the family of elliptical distributions are the multivariate normal and the multivariate t distributions. The dependence structure in an elliptical distribution can be characterized by its correlation matrix. It is this simplicity that makes elliptical distributions appealing in practice. Unfortunately they are not a very realistic model of the dependence structure between different financial assets. They main criticism of elliptical models is that the correlation is not an adequate description of the dependence structure. Papers by Blyth (1996), Shaw (1997) and Embrechts et al. (1999) demonstrate that models based on linear correlations can not accurately capture the non linear dependence that is present in financial data. The main reason is that they assume that the dependence between extreme returns is the same as between moderate returns. This assumption is wrong. We refer to the work of Longin and Solnik (1998), who show in an empirical study that the dependence between large negative returns is much closer than suggested by the correlation coefficient of the entire data. They also found that the dependence structure is not symmetric. This is, however, another feature of elliptical distributions. They imply that the dependence structure between positive and negative returns is the same.

In the light of the findings of Longin and Solnik (1998) the need for models that specifically address the dependence structure in the tails of a distribution becomes evident. Multivariate extreme value theory provides us with a tool to develop models that address this need. Several different models and methods have been introduced in the last 20 years. These models are based on multivariate extreme value distributions or multivariate regular varying distributions. A multivariate regular varying distribution can be characterized by its marginals and a finite measure on the unit sphere. Most commonly, the distribution of a random vector $\mathbf{X}$ is called multivariate regular varying, if there exists a constant $\alpha > 0$ such that the following limit exists for all $x > 0$

$$\frac{\mathbb{P}[\|\mathbf{X}\| > tx, \mathbf{X}/\|\mathbf{X}\| \in \cdot]}{\mathbb{P}[\|\mathbf{X}\| > t]} \longrightarrow^{\nu}_{t\to\infty} x^{-\alpha} S(\cdot), \tag{1.1}$$

where $\longrightarrow^{\nu}$ denotes vague convergence on $\mathbb{S}^{d-1}$, the d dimensional unit sphere, and S stands for the spectral measure.

The spectral measure describes the structure of the dependence in the tails between the marginal distributions. An important class of multivariate regular varying distributions are multivariate extreme value distributions. Existing models of multivariate regular varying distributions in general and multivariate extreme value distributions in particular do not describe the distribution via the spectral measure. Instead, they focus on closed form equations of the cumulative distribution function. Examples of such models can be found in Resnick (1986), Tawn (1988), Joe et al. (1992), de Haan and Resnick (1993), Stărică (1999), Embrechts et al. (1997), Klueppelberg and May (1998), Embrechts (2000), Embrechts et al. (2003) and Breymann et al. (2003) and others.

However, so far none of these proposed models is flexible enough to give a realistic description of the dependence structure of the tails of a distribution. They usually use one or two parameters to describe the dependence structure between their marginal components. As a consequence, their spectral measure, that can be calculated from the cumulative distribution function, has a very simple structure. Typically these spectral measures are therefore fairly simple. Consider in contrast the spectral measure of the joint distribution of the absolute values of the log returns of the stocks of BMW and Siemens. A scatter plot of these log returns is given in the right hand side of Figure 1.1. Since the data is positive and bivariate, its, spectral measure is a measure that lives in the first quadrant of the unit circle $\mathbb{S}^1$, that is on the

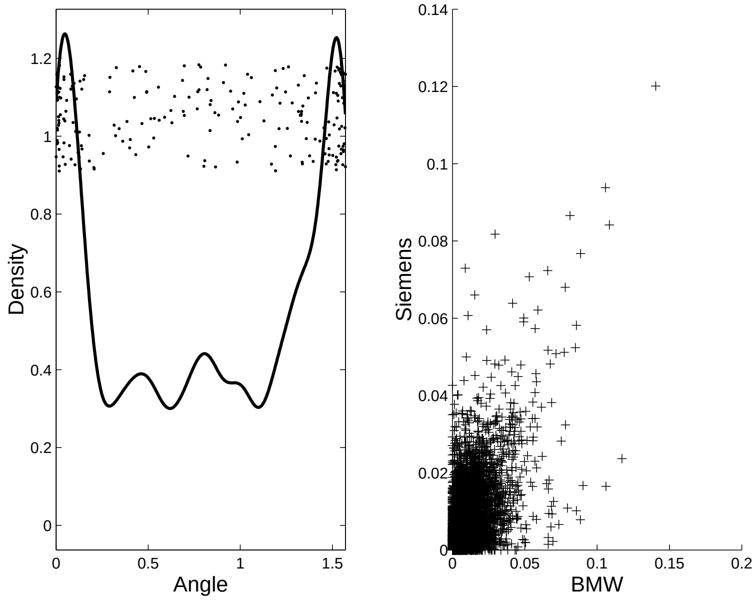

Figure 1.1: *Scatter plot of the absolute values of the log returns of the stocks of BMW and Siemens (right) and estimate of the spectral measure of the corresponding distribution(left).*

set $\{(x,y) \in \mathbb{S}^1 : x = \cos\theta, y = \sin\theta, \theta \in [0,\pi/2]\}$. A non parametrical estimate of the density of the spectral measure of the joint distribution is given in the left hand side of Figure 1.1. We explain in detail how we obtain estimates of the spectral measure in Chapter 2. We can clearly see that the density indicates that the dependence structure in the tails is too complicated to be described by a single parameter.

This is the motivation for the research presented in this thesis. We propose a new model for multivariate regular varying distributions. Instead of focusing on the joint cumulative distribution function, we focus on the spectral measure. Since the spectral measure is a measure on the unit sphere, we work with directional distributions. Directional distributions are distributions designed to model observations on the unit sphere. The topology of the unit sphere $\mathbb{S}^{d-1}$ is different from the one of the Euclidian space $\mathbb{R}^d$. Directional distributions reflect this different topology. We decided to use the von Mises-Fisher distributions on $\mathbb{S}^{d-1}$ as the corner stone of our models. The von Mises-Fisher distributions form a parametric family. It is parameterized by the mean direction, a point in $\mathbb{S}^{d-1}$, and a concentration parameter. It can be seen as an analogue of the normal distribution on $\mathbb{R}^d$. Additionally, we make use of the concept of finite mixture models. That is, we assume that the spectral measure has a density of the form

$$f(\mathbf{x}) = \sum_{i=1}^{m} p_i f_i(\mathbf{x}); \mathbf{x} \in \mathbb{S}^{d-1}. \tag{1.2}$$

The parameters p_i are called the weights of the mixture and satisfy $p_i > 0, i = 1, ..., m$ and $\sum_{i=1}^{m} p_i = 1$. The densities $f_i(\mathbf{x})$ are called the component densities. The concept of a finite mixture model provides us with the flexibility needed to model complex dependence structures. The number of components, m, is itself a parameter of the model. The drawback is that the estimation of the parameters not a trivial task. We used the EM algorithm to estimate the parameters of the model for a fixed number of components. The EM algorithm is an algorithm specifically designed for the calculation of maximum likelihood estimates in finite mixture models. We refer to Dempster et al. (1977), Redner and Walker (1984), Titterington et al. (1985) and McLachlan and Peel (2000) for references on mixture models and the EM algorithm. Additionally, we have to decide on how many components are needed to accurately describe the spectral measure. If we choose a number that is too small, we will miss important features of the spectral measure. On the other hand, having too many components renders the model too complicated. Traditionally this kind of problem is addressed with a likelihood ratio test. Unfortunately, the regularity conditions that guarantee the usual central chi-square distribution of the test statistic under the null-hypothesis do not hold in the framework of mixture models. Instead, we were able to use results based on work of Vuong (1989), White (1982) and Lo et al. (2001). They show that under certain conditions the asymptotical distribution of the test statistic follows a weighted sum of central chi-square distributions. We found that if the true spectral measure is not a finite mixture distribution of von Mises-Fisher distribution, we can apply these results to our model. This enables us to determine the number of components needed to accurately model the spectral measure, while avoiding models with too many components. We sometimes also consulted other statistics to decide on the number of components. These other statistics performed well in empirical studies but lack a theoretical justification.

We found that our model gives an accurate description of the spectral measure of bivariate and three dimensional datasets of financial assets. For higher dimensional data, we did not have datasets of sufficient sample size to perform a meaningful statistical analysis.

We develop a comprehensive model for a multivariate regular varying distribution that is based on our model of the spectral measure. This model consists of a part describing the tails of the distribution and a separate part describing the body of the distribution.

The model of tails utilizes our model of the spectral measure to describe the dependence between the marginal components. The model of the body consists of a simple multivariate normal distribution, although other choices are possible, without changing the tail behavior of the resulting distribution.

As an application of our new model we consider the problem of portfolio optimization. We concentrate on the bivariate case. We calculate the portfolios that minimize the expected shortfall while having a certain expected return. We compare the resulting portfolios with optimal portfolios based on two other models. The first is the bivariate normal distribution model and the other is a model based on a t copula.

While the optimal portfolios based on the normal model are fairly similar to the ones based on our model, the normal model severely underestimates the risk of the portfolio. The estimates and predictions based on our model were very accurate. The portfolios based on the t copula model suffer from problems related to the estimation of the parameters of that model. As a result, these portfolios do not achieve the expected return they are designed to have. In addition, despite having a much smaller average return than the portfolios based on our model and the normal model, they have am expected shortfall that is comparable in size to the ones of the portfolios based on our model and the normal model.

The thesis is organized as follows: in Chapter 2 we give an introduction to the extreme value theory and its related topics. Chapter 3 provides an introduction into directional distributions and their properties. Chapter 4 explains finite mixture models in general and finite mixtures of von Mises-Fisher models in particular. We also explain the parameter estimation using the EM algorithm, the likelihood ratio test and the other statistics used to decide on the number of components. In Chapter 5 we present the results of modelling the spectral measure of several different financial assets. In Chapter 6 we develop our comprehensive model for a multivariate distribution, based on the proposed mixture model of the spectral measure. Finally, Chapter 7 documents the results of the portfolio optimizations based on our model and the two selected alternative models.

Chapter 2

Extreme Value Theory

2.1 Univariate Extreme Value Theory

2.1.1 Asymptotic Behavior of Maxima

Let $(X_1, ..., X_n)$ be i.i.d. random variables with some distribution F. Extreme Value Theory describes the asymptotic behavior of the probability distribution of $M_n := max(X_1, ..., X_n)$. Of course, we have for any n

$$\mathbb{P}[M_n \leq x] = \mathbb{P}[X_1 \leq x, ..., X_n \leq x] = F^n(x). \tag{2.1}$$

Let x_F denotes the right endpoint of F, defined as $x_F := sup\{x \in \mathbb{R} : F(x) < 1\}$. One can show that

Proposition 2.1.1

$M_n \longrightarrow x_F$ with probability 1, as $n \to \infty$.

Proof: See Resnick (1986). ■

To illustrate the significance and use of extreme value theory, it is helpful to consider the better known result of the Central Limit Theorem. Recall, that if $(X_1, ..., X_n)$ are i.i.d. random variables following a distribution with finite mean μ and variance $\sigma^2 < \infty$ and n is sufficiently large, then the following approximation holds:

$$Z := \frac{S_n - n\mu}{\sqrt{n}\sigma}, \text{ is approximately distributed as } N(0,1), \tag{2.2}$$

where $S_n = \sum_{i=1}^{n} X_i$. Consider this result for the case of exponentially distributed random variables X_i, distributed as $exp(\lambda)$. Since $X_i \geq 0$, $S_n = \sum_{i=1}^{n} X_i$ converges to ∞ with probability 1. The analogous statement in the context of extreme value theory is made in Proposition 2.1.1. In the light of the degenerate limit of S_n, the central limit theorem quantifies the limit of $a_n^{-1}(S_n - b_n)$ with $a_n = \sqrt{n}\sigma$ and $b_n = n\mu$. This result is very useful in approximating the distribution of S_n for large n. In the same way, extreme value theory describes the convergence of $c_n^{-1}(M_n - d_n)$ for appropriate sequences c_n and d_n. The Fisher Tippet Theorem below is the basis of extreme value theory and its applications discussed in this thesis. It can be seen as an the counterpart of the central limit theorem in the field of extreme value theory.

Theorem 2.1.2 ***(Fisher-Tippet)***

Let (X_n) be a sequence of i.i.d. random variables and let $M_n := max(X_1, ..., X_n)$. If there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$, such that for a non-degenerate distribution H

$$c_n^{-1}(M_n - d_n) \Longrightarrow M, \textit{ with distribution } H(x), \tag{2.3}$$

then H is one of the following types of distributions:

$$\begin{aligned}
Fréchet: \quad \Phi_\alpha(x) \quad &= \begin{cases} 0 & x \leq 0 \\ exp(-x^{-\alpha}) & x > 0 \end{cases} & \alpha > 0 \\
Weibull: \quad \Psi_\alpha(x) \quad &= \begin{cases} exp(-(-x)^{\alpha}) & x \leq 0 \\ 1 & x > 0 \end{cases} & \alpha > 0 \\
Gumbel: \quad \Lambda(x) \quad &= exp(-e^{-x}), & x \in \mathbb{R}
\end{aligned}$$

We call two distribution functions F and G of the same type, if, for all $x \in \mathbb{R}$

$$F(x) = G(ax + b)$$

for two constants a and b.

Proof: See Resnick (1986). ■

The three distributions Φ_α, Ψ_α and Λ are called Extreme Value Distributions, EVD. If (2.3) holds for (X_i) with distribution F, we say that F is in the *maximum domain* *of attraction of H* and write $F \in MDA(H)$. The MDA's for the three EVDs are well understood. The extreme value distributions and the corresponding norming constants are known for the most common distributions. In the following, we give a very brief overview. A more detailed discussion can be found in Embrechts et al. (1997) or in Resnick (1986).

2.1.2 Domains of Attractions for Φ_α, Ψ_α and Λ and the GEV

Fréchet

In order to characterize the domain of attraction of the Fréchet distribution, it is useful to recall the definition of a regular varying function.

Definition 2.1.3

A measurable function g: $\mathbb{R}_+ \mapsto \mathbb{R}_+$ *is regular varying at* ∞ *with index* $\alpha \in \mathbb{R}$*, if for any* $x > 0$ *we have that*

$$\lim_{t\to\infty} \frac{g(xt)}{g(t)} = x^\alpha. \tag{2.4}$$

In this case we use the notation $g \in RV_\alpha$.

The classical example of a function that is regular varying at ∞ with tail index α is of course $g(x) = x^\alpha$. We say that a random variable with distribution function F is regular varying with tail index α, $\alpha > 0$, if its tail function $\bar{F} := 1 - F$ is regular varying at ∞ with index $-\alpha$. If the distribution F is regular varying with index α, then there is a slowly varying function $L(x)$, such that

$$1 - F = x^{-\alpha} L(x), x > 0. \tag{2.5}$$

A function $L(x)$ is called slowly varying with if

$$\lim_{t\to\infty} \frac{L(xt)}{L(t)} = 1, x > 0.$$

The relationship (2.5) expresses the fact that, asymptotically, the tail function behaves like a power function. This is in contrast to the behavior of the Exponential or the Normal distribution, whose tail functions approach zero at an exponential and superexponential rate, respectively. Typical examples of regular varying distributions are the Cauchy and the Pareto distributions. It is not hard to show that the Fréchet distribution Φ_α is regular varying with tail index α.

The following theorem says that all distributions with regular varying tail function $\overline{F}$ belong to the maximum domain of attraction of the Fréchet distribution with the same tail index α.

Proposition 2.1.4

The distribution with cdf F belongs to the maximum domain of attraction of Φ_α*, if and only if* $1 - F \in RV_{-\alpha}, \alpha > 0$.

Proof: See Embrechts et al. (1997), p. 131f. ■

It follows for example, that the Cauchy distribution is in $MDA(\Phi_1)$ and that $c_n = n/\pi$, $d_n = 0$, so that $\pi n^{-1} M_n \to \Phi_1$. Other prominent members of $MDA(\Phi_\alpha)$ are the stable distribution with $\alpha < 2$ and the Pareto distribution. It is widely accepted that the log returns of financial time series have marginal distributions with regular varying tails. For this reason, $MDA(\Phi_\alpha)$ has received more attention in research papers than the other two EVDs. See Section 2.1.5 for more results on $MDA(\Phi_\alpha)$.

Weibull

The most important fact about $MDA(\Psi_\alpha)$ is, that all its members have a finite right endpoint. Well known distributions in $MDA(\Psi_\alpha)$ are the Uniform and Beta distributions. The following result gives a mathematical description similar to Proposition 2.1.4.

Proposition 2.1.5

A distribution with cdf F belongs to $MDA(\Psi_\alpha)$ *if and only if* $x_F < \infty$ *and* $1 - F(x_F - x^{-1}) = x^{-\alpha}L(x)$ *for a slowly varying function L.*

Proof: See Embrechts et al. (1997) ■

Consider for example the Uniform distribution. Since $x_F = 1$ and $1 - F(1 - x^{-1}) = x^{-1}$, the Uniform distribution is in $MDA(\Psi_1)$. One finds that $c_n = n^{-1}$ and $d_n = 1$. Similarly, the Beta distribution with parameters $a > 0$ and $b > 0$, given by the density $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, $0 < x < 1$, is in $MDA(\Psi_b)$. We see that the parameter α of the Weibull distribution indicates "how fast" the distribution $F \in MDA(\Psi_\alpha)$ approaches its right endpoint.

Gumbel

The maximum domain of attraction of the Gumbel distribution contains most distributions with an infinite right endpoint with light right tails. We say that a distribution has a light right tail, if all positive moments $\mathbb{E}[(X^+)^k]$ exist and are finite. This is in contrast to the distributions in $MDA(\Phi_\alpha)$, which only have finite moments up to order α. The Gumbel distribution itself has the property that

$$\lim_{x \to \infty} \frac{1 - \Lambda(x)}{e^{-x}} = 1.$$

Therefore, all distributions with an exponential or a "close to" exponential tail are in $MDA(\Lambda)$. In particular, the Exponential, Gamma, Normal and Log-normal distributions all belong to the domain attraction of the Gumbel distribution. For a more formal discussion, see section 3.3.3 in Embrechts et al. (1997).

The Generalized Extreme Value Distribution, GEV

It turns out that the three parametric families Φ_α, Ψ_α and Λ are related to one type of distribution. The distribution is called *Generalized Extreme Value Distributions, GEV.* We will differentiate between a standard GEV, which has one parameter, and the general GEV. The general GEV has three parameters and can be used to represent the three parametric families Φ_α, Ψ_α and Λ in one family.

Definition 2.1.6 (GEV)

Define the standard generalized extreme value distribution as the distribution with cdf

$$H_\xi(x) = \begin{cases} exp\left(-\left(1+\xi\cdot x\right)^{-1/\xi}\right) & \xi \neq 0 \\ exp\left(-exp\left(-x\right)\right) & \xi = 0 \end{cases} \tag{2.6}$$

where $1+\xi\cdot x > 0.$

Related to this distribution is a three parameter location-scale family, consisting of all distributions that are of the same type as the standard generalized extreme value distribution. We refer to such distributions as generalized extreme value distributions. The cdf of such a distribution is given by

$$H_{\xi,\mu,\psi}(x) = \begin{cases} exp\left(-\left(1+\xi\frac{x-\mu}{\psi}\right)^{-1/\xi}\right) & \xi \neq 0 \\ exp\left(-exp\left(-\frac{x-\mu}{\psi}\right)\right) & \xi = 0 \end{cases} \tag{2.7}$$

for x *such that* $1+\xi\frac{x-\mu}{\psi} > 0$ *with* $\xi \in \mathbb{R}$, $\mu \in \mathbb{R}$ *and* $\psi \in \mathbb{R}_+$.

We will refer to both H_ξ *and* $H_{\xi,\mu,\psi}$ *as GEV.*

Note that H_ξ and $H_{\xi,\mu,\psi}$ are of the same type if and only if they have the same value for the parameter ξ. The two additional parameters μ and ψ in the location-scale family $H_{\xi,\mu,\psi}$ are called the location and scale parameter, respectively. The role of those parameters is made clear in next section. The crucial parameter is the so called *shape parameter* ξ:

- $\xi > 0$: The Fréchet distribution can be expressed as a distribution of the type of H_ξ with $\xi = 1/\alpha$. That is, we have $\Phi_\alpha = H_{\xi,\mu,\psi}$ for some constants $\mu \in \mathbb{R}$ and $\psi > 0$ and $\xi = 1/\alpha$.
- $\xi = 0$: The Gumbel distribution can be expressed as a distribution of the type of H_ξ with $\xi = 0$.
- $\xi < 0$: The Weibull distribution can be expressed as a distribution of the type of H_ξ with $\xi < 0$.

This inclusion of the three extreme value distributions in one parametric family with three parameters is important for the applications. Suppose, we would like to decide whether the data comes from a distribution with heavy or rather light tails. We could fit $H_{\xi,\mu,\psi}(x)$ to maxima from that dataset and observe whether ξ is significantly different from 0. For more details, see section 2.1.4

2.1.3 Maxima of Stationary Time Series

The results explained in the previous sections hold for i.i.d. data. A more reasonable assumption for real life data, like the one considered in this thesis, is that the observations are not from an i.i.d. time series, but rather from a stationary one. We therefore need to consider how the extreme value distribution of a stationary time series relates to the one of i.i.d. data with the same marginal distribution. How far away from the i.i.d. case can one go and still have the same distribution of the maximum ? The answer to that question is discussed in detail in Leadbetter et al. (1983). In the following, we give a brief summary:

Let $X_1, X_2, ..., X_n$ be a strictly stationary time series and $M_n = max(X_1, ..., X_n)$. Let furthermore $\widetilde{X}_1, ..., \widetilde{X}_n$ be an i.i.d. series with the same marginal distribution as

$X_1, X_2, ..., X_n$ and let $\widetilde{M}_n = max(\widetilde{X}_1, ..., \widetilde{X}_n)$. Finally, assume that

$$c_n^{-1}(\widetilde{M}_n - d_n) \Longrightarrow M, \text{ distributed as } H_{\xi,\mu,\psi}. \tag{2.8}$$

Define the sequence of functions $u_n(x)$ as $u_n(x) = c_n x + d_n$. Then (2.8) is equivalent to

$$\mathbb{P}[\widetilde{M}_n \leq c_n x + d_n] = \mathbb{P}[\widetilde{M}_n \leq u_n(x)] \longrightarrow H_{\xi,\mu,\psi}(x).$$

The constants c_n and d_n of the linear functions $u_n(x)$ are determined by the distribution of the strictly stationary sequence $X_1, ..., X_n$. If the distribution of $X_1, ..., X_n$ satisfies two technical conditions that can be expressed by means of $u_n(x)$, then we also have

$$\mathbb{P}[M_n \leq c_n x + d_n] = \mathbb{P}[M_n \leq u_n(x)] \longrightarrow H_{\xi,\mu,\psi}(x).$$

The two conditions are as follows:

Condition $D(u_n)$*: For any integers p,q and n*

$$1 \leq i_1 < ... < i_p < j_1 < ... < j_q \leq n$$

such that for $j_1 - i_p \geq l$ *we have that*

$$\left| P\left(\max_{i \in A_1 \cup A_2} X_i \leq u_n\right) - P\left(\max_{i \in A_1} X_i \leq u_n\right) P\left(\max_{i \in A_2} X_i \leq u_n\right) \right| \leq \alpha_{n,l},$$

where $A_1 = \{i_1, ..., i_p\}, A_2 = \{j_1, ..., j_q\}$ *and* $\alpha_{n,l} \to 0$ *as* $n \to \infty$ *for some sequence* $l = l_n = o(n)$.

$D(u_n)$ can be interpreted as stating that the sequence (X_i) should not have a too strong serial dependence. For example, if X_i is a Gaussian process, it is known that $D(u_n)$ is satisfied if the auto-covariance function $\gamma(h)$ satisfies $\gamma(h)\log(h) \to 0$, as $h \to \infty$. This conditions is very weak. It is satisfied by all ARIMA and even all fractional ARIMA processes. The latter are examples of processes having long range dependence in the sense that the sequence $\gamma(h)$ is not absolutely summable. We will assume, that the datasets considered in this thesis can be modelled by distributions for which $D(u_n)$ is satisfied.

Condition $D'(u_n)$*: The relation*

$$\lim_{k\to\infty} \limsup_{n\to\infty} n \sum_{j=2}^{[n/k]} P(X_1 > u_n, X_j > u_n) = 0.$$

$D'(u_n)$ says that extreme observations of X_i do not occur in clusters, but are isolated events. The distributions of the data under investigation in this thesis usually are not assumed to satisfy $D'(u_n)$. The reason is that the data exhibits behavior that makes the usage of models satisfying $D'(u_n)$ unreasonable. We refer to Leadbetter et al. (1983) and Embrechts et al. (1997) for more detailed discussions on this subject.

Because we do not assume that the condition $D'(u_n)$ holds, we cannot assume that the distribution of M_n converges to the same GEV distribution as the one of $\widetilde{M}_n$. It turns out that the limit distribution of the maximum M_n can be expressed via H, the limit distribution of $\widetilde{M}_n$ and the extremal index of the time series $X_1, ..., X_n$, if this one exists. The extremal index is a measure of the amount of clustering in the tails. Before giving the definition of the extremal index, we note that for an i.i.d. time series:

$$\begin{aligned} \mathbb{P}[\widetilde{M}_n \leq u_n] &= \mathbb{P}^n[\widetilde{X} \leq u_n] \\ &= exp[n \cdot ln(1 - \mathbb{P}[\widetilde{X} > u_n])] \\ &\approx exp[-n\bar{F}(u_n)] \end{aligned} \tag{2.9}$$

The last approximations follows from the Taylor Series extension approximation: $ln(1-x) \approx -x$ for small x. Based on this motivation we state that for a given $\tau \in [0, \infty]$ and a sequence u_n of real numbers we have

$$n\bar{F}(u_n) \to \tau \Longleftrightarrow \mathbb{P}[\widetilde{M}_n \leq u_n] \to e^{-\tau}. \tag{2.10}$$

If $D'(u_n)$ is violated, (2.9) usually does not hold. Instead, one may observe the following, for $\theta \in [0,1]$:

$$\begin{aligned}\mathbb{P}[M_n \le u_n] &\approx \mathbb{P}^{\theta}[\widetilde{M_n} \le u_n] \\ &= \mathbb{P}^{n\theta}[\widetilde{X} \le u_n] \\ &= exp[\theta \cdot n \cdot ln(1 - \mathbb{P}[\widetilde{X} > u_n])] \\ &\approx exp[-\theta \cdot n\bar{F}(u_n)]\end{aligned} \tag{2.11}$$

If (2.11) holds we get from (2.10):

$$n\bar{F}(u_n) \to \tau \Longleftrightarrow \mathbb{P}[M_n \le u_n] \to exp[-\theta\tau]. \tag{2.12}$$

Based on these observations we define:

Definition 2.1.7

Consider a stationary time series $(X_k)_{k\in\mathbb{N}}$ with marginal distribution F and let $M_n = max(X_1, ..., X_n)$. We say that $(X_k)_{k\in\mathbb{N}}$ has extremal index $\theta \in [0,1]$, if, for every τ, there exists a sequence (u_n), such that

$$\begin{aligned}\lim_{n\to\infty} n\bar{F}(u_n) &= \tau \\ \lim_{n\to\infty} \mathbb{P}[M_n \le u_n] &= e^{-\theta\tau}\end{aligned} \tag{2.13}$$

We refer to Embrechts et al. (1997) as a reference on the extremal index. The extremal index can be interpreted as the reciprocal of the average cluster size. The effect of clustering in the data is illustrated in Figure 2.1. The top plot shows 1000 realization of an AR(1) process $X_n = \alpha \cdot X_{n-1} + Y_n$ with Y_n i.i.d. with a student's t distribution with 2 degrees of freedom and $\alpha = .8$. Such an AR(1) process has extremal index $1 - \alpha^2$=.36. The bottom plot shows 1000 i.i.d. realizations with the same marginal distribution as in the top plot.

The influence of the extremal index on the limit distribution of the maxima is summarized in the following Theorem.

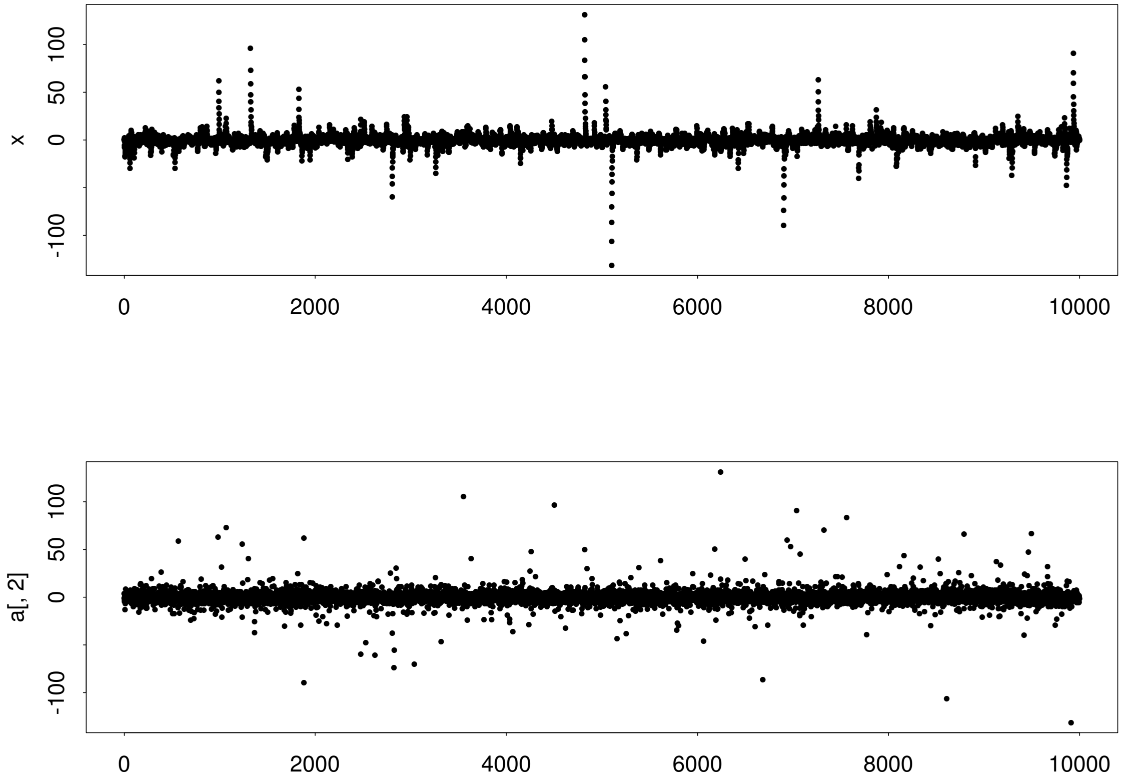

Figure 2.1: *Effect of clustering in the tails for an AR(1) process (top) compared to an i.i.d process (bottom).*

Theorem 2.1.8

Suppose that (X_n) *is a stationary time series with extremal index* θ *and define* $M_n = max(X_1, ..., X_n)$*. Furthermore, let* $(\widetilde{X}_n)$ *be an i.i.d. sequence of random variables with the same distribution as* (X_n) *and* $\widetilde{M}_n = max(\widetilde{X}_1, ..., \widetilde{X}_n)$*. Then*

$$\lim_{n\to\infty} \mathbb{P}[c_n^{-1}(\widetilde{M}_n - d_n) \leq x] = H(x) \tag{2.14}$$

for a GEV H*, if and only if*

$$\lim_{n\to\infty} \mathbb{P}[c_n^{-1}(M_n - d_n) \leq x] = H^{\theta}(x). \tag{2.15}$$

Proof: Embrechts et al. (1997) ■

In the light of the above equations it is important to note that if H is a GEV, so is H^{θ}. This point is made precise by the following equations:

$$H^{\theta}_{\xi,\mu,\psi}(x) = H_{\xi,\mu,\psi}(ax + b), \tag{2.16}$$

where $a = \theta^{-\xi}$, $b = (1 - \theta^{-\xi})(u - \frac{\psi}{\xi})$ if $\xi \neq 0$ and $a = 1$, $b = -\psi log(\theta)$, if $\xi = 0$. Moreover,

$$H_{\xi,\mu,\psi}(ax + b) = H_{\xi,\hat{\mu},\hat{\psi}}(x), \tag{2.17}$$

where $\hat{\mu} = \frac{\mu - b}{a}, \hat{\psi} = \psi a$. Equations (2.16) and (2.17) mean that the adjustments for the unknown extremal index θ are incorporated in the model by the parameters μ and ψ. Equation (2.17) also shows that the same is true for the norming constants. Equation (2.17) gives us the justification for working with the block wise maxima when fitting the GEV model to data. We do not have to scale the maxima with the norming constants from Theorem 2.1.2. The two equations (2.16) and (2.17) show that the generalized extreme value distribution (2.7) with its three parameters is flexible enough to incorporate these adjustments into the model by using the location and the scale parameters μ and ψ.

2.1.4 Fitting of a GEV Model to a Dataset

Maximum Likelihood Estimators

In order to fit a GEV to extremes of a dataset, we may proceed as follows. We divide the data set into blocks of the same sample size. Within each block, we determine the maximum. The set of the thus obtained block wise maxima is treated as an i.i.d. sample from a GEV. The parameter estimates are now determined using a maximum likelihood estimation based on this sample of block wise maxima. The obvious question in this context is: Into how many blocks are we to divide the data? Or in other words: How many observations should make one block?

The answer to that question depends on the data. On the one hand, we have to make sure that the blocks are large enough so that their maxima are nearly i.i.d. and their distribution is close enough to a GEV. On the other hand, we want to keep the block size as small as possible. If the chosen block size is too large, the number of blocks may not be sufficient to produce a reliable estimator. On the other hand, if the block size is too small, the distribution of the block wise maxima may not be close to a GEV and they may not be independent. We usually tried several different block sizes and then checked the quality of the fit do determine a good block size.

The three parameters are estimated using a maximum likelihood method. A numerical procedure is needed to find the solutions to the complex log-likelihood equations. We used EVIS 5.0, a software package on SPLUS, to carry out the calculations. If $\xi > -.5,$ Smith (1985) shows that the MLEs are consistent and asymptotically efficient estimators. That is, they are asymptotically normally distributed and their covariance matrix is the inverse of the Fisher-Information matrix. The goodness of the fit may be tested using QQ-plots and similar exploratory tools.

2.1.5 Estimating the Shape Parameter in $MDA(\Phi_\alpha)$

If we assume that the distribution function has regular varying tails, we have additional methods at hand for estimating the shape parameter ξ or, equivalently, the tail index $\alpha = 1/\xi$. The most prominent such method is the Hill estimator. See Resnick (2002) for a list of references.

Assume, that $X_1, ..., X_n$ is a sample of non-negative, i.i.d. random variables with a distribution with regular varying tails. Let $X_{(i)}$ be the i^{th} largest value, $1 \leq i \leq n$.

The Hill estimator

The Hill estimator of the tail index $\alpha = \xi^{-1}$ is given by $H_{k,n}^{-1}$, where

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} log\left[\frac{X_{(i)}}{X_{(k+1)}}\right]. \tag{2.18}$$

The Hill estimator is a consistent estimator of $\alpha = \xi^{-1}$ and, under second order conditions, asymptotically normally distributed:

$$\sqrt{k}(H_{k,n}^{-1} - \alpha) \Rightarrow N(0, \alpha^2), \text{ provided that } n \to \infty, k \to \infty, k/n \to 0$$

A summary of the consistency results for the Hill estimators as well as a good list of references is provided in Embrechts et al. (1997) on p.336 ff. Resnick and Stărică (1998) proved consistency of the Hill estimator for certain classes of dependent data. The quality of the estimator $H_{k,n}^{-1}$ depends on the choice of k. If k is chosen too large, the estimator becomes biased, because data that is not sufficiently far enough in the tails is used. On the other hand, if k is chosen too small, the estimator becomes unreliable due to the small sample size used in the estimation and useful information is wasted. In practice, it is customary to study what has become known as the ”Hill-plot” $\{k, H_{k,n}^{-1}, 1 \leq k \leq n\}$. One then looks for an area of k where the plot resembles a horizontal line. In some cases this works nicely, in other cases this may be very frustrating and difficult, as no such area is easily identifiable. A more detailed discussion of the performance of the Hill estimator in practice can be found in Embrechts et al. (1997).

The QQ-estimator

The QQ-estimator is sometimes a valuable alternative to the Hill estimator. It is based on the idea that if the distribution of the data is regular varying with tail index α and k is small compared to the sample size n, then the points of the set $\{-\log(i/(k+1)), \log(X_{(i)}), 1 \leq i \leq k\}$ should form a straight line with slope α^{-1}. Hence, the slope of a least squares line fitted to the set $\{-\log(i/(k+1)), \log(X_{(i)}), 1 \leq i \leq k\}$ should be a reasonable estimate of α^{-1}. Therefore, we define the QQ- estimator as

$$\widehat{\alpha}_{k,n}^{-1} := SL(\{-\log(i/(k+1)), \log(X_{(i)}), 1 \leq i \leq k\}), \tag{2.19}$$

where

$$SL(\{x_i, y_i\}, i = 1, .., n\}) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\frac{1}{n}\sum_{i=1}^{n} x_i)(\frac{1}{n}\sum_{i=1}^{n} y_i)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2}$$

is the slope of the line fitted to $\{x_i, y_i\}, i = 1, .., n\}$ by means of least squares. We have that

$$\widehat{\alpha}_{k,n}^{-1} \xrightarrow{P} \alpha^{-1}, \text{provided that } k \to \infty \text{ and } n/k \to \infty.$$

However, one is faced with the same problems of choosing an appropriate value for k as in the case of the Hill estimator. Similar to the case of the Hill estimator, a plot of $\{k, \widehat{\alpha}_{k,n}, 1 \leq k \leq n\}$ is studied and one tries to identify an area of k, where the plot resembles a horizontal line. These plots have a tendency to be easier to interpret than the Hill plots and it may be easier to find an reasonable estimate of α.

Other estimators have been proposed, see Embrechts et al. (1997), Section 6.4.

2.1.6 Peaks over Threshold

We first give the definition of a distribution related to the GEV family

Definition 2.1.9

Define the standard Generalized Pareto distribution as the distribution with cdf

$$G_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi} & \xi \neq 0 \\ 1 - e^{-x} & \xi = 0 \end{cases} \tag{2.20}$$

where

$$x \geq 0 \qquad if \quad \xi \geq 0$$

$$0 \leq x \leq -1/\xi \quad if \quad \xi < 0.$$

As in the case of the GEV, there is a three parameter location scale family associated with this distribution that is flexible enough to allow fits to datasets. It is constructed by replacing x in (2.20) by $(x-\nu)/\beta$:

$$G_{\xi,\beta,\nu}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-\nu}{\beta}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - exp\left(-\frac{x-\nu}{\beta}\right) & \xi = 0 \end{cases} \tag{2.21}$$

where $1 + \xi \frac{x-u}{\beta} \geq 0$ and $\xi \in \mathbb{R}, \mu \in \mathbb{R}$ and $\beta \in \mathbb{R}_+$. Note that all the members of this parametric family are of the same type as the standard GPD. We refer to these distributions as Generalized Pareto distributions, GPD. We will denote the special case $G_{\xi,\beta,0}(x)$ by $G_{\xi,\beta}(x)$. Similar to the GEV, the crucial parameter is the shape parameter ξ, while β and ν are chosen to make the distribution flexible enough for fitting to a data set.

Define the excess distribution function of a random variable X as

$$F_u(x) = \mathbb{P}[X - u \leq x | X > u].$$

Then we can write

$$F(x) = \mathbb{P}[X \leq x | X > u] \cdot \mathbb{P}[X > u] = F_u(x-u)(1 - F(u)).$$

The connection with the results about the distribution of maxima is given by the following equation, given in Embrechts et al. (1997). We have for all $\xi \in \mathbb{R}$:

$$F \in MDA(H_\xi) \Longleftrightarrow \lim_{u \to x_F} \sup_{0<x<x_F} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0 \tag{2.22}$$

for some positive function $\beta(u)$. This result says that the GPD $G_{\xi,\beta}$ appears as the limit distribution of scaled excesses over high thresholds of i.i.d. data in the domain of attraction of H_ξ. For high thresholds u we may thus use the approximation $F_u(x) \approx G_{\xi,\beta}(x)$. This leads to the following approximation for high quantiles of F. We have for $x > u$ and u large enough:

$$F(x) = F_u(x-u)(1-F(u)) \approx G_{\xi,\beta}(x-u)(1-F(u)) = G_{\xi,\beta,u}(x)(1-F(u)) \tag{2.23}$$

For estimation purposes, one chooses a high threshold u, sets $\nu = u$ and then estimates the parameters ξ and β, for example using maximum likelihood techniques. There is no obvious preferred choice for the threshold u. One faces similar problems as for the estimation of the tail index α or the parameters of a GEV. If u is chosen too high, only very few observation remain above the threshold and the estimates of ξ and β become unreliable due to their large variability. On the other hand, if u is chosen too low, too many points are above the threshold and one can no longer expect that a GPD is a good approximation of the distribution of the excesses. Hence, one would introduce a bias in the estimates of ξ and β. We usually consulted QQ-plots and other exploratory tools to assess the quality of a fit and subsequently chose the lowest threshold that resulted in good fits.

In this context, it is important to note that the class of GPDs is closed under changes of the threshold as explained in the following. We have

$$\frac{\bar{G}_{\xi,\beta,\nu}(w+u)}{\bar{G}_{\xi,\beta,\nu}(u)} = \bar{G}_{\xi,\beta+\xi\cdot u,u}(w), \tag{2.24}$$

where $\bar{G} = 1 - G$. This equality is important, because both the left hand side and the right hand side can be seen as an approximation of

$$\mathbb{P}[X > w + u | X > u] = \frac{\mathbb{P}[X > w + u]}{\mathbb{P}[X > u]} \tag{2.25}$$

if $u > \nu$ is large enough. To see this for the left hand side, choose a threshold $\nu \geq 0$. If we have $u > \nu$, we can use (2.23) as an approximation of $\mathbb{P}[X > w + u]$ and $\mathbb{P}[x > u]$ to get:

$$\mathbb{P}[X > w + u] \approx (1 - G_{\xi,\beta,\nu}(w + u))\mathbb{P}[X > \nu] =: \bar{G}_{\xi,\beta,\nu}(w + u)\mathbb{P}[X > \nu] \tag{2.26}$$

and similarly

$$\mathbb{P}[X > u] \approx (1 - G_{\xi,\beta,\nu}(u))\mathbb{P}[x > \nu] =: \bar{G}_{\xi,\beta,\nu}(u)\mathbb{P}[x > \nu]. \tag{2.27}$$

Hence we obtain the approximation

$$\mathbb{P}[X > w + u | X > u] = \frac{\mathbb{P}[X > w + u]}{\mathbb{P}[X > u]} \approx \frac{\bar{G}_{\xi,\beta,\nu}(w + u)}{\bar{G}_{\xi,\beta,\nu}(u)}. \tag{2.28}$$

For the right hand side of (2.24), we consider the application of (2.23) when choosing $\nu = u$. We get similar to (2.26):

$$\mathbb{P}[X > w + u] \approx \bar{G}_{\widetilde{\xi},\widetilde{\beta},u}(w + u)\mathbb{P}[X > u] \tag{2.29}$$

for two parameters $\widetilde{\xi}$ and $\widetilde{\beta}$. This leads to the following approximation of (2.25):

$$\mathbb{P}[X > w + u] \approx \bar{G}_{\widetilde{\xi},\widetilde{\beta},u}(w + u) \tag{2.30}$$

If the technique of approximating the distribution of high quantiles by means of a GPD is to be consistent for different choices of the threshold, the two right hand sides of (2.28) and (2.30) need to be the same. That is, we need to be able to express $\widetilde{\xi}$ and with $\widetilde{\beta}$ with ξ, β and u. This is exactly what (2.24) asserts us is true. It says that $\widetilde{\xi} = \xi$ and that $\widetilde{\beta} = \beta + \xi u$.

A nice discussion of the GPD and its properties, including the results in this section, can be found in Embrechts et al. (1997). It also provides a much deeper introduction into the Peaks over Threshold techniques. We used the EVIS 5.0 software for all our statistical analysis involving Peaks over Threshold.

2.2 Multivariate Extreme Value Theory

In the univariate case, the Fisher-Tippet Theorem, Theorem 2.1.2, describes the class of limiting distributions for extremes. In the multivariate case, the class of possible limit distributions for extremes is much wider, because of the dependence structure between the marginal components. Usually, the limit distribution of multivariate extremes is described by

- the marginal distributions, which are given by the Fisher-Tippet Theorem and were discussed in the previous section;
- a finite measure on the unit sphere, referred to as the spectral or angular measure, that describes the dependence structure between the different components.

We first describe the possible limit distributions of multivariate extremes. Then we show how multivariate regular variation can be used to characterize the MDA's. Finally, we show how the spectral measure can be consistently estimated. Good references on these topics were written by Resnick (1986), Resnick (2002), Stărică Stărică (1999), and Einmahl et al. (2001).

2.2.1 Limit Distributions for Multivariate Extremes

We first introduce the notation that we will use throughout this section. All operations on vectors are understood componentwise. For example, we have for two vectors $\mathbf{x}$ and

$\mathbf{y}$ and two points $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^d$:

$$\begin{array}{lll}
\mathbf{x} \leq \mathbf{y} & \text{means} & x^{(i)} \leq y^{(i)}, i = 1, ..., d, \\
\mathbf{x} < \mathbf{y} & \text{means} & x^{(i)} < y^{(i)}, i = 1, ..., d, \\
\mathbf{x} + \mathbf{y} & \text{means} & (x^{(1)} + y^{(1)}, ..., x^{(d)} + y^{(d)}), \\
\mathbf{x} \cdot \mathbf{y} & \text{means} & (x^{(1)} \cdot y^{(1)}, ..., x^{(d)} \cdot y^{(d)}), \\
\mathbf{x} \bigvee \mathbf{y} & \text{means} & (x^{(1)} \bigvee y^{(1)}, ..., x^{(d)} \bigvee y^{(d)}), \\
(\mathbf{a}, \mathbf{b}) & \text{means} & (a^{(1)}, b^{(1)}) \times ... \times (a^{(d)}, b^{(d)}) \subseteq \mathbb{R}^d, \ \textit{if } \mathbf{a} < \mathbf{b}
\end{array}$$

Let $\{\mathbf{X}_i\}_{i\in\mathbb{N}} = \{(X_i^{(1)}, ..., X_i^{(d)})\}_{i\in\mathbb{N}}$ be i.i.d. random vectors in $\mathbb{R}^d$. We are considering limit distributions for $\mathbf{M}_n = (M_n^{(1)}, .., M_n^{(d)}) = \left(\bigvee_{i=1}^n X_i^{(1)}, ..., \bigvee_{i=1}^n X_i^{(d)}\right)$. Denote the joint cdf of $\mathbf{X}_1$ with $F(\mathbf{x})$. Assume that there exist sequences of vectors $\mathbf{b}_n \in \mathbb{R}^d$ and $\mathbf{a}_n > \mathbf{0}$, such that

$$\mathbb{P}\left[\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \leq \mathbf{x}\right] = F^n(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n) \longrightarrow G(\mathbf{x}), \text{ as } n \to \infty, \tag{2.31}$$

where $G(\mathbf{x})$ has non-degenerate marginals $G_i(x), i = 1, \ldots, d$. By the results from the previous section, we know that each of the G_i is a GEV. However, the marginals need not be of the same type. To simplify the task of describing the class of possible limits distributions with non-degenerate marginals, it is helpful to standardize the marginals to a specified distribution. We chose the unit Fréchet distribution Φ_1, introduced in Section 2.1.1. That enables us to use results about multivariate regular variation. Different standardizations could be and have been considered. They lead to similar results as the one described in the following. The first result asserts that the standardization does not create any changes in the convergence behavior.

Proposition 2.2.1 *Define the random vectors $\{\mathbf{X}_i\}_{i\in\mathbb{N}}$ as above with joint distribution function F and marginal distribution functions F_i. Assume that (2.31) holds and that* *the marginals of the limit distribution are non-degenerate. Define for* $i = 1, .., d$

$$\psi_i(x) = (1/(-\log(G_i)))^{\leftarrow}(x), x > 0 \tag{2.32}$$

and

$$G_*(\mathbf{x}) = G(\psi_1(x^{(1)}), ..., \psi_d(x^{(d)})).$$

Then $G_*(\mathbf{x})$ *has* Φ_1 *marginals* $G_{*i}(x)$*. If* G *is a multivariate extreme value distribution, so is* G_**.*

Define $U_i(x_i) := 1/(1 - F_i(x_i)), i = 1, .., d$*, and let* F_* *be the distribution of* $\left(U_1\left(X_1^{(1)}\right), ..., U_d\left(X_1^{(d)}\right)\right)$*. That is, let*

$$F_*(\mathbf{x}) = F(U_1^{\leftarrow}(x^{(1)}), .., U_d^{\leftarrow}(x^{(d)})).$$

Then, if $F \in D(G)$*, we have that* $F_* \in D(G_*)$ *and*

$$P\left[\bigvee_{j=1}^{n} U_i(X_j^{(i)})/n \le x^{(i)}, i = 1, .., d\right] = F_*^n(n\mathbf{x}) \to G_*(\mathbf{x}), \text{ as } n \to \infty. \tag{2.33}$$

Conversely, if (2.33) holds and if for $i = 1, .., d$*:* $F_i^n(a_n^{(i)}x + b_n^{(i)}) \to G_i(x)$*, non-degenerate, we have that* $F \in D(G)$ *and that (2.31) holds.*

Proof: See Resnick (1986). ■

The following theorem gives the exact description of the class of limit distributions with Φ_1 marginals. Proposition 2.2.1 asserts that this is sufficient for describing the class of multivariate extreme value distributions, since for every extreme value distribution G there exist a standardized extreme value distribution G_*, obtained from G by the transformation given by (2.32).

Theorem 2.2.2 *The following are equivalent:*

1. G_* *is a multivariate extreme value distribution with* Φ_1 *marginals.*

2. *There is a Radon measure μ_* on $\mathbb{E} = [\mathbf{0}, \infty) \setminus \{\mathbf{0}\} \subseteq \mathbb{R}^d$ such that*

$$G_*(\mathbf{x}) = exp(-\mu_*([\mathbf{0}, \mathbf{x}]^c)), \tag{2.34}$$

such that for $r > 0$ and a Borel set $A \in \mathbb{S}^{d-1} = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\| = 1\}$

$$\mu_*\{\mathbf{y} \in \mathbb{E} : \|\mathbf{y}\| > r, \|\mathbf{y}\|^{-1}\mathbf{y} \in A\} = r^{-1}S_*(A), \tag{2.35}$$

where S_* is a finite measure on $\aleph = \mathbb{E} \cap \mathbb{S}^{d-1}$ satisfying the marginal conditions

$$\int_{\aleph} x^{(i)} S_*(d\mathbf{x}) = 1, i = 1, ..., d. \tag{2.36}$$

Proof: See Resnick (1986) ■

The finite measure S_* in (2.35) is referred to as the spectral measure or angular measure. The Radon measure μ_* is referred to as the exponent measure. Both measures completely describe the distribution function G_*. S_* can be interpreted as the description of the dependence structure of G_* and hence it describes the dependence of the extremes of $\mathbf{X}_1$. From the above two results we see that the extreme value distribution of $G(\mathbf{x})$ of $\mathbf{X}_1$ can be described by

- the marginal distributions G_i,
- the spectral measure S_* of the standardized extreme value distribution G_*.

In that sense, the spectral measure has a similar function as the copula in describing the dependence structure of the limit distribution. Recall that the copula C of a distribution function F with marginals $F_i, i = 1, ..., d$ is given by

$$F(x_1, .., x_d) = C(F_1(x_1), ..., F_d(x_d)) \Longleftrightarrow C(u_1, ..., u_d) = F(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)).$$

The copula, having standardized Uniform[0,1] marginals, describes the dependence structure, to which the desired marginal distributions are attached.

It is worthwhile to mention two specific cases of possible dependence structures described by S_*, or equivalently by μ_*.

1. The exponent measure concentrates on $\bigcup_i \{0 \times ... \times (0, \infty) \times ... \times 0\}$. In that case, the spectral measure is a discrete measure concentrating its mass on the axes $e_i = \{\mathbf{x} : x_j = 0, j \neq i\}, i = 1, ..., d$. In this case, if $\mathbf{X}_*$ is distributed as G_*, the marginal components X_i are independent. In the case $d = 2$ this means that S_* concentrates on the x and y axis. As a consequence of equation (2.36), in polar coordinates, S_* is a measure with mass 1 on the points 0 and $\pi/2$.

2. μ_* concentrates on $\{t\mathbf{1}, t > 0\}$ and hence S_* concentrates on $\|\mathbf{1}\|^{-1}\mathbf{1}$. In that case there is total dependence among the marginal components $X_*^{(i)}$ of $\mathbf{X}_*$. That is, we have $\mathbb{P}[X_*^{(1)} = ... = X_*^{(d)}] = 1$. In the case $d = 2$, this means that S_* puts all its mass in the point $\mathbf{x} \in \mathbb{S}^1 : x_1 = x_2$. Expressed in polar coordinates, S_* is a point mass concentrated on $\pi/4$.

2.2.2 Regular Variation and Domains of Attraction

The spectral measure can be used to identify and describe the domains of attraction of G_*, using regular variation. Regular variation of univariate random variables was introduced in Definition 2.1.3. In the multivariate setting a function $f : C \subset \mathbb{R}^d \rightarrow (0, \infty)$, where C is a cone, is called *regular varying with limit function* $\lambda(\mathbf{x})$, if and only if there exists a function $V : (0, \infty) \rightarrow (0, \infty)$ such that $V \in RV_\alpha$ and for all $\mathbf{x} \in C$ we have

$$\lim_{t \to \infty} \frac{f(t\mathbf{x})}{V(t)} = \lambda(\mathbf{x}).$$

The following theorem describes how the domains of attraction of a multivariate extreme value distribution with Φ_1 marginal distributions can be characterized.

Theorem 2.2.3 *Let F_*, G_*, μ_* and S_* be as in Proposition 2.2.1 and Theorem 2.2.2. Let $\mathbb{E} = [\mathbf{0}, \infty) \setminus \{\mathbf{0}\}$ and $\aleph = \mathbb{E} \cap \mathbb{S}^{d-1}$. The following are equivalent:*

1) $F_* \in D(G_*)$

2) $1 - F_*$ is regular varying on $\mathbb{E}$ with

$$\lim_{t\to\infty} \frac{1 - F_*(t\mathbf{x})}{1 - F_*(t\mathbf{1})} = \frac{-\log(G_*(\mathbf{x}))}{-\log(G_*(\mathbf{1}))} = \frac{\mu_*([\mathbf{0}, \mathbf{x}]^c)}{\mu_*([\mathbf{0}, \mathbf{1}]^c)}. \tag{2.37}$$

3) Let $M_+(\mathbb{E})$ denote the space of Radon measures on $\mathbb{E}$. Suppose $\mathbf{X}_1$ is distributed as F_*. Then

$$tF_*(t\cdot) = t\mathbb{P}[\frac{\mathbf{X}_1}{t} \in \cdot] \xrightarrow{\nu} \mu_* \text{ in } M_+(\mathbb{E}), \text{ as } t \to \infty. \tag{2.38}$$

Here $\xrightarrow{\nu}$ stand for vague convergence.

4) Define $(\mathbf{R}, \Theta) := (\|\mathbf{X}_1\|, \|\mathbf{X}_1\|^{-1}\mathbf{X_1})$. In $M_+((0, \infty] \times \aleph)$ we have that

$$t\mathbb{P}[(\frac{\mathbf{R}}{t}, \Theta) \in \cdot] \xrightarrow{\nu} r^{-2}dr \times S_*(d\theta). \tag{2.39}$$

5) Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be i.i.d. random vectors with joint distribution function F_*. For any sequence $k = k(n) \to \infty$ such that $n/k \to \infty$ and $k(n) \sim k(n+1)$

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{(\mathbf{X}_i/\frac{n}{k})} \Rightarrow \mu_* \tag{2.40}$$

in $M_+(\mathbb{E})$.

Proof: See Resnick (2002). ■

Remarks: The theorem shows that the extreme value distribution G_* in whose domain of attraction F_* is, can be found and described by the regular variation property (2.37). The extreme value distribution G_* is determined by the exponent measure μ_*. In polar coordinates, this exponent measure appears as a product measure on $(0, \infty] \times \aleph$ of $r^{-2}dr$ and the spectral measure $S_*(d\theta)$. Both the spectral measure or the exponent measure completely describe G_*.

To identify the extreme value distribution G in whose domain of attraction F lies, proceed as follows:

1. Compute the marginals F_i and then find the univariate extreme value distribution G_i such that $F_i^n(a_n^{(i)}x + b_n^{(i)}) \to G_i(x)$.
2. Compute F_* and use Theorem 2.2.3 to find G_*, such that $F_* \in D(G_*)$.
3. Calculate

$$G(\mathbf{x}) = G(x^{(i)}, ..., x^{(d)}) = G_* \left(\psi_1^{\leftarrow}(x^{(1)}), ..., \psi_d^{\leftarrow}(x^{(d)})\right).$$

In Theorem 2.2.3 we worked with the assumption that the all the marginal distributions are Φ_1. This assumptions is clearly unrealistic as far as real data is concerned. There is no reason why one should assume that the tail indexes of each marginal distribution should be the same, not to mention why they should be equal to one. We therefore have to make different, more general, assumptions about the joint regular variation of the distribution of $\mathbf{X}_1$ than the one given in the theorem above. We assume that the distribution satisfies the two regular variation conditions given below, found in Resnick (2002). As before, let $\mathbb{E} = [\mathbf{0}, \infty] \setminus \{\mathbf{0}\}$. Define the measures μ_{α_i} on $(0, \infty]$ by $\mu_{\alpha_i}(x, \infty] = x^{-\alpha_i}, \alpha_i > 0$. Define the sequences $\{b_n^{(i)}, n \geq 1, 1 \leq i \leq d\}$ such that

$$\lim_{n\to\infty} b_n^{(i)} = \infty, i = 1, ..., d.$$

Marginal Condition *For each i=1,...,d, we have in* $M_+((0, \infty])$

$$n\mathbb{P}\left[\frac{X_1^{(i)}}{b_n^{(i)}} \in \cdot\right] \xrightarrow{\nu} \mu_{\alpha_i}. \tag{2.41}$$

Global Condition *There exists a measure* μ *on Borel subsets of* $\mathbb{E}$*, such that in* $M_+(\mathbb{E})$

$$n\mathbb{P}\left[\frac{\mathbf{X_1}}{(b_n^{(1)}, ..., b_n^{(d)})} \in \cdot\right] \xrightarrow{\nu} \mu. \tag{2.42}$$

We say that the random vector $\mathbf{X_1}$ *is jointly regular varying*, if the both the Marginal and the Global Condition are met. Equation (2.41) is equivalent to the definition of regular variation for univariate random variables, given earlier. The marginal condition therefore states that the marginal distributions have regular varying tails. The global condition is a more general formulation of (2.38), given in Theorem 2.2.3 above. In that case we may choose $b_n^{(i)} = n, i = 1, ..., d$ and we have $\alpha_i = 1, i = 1, ..., d$. The global condition describes the dependence structure among the marginal components of $\mathbf{X_1}$. It is not hard to show that (2.41) and (2.42) are necessary and sufficient conditions for

$$\mathbb{P}\left[\bigvee_{j=1}^{n} \frac{\mathbf{X_j}}{\mathbf{b_n}} \leq \mathbf{x}\right] \rightarrow G(\mathbf{x}) = \exp(-\mu([\mathbf{0}, \mathbf{x}]^c)) \tag{2.43}$$

and the limit distribution $G(x)$ has marginal distributions Φ_{α_i}.

The following result states that this definition is consistent with results in Theorem 2.2.3, where we assumed that all marginal distributions are Φ_1. It essentially rephrases Proposition 2.2.1 in the language of regular variation.

Proposition 2.2.4 *Assume that X_1 is a jointly regular varying non-negative random vector. That is, assume that the Global and Marginal Conditions formulated above hold for some sequences $\mathbf{b}_n$, defined as above. Let $F_{(i)}(x)$ be the i$^t h$ marginal distribution function and define*

$$U_{(i)}(x) = \frac{1}{1 - F_{(i)}(\cdot)}(x), x > 1.$$

Then we have

1. ***Standard Global Convergence:***

$$nF_*(n\cdot) := n\mathbb{P}\left[\left(\frac{U_{(i)}(X_1^{(i)})}{n}, i = 1, ..., n\right) \in \cdot\right] \xrightarrow{\nu} \mu_* \text{ in } M_+(\mathbb{E}), \tag{2.44}$$

where

$$\mu_*(t\cdot) = t^{-1}\mu_*(\cdot) \tag{2.45}$$

on Borel subsets of $\mathbb{E}$.

2. ***Standard Marginal Convergence:***

$$n\mathbb{P}\left[\frac{U_{(i)}(X_1^{(i)})}{n} > x\right] \to x^{-1}, x > 0. \tag{2.46}$$

Proof: See Resnick (2002) ■

The Proposition essentially verifies that we can replace the convergence condition given in (2.31) with the regular variation conditions given above and still apply the transformations described in Proposition 2.2.1. As a Corollary to Proposition 2.2.4, we get the following important relationship between the exponent measures μ and μ_* from (2.43) and (2.44)

Corollary 2.2.5 *Let* μ *be as in (2.43) and let* μ_* *be as in (2.44). Then*

$$\mu_*([\mathbf{0}, \mathbf{x}]^c) = \mu([\mathbf{0}, \mathbf{x}^{\mathbf{1}/\alpha}]^c). \tag{2.47}$$

Proof: See Resnick (2002) ■

This relationship plays an important role in the estimation of the spectral measure, discussed in the next section.

2.2.3 Estimation of the Exponent and Spectral Measure

The Ranks method

For references on the following results we refer to Resnick (2002). Let $\mathbf{X}_i$, $i = 1, ..., n$ be a sequence of i.i.d. positive random vectors as above. Define the (anti)-ranks for $i = 1, ..., d$ as

$$r_j^{(i)} = \sum_{l=1}^{n} \mathbf{1}_{[X_l^{(i)} > X_j^{(i)}]} \text{ and } \mathbf{r_j} = (r_j^{(1)}, ..., r_j^{(d)}) \tag{2.48}$$

to be the number of i^{th} components bigger than $X_j^{(i)}$. Then we have, as $k \to \infty, n \to \infty, k/n \to 0$,

$$\frac{1}{k}\sum_{j=1}^{n} \epsilon_{\left(\frac{k}{\mathbf{r_j}}\right)} \Rightarrow \mu_* \text{ in } M_+(\mathbb{E}). \tag{2.49}$$

Applying the transformation into polar coordinates $T(\mathbf{x}) := (\|\mathbf{x}\|, \|\mathbf{x}\|^{-1}\mathbf{x}) =: (R, \theta)$ we get with $T(\frac{k}{\mathbf{r_j}}) =: (R_{j,k}, \theta_{j,k})$ and applying the continuous mapping theorem that

$$\frac{1}{k}\sum_{j=1}^{n} \epsilon_{(R_{j,k},\theta_{j,k})} \Rightarrow c\mu_1 \times S_* \text{ in } M_+((0,\infty] \times \aleph)$$

for a constant $c > 0$. Therefore, if our sample size n is large enough, we may use the approximation

$$\frac{1}{k}\sum_{j=1}^{n} \epsilon_{(R_{j,k},\theta_{j,k})}((1,\infty] \times A) \approx c\mu_1(1,\infty]) \times S_*(A),$$

for a Borel set $A \subset \aleph$ and a constant $c > 0$. This motivates the following estimator for the spectral measure:

$$\widehat{S_{k,n}}(\cdot) := \frac{\sum_{j=1}^{n} \mathbf{1}_{(R_{j,k}>1)}\epsilon_{(\theta_{j,k})}(\cdot)}{\sum_{j=1}^{n} \mathbf{1}_{(R_{j,k}>1)}} \Rightarrow S_* \tag{2.50}$$

This estimator depends on a good choice of k. We used the Stărică plot to make a choice of k, see below. The advantage of the ranks method is that we do not have to estimate the different tail indexes $\alpha_i > 0, i = 1.., d$. These estimations can be difficult, as explained in the section about the Hill estimator. However, the ranks, the data used to estimate S_* in (2.50), are not independent. For this reason, asymptotic properties of the estimator $\widehat{S_{k,n}}$ are hard to come by. It is also an open question whether the ranks statistics (2.48) is a sufficient statistic for the description of the exponent measure. Therefore, it may be desirable to consider a second approach that avoids these problems. However, it forces us to use the possibly unreliable estimates of the different tail indexes.

Direct approach adjusting the tail indexes

The following method for estimating the spectral measure is based on the results found in de Haan and Resnick (1993).

Recall from (2.40) that if $k \to \infty, n/k \to \infty$, we have that

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{(\mathbf{X}_i/\frac{n}{k})} \Rightarrow \mu_*,$$

if the data $\mathbf{X}_n$ is i.i.d. with distribution F_* as defined in Proposition 2.2.1. Similarly, if X_i has a regular varying distribution F_i with tail index $\alpha_i > 0$, equation (2.42) implies under the same conditions for k and n that

$$\frac{1}{k}\sum_{j=1}^{n} \epsilon_{(\mathbf{X}_j/\mathbf{b}(\frac{n}{k}))} \Rightarrow \mu.$$

We adjust the tails for their respective and possibly different tail indexes. From (2.41) we have that

$$n\mathbb{P}\left[\left(\frac{X_1^{(i)}}{b^{(i)}(\frac{n}{k})}\right)^{\alpha_i} \in \cdot\right] \xrightarrow{\nu} \mu_1.$$

Remembering that operations are carried out componentwise, we obtain

$$\frac{1}{k}\sum_{j=1}^{n} \epsilon_{\left(\mathbf{X}_j/\mathbf{b}(\frac{n}{k})\right)^{\boldsymbol{\alpha}}} \Rightarrow \mu_*. \tag{2.51}$$

Suppose that we had consistent estimators of $\boldsymbol{\alpha}$ and $\mathbf{b}(\frac{n}{k})$, denoted by $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\mathbf{b}}(\frac{n}{k})$. de Haan and Resnick (1993) showed that using these estimates, we have that

$$\widehat{\mu}_* := \frac{1}{k}\sum_{j=1}^{n} \epsilon_{\left(\mathbf{X}_j/\widehat{\mathbf{b}}(\frac{n}{k})\right)^{\widehat{\alpha}}}. \tag{2.52}$$

is a consistent estimator of μ_*.

In practice we have to:

1. Choose an appropriate k. We use the Stărică plot, see next section below.

2. Consistently estimate $\mathbf{b}(n/k)$. Since

$$b^{(i)}(n/k)/X^{(i)}_{(k+1)} \xrightarrow{P} 1,$$

see de Haan and Resnick (1993), we use the $(k+1)$st order statistic $\widehat{b^{(i)}}(n/k) = X^{(i)}_{(k+1)}$ as an estimator of $b^{(i)}(n/k)$.

3. Consistently estimate the tail indexes α_i. We use the Hill estimator, introduced in section $2.1.5$ for that purpose.

Proceeding in a similar fashion as with the ranks method, we obtain an estimator of the spectral measure by using a transformation into polar coordinates. Using the transformation to polar coordinates as above, namely $T(\mathbf{x}) := (\|\mathbf{x}\|, \|\mathbf{x}\|^{-1}\mathbf{x}) =: (R, \theta)$ and defining

$$(\mathcal{R}_{j,k}, \psi_{j,k}) := T\left(\left(\frac{\mathbf{X}_j}{\widehat{\mathbf{b}}(n/k)}\right)^{\widehat{\boldsymbol{\alpha}}}\right),$$

we have that

$$\widehat{S_{k,n}}(\cdot) := \frac{\sum_{j=1}^{n} \mathbf{1}_{(\mathcal{R}_{j,k}>1)} \epsilon_{(\psi_{j,k})}(\cdot)}{\sum_{j=1}^{n} \mathbf{1}_{\mathcal{R}_{i,k}}((1,\infty])} \tag{2.53}$$

estimates $S_*(\cdot)$ consistently, see de Haan and Resnick (1993). Essentially, a point X_j is considered extreme in the sense that it is used in the estimation of $S_*(\cdot)$, if the corresponding $\mathcal{R}_{j,k} > 1$.

Choosing k: The Stărică Plot

Both methods of estimating the spectral measure described above depend on choosing a k. The following idea, due to Stărică (1999), uses the scaling property of the exponent measure

$$t\mu_*(t\cdot) = \mu_*(\cdot) \tag{2.54}$$

to make a choice for k. Suppose that we have an estimator $\widehat{\mu}_* := \widehat{\mu}_{*,k,n}$ of μ_*. We use it to plot

$$\left\{ \frac{t\widehat{\mu}_*(t\mathcal{A})}{\widehat{\mu}_*(\mathcal{A})}, \text{ for } t \text{ in a neighborhood of } 1 \right\},$$

where $\mathcal{A} = \{\mathbf{x} \in \mathbb{E} : \|x\| > 1\}$. If k was chosen appropriately, $\widehat{\mu}_*$ will be a meaningful estimator of μ_* and the plot should be close to a horizontal line at 1. Different choices of k will result in different plots. We choose the k that results in a plot that most closely resembles the horizontal line at 1. Using the ranks method to as an estimator for μ_*, we obtain

$$\frac{t\widehat{\mu}_*(t\mathcal{A})}{\widehat{\mu}_*(\mathcal{A})} = \frac{t\sum_{j=1}^{n} \epsilon_{(\frac{k}{\mathbf{r_j}})}(t\mathcal{A})}{\sum_{j=1}^{n} \epsilon_{(\frac{k}{\mathbf{r_j}})}(\mathcal{A})} = \frac{t\sum_{j=1}^{n} \mathbf{1}_{(R_{j,k}>t)}}{\sum_{j=1}^{n} \mathbf{1}_{(R_{j,k}>1)}}, \tag{2.55}$$

where $\{R_{j,k}, j = 1, ..., n\}$ are the radial components of the polar coordinate representation of $\{\frac{k}{\mathbf{r_j}}, j = 1, ..., n;\}$.

Alternatively, we could also use equation (2.52) as an estimator for μ_*. In that case we plot

$$\frac{t\widehat{\mu}_*(t\mathcal{A})}{\widehat{\mu}_*(\mathcal{A})} = \frac{t\sum_{j=1}^{n} \epsilon_{\left(\mathbf{X}_j/\widehat{\mathbf{b}}(\frac{n}{k})\right)^{\widehat{\alpha}}}(t\mathcal{A})}{\sum_{j=1}^{n} \epsilon_{\left(\mathbf{X}_j/\widehat{\mathbf{b}}(\frac{n}{k})\right)^{\widehat{\alpha}}}(\mathcal{A})} = \frac{t\sum_{j=1}^{n} \mathbf{1}_{(\bar{R}_{j,k}>t)}}{\sum_{j=1}^{n} \mathbf{1}_{(\bar{R}_{j,k}>1)}}, \tag{2.56}$$

where $\{\bar{R}_{j,k}, j = 1, ..., n\}$ are the radial components of the polar coordinate representations of $\left(\mathbf{X}_j/\widehat{\mathbf{b}}(\frac{n}{k})\right)^{\widehat{\alpha}}$. We use the first $(k+1)^{st}$ order statistics of $\mathbf{X_j}$ and the Hill estimator for $\widehat{\mathbf{b}}(\frac{n}{k})$ and $\widehat{\alpha}$ respectively, for reasons explained above.

2.2.4 The Spectral Measure for non-positive Data

So far, we have only considered tail dependence for positive data. We explained how we describe the tail dependence of positive random vectors, with range $\mathbb{E} = [\mathbf{0}, \infty] \setminus \{\mathbf{0}\}$, with the spectral measure. We also introduced two different methods for estimating the spectral measure from data. However, in a number of applications one has to work with data that contains positive as well as negative observations. It is one of the goals of this thesis to describe the tail dependence between the log-returns of different stocks. It may be of interest to learn about the structure of the tail dependence of a bivariate distribution in all four quadrants and not just the first quadrant only. We may for example be interested in the tail dependence between large negative returns between two stocks. This poses the problem of how to define and estimate the spectral measure for data that is not non-negative.

Assume that the random variables $X^{(1)}$ and $X^{(2)}$ describe the log returns of two stocks respectively. If it is our intention to only focus on the tail dependence between large positive returns of the stocks, we do not need to consider the negative returns. We consider only the observations for which both stocks have a non-negative return. This way, we obtain a dataset of only non-negative observations. This allows us describe the tail dependence with the spectral measure. Consequently, we can use the techniques explained earlier in this chapter.

To study, say, the dependence between large negative returns, we can proceed in a similar way. We only consider observations for which both stocks have a non-positive return. We hence discard all observations for which either $X^{(1)}$ or $X^{(2)}$ is positive. This way we obtain a dataset consisting of only non-positive observations. By considering the absolute values of these observations, we obtain a non-negative dataset. This way, we can again make use of the concept of the spectral measure to describe the tail dependence. In a similar fashion, we can study the tail dependence between $X^{(1)}$ and $-X^{(2)}$ or $-X^{(1)}$ and $X^{(2)}$. Obviously, this solution is not limited to the two dimensional case and an extension to higher dimensions is straight forward, even though the number of different cases to be considered grows exponentially with the dimension d.

However, this approach is not satisfying. We would like to be able to describe the entire tail dependence of the considered random variables with the concept of the spectral measure. With the approach outlined above, we are only describing the tail dependence in certain quadrants by separate spectral measures. We need to define the spectral measure of the entire distribution. The definition has to be consistent with the definition of the spectral measure given earlier in this thesis. We then need to describe how we estimate this spectral measure. The following definition introduces the notion of a spectral measure for a distribution with both positive and negative observations.

Definition 2.2.6 *The distribution of a random vector* $\mathbf{X}$ *is called "multivariate regular varying" with tail index* α *and spectral measure S, if the following limit exists for all* $x > 0$*:*

$$\frac{\mathbb{P}[\|\mathbf{X}\| > tx, \mathbf{X}/\|\mathbf{X}\| \in \cdot]}{\mathbb{P}[\|\mathbf{X}\| > t]} \longrightarrow^{\nu}_{t\to\infty} x^{-\alpha} S(\cdot), \tag{2.57}$$

where $\longrightarrow^{\nu}$ *denotes vague convergence on* $\mathbb{S}^{d-1}$*, the d dimensional unit sphere.*

The definition is consistent with the definition that we gave earlier for the spectral measure of positive data. Recall that in Theorem 2.2.3 we had stated that

$$F_* \in D(G_*) \tag{2.58}$$

if and only if for $(\mathbf{R}, \Theta) := (\|\mathbf{X}_1\|, \|\mathbf{X}_1\|^{-1}\mathbf{X_1})$ we have that

$$t\mathbb{P}[(\frac{\mathbf{R}}{t}, \Theta) \in \cdot] \xrightarrow{\nu} r^{-2}dr \times S_*(d\theta). \tag{2.59}$$

In this framework, Definition 2.2.6 naturally extends the spectral measure as a tool to describe the tail dependence onto the entire unit sphere.

However, (2.57) assumes that the tail indexes of all marginal distributions are the same, namely α. It also assumes that the tail indexes of the left and the right tail of each marginal distribution are equal. Clearly, this is not a reasonable assumption. In the context of Theorem 2.2.3, we did not assume that the actual distribution F of the data satisfies the regular variation condition (2.59). Instead, we assume that the distribution of the data satisfied two conditions, called the “Marginal Condition” (2.41) and the “Global Condition” (2.42). Proposition 2.2.4 states, that if (2.41) and (2.42) are met, then there exists a transformation of the data, such that (2.57) holds for the transformed data. We need to adapt these results for the case of data that is not non-negative.

Suppose, that we have a random vector $\mathbf{X} = (X^{(1)}, ..., X^{(d)}) \in \mathbb{R}^d$. Define the random vector $\mathbf{Z} \in \mathbb{R}_+^{2d}$ as a function T of the random vector $\mathbf{X}$, as follows:

$$\begin{aligned} \mathbf{Z} \quad &= (Z^{(1)}, ..., Z^{(2d)}) = T(\mathbf{X}) = T((X^{(1)}, ..., X^{(d)})) \in \mathbb{R}_+^{2d}; \text{ where} \\ Z^{(2i-1)} &= X_+^{(i)} := \max(\mathbf{X}^{(i)}, 0), i = 1, ..., d \text{ and} \\ Z^{(2i)} &= X_-^{(i)} := \max(-\mathbf{X}^{(i)}, 0), i = 1, ..., d. \end{aligned} \tag{2.60}$$

The random vector $\mathbf{Z}$ is a non-negative random vector. We can therefore apply the results from Section 2.2.2. This motivates the following definition:

Definition 2.2.7 *We say that a random vector $\mathbf{X} \in \mathbb{R}^d$ is jointly regular varying, if the random vector $\mathbf{Z} = T(\mathbf{X})$, defined by (2.60), satisfies the “Marginal Condition” (2.41) and the “Global Condition” (2.42).*

It follows from Proposition 2.2.4 that $\mathbf{Z}$ has standard global convergence (2.44) and standard marginal convergence (2.46). Therefore it has a spectral measure. Due to the special nature of the random vector $\mathbf{Z}$, the spectral measure of $\mathbf{Z}$ can be translated into a spectral measure describing the tail dependence of the random vector $\mathbf{X}$.

Definition 2.2.8 *The spectral measure $S_{\mathbf{X}}$ of a jointly regular varying random vector $\mathbf{X} \in \mathbb{R}^d$ is defined as the map of the spectral measure $S_{\mathbf{Z}}$ of $\mathbf{Z} = T(\mathbf{X})$ under T. That is, we define*

$$S_{\mathbf{X}}(\cdot) = S_{\mathbf{Z}}(T(\cdot)). \tag{2.61}$$

To illustrate this definition, consider for simplicity the case $d = 2$. Assume, that we wish to study the spectral measure of the random vector $\mathbf{X} = (X^{(1)}, X^{(2)}) \in \mathbb{R}^2$. We have that

$$\mathbf{Z} = (Z^{(1)}, ..., Z^{(4)}) = ((X^{(1)})_+, (X^{(1)})_-, (X^{(2)})_+, (X^{(2)})_-).$$

The spectral measure of $\mathbf{Z}$ is a measure on the 4 dimensional unit sphere. However, the way we defined $\mathbf{Z}$, we have that either $Z^{(1)} = 0$ or $Z^{(2)} = 0$. At the same time, we have that either $Z^{(3)} = 0$ or $Z^{(4)} = 0$. In either case, at least two entries of the vector $\mathbf{Z}$ equal 0. The distribution of $\mathbf{Z}$ concentrates on 2 dimensional sub-planes of $\mathbb{R}^4$. Each of those planes corresponds to a quadrant in $\mathbb{R}^2$:

- If $\mathbf{Z}$ lies in the sub-plane $Z^{(2)} = 0$ and $Z^{(4)} = 0$, then $\mathbf{Z}$ corresponds to the point $(X^{(1)}, X^{(2)})$ with both $X^{(1)} \geq 0$ and $X^{(2)} \geq 0$.
- If $\mathbf{Z}$ lies in the sub-plane $Z^{(1)} = 0$ and $Z^{(4)} = 0$, then $\mathbf{Z}$ corresponds to the point $(X^{(1)}, X^{(2)})$ with $X^{(1)} \leq 0$ and $X^{(2)} \geq 0$.
- If $\mathbf{Z}$ lies in the sub-plane $Z^{(1)} = 0$ and $Z^{(3)} = 0$, then $\mathbf{Z}$ corresponds to the point $(X^{(1)}, X^{(2)})$ with both $X^{(1)} \leq 0$ and $X^{(2)} \leq 0$.
- If $\mathbf{Z}$ lies in the sub-plane $Z^{(2)} = 0$ and $Z^{(3)} = 0$, then $\mathbf{Z}$ corresponds to the point $(X^{(1)}, X^{(2)})$ with $X^{(1)} \geq 0$ and $X^{(2)} \leq 0$.

This way, the distribution of $\mathbf{Z}$ has a one to one relationship with the distribution of $\mathbf{X}$. The same relationship therefore also applies to the respective spectral measures. Since the distribution of $\mathbf{Z}$ is concentrated on 2 dimensional the sub-planes, the same is true for its spectral measure. The spectral measure in each sub-plane can hence be interpreted as the spectral measure of $\mathbf{X}$ in the corresponding quadrant. This idea is captured in Definition 2.2.6.

The estimation of the spectral measure of $\mathbf{X}$, based on a sample $\mathbf{X}_1, ..., \mathbf{X}_N$, follows naturally from the above definition of the spectral measure of $\mathbf{X}$. We obtain an estimate

$\widehat{S}_{\mathbf{Z}}$ of the spectral measure of the transformed sample $\mathbf{Z_i} = T(\mathbf{X_i}), i = 1, ..., N$, using the techniques described in Section 2.2.3. We then obtain the corresponding estimate of $\widehat{S}_{\mathbf{X}}$ from

$$\widehat{S}_{\mathbf{X}}(\cdot) = \widehat{S}_{\mathbf{Z}}(T(\cdot)). \qquad (2.62)$$

Chapter 3

Directional Distributions

Directional distributions model observations that are directions. The observations are usually recorded as points on the unit sphere. In the following, we will first concentrate on the relatively simple case of observations on the unit circle in $\mathbb{R}^2$, before describing the general case of the d dimensional unit sphere $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$. The problem of defining a distribution and its characteristics for distributions on a sphere is different from the problem of defining a distribution in the Euclidian space $\mathbb{R}^d$. The usual concepts of distributions in $\mathbb{R}^d$ are not appropriate, because the sphere has a very different topology. Consider the case of the unit circle in $\mathbb{R}^2$. If $\phi \rightarrow (cos(\phi), sin(\phi))$ is a parametrization of the unit circle, we know that the point $(cos(0), sin(0))$ is the same point as $(cos(2\pi), sin(2\pi))$. This periodicity is not present in the regular Euclidian space. This natural periodicity of the circle in particular and the sphere in general should be reflected in the description of distributions on the circle and the sphere. In the following, we will refer to distributions on the unit circle as circular distributions and distributions on the sphere $\mathbb{S}^{d-1}$, $d \geq 3$, as spherical distributions. The most common references on directional distributions are Jupp and Mardia (2000) and Mardia (1972).

3.1 Circular Distributions

3.1.1 Definitions and Descriptive Measures

Throughout our work we use the following parameterizations of the unit circle: $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 : x = cos(\phi), y = sin(\phi), \phi \in [0, 2\pi)\}$. For the discussion of certain properties it is more convenient to consider the complex unit circle, rather than the real unit circle. This allows the representation of a circular random variable X as $X = e^{i\Phi}$, $\Phi \in [0, 2\pi)$. Let $X = e^{i\Phi}$ be a random variable with values on the unit circle in $\mathbb{R}^2$. In a slight abuse of notation, we will refer to the random variable $X = e^{i\Phi}$ with both X and Φ, depending on which notation is more convenient.

Definition 3.1.1 *A function F with domain $\mathbb{R}$ is called circular cumulative distribution function (cdf) of a circular random variable $X = e^{i\Phi}$, if the following equations hold:*

1. $F(\varphi) = \mathbb{P}[0 \leq \Phi \leq \varphi], 0 < \varphi \leq 2\pi$

2. $F(\varphi + 2\pi) - F(\varphi) = 1, \forall\varphi \in \mathbb{R}$

The first property is similar to the definition of a cdf of a random variable on the real line. It implies $F(0) = 0$, unless there is a atom at 0, and $F(2\pi) = 1$. The second property describes how to extend the domain of F to the real line. In a similar fashion, we can define the density of an absolute continuous random variable.

Definition 3.1.2 *A non-negative function f with domain $\mathbb{R}$ is called probability density function (pdf) of $X = e^{i\Phi}$, if the following equation holds for a circular cdf F:*

$$F(\varphi) = \int_0^{\phi} f(\phi) d\phi, 0 < \phi \leq 2\pi$$

The two definitions imply the following properties for a density of a circular random variable:

1. $f(\phi + 2\pi) = f(\phi), 0 < \phi \leq 2\pi$ a.s.

2. $\int_0^{2\pi} f(\phi) d\phi = 1$

Conversely, any positive function $f(\phi)$ that satisfies the two properties above is a density function for a circular distribution.

We now give a definition of the characteristic function. We use the theory of Fourier series for periodic function, that implies that, in order to characterize a distribution, it is enough to consider integer values for p.

Definition 3.1.3 *Let $X = e^{i\Phi}$ be a circular random variable. Then the function*

$$\Psi_p := \Psi(p) = \mathbb{E}[X^p] = \mathbb{E}[e^{ip\Phi}] = \int_0^{2\pi} e^{ip\phi} F(d\phi), p \in \mathbb{Z} \tag{3.1}$$

is called the characteristic function (ch.f.) of X.

We write

$$\Psi_p = a_p + ib_p = \rho_p e^{i\alpha_p^0}, \tag{3.2}$$

where

$$a_p = \mathbb{E}[cos(p\Phi)] = \int_0^{2\pi} cos(p\phi)F(d\phi) \tag{3.3}$$

and

$$b_p = \mathbb{E}[sin(p\Phi)] = \int_0^{2\pi} sin(p\phi)F(d\phi). \tag{3.4}$$

The sequences a_p and b_p are referred to as the trigonometric moments of $X = e^{i\Phi}$. For the special case p=1 we use the notations $\rho_1 = \rho$ and $\alpha_1^0 = \alpha_0$. The key property of the ch.f. of circular distribution is that such distribution are determined by their ch.f., see Jupp and Mardia (2000).

Definition 3.1.4 Ψ_1 *is called the resultant.* α_0 *is called the mean direction of* $X = e^{i\Phi}$, *while* ρ *is called the resultant length.*

The mean direction takes the role that the mean has for a distribution on the line. One can show that the mean direction is the solution to the equations

$$\begin{aligned} \mathbb{E}[\sin(\Phi - \alpha_0)] &= 0, \alpha_0 \in [0, 2\pi) && (3.5) \\ \mathbb{E}[\cos(\Phi - \alpha_0)] &> 0. && (3.6) \end{aligned}$$

The resultant length is then given by

$$\rho = \mathbb{E}[\cos(\Phi - \alpha_0)]. \tag{3.7}$$

See Mardia (1972) for a reference. It is important to point out that the mean direction is only well defined if $\rho > 0$. The Lattice distribution and the Uniform distribution are examples of circular distributions for which resultant length is 0. The reason why we are not considering the usual mean is illustrated in the following example.

Example 3.1.5 *Let $X = e^{i\Phi}$ be concentrated on two points, $\pi/100$ and $199/100\pi$, each attained with probability 0.5. The mean, as calculated for a distribution on the line, would be π. Note that both values of Φ are close to 0. Obviously, a mean of π is not what we expect intuitively in this case. On the other hand, we have that $\Psi_1 = cos(\pi/100) \approx 0.9995$. Hence the resultant length is $\rho = cos(\pi/100)$ and the mean direction is $\alpha_0 = 0$.*

Now consider a change in the coordinate system, making the direction $\nu = -\pi/50$ the new zero direction. In the new coordinate system Φ has values $3/100\pi$ and $1/100\pi$. Therefore the mean, as calculated for a distribution on the line is now $1/50\pi$. On the other hand, we have $\Psi_1^{'} = cos(\pi/100)e^{i1/50\pi}$. The mean direction is therefore also $1/50\pi$. We see that the new mean direction α_0' satisfies $\alpha_0^{'} = \alpha_0 - \nu$. If we choose a new zero direction, we cannot expect the direction of the mean as calculated on the line to change by the angle between the new and old zero direction. For this reason, the new definition of a mean direction is needed.

The resultant length is used to define the circular variance, a measure of dispersion.

Definition 3.1.6 *Let X be a circular random variable with resultant length ρ. The circular variance of X, V_0, is defined as $V_0 = 1 - \rho = 1 - \mathbb{E}[cos(\Phi - \alpha_0)] \in [0, 1]$, using (3.7).*

Note that V_0 is invariant under changes of the zero direction. This is not true for the variance as calculated for a distribution on the line. This is illustrated by the following example.

Example 3.1.7 *Let $X = e^{i\Phi}$ be as in Example 3.1.5. The variance as calculated on the real line is 9.67. On the other hand, V_0 is equal to 0.0005. Note that since the distribution is concentrated on two points that are close together, the large value of the variance as calculated on the real line is not meaningful.*

Now consider again the change in the coordinate system, making the direction $\nu = -\pi/50$ the new zero direction. In the new coordinate system Φ has a different mean as calculated on the line and hence also a new variance, which now is 0.001. On the other hand, the length of the resultant and therefore the circular variance V_0 do not change. For this reason the new definition of a variance is needed.

Let $x_1 = e^{j\phi_1}, \ldots, x_n = e^{j\phi_n}$ be an i.i.d. sample of a circular random variable $X = e^{i\Phi}$. The sample trigonometric moments

$$\overline{C_p} = \tfrac{1}{n} \textstyle\sum_{j=1}^{n} cos(p\phi_j)$$
$$\overline{S_p} = \tfrac{1}{n} \textstyle\sum_{j=1}^{n} sin(p\phi_j)$$

are unbiased estimators of the trigonometric moments. Of particular interest are $\overline{C_1} =: \overline{C}$ and $\overline{S_1} =: \overline{S}$, as they are used to estimate the resultant length and the mean direction. The resultant length is estimated by the mean resultant

$$\overline{R} = \left(\overline{C}^2 + \overline{S}^2\right)^{1/2}.$$

The mean direction is estimated by the sample mean direction. The sample mean direction is the solution $\overline{\alpha_0}$ of the following system of equations, whenever $\overline{R} > 0$:

$$\begin{aligned} \overline{C} &= \overline{R}cos(\overline{\alpha_0}) \\ \overline{S} &= \overline{R}sin(\overline{\alpha_0}) \end{aligned} \tag{3.8}$$

The sample mean direction is a consistent estimator of the mean direction, see Jupp and Mardia (2000). In general, it needs not be an unbiased estimator, but it is unbiased in the case of a von Mises distribution, as we will see below. We want to point out that the sample mean direction and the mean resultant length can also be expressed by $x_1, ..., x_n$ in cartesian coordinates. To that end, define the *sample mean vector* as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.9}$$

Then we have

$$\overline{R} = \|\overline{x}\|, \text{ and } \overline{\alpha_0} = \|\overline{x}\|^{-1}\overline{x}. \tag{3.10}$$

3.1.2 Important Circular Distributions

We now present several important circular models. Of particular interest are the Wrapped Normal distribution and the von Mises distribution. They can be seen as the analogues of the Normal distribution on the circle. Neither of them have all the important characterizations that the Normal distribution on the line incorporates. Some of those characterizations are held by the Wrapped Normal distribution, while others are held by the von Mises distribution. It turns out that these two distributions can be seen as approximations of each other. We may therefore use either one of them as the circular analogue of the Normal distribution on the line.

Point Distribution

$X = e^{i\Phi}$ is said to have a point distribution, if there is an $\alpha \in [0, 2\pi)$, such that:

$$\mathbb{P}[\Phi = \alpha] = 1$$

In that case α is also the mean direction, the resultant length is 1, the circular variance is 0 and the ch.f. is given by $\Psi_p = e^{ip\alpha}$.

Lattice Distribution

A lattice distribution is a discrete circular distribution concentrating its mass on a countable number of equally spaced points. It has probability function

$$\mathbb{P}\left(\Phi = \nu + \frac{2\pi r}{m}(mod\,2\pi)\right) = p_r, \qquad \text{for } r = 1, \ldots, m \text{ and } \nu \in (0, 2\pi], \tag{3.11}$$

where $p_r > 0$ are the probabilities of the points of support $\{\nu + \frac{2\pi r}{m}, r = 1, ..., m\}$ with $\sum_{r=0}^{m} p_r = 1$. The points of support have equal distances from their neighbors on the circle. They are the vertices of an m-sided regular polygon. A special case of the lattice distribution is called the discrete uniform distribution with m points of support. It is the lattice distribution with $p_r = 1/m$ for all $r = 1, ..., m$. The characteristic function of this uniform distribution is given by

$$\Psi_p = \begin{cases} 1, & p = 0 \text{ (mod m)} \\ 0, & \text{otherwise.} \end{cases}$$

In particular, we see from the ch.f. that, if $m \geq 2$, then the resultant length is 0. This means that the mean direction is not defined.

Uniform Distribution

If $X = e^{i\Phi}$ has pdf

$$f(\phi) = \frac{1}{2\pi}, 0 < \phi \leq 2\pi,$$

we say that $X = e^{i\Phi}$ is uniformly distributed on the circle. Note that the resultant length is 0. Therefore, the mean direction is not defined and the circular variance is 1. The ch.f. is $\Psi_p = (e^{ip2\pi} - 1)/2\pi ip$, $p \neq 0$. Therefore we have $\Psi_p = 1$, if $p = 0$ and $\Psi_p = 0$, if $p \neq 0$. The Uniform distribution appears as the limit distribution of sums of i.i.d circular random variables. Let $X_j = e^{i\Phi_j}, (j \in \mathbb{N})$ be an i.i.d. sequence of circular random variables. If the distribution of X_1 is not a lattice distribution, then $S_n =$

$\prod_{j=1}^{n} X_j = e^{i\sum_{j=1}^{n}\Phi_j}$ converges weakly to a uniformly distributed random variable. The summation of the random variables Φ_j is understood modulo 2π. See section 4.3.1 Jupp and Mardia (2000) for a proof. In particular, if X_1 is itself uniformly distributed on the unit circle, then the distribution of S_n is also the Uniform distribution, for all $n \in \mathbb{N}$.

Wrapped Normal Distribution

A random variable whose distribution has the characteristic function given by

$$\Psi_p = e^{i\alpha_0 p - p^2\sigma^2/2}, \tag{3.12}$$

is said to have a wrapped normal distribution, $WN(\alpha_0, \rho)$. It's trigonometric moments are given by

$$a_p = e^{-p^2\sigma^2/2}\cos(p\alpha_0) \text{ and } b_p = e^{-p^2\sigma^2/2}\sin(p\alpha_0). \tag{3.13}$$

The distribution is unimodal and symmetric about α_0. As $\rho \to 0$, it tends to the Uniform distribution, while, as $\rho \to 1$, it tends to the Point distribution at α_0. The pdf of $WN(\alpha_0, \rho)$ is given by

$$f(\phi; \alpha_0, \rho) = \frac{1}{\sigma(\rho)\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left[\frac{-(\phi - \alpha_0 + 2k\pi)^2}{2\sigma(\rho)^2}\right]. \tag{3.14}$$

The distribution has its name because of the following property. Let X have a normal distribution with mean μ and variance σ^2, $N(\mu, \sigma^2)$, on the real line. Then the circular random variable $X = e^{i\Phi}$ with $\Phi = X(\text{mod } 2\pi)$ has a wrapped normal distribution. Its mean direction is given by $\alpha_0 = u(mod\, 2\pi)$ and its resultant length ρ has the following relationship with σ:

$$\sigma(\rho)^2 = -2log(\rho) \Leftrightarrow \rho(\sigma) = e^{-\sigma^2/2}.$$

We refer to Jupp and Mardia (2000) as a reference.

On the line, we have that the sum of independent normally distributed random variables has again a normal distribution. Not surprisingly, this property transfers to a similar property for the wrapped normal distribution. If $X_j = e^{i\Phi_j}, j = 1, ..., n$, are independent and X_j has a $WN(\alpha_j, \rho_j)$ distribution, then we have that

$$\prod_{j=1}^{n} X_j = exp\left(i\sum_{j=1}^{n}\Phi_j\right) \text{ is distributed as } WN\left(\sum_{j=1}^{n}\alpha_j(mod\, 2\pi), \prod_{j=1}^{n}\rho_j\right). \quad (3.15)$$

Several other wrapped distributions have been considered. See Jupp and Mardia (2000) for definitions of a wrapped Poisson and a wrapped Cauchy distribution.

Von Mises Distribution

$X = e^{i\Phi}$ is said to have a von Mises distribution with parameters α and κ, $\mathcal{M}(\alpha, \kappa)$, if it has density

$$f_M(\phi; \alpha, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\phi-\alpha)}, 0 < \phi \leq 2\pi, \kappa > 0, 0 \leq \alpha < 2\pi. \quad (3.16)$$

where $I_0(\kappa)$ denotes the modified Bessel function of the first kind of order zero:

$$I_0(\kappa) = \sum_{n=0}^{\infty}\frac{1}{(n!)^2}\left(\frac{\kappa}{2}\right)^{2n} = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos(t)}dt. \quad (3.17)$$

α is the mean direction, as we will see below, while κ is a concentration parameter, but not the resultant length.

Note that the density of the von Mises distribution can also be expressed in cartesian coordinates. If we define $\boldsymbol{\mu} = (\mu_1, \mu_2) := (\cos(\alpha), \sin(\alpha))$ and $\mathbf{x} = (x_1, x_2) := (\cos(\phi), \sin(\phi))$, then we can rewrite (3.16) as

$$f_M(\mathbf{x}; \boldsymbol{\mu}, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa(\mu_1 x_1 + \mu_2 x_2)}; \boldsymbol{\mu}, \mathbf{x} \in \mathbb{S}^1, \kappa > 0. \quad (3.18)$$

For this reason, we sometimes also use the notation $\mathcal{M}(\boldsymbol{\mu}, \kappa)$ when referring to the von Mises distribution.

The density $f_M(\phi; \alpha, \kappa)$ is strictly positive for all $\phi \in [0, 2\pi]$, as long as the concentration parameter κ is finite. The distribution function of the von Mises distribution cannot be expressed in an easy closed form. We used numerical integration with Matlab to evaluate $F_M(\varphi; \alpha, \kappa) = \int_0^{\varphi} f_M(\phi; \alpha, \kappa) d\phi$. Alternatively, one could work with tables of values of $F_M(\varphi; 0, \kappa)$. Such tables can for example be found in Mardia (1972). The distribution is unimodal and symmetric about α. If $\kappa = 0$, then $f_M(\phi; \alpha, \kappa) = \frac{1}{2\pi}$, the pdf of the Uniform distribution. As $\kappa \to \infty, \mathbb{P}[\Phi \in [\alpha - \epsilon, \alpha + \epsilon]] \to 1$, so that the distribution converges to the point distribution at α. Figure 3.1 shows a plot of the density of the von Mises distribution. Note how the distribution is concentrated much closer around the mean direction for $\kappa = 10$ than it is for κ =2 or 0.2.

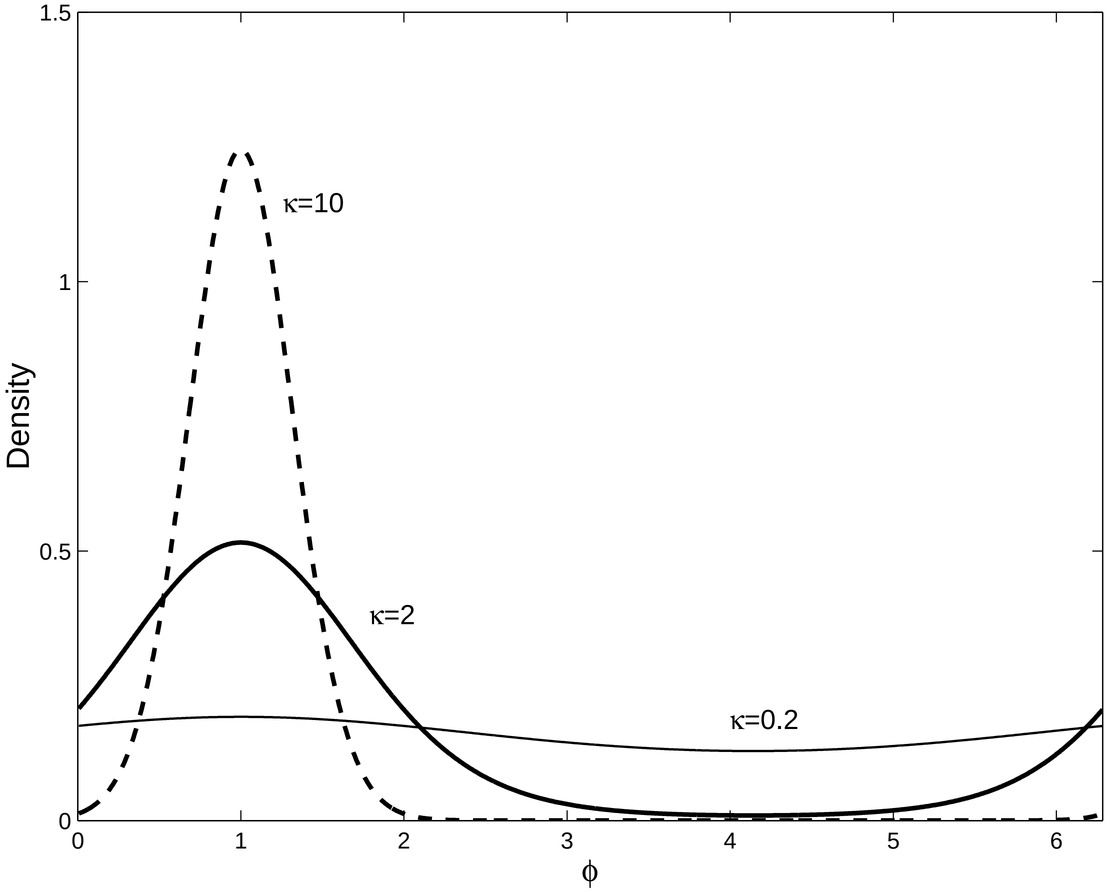

Figure 3.1: *The density $f_M(\phi; \alpha, \kappa)$ of the von Mises distributions with $\alpha = 1$ and $\kappa = 10, 2$ and 0.2, respectively.*

The ch.f. is given by

$$\Psi_p = e^{ip\alpha}\frac{I_p(\kappa)}{I_0(\kappa)}. \tag{3.19}$$

$I_p(\kappa)$ denotes the modified Bessel function of the first kind of order $p \in \mathbb{R}^+$, which is given by

$$I_p(\kappa) = \left(\frac{z}{2}\right)^p \sum_{j=0}^{\infty} \frac{(\kappa/2)^{2j}}{\Gamma(p+j+1)\Gamma(j+1)}, \tag{3.20}$$

where $\Gamma(x)$ denotes the Gamma function. For the purpose of calculating the ch.f. it is enough to consider $I_p(\kappa)$ only for integer values of p. However, for other purposes that we will discuss later in this thesis, we need to consider the function $I_p(\kappa)$ for non integer values of p. The following equation including an integral will prove to be helpful.

$$\begin{aligned} I_p(\kappa) &= \frac{(\frac{\kappa}{2})^p}{\sqrt{\pi}\Gamma(p+\frac{1}{2})}\int_{-1}^{1}(1-t^2)^{p-\frac{1}{2}}e^{\kappa t}dt \\ &= \frac{(\frac{\kappa}{2})^p}{\sqrt{\pi}\Gamma(p+\frac{1}{2})}\int_{0}^{1}(1-t^2)^{p-\frac{1}{2}}\left(e^{\kappa t}+e^{-\kappa t}\right)dt \end{aligned} \tag{3.21}$$

In case of $p \in \mathbb{N}$, we also have the following equation:

$$I_p(\kappa) = \frac{1}{\pi}\int_0^{\pi} e^{\kappa\cos(\phi)} cos(p\phi)d\phi.$$

As a consequence of this last equation we see that the trigonometric moments of order $p \in \mathbb{N}$ are

$$a_p = \frac{I_p(\kappa)}{I_0(\kappa)}cos(p\alpha) \text{ and } b_p = \frac{I_p(\kappa)}{I_0(\kappa)}sin(p\alpha). \tag{3.22}$$

In particular, we have for the resultant length that

$$\rho = A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}, \tag{3.23}$$

and that α is the mean direction. Hence the circular variance is $V_0 = 1 - A(\kappa)$.

The von Mises distribution can be related to other circular distributions. We already discussed the relations to the Uniform and the Point distributions. For large values of κ, one can furthermore show that the von Mises distribution $\mathcal{M}(\alpha, \kappa)$ can be approximated by a Wrapped Normal distribution with resultant length $\rho = A(\kappa)$. A result due to Kent (1978) states that

$$f_M(\phi;\alpha,\kappa) - f(\phi;\alpha,A(\kappa)) = O(\kappa^{-1/2}), \tag{3.24}$$

where $f_M(\phi;\alpha,\kappa)$ is the density of the von Mises distribution and $f(\phi;\alpha,A(\kappa))$ is the density of the approximating Wrapped Normal distribution. This could for example be used to obtain approximately $\mathcal{M}(\alpha,\kappa)$ distributed samples. A more sophisticated algorithm for simulating the von Mises distribution is given in Jupp and Mardia (2000).

While the sum of independent von Mises random variables is not a von Mises random variable again, it can be approximated by a von Mises random variable. One can show that as a consequence of the closeness of the von Mises and the Wrapped Normal distribution and (3.15), we have the following approximation. Assume that $X_1 = e^{i\Phi_1}$ is distributed as $\mathcal{M}(\alpha_1,\kappa_1)$ and that $X_2 = e^{i\Phi_2}$ is distributed as $\mathcal{M}(\alpha_2,\kappa_2)$. Then we have that $\Phi_1+\Phi_2$ is approximately distributed as $\mathcal{M}(\alpha_1+\alpha_2,\kappa_3)$ with $A(\kappa_3) = A(\kappa_1)A(\kappa_2)$. See again Mardia (2002) Jupp and Mardia (2000) for a proof.

The following two characterizations of the von Mises distribution are analogous to those of the Normal distribution on the line. We refer to Jupp and Mardia (2000) for a more detailed discussion.

The first characterization is the Maximum Entropy Characterization. The entropy of a distribution on the unit circle with pdf $f(\phi)$ is the defined as $-\int_0^{2\pi} f(\phi)\log f(\phi)d\phi$. The von Mises distribution has the maximum entropy of all distributions with given mean direction and circular variance. The Normal distribution maximizes the entropy on the line for fixed mean and variance.

Let $f(\phi-\alpha)$ be a pdf of a distribution on the circle belonging to a location family with varying mean direction α. If the maximum likelihood estimator of α is the sample mean direction, then f is the pdf of a von Mises distribution. Compare this to the situation on the real line: If $f(x-\alpha_0)$ is the pdf of a distribution on the line belonging to a location family, then f is pdf of the Normal distribution if and only if the maximum likelihood estimator of the mean α_0 is the sample mean.

Maximum Likelihood Estimation in a von Mises distribution Let $X_1,..,X_n$ be i.i.d., distributed as $\mathcal{M}(\alpha,\kappa)$. The corresponding log-likelihood function for the observations $\mathbf{x}=(x_1,...,x_n)$ of $X_1,..,X_n$ is

$$L(\alpha,\kappa;\mathbf{x}) = -n\log(2\pi) - n\log(I_0(\kappa)) + \kappa\sum_{i=1}^{n}\cos(x_i-\alpha). \tag{3.25}$$

It turns out that the MLE of the mean direction α can be determined without any knowledge about κ. We have

$$\begin{aligned}\frac{\partial L}{\partial\alpha} &= \kappa\sum_{i=1}^{n}\sin(x_i-\alpha) = \kappa\sum_{i=1}^{n}(\sin(x_i)\cos(\alpha)-\sin(\alpha)\cos(x_i)) \\ &= \kappa(S\cos(\alpha)-\sin(\alpha)C), \end{aligned}\tag{3.26}$$

where $C=\sum_{i=1}^{n}\cos(x_i)$ and $S=\sum_{i=1}^{n}\sin(x_i)$. The second derivative of the log-likelihood function is given by

$$\frac{\partial^2 L}{\partial\alpha^2} = -\kappa(S\sin(\alpha)+C\cos(\alpha)). \tag{3.27}$$

Let $R=\sqrt{(S^2+C^2)}$. Then by (3.26) and (3.27) the MLE $\widehat{\alpha}$ of α must satisfy

$$\begin{aligned} C &= R\cos(\widehat{\alpha}) \\ S &= R\sin(\widehat{\alpha}) \end{aligned}\tag{3.28}$$

The solution of (3.28) solves $\frac{\partial L}{\partial\alpha}=0$ and $\frac{\partial^2 L}{\partial\alpha^2}<0$, as long as $R>0$. Therefore, by comparing with (3.8), we see that $\widehat{\alpha}$ is just the mean direction.

Turning to the estimation of κ, we have:

$$\frac{\partial L}{\partial\kappa} = -n\frac{I_0'(\kappa)}{I_0(\kappa)} + \sum_{i=1}^{n}\cos(x_i-\alpha), \tag{3.29}$$

where $I_0'(\kappa)$ stands for the derivative of the $I_0(\kappa)$. Using the fact that $I_0'(\kappa) = I_1(\kappa)$ and recalling that $A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$, we therefore have:

$$\frac{\partial L}{\partial \kappa} = -nA(\kappa) + cos(\alpha)C + sin(\alpha)S. \tag{3.30}$$

Solving $\frac{\partial L}{\partial \kappa} = 0$ and replacing α with $\widehat{\alpha}$, we obtain the equation

$$-nA(\widehat{\kappa}) + \frac{C^2 + S^2}{R} = 0 \Leftrightarrow nA(\widehat{\kappa}) = R \Leftrightarrow A(\widehat{\kappa}) = \overline{R} \Leftrightarrow \widehat{\kappa} = A^{-1}(\overline{R}). \tag{3.31}$$

Thus the MLE of κ is well defined and unique, if the equation $A(\widehat{\kappa}) = \overline{R}$ has a unique solution for all $\overline{R} \in [0, 1)$. This is the case, if the function $A(z)$ has the following properties:

- $\lim_{z \to 0} A(z) = 0$,
- $\lim_{z \to \infty} A(z) = 1$,
- $A(z)$is strictly monotone increasing.

In the following section, we will consider a extension of the von Mises distribution to higher dimensions. We will need that a family of functions similar to $A(z)$ satisfies the three properties above. We therefore show that these three properties are not only satisfied by $A(z)$, but rather by a larger family of functions, referred to as $B_d(z)$. Note that $A(z) = B_1(z)$.

Proposition 3.1.8 *Let $d > 1$ be a real number. Define for $z > 0$*

$$B_d(z) := \frac{I_d(z)}{I_{d-1}(z)}$$

Then $B_d(z)$ has the following properties:

$$\lim_{z \to 0} B_d(z) = 0, \tag{3.32}$$

$$\lim_{z \to \infty} B_d(z) = 1, \tag{3.33}$$

$$B_d(z) \textit{ is a continuous, strictly monotone increasing function}. \tag{3.34}$$

Proof:

For the proof of (3.32) we make use of equation 9.6.7 in Abramowitz and Stegun (1972):

$$I_p(z) \sim \frac{(\frac{1}{2}z)^p}{\Gamma(p+1)}, \text{for fixed p and as } z \to 0.$$

Hence, we have immediately:

$$B_d(z) \sim \frac{1}{2}z\frac{\Gamma(d+1)}{\Gamma(d)}, \text{ as } z \to 0.$$

and therefore $\lim_{z\to 0} B_d(z) = 0$.

For the proof of (3.33), recall from (3.21), that

$$I_p(z) = \frac{(\frac{z}{2})^p}{\sqrt{\pi}\Gamma(p+\frac{1}{2})}\int_0^1 (1-t^2)^{p-\frac{1}{2}}\left(e^{zt}+e^{-zt}\right)dt.$$

Hence we can write

$$B_d(z) = \frac{(\frac{z}{2})^d\Gamma(d-\frac{1}{2})}{(\frac{z}{2})^{d-1}\Gamma(d+\frac{1}{2})}\frac{\int_0^1(1-t^2)^{d-\frac{1}{2}}\left(e^{zt}+e^{-zt}\right)dt}{\int_0^1(1-t^2)^{d-\frac{3}{2}}\left(e^{zt}+e^{-zt}\right)dt}.$$

Using that $\Gamma(d+\frac{1}{2}) = (d-\frac{1}{2})\Gamma(d-\frac{1}{2})$ this simplifies to

$$B_d(z) = \frac{(\frac{z}{2})}{(d-\frac{1}{2})}\frac{\int_0^1(1-t^2)^{d-\frac{1}{2}}\left(e^{zt}+e^{-zt}\right)dt}{\int_0^1(1-t^2)^{d-\frac{3}{2}}\left(e^{zt}+e^{-zt}\right)dt}.$$

We therefore need to show that, as $z \to \infty$,

$$\frac{\int_0^1(1-t^2)^{d-\frac{1}{2}}\left(e^{zt}+e^{-zt}\right)dt}{\int_0^1(1-t^2)^{d-\frac{3}{2}}\left(e^{zt}+e^{-zt}\right)dt} \sim \frac{2(d-\frac{1}{2})}{z}. \tag{3.35}$$

We have, as $z \to \infty$, that

$$\frac{\int_0^1(1-t^2)^{d-\frac{1}{2}}\left(e^{zt}+e^{-zt}\right)dt}{\int_0^1(1-t^2)^{d-\frac{3}{2}}\left(e^{zt}+e^{-zt}\right)dt} \sim \frac{\int_0^1(1-t^2)^{d-\frac{1}{2}}e^{zt}dt}{\int_0^1(1-t^2)^{d-\frac{3}{2}}e^{zt}dt}.$$

Furthermore, we have that $\forall\epsilon$, $\exists\delta = \delta(z,\epsilon)$, such that

$$(1-\delta)\left(\int_0^1(1-t^2)^{d-\frac{1}{2}}e^{zt}dt\right) = \int_{1-\epsilon}^1(1-t^2)^{d-\frac{1}{2}}e^{zt}dt. \tag{3.36}$$

As $z \to \infty$, we have $\forall \epsilon > 0$ that $\delta = \delta(z, \epsilon) \to 0$. Therefore, we get that

$$\frac{\int_0^1 (1-t^2)^{d-\frac{1}{2}} e^{zt} dt}{\int_0^1 (1-t^2)^{d-\frac{3}{2}} e^{zt} dt} \sim \frac{\int_{1-\epsilon}^1 (1-t^2)^{d-\frac{1}{2}} \left(e^{zt}\right) dt}{\int_{1-\epsilon}^1 (1-t^2)^{d-\frac{3}{2}} \left(e^{zt}\right) dt}.$$

as $z \to \infty$. A Taylor series approximation gives us $(1-t^2) \sim 2(1-t)$, as $t \to 1$. Therefore, we get

$$\frac{\int_{1-\epsilon}^1 (1-t^2)^{d-\frac{1}{2}} \left(e^{zt}\right) dt}{\int_{1-\epsilon}^1 (1-t^2)^{d-\frac{3}{2}} \left(e^{zt}\right) dt} \sim \frac{\int_{1-\epsilon}^1 2^{d-\frac{1}{2}} (1-t)^{d-\frac{1}{2}} \left(e^{zt}\right) dt}{\int_{1-\epsilon}^1 2^{d-\frac{3}{2}} (1-t)^{d-\frac{3}{2}} \left(e^{zt}\right) dt} = 2\frac{\int_{1-\epsilon}^1 (1-t)^{d-\frac{1}{2}} \left(e^{zt}\right) dt}{\int_{1-\epsilon}^1 (1-t)^{d-\frac{3}{2}} \left(e^{zt}\right) dt}.$$

With an argument analogue to (3.36) we get

$$2\frac{\int_{1-\epsilon}^1 (1-t)^{d-\frac{1}{2}} \left(e^{zt}\right) dt}{\int_{1-\epsilon}^1 (1-t)^{d-\frac{3}{2}} \left(e^{zt}\right) dt} \sim 2\frac{\int_0^1 (1-t)^{d-\frac{1}{2}} \left(e^{zt}\right) dt}{\int_0^1 (1-t)^{d-\frac{3}{2}} \left(e^{zt}\right) dt}.$$

We multiply the integrand in both the numerator and the denominator by the constant term e^{-z} and then use the change of variable $x = (1-t)$ to get

$$2\frac{\int_0^1 (1-t)^{d-\frac{1}{2}} e^{zt} dt}{\int_0^1 (1-t)^{d-\frac{3}{2}} e^{zt} dt} = 2\frac{\int_0^1 (1-t)^{d-\frac{1}{2}} e^{-z(1-t)} dt}{\int_0^1 (1-t)^{d-\frac{3}{2}} e^{-z(1-t)} dt} = 2\frac{\int_0^1 x^{d-\frac{1}{2}} e^{-zx} dx}{\int_0^1 x^{d-\frac{3}{2}} e^{-zx} dx}.$$

A second change of variable $y = xz$ gives us

$$2\frac{\int_0^1 x^{d-\frac{1}{2}} e^{-zx} dx}{\int_0^1 x^{d-\frac{3}{2}} e^{-zx} dx} = 2\frac{\int_0^1 (\frac{y}{z})^{d-\frac{1}{2}} e^{-y} \frac{1}{z} dy}{\int_0^1 (\frac{y}{z})^{d-\frac{3}{2}} e^{-y} \frac{1}{z} dy} = \frac{2}{z} \frac{\int_0^z y^{d-\frac{1}{2}} e^{-y} dt}{\int_0^z y^{d-\frac{3}{2}} e^{-y} dy}.$$

Remembering that

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,$$

we observe that

$$\frac{2}{z} \frac{\int_0^z y^{d-\frac{1}{2}} e^{-y} dt}{\int_0^z y^{d-\frac{3}{2}} e^{-y} dy} \sim \frac{2}{z} \frac{\int_0^\infty y^{d-\frac{1}{2}} e^{-y} dt}{\int_0^\infty y^{d-\frac{3}{2}} e^{-y} dy} \sim \frac{2}{z} \frac{\Gamma(d+\frac{1}{2})}{\Gamma(d-\frac{1}{2})} = \frac{2}{z} \left(d - \frac{1}{2}\right).$$

Together we have therefore shown that, as $z \to \infty$,

$$\frac{\int_0^1 (1-t^2)^{d-\frac{1}{2}} \left(e^{zt} + e^{-zt}\right) dt}{\int_0^1 (1-t^2)^{d-\frac{3}{2}} \left(e^{zt} + e^{-zt}\right) dt} \sim \frac{2}{z} \left(d - \frac{1}{2}\right).$$

This is exactly (3.35). This shows that $B_d(z) \to 1$, as $z \to \infty$.

To show that $B_d(z)$ is a strictly monotone increasing function, we first observe that

$$B_d(z) = c_1 \cdot z \frac{\int_{-1}^{1} (1-t^2)^{d-\frac{1}{2}} e^{zt} dt}{\int_{-1}^{1} (1-t^2)^{d-\frac{3}{2}} e^{zt} dt},$$

where c_1 is a constant. Using integration by parts, we obtain for the integral in the numerator:

$$(1-t^2)^{d-\frac{1}{2}} e^{zt} \,\Big|_{-1}^{1} + \int_{-1}^{1} \left(d - \frac{1}{2}\right) 2t(1-t^2)^{d-\frac{3}{2}} e^{zt} dt = c_2 \cdot \int_{-1}^{1} t(1-t^2)^{d-\frac{3}{2}} e^{zt} dt,$$

where c_2 stand for a constant. Define

$$f(t) := (1-t^2)^{d-\frac{3}{2}}.$$

We can now rewrite $B_d(z)$ as

$$B_d(z) = c_3 \cdot z \frac{\int_{-1}^{1} t f(t) e^{zt} dt}{\int_{-1}^{1} f(t) e^{zt} dt}$$

for a constant c_3. To show that the right hand side is strictly monotone increasing, we note that

$$\frac{\int_{-1}^{1} (-1+t+1) f(t) e^{zt} dt}{\int_{-1}^{1} f(t) e^{zt} dt} = -1 + \frac{\int_{-1}^{1} (t+1) f(t) e^{zt} dt}{\int_{-1}^{1} f(t) e^{zt} dt}.$$

The fraction on the right hand side can be written as

$$\frac{\int_{-1}^{1} f(t) e^{zt} dt \int_{-1}^{t} dx}{\int_{-1}^{1} f(t) e^{zt} dt} = \frac{\int_{-1}^{1} dx \int_{x}^{1} f(t) e^{zt} dt}{\int_{-1}^{1} f(t) e^{zt} dt}.$$

Hence, we can write $B_d(z)$ as

$$B_d(z) = c_3 \cdot z \left(-1 + \frac{\int_{-1}^{1} dx \int_{x}^{1} f(t) e^{zt} dt}{\int_{-1}^{1} f(t) e^{zt} dt} \right).$$

In order to show that $B_d(z)$ is strictly monotone increasing, it is enough to show that the function

$$f_2(z) = \frac{\int_{x}^{1} f(t) e^{zt} dt}{\int_{-1}^{1} f(t) e^{zt} dt}$$

is non-decreasing. We can split the integral in the denominator to obtain

$$f_2(z) = \frac{\int_x^1 f(t)e^{zt}dt}{\int_{-1}^x f(t)e^{zt}dt + \int_x^1 f(t)e^{zt}dt} = \left(\frac{\int_{-1}^x f(t)e^{zt}dt}{\int_x^1 f(t)e^{zt}dt} + 1 \right)^{-1} .$$

$f_2(z)$ is non-decreasing, if and only if

$$\frac{\int_{-1}^x f(t)e^{zt}dt}{\int_x^1 f(t)e^{zt}dt} = \frac{\int_{-1}^x f(t)e^{z(t-x)}dt}{\int_x^1 f(t)e^{z(t-x)}dt} \tag{3.37}$$

is non-increasing. Note that, for the integrand in the numerator, we have $t \leq x$. Therefore, $e^{z(t-x)}$ is a non-increasing function of z, since $(t - x) \leq 0$. As a consequence the numerator is a decreasing function of z. On the other hand, for the numerator, we have $t \geq x$ and hence $e^{z(t-x)}$ is an increasing function of z. Therefore, the denominator is a increasing function of z. Hence, we see that the right hand side in (3.37) is a decreasing function of z. This is turn implies that $f_2(z)$ is non-decreasing and hence $B_d(z)$ is strictly monotone increasing. Finally, the fact that $I_d(z)$ is a continuous, positive function for $z > 0$ and $d \geq 0$ implies that $B_d(z)$ is a continuous function on $z > 0$. Equation (3.32) proves continuity at 0. ■

Unfortunately, there is no explicit, closed form equation for the evaluation of $A^{-1}(\cdot)$. We used a numerical procedure implemented in Matlab to evaluate $A^{-1}(\cdot)$. Alternatively, one could use tables or polynomial approximations to evaluate $A^{-1}(\cdot)$. Suitable approximations can for example be found in Jupp and Mardia (2000). Further approximations could also be based on corresponding approximations for $I_d(z)$ found in sections 9.7 and 9.8 in Abramowitz and Stegun (1972).

3.1.3 Distributions on $(0, 2\pi/k)$

There are instances where one needs a circular random variable whose range is only a part of the unit circle. That is, we are interested in developing models for random variables $X^* = e^{i\Phi^*}$ with $\Phi^* \in [0, 2\pi/l)$ for a real number l. Typically l is an integer. For example, if one is attempts to model the angle, with which objects like asteroids enter the earth's atmosphere, one would want to work with a random variable $X^* = e^{i\Phi^*}$ with $\Phi^* \in [0, \pi/2)$. Such random variables are derived from circular random variables $X = e^{i\Phi}$, with $\Phi \in [0, 2\pi)$ by setting

$$\Phi^* = \Phi/l \Longleftrightarrow X^* = X^{1/l}. \tag{3.38}$$

This allows us to adapt any circular model to describe random variables whose angle Φ only has values in $([0, 2\pi/l)$. If $X = e^{i\Phi}, \Phi \in [0, 2\pi)$ has density $f(\phi)$, then

$$f^*(\phi) = f(\phi \cdot l) \cdot l, \phi \in (0, 2\pi/l) \tag{3.39}$$

is the density of $X^* = e^{i\Phi^*}$. Following this idea, we may define the ch.f. of $X^* = e^{i\Phi^*}$ as $\Psi_p = \mathbb{E}[e^{ilp\Phi^*}]$. As a consequence, it seems natural to define the mean direction of $X^* = e^{i\Phi^*}$ as $\alpha_0^* = \alpha_0/l$, where α_0 is the mean direction of $X = e^{i\Phi}$. It is not easy to obtain an appropriate definition of the circular variance of $X^* = e^{i\Phi^*}$ from the corresponding definition of V_0, the circular variance of $X = e^{i\Phi}$. See Section 3.5.2 Mardia (1972) for a discussion. They suggest that the variance V_0^* of $X^* = e^{i\Phi^*}$ be defined as

$$V_0^* = 1 - (1 - V_0)^{1/l^2}. \tag{3.40}$$

3.2 Spherical Distributions

3.2.1 Definitions and Descriptive Measures

Let $\{\Omega, (A), \mathbb{P}\}$ be a probability space. We say that the random variable X has a spherical distribution, if

$$X(\omega) \in \mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}, \forall \omega \in \Omega.$$

Due to the special topology of $\mathbb{S}^{d-1}$, the concept of a cdf is not widely used in the description of spherical distributions. It is customary to describe such distributions either using density functions or probability functions, depending on whether the distribution is absolute continuous or discrete. The definition of a pdf for a spherical distribution is as follows.

Definition 3.2.1 *A nonnegative function g with domain $\mathbb{S}^{d-1}$ is called a probability density function, pdf, of a spherical distribution, if*

$$\int_{\mathbb{S}^{d-1}} g(\mathbf{x}) do(\mathbf{x}) = 1, \tag{3.41}$$

where $do(\mathbf{x})$ denotes the surface measure on $\mathbb{S}^{d-1}$. That is, $do(\mathbf{x})$ is the Lebesgue measure restricted to $\mathbb{S}^{d-1}$, satisfying

$$\int_{\mathbb{S}^{d-1}} do(\mathbf{x}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}.$$

It is sometimes more convenient to express a pdf in spherical coordinates. The representation of a point $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ in spherical coordinates is as follows:

$$\begin{aligned}
x_1 &= r\cos(\phi)\prod_{i=1}^{d-2}\sin(\theta_i), && (3.42)\\
x_2 &= r\sin(\phi)\prod_{i=1}^{d-2}\sin(\theta_i), && (3.43)\\
x_j &= r\cos(\theta_{j-2})\prod_{i=j-1}^{d-2}\sin(\theta_i),\ \text{for } j = 3, \ldots, d-1, && (3.44)\\
x_d &= r\cos(\theta_{d-2}), && (3.45)
\end{aligned}$$

where

$$r = \|\mathbf{x}\|, \cos(\phi) = x_1, \sin(\phi) = x_2 \quad \text{and} \tag{3.46}$$

$$\tan(\theta_j) = \frac{\sqrt{\sum_{i=1}^{j+1} x_i^2}}{x_{j+2}}, j = 1, \ldots, d-2. \tag{3.47}$$

Here, $r > 0$, $\phi \in [0, 2\pi)$ and $\theta_j \in [0, \pi)$ for $j = 1, \ldots, d-2$. Using this definition of spherical coordinates, we can reformulate the definition of a pdf of a spherical distribution in $\mathbb{R}^d$.

Definition 3.2.2 *A nonnegative function f with domain $D = [0, 2\pi) \times [0, \pi)^{d-2}$ is called a probability density function, pdf, of a spherical distribution in d dimensions, if*

$$\int_D f(\phi, \theta_1, \ldots, \theta_{d-2}) d\phi d\theta_1, \ldots, d\theta_{d-2} = 1.$$

The domain can be extended as follows

$$f(\phi + 2\pi, \theta_1, \ldots, \theta_{d-2}) = f(\phi, \ldots, \theta_{d-2}), \forall \phi \in [0, 2\pi), \forall \theta_i = [0, \pi), i = 1, \ldots, d-2.$$

The connection between a pdf $g(\mathbf{x})$ in cartesian coordinates and the corresponding pdf $f(\phi, \theta_1, \ldots, \theta_{d-2})$ in spherical coordinates is given by the well known theorem describing the change of variables, see for example Billingsley (1995), p. 215ff or p. 225ff. Noting that the Jacobian determinant of the transformation given by (3.42) - (3.45) is

$$|J(r, \phi, \theta_1, \ldots, \theta_{d-2})| = r^{d-1} \prod_{i=1}^{d-2} (\sin(\theta_i))^i,$$

we get the following relationship

$$g(x_1, \ldots, x_d) = f(\phi, \theta_1, \ldots, \theta_{d-2}) \left(\prod_{i=1}^{d-2} (\sin(\theta_i))^i \right)^{-1}. \tag{3.48}$$

We will work with densities in both cartesian and spherical coordinates, depending on which notation is more useful.

The main characteristic of spherical distributions is, as for circular distributions, the resultant. It is easier to define the resultant using cartesian coordinates rather than spherical coordinates.

Definition 3.2.3 *Let $\mathbf{X}$ be a d dimensional spherical random vector whose distribution is given by the pdf $g(\mathbf{x})$, expressed in cartesian coordinates. Then the population mean resultant ρ of $\mathbf{X}$ is defined as*

$$\rho = \left(\sum_{i=1}^{d} (\mathbb{E}[X_i])^2 \right)^{\frac{1}{2}} =: (\mathbb{E}[\mathbf{X}]^T \mathbb{E}[\mathbf{X}])^{\frac{1}{2}}, \tag{3.49}$$

where

$$\mathbb{E}[X_i] = \int_{\mathbb{S}^{d-1}} x_i g(\mathbf{x}) do(\mathbf{x})), \textit{ for } i = 1, \ldots, d.$$

The population mean direction is defined by

$$\mu_0 = \rho^{-1} \mathbb{E}[\mathbf{X}]. \tag{3.50}$$

The definition of the resultant length and the population mean direction are higher dimensional analogues of the respective definitions for circular distributions, given in Definition 3.1.4. Also very similar to the circular case, we define for a sample of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on $\mathbb{S}^{d-1}$ the *sample mean vector* as

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i. \tag{3.51}$$

As in the circular case, we define the *mean resultant length* $\overline{R}$ and the *sample mean direction* $\overline{\mathbf{x}}_0$ as

$$\overline{R} = \|\overline{\mathbf{x}}\|, \text{ and } \overline{\mathbf{x}}_0 = \|\overline{\mathbf{x}}\|^{-1} \overline{\mathbf{x}}. \tag{3.52}$$

Another important measure of dispersion for spherical distributions, that we mention for completeness, is the *scatter matrix $\overline{T}$ about the origin,* defined by

$$\overline{T} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T.$$

It may be useful to note that $\overline{T}$ can be thought of as the inertia tensor about the origin of a group of particles with equal mass n^{-1} located at positions $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The use and interpretation of $\overline{T}$ is given in Section 10.2 in Jupp and Mardia (2000).

3.2.2 Important Spherical Distributions

In the following we present some of the most important spherical distributions. We mostly concentrate on the von Mises-Fisher distribution, since it is our preferred choice for modelling the distribution of directional data in the framework of finite mixture models.

The Uniform Distribution, $U(\mathbb{S}^{d-1})$

This is the most basic distribution on $\mathbb{S}^{d-1}$. If $\mathbf{X}$ is distributed as $U(\mathbb{S}^{d-1})$, the probability $\mathbb{P}[\mathbf{X} \in A]$ is proportional to the surface area of A on $\mathbb{S}^{d-1}$. Therefore, we have in cartesian coordinates

$$g(x_1, ..., x_d) = \frac{1}{c(d)} = \frac{\Gamma(d/2)}{2\pi^{d/2}}, \tag{3.53}$$

where

$$c(d) = \int_{\mathbb{S}^{d-1}} do(\mathbf{x}) = \frac{2\pi^{d/2}}{\Gamma(d/2)} \tag{3.54}$$

denotes the surface area of $\mathbb{S}^{d-1}$ and where $\Gamma(x)$ denotes the Gamma function. In spherical coordinates we therefore get

$$f(\phi, \theta_1, \ldots, \theta_{d-1}) = \frac{1}{c(d)} \prod_{i=1}^{d-2} (\sin(\theta_i))^i = \frac{\Gamma(d/2)}{2\pi^{d/2}} \prod_{i=1}^{d-2} (\sin(\theta_i))^i. \tag{3.55}$$

The population mean resultant ρ of the Uniform distribution is 0, as in the circular case. Therefore, the population mean direction is not defined.

The von Mises-Fisher distribution

The von Mises-Fisher distribution is the natural extension of the circular von Mises distribution into higher dimensions. Recall that the von Mises distribution has density

$$g_M(\mathbf{x}; \boldsymbol{\mu}, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \boldsymbol{\mu}^T \mathbf{x}} = \frac{1}{2\pi I_0(\kappa)} e^{\kappa(\mu_1 x_1 + \mu_2 x_2)} \tag{3.56}$$

expressed in cartesian coordinates. Based on this observation, we make the following definition.

Definition 3.2.4 *The von Mises-Fisher distribution on $\mathbb{S}^{d-1}$, denoted by $\mathcal{M}(\boldsymbol{\mu}, \kappa)$, is the distribution whose density in cartesian coordinates is given by*

$$g_M(\mathbf{x}; \boldsymbol{\mu}, \kappa) = c_d(\kappa) \exp(\kappa \cdot \boldsymbol{\mu}^T \mathbf{x}) = c_d(\kappa) \exp\left(\kappa \sum_{i=1}^{d} \mu_i x_i\right), \tag{3.57}$$

with $c_d(\kappa)$ as given below, $\kappa > 0$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d) \in \mathbb{S}^{d-1}$ the mean direction, expressed in cartesian coordinates and $\mathbf{x} \in \mathbb{S}^{d-1}$.

In their book, Jupp and Mardia (2000) give the following equation for $c_d(\kappa)$:

$$c_d(\kappa) = \frac{(\kappa/2)^{d/2-1}}{\Gamma(d/2) I_{d/2-1}(\kappa)},$$

where $I_p(z)$ denotes the modified Bessel function of the first kind of order, given by equation (3.20). Unfortunately, this is not the correct formula for $c_d(\kappa)$. The following Lemma gives the correct equation for the constant $c_d(\kappa)$.

Lemma 3.2.5 *We have*

$$c_d(\kappa) = \frac{(\kappa/2)^{d/2-1}}{2\pi^{d/2} I_{d/2-1}(\kappa)}. \tag{3.58}$$

Proof:

We need to show that

$$c_d(\kappa)^{-1} = \int_{\mathbb{S}^{d-1}} \exp(\kappa \cdot \boldsymbol{\mu}^T \mathbf{x}) do(\mathbf{x})).$$

We express $\boldsymbol{\mu}$ and $\mathbf{x}$ in spherical coordinates, as explained in (3.42) - (3.45). We express $\boldsymbol{\mu}$ with the angles $\alpha, \beta_1, ..., \beta_{d-2}$ and $\mathbf{x}$ with $\phi, \theta_, ..., \theta_{d-2}$. After applying a change of coordinates, we may assume, without loss of generality, that $\beta_{d-2} = 0$. Through equations (3.42) - (3.45) we see that this means $\mu_d = 1$ and since $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$, this implies

$\mu_1 = ... = \mu_{d-1} = 0$. The integrand therefore simplifies to

$$\exp(\kappa \cdot \boldsymbol{\mu}^T \mathbf{x}) = \exp(\kappa \cdot \mu_d x_d) = \exp(\kappa \cdot \cos(\theta_{d-2})).$$

Therefore, we have

$$c_d(\kappa)^{-1} = \int_0^{2\pi} \int_0^{\pi} \dots \int_0^{\pi} e^{\kappa \cdot \cos(\theta_{d-2})} \prod_{i=1}^{d-2} (\sin(\theta_i))^i d\theta_1 \dots d\theta_{d-2} d\phi$$

$$= 2\pi \int_0^{\pi} \dots \int_0^{\pi} \prod_{i=1}^{d-3} (\sin(\theta_i))^i d\theta_1 \dots d\theta_{d-3} \int_0^{\pi} e^{\kappa \cdot \cos(\theta_{d-2})} \sin(\theta_{d-2})^{d-2} d\theta_{d-2}.$$

Note, that

$$2\pi \int_0^{\pi} \dots \int_0^{\pi} \prod_{i=1}^{d-3} (\sin(\theta_i))^i d\theta_1 \dots d\theta_{d-3} = \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)}.$$

because the left hand side equals $c(d-1) = \int_{\mathbb{S}^{d-2}} do(\mathbf{x})$, which equals the right hand side by (3.54). Furthermore we have from equation 9.6.18 in Abramowitz and Stegun (1972) that

$$\int_0^{\pi} e^{\kappa \cos(\theta)} \sin(\theta)^{2\nu} d\nu = \frac{\sqrt{\pi}\Gamma(\nu + 1/2) I_\nu(\kappa)}{(\kappa/2)^\nu}.$$

Setting $\nu = d/2 - 1$ and combining the two equations, we get

$$c_d(\kappa)^{-1} = \frac{2\pi^{d/2} I_{d/2-1}(\kappa)}{(\kappa/2)^{d/2-1}}.$$

■

In particular, we see that for d=2 and d=3 we have:

$$c_2(\kappa) = \frac{1}{2\pi I_0(\kappa)} \tag{3.59}$$

$$c_3(\kappa) = \frac{k}{4\pi \sinh(\kappa)}, \tag{3.60}$$

where we used that $I_{1/2}(\kappa) = (\frac{2}{\kappa\pi})^{1/2} \sinh(x)$ for (3.60). For d=3, we get the following equation for the density of the von Mises-Fisher distribution, expressed in spherical coordinates:

$$f_M(\phi, \theta; (\alpha, \beta), \kappa) = \frac{k}{4\pi \sinh(\kappa)} e^{\kappa[\cos\beta \cos\theta + \sin\beta \sin\theta \cos(\phi - \alpha)]} \sin(\theta), \tag{3.61}$$

with $0 \leq \theta, \beta < \pi$ and $0 \leq \phi, \alpha < 2\pi$. For $d > 3$, the density is usually only expressed in cartesian coordinates, as the expressions in spherical coordinates become to complicated.

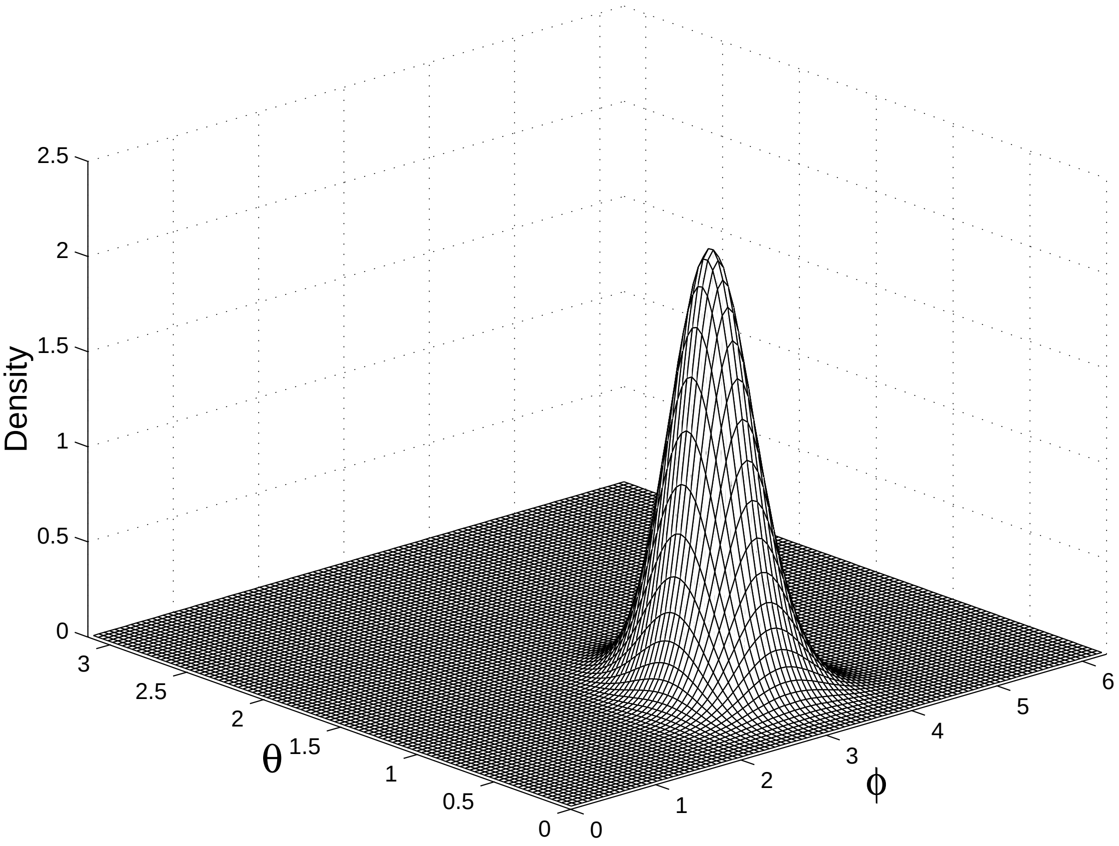

Figure 3.2: *The density of the von Mises-Fisher distribution with mean direction given by $\alpha = \pi$ and $\beta = \pi/4$ in spherical coordinates. The value of κ is 20.*

The density of the von Mises-Fisher distribution is unimodal with the mode at $\boldsymbol{\mu}$, provided that $\kappa > 0$. If $\kappa = 0$, the von Mises-Fisher distribution equals the Uniform distribution on $\mathbb{S}^{d-1}$. The larger the value of κ, the more the distribution is concentrated around $\boldsymbol{\mu}$. One can show that the density is rotationally symmetric about the mean direction $\boldsymbol{\mu}$. In that sense, the von Mises-Fisher distribution is comparable to a multivariate

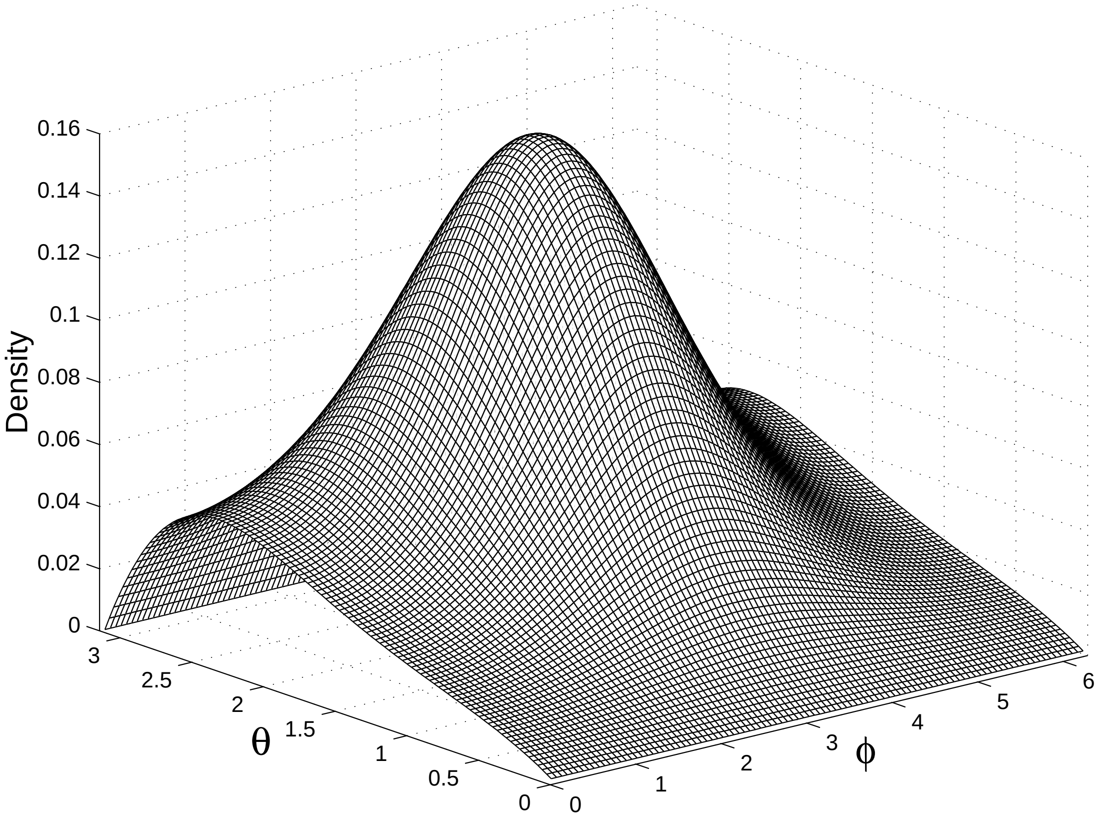

Figure 3.3: *The density of the von Mises-Fisher distribution with* $\alpha = \pi$ *and* $\beta = 3\pi/4$ *in spherical coordinates. The value of* κ *is 1.*

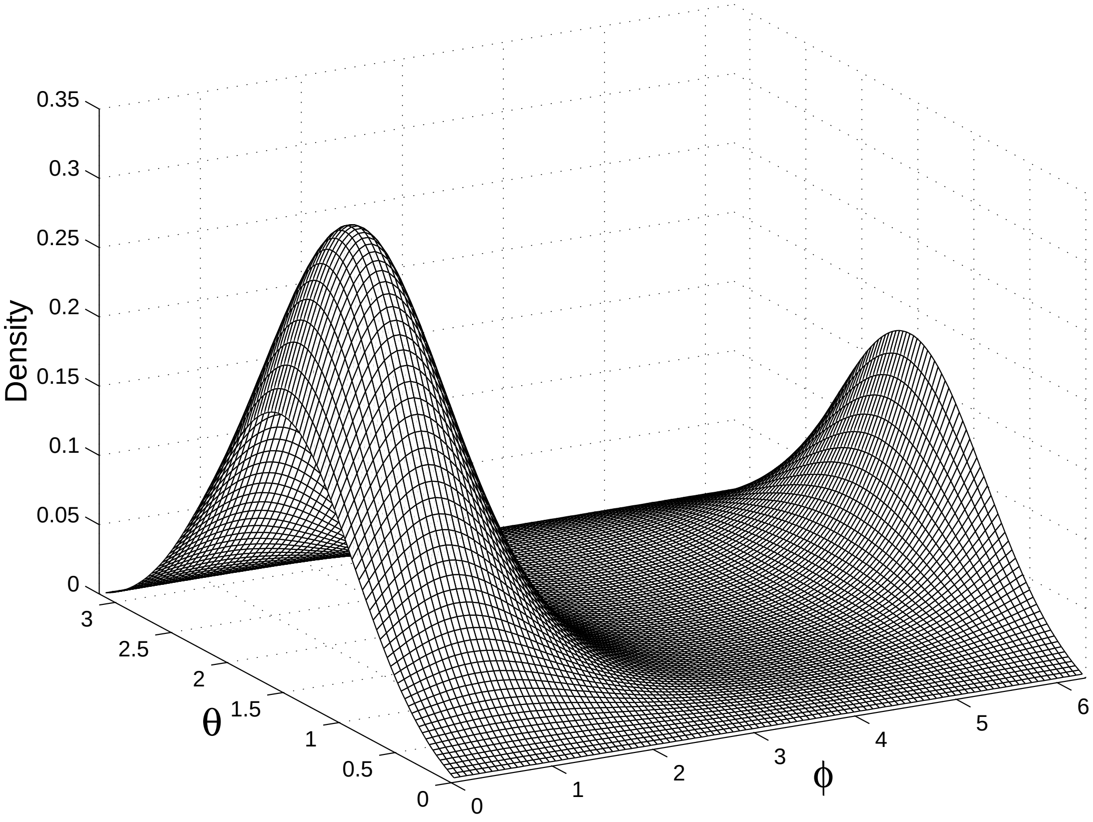

Figure 3.4: *The density of the von Mises-Fisher distribution with* $\alpha = \pi/4$ *and* $\beta = \pi/2$ *in spherical coordinates. The value of* κ *is 2.*

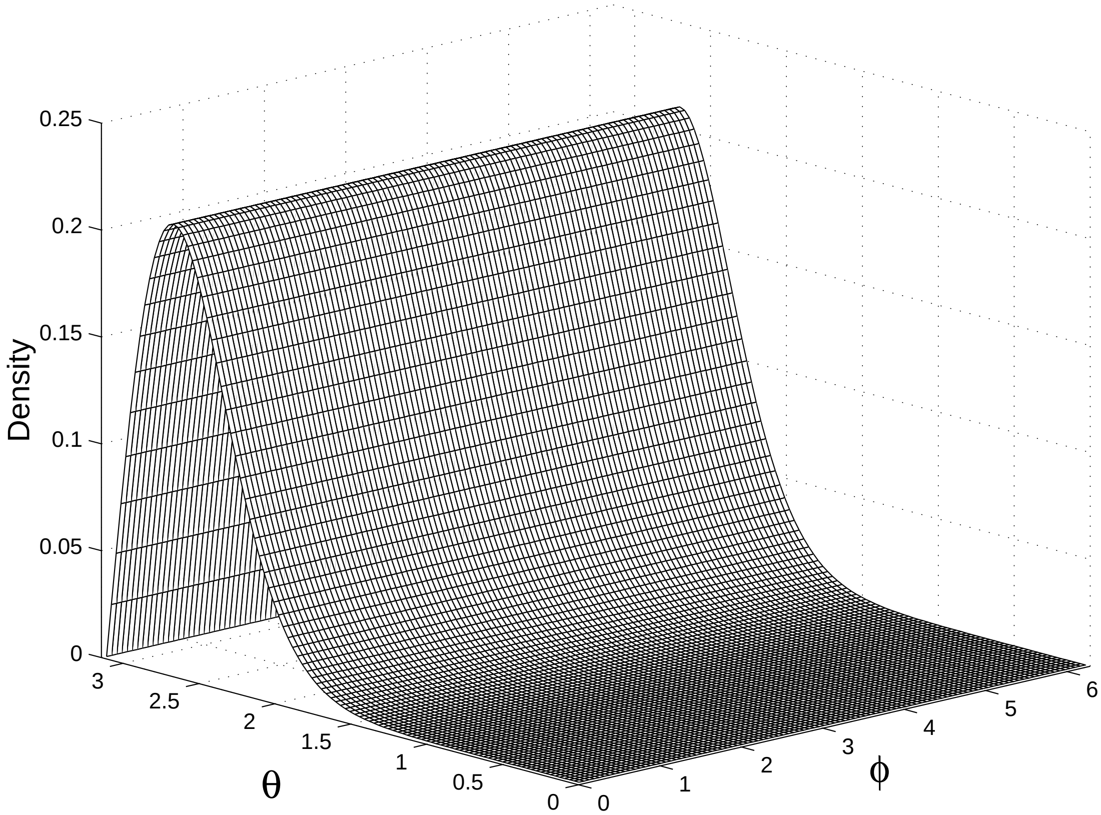

Figure 3.5: *The density of the von Mises-Fisher distribution with $\alpha = \pi$ and $\beta = \pi$ in spherical coordinates. The value of κ is 5.*

Normal distribution with a diagonal Variance-Covariance matrix.

As mentioned before, $\boldsymbol{\mu}$ is the population mean direction. The population resultant length ρ is given by

$$\rho = A_d(\kappa) := \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}. \tag{3.62}$$

Figures 3.2 - 3.5 exhibit the different shapes that the von-Mises Fisher distribution $\mathcal{M}(\boldsymbol{\mu}, \kappa)$ on $\mathbb{S}^2$ can have. The figures show the density of von Mises-Fisher distributions, given by (3.61) with various different choices of α, β and κ. Notice how the distribution is closely concentrated around the mean direction in Figure 3.2, where κ is fairly large, whereas in Figure 3.3 it is spread out over the entire unit sphere. Figure 3.4 clearly shows the periodicity of the density given by (3.61) in the first spherical coordinate, ϕ. Figure 3.5 shows a von Mises Fisher distribution with a mean direction of $(0, 0, 1)$, expressed in cartesian coordinates. The distribution is concentrated around the positive z-axis. Recall that the distribution has a rotational symmetry about the mean direction. Since the mean direction in this case is the z-axis, the variable ϕ is uniformly distributed, while the second variable, θ describes how concentrated the distribution is around the mean direction.

Notice that, with the exception of the density in Figure 3.5, the densities do not appear to be rotationally symmetric. This is due to distortions created by the change of variables, given in (3.42) - (3.45). We would also like to note that the family of von Mises-Fisher distributions is closed under orthogonal transformations. That is, if $\mathbf{U}$ is a orthogonal transformation, and $\mathbf{X} \overset{d}{=} \mathcal{M}(\boldsymbol{\mu}, \kappa)$, then $\mathbf{UX} \overset{d}{=} \mathcal{M}(\mathbf{U}\boldsymbol{\mu}, \kappa)$.

Maximum Likelihood Estimation in a von Mises Fisher distribution

Since the von Mises-Fisher distribution is an extension of the von Mises distribution, it is not surprising that the maximum likelihood estimators of the mean direction $\boldsymbol{\mu}$ and the concentration parameter κ are also analogues of their counterparts in the von Mises case. Let $\mathbf{x}_1, .., \mathbf{x}_n$ be a realization of the i.i.d. sequence of random variables $\mathbf{X}_1, .., \mathbf{X}_n$ distributed as $\mathcal{M}(\boldsymbol{\mu}, \kappa)$ on $\mathbb{S}^{d-1}$. The log-likelihood function, expressed in cartesian coordinates is

$$\begin{aligned} L(\boldsymbol{\mu}, \kappa; \mathbf{x}_1, .., \mathbf{x}_n) &= n \log(c_d(\kappa)) + \sum_{i=1}^{n} \kappa \boldsymbol{\mu}^T \mathbf{x}_i \\ &= n(\frac{d}{2} - 1) \log\left(\frac{\kappa}{2}\right) - n \log(2\pi^{d/2}) \\ &\quad -n \log(I_{d/2-1}(\kappa)) + \kappa \boldsymbol{\mu}^T (\sum_{i=1}^{n} \mathbf{x}_i). \end{aligned} \tag{3.63}$$

Concerning the MLE of the mean direction $\boldsymbol{\mu}$, we note that we can maximize the term involving $\boldsymbol{\mu}$, namely $\kappa \boldsymbol{\mu}^T (\sum_{i=1}^{n} \mathbf{x}_i)$, independently of the value of κ. This term is maximized by the vector in $\mathbb{S}^{d-1}$ with the same direction as $(\sum_{i=1}^{n} \mathbf{x}_i)$. That vector is of course the sample mean direction $\overline{\mathbf{x_0}}$, as defined in (3.52). We conclude that the MLE of the mean direction is

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x_0}}. \tag{3.64}$$

In that case we have

$$\kappa \widehat{\boldsymbol{\mu}}^T (\sum_{i=1}^{n} \mathbf{x}_i) = \kappa \overline{\mathbf{x_0}}^T (n \overline{\mathbf{x}}) = n \kappa \overline{R},$$

where $\overline{\mathbf{x}}$ is the sample vector mean and $\overline{R}$ is the mean resultant length.

Concerning the MLE of κ, we therefore need to maximize

$$\mathcal{L}(\kappa) := (\frac{d}{2} - 1) \log(\kappa) - \log(I_{d/2-1}(\kappa)) + \kappa \overline{R}$$

over the set $\{\kappa > 0\}$. The first derivative of $\mathcal{L}$ with respect to κ is

$$\frac{\partial \mathcal{L}}{\partial \kappa} = \frac{\frac{d}{2} - 1}{\kappa} - \frac{I'_{d/2-1}(\kappa)}{I_{d/2-1}(\kappa)} + \overline{R},$$

where $I'_{d/2-1}(\kappa) = \frac{\partial}{\partial \kappa} I_{d/2-1}(\kappa)$. From Abramowitz and Stegun (1972) we know that the following recurrence equation holds for $\nu > 0$:

$$I'_{\nu}(\kappa) = I_{\nu+1}(\kappa) + \frac{\nu}{\kappa} I_{\nu}(\kappa). \tag{3.65}$$

Therefore, we obtain

$$\frac{\partial \mathcal{L}}{\partial \kappa} = 0 \Longleftrightarrow \frac{\frac{d}{2}-1}{\kappa} - \frac{I_{d/2}(\kappa) + \frac{d/2-1}{\kappa} I_{d/2-1}(\kappa)}{I_{d/2-1}(\kappa)} = -\overline{R}.$$

Hence, $\widehat{\kappa}$ solves the equation

$$\frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} = \overline{R}.$$

If we define

$$A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}, \tag{3.66}$$

we see that the MLE of $\kappa, \widehat{\kappa}$ satisfies

$$A_d(\widehat{\kappa}) = \overline{R}. \tag{3.67}$$

Comparing $A_d(z)$ with the functions considered in Proposition 3.1.8, we see that $A_d(z) = B_{d/2}(z)$. Proposition 3.1.8 hence implies that $A_d(\cdot)$ is a monotone strictly increasing and continuous function and we have $\lim_{\kappa\to 0} A_d(\kappa) = 0$. In addition, we have $\lim_{\kappa\to\infty} A_d(\kappa) = 1$. Therefore, $\widehat{\kappa}$ is unique and well defined, since $\overline{R}$ is by definition a value in the interval $[0, 1]$. As for $A(\kappa) = A_2(\kappa)$, there is no explicit formula for $A_d^{-1}(\cdot)$. We again used a numerical procedure, implemented in Matlab, to evaluate $A_d^{-1}(\cdot)$. It should be noted that the maximum likelihood estimator is not unbiased, see Best and Fisher (1981). Modified estimators have been proposed to make the estimator of κ more robust. See Fisher (1982) for a reference on the 3 dimensional case. However, both the MLE for $\boldsymbol{\mu}$ and κ are consistent and asymptotically efficient estimators. See Jupp and Mardia (2000) for more properties of the estimators.

Generalizations of the von Mises-Fisher distribution

Recall from the definition of the von Mises-Fisher distribution that logarithm of the density is linear in $\mathbf{x}$. Generalizations of the von Mises-Fisher distribution typically add higher order polynomials to this linear term. The easiest case is given below, where quadratic terms have been added.

The *Fisher-Bingham* model (Mardia (1975)) has density

$$g(\mathbf{x};\boldsymbol{\mu},\kappa,\mathbf{A}) = \frac{1}{a(\kappa,\mathbf{A})} exp\{\kappa \cdot \boldsymbol{\mu}^T \cdot \mathbf{x} + \mathbf{x}^T\mathbf{A}\mathbf{x}\}, \tag{3.68}$$

where $\mathbf{A}$ is a symmetric $d \times d$ matrix. The constraint $\mathbf{x}^T\mathbf{x} = 1$ allows us to assume that $tr(\mathbf{A}) = 0$. Further models can be obtained by adding appropriate additional restrictions on the parameters of the Fisher-Bingham distribution. A variety of such models are listed in Section 9.3.3. of Jupp and Mardia (2000).

The *Kent distribution* has the same density as the Fisher-Bingham distribution, but with the additional constraint $\mathbf{A}\boldsymbol{\mu} = 0$.

The *Fisher Watson distribution* is obtained from (3.68) by replacing the restriction $tr(\mathbf{A}) = 0$ with the assumption that A is a diagonal matrix of full rank:

$$g(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\mu}_0,\kappa,\kappa_0) = \frac{1}{a(\kappa,\boldsymbol{\mu},\boldsymbol{\mu}_0,\kappa_0)} exp\{\kappa_0 \cdot \boldsymbol{\mu}_0^T \cdot \mathbf{x} + \kappa(\boldsymbol{\mu}^T\mathbf{x})^2\}. \tag{3.69}$$

A rotationally symmetric spherical distribution with a modal ridge along a small circle, instead of a mode at a single point, can be modelled by the *Bingham-Mardia distribution*. This ’small circle’ distribution has density

$$g(\mathbf{x};\boldsymbol{\mu},\kappa,\nu) = \frac{1}{a(\kappa)} exp\{\kappa(\boldsymbol{\mu}^T\mathbf{x} - \nu)^2\}. \tag{3.70}$$

The main problem for all those models is that the evaluation of the norming constants, $a(\kappa,\mathbf{A})$, $a(\kappa,\boldsymbol{\mu},\boldsymbol{\mu}_0,\kappa,\kappa_0)$ and $a(\kappa)$ respectively, is not easily done and may pose significant practical difficulties. This makes parameter estimation, using for example maximum likelihood methods, very difficult. This was the main reason that we decided to work with the simpler von Mises-Fisher model.

Chapter 4

Mixture Models of von-Mises distributions

4.1 Definition and Characteristic Function

Definition 4.1.1 *We define a finite mixture model of von Mises-Fisher distributions as the distribution with the pdf*

$$f_{mix}(\mathbf{x};\boldsymbol{\gamma}) = \sum_{i=1}^{m} p_i \cdot g_M(\mathbf{x};\boldsymbol{\mu_i},\kappa_i), \mathbf{x} \in \mathbb{S}^{d-1}, \tag{4.1}$$

where $g_M(\mathbf{x};\boldsymbol{\mu_i},\kappa_i)$ is the density of the von Mises-Fisher distribution with mean direction $\boldsymbol{\mu_i} \in \mathbb{S}^{d-1}$ and concentration parameter $\kappa_i \geq 0$, and $0 < p_i < 1$ are numbers satisfying $\sum_{i=1}^{m} p_i = 1$. Finally,

$$\boldsymbol{\gamma} = \{\boldsymbol{\mu_1},..,\boldsymbol{\mu_m},\kappa_1,..,\kappa_m,p_1,..,p_{m-1}\} \tag{4.2}$$

denotes the parameter matrix of the mixture model.

The p_i are referred to as the weights or mixing proportions. Note, that one of the weights is redundant because of the linear constraint $\sum_{i=1}^{m} p_i = 1$. We arbitrarily chose to omit the m^{th} weight p_m in the definition of the parameter $\boldsymbol{\gamma}$. The von Mises-Fisher densities $g_M(\mathbf{x};\boldsymbol{\mu_i},\kappa_i)$ are called the component densities. In (4.1) it is assumed that m, the number of components, is fixed. In practice, the choice of m is a part of the model and typically not known. By considering the number of components as yet another parameter of the model, the framework of finite mixture models (4.1) provides us with a very flexible method of modelling directional data. By choosing an appropriately large number of components, the density (4.1) can be made to provide an adequate fit to almost any data set. However, one has to be careful not to overfit the data and then end up with a meaningless model. For example, a seemingly perfect model for a data set $(\mathbf{x}_1, .., \mathbf{x}_N)$ of sample size N can be obtained with m components choosing $\kappa_i = \infty$, $\boldsymbol{\mu_i} = \mathbf{x_i}$ and $p_i = 1/m$. However, such a model has no predictive power for future observations and is obviously of little use. We will address the problem of choosing an adequate number of components in subsection 4.4.

Mixture models are especially useful in modelling heterogeneity in the data that stems from factors. Consider a categorical random variable Z with a distribution given by $\mathbb{P}[Z = i] = p_i$, $i = 1, .., m$. Assume, that there is another random variable Y that has conditional density $f_i(y)$, given $\{Z = i\}$. Then Y has unconditional density $f(x) = \sum_{i=1}^{m} p_i \cdot f_i(x)$. In this way, Y can be thought of as being drawn from m populations with densities $f_i(y)$ and proportions p_i. First, the categorical random variable Z chooses the population and then Y is drawn from the chosen distribution. The same framework also lets us interpret a mixture model as a case of incomplete data. We regard Y as the observable part of the random vector $X = (Z, Y)$, with Y and Z as above. However, we assume that the categorical random variable Z, thought of as the label of Y, has not been recorded or is not observable. Thus, we don't know which population generated Y. This idea of attaching missing labels to the observations is very useful in maximum likelihood estimation, as we will see in Section 4.2.2.

To calculate the characteristic function of a mixture of von-Mises distributions in the special case of $d = 2$, recall from (3.19) that if $X = e^{i\Phi}$ has a von Mises distribution, $\mathcal{M}(\alpha, \kappa)$, we have for its ch.f. that

$$\Psi_p = e^{ip\mu} \frac{I_p(\kappa)}{I_0(\kappa)}.$$

Therefore, if a random variable has a density given by (4.1), it has ch.f.:

$$\begin{aligned} \Psi_p &= \int_{\mathbb{S}^1} e^{ip\mathbf{x}} f_{mix}(\mathbf{x}; \boldsymbol{\gamma}) d\mathbf{x} \\ &= \sum_{j=1}^{m} p_j \int_0^{2\pi} e^{ip\varphi} f_M(\varphi; \alpha_j, \kappa_j) d\varphi = \sum_{j=1}^{m} p_j e^{ip\alpha_j} \frac{I_p(\kappa_j)}{I_0(\kappa_j)} \end{aligned} \qquad (4.3)$$

Hence the trigonometric moments are:

$$a_p = \mathbb{E}[\cos(p\Phi)] = \sum_{j=1}^{m} p_j \cos(p\mu_j)\frac{I_p(\kappa_j)}{I_0(\kappa_j)} \tag{4.4}$$

$$b_p = \mathbb{E}[\sin(p\Phi)] = \sum_{j=1}^{m} p_j \sin(p\mu_j)\frac{I_p(\kappa_j)}{I_0(\kappa_j)} \tag{4.5}$$

Recall that for spherical distributions on $\mathbb{S}^{d-1}$, for $d > 2$ we did not define a characteristic function or trigonometric moments. However we can give the population mean direction and the population resultant length. Recall from (3.62) that the resultant length of a von Mises-Fisher distribution with concentration parameter κ is $\rho = A_d(\kappa)$. Hence, we have that if $\mathbf{X}$ has a $\mathcal{M}(\boldsymbol{\mu}, \kappa)$ distribution, then

$$\mathbb{E}[\mathbf{X}] = \int_{\mathbb{S}^{d-1}} \mathbf{x} g_M(\mathbf{x}; \boldsymbol{\mu}, \kappa) d\mathbf{o}(\mathbf{x}) = \rho \cdot \boldsymbol{\mu} = A_d(\kappa)\boldsymbol{\mu}.$$

Therefore, we have for a finite mixture of von Mises-Fisher distributions (4.1)

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= \int_{\mathbb{S}^{d-1}} \mathbf{x} f_{mix}(\mathbf{x}; \boldsymbol{\gamma}) do(\mathbf{x}) \\ &= \int_{\mathbb{S}^{d-1}} \mathbf{x} \left(\sum_{i=1}^{m} p_i \cdot g_M(\mathbf{x}; \boldsymbol{\mu_i}, \kappa_{\mathbf{i}}) \right) do(\mathbf{x}) \\ &= \sum_{i=1}^{m} p_i A_d(\kappa_i) \boldsymbol{\mu_i}. \end{aligned} \tag{4.6}$$

We see that the expectation $\mathbb{E}[\mathbf{X}]$ is a linear combination of the mean directions $\boldsymbol{\mu_i}$ of the components with coefficients $p_i A_d(\kappa_i)$. As a consequence, we get for the resultant length and the mean direction

$$\rho = \left\| \sum_{i=1}^{m} p_i A_d(\kappa_i) \boldsymbol{\mu_i} \right\| \text{ and } \boldsymbol{\mu}_0 = \sum_{i=1}^{m} p_i A_d(\kappa_i) \boldsymbol{\mu_i} \cdot \rho^{-1}, \tag{4.7}$$

unless $\rho = 0$.

4.2 Parameter Estimation

4.2.1 Identifiability

Consider a parametric family represented by densities $f(x;\boldsymbol{\gamma})$. Estimation of the parameter $\boldsymbol{\gamma}$, based on a sample $\mathbf{x} = (x_1, .., x_N)$ is only meaningful, if the parameter $\boldsymbol{\gamma}$ of the density $f(x;\boldsymbol{\gamma})$ is identifiable. $\boldsymbol{\gamma}$ is called identifiable, if $f(x;\boldsymbol{\gamma}_1) \equiv f(x;\boldsymbol{\gamma}_2), \forall x$ implies $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$. Put in words, this means that distinct parameter values result in distinct densities. This is not true for finite mixture densities. A permutation of the component densities leaves $f(x;\boldsymbol{\gamma})$ invariant. Assume that the component densities $f_i(x;\xi_i)$, $i = 1, \ldots, m$ belong to parametric families. We have:

$$f(x;\boldsymbol{\gamma_1}) := \sum_{j=1}^{m} p_j \cdot f_j(x;\xi_j) = \sum_{j=1}^{m} p_{\pi(j)} \cdot f_{\pi(j)}(x;\xi_{\pi(j)}) =: f(x;\boldsymbol{\gamma_2}), \tag{4.8}$$

where π is a permutation of the numbers $1, .., m$ and $\boldsymbol{\gamma}_1 = (\xi_1, ..., \xi_m, p_1, ..., p_{m-1})$ and $\boldsymbol{\gamma}_2 = (\xi_{\pi(1)}, ..., \xi_{\pi(m)}, p_{\pi(m)}, ..., p_{\pi(m-1)})$ denote the parameter of the right hand side and left hand side of (4.8), respectively. Then we have in general that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_2$, but, nevertheless $f(x;\boldsymbol{\gamma}_1) = f(x;\boldsymbol{\gamma}_2)$. For this reason, the parameter vector $\boldsymbol{\gamma}$ of a finite mixture is not identifiable.

Fortunately this problem does usually not pose problems in practice for maximum likelihood estimation. The important exception to this statement is encountered when one uses Bayesian techniques using reversible Jump Markov Chain Monte Carlo techniques to determine the maximum likelihood estimators. Good references on that topic include Green (1995) and Green and Richardson (1997).

In order to obtain an identifiable model, we may for example impose restrictions on the parameters. In the von Mises-Fisher case with parameter $\boldsymbol{\gamma}$ given by (4.2), we might for example impose the following conditions on the parameters $\boldsymbol{\mu_1}, .., \boldsymbol{\mu_m}$, $\kappa_1, .., \kappa_m$, and $p_1, .., p_{m-1}$:

1. $\kappa_1 \leq \kappa_2 \leq ... \leq \kappa_m$

2. In case of a tie, that is, in case for any i, j $\kappa_i = \kappa_j$, we have, then we have $\boldsymbol{\mu_{i}}_1 \leq \boldsymbol{\mu_{j}}_1$.

3. In case this does not resolve the tie, that is, if $\kappa_i = \kappa_j$, and $\boldsymbol{\mu_{i}}_1 = \boldsymbol{\mu_{j}}_1$, we have $\boldsymbol{\mu_{i}}_2 \leq \boldsymbol{\mu_{j}}_2$. If the tie is still not resolved, we compare $\boldsymbol{\mu_{i}}_3 \leq \boldsymbol{\mu_{j}}_3$ and so forth.

These three conditions define a complete order on components of the mixture model, unless two components are identical, that is unless we have $\kappa_i = \kappa_j$, and $\boldsymbol{\mu_i} = \boldsymbol{\mu_j}$. If the above restrictions are placed on the parameter space Γ, the parameter of a finite mixture model becomes identifiable, provided that no two components are identical. We will for the remainder of the thesis assume that the parameter space has been restricted by a set of conditions like the one listed above in order to make the parameter identifiable and that not two components are identical. However, we were able to carry out the maximum likelihood estimation without this restrictions, using an EM algorithm, as we will describe below.

The case of two identical components is more problematic. It arises for example from attempts of fitting a model with too many components. We may fit a mixture model of $m + 1$ components to data that stems from a mixture density with m components by either

(i) setting one the weights $p_i = 0$, or

(ii) splitting a component into identical components.

We encountered this phenomenon in practice. In particular, if we worked with data that was simulated from a von Mises-Fisher mixture model with m components and tried to fit a model with $m + 1$ components, the EM algorithm returned parameter estimates with $\mu_i = \mu_j$, $\kappa_i = \kappa_j$, for two components $i \neq j$. While working with our implementation of the EM algorithm we observed that it splits components when the model is not identifiable because it has too many components.

However, suppose that we are working with the right number of components and add constraints like the ones listed above to the parameters to avoid an unidentifiable model due to permutations. In that case Titterington et al. (1985) shows that the parameters of finite mixtures of a large class of continuous densities are identifiable. The identifiability of the parameter of a finite mixture of von Mises distribution was proved in Fraser et al. (1981). The identifiability of a larger class of directional distributions, including the von Mises-Fisher distribution follows from a result in Kent (1983).

4.2.2 The EM Algorithm for General Mixture Models

It turns out that explicit formulas for the parameter estimates for mixture models are usually not available. The estimates of von Mises-Fisher mixture models in general and von Mises mixture models in particular are no exception. There is a wide spectrum of literature listing a variety of methods that have been used to obtain parameter estimates of various mixture models. They include Maximum Likelihood (ML) estimation Redner and Walker (1984) Dempster et al. (1977) , Bayesian estimation Green (1995), Green and Richardson (1997), Method of Moments Lindsay and Basak (1993), Minimum Distance methods Chen and Kalbfleisch (1996) and graphical methods. For a detailed overview of early work done on the estimation of finite mixture models we recommend Redner and Walker (1984) and the book Titterington et al. (1985). For an overview over later results, please consult McLachlan and Peel (2000).

We decided to use ML estimation. We employed the EM algorithm to iteratively compute the ML estimates. The main reason for the use of ML estimation are Theorems 4.2.4 and 4.2.6 given below. They state that in the framework of finite mixture models, the maximum likelihood estimator is asymptotical efficient and that the EM algorithm, started with proper starting values, converges to the MLE. In this section we describe the EM algorithm in a general framework. We will consider the special case of von Mises-Fisher distributions in Section 4.3.

Consider a finite mixture model involving parametric densities $f_i(x;\xi_i)$. We adopt the interpretation of a mixture measure as a case of incomplete data, mentioned in Section 4.1. Let $\mathcal{Y} = \mathcal{X} \times \{1,..,m\}$, where $\mathcal{X}$ is a measure space. Consider a sample of realizations $\mathbf{y} = (y_1,..,y_N)$ with $y_j = (x_j,i_j) \in \mathcal{Y}$, where $x_j \in \mathcal{X}$ are referred to as the observations and i_j are the unobservable labels. Assume that the joint density of the realizations $\mathbf{y}$, with respect to the product measure of the Lebesgue measure on $\mathcal{X}$ and the counting measure on $\{1,..,m\}$, is given by

$$\mathbf{f^c}(\mathbf{y};\boldsymbol{\gamma}) = \prod_{j=1}^{N} f^c((x_j,i_j);\boldsymbol{\gamma}) = \prod_{j=1}^{N} p_{i_j} \cdot f_{i_j}(x_j;\xi_{i_j}). \tag{4.9}$$

Here $\boldsymbol{\gamma} = \{\xi_1,..,\xi_m,p_1,..,p_{m-1},\} \in \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is the parameter space and $p_m = 1 - \sum_{j=1}^{m} p_j$. We assume that $\boldsymbol{\Omega}$ is a subset of the Euclidian space $\mathbb{R}^{m(q+1)-1}$. That is, we assume that $\xi_i \in \overline{\Omega}$, with $\overline{\Omega} \subseteq \mathbb{R}^q$. As explained in Section 4.1, a categorical random variable Z chooses the population and then the observation is drawn from the chosen distribution, independent of Z. An alternative notation for this model makes use of a matrix to label the observations. The matrix, denoted with $\mathbf{z}$ is defined by

$$z_{ij} = (\mathbf{z})_{ij} = \begin{cases} 1, & \text{if } i_j = i \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

Then we can express the density introduced in (4.9) as

$$\mathbf{f^c}(\mathbf{x},\mathbf{z};\boldsymbol{\gamma}) = \prod_{j=1}^{N}\prod_{i=1}^{m} p_i^{z_{ij}} \cdot f_i(x_j;\xi_i)^{z_{ij}}. \tag{4.11}$$

This is referred to as the complete model, because we know for each observation from which density $f_i(x;\xi_i)$ it was drawn. However, this is the information that we assume to be missing in the mixture model context. We do not know the label i_j that belongs to the observed value $x_j, j = 1, \ldots, m$. The recorded observations $\mathbf{x} = (x_1, .., x_N)$ thus have the following joint density, induced by (4.9):

$$\mathbf{f_{mix}}(\mathbf{x}; \boldsymbol{\gamma}) = \prod_{j=1}^{N} f_{mix}(x_j; \boldsymbol{\gamma}) = \prod_{j=1}^{N} \left[\sum_{i=1}^{m} p_i \cdot f_i(x_j; \xi_i) \right]. \tag{4.12}$$

This is referred to as the incomplete model.

While we do not know from which population a specific observation originated, we are able to make some inferences about the lost label. For an observation $x \in \mathcal{X}$ define $\mathcal{Y}(x) = \{y \in \mathcal{Y} : y = (x, i), i \in \{1, .., m\}\}$. The complete model (4.9) and the incomplete model (4.12) induce a conditional density with respect to the counting measure on $\mathcal{Y}(x)$. We denote that counting measure in the following by $c(dy)$. The conditional density on $\mathcal{Y}(x)$, given x, induced by (4.9) and (4.12) can be given in the notation $f^c(y; x, \boldsymbol{\gamma}) = k(y; x, \boldsymbol{\gamma}) \cdot f_{mix}(x; \boldsymbol{\gamma})$, where

$$k(y; x, \boldsymbol{\gamma}) = \frac{f^c(y; x, \boldsymbol{\gamma})}{f_{mix}(x; \boldsymbol{\gamma})} = \frac{p_i \cdot f_i(x; \xi_i)}{f_{mix}(x; \boldsymbol{\gamma})}. \tag{4.13}$$

$k(y; x, \boldsymbol{\gamma})$ can be interpreted as the posteriori probability that the observation x originated from the i^{th} population. In a similar fashion, we define the space $\mathcal{Y}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y}^N : y_j = (x_j, i_j), i \in \{1, .., m\}, j = 1, .., N\}$. Assuming that the realizations $\mathbf{y}$ are i.i.d., we define

$$\mathbf{k}(\mathbf{y}; \mathbf{x}, \boldsymbol{\gamma}) = \frac{\mathbf{f^c}(\mathbf{y}; \mathbf{x}, \boldsymbol{\gamma})}{\mathbf{f_{mix}}(\mathbf{x}; \boldsymbol{\gamma})} = \prod_{j=1}^{N} \frac{p_{i_j} \cdot f_{i_j}(x_j; \xi_{i_j})}{f_{mix}(x_j; \boldsymbol{\gamma})} \tag{4.14}$$

as a density on $\mathcal{Y}(\mathbf{x})$ with respect to the counting measure on $\mathcal{Y}(\mathbf{x})^N$.

This provides the framework that we use to maximize the log-likelihood function of the incomplete data:

$$L_N(\boldsymbol{\gamma}; \mathbf{x}) = \log((\mathbf{f_{mix}}(\mathbf{x}; \boldsymbol{\gamma})) = \sum_{j=1}^{N} \log(f_{mix}(x_j; \boldsymbol{\gamma})) \tag{4.15}$$

Consider a fixed parameter value $\widetilde{\gamma}$ of $\gamma \in \Omega$. We express the log-likelihood function as an expectation using the kernel density (4.14) with the fixed parameter $\widetilde{\gamma}$. That is, we write:

$$\begin{aligned}
L_N(\boldsymbol{\gamma};\mathbf{x}) &= \int_{\mathcal{Y}(\mathbf{x})} log((\mathbf{f_{mix}}(\mathbf{x};\boldsymbol{\gamma}))\mathbf{k}(\mathbf{y};\mathbf{x},\widetilde{\boldsymbol{\gamma}})c(d\mathbf{y}) \\
&= \sum_{j=1}^{N}\int_{\mathcal{Y}(x_j)} \log(f_{mix}(x_j;\boldsymbol{\gamma}))k(y;x_j,\widetilde{\boldsymbol{\gamma}})c(dy) \\
&= \sum_{j=1}^{N}\sum_{i=1}^{m} \log(f_{mix}(x_j;\boldsymbol{\gamma}))\frac{\widetilde{p_i}\cdot f_i(x_j;\widetilde{\xi_i})}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})} \qquad (4.16)
\end{aligned}$$

Using (4.13), we substitute $\log(f_{mix}(x_j;\boldsymbol{\gamma}))$ with $\log(f^c(y_j;\boldsymbol{\gamma})) - \log(k(y_j;x_j,\boldsymbol{\gamma}))$ and using that in the term

$$\log(f_{mix}(x_j;\boldsymbol{\gamma}))\frac{\widetilde{p_i}\cdot f_i(x_j;\widetilde{\xi_i})}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}$$

y_j stands for (x_j, i), we get:

$$\begin{aligned}
L_N(\boldsymbol{\gamma};\mathbf{x}) &= \sum_{j=1}^{N}\sum_{i=1}^{m} \log(f^c(y_j;\boldsymbol{\gamma}))\frac{\widetilde{p_i}\cdot f_i(x_j;\widetilde{\xi_i})}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})} \\
&\quad -\sum_{j=1}^{N}\sum_{i=1}^{m} \log(k(y_j;x_j,\boldsymbol{\gamma}))\frac{\widetilde{p_i}\cdot f_i(x_j;\widetilde{\xi_i})}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})} \\
&= \sum_{j=1}^{N}\int_{\mathcal{Y}(x_j)} \log(f^c(y_j;\boldsymbol{\gamma}))k(y;x_j,\widetilde{\boldsymbol{\gamma}})c(dy) \\
&\quad -\sum_{j=1}^{N}\int_{\mathcal{Y}(x_j)} \log(k(y_j;x_j,\boldsymbol{\gamma}))k(y;x_j,\widetilde{\boldsymbol{\gamma}})c(dy) \\
&= \sum_{j=1}^{N}\mathbb{E}[\log(f^c(y_j;\boldsymbol{\gamma}))|x_j,\widetilde{\boldsymbol{\gamma}}] - \sum_{j=1}^{N}\mathbb{E}[\log(k(y_j;x_j,\boldsymbol{\gamma}))|x_j,\widetilde{\boldsymbol{\gamma}}] \\
&= \mathbb{E}[\log(\mathbf{f^c}(\mathbf{y};\boldsymbol{\gamma}))|\mathbf{x},\widetilde{\boldsymbol{\gamma}}] - \mathbb{E}[\log(\mathbf{k}(\mathbf{y};\mathbf{x},\boldsymbol{\gamma}))|\mathbf{x},\widetilde{\boldsymbol{\gamma}}] \\
&=: Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) - H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) \qquad (4.17)
\end{aligned}$$

Consider the term $H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$. By Jensen's inequality, we have that for all $\boldsymbol{\gamma} \in \Omega$:

$$\begin{aligned} H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) - H(\widetilde{\boldsymbol{\gamma}}|\widetilde{\boldsymbol{\gamma}}) &= \mathbb{E}\left[\log\left(\frac{\mathbf{k}(\mathbf{y}|\mathbf{x},\boldsymbol{\gamma})}{\mathbf{k}(\mathbf{y}|\mathbf{x},\widetilde{\boldsymbol{\gamma}})}\right)|\mathbf{x},\widetilde{\boldsymbol{\gamma}}\right] && (4.18) \\ &\leq \log\left(\mathbb{E}\left[\frac{\mathbf{k}(\mathbf{y};\mathbf{x},\boldsymbol{\gamma})}{\mathbf{k}(\mathbf{y};\mathbf{x},\widetilde{\boldsymbol{\gamma}})}|\mathbf{x},\widetilde{\boldsymbol{\gamma}}\right]\right) && (4.19) \\ &= \log\left(\int_{\mathcal{Y}(\mathbf{x})}\mathbf{k}(\mathbf{y};\mathbf{x},\boldsymbol{\gamma})c(d\mathbf{y})\right) = 0. && (4.20) \end{aligned}$$

The last equality follows from the fact that $\mathbf{k}(\mathbf{y};\mathbf{x},\boldsymbol{\gamma})$ is a density on $\mathcal{Y}(\mathbf{x})$ and hence its integral over $\mathcal{Y}(\mathbf{x})$ equals one. We conclude that for all $\boldsymbol{\gamma},\widetilde{\boldsymbol{\gamma}} \in \Omega$

$$H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) \leq H(\widetilde{\boldsymbol{\gamma}}|\widetilde{\boldsymbol{\gamma}}). \tag{4.21}$$

Algorithm 4.2.1 ***(The general EM algorithm)***

Given a current estimate $\widetilde{\boldsymbol{\gamma}}$, obtain the next approximation $\boldsymbol{\gamma}^+$ as follows:

1. E Step: Determine $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$

2. M Step: Choose $\boldsymbol{\gamma}^+ =$ argmax$_{\boldsymbol{\gamma}\in\Omega}$ $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$

Equation (4.21) suggests that in each step, in order to obtain the next approximation to the MLE of $\boldsymbol{\gamma}$, it is enough to find a new estimate that maximizes $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$. Any value $\boldsymbol{\gamma}^+$ that maximizes $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ will reduce the value of $H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$. Therefore, we have that (4.21) and the definition of $\boldsymbol{\gamma}^+$ imply that

$$L(\boldsymbol{\gamma}^+;\mathbf{x}) = Q(\boldsymbol{\gamma}^+|\widetilde{\boldsymbol{\gamma}}) - H(\boldsymbol{\gamma}^+|\widetilde{\boldsymbol{\gamma}}) \geq Q(\widetilde{\boldsymbol{\gamma}}|\widetilde{\boldsymbol{\gamma}}) - H(\widetilde{\boldsymbol{\gamma}}|\widetilde{\boldsymbol{\gamma}}) = L(\widetilde{\boldsymbol{\gamma}};\mathbf{x}). \tag{4.22}$$

In others words, if we search for the maximum likelihood estimator of $\boldsymbol{\gamma}$ by means of the EM algorithm, the value of the likelihood increases with each iteration. It is this monotonicity property that makes the EM algorithm very attractive. It is also the property behind the convergence theorems given below.

The practicability of the EM algorithm heavily depends on how easy the maximization in the M-step is. We have

$$
\begin{aligned}
Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) &= \sum_{j=1}^{N}\sum_{i=1}^{m}\log(f^c(y_j;\boldsymbol{\gamma}))\frac{\widetilde{p}_i\cdot f_i(x_j;\widetilde{\xi}_i)}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}\\
&= \sum_{i=1}^{m}\log(p_i)\sum_{j=1}^{N}\frac{\widetilde{p}_i\cdot f_i(x_j;\widetilde{\xi}_i)}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}\\
&\quad+\sum_{i=1}^{m}\sum_{j=1}^{N}\log(f_i(x_j;\xi_i))\frac{\widetilde{p}_i\cdot f_i(x_j;\widetilde{\xi}_i)}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}. \qquad (4.23)
\end{aligned}
$$

This allows us to maximize the two terms separately. The maximization of the first term will give the new approximation p_i^+ of the component weight p_i, while the second term will give the new approximation ξ_i^+ of ξ_i. One can easily verify that the maximizer $\boldsymbol{\gamma}^+ = (p_1^+,..,p_m^+,\xi_1^+,..,\xi_m^+)$ satisfies

$$
\begin{aligned}
p_i^+ &= \frac{1}{N}\sum_{j=1}^{N}\frac{\widetilde{p}_i\cdot f_i(x_j;\widetilde{\xi}_i)}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}, \qquad (4.24)\\
\xi_i^+ &= \text{argmax}_{\xi_i\in\overline{\Omega}}\sum_{j=1}^{N}\log(f(x_j;\xi_i))\frac{\widetilde{p}_i\cdot f_i(x_j;\widetilde{\xi}_i)}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}. \qquad (4.25)
\end{aligned}
$$

The difficulty of solving equation (4.25) depends on the parametric family $f(x;\boldsymbol{\gamma})$ considered. It turns out that usually each ξ_i is easily and often uniquely and explicitly determined by (4.25). This is the case for example for exponential families and also the von Mises-Fisher distribution.

Note that

$$\frac{\widetilde{p}_i\cdot f_i(x_j;\widetilde{\xi}_i)}{f_{mix}(x_j;\widetilde{\boldsymbol{\gamma}})}$$

is the posterior probability that x_j was drawn from the i^{th} component population, based on the current estimate $\widetilde{\boldsymbol{\gamma}}$. p_i^+ is just the sample mean of those posterior probabilities.

Stopping Criteria The easiest stopping criteria involve the size of the change in either the parameter or the log-likelihood $L_N(\boldsymbol{\gamma};\mathbf{x})$. According to such a criteria, we would stop the algorithm as soon as the change in the value of the $L_N(\boldsymbol{\gamma};\mathbf{x})$ or the change of the parameters falls below a certain threshold. However, these are measures of lack of progress and not measures of convergence. We often observed that for many iterations of the EM algorithm the change in $L_N(\boldsymbol{\gamma}; \mathbf{x})$ remained small only to consequently grow to significant proportions again. This happened as the EM algorithm struggled through a sequence of approximations $\boldsymbol{\gamma}^{(k)}$ of the parameter $\boldsymbol{\gamma}$ that brought little change in $L(\boldsymbol{\gamma}^{(k)}; \mathbf{x})$ or $\boldsymbol{\gamma}^{(k)}$ itself before finding significantly better estimates again.We observed that the rate of convergence of $L^{(k)} = L(\boldsymbol{\gamma}^{(k)}; \mathbf{x})$ appeared to be very slow. Unfortunately this is known as the biggest drawback of the EM algorithm. See Redner and Walker (1984), McLachlan and Peel (2000), Titterington et al. (1985) or Lindsay and Basak (1993) for references on what they call ”linear“ convergence behavior of the sequence $L^{(k)}$. What they mean by ”linear“ is made precise in the following in equation (4.26). In this situation a stopping criteria, called the Aitken stopping criteria (ASC) is more adequate than the simple criteria mentioned in the beginning. Assume that the sequence $L^{(k)}$ converges to some value L^* as follows:

$$L^{(k+1)} - L^* \approx a(L^{(k)} - L^*) \Longleftrightarrow L^{(k+1)} - L^{(k)} \approx (1-a)(L^* - L^{(k)}). \qquad (4.26)$$

Even though the referenced authors refer to this convergence as linear, we feel more comfortable characterizing this form of convergence as a geometric convergence. Under (4.26), if a is close to one, a small difference $L^{(k+1)} - L^{(k)}$ does not imply that $L^{(k)}$ is close to L^*. We rather have that

$$L^* \approx L^{(k)} + \frac{1}{1-a}(L^{(k+1)} - L^{(k)}). \qquad (4.27)$$

Hence we obtain an estimate of L^*, called $L_A^{(k+1)}$, by replacing a in (4.27) with an estimate, say

$$a_{(k)} = \frac{L^{(k+1)} - L^{(k)}}{L^{(k)} - L^{(k-1)}}.$$

We obtain better stopping criteria

$$|L_A^{(k+1)} - L^{(k+1)}| \;<\; \epsilon \text{ or} \tag{4.28}$$

$$|L_A^{(k+1)} - L_A^{(k)}| \;<\; \epsilon, \tag{4.29}$$

where $\epsilon > 0$ is a chosen tolerance.

4.2.3 Properties of the MLE and the EM Algorithm in Finite Mixture Models

The results in this section are taken from Redner and Walker (1984), who summarize earlier results by Wald (1949) and Redner (1981). They address the consistency of the MLE and the convergence of the EM algorithm under the regularity assumptions below. In the following, we denote the true parameter vector by $\boldsymbol{\gamma}^*$ and the MLE of $\boldsymbol{\gamma}^*$ based on N observations by $\widehat{\boldsymbol{\gamma}}_N$. For this section only, we write $\boldsymbol{\gamma} = (\xi_1, ..., \xi_\nu)$ with $\xi_j \in \mathbb{R}^1$. ν denotes the dimension of the parameter vector.

Assumption 4.2.2 *For all $\boldsymbol{\gamma} \in \Omega$, for almost all $x \in \mathbb{R}^d$ and for $i, j, k = 1, ..., \nu$, the partial derivatives $\partial g/\partial\xi_i$, $\partial^2 g/\partial\xi_i\partial\xi_j$, and $\partial^3 g/\partial\xi_i\partial\xi_j\partial\xi_k$, exist and satisfy*

$$\left|\frac{\partial f_{mix}(x;\boldsymbol{\gamma})}{\partial\xi_i}\right| \leq f^i(x), \left|\frac{\partial^2 f_{mix}(x;\boldsymbol{\gamma})}{\partial\xi_i\partial\xi_j}\right| \leq f^{ij}(x), \left|\frac{\partial^3 f_{mix}(x;\boldsymbol{\gamma})}{\partial\xi_i\partial\xi_j\partial\xi_k}\right| \leq f^{ijk}(x)$$

where f^i and f^{ij} are integrable and f^{ijk} satisfies

$$\int_{\mathbb{R}^d} f^{ijk}(x) f_{mix}(x;\boldsymbol{\gamma}^*)dx < \infty$$

Assumption 4.2.3 *The Fisher Information matrix*

$$I(\boldsymbol{\gamma}) = \int_{\mathbb{R}^d} [\nabla_{\boldsymbol{\gamma}} \log(f_{mix}(x;\boldsymbol{\gamma}))][\nabla_{\boldsymbol{\gamma}} \log(f_{mix}(x;\boldsymbol{\gamma}))]^T f_{mix}(x;\boldsymbol{\gamma})dx$$

is well defined and positive definite at $\boldsymbol{\gamma}^*$.

Theorem 4.2.4 *If Assumptions 4.2.2 and 4.2.3 are satisfied and any sufficiently small neighborhood of $\boldsymbol{\gamma}^*$ in Ω is given, then with probability 1, there is for sufficiently large sample size N a unique solution $\widehat{\boldsymbol{\gamma}}_N$ of the likelihood equations $\nabla_{\boldsymbol{\gamma}} L_N(\boldsymbol{\gamma}; \mathbf{x}) = 0$ in that neighborhood, and this solution locally maximizes the log-likelihood function. Furthermore, $\sqrt{N}(\widehat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}^*)$ is asymptotically normally distributed with mean zero and covariance matrix $I(\boldsymbol{\gamma}^*)^{-1}$. Furthermore, if $H(\boldsymbol{\gamma}) = \sum_{j=1}^{N} \nabla_{\boldsymbol{\gamma}} \nabla_{\boldsymbol{\gamma}}^T \log(f_{mix}(x_j; \boldsymbol{\gamma}))$ is the Hessian of the log-likelihood function, with probability 1,*

$$\lim_{N \to \infty} \frac{1}{N} H(\widehat{\boldsymbol{\gamma}}_N) = -I(\boldsymbol{\gamma}^*)$$

While assuring us that the MLE $\widehat{\boldsymbol{\gamma}}_N$ is an asymptotically efficient estimator for $\boldsymbol{\gamma}$, the theorem still leaves two questions unresolved: Is $\widehat{\boldsymbol{\gamma}}_N$ really the largest local maximum of the log-likelihood function? Does a sequence of parameter estimates $\boldsymbol{\gamma}^{(j)}$ generated by the EM algorithm converge to $\widehat{\boldsymbol{\gamma}}_N$? The answer is given in the next theorem. We need Assumptions 3 and 4 given below. For $\boldsymbol{\gamma} \in \Omega$ and sufficiently small $r > 0$, let $N_r(\boldsymbol{\gamma})$ denote the closed ball of radius r about $\boldsymbol{\gamma}$ in Ω and define

$$f_{mix}(x; \boldsymbol{\gamma}, r) = \sup_{\widetilde{\boldsymbol{\gamma}} \in N_r(\boldsymbol{\gamma})} f_{mix}(x; \widetilde{\boldsymbol{\gamma}})$$

and

$$f^*(x; \boldsymbol{\gamma}, r) = max\{1, f_{mix}(x; \boldsymbol{\gamma}, r)\}$$

Assumption 4.2.5 *For each $\boldsymbol{\gamma} \in \Omega$ and sufficiently small $r > 0$,*

$$\int_{\mathbb{R}^d} f^*(x; \boldsymbol{\gamma}, r) f_{mix}(x; \boldsymbol{\gamma}^*) dx < \infty$$

Theorem 4.2.6 *Suppose that Assumptions 4.2.2 through 4.2.5 hold in Ω, and let Ω' be a compact subset of Ω which contains $\boldsymbol{\gamma}^*$ in its interior and such that $f_{mix}(x; \boldsymbol{\gamma}) = f_{mix}(x; \boldsymbol{\gamma}^*)$ almost everywhere in x for $\boldsymbol{\gamma} \in \Omega'$ only if $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. Suppose further that with* *probability 1, the function $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}))$ of the E-step of the EM algorithm is continuous in $\boldsymbol{\gamma}$ and $\widetilde{\boldsymbol{\gamma}}$ in Ω' and both $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ and the log-likelihood function $L_N(\boldsymbol{\gamma};\mathbf{x})$ are differentiable in $\boldsymbol{\gamma}$, for $\boldsymbol{\gamma} \in \Omega'$. Finally, for $\boldsymbol{\gamma}^{(0)}$ in Ω' denote by $\{\boldsymbol{\gamma}^{(j)}\}_{j=0,1,2..}$ a sequence generated by the EM algorithm in Ω', i.e., a sequence in Ω' satisfying*

$$\boldsymbol{\gamma}^{(j+1)} = arg \max_{\boldsymbol{\gamma}\in\Omega'} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(j)}), j = 0, 1, 2, \ldots$$

Then, with probability 1, whenever N is sufficiently large, the unique strongly consistent maximum-likelihood estimate $\widehat{\boldsymbol{\gamma}}_N$ is well defined in Ω' and $\widehat{\boldsymbol{\gamma}}_N = \lim_{j\to\infty} \boldsymbol{\gamma}^{(j)}$ whenever $\boldsymbol{\gamma}^{(0)}$ is sufficiently near $\widehat{\boldsymbol{\gamma}}_N$.

Theorems 4.2.4 and 4.2.6 assure of existence and uniqueness of a strongly consistent maximum likelihood estimate that can be obtained as the solution of the likelihood equations. We can find that estimate using the EM algorithm, if we have a starting point that is good enough. The two theorems provide the theoretical basis needed to justify the use of maximum likelihood estimation and the EM algorithm. However, many practical problems remain. Typically, the log-likelihood function will have many local maxima and may even be unbounded as $\boldsymbol{\gamma}$ approaches the boundary of the parameter space Ω. The likelihood equation may have solutions that are not local maxima of the log-likelihood function. In addition, the EM algorithm exhibits a very slow convergence behavior. It often takes several hundred iterations before the convergence criterion is met. It is therefore crucial to have a good starting point for the algorithm. We explain in Section 4.3.1 how we obtain good starting values.

4.3 The EM Algorithm for Finite von Mises-Fisher Mixture Models

It is easy to see that Assumptions 4.2.2 through 4.2.5 of the previous section hold for a finite mixture of von Mises-Fisher distributions, because the support of the densities is compact and each component density is in $C^\infty(\mathbb{S}^{d-1})$. We will make this point more precise below in Sections 4.4.1 and 4.4.2. In those sections we carefully check the validity of a set assumption that include or guarantee the validity of Assumptions 4.2.2, 4.2.3, and 4.2.5. Therefore, we can apply the results of Theorems 4.2.4 and 4.2.6.

The M-step in a finite von Mises-Fisher mixture model In the execution of the EM algorithm, let $\widetilde{\boldsymbol{\mu_i}}, \widetilde{\kappa_i}$ and $\widetilde{p_i}$ be the current approximation to the MLE of the parameters of the i^{th} component of the mixture model. Recall from (4.24) that the new approximation $\mathbf{p}^+ = (p_1^+, \ldots, p_m^+)$ of the weights $\mathbf{p} = (p_1, \ldots, p_m)$ is given by

$$p_i^+ = \frac{1}{N} \sum_{j=1}^{N} \frac{\widetilde{p_i} g_M(\mathbf{x}_j; \widetilde{\boldsymbol{\mu_i}}, \widetilde{\kappa_i})}{\sum_{k=1}^{m} \widetilde{p_k} g_M(\mathbf{x}_j; \widetilde{\boldsymbol{\mu_k}}, \widetilde{\kappa_k})} =: \frac{1}{N} \sum_{j=1}^{N} P_i(\mathbf{x}_j). \tag{4.30}$$

To find the new approximation of $\boldsymbol{\xi_i} = (\boldsymbol{\mu_i}, \kappa_i), i = 1, \ldots, m$, in the following denoted by $\boldsymbol{\xi_i}^+ = (\boldsymbol{\mu_i}^+, \kappa_i^+), i = 1, \ldots, m$, we need to solve equation (4.25). We need to find the pairs $(\boldsymbol{\mu_i}, \kappa_i)$, that maximize the equations

$$\sum_{j=1}^{N} \log(g_M(\mathbf{x}_j; \boldsymbol{\mu_i}, \kappa_i)) P_i(\mathbf{x}_j), i = 1, \ldots, m. \tag{4.31}$$

Recalling the definition of the von Mises-Fisher density $g_M(\mathbf{x}_j; \mu_i, \kappa_i)$ from (3.57), we have, after the simplifying and dropping the constant terms, that,

$$\left[\left(\frac{d}{2} - 1\right) \log(\kappa_i) - \log(I_{d/2-1}(\kappa_i))\right] \sum_{j=1}^{N} P_i(\mathbf{x}_j) + \kappa_i \boldsymbol{\mu_i}^T \left(\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right). \tag{4.32}$$

We see that $\boldsymbol{\mu_i}$ only appears in the second term. As in the case of the simple von Mises-Fisher distribution, we can therefore calculate the new approximations of κ_i and $\boldsymbol{\mu_i}$ separately. The second term, which needs to be maximized over $\boldsymbol{\mu_i} \in \mathbb{S}^{d-1}$, is the inner product of the two vectors $\boldsymbol{\mu_i}$ and $(\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j))$. Therefore, we conclude that

$$\boldsymbol{\mu_i}^+ = \left\|\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right\|^{-1} \sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j). \tag{4.33}$$

Studying the first term of (4.32), we see that we can also obtain the new approximation κ_i^+ for κ_i in a similar fashion as we obtained the MLE for the concentration parameter of a single von Mises-Fisher distribution. We have with the above notation, using (3.65) and (4.33):

$$\frac{\partial}{\partial \kappa_i}\left(\sum_{j=1}^{N} \log(M(\mathbf{x}_j; \boldsymbol{\mu_i}^+, \kappa_i))P_i(\mathbf{x}_j)\right) = 0 \tag{4.34}$$

$$\iff \quad \frac{\partial}{\partial \kappa_i}\left(\left[(\frac{d}{2} - 1)\log(\kappa_i) - \log(I_{d/2-1}(\kappa_i))\right]\sum_{j=1}^{N} P_i(\mathbf{x}_j)\right) \tag{4.35}$$

$$+\frac{\partial}{\partial \kappa_i}\left(\kappa_i \boldsymbol{\mu_i}^+ \left[\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right]\right) = 0$$

$$\iff \quad \left[\frac{\frac{d}{2} - 1}{\kappa_i^+} - \frac{I'_{d/2-1}(\kappa_i^+)}{I_{d/2-1}(\kappa_i^+)}\right]\sum_{j=1}^{N} P_i(\mathbf{x}_j) + \boldsymbol{\mu_i}^+ \sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j) = 0$$

$$\iff \quad -\frac{I_{d/2}(\kappa_i^+)}{I_{d/2-1}(\kappa_i^+)}\sum_{j=1}^{N} P_i(\mathbf{x}_j) + \left|\left|\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right|\right| = 0$$

$$\iff \quad A_d(\kappa_i^+) = \frac{\left|\left|\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right|\right|}{\sum_{j=1}^{N} P_i(\mathbf{x}_j)} \tag{4.36}$$

$$\iff \quad \kappa_i^+ = A_d^{-1}\left(\frac{\left|\left|\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right|\right|}{\sum_{j=1}^{N} P_i(\mathbf{x}_j)}\right) \tag{4.37}$$

We already mentioned that $A_d(\kappa)$ is a monotone strictly increasing function satisfying $\lim_{\kappa\to 0} A_d(\kappa) = 0$ and $\lim_{\kappa\to\infty} A_d(\kappa) = 1$. Therefore, (4.37) is meaningful, if

$$0 \leq \frac{\left|\left|\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)\right|\right|}{\sum_{j=1}^{N} P_i(\mathbf{x}_j)} \leq 1, \tag{4.38}$$

for all samples $\{\mathbf{x}_j \in \mathbb{S}^{d-1}, j = 1, \dots, N\}$ and for all choices of parameters $\widetilde{p_i}, \widetilde{\boldsymbol{\mu_i}}$ and $\widetilde{\kappa_i}$ that impact $P_i(\mathbf{x}_j)$. Indeed, $\sum_{j=1}^{N} \mathbf{x}_j P_i(\mathbf{x}_j)$ is a linear combination of the vectors $\mathbf{x}_j \in \mathbb{S}^{d-1}$ with parameters $P_i(\mathbf{x}_j)$. The length of the resulting vector is less or equal to $\sum_{j=1}^{N} P_i(\mathbf{x}_j)$, with equality if and only if for all j all $\mathbf{x}_j = \mathbf{x}$ for some $\mathbf{x} \in \mathbb{S}^{d-1}$. Hence, equation (4.38) is satisfied.

Summarizing, we can write the EM-Algorithm for a finite mixture of von Mises-Fisher distributions as follows:

Given the current values $\widetilde{\boldsymbol{\mu_1}}, \ldots \widetilde{\boldsymbol{\mu_m}}, \widetilde{\kappa_1}, \ldots, \widetilde{\kappa_m}, \widetilde{p_1}, \ldots, \widetilde{p_m}$, we obtain the updated value for p_i via (4.30), get the new values for $\boldsymbol{\mu_i}$ from (4.33) and the new values for κ_i from (4.37), $i = 1, \ldots, m$, until the Aitken stopping criterium (4.28) is met.

We see that for each component carrying out an iteration of the EM algorithm is no more difficult than obtaining the MLE for a single von Mises-Fisher distribution. Thus the speed the algorithm depends on the number of components and the efficiency of calculating the MLE of the parameters of a von Mises-Fisher distribution. Special care should be devoted to program an efficient version of the inversion of $A_d(\kappa)$.

4.3.1 Obtaining Good Starting Values: Method Of Moments

This approach to finding good starting values for the EM algorithm is based on results for finite mixture models of univariate normal distributions. While it is easy to implement and fast, it suffers from the drawback that it can only be used for von Mises mixture models. In other words, it is not useful in finding starting values for finite mixture models in higher dimensions than $\mathbb{S}_1$. It's use is therefore limited in practice.

Suppose that we wish to run the algorithm to fit a mixture model of von Mises distribution with m components to a dataset on the unit circle $\mathbb{S}_1$. We have to find starting values for

1) the concentration parameters, $\kappa_1, \ldots, \kappa_m$

2) the mean directions $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_m}$

3) the population weights $p_1, \ldots, p_m$ such that $\sum_{i=1}^{m} p_i = 1$.

Estimation of the location parameters

In Lindsay and Basak (1993), a fast method of moments is introduced to obtain starting values for the EM Algorithm in the case of finite mixtures of multivariate normal distributions. The paper is based on results of moments matrices found in the Appendix II of Uspensky (1937). His results describe how one can identify the n points of support of a discrete distribution and their weights. We adapt some of the results to the situation of discrete distributions on the unit circle and then explain how they can be used to find starting values for the location parameters.

Let as before $Z = e^{i\Theta}$ be a circular random variable with distribution function $F(d\phi)$. Let A_p be $p \times p$ matrix defined as

$$(A_p)_{i,j} = \Psi_{i+j-2} = \mathbb{E}[Z^{i+j-2}]$$

for $1 \leq i, j, \leq p$.

Finally let $\Delta_p = \det(A_p)$, $p \geq 0$. We set $\Psi_0 = \Delta_0 = 1$.

Assumption 4.3.1 *We have*

$$\Delta_0 \neq 0, \Delta_1 \neq 0, \ldots, \Delta_n \neq 0 \tag{4.39}$$

except on a set of parameters $\{p_1, \ldots, p_{s-1}, \alpha_1, \ldots, \alpha_s\}$ *of Lebesgue measure 0 in* $[0,1]^{s-1} \times [0, 2\pi)^s$.

Assume that $Z = e^{i\Theta}$ is a discrete circular random variable. Assumption 4.3.1 enables us to identify the points of support $\mu_j = e^{i\alpha_j}$, $j = 1, \ldots, n$ and their weights p_j of Z. Since the distribution of Z is discreet, it is entirely concentrated on the points of support. These points of support appear as the atoms of the distribution function F of Z and the corresponding weights satisfy the linear constraint

$$\sum_{j=1}^{s} p_j = 1.$$

Evaluating the first 2n-1 moments and the linear constraint on the weights p_j yields the following system of equations:

$$\begin{array}{c} \sum_{j=1}^{n} p_j = 1 \\ \sum_{j=1}^{n} p_j \mu_j^p = \Psi_p, \quad p = 1, \dots, 2n-1 \end{array} \tag{4.40}$$

(4.40) can be replaced by the more general but equivalent requirement that

$$\mathbb{E}[T(Z)] = \int_0^{2\pi} T(e^{i\phi}) F(d\phi) = \sum_{j=1}^{n} p_j T(\mu_j) \text{ for all functions } T. \tag{4.41}$$

It is in particular true for all polynomials with with $\deg T \leq 2n-1$. Suppose that such a polynomial $T(x)$ can be factorized as follows: $T(x) = a(x) \cdot Q(x)$, where $Q(x) = \prod_{j=1}^{n}(x - \mu_j) =: \sum_{k=0}^{n} q_k x^k$ and $a(x) = \sum_{j=0}^{n-1} a_j x^j$ is any polynomial of degree no more than $n-1$. Since the points of support of $F(d\phi)$ are exactly the roots of $Q(x)$, we have

$$\mathbb{E}[a(Z)Q(Z)] = \int_0^{2\pi} a(e^{i\phi})Q(e^{i\phi})F(d\phi) = \sum_{j=1}^{n} p_j a(\mu_j) Q(\mu_j) = 0.$$

On the other hand we have that

$$\begin{aligned} \int_0^{2\pi} a(e^{i\phi})Q(e^{i\phi})F(d\phi) &= \int_0^{2\pi} \sum_{j=0}^{n-1} \sum_{k=0}^{n} a_j q_k e^{i\phi(k+j)} F(d\phi) \\ &= \sum_{j=0}^{n-1} a_j \left(\sum_{k=0}^{n} q_k \phi_{k+j} \right) = 0. \end{aligned}$$

Since this must hold for arbitrary a_j, we must have that $\sum_{k=0}^{n} q_k \phi_{k+j} = 0$ for all $j = 0, \dots, n-1$. In matrix notation this is written as:

$$\underbrace{\begin{pmatrix} \phi_0 & \dots & \phi_{n-1} & \phi_n \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{n-1} & \dots & \phi_{2n-2} & \phi_{2n-1} \\ 0 & \dots & 0 & 1 \end{pmatrix}}_{=:B} \begin{pmatrix} q_0 \\ \vdots \\ q_{n-1} \\ q_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ c \end{pmatrix}, \tag{4.42}$$

by adding the additional condition $q_n = c$ to make B a possibly regular matrix, which would guarantee uniqueness of the solution. If Assumption 4.3.1 holds, we have that except on a set of parameters of Lebesgue measure 0, $\det(B) = \Delta_{n-1} \neq 0$. Using Cramer's rule we express the unique solution as follows:

$$\begin{aligned} q_j &= \frac{det\begin{pmatrix} \phi_0 & \dots & \phi_{j-1} & 0 & \phi_{j+1} & \dots & \phi_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{n-1} & \dots & \phi_{n+j-2} & 0 & \phi_{n+j} & \dots & \phi_{2n-1} \\ 0 & \dots & 0 & c & 0 & \dots & 0 \end{pmatrix}}{det(B)} \\ &= \frac{(-1)^{j+n}c}{\Delta_{n-1}}\ det\begin{pmatrix} \phi_0 & \dots & \phi_{j-1} & \phi_{j+1} & \dots & \phi_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{n-1} & \dots & \phi_{n+j-2} & \phi_{n+j} & \dots & \phi_{2n-1} \end{pmatrix} \end{aligned} \tag{4.43}$$

Therefore, we can write $Q(z)$ elegantly as the following determinant:

$$Q(z) = \frac{(-1)^n c}{\Delta_{n-1}}\ det\begin{pmatrix} 1 & z & \dots & z^n \\ \phi_0 & \phi_1 & \dots & \phi_n \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{n-1} & \phi_n & \dots & \phi_{2n-1} \end{pmatrix} \tag{4.44}$$

Note that if Assumption 4.3.1 holds, $Q(z)$ is indeed a polynomial of degree n, since the highest coefficient of $Q(z)$, $q_n = (-1)^n c$, is nonzero. We have proved the following Proposition.

Proposition 4.3.2 *Suppose that $Z = e^{i\Theta}$ is a discrete circular random variable with n points of support, called $\mu_j = e^{i\alpha_j}, j = 1, \dots, n$, such that Assumption 4.3.1 holds. Let $\Psi_p = \mathbb{E}[Z^p]$. Then the points of support are the simple and distinct roots of the polynomial are given by (4.44).*

We apply this result to the problem of obtaining starting values for the mean directions of the von Mises mixture model. Assume for the moment that all the components in the mixture have the same concentration parameter. We can write a random variable $Y = e^{iX}$ with that distribution stochastically as $Y = e^{i(\Theta+\mathcal{M})}$. $Z = e^{i\Theta}$ is a discrete random variable with n points of support $\mu_1 = e^{i\alpha_1}, \ldots, \mu_n = e^{i\alpha_n}$ and $\mathbb{P}[Z = \mu_j] = p_j$ and $M = e^{i\mathcal{M}}$ is a von Mises random variable with mean direction 0 and concentration parameter κ, independent of Z. Remembering that for a von Mises random variable we have that

$$\mathbb{E}[M^p] = \frac{I_p(\kappa)}{I_0(\kappa)},$$

we get

$$\mathbb{E}[Y^p] = \mathbb{E}[e^{ip\Theta}e^{ip\mathcal{M}}] = \mathbb{E}[e^{ip\Theta}]\mathbb{E}[M^p] = \Psi_p \cdot \frac{I_p(\kappa)}{I_0(\kappa)} =: \Psi_p(\kappa), \tag{4.45}$$

where $\Psi_p = \sum_{j=1}^n p_j e^{ip\alpha_j}$. Given an estimate $\widehat{\kappa}$ of κ we can therefore estimate Ψ_p by

$$\widehat{\Psi}_p = \frac{\widehat{\Psi_p(\kappa)} I_0(\widehat{\kappa})}{I_p(\widehat{\kappa})},$$

where $\widehat{\Psi_p(\kappa)}$ is an estimator for $\Psi_p(\kappa)$. We use the pth sample mean of the respective data set. We use $\widehat{\Psi}_p$ instead of the true and unknown moments Ψ_p in (4.44). We calculate the roots of the resulting polynomial $Q(x)$ and would like to use them as starting values for the location parameters in the EM algorithm. However, the data does not really follow a von Mises mixture with equal concentration parameters. Therefore, the roots of $Q(z)$ typically do not lie on the unit circle. However, we found that if $\breve{\mu}_j = r_j e^{i\widehat{\alpha}_j}, j = 1, \ldots, m$ are the roots of the polynomial, the values $\widehat{\mu}_j = e^{i\widehat{\alpha}_j}, j = 1, \ldots, m$ provide good starting values.

If Assumption 4.3.1 is violated, we may in particular have that $\det(B) = \Delta_{n-1} = 0$. In that case the matrix B is not regular and the coefficients of the polynomial $Q(x)$ cannot be determined from equation (4.42). In that case, equations (4.43) and (4.44) are meaningless because of the division by $\det(B) = \Delta_{n-1} = 0$. However, this was a case that we never experienced in our implementation of this method. Assumption 4.3.1 was never contradicted by empirical evidence. This allowed us to obtain good starting values by means of a method of moments.

Estimation of the concentration parameters

If we wish to apply the method of moments technique to find starting values for the mean directions of the EM algorithm, we need to obtain a good estimate of κ, the concentration parameter, that we assume to be equal for all components. The quality and usefulness of the starting values for the mean directions, is expected to depend on how good our estimate of κ is. Since the results in the previous paragraph assume that all $\kappa_j, j = 1, \ldots, m$ have the same value, the quality of the starting values will also depend on accurate that assumption is. If the actual concentration parameters κ_j are close to each other, we can expect to get fairly good starting values. However, if the true values for κ_j are very different, we might get starting values for the mean direction that do not lead the EM algorithm to the global maximum of the log-likelihood function, but rather only to a local maximum. We therefore try to identify a single value κ that is best used as the starting value for all κ_j, $j = 1, \ldots, n$. For a simple von Mises distribution, the concentration parameter κ is a function of the resultant length. We therefore concluded in (3.67) that the MLE of κ is a function of the mean resultant length. In our situation the situation is much trickier, since we have several components that influence the resultant length. The resultant length might even be 0. This is for example the case for a two component model with $\kappa_1 = \kappa_2 > 0$, $\mu_1 = \mu_2 + \pi$ and weights $p_1 = p_2 = .5$. This is however not a situation that we expect to see in practice. But we do expect that different components of the mixture that have different mean direction will have a resultant length that is smaller than the resultant length of each component alone. There is no easy and reliable way of separating the different components, before making some assumptions about the components of the model.

The following approach is therefore not expected to result in a reliable estimator for κ. It does, however, provide us with a reasonable starting value for κ, in the sense that it resulted in reasonable starting values for the mean directions. We need to make a number of simplifying assumptions about the nature of the mixture components. The first assumption is that we assume that each of the m components of the mixture is a random variable $Z'_j = e^{i\Theta'}$, where Θ' has range $[\alpha_j - \frac{2\pi}{m}, \alpha_j + \frac{2\pi}{m}]$ $i = 1, \ldots, n$. Here α_j stand for the mean direction of the j^{th} component. To get an estimate of κ we therefore essentially consider a circular random variables $Z' = e^{i\Theta'}$ with $\Theta' \in [0, \frac{2\pi}{m}]$. In Mardia (1972) it is argued that a reasonable definition of the circular variance V'_0 of Θ' could be defined as:

$$V'_0 = 1 - (1 - V_0)^{1/m^2},$$

where V_0 is the circular variance of the random variable $\Theta = m \cdot \Theta'$ with range $[0, 2\pi)$. In Chapter 3, we saw defined V_0 as $V_0 = 1 - \rho$, where ρ is the resultant length. We introduced the mean resultant length $\overline{R}$ as an estimator of ρ. Unfortunately there is no easy way of estimating the resultant length of $\Theta = m \cdot \Theta'$ of each component. We therefore make another simplifying assumption, namely that the mean resultant $\overline{R}$ of the entire data can be used. We therefore use

$$\widehat{V_0}' = 1 - (1 - \widehat{V_0})^{1/m^2} = 1 - \overline{R}^{1/m^2}$$

as an estimator of V'_0, since we need to obtain an estimate of V_0 of a generic component. Now recall that for a von Mises distribution, we have that the circular variance is $1 - A(\kappa) = 1 - \rho$, where ρ is the resultant length, estimated by the mean resultant length.

Therefore an estimator of the concentration parameter of $Z = e^{i\Theta}$ is the solution of

$$A(\widehat{\kappa}) = 1 - \widehat{V_0'} = \overline{R}^{1/m^2}. \tag{4.46}$$

We obtain the starting value for the concentration parameter κ of the von Mises mixture model by using the mean resultant length of the entire dataset as our choice for $\overline{R}$ in (4.46) and then solve for $\widehat{\kappa}$.

This is a similar equation as the one solved in the maximum likelihood estimation of the concentration parameter of a single von Mises distribution. Of course we are well aware that this method is fairly crude. As stated before it assumes that the concentration parameters κ_j have the same values. The interpretation of each component as a random variable on only a part of the circle is also only valid for large values of κ. In that case the corresponding random variable will be closely concentrated around its mean direction and can therefore essentially be regarded as a random variable on only a part of the unit circle. Clearly, this is not true if κ is fairly small. In addition, our technique implies the assumption that the resultant length of a mixture of m components with equal resultant length ρ is given by ρ^{1/m^2}. This need not be the case as pointed out by the example above with the two components placed on opposite places of the unit circle.
However, we only use this technique to obtain starting values and not actual estimates of κ_j, $j = 1, \ldots, n$.

Estimation of the weights

Given the starting values for the mean directions and the concentration parameter, we need to obtain starting values of the component weights. Ideally, we would like to use equations (4.40), replacing $\mu_j = e^{i\alpha_j}$ with the starting values for the mean directions described above. However, the data does not really follow a von Mises mixture with equal concentration parameters. Therefore the roots of the empirical version $Q(z)$ typically do not lie on the unit circle. Therefore the solutions $(p_1, \ldots, p_n)$ of (4.40) need not be real. This would likely even be true if the true distribution were a von Mises mixture model with equal concentration parameters, because of the noise in the data. However, we may take the real parts of that solution and treat them as starting values. Unfortunately, sometimes we even found that not all real parts are positive.

As an alternative, we consider a maximum likelihood estimator approach to obtain starting values of $p_1, \ldots, p_m$. We first obtain starting values for $\mu_1 = e^{i\alpha_1}, \ldots, \mu_m = e^{i\alpha_m}$ and κ. We then find the values $p_1, \ldots, p_n$ that maximize the log likelihood function, where $\mu_1, \ldots, \mu_m$ and $\kappa_1 = \cdots = \kappa_m = \kappa$ are considered parameters and not variables. That is, we treat the starting values for the location and concentration parameters as the true values in the execution of the EM-Algorithm and only maximize the log-likelihood function over the possible values of the component weights. Recall from (4.23) that in the E-Step we calculate

$$\begin{aligned} Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) &= \sum_{k=1}^{m} \log(p_k) \sum_{j=1}^{N} \frac{\widetilde{p}_k \cdot f(\mathbf{x}_j; \widetilde{\xi}_k)}{g(\mathbf{x}_j; \widetilde{\boldsymbol{\gamma}})} \\ &\quad + \sum_{k=1}^{m} \sum_{j=1}^{N} \log(f(\mathbf{x}_j; \xi_k)) \frac{\widetilde{p}_k \cdot f(x_j; \widetilde{\xi}_k)}{g(x_j; \widetilde{\boldsymbol{\gamma}})}, \end{aligned} \tag{4.47}$$

where $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_m, \widetilde{\kappa}_1, \ldots, \widetilde{\kappa}_m, \widetilde{p}_1, \ldots, \widetilde{p}_m)$ denotes the current estimate. The new estimate are found in the M-step in maximizing $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$. Assuming that $\mu_1 = \widehat{\mu}_1, \ldots, \mu_m = \widehat{\mu}_m$, and $\kappa_1 = \cdots = \kappa_m = \widehat{\kappa}$ are fixed at the values that we obtained by the methods described in the previous paragraphs, we only maximize $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ over $p_1, \ldots, p_m$. Given our current estimate $\widetilde{p}_1, \ldots, \widetilde{p}_m$, we find the new estimates according to the M-Step as

$$p_j^+ = \frac{1}{N} \sum_{k=1}^{N} \frac{\widetilde{p}_j f_j(x_k; \widehat{\mu}_j, \widehat{\kappa})}{f(x_k; \widehat{\boldsymbol{\gamma}})}, \tag{4.48}$$

where $\widehat{\boldsymbol{\gamma}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_m, \widehat{\kappa}, \ldots, \widehat{\kappa}, \widetilde{p}_1, \ldots, \widetilde{p}_m)$. The value of the log likelihood function increases in each iteration. The algorithm stops when the Aitken convergence criterion (4.28) is met. Since (4.48) provides an explicit formula for the new estimates, the algorithm usually is efficient and fast. The returned estimates $(\widehat{p_1}, \ldots, \widehat{p_m})$ can subsequently be used as starting values for the EM algorithm.

Performance in practice

In practice, the starting values, $\boldsymbol{\gamma}_0$, obtained by this method proved to be good if the number of components of the fitted model was small, typically not larger than 5. The value of the log-likelihood function $L(\boldsymbol{\gamma}_0; \mathbf{x})$ is reasonable close to the one at the MLE $\widehat{\boldsymbol{\gamma}}$, $L(\widehat{\boldsymbol{\gamma}}; \mathbf{x})$. The EM algorithm, started at $\boldsymbol{\gamma}_0$, usually converges to the largest of the local maxima of $L_N(\boldsymbol{\gamma}; \mathbf{x})$ in a reasonable number of iterations.

However, if a model with a larger number of components was fitted, problems with the starting values of the weights arose. The restricted EM algorithm used to obtain starting values for the weights $p_1, \ldots, p_m$ often converges to a vector $\widehat{p_m}, \ldots, \widehat{p_m}$, with one or more of the estimates $\widehat{p_j}$ very close to 0. This makes the corresponding component, and hence its mean direction and concentration parameter, insignificant in its influence on the value of the log likelihood function. In most of these cases the maximum likelihood estimates of those weights were distinctively different from 0, indicating that the starting values were very poor. It usually took the EM algorithm many iterations to recover from the bad starting values of the weights, if it did so at all. Oftentimes, the real parts of the solution $(p_1, \ldots, p_m)$ of (4.40) or even the crude estimates $p_j = \frac{1}{m}, j = 1 \ldots m$ provided better starting values. A possible reason for the poor performance of the method of moments with a relatively large number of components is that the differences between the different concentration parameters leads to a significant bias in the estimates of the mean direction. This in turn results in unreliable estimates of the weights. We often observed that for a small number of components the estimates for the concentration parameters were in the same range. However, when more than 5 component models were fitted, estimates for some κ_j were in the range of over 500-700, while others were well below 10.

Especially for models with a large number of components, the procedure presented in the next section proved superior to the method of moments, while the latter proved very helpful for models with a small number of components. Even more important is the fact that the procedure to be introduced below is applicable for data of any dimension, unlike the method of moments that we only implemented for the two dimensional case.

4.3.2 Starting Values Based on a Smaller Model

The need to fit a mixture model with a large number of components often arises because a reduced model does not provide a satisfactory fit. One might also try to justify the current model by fitting a model with an increased number of components and then showing that the new model provided no significant improvement over the current model. In both cases, the parameter estimates of the current model may already give us good information about the parameter estimates of some of the components of the larger model. This is especially true for models with a large number of components, because in that case the current model usually already provides us with a moderately good fit of the data. Therefore, we need not obtain starting values for all parameters using the method of moments described in the previous subsection. Instead, we can use the maximum likelihood estimates of the parameters of the smaller model as starting values for the parameters of the first components of the larger model. Assume that we already obtained a maximum likelihood estimate of the parameters of a mixture model with m components. We wish to fit a mixture model with $m + 1$ components. We assume that the MLE's of the mean directions, $\widehat{\boldsymbol{\mu}_1}, \ldots, \widehat{\boldsymbol{\mu}_m}$ and concentration parameters, $\widehat{\kappa}_1, \ldots, \widehat{\kappa}_m$ of the m components provide adequate starting values for the first m components of the larger model. We are therefore left with the problem of finding starting values for the weights $p_1, \ldots, p_{m+1}$ and the parameters $\boldsymbol{\mu}_{m+1}$ and κ_{m+1}.

We choose the values that maximize the log likelihood function of the larger model, where $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$, $\kappa_1, \ldots, \kappa_m$ have been fixed and are considered parameters and not variables. We therefore consider the log-likelihood function only as a function in $\boldsymbol{\mu}_{m+1}$, κ_{m+1} and $(p_1, \ldots, p_{m+1})$. That is, we attempt to maximize the following function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}_{m+1}, \kappa_{m+1}, p_1, \ldots, p_{m+1}; \mathbf{x_1}, \ldots, \mathbf{x_N}, \widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_m \widehat{\kappa}_1, \ldots, \widehat{\kappa}_m) = \\ \sum_{i=1}^{N} \log\left(\sum_{j=1}^{m} p_j \cdot M(\mathbf{x}_i; \widehat{\boldsymbol{\mu}}_j, \widehat{\kappa}_j) + p_{m+1} M(\mathbf{x}_i; \boldsymbol{\mu}_{m+1}, \kappa_{m+1})\right),\end{aligned} \tag{4.49}$$

where $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_m$ and $\widehat{\kappa}_1, \ldots, \widehat{\kappa}_m$ are the maximum likelihood estimate of the respective parameters in the smaller model with m components.

To find the desired starting values, we run a restricted EM algorithm similar to the case of determining the starting values of the weights in the method of moments technique, described in the last subsection. In each step, we only update the estimates of the values of the weights $p_1, \ldots, p_{m+1}$ and the parameters $\boldsymbol{\mu}_{m+1}$ and κ_{m+1}. Let $(\widetilde{p}_1, \ldots, \widetilde{p}_{m+1}, \widetilde{\boldsymbol{\mu}}_{m+1}, \widetilde{\kappa}_{m+1})$ denote the current approximations to the restricted MLE of $(p_1, \ldots, p_{m+1}, \boldsymbol{\mu}_{m+1}, \kappa_{m+1})$, let $(\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_m, \widehat{\kappa}_1, \ldots, \widehat{\kappa}_m)$ denote the fixed MLE's of the parameters $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, \kappa_1, \ldots, \kappa_m)$ of the first m components, and finally define $\widetilde{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_m, \widetilde{\boldsymbol{\mu}}_{m+1}, \widehat{\kappa}_1, \ldots, \widehat{\kappa}_m, \widetilde{\kappa}_{m+1}, \widetilde{p}_1, \ldots, \widetilde{p}_{m+1})$. Then we obtain our new approximations as:

$$p_k^+ = \frac{1}{N} \sum_{i=1}^{N} \frac{\widetilde{p}_k g_M(\mathbf{x}_i; \widetilde{\boldsymbol{\mu}}_k, \widetilde{\kappa}_k)}{f_{mix}(\mathbf{x}_i; \widetilde{\boldsymbol{\gamma}})}, \text{ for k=1,\ldots,m+1} \tag{4.50}$$

$$\boldsymbol{\mu}_{m+1}^+ = \left\|\sum_{i=1}^{N} \mathbf{x}_i P_{m+1}(\mathbf{x}_i)\right\|^{-1} \sum_{i=1}^{N} \mathbf{x}_i P_{m+1}(\mathbf{x}_i) \tag{4.51}$$

$$\kappa_{m+1}^{+} = A_d^{-1}\left(\frac{\left|\left|\sum_{i=1}^{N} \mathbf{x}_i P_{m+1}(\mathbf{x}_i)\right|\right|}{\sum_{i=1}^{N} P_{m+1}(\mathbf{x}_i)}\right) \tag{4.52}$$

where $P_{m+1}(\mathbf{x}_i)$ is as in (4.31), using the current approximations to the parameters. It is the calculation of the update of the concentration parameters that slows down the EM algorithm for the von Mises-Fisher model. This algorithm usually converges in a short time compared to the full fledged EM algorithm, since only the parameters of one component and the weights have to be updated in each iteration. Even though many iterations may be needed to find the desired starting values for $(p_1, \ldots, p_{m+1}, \boldsymbol{\mu}_{m+1}, \kappa_{m+1})$ the procedure proved to be very efficient in practice.

We usually started the algorithm with several different initial guesses for $(p_1, \ldots, p_{m+1}, \boldsymbol{\mu}_{m+1}, \kappa_{m+1})$. Typically these different initial values resulted in several different possible starting values for the EM algorithm. Among those possible starting values we typically preferred the one with the largest log-likelihood value. We observed however exceptions to this rule. Therefore, we usually ran the EM algorithm from all obtained possible starting values.

This method proved very valuable in practice, especially for a larger number of components when the method of moments estimates for the weights suffered from deficiencies described above. In higher dimensions it was our only tool to obtain good starting values.

4.4 Deciding on the Number of Components

The problem of determining the number of components in a finite mixture model has proven to be surprisingly tricky. A commonly used tool to determine the dimensionality of a model is the Likelihood Ratio (LR) test. Under certain regularity conditions, the test statistics asymptotically has a central χ^2 distribution with a known number of degrees of freedom, see Shao (1998) for a reference. Unfortunately these regularity condition are not met in the context of mixture models. Assume that we wish to test

H_0: The data arises from a mixture distribution with m_0 components.

H_1: The data arises from a mixture distribution with $m_1 > m_0$ components.

Recall from Section 4.2.1 that we can fit a model with m_0 components to data that stems from a mixture density with $m_1 < m_0$ components by either setting one the weights $p_i = 0$, or splitting a component into identical components. This means that under H_0 the parameters of the H_1 model are not identifiable or may lie on the boundary of the parameter space. It is not meaningful to estimate parameters that are not identifiable since the maximum likelihood function does not have a global maximum. It is therefore not meaningful to conduct likelihood ratio tests comparing the two models. Furthermore, the fact that the parameter estimates of the model may lie on the boundaries of the parameter space is a violation of the conditions necessary for the test statistic to have a central χ^2 distribution. We refer to McLachlan and Peel (2000), who discuss the problem of likelihood ratio testing in this framework in more detail. They note that the distribution of the usual likelihood ratio test function depends on the unknown parameter.

However, if we relax the assumptions about the true distribution, we can apply a result by Lo et al. (2001), presented for normal mixture models, that is based on earlier papers by White (1982) and Vuong (1989). This is our approach, which we explain it in more detail in this section. We assume that the true and unknown distribution of our observations is not part of our parametric model. To make this point more precise: We assume that the true distribution is not a finite mixture model of von Mises-Fisher distributions.

Before we discuss this approach in more detail, we address another practical concern in deciding on the number of components of a von Mises-Fisher mixture. It is the existence of spurious maxima of the log-likelihood function. These are local maxima that occur as a consequence of a cluster of a few data points that are relatively close together. These local maxima typically have at least one component with a very large concentration parameter κ and a very small component weight p. The models associated with these local maxima may have a high log-likelihood and therefore appear as a significant improvement over a reduced model in which the spurious component has been omitted. However, they are of little practical use and do not have a meaningful real world interpretation.

The following guidelines help identify spurious maxima and ignore them, even though they may seem as significant based on the model selection criteria explained in this section. Typically, the spurious component is not well isolated from the other components. It usually features a concentration parameter that is much larger than the ones from the other components and at the same time a weight that is much smaller, compared with the other weights. We often see $\kappa > 200$ and $p < 0.01$ for such a component. On the other hand, if a component is well separated from the other components it may have a meaningful real world interpretation, even though it shows a small p and a large κ. In addition, the EM algorithm usually only converges to a spurious maximum from a particular starting point. If even a moderately different starting point is chosen, convergence to another local maxima is observed. Isolated components with a large κ and low p do not have that property. This observation is useful in deciding on whether a solution is spurious.

4.4.1 MLE and Likelihood Ratio Testing in Misspecified Models

In this section we present a summary of the results about maximum likelihood estimation and likelihood ratio testing in misspecified models. The results first explain properties of the maximum likelihood estimators of the parameters of parametric models, if the true distribution of the observations is not included in the parametric model considered. They then continue to explain how to compare different such misspecified models in order to determine which one is closer to the true distribution. What exactly "closer to the true distribution" means will be made clear in the following. We will show later how these results can be applied to finite von Mises-Fisher mixture models.

Consider two different parametric models for the distribution of a random variable X. Following Vuong (1989), we assume that X has values in a Polish space $\mathcal{X}$.

$$\mathbb{F}_{\boldsymbol{\gamma}} = \{F(x;\boldsymbol{\gamma}), \boldsymbol{\gamma} \in \Gamma\} \subset \mathbb{R}^{n_1}, \tag{4.53}$$

and

$$\mathbb{G}_{\boldsymbol{\delta}} = \{G(x;\boldsymbol{\delta}), \boldsymbol{\delta} \in \Delta\} \subset \mathbb{R}^{n_0}. \tag{4.54}$$

We assume that $n_0 < n_1$. During this general discussion, the two families may or may not contain the true distribution $H(x)$ with density $h(x)$ with respect to a σ finite measure $\mu_{\mathcal{X}}$ on $\mathcal{X}$. It is convenient to think of $\mathcal{X}$ as the d-dimensional Euclidian space $\mathbb{R}^d$ and to assume that $\mu_{\mathcal{X}}$ is the Lebesgue measure.

It is our goal to decide which of the two parametric models is superior over the other one as explained in the following, based on a statistical test. We make the assumptions given below about the two competing families. The assumptions and results are stated only in terms of members of $\mathbb{F}_{\boldsymbol{\gamma}}$, but it is assumed throughout that analogous statements and results also hold for members of $\mathbb{G}_{\boldsymbol{\delta}}$.

Assumption 4.4.1

The random variables $X_1, .., X_N$ are independent and identically distributed with the density function $h(x)$, which is strictly positive for almost all $x \in \mathcal{X}$.

Assumption 4.4.2

(a) For every $\boldsymbol{\gamma}$ in Γ, $F(x;\boldsymbol{\gamma})$ has a density $f(x;\boldsymbol{\gamma})$ that is strictly positive for almost all $x \in \mathcal{X}$.

(b) The parameter space Γ is a compact subset of $\mathbb{R}^{n_1}$ and $f(x;\boldsymbol{\gamma})$ is continuous in $\boldsymbol{\gamma}$ for almost all x.

Assumption 4.4.3

(a) For almost all x, $|\log(f(x;\boldsymbol{\gamma}))|$ is bounded above by a function of x, independent of $\boldsymbol{\gamma}$, integrable with respect to H.

(b) The function $\mathbb{E}_h[\log(f(x;\boldsymbol{\gamma}))] = \int \log(f(x;\boldsymbol{\gamma}))h(x)\mu_{\mathcal{X}}(dx)$ has a unique maximum at $\boldsymbol{\gamma}^*$ in Γ.

(c) $\mathbb{E}_h[\log(h(x))] = \int \log(h(x))h(x)\mu_{\mathcal{X}}(dx)$ is well defined and finite.

Definition 4.4.4 *Define*

$$\begin{aligned} I(h:f|\boldsymbol{\gamma}) &:= \mathbb{E}_h\left[log\left(\frac{h(X)}{f(X;\boldsymbol{\gamma})}\right)\right] \\ &= \int_{-\infty}^{\infty}\log(h(x))h(x)dx - \int_{-\infty}^{\infty}\log(f(x;\boldsymbol{\gamma}))h(x)\mu_{\mathcal{X}}(dx). \quad (4.55) \end{aligned}$$

The function $I(h:f|\boldsymbol{\gamma})$ is called the Kullback-Leibler Information criterion (KLIC) statistic.

We refer to Kullback and Leibler (1951) for a discussion of the of the KLIC and its properties. $I(h:f|\boldsymbol{\gamma})$ can be understood as a measure of the distance between the model $F(x;\boldsymbol{\gamma})$ and the true distribution $H(x)$, see Akaike (1973) and Akaike (1974).

Assumptions 4.4.3 (a) and (c) assure that the KLIC is well defined. Define $\boldsymbol{\gamma}^*$ as the value $\boldsymbol{\gamma} \in \Gamma$ that minimizes the KLIC statistic over the parametric family $\mathbb{F}_{\boldsymbol{\gamma}}$. $\boldsymbol{\gamma}^*$ is called the quasi true value of $\boldsymbol{\gamma}$. Assumptions 4.4.3 (b) and (c) ensure that $\boldsymbol{\gamma}^*$ is globally identifiable. Since we interpret the KLIC as a measure of the distance of the model from the true distribution, we can use it to compare two competing models. We say that $\mathbb{F}_{\boldsymbol{\gamma}}$ is a better approximation to H than $\mathbb{G}_{\boldsymbol{\delta}}$, if

$$I(h : f|\boldsymbol{\gamma}^*) < I(h : g|\boldsymbol{\delta}^*). \tag{4.56}$$

To use this idea in practice, we need to find a test statistics based on a sample. We especially need to estimate $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$. To that end, define the quasi log-likelihood function of the sample $\mathbf{X} = (X_1, .., X_N)$ as

$$L_N(\boldsymbol{\gamma}; \mathbf{X}) = \sum_{i=1}^{N} \log(f(X_i; \boldsymbol{\gamma})) \tag{4.57}$$

and define the *quasi log-likelihood estimator* $\widehat{\gamma}_N$ (QMLE) as a parameter that solves

$$\max_{\boldsymbol{\gamma} \in \Gamma} L_N(\boldsymbol{\gamma}; \mathbf{X}). \tag{4.58}$$

The reason that we refer to $\widehat{\gamma}_N$ as the QMLE, rather than the MLE, is that we do not necessarily assume that the true distribution is a part of the parametric family $\mathbb{F}_{\boldsymbol{\gamma}}$. Therefore, $\boldsymbol{\gamma}$ does not necessarily estimate the true parameter, since there may not be a true parameter. But the QMLE is a natural estimator for $\boldsymbol{\gamma}^*$. This is made clear by the result below, addressing the consistency of the QMLE. Furthermore, if the true distribution is indeed part of the parametric family $\mathbb{F}_{\boldsymbol{\gamma}}$, then the QMLE is just the MLE and the quasi true value $\boldsymbol{\gamma}^*$ is of course the true value of $\boldsymbol{\gamma}$. That is the reason why we use the notation $\widehat{\gamma}_N$ for both the QMLE and the MLE.

Theorem 4.4.5 *If Assumptions 4.4.1 through 4.4.3 hold, then for all N there exists a measurable QMLE $\widehat{\boldsymbol{\gamma}}_N$ and $\widehat{\boldsymbol{\gamma}}_N \to \boldsymbol{\gamma}^*$ holds with probability 1, as $N \to \infty$. Furthermore, we have that with probability 1:*

$$\frac{1}{N} L_N(\mathbf{X}, \widehat{\boldsymbol{\gamma}}_N) \to \mathbb{E}_h[log(f(X; \boldsymbol{\gamma}^*))] \tag{4.59}$$

Proof: See Vuong (1989) or White (1982). ■

A direct consequence of Theorem 4.4.5 is that we have with probability 1

$$\begin{aligned} \frac{1}{N} LR_N \quad &:= \quad \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{f(X_i; \widehat{\boldsymbol{\gamma}_N})}{g(X_i; \widehat{\boldsymbol{\delta}_N})}\right) && (4.60) \\ &\longrightarrow \quad \mathbb{E}_h\left[\log\left(\frac{f(X; \boldsymbol{\gamma}^*)}{g(X; \boldsymbol{\delta}^*)}\right)\right] = I(h : g|\boldsymbol{\delta}^*) - I(h : f|\boldsymbol{\gamma}^*) && (4.61) \end{aligned}$$

Therefore, the likelihood ratio test appears as the natural test statistic for testing the null hypothesis that

$$I(h : f|\boldsymbol{\gamma}^*) = I(h : g|\boldsymbol{\delta}^*) \tag{4.62}$$

against the alternative hypothesis that

$$I(h : f|\boldsymbol{\gamma}^*) < I(h : g|\boldsymbol{\delta}^*) \tag{4.63}$$

We cannot expect that the asymptotic distribution of the LR test statistics will be the usual central χ^2 distribution, since the true distribution may not be included in any of the two parametric families. In order to get a more general result describing a non-degenerate limit distribution, we need to make the following further assumptions. We first introduce the following notation. Let

$$\left(\frac{\partial \log(f(x; \boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}}\right) \text{ and } \left(\frac{\partial \log(f(x; \boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}}\right)^T$$

be the vector with entries

$$\left(\frac{\partial \log(f(x; \boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}}\right)_j = \frac{\partial \log(f(x; \boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}_j}, j = 1, ..., n_1$$

and its transposed, respectively. Let

$$\left(\frac{\partial^2 \log(f(x;\boldsymbol{\gamma}))}{(\partial\boldsymbol{\gamma})\cdot(\partial\boldsymbol{\gamma})^T}\right)$$

be the matrix containing the second derivatives

$$\frac{\partial^2 \log(f(x;\boldsymbol{\gamma}))}{(\partial\boldsymbol{\gamma}_i)\cdot(\partial\boldsymbol{\gamma}_j)}, i = 1,\dots,n_1; j = 1,\dots,n_1$$

of $log(f(x;\boldsymbol{\gamma}))$.

Assumption 4.4.6

(a) For H almost all x, $log(f(x;\boldsymbol{\gamma}))$ *is twice continuously differentiable in* $\boldsymbol{\gamma}$*.*

(b) For H almost all x, the functions

$$\left|\left(\frac{\partial \log(f(x;\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}\right)^T\left(\frac{\partial \log(f(x;\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}\right)\right|_{(i,j)}$$

and

$$\left|\left(\frac{\partial^2 \log(f(x;\boldsymbol{\gamma}))}{(\partial\boldsymbol{\gamma})\cdot(\partial\boldsymbol{\gamma})^T}\right)\right|_{(i,j)}, i = 1,\dots,n_1; j = 1,\dots,n_1$$

are dominated by H-integrable functions that are independent of $\boldsymbol{\gamma}$*.*

Assumption 4.4.6 ensures the existence of the following matrices:

$$\begin{aligned} A_f(\boldsymbol{\gamma}) &= \mathbb{E}_h\left[\frac{\partial^2 \log(f(X;\boldsymbol{\gamma}))}{(\partial\boldsymbol{\gamma})\cdot(\partial\boldsymbol{\gamma})^T}\right] && (4.64)\\ B_f(\boldsymbol{\gamma}) &= \mathbb{E}_h\left[\left(\frac{\partial \log(f(X;\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}\right)\left(\frac{\partial \log(f(X;\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}\right)^T\right] && (4.65)\\ B_{fg}(\boldsymbol{\gamma},\boldsymbol{\delta}) &= \mathbb{E}_h\left[\left(\frac{\partial \log(f(X;\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}\right)\left(\frac{\partial \log(g(X;\boldsymbol{\delta}))}{\partial\boldsymbol{\delta}}\right)^T\right] && (4.66)\end{aligned}$$

Note, that we have $B_{gf}^T(\boldsymbol{\delta},\boldsymbol{\gamma}) = B_{fg}(\boldsymbol{\gamma},\boldsymbol{\delta})$. If the true distribution is indeed in the parametric family $F(x;\boldsymbol{\gamma})$, we have under certain regularity conditions that $-A_f(\boldsymbol{\gamma}^*) = B_f(\boldsymbol{\gamma}^*) = I(\boldsymbol{\gamma}^*)$, where $I(\boldsymbol{\gamma}^*)$ denotes the Fisher Information matrix. This is made precise in Corollary 4.4.9 below. Before stating the main result in this section, we need to make one more assumption.

Assumption 4.4.7

$\boldsymbol{\gamma}^*$ is an interior point of Γ and a regular point of $A_f(\boldsymbol{\gamma})$, that is, $A_f(\boldsymbol{\gamma})$ has constant rank in a neighborhood of $\boldsymbol{\gamma}^*$.

A result in White (1982) states that under Assumptions 4.4.1 - 4.4.6, Assumption 4.4.7 implies that $A_f(\boldsymbol{\gamma}^*)$ is negative definite and hence of full rank. We can now state the main results of this section.

Proposition 4.4.8 *Assume that the Assumptions 4.4.1 through 4.4.7 hold. Then we have, as $N \to \infty$:*

$$\sqrt{N}(\widehat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}^*) \Longrightarrow \mathcal{N}(0, C_f(\boldsymbol{\gamma}^*)) \tag{4.67}$$

where $C_f(\boldsymbol{\gamma}^*) = A_f^{-1}(\boldsymbol{\gamma}^*)B_f(\boldsymbol{\gamma}^*)A_f^{-1}(\boldsymbol{\gamma}^*)$

Proof: White (1982). ■

In order to understand Proposition 4.4.8 it is helpful to consider its statement in the case where the model is not misspecified. That is, consider the case where the true distribution is part of the parametric family $F(x; \boldsymbol{\gamma})$.

Corollary 4.4.9 *Given Assumptions 4.4.1 - 4.4.7 and if $h(x) = f(x; \boldsymbol{\gamma}_0)$, for some $\boldsymbol{\gamma}_0 \in \Gamma$, we have that*

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0 \textit{ and } A_f(\boldsymbol{\gamma}_0) = -B_f(\boldsymbol{\gamma}_0) \textit{ so that } C_f(\boldsymbol{\gamma}_0) = B_f^{-1}(\boldsymbol{\gamma}_0) = -A_f^{-1}(\boldsymbol{\gamma}_0). \tag{4.68}$$

In that case $C_f(\boldsymbol{\gamma}_0)$ is the Fisher Information matrix.

Proof: White (1982) ■

We see that the interpretation of the matrix $C_f(\boldsymbol{\gamma}_0)$ is analogue to that of the Fisher Information matrix. Assumptions 4.4.1 - 4.4.7 can be seen as the 'regular' maximum likelihood conditions. However, unless the model is correctly specified, we cannot expect that $A_f(\boldsymbol{\gamma}_0) = -B_f(\boldsymbol{\gamma}_0)$ and hence the asymptotic variance-covariance matrix may not equal the Fisher Information matrix.

In the framework of testing whether the larger model $F(x;\boldsymbol{\gamma})$ is significantly better than the smaller model $G(x;\boldsymbol{\delta})$, we need to make an assumption of how the smaller model can be seen as a special case of the larger model. This assumption is as follows:

Assumption 4.4.10

There exists a function $\xi(\cdot)$ from Δ to Γ such that, for almost all x, $g(x;\boldsymbol{\delta}) = f(x;\xi(\boldsymbol{\delta}))$, for every $\boldsymbol{\delta} \in \Delta$.

Given Assumption 4.4.10, together with Assumptions 4.4.1 through 4.4.3, we have that

$$\mathbb{E}_h[\log(f(X;\boldsymbol{\gamma}^*))] = \mathbb{E}_h[\log(g(X;\boldsymbol{\delta}^*))] \Longleftrightarrow I(h:f|\boldsymbol{\gamma}^*) = I(h:g|\boldsymbol{\delta}^*),$$

implies that $f(x;\boldsymbol{\gamma}^*) \equiv g(x;\boldsymbol{\delta}^*)$ for almost all x.

The following definition introduces the distribution that appears as the limiting distribution of the LR test statistic (4.60).

Definition 4.4.11 *Let $Z_1, ..., Z_n$ be i.i.d. standard normal random variables. Let $\lambda_1, \ldots,$ λ_n be real numbers. Then the distribution of the random variable $\sum_{i=1}^n \lambda_i Z_i^2$ is called weighted sum of chi-squared random variables with parameters n, λ. We use the notation: $\mathbb{P}[\sum_{i=1}^n \lambda_i Z_i^2 \le x] = M_n(x;\lambda)$, $x \in \mathbb{R}$.*

The distribution function $M_n(x;\lambda)$ is not available in closed form. We can however write it as an integral:

$$M_n(x;\lambda) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\sin(\delta(u))}{u\rho(u)}du, \tag{4.69}$$

where

$$\delta(u) = \frac{1}{2}\sum_{i=1}^{n} \arctan(\lambda_i u) - \frac{1}{2}xu,\ \rho(u) = \prod_{i=1}^{n}(1+\lambda_i^2 u^2)^{1/4}.$$

Theorem 4.4.12 *Assume that Assumptions 4.4.1-4.4.10 hold and that for almost all x we have $f(x;\boldsymbol{\gamma}^*) = g(x;\boldsymbol{\delta}^*)$. Then $2LR_N$ converges weakly to a weighted sum of chi-squared random variables:*

$$\mathbb{P}[2LR_N \leq y] \longrightarrow M_{n_1+n_0}(y;\lambda), \tag{4.70}$$

where λ is the vector of eigenvalues of the matrix

$$W = \begin{pmatrix} -B_f(\boldsymbol{\gamma}^*)A_f^{-1}(\boldsymbol{\gamma}^*) & -B_{fg}(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*) \\ B_{gf}(\boldsymbol{\delta}^*,\boldsymbol{\gamma}^*) & B_g(\boldsymbol{\delta}^*)A_g^{-1}(\boldsymbol{\delta}^*) \end{pmatrix}. \tag{4.71}$$

If on the other hand $\mathbb{E}_h[f(x;\boldsymbol{\gamma}^*)] > \mathbb{E}_h[g(x;\boldsymbol{\delta}^*)]$ then

$$2LR_N \longrightarrow \infty \text{ with probability 1.} \tag{4.72}$$

Proof: Lo et al. (2001), Vuong (1989). ■

In practice, λ has to be consistently estimated by the vector of the eigenvalues $\widehat{\lambda}$ of the matrix $\widehat{W}$, which is an estimate of W, obtained by replacing the expectation in the equations (4.64)-(4.66) by sample means and replacing $\boldsymbol{\gamma}^*$, $\boldsymbol{\delta}^*$ by their respective QMLE's.

4.4.2 Testing the Number of Components in a von Mises-Fisher Mixture Model

In order to use Theorem 4.4.12 we need to make sure that Assumptions 4.4.1 through 4.4.10 are satisfied. In the following, we assume specifically that $\mathbb{F}_{\boldsymbol{\gamma}}$ is the family of von Mises-Fisher mixtures with m components and that $\mathbb{G}_{\boldsymbol{\delta}}$ is the family of von

Mises-Fisher mixtures with $m-1$ components, $m \geq 2$. It is useful and sometimes even necessary during this section to work in spherical coordinates. We will especially need to express the mean direction of the components in spherical coordinates, when considering the derivatives mentioned in Assumptions 4.4.3 and 4.4.7. The reason is when we are taking derivatives, we need to make sure that there are no hidden constraints among the entries of the parameter vector. The mean directions of the components of the mixture appear as d dimensional vectors when expressed in cartesian coordinates. However the condition that they are vectors of unit sphere results in the fact that they only have $d-1$ degrees of freedom. If we were to take derivatives with respect to the mean direction of a certain component, expressed in cartesian coordinates, $\boldsymbol{\mu_i} = (\mu_1^{(1)}, \ldots, \mu_i^{(i)})$, we would have to consider the constraint

$$\sum_{j=1}^{d} \left(\mu_i^{(j)}\right)^2 = 1.$$

If we work in spherical coordinates and express $\boldsymbol{\mu_i}$ as $\boldsymbol{\mu_i} = (\alpha_i, \beta_i^{(1)}, \ldots, \beta_i^{(d-2)}) \in [0, 2\pi) \times [0, \pi]^{(d-2)} \subset \mathbb{R}^{d-1}$ we do not have constraints among the parameters components. In the following we mostly think in spherical coordinates. We also implemented the ratio likelihood ratio test described in this section in programs that carry out the calculations in spherical coordinates.

Concerning Assumption 4.4.1: We do not know what the true distribution of the data is, therefore we do not know whether its density is strictly positive for all $\mathbf{x} \in \mathbb{S}^{d-1}$. We will assume that this is true.

Concerning Assumption 4.4.2: The von Mises-Fisher density is strictly positive on the unit sphere as long as the concentration parameter κ is finite. Therefore a finite mixture of such densities is strictly positive, if at least one of the concentration parameters is finite. Below, we impose a finite upper bound on the concentration parameters to ensure compactness of the parameter space. Therefore, 4.4.2 a) is true. In order to address Assumption 4.4.2 b), we consider the mean directions in spherical coordinates. That is we have $\boldsymbol{\gamma} = \{\boldsymbol{\mu_1}, .., \boldsymbol{\mu_m}, \kappa_1, .., \kappa_m, p_1, .., p_{m-1}\}$ with $\boldsymbol{\mu_i} = (\alpha_i, \beta_i^{(1)}, \ldots, \beta_i^{(d-2)}) \in [0, 2\pi) \times [0, \pi]^{d-2}$, $\kappa_i \geq 0$, and $0 < p_i < 1$ are numbers satisfying $\sum_{i=1}^{m} p_i = 1$. We need to impose certain restrictions on the values of the parameters in order to obtain a compact parameter space. We start by making the range of α_i compact, by restricting its parameter values to the closed interval $[0, 2\pi - \epsilon]$, where $\epsilon > 0$ is chosen small enough so that this restriction is not of practical importance. We further introduce a maximal admissible value for the concentration parameters κ_j. This is necessary to obtain a compact parameter space. In practice, we never saw an estimate of a concentration parameter that exceeded a value of a 1000. We may therefore safely add a constraint of the form $0 \leq \kappa_i \leq \epsilon^{-1}$, $i = 1, .., m$, where ϵ is as above. Finally, we need to make the range of permissible values of p_i, $i = 1, \ldots, m$ compact. We therefore demand that for all i $p_i \in [\epsilon, 1 - \epsilon]$. Again ϵ is chosen small enough so that the restriction is not of practical importance. In practice, we never saw a parameter of the weights that was smaller than 10^{-4}, not even for spurious components. Note with this restriction the space of permissible values of the weights $\{\mathbf{p} = (p_1, \ldots, p_m) \in [\epsilon, 1 - \epsilon]^m : \sum_i p_i = 1\}$ is compact. Together we have that the space of possible values of $\boldsymbol{\gamma} = \{\boldsymbol{\mu_1}, .., \boldsymbol{\mu_m}, \kappa_1, .., \kappa_m, p_1, .., p_{m-1}\}$ is compact. In the following, we denote this compact parameter space with Γ_c.

Finally, it is easy to see that the density of a von Mises-Fisher model is a continuous function in each of the parameters.

Concerning Assumption 4.4.3: As we mentioned before, $\mathbb{S}^{d-1}$ is a compact subspace in $\mathbb{R}^d$. Therefore the density of a finite von Mises-Fisher mixture distribution, is a continuous function in $\mathbf{x} \in \mathbb{S}^{d-1}$ on a compact set. We just mentioned above that it is also a continuous function in the parameter $\boldsymbol{\gamma} \in \Gamma_c$, also a compact space. Therefore, the density is a continuous function in $(\mathbf{x}, \boldsymbol{\gamma}) \in \mathbb{S}^{d-1} \times \Gamma_c$. As a continuous function with a compact domain it is a bounded function. Combining this with the fact that the density is strictly positive, we get that $|\log(f_{mix}(\mathbf{x}, \boldsymbol{\gamma}))|$ is a bounded function and that the bound is independent of the parameter $\boldsymbol{\gamma}$.

In order to make the $\boldsymbol{\gamma}$ identifiable, we first impose the constraints introduced in Section 4.2.1. We denote the parameter space obtained from Γ_c by imposing the constraints 1. through 3. from Section 4.2.1 as Γ. However, if the true, unknown distribution H is indeed a finite von Mises-Fisher mixture distribution, Assumption 4.4.3(b) is still violated whenever the true distribution has less than m components. In that case $\boldsymbol{\gamma}$ is not identifiable, as we mentioned earlier in Section 4.2.1. We therefore need to assume that the true distribution of the data is either von Mises-Fisher mixture with at least m components or that it is not a finite von Mises-Fisher mixture at all. We worked with the second alternative. We assume that the true distribution is such that the parameter of the von Mises-Fisher mixture distribution is globally identifiable in the sense that $\mathbb{E}_h[\log(f_{mix}(\mathbf{x}, \boldsymbol{\gamma})]$ has a unique maximum at a parameter $\boldsymbol{\gamma}^*$ in the parameter space Γ. We need to stress that this assumption is stronger than the assumption that the true distribution is not a finite von Mises-Fisher model. There is no guarantee that the parameter $\boldsymbol{\gamma}$ of the mixture models is identifiable if we permit any distribution other than finite von Mises-Fisher distributions as the true distribution. Since we do not know the true distribution, how do we justify this assumption? We report in Section 4.2.1 how our implementation of the EM Algorithm handles the attempt to estimate a non identifiable parameter. We attempted to fit a von Mises-Fisher model with $m + 1$ to data that we simulated from a finite von Mises-Fisher mixture model with only m components. As described before, the parameter of the mixture with $m + 1$ components is not identifiable. We observed that in this situation, the EM Algorithm converges to a parameter estimate with two identical components. That is, it returns a parameter estimate $\widehat{\boldsymbol{\gamma}} = \{\widehat{\boldsymbol{\mu}}_{\boldsymbol{1}}, .., \widehat{\boldsymbol{\mu}}_{\boldsymbol{m}}, \widehat{\kappa}_1, .., \widehat{\kappa}_m, \widehat{p}_1, .., \widehat{p}_{m-1}\}$ with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{j}}$, and $\widehat{\kappa}_i = \widehat{\kappa}_j$ for $i \neq j$. If we would observe the EM algorithm return such estimates for real life data, this would indicate that the parameter we are trying to estimate is not identifiable. However, we did not observe this phenomenon, when working with real life data. This gives us confidence to work with the assumption that the underlying, true distribution is such that the parameters of the finite mixture models of von Mises-Fisher distributions is identifiable, regardless of the number of components in the mixture. Furthermore, we have to remember that the finite von Mises-Fisher mixture model is only a model. We cannot expect the data to originate precisely from any particular model we choose. Therefore the assumption that the true distribution is not captured in our model is not an unreasonable. We maintain however, that a finite von Mises-Fisher mixture is a good approximation to the true unknown distribution. The results in the previous section give the theoretical background of using the EM algorithm to obtain maximum likelihood estimates of the parameters under Assumption 4.4.3 b).

Finally, since we do not know the true distribution, we cannot verify whether Assumption 4.4.3 c) holds. Since we assume that the density h is strictly positive on the compact set $\mathbb{S}^{d-1}$, it is reasonable to assume that Assumption 4.4.3 c) holds.

Recall that based on these assumptions, Theorem 4.4.5 assures that the likelihood ratio test statistic converges almost surely to the difference of the KLIC statistics for two competing models.

Concerning Assumption 4.4.6: It is not hard to see that the function $\log(f_{mix}(\mathbf{x}, \boldsymbol{\gamma})$ is twice continuously differentiable. An examination of the resulting first and second derivatives reveals that they are all not only continuous functions of $\mathbf{x}$, but also of the parameter $\boldsymbol{\gamma}$. Therefore, we can repeat our argument from Assumption 4.4.3 to conclude that, as a continuous function on a compact set, the derivatives of $|\log(f_{mix}(\mathbf{x}, \boldsymbol{\gamma})|$ are bounded by a constant and hence Assumption 4.4.6 b) holds.

Concerning Assumption 4.4.7: The quasi true parameter $\boldsymbol{\gamma^*} \in \Gamma$ of the mixture models considered is in the interior of the respective parameter space, as long as the representation in spherical coordinates of the mean directions all have only angles that are in the interior of their permissible ranges. That we need to have that $\beta_i^{(j)*} \in (0, \pi)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, d-2$ and that $\alpha_i^* \in (0, 2\pi - \epsilon)$ for $i = 1, \ldots, m$. We can assume that this is true, otherwise we can apply a rotation of the coordinate system. Since we have from Assumption 4.4.3 that $\boldsymbol{\gamma}$ is identifiable, we have that all weights satisfy $p_i^* \in (\epsilon, 1-\epsilon)$ as long as ϵ has been chosen small enough. Finally we need to note that the condition $\kappa_i^* \in (0, \epsilon^{-1})$ only excludes uniform components. We can therefore safely assume that Assumption 4.4.7 a) is satisfied.

Checking that $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$ are regular points of $A_f(\boldsymbol{\gamma})$ and $A_g(\boldsymbol{\delta})$ respectively is not possible without knowledge of the unknown true distribution H. In practice we consider the corresponding estimates and check that they are indeed regular. We never encountered a instance, where one of those matrices was not regular.

Finally, assumption 4.4.10 is trivially satisfied.

After convincing ourselves that Assumptions 4.4.1-4.4.10 hold, we can apply Theorem 4.4.12 to perform likelihood ratio tests to compare finite mixture models with different number of components. As a result of Theorem 4.4.12, the following statistical test has asymptotical significance level α:

Likelihood ratio test for von Mises-Fisher mixture models:

Let $f(x; \boldsymbol{\gamma})$ be the density of a von Mises-Fisher mixture with m_1 components and let $g(x; \boldsymbol{\delta})$ be the density of a von Mises-Fisher mixture with m_0 components. Of course, *we assume that $m_0 < m_1$. The statistical test considers*

$H_0 : \mathbb{E}_h[log(f(x;\boldsymbol{\gamma}^*))] = \mathbb{E}_h[log(g(x;\boldsymbol{\delta}^*))]$ *i.e the two models are equivalent, versus*

$H_1 : \mathbb{E}_h[log(f(x;\boldsymbol{\gamma}^*))] > \mathbb{E}_h[log(g(x;\boldsymbol{\delta}^*))]$*, i.e. the larger model provides a significant improvement.*

We reject H_0, based on a data sample $X_1, ..., X_N$, if

$$2LR_N > M_m^{\leftarrow}(1-\alpha;\lambda), \tag{4.73}$$

where $M_m^{\leftarrow}(\cdot;\lambda)$ denotes the quantile function of the distribution $M_m(\cdot|\lambda)$, and m is the total number of parameters from both models.

We use this test as a tool in an algorithm to determine the number of components in a von Mises-Fisher mixture. We proceeded as follows, starting with m=2:

Algorithm 4.4.13 ***(Determining the number of components)***

1. *Estimate the parameters of a von Mises-Fisher mixture model with m and $m+1$ components.*

2. *Perform the likelihood ratio test (4.73) to compare the two models.*

3. *If the Null hypothesis is rejected, repeat steps 1. and 2. with m replaced by $m+1$, else accept m as an adequate number of components and $g(x;\widehat{\boldsymbol{\delta}})$ as the best fitted model.*

This is an automated procedure to determine an adequate number of components. Why then do we not compare a model with a certain number of components, say m with all reduced models with $2, 3, .., m-1$ number of components? The reason is that some of those tests would fail to reject the null hypothesis, while others would reject it. How would we decide which model is the best?

For example, the model with 4 components could appear statistically significantly superior over the model having 2 components. At the same time, it may not appear significantly superior compared with the model having 3 components. That model in return may or may not be significantly superior compared to the model with 2 components. Should we conclude that the model with 4 components is the most adequate model, based on its superiority over the 2 component model? Or should we choose the 3 component model, because it is statistically significantly superior over the 2 component model, while not being significantly inferior to the 4 component model? Our approach resolves these questions by only comparing each model only with a model that has either one component more or one component less. The procedure is motivated and justified, at least to some extend, by the following result, found in Cadez and Smyth (2000):

Proposition 4.4.14 *Denote with f_k the density of the mixture density*

$$f_k(x) = \sum_{i=1}^{k} \widehat{p}_i M(x_j; \widehat{\mathbf{u}}_\mathbf{i}, \widehat{\kappa}_i)$$

Denote with L_k the log-likelihood value of the mixture model with k components evaluated at the maximum likelihood estimates $(\widehat{p}_i, \widehat{\mathbf{u}}_\mathbf{i}, \widehat{\kappa}_i); i = 1, \ldots, k$. If for k_1 and k_2 we have that

$$L_{k_1} - L_{k_2} = \alpha \sum_{j=1}^{N} \frac{f_{k_2} - f_{k_1}}{f_{k_1}} \tag{4.74}$$

for a constant α. Then we have that

$$L_{k+1} - 2L_k + L_{k-1} \leq 0,$$

where

$$L_k = \sum_{j=1}^{N} \log \left(\sum_{i=1}^{k} \widehat{p}_i M(x_j; \widehat{\mathbf{u}}_\mathbf{i}, \widehat{\kappa}_i) \right)$$

stands for the log-likelihood value of the mixture model with k components evaluated at the maximum likelihood estimates $(\widehat{p}_i, \widehat{\mathbf{u}}_\mathbf{i}, \widehat{\kappa}_i); i = 1, \ldots, k$.

In other words, the log-likelihood function, evaluated at the corresponding MLE's is a concave function in the number of components used, under certain technical conditions. Cadez and Smyth (2000) note that if condition (4.74) holds approximately, the log likelihood is approximately concave. We refer to Cadez and Smyth (2000) for a more detailed discussion. As a consequence, the likelihood ratio test statistic $2LR_N$ is approximately monotone decreasing in the number of components. This does not imply that the p values of the corresponding likelihood ratio tests, described in this chapter, will also be monotone decreasing. Remember that the distribution of the statistic depends on the vector of parameters λ, defined as the vector of eigenvalues of the matrix given in Theorem 4.4.12. This means that because of different values associated with a likelihood ratio test, a test with a lower value of the test statistic than that of another test may reject the null hypothesis, while the later does not. This is however not very common.

We stop when the first likelihood ratio test comparing a mixture model with $m-1$ components with a model with m components fails to reject the null hypothesis. Even though there is no guarantee that a subsequential test comparing models with m and $m+1$ components will not reject the null hypothesis, Proposition 4.4.14 tells us that the value of the test statistic is monotone decreasing and hence that future significant values become fairly unlikely. In practice, we rarely saw this happening. When it happened, it was due to components that appeared to be spurious.

4.4.3 Information Criteria

As an alternative to the likelihood ratio test we also considered a variety of so called ”information criteria”. They are based on the Kullback Leibler information criterion of a parametric family with density $f(x;\boldsymbol{\gamma})$, introduced in (4.55):

$$I(h:f|\boldsymbol{\gamma}) := \mathbb{E}_h\left[log\left(\frac{h(X)}{f(X;\boldsymbol{\gamma})}\right)\right] = \int \log(h(x))h(x)dx - \int \log(f(x;\boldsymbol{\gamma}))h(x)dx.$$

As before, $h(x)$ stands for the density of the true distribution that may or may not be included in the parametric family $f(x;\boldsymbol{\gamma})$. If we had an estimator of the KLIC, we could therefore pick the model that minimizes said estimator. We fit models with a different number of components and choose the one that seems to minimize $I(h:f|\boldsymbol{\gamma})$. Recall that under certain regularity conditions, discussed in Section 4.4.1, we had in Theorem 4.4.5:

$$\frac{1}{N}L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)=\frac{1}{N}\sum_{i=1}^{N}\log(f(X_i;\widehat{\boldsymbol{\gamma}}_N))\rightarrow\mathbb{E}_h[log(f(X;\boldsymbol{\gamma}^*))],\text{ as }N\rightarrow\infty.$$

Recall that $\boldsymbol{\gamma}^*$ stands for the quasi true value of $\boldsymbol{\gamma}$, while $\widehat{\boldsymbol{\gamma}}_N$ stands for the QMLE, based on a sample $\mathbf{x}=(x_1,...,x_N)$ of $\mathbf{X}=(X_1,...,X_N)$ of sample size N. Hence we would choose the model that maximizes $L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)$. Unfortunately, the idea suffers from the problem that $L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)$ is a monotone increasing function in the number of components of the model, leading to over-parametrization. As a solution, we consider criteria based on functions that subtract a penalty term from $L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)$. The motivation for this approach is given by the fact, that even though we know that $\frac{1}{N}L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)$ is a consistent estimator for $\mathbb{E}_h[log(f(X;\boldsymbol{\gamma}^*))]$, it needs not be unbiased. Indeed, McLachlan and Peel (2000) mention, that the log likelihood usually has a positive bias. The bias is given by

$$b(h)=\mathbb{E}_h\left[\frac{1}{N}\sum_{i=1}^{N}\log(f(X_i;\widehat{\boldsymbol{\gamma}}_N))\right]-\int\log(f(x;\boldsymbol{\gamma}))h(x)\mu_X(dx).$$

This leads to the idea of estimating $\mathbb{E}_h[log(f(\mathbf{X};\boldsymbol{\gamma}^*))]$ by a term of the form

$$\frac{1}{N}L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)-\widehat{b(h)},\tag{4.75}$$

where $\widehat{b(h)}$ is an appropriate estimate of the bias $b(h)$. In the framework of mixture models, this motivates a new criterion for selecting the number of components of the model. Since the value of $L_N(\mathbf{X},\widehat{\boldsymbol{\gamma}}_N)$ is strictly monotone increasing in the number of components, we choose the model that maximizes a function of the form of equation (4.75). In literature, such functions are referred to as information criteria, since they aim to find the model that minimizes a modified version of the KLIC. They are typically expressed in the following form:

$$-2L_N(\mathbf{X}, \widehat{\boldsymbol{\gamma}}_N) + 2C(\widehat{\boldsymbol{\gamma}}_N), \tag{4.76}$$

where $2C(\widehat{\boldsymbol{\gamma}}_N)$ represents an appropriate penalty term. After fitting models with different numbers of components, we choose the model that minimizes a function of the form (4.76).

Obviously, the choice of $C(\widehat{\boldsymbol{\gamma}}_N)$ is critical to the sensibility of the criterion. Therefore, considerable effort has been devoted to an appropriate choice of $C(\widehat{\boldsymbol{\gamma}}_N)$. In the following, we present some proposed penalty terms that we considered for our work.

Akaike's Information Criterion

Akaike (1974) shows that, under certain regularity conditions, the bias term $b(h)$ asymptotically tends to d, the total number of parameters in the model, as N, the sample size, tends to ∞. This motivates the *Akaike's Information Criterion, AIC*:

$$AIC(\mathbf{X}; \widehat{\boldsymbol{\gamma}}_N) = -2L_N(\mathbf{X}, \widehat{\boldsymbol{\gamma}}_N) + 2d. \tag{4.77}$$

However, according to Titterington et al. (1985), the regularity conditions used by Akaike and other authors to derive the AIC are the same as the ones needed for the classical likelihood ratio test. As mentioned in the beginning of Section 4.4 these conditions break down in the framework of finite mixture models. However the AIC is still frequently used in deciding the number of components in various mixture models. In an empirical study we observed that the AIC tends to overestimate the true number of components, see Section $4.4.4$. This is in line of what other researchers reported as well.

Ishiguro et al. (1997) proposed using a bootstrap method to estimate the unknown bias term. See also McLachlan and Peel (2000) for a brief discussion of the resulting *Efron Information Criterion, EIC.*

Bayesian Information Criterion

The AIC and EIC are directly motivated by estimating the bias term in (4.75). The following criterion originated in the framework of Bayesian analysis, but has a similar form. Since it can be used in a non-Bayesian framework and is not harder to implement than the AIC, we found it to be very useful. In a Bayesian framework, assume that the prior distribution of the parameter $\boldsymbol{\gamma}$ is given by the density $f_p(\boldsymbol{\gamma})$. The integrated likelihood is then defined as

$$f_I(\mathbf{x}) = \int f_p(\boldsymbol{\gamma}) L_N(\mathbf{x}, \boldsymbol{\gamma}) d\boldsymbol{\gamma}.$$

Define the posterior mode $\widetilde{\boldsymbol{\gamma}}_N$ as the value of $\boldsymbol{\gamma}$ that maximizes $\log(f_p(\boldsymbol{\gamma}) L_N(\mathbf{x}, \boldsymbol{\gamma}))$. It solves the equation

$$\frac{\partial \log(f_p(\boldsymbol{\gamma}) L_N(\mathbf{x}, \boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}} = 0. \tag{4.78}$$

Using a second order Taylor approximation about the posterior node $\widetilde{\boldsymbol{\gamma}}_N$ we can approximate the integrated log likelihood with

$$\log(f_I(\mathbf{x})) = L_N(\mathbf{x}, \widetilde{\boldsymbol{\gamma}}_N) + \log(f_p(\widetilde{\boldsymbol{\gamma}}_N)) - \frac{1}{2}|I(h : f|\widetilde{\boldsymbol{\gamma}}_N)| + \frac{1}{2} d \log(2\pi). \tag{4.79}$$

Schwarz (1978) essentially obtained his *Bayesian Information Criterion, BIC*

$$BIC(\mathbf{x}; \widehat{\boldsymbol{\gamma}}_N) = -2L_N(\mathbf{x}, \widetilde{\boldsymbol{\gamma}}_N) + d \log(N) \tag{4.80}$$

from (4.79) by ignoring the terms $\log(f_p(\widetilde{\boldsymbol{\gamma}}_N))$ and $\frac{1}{2} d \log(2\pi)$ and using that $|I(h : f|\widetilde{\boldsymbol{\gamma}}_N)| = O(d \log N)$.

Comparing with the AIC, we see that as soon as $\log(N) > 2$, the penalty factor of the BIC is larger than the one of the AIC. Because of the larger penalty, the BIC has a smaller risk of choosing a too complicated model than the AIC. In our simulation study, presented in Section 4.4.4, we found that the BIC indeed performed better than the AIC. Other researchers reported similar findings in the context of mixture models, see McLachlan and Peel (2000), p. 209.

McLachlan and Peel (2000) mention however, that the regularity conditions needed for the Taylor approximation, as well as other approximations leading to (4.80), are not satisfied by mixture models. In particular, the approximation (4.79) requires that the parameters of the model be identifiable. As for the AIC, there is hence no theoretical justification for using the BIC in a mixture model context to decide on the number of components. As explained in the introduction to Section 4.4, if the true distribution is part of the considered mixture family, and we are considering a model with more components than the true distribution, the parameters of the model are not identifiable.

However, Leroux (1992) has shown that asymptotically, for large sample sizes, both the AIC and the BIC do not underestimate the true number of components. This is reassuring. It means that when using the BIC and/or the AIC for deciding on the number of components in the model, we will likely not choose a model that is too simple and therefore miss important information about the tail dependence in the distribution.

McLachlan and Peel (2000) mention two more complicated criteria that are based on Bayesian methods, the Laplace Metropolis Criterion and the Laplace Empirical Criterion. It should also be noted, that Green (1995) present a Bayesian approach to the estimation of the parameters of a model that Green and Richardson (1997) applied to finite mixtures. In that approach the number of components is treated just like another parameter. With the help of a Monte Carlo method, discussed in Green (1995), a posterior distribution on the number of components is derived. However the computational requirements are significant, even for univariate data and would increase dramatically for multivariate data.

Classification-Based Information Criterion

We introduce two criteria that are based on the idea that the true model should be able to classify the observations using the different components of the model. It should be possible, with the help of the model, to determine from which component a particular observation originated. Recall, that the complete model, introduced in Section 4.2.2, refers to the case where we know for each observation from which component it comes from. Its density is given by equation (4.9):

$$\mathbf{f^c}(\mathbf{y};\boldsymbol{\gamma}) = \prod_{j=1}^{N} f^c((x_j, i_j);\boldsymbol{\gamma}) = \prod_{j=1}^{N} p_{i_j} \cdot f(x_j;\xi_{i_j}),$$

where ξ_i are the parameters of the i^{th} component density. To express the log likelihood of the complete model, L_N^c, recall from (4.10) the definition of the matrix

$$z_{ij} = \begin{cases} 1, & \text{if } i_j = i \\ 0 & \text{otherwise} \end{cases}$$

Then we have

$$L_N^c(\mathbf{x};\mathbf{z};\boldsymbol{\gamma}) = \sum_{i=1}^{m}\sum_{j=1}^{N} z_{ij}\left[\log(p_i) + \log(f(x_j;\xi_i))\right]. \tag{4.81}$$

The connection between $L_N^c(\mathbf{x};\mathbf{z};\boldsymbol{\gamma})$ and the log-likelihood function $L_N(\mathbf{x};\boldsymbol{\gamma})$ of the incomplete model is given by the equation

$$L_N^c(\mathbf{x};\mathbf{z};\boldsymbol{\gamma}) = L_N(\mathbf{x};\boldsymbol{\gamma}) + \log(k_N(\mathbf{x};\mathbf{z};\boldsymbol{\gamma})), \tag{4.82}$$

where

$$\log(k_N(\mathbf{x};\mathbf{z};\boldsymbol{\gamma})) := \sum_{i=1}^{m}\sum_{j=1}^{N} z_{ij}\log(\tau_{ij}),$$

see McLachlan and Peel (2000), and

$$\tau_{ij} := \mathbb{E}[z_{ij}|x_j] = \frac{p_i f(x_j;\xi_i)}{\sum_{k=1}^{m} p_k f(x_j;\xi_k)} \tag{4.83}$$

is the posterior probability that x_j belongs to the i^{th} component of the mixture. We would like to choose the model whose complete form has the largest log-likelihood value $L^c(\mathbf{x};\boldsymbol{\gamma})$. To estimate the complete log-likelihood function, we could use (4.82). The term $L_N(\mathbf{x};\boldsymbol{\gamma})$ is estimated by $L_N(\mathbf{x};\widehat{\boldsymbol{\gamma}}_N)$. Since we do not know the matrix z_{ij}, we approximate $\log(k_N(\mathbf{x};\boldsymbol{\gamma}))$ by its expectation, given by

$$\mathbb{E}[\log(k_N(\mathbf{x};\mathbf{z};\boldsymbol{\gamma}))|\mathbf{x}] = \sum_{i=1}^{m}\sum_{j=1}^{N} \tau_{ij}\log(\tau_{ij})$$

The posterior probabilities τ_{ij} can be estimated using the MLE $\widehat{\boldsymbol{\gamma}}_N$ of $\boldsymbol{\gamma}$:

$$\widehat{\tau}_{ij} := \frac{\widehat{p}_i f(x_j;\widehat{\xi}_i)}{\sum_{k=1}^{m} \widehat{p}_k f(x_j;\widehat{\xi}_k)}. \tag{4.84}$$

The model is able to clearly classify the observations according to their components, if the posterior probabilities clearly indicate from which component each, or at least most, observations originated.

Define

$$EN(\widehat{\boldsymbol{\tau}}) = -\sum_{i=1}^{m}\sum_{j=1}^{N} \widehat{\tau}_{ij}\log(\widehat{\tau}_{ij}). \tag{4.85}$$

This motivates the *classification likelihood information criterion, CLC*

$$CLC(\mathbf{x};\widehat{\boldsymbol{\gamma}}_N) := -2L_N(\mathbf{x},\widetilde{\boldsymbol{\gamma}}_N) + 2EN(\widehat{\boldsymbol{\tau}}). \tag{4.86}$$

The size of the penalty factor $EN(\widehat{\boldsymbol{\tau}})$ depends on how well the model is able to classify the observations. If, for a particular observation x_j, the estimated posterior probabilities $\widehat{\tau}_{ij}$ are large for one particular component and close to zero for all other components, we say that this observation has been clearly classified. Since the terms $\widehat{\tau}_{ij}\log(\widehat{\tau}_{ij})$ are close to zero, if the corresponding values of $\widehat{\tau}_{ij}$ are close to zero or one. Therefore, if most observations can be clearly classified, the penalty factor will be small. If the number of components in the model is either too small or accurate, most observations should be clearly classified, as the components are clearly separated. We observed this in most instances when applying the von Mises-Fisher model to directional data. However, if the model has too many components, the observations cannot be clearly classified. The posteriori probabilities $\widehat{\tau}_{ij}$ of a large number of observations may be significant for more than one component. In that case, many of the terms $\widehat{\tau}_{ij}\log(\widehat{\tau}_{ij})$ will not be close to zero and the size of the penalty term can be considerable. The CLC states that an additional component should only be added to the mixture model, if the decrease in the clarity of the classification of the points is not greater than the increase in the log likelihood function. One of the drawbacks of the CLC that we observed is that it is not even monotone in the number of components. As a consequence, if the CLC of a model with $m+1$ components is larger than the one of the model with m components, there is no guarantee that model that minimizes the CLC has more than $m+1$ components. A study done by Biernacki et al. (1996) states that the CLC tends to overestimate the correct number of components. For these reasons, we usually did not consult the CLC when deciding the number of components, but rather worked with an improved version, called the ICL-BIC.

Integrated Classification Likelihood Criterion

The *Integrated Classification Likelihood Criterion, ICL,* attempts to improve the shortcomings of the CLC. In the following we give a brief outline of the motivation for the ICL, essentially found in McLachlan and Peel (2000) and Biernacki and Govaert (2000).

Define the integrated classification likelihood as

$$f_{icl}(\mathbf{x},\mathbf{z}) = \int f_N^c(\mathbf{x};\mathbf{z};\boldsymbol{\gamma})f_p(\boldsymbol{\gamma})d\boldsymbol{\gamma},$$

where $f_N^c(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$ denotes the complete likelihood function given by

$$f_N^c(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma}) = \prod_{j=1}^{N} p_i^{z_{ij}} \cdot f_i(x_j; \xi_i)^{z_{ij}}$$

and $f_p(\boldsymbol{\gamma})$ is a prior density on the model parameter $\boldsymbol{\gamma}$. Assume that the prior density can be factorized as

$$f_p(\boldsymbol{\gamma}) = f_{p1}(\mathbf{p}) f_{p2}(\boldsymbol{\xi}),$$

where $\mathbf{p} = (p_1, ..., p_m) \in \mathcal{P}$ denotes the vector of the component weights and $\boldsymbol{\xi} = (\xi_1, ..., \xi_m) \in \Xi$ is the vector of the parameters of the component densities in a finite mixture with m components. f_{p1} and f_{p2} denote the respective prior densities. In that case the integrated likelihood function $f_{icl}(\mathbf{x}, \mathbf{z})$ factorizes as

$$f_{icl}(\mathbf{x}, \mathbf{z}) = f_{icl}(\mathbf{x}|\mathbf{z}) f_{icl}(\mathbf{z}), \tag{4.87}$$

where

$$f_{icl}(\mathbf{x}|\mathbf{z}) = \int_{\Xi} f_N^c(\mathbf{x}, \boldsymbol{\xi}|\mathbf{z}) f_{p2}(\boldsymbol{\xi}) d\boldsymbol{\xi},$$

with

$$f_N^c(\mathbf{x}, \boldsymbol{\xi}|\mathbf{z}) = \prod_{j=1}^{N} f(x_j; \xi_i)^{z_{ij}}$$

and

$$f_{icl}(\mathbf{z}) = \int_{\mathcal{P}} \left(\prod_{j=1}^{N} p_i^{z_{ij}} \right) f_{p1}(\mathbf{p}) d\mathbf{p}.$$

Biernacki and Govaert (2000) assume that the prior distribution $f_{p1}(\mathbf{p})$ is the Dirichlet distribution $D(\alpha_1, ..., \alpha_m)$, given by density

$$f_D(\mathbf{p}) = \Gamma \left(\sum_{i=1}^{m} \alpha_i - m \right) \prod_{i=1}^{m} p_i^{\alpha_i - 1} \Gamma(\alpha_i)^{-1}.$$

with $\Gamma(x)$ denoting the Gamma function. They work with $\alpha_1 = ... = \alpha_m = \alpha$ and show that under these assumptions, we have that

$$\log(f_{icl}(\mathbf{z})) \approx K(N_1, ..., N_m). \tag{4.88}$$

In the above, $N_i = \sum_{j=1}^{N} z_{ij}$; $i = 1, ..., m$, are the number of observations in the i^{th} component and $K(N_1, ..., N_m, \alpha)$ is the function

$$\begin{aligned} K(N_1, ..., N_m, \alpha) &= \sum_{i=1}^{m} \log(\Gamma(N_i + \alpha)) - \log(\Gamma(N + m \cdot \alpha)) \\ &\quad -m \log(\Gamma(\alpha)) + \log(\Gamma(m \cdot \alpha)). \end{aligned}$$

They also show that the following approximation holds for $f_{icl}(\mathbf{x}|\mathbf{z})$:

$$f_{icl}(\mathbf{x}|\mathbf{z}) \approx \max_{\boldsymbol{\xi}} \log(f_N^c(\mathbf{x}, \boldsymbol{\xi}|\mathbf{z})) - \frac{d_1}{2} \log(N), \tag{4.89}$$

where d_1 is the total number of the parameters except $\widehat{p}_i$, $i = 1, ..., m$. McLachlan and Peel (2000) note that if we estimate the unknown matrix $\mathbf{z}$ with $\widehat{\boldsymbol{\tau}}$, we have that

$$\max_{\boldsymbol{\xi}} \log(f_N^c(\mathbf{x}, \boldsymbol{\xi}|\mathbf{z})) = L_N(\mathbf{x}, \widehat{\boldsymbol{\gamma}}_N) - EN(\widehat{\boldsymbol{\tau}}) - N \sum_{i=1}^{m} \widehat{p}_i \log(\widehat{p}_i), \tag{4.90}$$

where $\widehat{p}_i$ is the MLE for the weights of the components of the mixture model and $EN(\widehat{\boldsymbol{\tau}})$ is as in (4.85). Combining (4.88) to (4.90) we have from (4.87) that

$$\begin{aligned} \log(f_{icl}(\mathbf{x}, \mathbf{z})) &\approx L_N(\mathbf{x}, \widehat{\boldsymbol{\gamma}}_N) - EN(\widehat{\boldsymbol{\tau}}) - N \sum_{i=1}^{m} \widehat{p}_i \log(\widehat{p}_i) \\ &\quad -d_1/2 \log(N) + K(N\widehat{p}_1, ..., N\widehat{p}_m), \end{aligned} \tag{4.91}$$

where $\widehat{p}_i$ is the MLE for the weights of the components of the mixture model, d_1 is the total number of the parameters except $\widehat{p}_i$, $i = 1, ..., m$ and $EN(\widehat{\boldsymbol{\tau}})$ is as in (4.85). This motivates the following definition of the *Integrated Classification Likelihood Criterion, ICL*

$$\begin{aligned} ICL(\mathbf{x}, \widehat{\boldsymbol{\gamma}}_N) &:= -2L_N(\mathbf{x}, \widehat{\boldsymbol{\gamma}}_N) + 2EN(\widehat{\boldsymbol{\tau}}) + 2N \sum_{i=1}^{m} \widehat{p}_i \log(\widehat{p}_i) \\ &\quad +d_1 \log(N) - 2K(N\widehat{p}_1, ..., N\widehat{p}_m). \end{aligned} \tag{4.92}$$

We see that the ICL incorporates elements from the CLC as well as from the BIC. Biernacki and Govaert (2000) derived the following approximation to (4.92), based on

Stirling's formula and therefore only valid, when the terms $N\widehat{p}_i$ are large. It is referred to as the *ICL-BIC criterion*:

$$\textit{ICL-BIC}(\mathbf{x};\widehat{\boldsymbol{\gamma}}_n) := -2L_N(\mathbf{x},\widehat{\boldsymbol{\gamma}}_N) + 2EN(\widehat{\boldsymbol{\tau}}) + d\log(N), \tag{4.93}$$

where d is the total number of parameters in the model. We see that the ICL-BIC combines the penalty terms from both the BIC and the CLC. Biernacki et al. (1996) report that the performance of the ICL-BIC differs little from the ICL, even if the estimated cluster sizes $N\widehat{p}_i$ are not large.

Even though the ICL-BIC is also not necessarily concave in the number of components of the model, it performs much better than the CLC. We saw that the growth of BIC term $d\log(N)$ outweighed the fluctuations of the CLC term, $2EN(\widehat{\boldsymbol{\tau}})$, as the number of components increased. Therefore we observed that for most datasets considered, the ICL-BIC is a concave function in the number of components in the model. We therefore worked with the easier ICL-BIC rather than the CLC.

McLachlan and Peel (2000) report an empirical study, comparing the performance of the criteria introduced in this section. They used multivariate normal distributions as the component distributions. The study concludes that only the ICL and the ICL-BIC are able to correctly pick the right number of components for the three different datasets they considered. The AIC, and to a lesser extend, the CLC as well as the BIC overestimated the complexity of the model. However, our situation is very different from the one considered in McLachlan and Peel (2000). Not only are we considering a different class of distributions as component distributions, we also consider distributions on a different space. McLachlan and Peel (2000) consider distributions in $\mathbb{R}^d$, we are considering distributions on $\mathbb{S}^{d-1}$. Therefore the criteria may perform very different than McLachlan and Peel (2000) reported and their results may not be valid. For these reasons, we conducted our own empirical study. We present our results and conclusions in the following section.

4.4.4 Empirical Comparison of the LR Test and the Information Criteria

In order to compare the different information criteria and the likelihood ratio test procedure, introduced in the last sections, we conducted an empirical study. We simulated datasets from 6 different settings of dimension, number of components and sample size. In all instances, data from a finite mixture of von Mises-Fisher distributions was generated. For each of the 6 choices we created 5 to 10 datasets. For each dataset, we proceeded to calculate the maximum likelihood estimates via the EM algorithm. We usually started by estimating the parameters of a 2 component model. We then proceeded to repeatedly increase the number of components in the fitted model by 1, until the information criteria and the likelihood ratio procedure indicated that we had passed the optimal number of components. Starting values for the EM algorithm were usually obtained by the method of adding a components, described in Section 4.3.2 or from randomized starting points. The sample size was typically between 200 and 500, as those were the sample sizes that we worked with for real life datasets.

Case 1: A 6 component mixture on $\mathbb{S}^2$

The true parameters of the model are given in Table 4.1. The left table shows the true parameters of the model considered, while the right table gives an overview of the performance of the criteria considered. The mean direction of each component is given in spherical coordinates by $\alpha \in [0, 2\pi)$ and $\beta \in [0, \pi]$. κ denotes the concentration parameter of the component and p lists the weight of the components. Components

Table 4.1: *Overview of case 1 of the simulation study.*

The true parameters of the model

	α	β	κ	p
1)	0	7/10π	20	.15
2)	π/2	3/4π	60	.05
3)	1	π/2	10	.30
4)	4	2	24	.10
5)	5	1	30	.15
6)	4.5	π/2	10	.25

The number of components

Dataset Number:	1	2	3	4	5
AIC:	9	7	8	7	7
BIC:	6	6	5	6	6
ICL-BIC:	5	5	6	5	5
LR Test 5$\%$:	9	7	6	7	7
LR Test 1$\%$:	8	6	6	7	7

4 through 6 are not very well separated, whereas the first three components are fairy well separated. This is made clear in Figure 4.1, which shows a plot of the density $f(\phi, \theta), \phi \in [0, 2\pi); \theta \in [0, \pi]$ of the distribution with parameters as in Table 4.1. We created 5 different datasets, each with a sample size of 500. The number of components, m, as estimated by the different criteria, for each dataset is given in the right portion of Table 4.1. We see that the AIC overestimates m in each dataset. The BIC estimates m correctly in 4 out of the 5 datasets, underestimating it by 1 only in the 3^{rd} dataset. The ICL-BIC also performs fairly well, although its estimate of m is correct only in dataset 3. But it only underestimates m by 1 in all other datasets. The likelihood ratio test seems to perform better than the AIC, but also has a tendency to overestimate the number of components. Testing at the high significance level of 1% improved the precision of the estimates of m, compared to testing at 5%. It reduced the number of components chosen by the likelihood ratio method by 1 component in both the first and the second dataset. The results are well in line with what other authors reported as far as the AIC goes. We

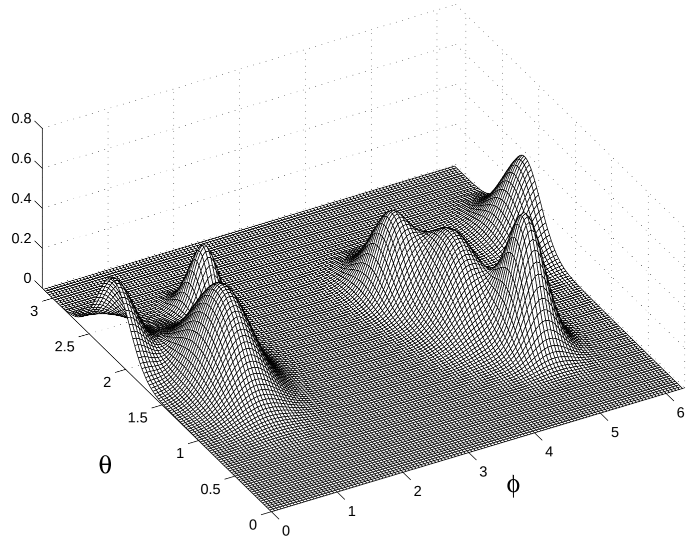

Figure 4.1: *The density of the von Mises mixture distribution from which the datasets of case 1 were created.*

found the rather poor performance of the likelihood ratio test disappointing, since the likelihood ratio test procedure was given theoretical justification in previous sections of this chapter, whereas the AIC, the BIC and the ICL-BIC lack this justification and were only considered because other authors mentioned in McLachlan and Peel (2000) had commented on their usefulness.

Case 2: A 5 component mixture on $\mathbb{S}^4$

Dataset 2 had a higher dimension, but only 5 components. 5 datasets with a sample size of 300 each, were created. This is less than for the previous datasets. An overview over the true parameters of the distribution and the performance of the criteria is given in Table 4.2. The left side of the table shows the true parameters of the model considered, while the right hand side table gives an overview of the performance of the criteria considered. The mean direction for each component is again given in spherical coordinates, represented by the angles $\alpha \in [0, 2\pi)$ and $\beta_i \in [0, \pi], i = 1, ..., 3$. It is nearly impossible to get a good impression of the shape of the distribution, since even in spherical coordinates, its density has a 4 dimensional domain. We studied 2 and 3 dimensional scatter plots of the datasets. It appears from those plots that the first 3 components are fairly well separated from each other, while the last two seemed to be closer together.

Table 4.2: *Overview of case 2 of the simulation study.*

The true parameters of the model

	α	β_1	β_2	β_3	κ	p
1)	0	1.5	1.5	1.5	10	.30
2)	1	1	1.5	2	50	.10
3)	4	3	2	2	20	.25
4)	5	2.5	1	1	10	.30
5)	5	2	1.5	1.5	100	.05

The number of components

Dataset Number:	1	2	3	4	5
AIC:	5	5	5	5	6
BIC:	5	5	5	5	5
ICL-BIC:	5	5	5	5	5
LR Test 5$\%$:	5	5	5	5	6
LR Test 1$\%$:	5	5	5	5	6

The performance of the different criteria is amazingly good. Both the ICL-BIC and the BIC estimate the correct number $m = 5$ in each dataset. The likelihood ratio test and the AIC both overestimate m in the last dataset, but provide a correct estimate of the number of components otherwise as well. A possible explanation is that the components are sufficiently separated so that each of them is clearly recognizable in a sample of the size considered here. Therefore, a model with less than 5 components will omit at least one of those components, resulting in a much lower log likelihood value compared to the 5 component model. This makes the 5 components model significant compared to a model with a lesser number of components. On the other hand, since each of the components is simulated from a von Mises-Fisher distribution, it is very hard to fit a model with 6 components and a significantly higher log-likelihood value to the dataset.

In order to succeed, we would need to fit two von Mises-Fisher components to a subset of the data representing one component. Since that subset was simulated from a single von Mises-Fisher distribution, this is unlikely to produce a large increase in the value of the log-likelihood function. The criteria never saw such a 6 component model as significant over the 5 component model, with the exception of the AIC and the likelihood ratio test procedure in the last dataset.

Case 3: A 4 component mixture on $\mathbb{S}^1$

Dataset 3 is a mixture model with 4 components in only 2 dimension. We created 10 datasets, each with a sample size of 400. The reason we created 10 rather than 5 datasets is that an earlier analysis of a similar model had not been conclusive enough based on only 5 different datasets. The true parameters of the model are found in Table 4.3. The

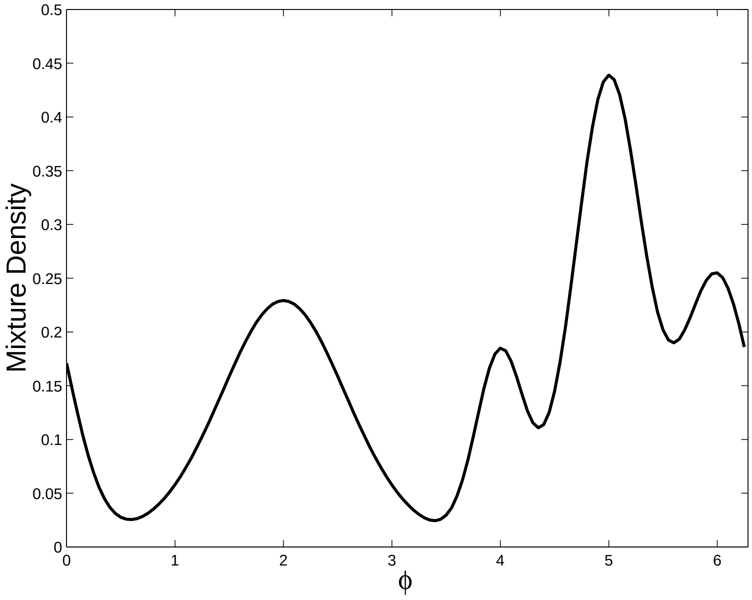

Figure 4.2: *The density of the von Mises mixture distribution from which the dataset of case 3 were created.*

Table 4.3: *The true parameters used in case 3.*

	α	κ	p
1)	2	3	.35
2)	5	10	.35
3)	4	20	.10
4)	6	10	.20

Table 4.4: *The number of components estimated in case 3.*

Dataset Number:	1	2	3	4	5	6	7	8	9	10
AIC:	5	5	5	6	4	4	4	4	4	4
BIC:	4	4	4	2	2	4	3	3	2	3
ICL-BIC:	4	2	4	2	2	2	2	2	2	2
LR Test 5%:	5	5	4	3	3	4	4	4	3	4
LR Test 1%:	5	4	4	2	2	4	4	3	3	4

density of the mixture is shown in Figure 4.2. We can see that the first component is clearly separated from the other three components. These three other components are not very well separated, but they are still clearly distinguishable.

The AIC is able to correctly estimate the number of components, $m = 4$, for 6 out of the 10 datasets. In the other instances it overestimates the number of components, in the case of the 4^{th} even by two components.

The BIC on the other hand shows a tendency to underestimate the number of components. Only in 4 out of the 10 datasets is it able to correctly estimate $m = 4$. For 3 dataset it even settles for 2 components, not being able to distinguish components 2,3 and 4. In the other 3 cases it picked a model with 3 components, because it was not able to clearly distinguish the last 3 components as well.

As the ICL-BIC has an ever greater penalty term, the underestimation of m is is even more severe. With the exception of two datasets, the ICL-BIC is not able to see that there are 4 rather than just 2 components.

The results for the likelihood ratio test are mixed. For both the significance level of 5% and 1%, we see instances where m is overestimated and instances where it is underestimated. The likelihood ratio procedure introduced in Algorithm 4.4.13 performs better here when using the lower significance level of 5%. For both the 5% and the 1% significance level the likelihood ratio test estimates m correctly for 5 of the 10 datasets. But the underestimation for datasets 4 and 5 is again severe, as only significant 2 components are identified. Overall, the AIC and the likelihood ratio test at 5% seemed to perform best here.

Case 4: A 6 component mixture on $\mathbb{S}^3$

Case 4 is a mixture with 6 components in 4 dimension. We created 5 datasets, each with a sample size of 300. The true parameters are found in the left hand side of Table 4.4, while the right hand side table gives an overview of the performance of the criteria considered. Since the distribution is on the 4 dimensional unit sphere, it is hard to determine to what degree the components are separated. However, looking at the parameter values, it appears that components 4 and 5 might not be clearly separated. Because their second spherical coordinate is close to π, the difference in the first spherical coordinate does not mean the points are far apart. In cartesian coordinates their mean direction are given by $(0.1140, -0.0332, -0.8330, 0.5403)$ and $(-0.1176, 0.0168, -0.8330, 0.5403)$, respectively. The other components are very different in at least one coordinate. We confirmed this idea by looking at 3 dimensional scatter plots of the simulated datasets.

Table 4.5: *Overview of case 4 of the simulation study.*

The true parameters of the model

	α	β_1	β_2	κ	p
1)	2	2	2	6	.25
2)	4	2	2	10	.15
3)	5	1	1	20	.10
4)	3	3	1	10	.20
5)	6	3	1	5	.20
6)	2	0.5	2	20	.01

The number of components

Dataset Number:	1	2	3	4	5
AIC:	7	5	5	5	5
BIC:	5	5	5	5	5
ICL-BIC:	5	5	5	5	5
LR Test 5%:	8	5	5	5	5
LR Test 1%:	7	5	5	5	5

Looking at the results of the estimation of m by the various criteria, we see that, expect for the first dataset, all criteria considered incorrectly estimate $m = 5$. In the first dataset the AIC and the likelihood ratio tests clearly overestimate m, while the BIC and the ICL-BIC still estimate $m = 5$. This is most likely due the fact that the criteria were not able to separate components 4 and 5. The 3 dimensional scatter plots that we considered also indicated that the choices of the concentration parameters, $\kappa = 5$ and $\kappa = 10$, respectively lead to fairly far spread out components. Additionally, as me mentioned above, the mean directions are very similar.

Case 5: A 10 component mixture on $\mathbb{S}^2$

Case 5 was motivated by the study of the spectral measure of the log returns of the three stocks IBM, Intel and Apple, see Section 5.1. The parameter values in the right table of Table 4.6 are the parameter of a 10 component von Mises-Fisher mixture model fitted to the spectral measure of the distribution of the daily log returns of the three stocks. See Section 5.1 for details. The right table gives an overview of the performance of the criteria considered. Notice that components 1, 4, 6, 7, 9 and 10 have a very high concentration parameter κ. Those components are very closely concentrated about their mean direction. Those mean directions turn out to be the axis directions. For example, the mean direction of the first component in cartesian coordinate is $\mu_1 = (0.9999, 0.0072, 0.0098)$, which is almost the direction of the x-axis pointing in positive direction. Similarly, components 4, 6, 7, 9 and 10 have mean directions that closely follow one of the axis. Compared to those 6 components, the remaining components 2,3,5 and 8 are fairly spread out. We present a contourplot below in Figure 4.3. We see that while the 6 highly concentrated components are very well separated from each other, components 1 and 2 and components 5 and 6 are not very well separated. The components 3 and 8 appear isolated, but they are so far spread out, that points from those components might get mixed with points from other components.

Table 4.6: *Overview of case 5 of the simulation study.*

The true parameters of the model

	α	β_1	κ	p
1)	0.01	1.56	329.8	0.09
2)	0.17	1.47	39.7	0.05
3)	0.78	0.32	13.2	0.10
4)	1.31	0.01	575.4	0.09
5)	1.35	1.45	20.4	0.09
6)	1.55	1.54	491.5	0.09
7)	3.20	1.61	188.5	0.10
8)	3.87	2.29	3.8	0.19
9)	4.07	3.09	477.3	0.09
10)	4.67	1.62	85.7	0.11

The number of components

Dataset Number:	1	2	3	4	5
AIC:	11	10	10	11	11
BIC:	10	10	10	10	10
ICL-BIC:	9	10	10	8	9
LR Test 5%:	11	10	10	11	10
LR Test 1%:	10	10	10	11	10

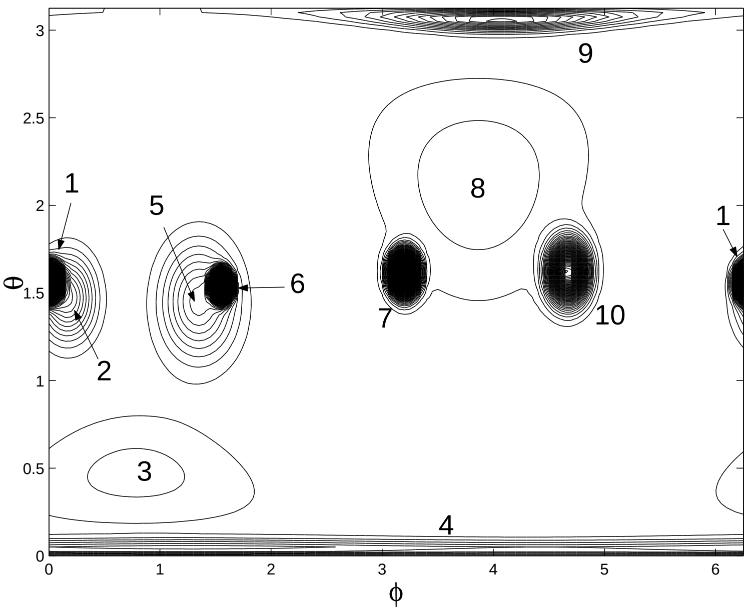

Figure 4.3: *A contour plot of the mixture discussed in case 5.*

Overall, the criteria do a good job of estimating m. The AIC overestimates m in three out of the five datasets by one component. The BIC is flawless and gets the correct estimate $m = 10$ for every dataset. The ICL-BIC seems to be penalizing too harshly and therefore underestimates m in exactly those datasets where the AIC is overestimating it. For the 4^{th} dataset it even claims that a 8 component mixture is the best model. The likelihood ratio tests perform better than AIC, but they also overestimate m. Comparing the model with $m = 10$ with the model with $m = 11$, the p value of the likelihood ratio test statistic under H_0 for the first dataset was 1.58%. Therefore, at 5%, the 11^{th} component is significant, while at 1% it is not.

Conclusions

None of the criteria that we considered performed flawless. We found that there is no single criterion that outperforms the others and should therefore given clear preference.

The BIC showed the most consistent performance, especially on the dataset on $\mathbb{S}^2$ and in higher dimensions. It showed a tendency to underestimate the number of components in $\mathbb{S}^1$, as is made clear in case 3. Since ICL-BIC has a greater penalty than the BIC, it's tendency to underestimate the number of components was even more pronounced. As for the BIC, the performance improved with the growing number of the dimension in the dataset. While the AIC performed as well as the other criteria in case 3, it showed a tendency to overestimate the number of components in higher dimensions. It was almost always the criterion that selected the largest number of components in each mixture. The algorithm 4.4.13, based on the likelihood ratio test had a fairly consistent performance in case 3. However it showed a tendency to overestimate the number of components in the higher dimensional cases. We conclude that the likelihood ratio test procedure should rather be used with the significance level of 1%, rather than the customary 5%. This helps reduce the danger of overestimating the complexity of the model. Based on our observations, it seems reasonable to use the likelihood ratio test procedure for datasets in $\mathbb{S}^1$. For the datasets of higher dimension we recommend considering the BIC as the preferred choice for determining the number of components.

In the next chapter we describe the results of fitting mixture models to various financial datasets. We see a much greater disagreement about the optimal number of components in the model indicated by the various criteria. Based on the results of our empirical study, we mostly consulted the BIC and the likelihood ratio test procedure with 1% significance to decide on the complexity of the model. However we also considered other factors and aspect of the various models as explained in the next chapter.

Chapter 5

Analysis of Datasets

In this chapter we present the results of modelling the spectral measure of several different financial time series with finite von Mises-Fisher mixture models. In each case, we first calculated the log returns of each of the time series. The log returns of a time series $X_1, ..., X_n$ are defines as

$$R_i = \log(X_{i+1}) - \log(X_i); i = 1, ..., n-1. \tag{5.1}$$

We obtained a non-parametrical estimate of the spectral measure of the log returns by means of the ranks method, introduced in Section 2.2.4. Recall from (2.50), that an observation is chosen by the ranks method, if and only if

$$R_{j,k} > 1,$$

where $R_{j,k}$ is the norm of $\frac{k}{\mathbf{r_j}}$ and $\mathbf{r_j} = (r_j^{(i)}, i = 1, ..., d)$ is the vector of the ranks of the observation $(X_j^{(1)}, ..., X_j^{(d)})$. The non-parametric estimate of the spectral measure consists of the angular components $\boldsymbol{\theta}_{j,k}$ of the points $\frac{k}{\mathbf{r_j}}$ chosen by the ranks method. We determined the number k, denoting the number of upper order statistics, with the help of the Stărică plot, explained in Section 2.2.3. We refer to observations that get selected by the ranks method as extreme observations. These are the observations that we use in the estimation of the parameters of a parametric model for the spectral measure. In this chapter, we discuss the results of fitting a von Mises-Fisher mixture model to the points $\boldsymbol{\theta}_{j,k} \in \mathbb{S}^{d-1}$. The number of components was determined with the help of the criteria introduced in Section 4.4.

5.1 Log Returns of IBM, Intel and Apple

The dataset under consideration consists of the daily closing prices of the stocks of IBM, Intel and Apple between 1/1/1986 and 10/6/2000. In our analysis we work with the time series of the log returns of these prices. The resulting dataset contained 3612 daily log returns for each of the three stocks.

5.1.1 Preliminary Analysis of the Spectral Measure

We started our analysis by estimating the tail indexes of the log returns with the Hill estimator and the QQ-estimator. We obtained the estimates presented in Table 5.1. These values are fairly typical for financial data. It is usually assumed that the tail indexes for financial time series are between 2 and 4. Based on our estimates, we created and studied Stărică plots. We determined that $k = 80$ is an acceptable choice for the purpose of estimating the spectral measure. We used the ranks method with this value and found that 424 observations were chosen for the estimation of the spectral measure. Figure 5.1 shows a scatter plot of the points that are selected by the ranks method. The plot shows the directional arguments of the selected points. That is, each point in Figure 5.1 gives the angular part $(\phi_{j,k}, \theta_{j,k}) \in [0, 2\pi) \times [0, \pi]$ of the spherical coordinates of a point $(R_{j,k}, (\phi_{j,k}, \theta_{j,k}))$ with $R_{j,k} > 1$. This can be seen as a non-parametrical estimate of the spectral measure. For 257 of the 424 points selected, the corresponding log returns all

Table 5.1: *The estimates of the tail indexes of the log returns of the three stocks considered in this section.*

IBM:	Right Tail	3.5	INTEL	Right Tail	4.0	APPLE	Right Tail	3.2
	Left Tail	2.8		Left Tail	3.0		Left Tail	3.0

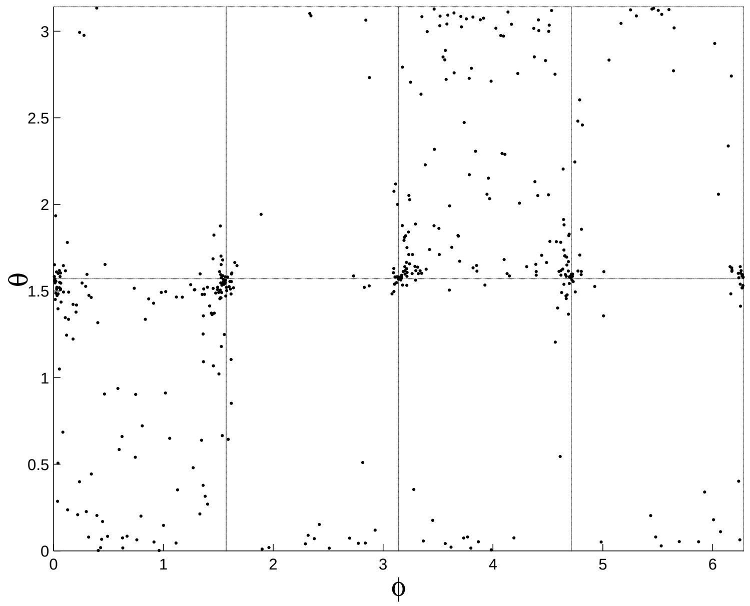

Figure 5.1: *Estimate of the spectral measure of the joint distribution of the daily log returns of the stock prices of IBM, Intel and Apple. See the text for more details.*

had the same sign. This is a first indicator that there is dependence among the extreme observations. A significant number of observations that are extreme, consist of returns that are either all positive or all negative. We see an indication of this in Figure 5.1 by the points in the areas $(\phi, \theta) \in [0, \pi/2] \times [0, \pi/2]$ and $(\phi, \theta) \in [\pi, 3\pi/2] \times [\pi/2, \pi]$. The number in those areas is significantly larger than the number of points in the other areas. The points in $[0, \pi/2] \times [0, \pi/2]$ represent observations where all three returns were positive, while the points in $[\pi, 3\pi/2] \times [\pi/2, \pi]$ represent the observations with negative returns.

We note that the a significant portion of the points is close to one of the following points: $(\phi_1, \theta_1) = (0, \pi/2)$, $(\phi_2, \theta_2) = (\pi/2, \pi/2)$, $(\phi_3, \theta_3) = (\pi, \pi/2)$, $(\phi_4, \theta_4) = (3\pi/2, \pi/2)$. We also see that two more clusters are grouped around $\theta_5 = 0$ and $\theta_6 = \pi$. These six coordinates represent the axes of the cartesian coordinate system.

$(\phi_1, \theta_1) = (0, \pi/2)$ represents the point $(1, 0, 0)$. Points that lie close to that point are observations for which the log-return of IBM is extreme and positive, whereas the corresponding returns of Intel and Apple are comparatively moderate. Similarly, points close to $(\phi_2, \theta_2) = (\pi/2, \pi/2)$, which corresponds to the point $(0, 1, 0)$, correspond to observations for which the return of Intel is extreme and positive, while the returns for IBM and Apple are moderate. The interpretation of the other 4 clusters is similar.

It also appears that a significant portion of the points are located near one of the planes $\{(\phi, \theta) : \phi \in \{0, \pi\}\}$, $\{(\phi, \theta) : \phi \in \{\pi/2, 3\pi/2\}\}$ and $\theta = \pi/2$. These are the planes that are spanned by either two of the axes of the cartesian coordinate system. The points close to those planes correspond to observations where two of the three stocks have a extreme return, while the third one only has a moderate one. The plane represented by $\theta = \pi/2$ is referred to as the ”IBM-Intel” plane, since it contains the observations for which only the returns of IBM and Intel were extreme. Similarly, the plane $\{(\phi, \theta) : \phi \in \{0, \pi\}\}$ is referred to as the ”IBM-Apple” plane and the plane $\{(\phi, \theta) : \phi \in \{\pi/2, 3\pi/2\}\}$ is referred to as the Intel-Apple plane in Figure 5.1.

We created a program in an attempt to separate points close to an axis and points close one of the planes mentioned above. First, observations that are closer to one of the axes than a certain tolerance are filtered out. From among the remaining points we then filter out the ones that are closer to one of the planes than a second tolerance. This gives us a preliminary picture of the structure of the dependence. No choice of the tolerances can be the only correct one. If they are chosen too small, not enough points will be deemed as close to an axis or a plane. Clusters around the axes would still be visible after removing the points selected as being close to the axes. If on the other hand, the tolerance is too large, points that are not really close to an axis will be included in that category. We tried several different values and found that it was reasonable to consider a point as close to an axis, if, in cartesian coordinates, one of his coordinate had a value of greater than 0.99. We considered a point that was not close to an axis, as close to a plane, if, in cartesian coordinates, the absolute value of one of his three components was not larger than 0.1.

We found that 216 observations were close to an axis and 118 additional points were close to a plane. 90 points were of full dimension. Figure 5.2 shows the separated dataset.

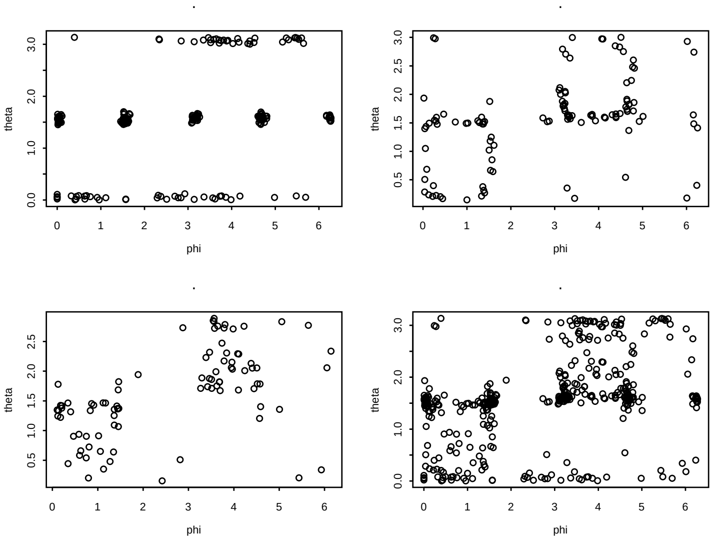

Figure 5.2: *Top left: Scatter plot of the observations close to an axis. Top right: Scatter plot of the observations close to a plane. Bottom left: Scatter plot of the points that are neither close to a plane nor close to a plane. Bottom right: The full dataset representing the spectral measure of the log returns of IBM, Intel and Apple. This is the same plot as Figure 5.1.*

Recall from Section 2.2 that the components of a random vector are said to be asymptotically independent, if the corresponding spectral measure concentrates on the axes. For such a distribution, extreme observations do not happen in more than one coordinate at the time. For the dataset of this section we more than half of the points close to an axis, but we see that there is also a significant number of points far away from an axis. These points correspond to observations where either two or all three of the log returns are extreme. From this preliminary analysis we have strong evidence that the log returns of IBM, Intel and Apple are not asymptotically independent.

Extreme changes in stock prices, positive and negative, are usually caused by some type of shock in the economy. Examples of these shocks are news about the company, the industry that the company operates in, or the economy of the US. Some shocks affect a large number of stocks at the same time, while others affect only certain stocks at a time. An increase or decrease in the federal interest rate or new numbers on the US economy would however affect most stocks at the same time. Certain news might be related to a certain industry, thus only affecting companies in that industry. IBM, Intel and Apple are companies that operate in similar, but not the same industry. This helps to explain some of the structure that we see in the spectral measure of the three stocks. As explained before, points that are close to an axis refer to observations where only one of the three stocks experienced a extreme return. These observations could have been caused by events or news only concerning that particular company. Other observations reflect shocks that affected more than one of those companies. If there are no shocks that affect more than one company at the same time, they would be asymptotically independent. The corresponding spectral measure would be concentrated on the axes. We see that this does not seem to be the case.

In the following, we will make our claim that the three stocks are not asymptotically independent, more precise by fitting a von Mises Fisher mixture model to the points selected by the ranks method. If the log returns of IBM, Intel and Apple would be asymptotically independent, a model with 6 components would provide an adequate fit.

The 6 components would have mean directions that, expressed in cartesian coordinates, approximately equal $(1, 0, 0)$, $(-1, 0, 0)$, $(0, 1, 0)$, $(0, -1, 0)$, $(0, 0, 1)$ and $(0, 0, -1)$. They would also have fairly large concentration parameters. Such a mixture distribution would be an approximation to a distribution on $\mathbb{S}^2$ that concentrates all its mass on the axes. Since our preliminary analysis, based on Figures 5.1 and 5.2, indicates that is not the case, we are not surprised that we need a more complex model to describe the spectral measure.

5.1.2 A von Mises-Fisher Model of the Spectral Measure

Based on the preliminary analysis, we decided that a 6 component mixture model was the simplest model that we fitted to the data. We then continued to consider more complicated models. For each given number of components we determined the (quasi) maximum likelihood estimates. We checked that the estimates do not describe spurious components. We compared the models of increasing complexity using the criteria explained in Section 4.4. We present an overview of the values of the criteria in Table 5.2. Each entry represents the value of the corresponding criteria for the best model with the corresponding number of components. The highlighted values indicate the estimate of m by each criterion. We see that the criteria indicate that the number of components of the model, m, is between 10 and 13. The ICL-BIC gives the smallest estimate, $m = 10$. The values of the BIC indicate that m equals 11. Note however, that value of the BIC for the 11 component model, 887.09, is only slightly lower than the corresponding value for the 10 component model, reported as 888.11. The BIC therefore states that the 11 component model is just barely more significant than the 10 component model. The likelihood ratio test procedure and the AIC both indicate that we need 13 components. In addition, we see that the value of the AIC for 13 components is just barely lower than

Table 5.2: *Overview of the model selection criteria.*

# of components	AIC	BIC	ICL-BIC	P value LR test
6	1140.5	1233.6	1282.5	-
7	966.22	1075.6	1168.4	0.28%
8	799.73	925.27	1036.5	0.27%
9	778.24	919.98	1056.8	0.043%
10	730.17	888.11	**1023.65**	0.084%
11	711.95	**887.09**	1041.85	0.063%
12	701.49	891.83	1088.8	0.153%
13	**684.09**	890.63	1074.85	**0.107%**
14	684.45	907.18	1089.8	6.3689%

the one for 14 components. At the same time, the value for 12 components is significantly higher. Similarly, the likelihood ratio test comparing the models with 12 and 13 components clearly rejects the null hypothesis that data has a mixture distribution with only 12 components. The corresponding p-value is about 0.1%. The p-value of the test comparing the models with 13 and 14 components is also just above the 5% threshold. This indicates, that there is some evidence there may be even more than 13 components.

The results are in line with what we would expect from our empirical study in Section 4.4.4. We had seen that the AIC and the likelihood ratio test tend to return larger estimates for the number of components. However, the difference between the estimates was less significant compared with what we observe here. This is due to the fact that, in the empirical study, the data actually had a von Mises-Fisher mixture distribution. The dataset under consideration here is real life data and we cannot expect that its distribution is a von Mises-Fisher mixture distribution.

In the empirical study we had concluded that the BIC and the likelihood ratio test procedure with a significance level of 1% are the most consistent criteria. For the IBM-Intel-Apple dataset they pick two different models, the BIC chooses the one with 11 components and the likelihood ratio test the one with 13 components. We noted in the empirical study, that if the dimension is higher than 2, the likelihood ratio test procedure showed a tendency to overestimate the number of components. The BIC on the other hand, showed a very consistent performance. We are therefore inclined to rely on the BIC rather than on the likelihood ratio test. Before we make that decision, we want to compare the two selected models. The parameter estimates of the model with 11 components is given in Table 5.3. The mean direction is given in spherical coordinates $(\phi, \theta) \in [0, 2\pi) \times [0, \pi]$. The column "Points" indicates how may points belong to each component. The first six components of the model have mean directions that are very close to the six axes points on the unit sphere. Each of these components also has a large concentration parameter. These six components describe the clusters of points around the axes, that we detected in the preliminary analysis. The remaining five components describe the remainder of the data. With the exception of component 11, they have a much smaller concentration parameter κ than the first six components. A closer look at the components reveals that component 7 is fairly close to component 3, component 9 is close to component 5 and that component 11 is close to component 6. Essentially, these components are adding more structure to the modelling of the clusters around the axes. The structure of those clusters seems to be too complicated to be described by a single von Mises-Fisher component. Component 10 models the points that represent the observations where all three log returns are positive. Component 8 models the points representing the observations where all three log returns are negative. It is the presence of these two components in both the models with 11 and 13 components, that allow us

Table 5.3: *Parameter estimates for the model with 11 components. See text for details.*

Component	Mean Direction		κ	weight	Points
1	(1.3062,	0.0048)	575.51	0.08529	38
2	(4.6713,	1.6188)	85.80	0.10996	50
3	(0.0071,	1.5610)	329.49	0.09430	43
4	(4.0730,	3.0861)	478.07	0.08712	38
5	(1.5521,	1.5401)	492.37	0.08627	39
6	(3.1486,	1.5727)	699.51	0.05306	24
7	(0.1673,	1.4691)	39.91	0.05568	21
8	(3.9103,	2.3116)	3.82	0.18647	73
9	(1.3506,	1.4467)	20.32	0.08988	35
10	(0.7845,	0.3195)	13.21	0.09741	40
11	(3.2628,	1.6766)	144.40	0.05453	23

to state that, based on our model, the log returns of the stocks are not asymptotically independent.

Recall from the definition of the CLC and ICL-BIC criterion, that a finite mixture model allows us to calculate the posterior probability that a particular point belongs to a particular component. The posterior probabilities are given by (4.84). It turns out that the points in this dataset can be clearly classified this way. We found that for every point, there is one component for which the posterior probability is greater than 0.5. We use these probabilities to classify the points according to the corresponding component. The last column in Table 5.3 shows how many points can be associated this way with each component. Figure 5.3 shows a scatter plot of the points, grouped by their respective component. All points belonging to the same component are pictured in the same color and style. The number next to the group gives the number of the corresponding component in Table 5.3 We now turn our attention to the model selected by the likelihood ratio tests and the AIC. The parameter estimates of the 13 components are given in Table 5.4. The mean direction is given in spherical coordinates $(\phi, \theta) \in [0, 2\pi) \times [0, \pi]$. The column “Points” indicates how may points belong to each component. “Difference” lists how many points each component has lost or gained compared to the 11 component model. The 13 component model is essentially an extended version of the 11 component model. The first 11 components are very similar to the components of the smaller model. The most significant change of the parameter estimates occurs in the 8^{th} component. We also note that this component now only has 30 points attributed to it. It used to have 73 points associated with it in the model with 11 components. A scatter plot, similar to Figure 5.3, reveals that the points that used to be associated with component 8 are now associated with 3 different components in the 13 component model. A more detailed analysis revealed the following passing of points between components:

Table 5.4: *Parameter estimates for the model with 13 components. See text for details.*

Component	Mean Direction		κ	weight	Points	Difference
1	(1.2975,	0.0048)	577.7734	0.0851	38	
2	(4.6915,	1.5879)	349.9370	0.0642	32	-18
3	(0.0100,	1.5591)	339.9207	0.0932	43	
4	(4.0438,	3.0880)	530.8783	0.0815	37	-1
5	(1.5519,	1.5393)	508.6555	0.0853	39	
6	(3.1452,	1.5715)	700.6945	0.0513	23	-1
7	(0.1508,	1.4963)	27.4700	0.0606	22	+1
8	(4.2573,	2.8729)	14.2850	0.0763	30	-43
9	(1.3637,	1.4546)	19.4902	0.0920	36	+1
10	(0.7855,	0.3120)	13.1067	0.0979	40	
11	(3.2644,	1.6313)	391.0939	0.0379	16	-7
12	(3.4617,	1.8748)	11.8493	0.0932	37	+37
13	(4.5716,	1.7170)	19.6770	0.0816	31	+31

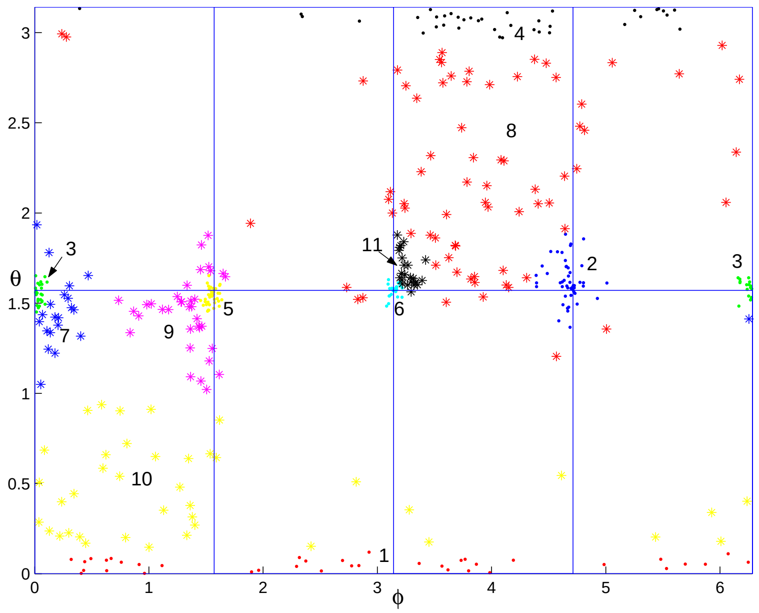

Figure 5.3: *Classification of the points according to the posterior probabilities (4.84) using the 11 components of the mixture model with parameters given in Table 5.3.*

-The 12^{th} component contains 29 points that were in the 8^{th} component and 8 points that were in the 11^{th} component in the smaller model.

-The 13^{th} component has 13 points that were in the 8^{th} component and 18 points that were in the 2^{nd} component in the smaller model.

-The 11^{th} component contains one point that was in the 6^{th} component, the 7^{th} and 9^{th} components now each contain a point that was in the 8^{th} component, which in turn acquires a point from the 4^{th} component.

The split of component the 8^{th} by adding two components in its neighborhood is deemed significant by the AIC and the likelihood ratio test, but not by the BIC. To answer the question of whether 11 or 13 components are needed to accurately model the data, we need to decide whether the 8^{th} component of the smaller model is sufficient to

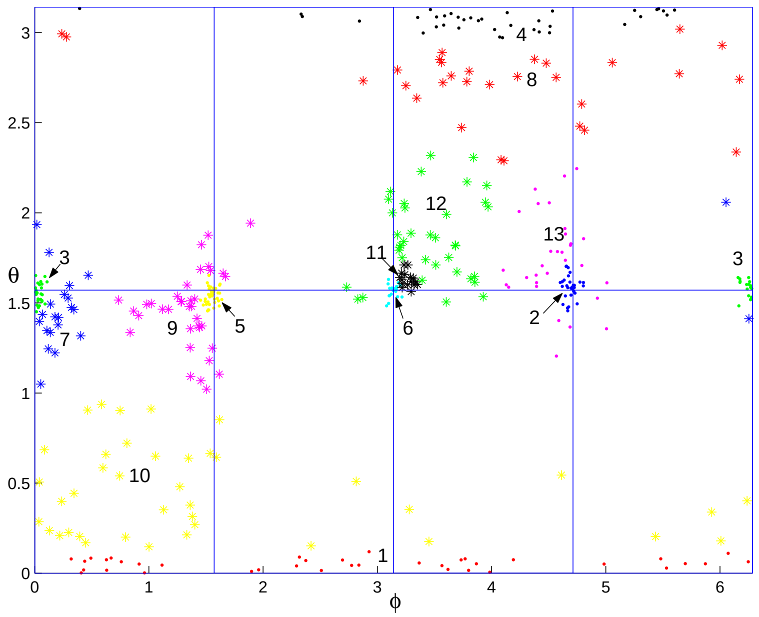

Figure 5.4: *Classification of the points using the 13 components of the mixture model with parameters given in Table 5.4.*

describe the dependence in the area representing extreme negative returns of all three stocks. The criteria and our more careful analysis do not indicate a clear and objective answer. We may note that in the area representing simultaneous extreme positive returns of all three stocks, one component was sufficient. This can be seen as a motivation to conclude that an 11 component mixture model with the parameters given in Table 5.3 is an adequate description of the spectral measure. However, the different clusters have different structures. This may result in a more complex model for one cluster than for another one, thus favoring the model with 13 components.

We want to stress that despite having a different number of components, the two model are fairly similar in their description of the spectral measure. Both models acknowledge the presence of clusters around the six axes. Both models acknowledge the presence of dependence among extreme negative and extreme positive returns. They differ in how to describe the dependence between extreme negative returns. In our opinion both models offer a valid and insightful option to describe the tail dependence among the log returns among the three stocks. Both models could be used to develop a holistic model of the distribution of the three stocks that could for example be used in assessing the risk of a portfolio of these stocks.

5.2 Log Returns of IBM and Intel

The dataset used in this analysis is the same as in the previous section. However, we concentrate on the daily log returns of the two stocks of IBM and Intel only, thus ignoring the log returns of Apple.

Before we describe the results of our analysis of this data with the help of our von Mises-Fisher mixture model, we want to consider the following question: How does the spectral measure of the joint distribution of IBM, Intel and Apple compare to the one of the distribution of IBM and Intel? Is there an easy way to obtain a consistent estimate of the spectral measure of the returns of IBM and Intel from the corresponding estimate of the spectral measure of IBM, Intel and Apple?

Recall that we had chosen Apple to be the third coordinate in the previous section. We could therefore just use the first angular component $\phi_{j,k}$ of the points $(R_{j,k}, (\phi_{j,k}, \theta_{j,k}))$, selected by the ranks method, as an estimate for the spectral measure of IBM and Intel. However, in doing so, we would keep the points that correspond to observations that were chosen only because of the extreme return of Apple. The same observations would not be chosen when we're using the ranks method on the two dimensional dataset of the log returns of IBM and Intel. We would hence include too many observations and obtain a biased estimate of the spectral measure. Only by estimating the spectral measure using the ranks method (2.50) or the direct approach (2.53) from the dataset of

IBM and Intel can we consistently estimate the spectral measure.

We used the tail index estimates listed in Table 5.1 in the creation of Stărică plots. These plots indicated that $k = 80$ is an acceptable choice and the ranks method selected 302 points. Recall that for the three dimensional dataset in the last section we had also used $k = 80$, but obtained 424 points. This indicates that 122 points were selected only because of the extreme return of Apple in these observations. These observations naturally did not get selected by the ranks method run on the log returns of IBM and Intel only. We present a scatter plot of the angular part $\phi_{j,k} \in [0, 2\pi)$ of the polar coordinates of the points $(R_{j,k}, \phi_{j,k}); j = 1, ..., 302$ with $R_{j,k} > 1$ in Figure 5.5. We add a non - parametrical estimate of the corresponding density.

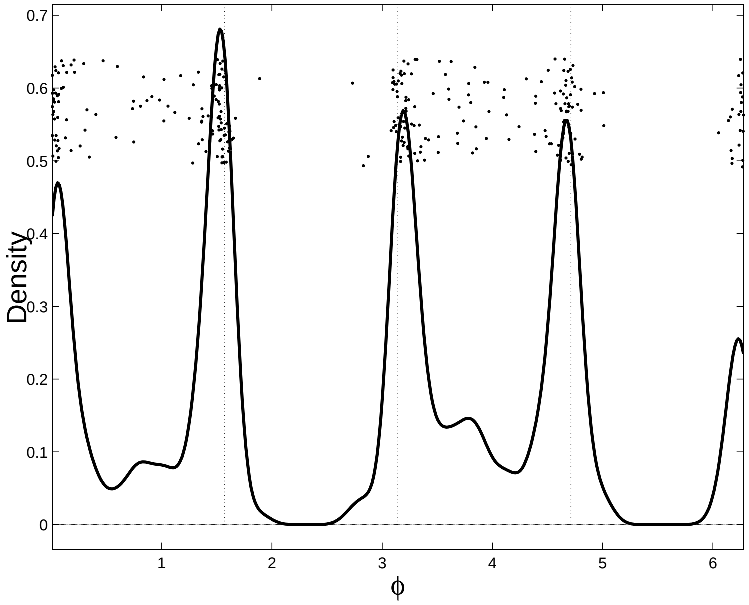

Figure 5.5: *A scatter plot of the points $\phi_{j,k}, j = 1, ..., 302$ selected by the ranks method and a non parametric estimate of the spectral measure of IBM and Intel.*

It is not a surprise to see 4 significant clusters, concentrated at the coordinates $\phi = 0 (= 2\pi), \pi/2, \pi$ and $3\pi/2$. Points in these clusters correspond to observations where only one of the two stocks has a extreme return. We also notice that there is a significant number of points with $\phi_{j,k} \in (0, \pi)$ or $\phi_{j,k} \in (\pi, 3\pi/2)$. We will refer to the area $(0, \pi/2)$ as the first quadrant and to the area $(\pi, 3\pi/2)$ as the third quadrant. Points in these areas correspond to observations where both the returns of IBM and Intel's stocks were large. This shows that there is a good chance that extreme positive returns as well as extreme negative returns of IBM and Intel occur at the same time. This is a clear indication that the returns of the two stocks are not asymptotically independent. Furthermore, we see that there are basically no points with $\phi_{j,k} \in (\pi/2, \pi)$ and $\phi_{j,k} \in (3\pi/2, 2\pi)$. These areas are referred as the second and fourth quadrant, respectively. The fact that we see no points in these quadrants means that extreme negative returns of IBM and extreme positive returns of Intel (and vice-versa) do not occur at the same time.

Similar to our analysis in the previous section, we fitted a sequence of von Mises mixture models with increasing complexity to the points selected by the ranks method. An overview over the value of the criteria estimating the appropriate number of components is given in Table 5.5. Each entry represents the value of the corresponding criteria for the best model with the corresponding number of components. Highlighted are the estimates of m by each criterion. As usual, the ICL-BIC chooses the smallest number of components. In this case it picks a 5 component model. That model captures the 4 clusters close to one of the axes as well the structure visible in the third quadrant. However, the non-parametric density plot in Figure 5.5 indicates that a sixth component modelling the data in the first quadrant is needed. The ICL-BIC showed a serious tendency to underestimate the number of components of a von Mises mixture model of similar sample size in the empirical study. We therefore dismiss the suggestion of the ICL-BIC and concentrate on the other criteria. The BIC suggests a 6 component mix-

Table 5.5: *Overview of the model selection criteria.*

# of components	AIC	BIC	ICL-BIC	P value LR test
2	1073.6	1092.1	1194.8	-
3	838.68	868.37	929.96	0.38%
4	756.85	797.67	805.33	0.14%
5	669.94	721.88	**783.10**	0.14%
6	611.50	**674.58**	801.5	0.10%
7	606.19	680.40	850.59	**0.29**%
8	**602.26**	687.6	894.81	1.55%
9	603.14	699.61	935.60	6.37%

ture model whose parameter estimates are presented in Table 5.6. The mean direction is given in polar coordinates $\phi \in [0, 2\pi)$. The column "Points" indicates how may points are associated with each component. The model includes 4 components close to an axis and one component for the points representing the observations were both log returns are positive and negative, respectively. It thus includes the component that we were

Table 5.6: *Parameter estimates for the model with 6 components. See text for details.*

Component	Mean Direction	κ	weight	Points
1	0.0203	189.0630	0.1900	59
2	1.5263	194.1920	0.1974	63
3	3.1922	217.5353	0.1548	51
4	4.6760	216.1783	0.1527	49
5	0.8102	4.4682	0.1188	31
6	3.8834	3.7299	0.1863	49

missing in the 5 component model.

The likelihood ratio test procedure with the 1% significance level indicates a 7 component model whose parameters are given in Table 5.7. The mean direction is given in polar coordinates $\phi \in [0, 2\pi)$. The column "Points" indicates how may points belong to each component. The column "Difference" lists how many points each component has lost or gained compared to the 6 component model. Compared to the 6 component model, a new component, close to component 1, has been added. The posterior probabilities indicate that 24 points are attributed to that new component. Most of those points belonged to component 1 before. This models implies that the structure of the points with values of $\phi_{j,k}$ close to $0 (= 2\pi)$ should not be modelled by a single von Mises distribution. Instead, a second component is needed.

Table 5.7: *Parameter estimates for the model with 7 components. See text for details.*

Component	Mean Direction	κ	weight	Points	Difference
1	0.0193	690.9928	0.1155	44	-15
2	1.5370	281.6753	0.1792	58	-5
3	3.1922	217.5969	0.1548	51	-
4	4.6760	215.9547	0.1528	49	-
5	1.1129	8.2541	0.1037	27	-4
6	3.8829	3.7424	0.1863	49	-
7	0.0755	39.7497	0.1076	24	+24

The AIC and the likelihood ratio test procedure with a 5% significance level indicate a model with 8 components. Table 5.8 lists the corresponding parameter estimates. While the model with 7 components added complexity to the modelling of the dependence structure of extreme positive returns, the model with 8 components adds to the

Table 5.8: *Parameter estimates for the model with 8 components. See text for details.*

Component	Mean Direction	κ	weight	Points	Difference
1	0.0193	690.9928	0.1155	44	-
2	1.5371	283.0930	0.1790	58	-
3	3.1871	250.4561	0.1448	49	-2
4	4.6761	691.1801	0.0900	35	-14
5	1.1151	8.1819	0.1045	27	-
6	3.6220	6.8114	0.1488	38	-11
7	0.0753	39.7523	0.1077	24	-
8	4.6417	32.6825	0.1097	27	+27

complexity of the dependence for extreme negative returns. It adds a component very close to component 4. It contains 27 points, most of which come from components 4 and 6. It has thus a similar role and interpretation as component 7 does.

As was the case for the models considered in the previous sections, there does not appear to be a single correct model. All the three models with 6, 7 or 8 components are valid models for the spectral measure of the two log returns of the two stocks. We recall from the empirical study in Section 4.4.4, that for bivariate data the BIC has a tendency to underestimate the number of components. On the other hand, the study indicated that the AIC and the likelihood ratio test with a 5% significance level tend to overestimate the number of components. We are therefore tend to favor the proposition of the likelihood ratio test with a 1% significance level that a model with 7 components accurately describes the spectral measure of the log returns of IBM and Intel.

We can use the model of the spectral measure to show that the points selected by the ranks method fall into two categories. The first category contains the points in components 1 through 4 in the model. If we work with a model that contains 7 or 8 components, the points in components 7 and 8 also fall into the first category. These points correspond to observations where only one of the two stocks showed an extreme return. If those were all the points selected by the ranks method, we would have strong evidence to conclude that extreme returns of IBM and Intel do not occur at the same time and that the stocks are therefore asymptotically independent. It is the presence of the points in the second category that shows that there is indeed tail dependence between the two stocks. These are the points in components 5 and 6. These two components together contain 76 out of the total of 302 points, that is 25.6%. This means that about one out of four extreme observations of the vector of the returns of IBM and Intel is caused by simultaneous extreme returns of the two stocks. For the majority of the 302 observations considered extreme, only one of the stocks had an extreme return. Nevertheless, the number of extreme observations were both stocks had a extreme return is significant. All the three mixture models analyzed in this section recognize these points by attributing two components to them. They therefore reject the notion that the two stocks are asymptotically independent.

5.3 Log Returns of BMW and Siemens

This dataset is available with the EVIS package for the SPLUS software. The software package is available at http://www.math.ethz.ch/∼mcneil/software.html. It consists of the daily closing prices for the stocks of BMW and Siemens from January 1973 to July 1996. The sample size of the dataset after calculating the log returns is 6146.

As for the previous dataset, we started our analysis by estimating the tail indexes of the right and left tail of the marginal distributions. The estimates of the tail index are similar in size to the estimates obtained for the dataset of the stock prices of IBM, Intel and Apple. Based on these estimates, given Table 5.9, we produced Stărică plots

Table 5.9: *The estimates of the tail indexes of the log returns of the daily closing prices of BMW and Siemens.*

BMW:	Right Tail:	3.5	Siemens:	Right Tail:	4.6
	Left Tail:	3.4		Left Tail:	3.2

to decide on an optimal value of k. We concluded that $k = 65$ was the best choice. The ranks method selected 225 observations in its estimation of the spectral measure. A plot of the selected points in polar coordinates together with a non-parametrical estimate of the corresponding density is given in Figure 5.6. Similar to the IBM-Intel case, we see that most of the points $\phi_{j,k}$ are concentrated in the first and third quadrant. Of the 225 points selected, 100 points were located in the first quadrant and 104 more were located

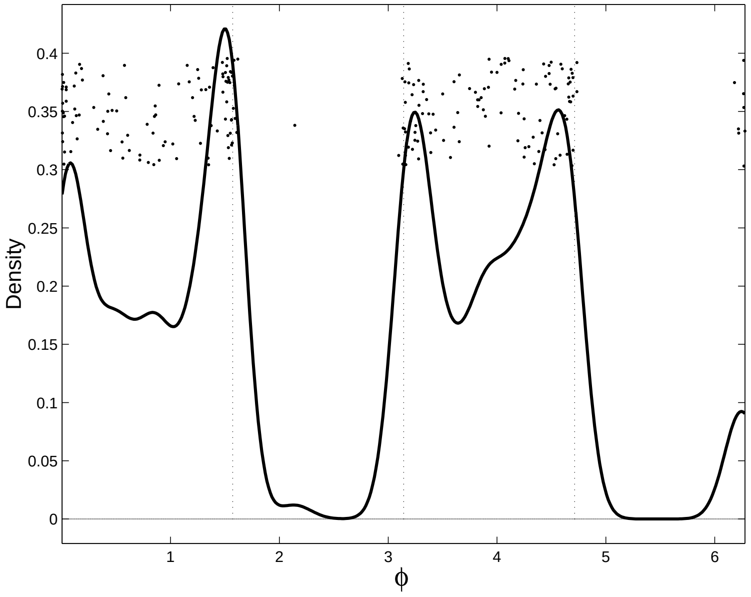

Figure 5.6: *A scatter plot of the points* $\phi_{j,k}, j = 1, ..., 225$ *selected by the ranks method and a non parametric estimate of the spectral measure of BMW and Siemens.*

in the third quadrant. Only 21 points were found in the second and fourth quadrant, mostly close to one of the axis points $\phi = 0$, $\phi = \pi/2$, $\phi = \pi$ and $\phi = 3\pi/2$. We also see that the points seem to form clusters around the axis points. The clusters seem to be less pronounced compared with the IBM-Intel case. This is an indication that the dependence between extreme positive returns of both stocks or extreme negative returns of both stocks is stronger than in the case of IBM and Intel.

We proceeded to fit a von Mises-Fisher mixture model to the points selected by the ranks method. The values of the criteria considered for estimating the correct number of components are given in Table 5.10. Each entry represents the value of the corresponding criteria for the best model with the corresponding number of components. We highlighted the values indicating the optimal number of components chosen by the corresponding criterion. Based on the non-parametrical estimate of the spectral measure in Figure 5.6, we do not think that a model with less than 6 component will adequately describe the spectral measure. However, we wanted to confirm this intuition. Therefore,

Table 5.10: *Overview of the model selection criteria.*

# of components	AIC	BIC	ICL-BIC	P value LR test
2	734.2737	751.3542	768.9404	0
3	663.9369	691.2657	708.8292	0.1193
4	609.5200	647.0971	673.5731	0.0931
5	550.2912	598.1166	**648.2170**	0.1032
6	530.6834	**588.7571**	671.6614	**0.1259**
7	**525.7542**	594.0762	688.8775	2.4840
8	526.5636	605.1339	708.4305	**4.4340**
9	529.4050	618.2236	722.5192	8.5537

we estimated the parameters of models with a smaller number of components. The estimates for the number of components that we thus obtained were very similar to the case of IBM-Intel. Again, the ICL-BIC gives the smallest estimate. It indicates that 5 components are enough. The BIC and the likelihood ratio test procedure with the 1% significance level both estimate that 6 components are needed. The estimate of the AIC of the number of components is 7 and the likelihood ratio test with a significance level of 5% even returns an estimates of 8 components. Since, as mentioned before, the BIC and the likelihood ratio test procedure at 1% are the criteria we trust most, we are inclined to conclude that a mixture model with 6 components is the optimal choice. The parameter estimates of the model with 6 components are given in Table 5.11. The mean direction is given in polar coordinates $\phi, \in [0, 2\pi)$. The column "Points" indicates how may points are associated with each component. As expected, there are four components, numbered 1-4 in Table 5.11, whose mean directions are close to the axis points $\phi = 0$, $\phi = \pi/2$, $\phi = \pi$ and $\phi = 3\pi/2$. Each of those components has a large concentration parameter, indicating that the component is very narrowly concentrated around the mean direction. Component 5 models the dependence in the first quadrant while component 6 models the

Table 5.11: *Parameter estimates for the model with 6 components. See text for details.*

Component	Mean Direction	κ	weight	Points
1	0.0375	216.2548	0.1417	35
2	1.5404	693.8882	0.1264	31
3	3.2333	139.3197	0.1388	33
4	4.6104	125.9612	0.1179	29
5	0.8883	5.7989	0.2431	49
6	4.0226	9.2655	0.2321	48

dependence in the third quadrant. As mentioned before, the model with 6 components is the smallest model that we are willing to accept after studying Figure 5.6. Similar to the case of IBM and Intel, the more complicated models with 7 and 8 components, respectively, add components close to one of the four components modelling points close to an axis. Since both the likelihood ratio test at 1% and the BIC do not consider the additional components as significant, we decided to work with the simpler model with 6 components.

As in the case of IBM and Intel, the model of the spectral measure allows us to categorize the points selected by the ranks method. The points in components 1 through 4 represent observations where only one the two stocks experienced a extreme return. We see from Table 5.11 that 128 of the 225 points belong to one of those components, while 97, or 43.1% of all points, belong to either component 5 or 6. Remember that for the spectral measure of IBM and Intel, we concluded that only 25.6% of the extreme observations were due to simultaneous extreme returns of both stocks. This indicates that the dependence between extreme events seems to be stronger for BMW and Siemens than it is for IBM and Intel. The spectral measure of BMW and Siemens is less concentrated around the axes than the one of IBM and Intel. It is to a larger degree concentrated in the first and third quadrant. This indicates, that if we have an extreme observation, in the sense that it gets selected by the ranks method, there is a larger probability that both stocks are affected than in the case for IBM and Intel.

5.4 Log Returns of Foreign Currencies

5.4.1 Preliminary Analysis

The dataset contains the daily exchange rates of five foreign currencies to the US $ from June 1973 to May 1987. The currencies are the British Pound (BP), the Canadian Dollar (CD), the German Mark (DM), the Swiss Franc (SF), and the Japanese Yen (JY). The time frame is well before the rates of the currencies replaced by the Euro were irrevocably fixed. The resulting dataset containing the log returns of the exchange rate had 3508 observations for each currency. We expect to see different dependence structures for different pairs of the currencies. The DM, SF and BP are currencies of European countries. We can expect a fairly close dependence among the returns of these currencies, since a lot of the underlying factors driving the exchange rates will be the same for all three currencies. On the other hand, the dependence between the CD and the JY will probably be much weaker. The two countries are on separate continents and therefore the factors underlying the exchange rates of the two currencies are fairly different. We will analyze the tail dependence of the five exchange rates by studying their spectral measure. We also take a closer look at selected pairs of the 5 currencies. The estimation and the analysis of the spectral measure of all five exchange rates turned out to be very difficult because of what we refer to as the ”curse of dimensionality”. The spectral measure is a measure that lives on $\mathbb{S}^4$. Even in polar coordinates it is a measure with a 4 dimensional domain. Fitting a parametric model to a selection of points is a formidable task. We discuss the problems that arose and present possible solutions for this problem.

As in the previous sections, we started by estimating the tail indexes of the marginal distributions. As before we used the Hill estimator and the QQ-estimator. The estimates of the tail indexes of the log returns of the daily exchange rates of the five exchange rates are listed in Table 5.12. The values in brackets represent alternate estimates that are also justifiable from both the Hill plots and the QQ estimator. It is important to point out, that no single estimate for a tail index can be considered the only correct one. Other estimates of the tail indexes could also be justified based on the Hill plots that we studied. This is of importance, because the Stărică plots depend on the estimates of the tail indices. For the bivariate distributions considered in the previous datasets this is only a moderate problem. We only have to estimate four different tail indices. A different choice for one or two of these estimates results in only small changes of the Stărică plots. For the foreign currencies we found that the range of possible estimates for each tail index is larger than for the tail indexes of the stocks. Additionally, we now have to estimate 10 different tail indexes. For the values presented in Table 5.12, the Stărică plots indicate that $k = 35$ or maybe even $k = 40$ are acceptable choices. The ranks method selects 287 points, if $k = 35$ is used and 318 points, if $k = 40$ was used. However, for the alternative values of the tail indexes, given in the brackets in Table 5.12, we found that Stărică plots indicate that $k = 15$ and maybe $k = 20$ are

Table 5.12: *The estimates of the tail indexes.*

BP:	Right Tail:	3.4 (3.5)	CD:	Right Tail:	3.1 (3.4)
	Left Tail:	3.8 (4)		Left Tail:	3.0 (3)
DM:	Right Tail:	4.5 (4)	JY:	Right Tail:	4.2 (4.5)
	Left Tail:	3.5 (4)		Left Tail:	3.75 (4)
SF:	Right Tail:	4.75 (5)			
	Left Tail:	3.4 (3.5)			

acceptable values. For the following analysis we chose to be conservative and decided to use $k = 20$. This way, we felt safe that we would not introduce a bias in the estimate of the spectral measure by using too many observations. We may, however, have omitted numerous observations that could have been included. As a result of the final choice of $k = 20$, 170 observations were chosen by the ranks method. As before, we refer to these observations as extreme observations.

We saw in the analysis of the bivariate stock data in the previous sections, that most points representing the spectral measure were either in the first or the third quadrant. That is, for most observations chosen by the ranks method, either both returns were positive or both returns were negative. We observed something very similar for the points representing the spectral measure of the five exchange rates.

- 49 points correspond to observations where the returns of all five exchange rates are positive.
- 44 points correspond to observations where the returns of all five exchange rates are negative.
- 20 points correspond to observations where the return of the CD is negative and the return of the other 4 exchange rates is positive.
- 20 points correspond to observations where the return of the CD is positive and the return of the other 4 exchange rates is negative.

The remaining 37 points were spread out over various of the other 28 possible ”quadrants”, that is, combinations of positive and negative returns of the different exchange rates. The fact that the majority of the points represent observations where the returns of all currencies are extreme is a first indication that there is tail dependence among the exchange rates of the five currencies.

In a next step, we tried to separate points who correspond to extreme returns of only one, two, three or four of the exchange rates. We used the same procedure as in the identification of points near an axis in the data of IBM, Intel and Apple, see Section 5.1. We are aware that this procedure is fairly crude. Nevertheless, it gives us important insights in the structure of the tail dependence of the different exchange rates. We call the return of an exchange rate extreme, if the corresponding observation was primarily selected because of the return of that particular exchange rate. That is, if a point is near an axis associated with positive returns of the Swiss Franc, we call the corresponding return of the Swiss Franc extreme. If the data point is near the axis spanned by the Swiss Franc and the British Pound axes, we call the corresponding returns of the Swiss Franc and the British Pound extreme.

We found that 72 of the 170 points correspond to an extreme return of only one exchange rate. They can be categorized as follows:

- 28 of those 72 points are due to extreme returns in the CD,
- 17 points are due to extreme movements of the JY,
- 11 points are due to extreme returns of the SF,
- 8 points are due to extreme returns of the BP,
- 7 points are due to extreme returns of the DM.

We observed only 15 points where two of the five exchange rates have an extreme return. 10 of those points come from the pair (DM, SF). The pair (BP, DM) contributes 2 points, while the pairs (BP, CD), (BP, SF) and (CD, SF) each contribute one point.

We found 26 points for which three of the five exchange rates have an extreme return. 13 of those observations come from the triple (BP, DM, SF). 8 observations come from the triple (DM, JY, SF). 2 points come from the two triples (BP, CD, JY) and (CD, DM, SF) respectively. Finally, the triples (BP, JY, SF) and (CD, JY, SF) contribute on point each.

41 points can be attributed to extreme returns in all but one exchange rate. For 20 of those points, the CD is the exception, for 16 it is the JY, for 4 it is the BP and for 1 it is the SF. Finally, we observed 16 points were all 5 exchange rates have an extreme return.

Based on this analysis, it appears that the SF and the DM have the strongest tail dependence among the five currencies. This is evident from the fact 10 of the 15 points that are due to extreme returns of two exchange rates are from the pair (DM, SF). Moreover, 23 of the 26 points for which three of the five exchange rates have returns that are extreme, also contain the pair SF and DM. The exchange rates of the CD and the JY seem to have much less tail dependence with the exchange rates of the other currencies. An indication of this is that they are only responsible for a small number of the extreme observations involving extreme returns of more than one currency, compared to the SF, DM or the BP. For example, there is not a single point with extreme returns of only two exchange rates involving the JY, and only 2 such points involving the CD. On the other hand, in most cases where all but one of the five exchange rates were extreme, they were the exception.

5.4.2 The von Mises-Fisher Mixture Model and the Curse of Dimensionality

We attempted to fit a von Mises-Fisher mixture model to the points selected by the ranks method. As was the case for the case of the IBM-Intel-Apple, we observed that a significant number of points were only selected because of the extreme return of only one of its marginal components. These points appear as clusters close to an axis. In the case of IBM Intel and Apple, this lead us to the conclusion that we need a model with at least 6 components. For the dataset of the exchange rates, we observed 10 clusters, one around each of the points representing the axes of the cartesian coordinate system on $\mathbb{S}^4$. For this reason we cannot expect a model with less than 10 components to be an accurate description of the spectral measure. We proceeded to increase the number of components. However it soon became clear that a much larger number of components is needed to obtain an adequate description of the spectral measure. The values of the criteria that we use to estimate the correct number of components is given in Table 5.13. Each entry represents the value of the corresponding criteria for the best model with the corresponding number of components. The criteria indicate that the best number of components is at least 15, because the criteria achieve the smallest value for the model with 15 components. However, recall that we are only using 170 points for our estimates. As for the previously studied datasets, we classified the points according to what component they are associated with, using the posterior probabilities (4.84). Already for the model with only 12 components, we saw that 2 components only had 6 and 5 points associated with them, respectively. For the model with 15 components,

Table 5.13: *Overview of the model selection criteria.*

# of components	AIC	BIC	ICL-BIC	P value LR test
10	-235.9677	-50.9555	-43.5588	
11	-282.7970	-78.9701	-72.3036	0.0868
12	-331.1082	-108.4665	-100.1527	0.0992
13	-351.9816	-110.5252	-104.4242	0.0984
14	-404.4313	-144.1600	-136.2359	0.0801
15	-426.3127	-147.2267	-140.7077	0.0836

we saw that three of those components had less than 5 points associated with them. 4 more components had less than 10 points associated with them. In our opinion, it is senseless to try to estimate the mean direction and the concentration parameter of a von Mises-Fisher component, based on less than 10 points. On the other hand, based on our preliminary analysis and the values in Table 5.13, we do not believe that a model with even 12 components is an accurate description of the spectral measure of the dataset.

This is what we referred to as the ”curse of the dimensionality” in the introduction to this section. The structure that the 4 dimensional data representing the spectral measure of all five currencies exhibits is very complicated. There are several small clusters of points scattered on $\mathbb{S}^4$, especially around the points of the axes. A von Mises-Fisher mixture model sees most of these clusters as significant and attributes a component to them. This results in a model with a large number of components, even is the sample size is rather small.

As we saw, the problem is already very challenging for a dataset of 5 different risk factors. For datasets of even higher dimension, we expect that problem to be even worse. We suggest two possible solutions to this problem.

On one hand we could work with a dataset with more observations. This can be achieved by working with data of higher frequency. Instead of using daily log returns, we could use hourly or data of even higher frequency. This would dramatically increase the sample size and hence allow us to consider more observations for the estimation of the spectral measure. Since the daily log returns are an aggregation the hourly log returns, the questions arises under what conditions the spectral measures of the different log returns are the same. The answer to that question is found in Hauksson et al. (2001).

They consider a high frequency process $(\mathbf{X}_k) \in \mathbb{R}^d$ and an aggregated process

$$\mathbf{Y}_i = \sum_{k=im}^{(i+1)m-1} \mathbf{X}_k$$

and prove the following result.

Theorem 5.4.1 *Let $(\mathbf{X}_k)$ be a stochastic process in $\mathbb{R}^d$, such that all $\mathbf{X}_k$ have the same distribution. Assume that the distribution is multivariate regular varying with tail index α. That is, we assume that*

$$\lim_{t\to\infty} \frac{\mathbb{P}[\|\mathbf{X}\| > tx, \|\mathbf{X}\|^{-1}\mathbf{X} \in A]}{\mathbb{P}[\|\mathbf{X}\| > t]} = x^{-\alpha} S_*(A).$$

for a finite measure S_* on $\mathbb{S}^{d-1}$. Let $(\mathbf{Y}_i)$ be as above. If the condition

$$\lim_{r\to\infty} \mathbb{P}[\|\mathbf{X}_i\| > r| \ \|\mathbf{X}_j\| > r] = 0, \text{ for } i \neq j \tag{5.2}$$

is satisfied, then $\mathbf{Y}_i$ is multivariate regular varying with tail index α and has the same spectral measure as $\mathbf{X}_k$.

Hauksson et al. (2001) furthermore argue in an empirical study that the bi-hourly and hourly returns of exchange rates of selected currencies seem to satisfy condition (5.2). Their study also indicates that 10 minutes and 30 minutes returns probably do not satisfy (5.2). Nevertheless this indicates a possibility to use higher frequency data to estimate the spectral measure. This would increase the number of points available for parameter estimation of a von Mises-Fisher mixture model or a similar model.

A second possibility is to try to show that certain marginal components of the dataset are asymptotically independent of the other components in the dataset. Assume for example that for the IBM-Intel-Apple dataset, we could have shown that the log returns of Apple are independent of the log returns of IBM and Intel. In that case the spectral measure of the three stocks would be concentrated on the set $\{(x, y, z) \in \mathbb{S}^2 : z \in$

$\{-1,0,1\}\}$. That is, the points of the spectral measure would be concentrated on the big circle of the equator $z = 0$ and the north and south pole of the sphere $\mathbb{S}^2$. We could then have fitted a lower dimensional model to the points describing the asymptotic dependence of IBM and Intel. In a dataset of higher dimension, this approach could prove very valuable. We could first identify the marginal components that are asymptotically independent of the other marginal components. These components could then be excluded from the dataset before attempting to estimate the spectral measure. For the dataset under consideration in this section, we can try to show that the JY or the CD are asymptotically independent of the other three currencies. We would then only have to estimate the spectral measure of the dataset of the three European currencies. This would greatly simplify the task of finding an adequate model of the spectral measure. Unfortunately, there are many problems that prevent us from doing this. Most importantly, there is to this date no statistical test for the asymptotic independence of two random variable available. Furthermore, it is often the case that there are no such independent marginal components. In the next section, we will argue that returns of the CD and the JY do not appear to be asymptotically independent of the returns of the other three currencies.

5.4.3 Are the CD and the JY Asymptotically Independent ?

When investigating asymptotical dependence or independence of the different marginal components of a random vector, it is enough to consider pairwise asymptotic independence. The reason is the following proposition, found in Resnick (1986).

Proposition 5.4.2 *Suppose* $\mathbf{X}$ *has a multivariate regular varying distribution with exponent measure* μ *concentrating on* $E := [-\infty,\infty]\backslash\{-\infty\}$. *The following are equivalent:*

1. *The components of* $\mathbf{X}$*, namely* $X^{(1)}, \ldots, X^{(d)}$ *are asymptotically independent random variables.*

2. *The components of* $\mathbf{X}$ *are pairwise asymptotically independent. For every* $1 \leq i < j \leq d$*,* $X^{(i)}$ *and* $X^{(j)}$ *are asymptotically independent random variables.*

Proof: See Resnick (1986). ■

This allows us to check whether the JY and the CD are asymptotically independent of the other currencies by checking pairwise asymptotical independence of these currencies. This is in contrast to checking ”classical” independence between random variables, where pairwise independence does not imply independence in general. In order to establish that the JY and the CD are not asymptotically independent of the other currencies, it is therefore enough to establish that they are not pairwise asymptotically independent of the other currencies.

We first focus the on returns of the JY. A preliminary analysis of non parametrical estimates of the spectral measures of the JY and the other currencies revealed that the tail dependence between the JY and the DM is weaker than the tail dependencies between the JY and the other currencies. Therefore, we especially focus on the relationship between the DM and the JY. If we find evidence against the hypothesis that the two currencies are not asymptotically independent, we also have evidence that the same is true for the JY and the other currencies.

Using the tail index estimates of Table 5.12 for the DM and the JY, we consulted Stărică plots to decide on an acceptable value of k. We concluded that $k = 60$ was the best choice. The ranks method selected 217 observations in its estimation of the spectral measure. We estimated the parameters of various von Mises-Fisher mixture models with different numbers of components. The BIC suggested a model with 6 components while the likelihood ratio test procedure suggests a model with 7 components, both for the 5%

and the 1% significance level. We decided to trust the likelihood ratio test procedure and hence conclude that the model with the parameter estimates in Table 5.14 is an adequate description of the spectral measure. The mean direction is given in polar coordinates $\phi \in [0, 2\pi)$. The column "Points" indicates how may points are associated with each component. The picture that emerges from studying the parameter estimates in Table 5.14 is fairly similar to what we have previously seen for bivariate data. We see four components with mean directions close $0, \pi/2, \pi, 3\pi/2$ and very large concentration parameters. These components describe the points close to an axis that correspond to observations where only one of the currencies experienced a extreme return. We also see three more components, containing a total of 65 points. Components 5 and 6 model the dependence between observations were both the DM and the JY had extreme positive returns. Component 7 models the dependence between observation were both currencies had extreme negative return. It is the presence of these significant components that leads us to reject the hypothesis that the DM and the JY are asymptotically independent.

Table 5.14: *Parameter estimates for the model with 7 components of the spectral measure of the log returns of DM and JY. See text for details.*

Component	Mean Direction	κ	weight	Points
1	0.0446	348.58	0.1271	30
2	1.5308	238.77	0.1780	40
3	3.1972	412.15	0.1568	36
4	4.6135	134.57	0.2075	46
5	0.2328	90.42	0.0918	18
6	1.0101	14.12	0.0953	19
7	3.8021	5.47	0.1425	28

We now turn our attention to the CD. By proceeding analogue to the analysis of the JY, we argue against the independence of the return of the CD by showing that its returns are not independent of the BP. We chose to investigate the pair of the CD and the BP because this was the pair that seemed to have the least tail dependence among all pairs involving the CD. If we find that our mixture model rejects the idea that the two currencies have asymptotically independent returns, this would give us confidence that the same is true for all the other pairs involving the CD as well.

Based on the tail index estimates in Table 5.12 we decided, by consulting Stărică plots, that $k = 60$ was a good choice. The ranks method selected 229 observations. After fitting von Mises-Fisher mixture models with several different number of components, we consulted the usual criteria to choose the best number of components. Almost all the criteria indicated that $m = 6$ is the best number of components. The only exception was the likelihood ratio test procedure with the 5% significance level. It indicated 7 components, as the p value of the test comparing the models with 6 and 7 components was 4.13%. Table 5.15 shows the parameter estimates of the model with 6 components.

Table 5.15: *Parameter estimates for the model with 6 components of the spectral measure of the log returns of BP and CD.*

Component	Mean Direction	κ	weight	Points
1	0.0635	123.3628	0.2262	52
2	1.5613	234.5120	0.1715	43
3	3.1957	188.3905	0.2268	53
4	4.7067	465.9547	0.1775	43
5	4.2586	4.1859	0.1024	20
6	1.1745	9.3842	0.0956	18

As before, we see four components modelling extreme returns by one, but not the other currency. We also see that there are two additional components, to which a total 38 points can be attributed. They describe the dependence of extreme simultaneous negative and positive returns of both currencies, respectively. Similar to the case of the JY and the DM, we see the fact that these components were deemed significant as strong evidence that the CD is not asymptotically independent from the other currencies.

Chapter 6

From the Spectral Measure to a Bivariate Distribution

In the previous chapter we discussed different examples of how we use the von Mises-Fisher mixture distribution as a model of the spectral measure of various datasets from finance. In this chapter, we present a model of the joint distribution of random variables, that is based on our model. We focus on modelling the dependence between the marginal components. The model of the dependence consists of two separate models, one that we refer to as the "model of the body of the distribution" and another one that we refer to as the "model for the tails of the distribution". We will concentrate on the description of the model for the tails, while we use a standard multivariate normal distribution as a model of the body. Other possible choices for the model of the body are briefly mentioned. The model of the tails uses what we call the "raw model". That model is based on von Mises-Fisher mixture model of the spectral measure to describe the tail dependence between the marginal components. We then combine this raw model with appropriate marginal distributions. In that sense, the raw model serves us like a copula. It focuses on the description of the tail dependence structure in the distribution, to which desired marginals can be attached. The chapter is organized as follows: We first present the raw model. Then we explain how the marginals of the raw model can be transformed to obtain a model with desired marginals. We present our model of the marginal distribution. Finally we show how we combine the model for the tails and the model for the body. We only describe the bivariate case, but higher dimensional extensions of our approach are straightforward. However, the notation would be much more complicated, which is the main reason that we restrict the discussion to the two dimensional case.

6.1 The Raw Model

The raw model is motivated by the following result, stated in Theorem 2.2.3 in Section 2.2.2. Let $\mathbf{X}_1$ be distributed as F_*, where F_* is as in Section 2.2.2. Let $(\mathbf{R}, \Theta) := (\|\mathbf{X}_1\|, \|\mathbf{X}_1\|^{-1}\mathbf{X_1})$. If we have that in $M_+((0,\infty] \times \aleph)$

$$t\mathbb{P}[(\frac{\mathbf{R}}{t}, \Theta) \in \cdot] \xrightarrow{\nu} r^{-2}dr \times S_*(d\theta), \tag{6.1}$$

then $F_* \in D(G_*)$, where

$$G_*(\mathbf{x}) = \exp(-\mu_*([\mathbf{0}, \mathbf{x}]^c))$$

and

$$\mu_*\{\mathbf{y} \in \mathbb{E} : \|\mathbf{y}\| > r, \|\mathbf{y}\|^{-1}\mathbf{y} \in A\} = r^{-1}S_*(A).$$

In the light of (6.1), let

$$s_0(\phi) = \sum_{i=1}^{m} p_i f_M(\phi; \alpha_i, \kappa_i)$$

be the density of a finite von Mises-Fisher mixture model in $d = 2$ dimensions with m components. The densities of the components are

$$f_M(\phi; \alpha_i, \kappa_i) = \frac{1}{2\pi I_0(\kappa_i)} e^{\kappa_i cos(\phi - \alpha_i)}, 0 < \phi \leq 2\pi, \kappa > 0, 0 \leq \alpha_i < 2\pi.$$

Definition 6.1.1 *The raw model is the distribution with range $\mathbb{D} = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \geq 1\}$, whose density, expressed in polar coordinates, is given by*

$$\hbar_0(r, \phi) = r^{-2} s_0(\phi) \mathbf{1}_{\{r>1\}}(r). \tag{6.2}$$

Figure 6.1 shows the density of a raw model. The spectral measure used was the 6 component von Mises mixture model, fitted to the log returns of the BMW-Siemens dataset. See Table 5.11 for the parameters of the model. To relate the definition of the

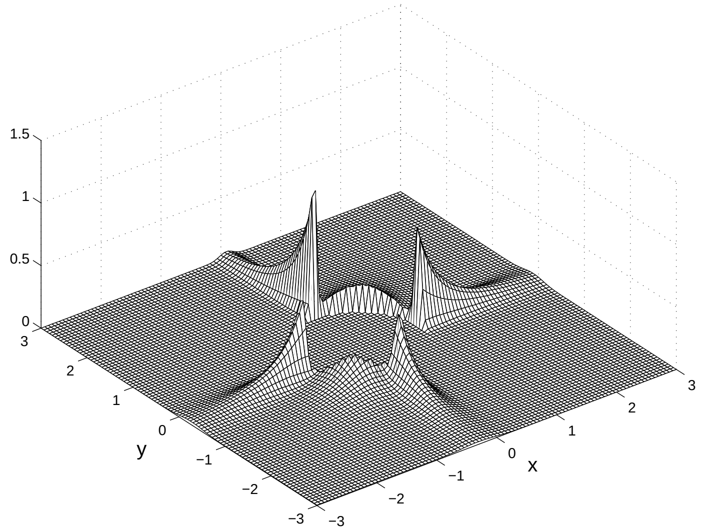

Figure 6.1: *The density plot of an example of the raw model.*

raw model with (6.1), note that if $\mathbf{X}$ is distributed as $\hbar_0$, then

$$\begin{aligned} n\mathbb{P}[\frac{\|\mathbf{X}\|}{n} > r, \|\mathbf{X}\|^{-1}\mathbf{X} \in A] &= n\mathbb{P}[\|\mathbf{X}\| > rn, \|\mathbf{X}\|^{-1}\mathbf{X} \in A] \\ &= n \cdot \big((nr)^{-1} S_0(A)\big) = r^{-1} S_0(A), \end{aligned}$$

where $S_0(A) = \int_A s_0(\phi) d\phi$. Therefore, we have analogue to (6.1), that the distribution with density $\hbar_0$ is in the domain of attraction of a extreme value distribution

$$G_0(\mathbf{x}) = \exp(-\mu_0([\mathbf{0}, \mathbf{x}]^c))$$

with

$$\mu_0\{\mathbf{y} \in \mathbb{E} : \|\mathbf{y}\| > r, \|\mathbf{y}\|^{-1}\mathbf{y} \in A\} = r^{-1} S_0(A).$$

Using the well known theorem describing the change of variables, we can express $\hbar_0$ in cartesian coordinates. Let $x = r\cos\phi$ and $y = r\sin\phi$ and denote with $h_0(x,y)$ the density expressed in cartesian coordinates. Then, we have

$$\hbar_0(r,\phi) = r^{-2} s_0(\phi) \mathbf{1}_{\{r>1\}}(r) = r h_0(r\cos\phi, r\sin\phi) = r h_0(x,y).$$

Therefore, we have

$$\begin{aligned} h_0(x,y) &= h_0(r\cos\phi, r\sin\phi) = r^{-3}s_0(\phi)\mathbf{1}_{r>1}(r) \\ &= (x^2+y^2)^{-3/2}s_0(Alan(x,y))\mathbf{1}_{(x^2+y^2>1)}(x,y). \qquad (6.3) \end{aligned}$$

In the above equation we denote with $atan(x,y)$ the angle ϕ such that $x = r\cos(\phi)$ and $y = r\sin(\phi)$. Let

$$H_0(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h_0(s,t)dsdt$$

be the bivariate distribution function connected to the density $h_0(x,y)$. Denote the marginals distributions of $H_0(x,y)$ by $H_1(x)$ and $H_2(y)$. That is, define

$$H_1(x) = \lim_{y\to\infty} H_0(x,y) \text{ and } H_2(y) = \lim_{x\to\infty} H_0(x,y).$$

We can express the cdf and the pdf of the marginal distribution, by the spectral measure density $s_0(\phi)$. For each marginal cdf we need to consider four different cases.

Proposition 6.1.2 *Let $h_0(x,y), H_0(x,y), H_1(x)$ and $H_2(y)$ be as above. Let*

$$c_1^- = \int_{\pi/2}^{3\pi/2} \cos\phi s_0(\phi)d\phi \textit{ and } c_1^+ = \int_{-\pi/2}^{\pi/2} \cos\phi s_0(\phi)d\phi. \qquad (6.4)$$

Then we have

$$\begin{aligned} H_1(x) &= \frac{c_1^-}{x}, \textit{ if } x \le -1 && (6.5) \\ H_1(x) &= \int_{\pi/2}^{3\pi/2}\left(\frac{\cos\phi}{x}\wedge 1\right)s_0(\phi)d\phi, \textit{ if } -1 < x \le 0 && (6.6) \\ H_1(x) &= 1 - \int_{-\pi/2}^{\pi/2}\left(\frac{\cos\phi}{x}\wedge 1\right)s_0(\phi)d\phi, \textit{ if } 0 < x \le 1 && (6.7) \\ H_1(x) &= 1 - \frac{c_1^+}{x}, \textit{ if } 1 < x && (6.8) \end{aligned}$$

Define

$$c_2^- = \int_{\pi}^{2\pi} \sin\phi s_0(\phi)d\phi \textit{ and } c_2^+ = \int_{0}^{\pi} \sin\phi s_0(\phi)d\phi. \qquad (6.9)$$

Then we have

$$\begin{aligned} H_2(y) &= \frac{c_2^-}{y},\ \textit{if}\ y \leq -1 && (6.10)\\ H_2(y) &= \int_{\pi}^{2\pi} \left(\frac{\sin\phi}{y} \wedge 1\right) s_0(\phi)d\phi,\ \textit{if}\ \ -1 < y \leq 0 && (6.11)\\ H_2(y) &= 1 - \int_0^{\pi} \left(\frac{\sin\phi}{y} \wedge 1\right) s_0(\phi)d\phi,\ \textit{if}\ 0 < y \leq 1 && (6.12)\\ H_2(y) &= 1 - \frac{c_2^+}{y},\ \textit{if}\ 1 < y && (6.13) \end{aligned}$$

Proof:

We have that

$$H_1(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} h_0(s,t)dsdt.$$

Recall from Definition 6.1.1 that $\mathbb{D} = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \geq 1\}$. We make a change of variables to polar coordinates and note that for $x < -1$ the set $\{(s,t) \in \mathbb{D} : s < x\}$ equals the set $\{(r,\phi) : r > x(\cos\phi)^{-1}, \phi \in [\frac{\pi}{2}, \frac{3\pi}{2}]\}$. Therefore, we have from $\int_s^{\infty} r^{-2}dr = s^{-1}$ that

$$H_1(x) = \int_{\pi/2}^{3\pi/2} \int_{\frac{x}{\cos\phi}}^{\infty} r^{-2} s_0(\phi)drd\phi = \frac{1}{x}\int_{\pi/2}^{3\pi/2} \cos\phi s_0(\phi)d\phi = \frac{c_1^-}{x}.$$

The calculations for $-1 < x \leq 0$ are very similar. Fix a value of $x \in (-1, 0]$. Then we have that the set $\{(s,t) \in \mathbb{D} : s < x\}$ equals the set $\{(r,\phi) : r > (x(\cos\phi)^{-1}) \vee 1, \phi \in [\frac{\pi}{2}, \frac{3\pi}{2}]\}$. We therefore get for $-1 < x \leq 0$

$$H_1(x) = \int_{\pi/2}^{3\pi/2} \int_{\frac{x}{\cos\phi}\bigvee 1}^{\infty} r^{-2} s_0(\phi)drd\phi = \int_{\pi/2}^{3\pi/2} \left(\frac{\cos\phi}{x} \wedge 1\right) s_0(\phi)d\phi.$$

For the case $0 \leq x < 1$, note that

$$H_1(x) = 1 - \int_{x}^{\infty} \int_{-\infty}^{\infty} h_0(s,t)dsdt.$$

Since we have $\{(s,t) \in \mathbb{D} : s > x\} = \{(r,\phi) : r > (x(\cos\phi)^{-1}) \vee 1, \phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]\}$, we get that

$$H_1(x) = 1 - \int_{-\pi/2}^{\pi/2} \int_{\frac{x}{\cos\phi}\bigvee 1}^{\infty} r^{-2} s_0(\phi)drd\phi = 1 - \int_{-\pi/2}^{\pi/2} \left(\frac{\cos\phi}{x} \wedge 1\right) s_0(\phi)d\phi.$$

Finally, we have for $x > 1$ that

$$H_1(x) = 1 - \int_{-\pi/2}^{\pi/2} \int_{\frac{x}{\cos\phi}}^{\infty} r^{-2} s_0(\phi) dr d\phi = 1 - \int_{-\pi/2}^{\pi/2} \left(\frac{\cos\phi}{x}\right) s_0(\phi) d\phi = 1 - \frac{c_1^+}{x},$$

remembering the definition of c_1^+ from (6.4). The proof of the equations for $H_2(y)$ are analogue.

Using the equations for the marginal distributions in Proposition 6.1.2, we obtain the density functions of the marginals by calculating the derivatives.

Proposition 6.1.3 *Let $h_0(x,y), H_0(x,y), H_1(x)$ and $H_2(y)$ be as above. The densities of the marginal distributions of $H_0(x,y)$ are given by*

$$\begin{aligned}
h_1(x) &= -c_1^- x^{-2}, x \leq -1 \\
h_1(x) &= \frac{-1}{x^2}\left[\int_{\pi/2}^{\arccos(x)} \cos\phi s_0(\phi) d\phi + \int_{-\arccos(x)}^{3\pi/2} \cos\phi s_0(\phi) d\phi\right], -1 < x \leq 0 \\
h_1(x) &= \frac{1}{x^2}\left[\int_{\arccos(x)}^{\pi/2} \cos\phi s_0(\phi) d\phi + \int_{-\pi/2}^{-\arccos(x)} \cos\phi s_0(\phi) d\phi\right], 0 < x \leq 1 \\
h_1(x) &= c_1^+ x^{-2}, 1 < x
\end{aligned}$$

and

$$\begin{aligned}
h_2(y) &= -c_2^- y^{-2}, y \leq -1 \\
h_2(y) &= \frac{-1}{y^2}\left[\int_{\pi}^{-\arcsin(y)} \sin\phi s_0(\phi) d\phi + \int_{\arcsin(y)}^{2\pi} \sin\phi s_0(\phi) d\phi\right], -1 < y \leq 0 \\
h_2(y) &= \frac{1}{y^2}\left[\int_{0}^{\arcsin(y)} \sin\phi s_0(\phi) d\phi + \int_{-\arcsin(y)}^{\pi} \sin\phi s_0(\phi) d\phi\right], 0 < y \leq 1 \\
h_2(y) &= c_2^+ y^{-2}, 1 < y
\end{aligned}$$

Proof:

The equations for $h_1(x)$ for $x > 1$ and $x \leq -1$ follow immediately from the corresponding equations (6.5) and (6.8) in Proposition 6.1.2 of the cdf by taking derivatives.

Consider now $0 < x \leq 1$. We have from (6.7), that

$$\begin{aligned} H_1(x) &= 1 - \int_{-\pi/2}^{\pi/2} \left(\frac{\cos\phi}{x} \wedge 1 \right) s_0(\phi) d\phi \\ &= 1 - \int_{-\arccos(x)}^{\arccos(x)} s_0(\phi) d\phi - \int_{\arccos(x)}^{\pi/2} \frac{\cos\phi}{x} s_0(\phi) d\phi \\ &\quad - \int_{-\pi/2}^{-\arccos(x)} \frac{\cos\phi}{x} s_0(\phi) d\phi. \end{aligned}$$

Therefore, we have that

$$\begin{aligned} h_1(x) &= \frac{\partial}{\partial x} H_1(x) \\ &= -\frac{\partial}{\partial x} \left[\int_{-\arccos(x)}^{\arccos(x)} s_0(\phi) d\phi \right] - \frac{\partial}{\partial x} \left[\int_{\arccos(x)}^{\pi/2} \frac{\cos\phi}{x} s_0(\phi) d\phi \right] \\ &\quad -\frac{\partial}{\partial x} \left[\int_{-\pi/2}^{-\arccos(x)} \frac{\cos\phi}{x} s_0(\phi) d\phi \right]. \end{aligned}$$

Using

$$\frac{\partial}{\partial x} \arccos(x) = \frac{-1}{\sqrt{(1-x^2)}},$$

we get

$$\begin{aligned} h_1(x) &= -\left[-s_0(\arccos(x)) \frac{1}{\sqrt{(1-x^2)}} - s_0(-\arccos(x)) \frac{1}{\sqrt{(1-x^2)}} \right] \\ &\quad - \left[\frac{-1}{x^2} \left(\int_{\arccos(x)}^{\pi/2} \frac{\cos\phi}{x} s_0(\phi) d\phi \right) + \frac{1}{x} \cdot x s_0(\arccos(x)) \frac{1}{\sqrt{(1-x^2)}} \right] \\ &\quad - \left[\frac{-1}{x^2} \left(\int_{-\pi/2}^{-\arccos(x)} \frac{\cos\phi}{x} s_0(\phi) d\phi \right) + \frac{1}{x} \cdot x s_0(-\arccos(x)) \frac{1}{\sqrt{(1-x^2)}} \right] \\ &= \frac{1}{\sqrt{(1-x^2)}} \left[s_0(\arccos(x)) + s_0(-\arccos(x)) \right] \\ &\quad + \frac{1}{x^2} \left[\int_{\arccos(x)}^{\pi/2} \cos\phi s_0(\phi) d\phi + \int_{-\pi/2}^{-\arccos(x)} \cos\phi s_0(\phi) d\phi \right] \\ &\quad - \frac{1}{x} \left[\frac{x}{\sqrt{(1-x^2)}} \{ s_0(\arccos(x)) + s_0(-\arccos(x)) \} \right] \\ &= \frac{1}{x^2} \left[\int_{\arccos(x)}^{\pi/2} \cos\phi s_0(\phi) d\phi + \int_{-\pi/2}^{-\arccos(x)} \cos\phi s_0(\phi) d\phi \right]. \qquad (6.14) \end{aligned}$$

The proof for the equations of $h_1(x)$ for $-1 < x \leq 0$ and for the equations of $h_2(y)$ are analogue. ■

Figure 6.2 shows the marginal density of the raw model pictured in Figure 6.1. Notice that the density is proportional to x^{-2} for $1 \leq x$ and $x < -1$. On the interval $(-1, 1)$ the structure of the density is determined by the shape of the spectral measure.

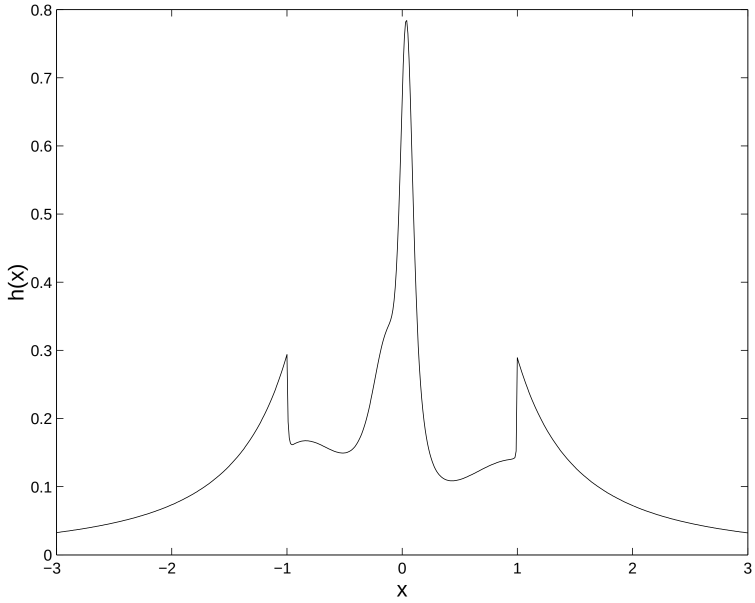

Figure 6.2: *The density $h_1(x)$ of the raw model pictured in Figure 6.1.*

6.2 From the Raw Model to Correct Marginals

6.2.1 Adjusting the Marginals of the Raw Model

The marginal distributions of the raw model, given in the last section, is of course not a reasonable choice for the marginal distributions for a model of tail dependence. The purpose of the raw model is only to describe the tail dependence between the marginals, not the distribution of the marginals themselves or the distribution of the body. Suppose, that we wish the model to have marginal distributions represented by the cumulative distribution functions $F_1(x)$ and $F_2(y)$. We assume that they are absolute continuous with densities $f_1(x)$ and $f_2(y)$. The following is the obvious procedure for obtaining a model with these marginals starting from the raw model.

Recall from (6.3) that the density of the raw model is given by

$$h_0(x,y) = (x^2+y^2)^{-3/2} s_0(atan(x,y)) \mathbf{1}_{\{x^2+y^2>1\}}(x,y).$$

As before, let

$$H_0(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} h_0(s,t) ds dt$$

be the bivariate distribution function connected to the density $h_0(x,y)$. Define for a monotone nondecreasing function $H(x)$ on $\mathbb{R}$ the left continuous inverse as

$$H^{\leftarrow}(y) := \inf\{s : H(s) \geq y\}. \tag{6.15}$$

Define the bivariate cdf $F(x,y)$ as

$$F(x,y) := H_0(H_1^{\leftarrow}(F_1(x)), H_2^{\leftarrow}(F_2(y))). \tag{6.16}$$

To check that $F(x,y)$ indeed has marginals $F_1(x)$ and $F_2(y)$, note that

$$\lim_{y\to\infty} F(x,y) = \lim_{y\to\infty} H_0(H_1^{\leftarrow}(F_1(x)), H_2^{\leftarrow}(F_2(y))) = H_1(H_1^{\leftarrow}(F_1(x))) = F_1(x).$$

The last equality holds since $H_1(x)$ is absolutely continuous with a density $h_1(x)$. Obviously, the same argument also shows that the second marginal distribution is indeed $F_1(x)$.

For the calculation of the density of $F(x,y)$ note that

$$\begin{aligned} \frac{\partial H_1^{\leftarrow}(F_1(x))}{\partial x} &= \frac{\partial H_1^{\leftarrow}(F_1(x))}{\partial F_1(x)} \cdot \frac{\partial F_1(x)}{\partial x} = \frac{f_1(x)}{\frac{\partial}{\partial z}(H_1(z))|_{z=H_1^{\leftarrow}(F_1(x))}} \\ &= \frac{f_1(x)}{h_1(H_1^{\leftarrow}(F_1(x)))}. \end{aligned} \tag{6.17}$$

The second equality is a consequence of the inverse function theorem. We obtain, with the help of (6.17) and by setting $z_1 = z_1(x) = H_1^{\leftarrow}(F_1(x)))$ and $z_2 = z_2(y) = H_2^{\leftarrow}(F_2(y)))$:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = h_0\left(z_1, z_2\right) \frac{f_1(x)}{h_1(z_1)} \frac{f_2(x)}{h_2(z_2)} \tag{6.18}$$

Recalling $h_0(x,y)$ in cartesian coordinates from (6.3), we finally get the following result.

Proposition 6.2.1 *Let s_0 be the density of a finite von Mises mixture model of the spectral measure. Let H_1 and H_2 be given by (6.5)-(6.8) and let (6.10)- (6.13) be the distribution functions of the raw model from Definition 6.1.1. Let h_1 and h_2 be the densities of H_1 and H_2, given by Proposition 6.1.3. Let $z_1 = H_1^{\leftarrow}(F_1(x)))$ and $z_2 = H_2^{\leftarrow}(F_2(y)))$. Denote by f_1 and f_2 two arbitrary density functions on $\mathbb{R}$. If we define $atan(z_1, z_2)$ is as in 6.3, then the bivariate density*

$$f_{Tail}(x,y) = (z_1^2 + z_2^2)^{-3/2} \cdot s_0\left(atan\left(z_1, z_2\right)\right) \cdot \frac{f_1(x)}{h_1(z_1)} \frac{f_2(x)}{h_2(z_2)} \mathbf{1}_{\mathbb{D}}(z_1, z_2), \tag{6.19}$$

has marginal distributions with densities f_1 and f_2.

The distribution given by the density $f_{Tail}(x,y)$ is determined by the spectral measure and its two marginals. The spectral measure describes the dependence between the marginal components, which in turn have distributions given by the densities $f_1(x)$ and $f_2(y)$.

6.2.2 A Model of the Marginal Distribution using the GPD

The choice of the marginal distribution is a crucial part of the model (6.19). The marginal distribution needs to be a reasonable approximation of the features of the data. Remember, that we are developing a model for the data that is in the tails of the distribution. We call an observation “in the tails”, if it is selected by the ranks method. We hence work with the observations selected by the ranks method. These observations will therefore contain the extreme observations for each marginal component. Recall from Section $2.1.6$ the definition of the Generalized Pareto distributions, given by

$$G_{\xi,\beta,\nu}(x) = \begin{cases} 1 - \left(1 + \xi\frac{x-\nu}{\beta}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - exp\left(-\frac{x-\nu}{\beta}\right) & \xi = 0, \end{cases}$$

where $\frac{x-\nu}{\beta} \geq 0$, if $\xi \geq 0$ and $1 + \xi\frac{x-\nu}{\beta} > 0$, if $\xi < 0$. We explained in Section $2.1.6$ that the Generalized Pareto distribution approximates excesses over high thresholds. In particular, if ν denotes a high threshold, we have for $x > \nu$ and a random variable X with distribution function $F \in D(H_\xi)$:

$$\mathbb{P}[X > x] \approx (1 - G_{\xi,\beta,\nu}(x))\mathbb{P}[X > \nu].$$

For this reason the GPD appears as the natural model for the left and the right tail of the marginal distributions $F_1(x)$ and $F_2(y)$. The GPD describes the tails beyond the thresholds $-\nu_l < 0$ and $\nu_r > 0$. We additionally need a model for the body of the marginal distributions $F_1(x)$ and $F_2(y)$. We use a normal distribution. Other choices, like a linear transformation of a Beta distribution also give reasonable models of the marginal distribution between $-\nu_l < 0$ and $\nu_r > 0$. We use a mixture model to combine the normal distribution of the body with the GPD of the tails. We hence assume that the marginal distributions have the following densities, $i = 1, 2$:

$$\begin{aligned} f_i(x) &= p_1^{(i)} g_r(x; \xi_r^{(i)}, \beta_r^{(i)}, \nu_r^{(i)}) + p_2^{(i)} g_l(x; \xi_l^{(i)}, \beta_l^{(i)}, \nu_l^{(i)}) \\ &\quad + (1 - p_1^{(i)} - p_2^{(i)}) \phi(x; \mu_T^{(i)}, \sigma_T^{(i)}), \end{aligned} \tag{6.20}$$

where

$$g_r(x;\xi,\beta,\nu) = \frac{1}{\beta}\left(1+\xi\frac{x-\nu}{\beta}\right)^{-\frac{1}{\xi}-1}\mathbf{1}_{[\nu,\infty)}(x) \tag{6.21}$$

$$g_l(x;\xi,\beta,\nu) = \frac{1}{\beta}\left(1+\xi\frac{-x-\nu}{\beta}\right)^{-\frac{1}{\xi}-1}\mathbf{1}_{(-\infty,-\nu)}(x) \tag{6.22}$$

$$\phi(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6.23}$$

The corresponding distribution functions are

$$\begin{aligned} F_i(x) &= p_1^{(i)}G_r(x;\xi_r^{(i)},\beta_r^{(i)},\nu_r^{(i)}) + p_2^{(i)}G_l(x;\xi_l^{(i)},\beta_l^{(i)},\nu_l^{(i)}) \\ &\quad +(1-p_1^{(i)}-p_2^{(i)})\Phi(x;\mu_T^{(i)},\sigma_T^{(i)}), \end{aligned} \tag{6.24}$$

where

$$G_r(x;\xi,\beta,\nu) = 1-\left(1+\xi\frac{x-\nu}{\beta}\right)^{-\frac{1}{\xi}}\mathbf{1}_{[\nu,\infty)}(x) \tag{6.25}$$

$$G_l(x;\xi,\beta,\nu) = \left(1+\xi\frac{-x-\nu}{\beta}\right)^{-\frac{1}{\xi}}\mathbf{1}_{(-\infty,-\nu)}(x) + \mathbf{1}_{[-\nu,\infty)}(x) \tag{6.26}$$

$$\Phi(x;\mu,\sigma) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right)dt \tag{6.27}$$

Estimation of the parameters

The estimation of the parameters of the marginal model with pdf (6.20) and cdf (6.24) is not very easy. We are using the observations selected by the ranks method to estimate these parameters, since that is the data whose distribution we are modelling. We found that an algorithm that maximizes the log likelihood function over all 10 parameters of the model is not practical. We therefore first obtain estimates of the parameters of the two GPD components and then estimate the parameters of the normal component and the weights p_1 and p_2 in a separate maximum likelihood procedure.

The ranks method essentially uses two criteria to decide which observations are to be selected for the estimation of the spectral measure. The first is the ranks of each coordinate and the second is a choice of the number of upper order statistics, referred to as k. As we explained before, if

$$\mathbf{r_j} = (r_j^{(i)}, i = 1, ..., d)$$

is the vector of the ranks

$$r_j^{(i)} = \sum_{l=1}^{N} 1_{[X_l^{(i)} > X_j^{(i)}]}$$

of the observation $(X_j^{(1)}, ..., X_j^{(d)})$, and if $R_{j,k}$ is the norm of $\frac{k}{\mathbf{r_j}}$, the ranks method selects observation $\mathbf{X}_j$, if and only if $R_{j,k} > 1$. In particular, in the bivariate case $d = 2$, any observation with a marginal component i, such that either

$$r_j^{(1)} < k \text{ or } r_j^{(2)} < k$$

will be selected. That is, if there is a marginal component $i = 1, 2$ of observation $\mathbf{X}_j$, such that $X_j^{(i)}$ is among the k largest of the observations $(X_1^{(i)}, ..., X_N^{(i)})$, observation $\mathbf{X}_j$ gets selected. We therefore found it natural to use the k largest order statistics of each marginal for the estimation of the GPD components.

Denote for the remainder of the section with $\mathbf{Z} = (Z_1, .., Z_N)$ the i^{th} marginal component of the observations $\mathbf{X}_1, ..., \mathbf{X}_N$. That is $\mathbf{Z}$ is contains the observations $(X_j^{(i)}; j = 1, ..., N)$ for which $R_{j,k} > 1$. Denote with $Z_{(k)}$ the order statistics of $\mathbf{Z}$: $Z_{(1)} < Z_{(2)} < ... < Z_{(N)}$. Then the estimators for the parameters $-\nu_l := -\nu_l^{(i)} < 0$ and $\nu_r := \nu_r^{(i)} > 0$ are as follows:

$$\widehat{\nu}_r = Z_{(N-k)} \quad (6.28)$$

$$\widehat{\nu}_l = Z_{(k+1)} \quad (6.29)$$

Based on (6.28) and (6.29), we obtain the maximum likelihood estimates of $\xi_r^{(i)}, \beta_r^{(i)}$, based on $(Z_{(N-k+1)}, ..., Z_{(N)})$. Similarly, we find the maximum likelihood estimators of

$\xi_l^{(i)}, \beta_l^{(i)}$ based on $(-Z_{(1)}, ..., -Z_{(k)})$. We use SPLUS, more specifically the EVIS 5.0 software package, to carry out the maximum likelihood estimation.

We then use the estimated parameters of the GPD in the estimation of the component weights and the parameters of the normal distribution components. We find the restricted maximum likelihood estimators $\widehat{\mu}, \widehat{\sigma}, \widehat{p}_1, \widehat{p}_2$ of $\mu_T := \mu_T^{(i)}, \sigma_T := \sigma_T^{(i)}, p_1 := p_1^{(i)}$ and $p_2 := p_2^{(i)}$ by maximizing the likelihood function

$$\begin{aligned} & L(\mu_T, \sigma_T, p_1, p_2; \mathbf{Z}) \qquad (6.30) \\ & = \sum_{i=1}^{n} \log\Big(p_1 g_r(Z_i; \widehat{\xi}_r, \widehat{\beta}_r, \widehat{\nu}_r) + p_2 g_l(Z_i; \widehat{\xi}_l, \widehat{\beta}_l, \widehat{\nu}_l) + (1 - p_1 - p_2)\phi(Z_i; \mu_T, \sigma_T)\Big) \end{aligned}$$

over $(\mu_T, \sigma_T, p_1, p_2) \in \mathbb{R} \times \mathbb{R}^+ \times \{(p_1, p_2) \in (0,1)^2 : p_1 + p_2 < 1\}$. We refer to the resulting estimates as restricted maximum likelihood estimates, rather than maximum likelihood estimates, because we obtain them by maximizing the log likelihood function only over μ_T, σ_T, p_1, and p_2, and not over all parameters. We find the values $(\widehat{\mu_T}, \widehat{\sigma_T}, \widehat{p}_1, \widehat{p}_2)$ that maximize (6.30) using the optimization toolbox in Matlab.

The marginal model in the case of IBM

To illustrate the shape and nature of the marginal model introduced in this section, we consider the case of the parameters values that we obtained as the estimates for the IBM dataset. We mentioned in Section 5.2 that we used the ranks method with $k = 80$, resulting in $n = 302$ observations being chosen. For the right tail, we find that $\widehat{\nu}_r = 0.0361$ and as a consequence we have that $\widehat{\xi}_r = 0.2175$ and $\widehat{\beta}_r = 0.0130$. Note that since $\widehat{\xi}_r > 0$, the GPD model indicates, that the right tail of the marginal distribution of IBM is heavy tailed. The corresponding estimate of the tail index is $\widehat{\alpha}_r = 1/\widehat{\xi}_r = 4.5977$. Recall that in Section 5.1 we estimated the tail index of the right tail of the distribution of IBM with 3.5, based on Hill plots. The difference in the estimates illustrates the difficulty of estimating the tail indices of heavy tailed distributions.

For the left tail, we obtain the following estimates: $\widehat{\nu_l} = 0.0344$, $\widehat{\xi_l} = 0.4261$ and $\widehat{\beta_l} = 0.0097$. Similar to the model of the right tail, we have that the estimate of ξ_l corresponds to a heavy tailed distribution with a tail index estimate of $\widehat{\alpha_l} = 1/\widehat{\xi_l} = 2.3469$. Recall that we obtained an estimate of $\widehat{\alpha} = 2.8$ for the tail index of the left tail in Section 5.1.

Based on these estimates for the parameters of the tail components, we obtain the following estimates for the weights and the parameters of the normal components:

$$\widehat{\mu_T} = 0.0015, \widehat{\sigma_T} = 0.0233, \widehat{p_1} = 0.2458 \text{ and } \widehat{p_2} = 0.2613.$$

Figure 6.3 shows the density of the marginal model for the tails of IBM with these parameters. The upper half of the figure shows a scatter plot of the data used in the estimation of the parameters and a non-parametrical estimate of the density. The lower half of the figure shows the density of the marginal model fitted to the returns of IBM. Note, that the density of the model seems to capture the structure of the data very well. In particular the two spikes of the density, that are visible at $\widehat{\nu_r} = 0.0361$ and $-\widehat{\nu_l} = -0.0344$ are also clearly visible in the data in the top plot in Figure 6.3. The reason for the presence of those spikes becomes clear when we study the scatter plot of the observations that were selected by the ranks method. That plot is given in Figure 6.4. We see that these observations seem to be located on the outside of a rectangle. A large number of these observations are close to the borders of that rectangle. Therefore, the marginal distribution appears to have two spikes, approximately at $\widehat{\nu_r}$ and $-\widehat{\nu_l}$.

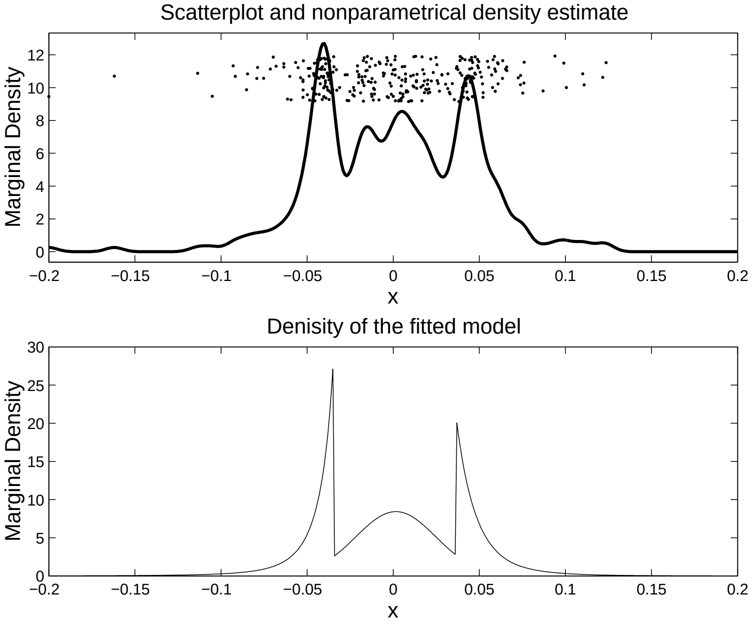

Figure 6.3: *The density of the marginal model (6.19) for IBM. See text for details.*

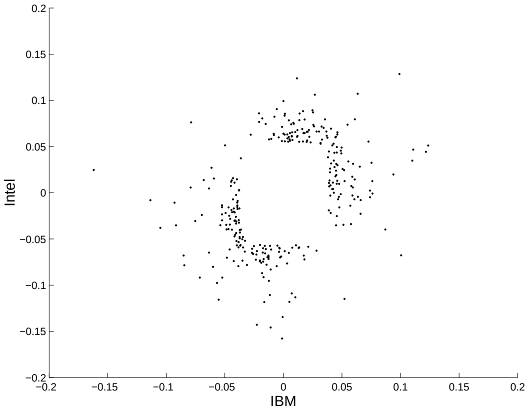

Figure 6.4: *Scatter plot of the points selected by the ranks method with* $k = 80$.

6.3 Body and Tails Combined

Remember that the model presented so far is only a model for the tail region of the distribution. Also recall that the density of the raw model has support

$$\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 > 1\}.$$

After adjusting the tails, the tail distribution density has support

$$\mathbb{D} = \{(x,y) \in \mathbb{R}^2 : (H_1^{\leftarrow}(F_1(x)))^2 + (H_2^{\leftarrow}(F_2(x)))^2 > 1\}.$$

In order to develop a model that describes the entire bivariate distribution of two random variables, and not just the distribution of the tail region, we need to introduce a component describing the ”body” of the distribution. That is, we need a model for the distribution in $\mathbb{D}^c$. That model should be based on the observations that were not selected by the ranks method. There are several different choices for such a model. We decided to use a bivariate normal distribution. We hence assume that the distribution of the body has the following density

$$f_{Body}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho^2)}} \exp\left[-\frac{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2}{2(1-\rho^2)}\right] \tag{6.31}$$

where μ_x and μ_y stand for the marginal expectations, σ_x and σ_y stand for the corresponding standard deviations and finally ρ stands for the correlation between the two marginal components.

More sophisticated models, for example models based on copulas, could be considered and they would probably be more accurate. Breymann et al. (2003) propose a number of dependence structures for high frequency data in finance. They focus on modelling the dependence structure of the entire distribution of the log returns of two currency exchange rates. They considered the Gaussian, the t, the Frank, the Gumbel and the Clayton copulas. They did not specify any marginal models. Instead, they used the empirical distributions to transform the data before fitting the respective copula. They found that for the dependence structure of the entire data the t-copula gave the best description among the considered models. However, they also found that different copulas best describe the lower and upper tail dependence. Models based on some of these copulas may be more realistic than the bivariate model that we chose. However, they would also be more challenging to implement. Also keep in mind that the normal distribution is only serving us as an appropriate model of the body of the data and not as a model of the entire distribution. We describe the tail dependence with great care separately with the help of our mixture model of the spectral measure. The results of Breymann et al. (2003) may not apply to our case, since they are based on research concentrated on the entire distribution and not just the body. Furthermore, our focus in this thesis is concentrated on developing a realistic model of the dependence in the tails of the distribution. It is not our goal to develop an optimal model for the body of the distribution. We are not aware of such research focused on modelling the dependence structure of only the body of a distributions recommending a specific model. In the absence of such research, we decided to use the most common model for describing multivariate data, the multivariate normal distribution.

In the following we describe how we combine the model of the body and the model of the tails to obtain a comprehensive model of the entire distribution. To unite the two models, $f_{Body}(x, y)$, given in (6.31) and $f_{Tail}(x, y)$, given by (6.19), we make again use of the concept of a mixture model. That is, we assume the bivariate distribution has density

$$f(x, y) = pf_{Tail}(x, y) + (1 - p)f_{Body}(x, y). \tag{6.32}$$

An advantage of this approach is that our mixture model can easily be combined with any particular model of the body that a researcher may see fit.

The marginal distributions of model (6.32) are easily obtained from the corresponding marginal distributions of $f_{Tail}(x, y)$ and $f_{Body}(x, y)$. The marginal densities are of the form

$$\begin{aligned} f_i(x) &= \alpha_1^{(i)} g_r(x; \xi_r^{(i)}, \beta_r^{(i)}, \nu_r^{(i)}) + \alpha_2^{(i)} g_l(x; \xi_l^{(i)}, \beta_l^{(i)}, \nu_l^{(i)}) \\ &\quad + \alpha_3^{(i)} \phi(x; \mu_T^{(i)}, \sigma_T^{(i)}) + \alpha_4^{(i)} \phi(x; \mu_B^{(i)}, \sigma_B^{(i)}), \end{aligned} \tag{6.33}$$

where $\alpha_1^{(i)} = p \cdot p_1^{(i)}$, $\alpha_2^{(i)} = p \cdot p_2^{(i)}$, $\alpha_3^{(i)} = p \cdot (1 - p_1^{(i)} - p_2^{(i)})$, and $\alpha_4^{(i)} = (1 - p)$. Furthermore, $\mu_T^{(i)}$ and $\sigma_T^{(i)}$ denote the mean and standard deviation of the normal components of the tail model, respectively. Finally $\mu_B^{(i)}$ and $\sigma_B^{(i)}$ stand for the mean and standard deviation of the corresponding marginal component of f_{Body}.

We estimate the parameters of the tail components $g_r(x; \xi, \beta, \nu)$, $g_l(x; \xi, \beta, \nu)$ and $\phi(x; \mu, \sigma)$ of the marginal distributions as mentioned above in Section 6.2.2. Since (6.31) is acting as the model for the distribution of the body, we only use the points not selected by the ranks method for the estimation of the parameters of $f_{Body}(x, y)$. We estimated the means, standard deviations and the correlation by the corresponding sample means, sample standard deviations, and the sample correlation, respectively. Finally, we estimate the weight p of the tail component, f_{Tail}, by the percentage of the points that were selected by the ranks method. Figure 6.5 shows a plot of the density (6.33) with the parameters that we estimated from the log returns of IBM. We obtained

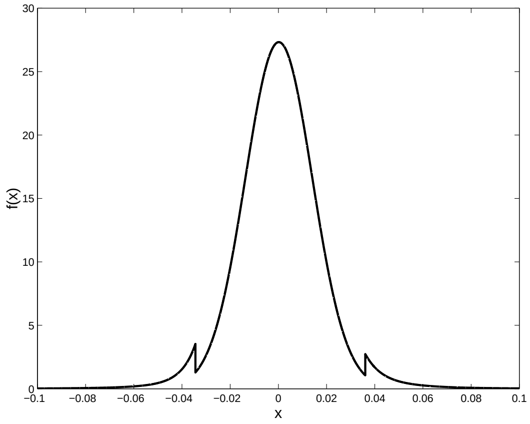

Figure 6.5: *The density of the marginal model (6.33) fitted to the log returns of IBM.*

the following estimates of $f_{Body}(x, y)$ for IBM:

$$\widehat{\mu}_B = 1.9514^{-4}; \widehat{\sigma}_B = 0.0138, \widehat{p} = 0.0836$$

Since the marginal densities of the tail component f_{Tail} showed two clear spikes at $\widehat{\nu}_r$ and $-\widehat{\nu}_l$, we also see the same spikes in the marginal distribution of the combined model. We furthermore see from Figure 6.5 that, for the parameters values that we estimated from the log-returns of IBM, the mixture of the normal distributions $\phi(x; \mu_B, \sigma_B)$ and $\phi(x; \mu_T, \sigma_T)$ is unimodal. The estimated standard deviations are of the same order and the means are very close to each other.

Another important fact is that the two normal components have very little influence over the tails of the marginal distribution. The two GPD components of the marginal distribution have much heavier tails than the two normal components. In addition, both normal components have fairly small standard deviations, thus they are closely concentrated around their respective means. For the marginal distribution with parameter values as estimated for the log returns of IBM, we observed the following: The two normal components together only have 1.29% of their total mass outside the interval $(-\nu_l, \nu_r)$. Remember that the outside of that interval is the domain of the two GPD components. This means that the two normal components have very little to do with the modelling of the tails of the log returns of IBM. We furthermore found that the fraction of the mass of the two normal components that lies outside of $(-2\nu_l, 2\nu_r)$ is only about $8 \cdot 10^{-6}$. At the same time the mass of the two GPD components have their entire mass outside $(-\nu_l, \nu_r)$ and still 11.26% of that mass outside of $(-2\nu_l, 2\nu_r)$. This means, that the influence of the normal distributions in the tails, that is beyond the points $-\nu_l$ and ν_r, is very small and that it is indeed the GPD components who are essentially describing the tails. This was typical of what we saw for other fitted marginal distributions as well.

Chapter 7

Portfolio Optimization

In this chapter we present an important application of the model developed in the last chapter. We show how our model can be used to optimize portfolios of different financial instruments. We calculate, based on our comprehensive model, the portfolio that minimizes a measure of risk for a given level of expected log return. There are many different definitions of risk and measures thereof. We give a brief overview over the different concepts of risk and motivate our particular choice, called the expected shortfall. We discuss and interpret the results from our optimization and compare the performance of our model with the performance of two other, simpler models. In order to keep the computations feasible we concentrated on the case of a portfolio that consists of two financial instruments.

7.1 Measures of Risk

Assume that X denotes the future log return over a certain time horizon of a financial instrument. We assume that X is a random variable on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. In risk management, we are concerned with the estimation of the distribution of X. We are specifically interested in measuring the risk of losses associated with X. Different distributions of X lead to different risks. The risk is usually assessed by a so called risk measure. We will concentrate our attention on risk measures that only depend on the distribution of X, and not on X itself. In this section we discuss some desirable properties of risk measures followed by an overview over some commonly used risk measures.

Definition 7.1.1 ***(Risk Measure)*** *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space. Let V be a non-*

empty set of $\mathcal{A}$ measurable, real-valued random variables. A risk measure is a mapping

$$\begin{aligned} \rho: \quad V \rightarrow & \quad \mathbb{R} \cup \{\infty\} \\ X \longmapsto & \quad \rho(X) \end{aligned} \tag{7.1}$$

This is a very general definition that allows for very different measures of risk. Artzner et al. (1999) introduced the notion of coherent risk measures. They postulated four properties that a reasonable, or coherent, risk measure should have.

Definition 7.1.2 ***(Coherent Risk Measure)*** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let V be a non-empty set of $\mathcal{A}$ measurable, real-valued random variables. A risk measure ρ is called a coherent risk measure, if it satisfies the following properties:*

1. *Monotonicity:* $X, Y \in V, X \leq Y \Rightarrow \rho(X) \geq \rho(Y)$
2. *Positive Homogeneity:* $\forall \lambda \geq 0, \forall X \in V$*, such that* $\lambda X \in V : \rho(\lambda X) = \lambda \rho(X)$
3. *Translation Invariance:* $X \in V, a \in \mathbb{R}, X + a \in V \Rightarrow \rho(X + a) = \rho(X) - a$
4. *Subadditivity:* $X, Y \in V, X + Y \in V \Rightarrow \rho(X + Y) \leq \rho(X) + \rho(Y)$

The four conditions are easy to interpret. The property "Subadditivity" represents the reduction of risk associated with diversification. It states that the risk of the portfolio obtained by adding two positions of financial instruments is not greater than the sum of the risk of the two positions.

We now introduce some examples of risk measures. They all share the property that they only depend on the distribution of X and not on X itself in the following sense: Let X and Y denote two random variables satisfying $\mathbb{P}[X \leq t] = \mathbb{P}[Y \leq t]$ for all $t \in \mathbb{R}$. Then we have that $\rho(X) = \rho(Y)$.

Standard Deviation

The standard deviation of the portfolio log return X is a risk measure. Of course the standard deviation is not coherent, since it is not monotone nor is it translation invariant. Nevertheless it is often used to measure the risk of a portfolio.

Value At Risk, VaR$_\alpha$

Value at Risk is a very popular risk measure in the finance industry. For a random variable X with distribution function F, define the *quantile of X at level α* as

$$q_\alpha(X) = \inf\{x \in \mathbb{R} : \mathbb{P}[X \leq x] \geq \alpha\} = F^{\leftarrow}(\alpha) \tag{7.2}$$

We call

$$VaR_\alpha(X) = q_{1-\alpha}(-X) \tag{7.3}$$

the *Value at Risk at confidence level α of X*. Usually, α is close to zero. Typical values for α are $\alpha = 0.01$ or $\alpha = 0.05$. Despite its popularity, Value at Risk is in general not a coherent risk measure, since it is not subadditive. Examples of violations of subadditivity of the Value at Risk can for example be found in Embrechts (2000), Tasche (2002), Acerbi et al. (2001) and Artzner et al. (1999). VaR$_\alpha$ is however a coherent risk measure on certain sets V of random variables. For example, if V only contains random variables with elliptical distributions, Embrechts et al. (2002) show that VaR$_\alpha$ is indeed a coherent risk measure. We refer to Embrechts et al. (2002) for precise statement of the result and a proof thereof.

Expected Shortfall, ES$_\alpha$, and related measures

Intuitively speaking, the Expected Shortfall with level α, ES$_\alpha$, is the average size of the loss encountered, given that the loss is worse than the VaR$_\alpha$. For that reason it has also been referred to as “conditional value at risk” or “tail value at risk”. It has been advocated in several variants as a coherent improvement over VaR$_\alpha$. The ES$_\alpha$ can be understood as an improvement over VaR$_\alpha$, because it describes how big your loss will be, given that it is severe. Two different financial instruments can have the same VaR$_\alpha$, but very different ES$_\alpha$. Most definitions of ES$_\alpha$ lead to the same risk measure, depending on set V of random variables considered. If V contains only random variables X that satisfy $\mathbb{P}[X = x] = 0$, for all $x \in \mathbb{R}$ most definitions of ES$_\alpha$ are indeed coherent risk measures. However, if we expand V to include random variables X whose distribution is not continuous, not all variants are coherent and they are different risk measures. In the following we give the definition of a coherent variant and mention some of the alternatives. For a detailed discussion we refer to Acerbi and Tasche (2002).

Assume throughout this paragraph that $\mathbb{E}[X^-] < \infty$. Then we call

$$\text{TCE}_\alpha(X) = -\mathbb{E}[X|X \leq q_\alpha(X)] \tag{7.4}$$

the *tail conditional expectation at level α of X*. It is an intuitive measure of the average loss that can be expected, given that the loss is bigger than the VaR$_\alpha$. However it is not necessarily a subadditive risk measure, see example 5.4 in Acerbi and Tasche (2002). The example is based on a distribution with discontinuities.

To avoid a violation of the subadditivity of the risk measure because of a lack of strict monotonicity of the distribution function, the following alternative to 7.4 has been adopted.

We define the *tail mean at level α of X* as

$$TM_\alpha(X) = \alpha^{-1}\left(\mathbb{E}[X\mathbf{1}_{\{X<q_\alpha(X)\}}] + q_\alpha(X)(\alpha - \mathbb{P}[X < q_\alpha(X)])\right) \tag{7.5}$$

We then define the *Expected Shortfall at level α of X* as

$$ES_\alpha(X) = -TM_\alpha(X). \tag{7.6}$$

The following proposition summarizes the most important properties of the ES$_\alpha$ and its relation to the TUE.

Proposition 7.1.3 *Let X be a real random variable on some probability space* $(\Omega, \mathcal{A}, \mathbb{P})$ *with* $\mathbb{E}[X^-] < \infty$ *and fix* $\alpha \in (0,1)$. *Then the ES*$_\alpha$*, given by (7.6), is a coherent risk measure. Furthermore, we have that*

$$TCE_\alpha(X) \leq ES_\alpha(X). \tag{7.7}$$

We have $TCE_\alpha(X) = ES_\alpha(X)$*, if and only if* $\mathbb{P}[X \leq q_\alpha(X)] = \alpha$ *or* $\mathbb{P}[X < q_\alpha(X)] = 0$.

Furthermore, the ES$_\alpha$ *has the following representation*

$$ES_\alpha(X) = -\alpha^{-1} \int_0^\alpha q_t(X)dt. \tag{7.8}$$

As a consequence, the mapping $\alpha \longmapsto ES_\alpha(X)$ *is continuous on (0,1).*

Proof: See Acerbi and Tasche (2002), Proposition 3.1, Proposition 3.2, Corollary 3.3 and Corollary 5.3. ■

An alternative version of the ES$_\alpha$ is the conditional value at risk, given by

$$CVaR_\alpha(X) = \inf\{\frac{\mathbb{E}[(X-s)^-]}{\alpha} - s : s \in \mathbb{R}\} \tag{7.9}$$

As shown in Acerbi and Tasche (2002), the ES$_\alpha$ is equal to the CAR, if X is integrable.

Spectral Measures Of Risk

Spectral measures of risk are motivated the integral representation of the ES$_\alpha$ given in Proposition 7.1.3:

$$ES_\alpha(X) = -\alpha^{-1} \int_0^\alpha q_p(X)dp = -\alpha^{-1} \int_0^\alpha F^{\leftarrow}(p)dp$$

This equation can be rewritten in the form

$$ES_\alpha(X) = -\int_0^1 F^{\leftarrow}(p)\zeta(p)dp. \tag{7.10}$$

with

$$\zeta(p) = \alpha^{-1}\mathbf{1}_{\{[0,\alpha]\}}(p).$$

This motivates the following definition

Definition 7.1.4 *A spectral measure of risk is a risk measure of the form*

$$M_\zeta(X) = -\int_0^1 F^{\leftarrow}(p)d\zeta(p). \tag{7.11}$$

The measure $\zeta(p)$ is referred the *risk aversion measure*. Not every choice of a risk aversion measure results in a coherent risk measure. We have to impose certain conditions on the possible risk aversion measure. We call a risk aversion measure an *admissible risk aversion measure*, if it is of the kind

$$d\zeta(p) = c \cdot d\delta\{p\} + \widetilde{\zeta}(p)dp, \tag{7.12}$$

where δ is the Dirac delta measure, $c \in [0,1]$ and $\widetilde{\zeta}(p) : [0,1] \to \mathbb{R}$ satisfies:

$$\begin{aligned} \widetilde{\zeta}(p) &\geq 0, \forall p && (7.13)\\ p_1 < p_2 &\Rightarrow \widetilde{\zeta}(p_1) \geq \widetilde{\zeta}(p_2) && (7.14)\\ \int_0^1 \widetilde{\zeta}(p)dp &= 1 - c && (7.15) \end{aligned}$$

Using this definition, we have

Proposition 7.1.5 *Let*

$$M_\zeta(X) = -\int_0^1 F^{\leftarrow}(p)d\zeta(p).$$

be a spectral measure of risk. Then M_ζ is a coherent measure of risk if and only if ζ is an admissible risk aversion function.

Proof: See Acerbi (2002). ■

From the definition of spectral measures of risk it is clear that they only depend on the distribution function F of the random variable X and not on X itself. That is, $M_{\zeta}(X)$ depends only on the distribution F of X. However, it is not true that all risk measures that only depend on the distribution of the random variables, and not the random variable itself, are spectral measures of risk.

The risk aversion measure expresses the subjective risk aversion of the risk manager. It expresses how much weight should be given to the quantiles $F^{\leftarrow}(p)$. In that sense they are a intuitive extension of the ES$_{\alpha}$. The risk measure is coherent if it assigns larger weights to larger negative quantiles. Larger negative quantiles represent worse scenarios. The ES$_{\alpha}$ assigns weight $1/\alpha$ to all scenarios that are worse than the VaR$_{\alpha}$ and no weight to quantiles that are smaller than VaR$_{\alpha}$.

The Dirac delta measure part allows us to include a factor for the worst case scenario $F^{\leftarrow}(0) = -\text{ess}\inf\{X\}$. We have for example that

$$ES_0(X) := -F_X^{\leftarrow}(0) = -\text{ess}\inf\{X\}$$

is a spectral measure of risk with risk aversion measure

$$d\zeta(p) = d\delta\{p\} + \widetilde{\zeta}(p)dp$$

with $c = 1$, $\widetilde{\zeta} = 0$. Hence, ζ is an admissible risk aversion measure and therefore $ES_0(X)$ is a coherent risk measure.

7.2 Managing Risk, Optimizing Portfolios

Assume that $\mathbf{Z} = (Z^{(1)}, ..., Z^{(d)})$ denotes the random vector of the log returns over a certain time horizon of d financial instruments $(\mathcal{Z}^{(1)}, ..., \mathcal{Z}^{(d)})$. That is, if we denote with $\mathrm{z}_1^{(i)}, \ldots, \mathrm{z}_N^{(i)}$ N observations of $Z^{(i)}$, we have for $t = 1, \ldots, N$:

$$\mathrm{z}_t^{(i)} = \log(z_t^{(i)}) - \log(z_{t-1}^{(i)})$$

where $z_t^{(i)}, t = 0, \ldots, N$, denote the observations of the random variable $\mathcal{Z}^{(i)}$. We will also refer to $(Z^{(1)}, ..., Z^{(d)})$ as the risk factors. Consider a linear portfolio, containing ω_i units of the instrument $Z^{(i)}$. The log return of that portfolio is a linear combination of the log returns of $\left(Z^{(1)}, ..., Z^{(d)}\right)$:

$$X := X(\boldsymbol{\omega}) = X(\omega_1, ..., \omega_d) = \sum_{i=1}^{d} \omega_i Z^{(i)}.$$

Different choices of the weights $\boldsymbol{\omega} = (\omega_1, ..., \omega_d)$ of the different instruments result in different distributions of the random variable X. Given a risk measure, we can compare different portfolios by comparing the expected log returns $\mu = \mathbb{E}[X(\boldsymbol{\omega})] = \sum_{i=1}^{d} \omega_i \mathbb{E}[Z^{(i)}]$, and the associated risks $\rho(X) := \rho(X(\boldsymbol{\omega}))$, assuming that all the expectations $\mathbb{E}[Z^{(i)}]$ exist and are finite. Typically, we seek to find a portfolio that minimizes $\rho(X(\boldsymbol{\omega}))$ compared to all possible portfolios with expected log return $\mu = \mathbb{E}[X(\boldsymbol{\omega})]$ under certain constraints. That is, we attempt to solve the following minimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\omega} \in \mathcal{W}} \rho(X(\boldsymbol{\omega})) \\ &\text{s. t. } \mathbb{E}[X(\boldsymbol{\omega})] = \mu \end{aligned} \tag{7.16}$$

The domain $\mathcal{W}$ reflects possible trade restrictions. A typical example of such a trade restriction is a limit on the value of short sales. It may also reflect budget constraints, such as the maximum cost associated with the portfolio. Alternatively, we might define a certain level of risk ϱ deemed admissible and then attempt to find a portfolio $\boldsymbol{\omega} \in \mathcal{W}$ that maximizes the expected log return compared to all possible portfolios whose risk measure equals ϱ:

$$\begin{aligned} &\max_{\boldsymbol{\omega}\in\mathcal{W}} \mathbb{E}[X(\boldsymbol{\omega})] \\ &\text{s. t. } \rho(X(\boldsymbol{\omega})) = \varrho \end{aligned} \tag{7.17}$$

These two problems are usually referred to as the risk-log return optimization problem.

Definition 7.2.1 *In the framework of the optimization problems (7.16) and (7.17) with domain $\mathcal{W}$ we say that a portfolio $\boldsymbol{\omega_1}$ dominates a portfolio $\boldsymbol{\omega_2}$, if*

$$\mathbb{E}[X(\boldsymbol{\omega}_1)] \geq \mathbb{E}[X(\boldsymbol{\omega}_2)] \text{ and } \rho(X(\boldsymbol{\omega}_1)) \leq \rho(X(\boldsymbol{\omega}_2)) \tag{7.18}$$

and at least one of the two inequalities is strict.

We say that a portfolio $\boldsymbol{\omega}$ is optimal, if there is no portfolio that dominates $\boldsymbol{\omega}$. The geometrical set of all optimal portfolios is called the efficient frontier in the plane $(\rho(X(\boldsymbol{\omega})), \mathbb{E}[X(\boldsymbol{\omega})])$

In order to compare the risk and log returns of different portfolios we need a model of the joint distribution of the risk factors $(Z^{(1)}, ..., Z^{(d)})$. Based on such a model, we can then calculate the expected log return of the portfolio, as well as its risk measure. In practice, the difficulty of the calculation of the risk measure of a portfolio depends on the model of the joint distribution of $(Z^{(1)}, ..., Z^{(d)})$. For simple models and risk measures, such as the standard deviation, the corresponding calculation is fairly straightforward and easy. However, we will see that for more sophisticated models the calculation of spectral measures of risk, such as ES_α, can be very time consuming and challenging. This can make the task of finding optimal portfolios a very hard one.

In the special case where we do not have any constraints, such as budget constraints or limitations on the short sales, we need to solve the optimization problem only once, provided that we are working with a coherent risk measure. The optimal portfolios for different levels of expected log returns all have the same proportions between the positions of the different risk factors.

To see this, suppose that we have an optimal portfolio with a certain expected log return μ and risk measure ρ^*. Denote the positions in the risk factors of the optimal portfolio with the vector $(\omega_1^*, ..., \omega_d^*)$. Assume that we wish to find the optimal portfolio for a different expected log return that is, say $\lambda\mu$. A candidate is the portfolio with positions $\lambda\omega_1^*, ..., \lambda\omega_d^*$. This portfolio has risk measure $|\lambda|\rho^*$, because of the positive homogeneity of the coherent risk measure and the linearity of the portfolio. This means its risk and expected log return are linear functions of $|\lambda|$. However, the same is true for every other portfolio $(\omega_1, ..., \omega_d)$ with expected log return μ and risk measure ρ. After inflating the positions by the factor λ, we have a portfolio with expected log return $\lambda\mu$ and risk measure $|\lambda|\rho$. But the portfolio $(\omega_1^*, ..., \omega_d^*)$ was the portfolio with the smallest risk measure among all portfolios with expected log return μ. That is, we have $\rho^* \leq \rho$. Therefore we also have $|\lambda|\rho^* \leq |\lambda|\rho$ for the risk measure $|\lambda|\rho$ of any portfolio $(\lambda\omega_1, ..., \lambda\omega_2)$. Therefore the optimal portfolio with expected log return $\lambda\mu$ is indeed $(\lambda\omega_1^*, ..., \lambda\omega_d^*)$. This shows that the proportions between the positions $\omega_1^*, ..., \omega_d^*$ of the optimal portfolio are the same for all expected level of log returns. The presence of budget or short sale constraints oftentimes complicate the calculation of the optimal portfolios significantly. While the portfolio $(\omega_1^*, ..., \omega_d^*)$ may satisfy these constraints, the same need not be true for the portfolio $(\lambda\omega_1^*, ..., \lambda\omega_d^*)$. Hence the optimal portfolio with expected log return $\lambda\mu$ is not $(\lambda\omega_1^*, ..., \lambda\omega_d^*)$. For certain levels of expected log return, there may not even be a portfolio $\boldsymbol{\omega} \in \mathcal{W}$ that achieves that level.

This illustrates the main reason why we worked with linear portfolios. However, linear portfolios are portfolios of log returns of financial instruments and not the log return of the portfolio of the financial instruments itself. In reality the investor would be concerned with the return of a linear portfolio of the financial instruments $(\mathcal{Z}^{(1)}, ..., \mathcal{Z}^{(d)})$ rather than a linear portfolio of the log returns $(Z^{(1)}, ..., Z^{(d)})$. He would concentrate on the log return of

$$\mathcal{X}(\boldsymbol{\omega}) = \sum_{i}^{d} \omega_i \mathcal{Z}^{(i)}.$$

Therefore, he would consider the expected value and the risk measure of the random variable

$$\log(\mathcal{X}(\boldsymbol{\omega})) = \log\left(\sum_{i}^{d} \omega_i \mathcal{Z}^{(i)}\right).$$

This is just one of many possible examples where the relationship between the risk factors and the log returns of the instruments in the portfolios is nonlinear. This nonlinear relationship complicates the calculation of the log return of the portfolio from the model of the joint distribution of the risk factors. This in turn makes the search for optimal portfolios much more involved. Glasserman et al. (2002) describe methods for computing portfolio VaR$_\alpha$ with heavy tailed risk factors and nonlinear relationships between risk factors and portfolios log returns.

We furthermore only considered the case of a portfolio consisting of two instruments. The reason was that for portfolios with more than two instruments the minimization problem (7.16) would become computationally too extensive to solve directly with the numerical methods that we employed, even for linear portfolios. To give the reader a taste of the difficulties involved, we consider the calculations involved for the case of a portfolio consisting of two instruments.

We calculated the optimal portfolios with respect our model using the Matlab optimization toolbox. The bottleneck in our computations was the calculation of the portfolio quantiles. Remember that the ES$_\alpha$ is a spectral measure of risk that is calculated as

$$ES_\alpha(X) = -\alpha^{-1} \int_0^\alpha F^{\leftarrow}(p)dp \tag{7.19}$$

where $F^{\leftarrow}(p)$ is the quantile of the distribution. Both for our model as well as the model based on the t copula there are no explicit equations for the distribution function of linear portfolio that could be easily evaluated. To calculate the distribution function of the linear portfolio $X = \omega_1 Z^{(1)} + \omega_2 Z^{(2)}$, given by

$$\mathbb{P}[X \leq s] = \mathbb{P}[\omega_1 Z^{(1)} + \omega_2 Z^{(2)} \leq s],$$

from the joint density $f(x,y)$ of $Z^{(1)}$ and $Z^{(2)}$, we need to calculate integrals of the form

$$\mathbb{P}[\omega_1 Z^{(1)} + \omega_2 Z^{(2)} \leq s] \quad = \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{s-\omega_2 y}{\omega_1}} f(x,y) doxy, \text{ if } \omega_1 > 0 \qquad (7.20)$$

$$\mathbb{P}[\omega_1 Z^{(1)} + \omega_2 Z^{(2)} \leq s] \quad = \quad \int_{-\infty}^{\infty} \int_{\frac{s-\omega_2 y}{\omega_1}}^{\infty} f(x,y) doxy, \text{ if } \omega_1 < 0 \qquad (7.21)$$

Since these integrals cannot be calculated analytically, we had to resort to numerical methods, which turned out to be very time and resource consuming. Had we attempted to calculate optimal portfolios for portfolios with $d > 2$ instruments, we would have had to calculate d dimensional analogues of the double integrals (7.20). While an extension of our model to higher dimensional portfolios is straightforward, the numerical calculations of the corresponding d dimensional integrals exceeded the capabilities our resources. For every calculation of the ES$_\alpha$ via (7.19) we needed to calculate a large number of quantiles of the corresponding portfolio distribution in order to get a good numerical approximation of the integral. The numerical integration was carried out with the numerical integration tool provided in Matlab. Typically, it involved the calculation of between 150 to 250 different quantiles. These in turn had to be calculated from the corresponding distribution function by numerically finding the solution of equations of the type

$$\mathbb{P}[\omega_1 Z^{(1)} + \omega_2 Z^{(2)} \leq s] = p$$

We used a bisection algorithm to carry out the calculation of several different quantiles at the same time. The algorithm typically needed between 30 to 40 evaluations of the distribution function $\mathbb{P}[\omega_1 Z^{(1)} + \omega_2 Z^{(2)} \leq s]$ of the portfolio to find the corresponding quantile. This means that in order to calculate the ES$_\alpha$, we needed to numerically calculate about 5,000 to 10,000 numerical double integrals of the form of (7.20). The time it took to carry out these calculations on a PC with a Pentium II processor averaged around 15 to 25 minutes.

The algorithm that we used to find the optimal portfolio with a certain expected log return usually needed 14 to 20 calculations of the ES$_\alpha$ for different positions in the risk factors in order find the portfolio of minimal risk. This means that it typically took us somewhere between 3 and 8 hours to find an optimal portfolio for both the model based on the spectral measure and the model based on the t copula, introduced below. We conclude that while, from a theoretical point of view, there is no difference describing the optimization problems (7.16) and (7.17) for our model and portfolios containing many instruments, in practice the computational resources needed to solve (7.16) and (7.17) forced us to work with portfolios with only two instruments. This clearly demonstrates the need for more efficient algorithms than the ones that we used to calculate portfolio quantiles. It is also the motivation for the development of Monte Carlo methods and approximations used in Glasserman et al. (2002).

In the following sections we discuss the result of solving the politicization problem (7.16), using our model presented in Chapter 6 as the model for the joint distribution of the log returns of the risk factors. We compare the results with two other, simpler models. The first of the alternative models that we considered is the easiest and most popular model for the joint distribution of the risk factors, proposed by the Isometrics (http://www.riskmetrics.com/) group. It assumes that the joint distribution of the log returns of the risk factors $(Z^{(1)}, ..., Z^{(d)})$ is a multivariate normal distribution. As a consequence the log returns of the linear portfolio are also normally distributed. Rockafellar and Uryasev (2000) considered the optimization problem (7.16) based on that model with respect to the ES$_\alpha$, the VaR$_\alpha$ and the standard deviation. They showed that optimal portfolios for all three optimization problems are the same. In other words the portfolios, that, with a given expected log return $\mathbb{E}[X(\boldsymbol{\omega})] = \mu$, minimize the ES$_\alpha$, the VaR$_\alpha$ and the standard deviation are in fact identical. This is true for all significance levels α. In particular the optimal portfolios with respect to, say, the ES$_{5\%}$ is the same as the optimal portfolio with respect to the ES$_{1\%}$. This makes the task of finding optimal portfolios very easy. It is also what makes the multivariate normal approach so attractive. However, the model assumption is unrealistic for two reasons. Firstly, it is widely accepted that the distribution of the log returns of financial time series has regular varying tails. The normal distribution does not have regular varying tails. We saw in Chapter 5, that there is clear indication that the distributions of the datasets under consideration in this thesis have regular varying tails. Secondly, one can show that the multivariate normal distribution has asymptotically independent marginals. See Chapter 5 of Resnick (1986) for a proof. We saw in Chapter 5 that we have clear and convincing evidence against the asymptotical independence of the marginal components in the datasets that we investigate.

Several new approaches and models have been proposed to overcome these obvious shortfalls of the simple multivariate normal model. Most recently the concept of the copula has received significant attention. The copula C of a distribution function F with continuous marginals $F_i, i = 1, ..., d$ is given by

$$F(x_1, .., x_d) = C(F_1(x_1), ..., F_d(x_d)) \Longleftrightarrow C(u_1, ..., u_d) = F(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)).$$

The copula has standardized Uniform[0,1] marginals and describes the dependence structure of the distribution. A comprehensive overview over copulas can be found in Embrechts et al. (2003). As mentioned before, Breymann et al. (2003) compare the quality of the fit of several copulas to certain bivariate financial time series. They used the empirical distributions as approximations for the unknown marginal distributions of the times series under investigation. They found that the best description of the dependence structure of the financial time series under consideration appears to be the t copula, $C^t_{\nu,P}(\mathbf{u})$. The t copula is the copula of the multivariate t distribution. The d dimensional t distribution with ν degrees of freedom, mean vector $\boldsymbol{\mu}$ and positive definite and symmetric dispersion matrix Σ is given by the density

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^d|\Sigma|}}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}}, \mathbf{x}\in\mathbb{R}^d \tag{7.22}$$

As a consequence, the t copula is given by

$$C^t_{\nu,P}(\mathbf{u}) = \int_{-\infty}^{T_\nu^{-1}(u_1)}\cdots\int_{-\infty}^{T_\nu^{-1}(u_d)}\frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^d|P|}}\left(1+\frac{\mathbf{x}^TP^{-1}\mathbf{x}}{\nu}\right)^{-\frac{\nu+d}{2}}d\mathbf{x} \tag{7.23}$$

where P is the matrix with entries $P_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ and $T_\nu^{-1}(\cdot)$ is the quantile of the univariate Student's t distribution with ν degrees of freedom. For a reference, see Embrechts et al. (2003) or Demarta and McNeil (2004). Remember that the univariate Student's t distribution has density

$$t_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{7.24}$$

and its cumulative distribution function is given by

$$T_\nu(x) = \int_{-\infty}^{x} t_\nu(s)ds.$$

In the bivariate case, (7.23) simplifies to

$$C^t_{\nu,\rho}(u_1,u_2) = \int_{-\infty}^{T_\nu^{-1}(u_1)}\int_{-\infty}^{T_\nu^{-1}(u_2)}\frac{1}{2\pi\sqrt{1-\rho^2}}\left(1+\frac{s^2-2\rho st+t^2}{\nu(1-\rho^2)}\right)^{-\frac{\nu+2}{2}}dsdt, \tag{7.25}$$

see Embrechts et al. (2003). Here ρ is the non diagonal element, referred to as the correlation coefficient of P. We should mention that ρ is not the linear correlation of the marginal components. The linear correlation coefficient depends on the marginals that are attached to the copula.

We decided to use the so called meta t distribution as our second alternative to the model developed in Chapter 6. A meta t distribution is a distribution with a t copula $C^t_{\nu,P}(\mathbf{u})$, but its marginal distributions are not necessarily the t_ν. To reflect the fact that the marginal distributions have regular varying tails, we assume that the marginal distributions have the following density:

$$t_{\nu,\mu,\sigma}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma\sqrt{(\pi\nu)}}\left(1 + \frac{(\frac{x-\mu}{\sigma})^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{7.26}$$

The distribution with density $t_{\nu,\mu,\sigma}(x)$ is called a Pearson Type VII distribution. It is the distribution of a linearly transformed Student's t distributed random variable with ν degrees of freedom. If Y is a real random variable with a Pearson Type VII distribution, then we can write $Y = \sigma X + \mu$, where X has Student's t distribution. The additional parameters μ and σ are referred to as the location and scale parameter, respectively. The distribution is symmetric and unimodal. The mode of the distribution is at μ. The tails of the distribution function are regular varying with tail index ν, see for example Embrechts et al. (1997). If $\nu > 1$ the distribution has a finite first moment. In that case, the expectation equals μ. We denote the distribution function associated with the density (7.26) with $T_{\nu,\mu,\sigma}(x)$.

We assume that the joint distribution of the financial instruments has the following form:

$$F(x_1, ..., x_d) = C^t_{\nu,P}(T_{\nu_1,\mu_1,\sigma_1}(x_1), ..., T_{\nu_d,\mu_d,\sigma_d}(x_d)). \tag{7.27}$$

If we denote with

$$c^t_{\nu,P}\left(u_1, ..., u_d\right) = \frac{\partial^d}{\partial u_1 ... \partial u_d} C^t_{\nu,P}(u_1, ..., u_d) = \frac{t^t_{\nu,P}\left(T_\nu^{-1}(u_1), ..., T_\nu^{-1}(u_d)\right)}{\prod_{i=1}^d t_\nu(T_\nu^{-1}(u_i)}$$

the density of the copula $C^t_{\nu,P}(u_1, ..., u_d)$, we see that the joint distribution has the following density

$$f(x_1, ..., x_d) = c^t_{\nu,P}\left(T_{\nu_1,\mu_1,\sigma_1}(x_1), ..., T_{\nu_d,\mu_d,\sigma_d}(x_d)\right)\prod_{i=1}^{d} t_{\nu_i,\mu_i,\sigma_i}(x_i) \tag{7.28}$$

We used maximum likelihood techniques in order to estimate the parameters of the model. We first separately estimate the parameters of the marginal models with the maximum likelihood estimators of the parameters of the Pearson Type VII distribution. We employed a numerical procedure to find these estimates. We used that fact that if a continuous random variable X has cumulative distribution function F, then $Y = F(X)$ has a Uniform[0,1] distribution. Assume that $\widehat{\nu}, \widehat{\mu}, \widehat{\sigma}$ are the estimates of the Pearson Type VII distribution, obtained from the i.i.d. vector of observations $\mathbf{x}$ of the random variable X. Denote

$$\mathbf{q} = T_{\widehat{\nu},\widehat{\mu},\widehat{\sigma}}(\mathbf{x}) \tag{7.29}$$

If the Pearson Type VII distribution with parameters $\widehat{\nu}, \widehat{\mu}$ and $\widehat{\sigma}$ is a reasonable approximation of the distribution of $\mathbf{x}$, then $\mathbf{q}$ is approximately uniformly distributed on the interval [0,1]. We found that, for the data sets considered in this chapter, the Pearson Type VII distribution provided a reasonable fit of the data. We therefore used transformation (7.29) to transform the data into data with an approximate Uniform distribution on [0,1]. We then used the transformed data to obtain parameter estimates for the t copula, given by (7.25).

The problem of finding optimal portfolios based on this model and with respect to a coherent risk measure, such as ES$_\alpha$, is much more difficult than for the multivariate normal distribution. We found that the complexity of the problem is similar to the one that we faced when finding optimal portfolios for our model, based on the spectral measure.

In contrast Rockafellar and Uryasev (2000) showed that finding optimal portfolios with respect to the ES$_\alpha$, working with no particular model for the data, but rather using historical data as an approximation for the true joint distribution of the risk factors, can be achieved using linear programming methods. They found in several case studies, that finding optimal portfolios this way takes less than one minute on an ordinary PC, even for large portfolios of up to 1000 instruments and large sample sizes of up to 20000 observations. Their work was extended by Acerbi and Simonetti (2002) to include portfolio optimization with respect to any spectral measure of risk. Our procedure was much more time consuming compared to the algorithm employed by Rockafellar and Uryasev (2000) because we worked with a particular model, rather than just historical observations. They used empirical quantiles as estimators for the portfolio quantiles in the calculation of

$$ES_\alpha(X) = -\alpha^{-1}\int_0^\alpha F^{\leftarrow}(p)dp,$$

while we based our estimates on our numerical integrations based on our model, as described above.

7.3 Application to Datasets

In this section, we discuss the results of optimizing portfolios with respect to the ES$_\alpha$ using our model, based on the spectral measure, as well as the bivariate normal distribution and the meta t distribution model given by (7.27). We chose the ES$_\alpha$ for several reasons. It is a coherent and spectral risk measure. It has been advertised as the coherent measure that should be used instead of the still popular VaR. However, all our calculations could also be carried out with respect to any other spectral and coherent measure of risk. We simply chose the ES$_\alpha$ because it has already received significant attention in literature. We worked with a significance level of 5% and in one instance with 1%.

7.3.1 Exchange Rates of the Deutsche Mark and the Swiss Franc

The dataset used in this section is part of the Foreign currency dataset studied in Section 5.4. Here, we consider the log returns of the exchange rates of the Deutsche Mark and the Swiss Franc to the US $ from June 1973 to May 1987. Figure 7.1 shows a scatter plot of the log returns of the two currencies. A strong dependence between the log returns is visible. Based on Stărică plots, we determined that k=50 is an appropriate number of upper order statistics to be used in the estimation of the spectral measure and the ranks method selected 158 observation. With the help of the criteria introduced in Section 4.4 and previously used in Section 5, we decided that a 5 component mixture model is an adequate description of the spectral measure. An overview over the estimates of the parameters of the model of the spectral measure is given in Table 7.1. The mixture model has two components in the first quadrant and three components in the third quadrant. Almost all its mass is concentrated in the either the first or the third quadrant, reflecting

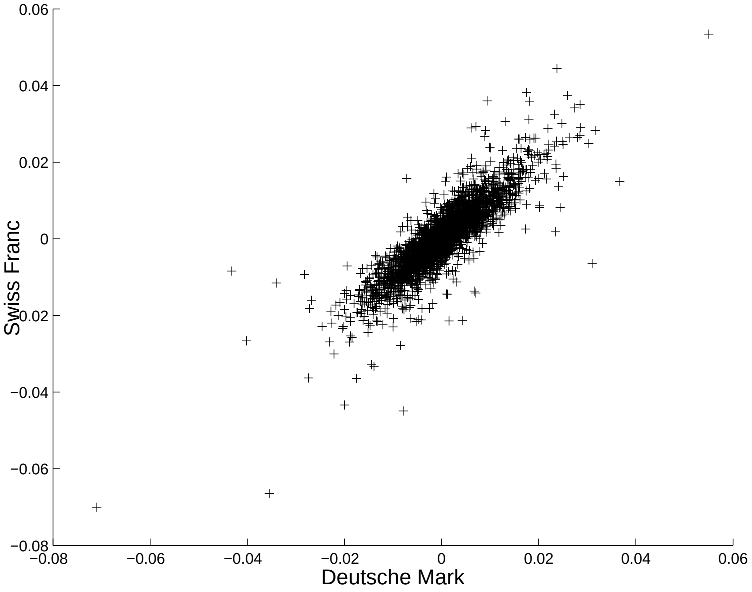

Figure 7.1: *Scatter plot of the log returns of the Deutsche Mark and the Swiss Franc.*

Table 7.1: *Parameters of the Spectral Measure of the Deutsche Mark and the Swiss Franc*

Mean Direction	κ	weight
0.6646	7.11	0.4235
1.4942	494.84	0.0638
3.3562	58.20	0.1405
3.9860	11.82	0.2640
4.6044	110.53	0.1081

the tight dependence between the log returns of two currencies.

The parameters of the marginal distribution of the tails, introduced in Chapter 6 is given in Table 7.2. Concerning the parameters of the GPD models for the tails, we see that for both currencies the thresholds ν_l and ν_r are approximately of the same size. The estimates of the shape parameters ξ_l of the left tails indicate regular varying tails with

Table 7.2: *Parameters of the marginal model of the tails for the log returns of the Deutsche Mark and the Swiss Franc*

Deutsche Mark:	ν_r	ξ_r	β_r	ν_l	ξ_l	β_l
	0.0182	0.0498	0.0054	0.0162	0.4347	0.0035
	μ_T	σ_T	p_1	p_2		
	8.08$10^{-4}$	0.0186	0.2350	0.2371		
Swiss Franc:	ν_r	ξ_r	β_r	ν_l	ξ_l	β_l
	0.0213	-0.0332	0.0066	0.0185	0.4118	0.0044
	μ_T	σ_T	p_1	p_2		
	-0.0019	0.0219	0.2556	0.2246		

tail indexes of 2.3 for the Deutsche Mark and 2.4 in case of the Swiss Franc. On the other hand, it is surprising to see that the estimates of the shape parameters of the right tails indicate that the tails are not very heavy. For the DM, we see that the estimate is ξ_r =0.0498, which corresponds to a tail index estimate of about 20. For the SF, the estimate of the shape parameter of the left tail, ξ_r= -0.0332, is even negative, indicating a distribution with a finite endpoint! It is however important to keep in mind that the GPD is described by all three parameters and not just the shape parameter. The right endpoint of the GPD with the parameter values of the right tail of the Swiss Franc log returns is approximately 0.221. This is well outside of the data range, as the largest log return of the Swiss Franc is approximately 0.053. In Table 5.12 we estimated the tail index of the right the of the Swiss Franc as somewhere between 4.75 and 5 and the tail index of the right tail of the Deutsche Mark between 4 and 4.5. These estimates where mostly based on the results of the QQ-estimator. Based on the Hill plot, estimates as high as 6 are justifiable for both tail indexes. This also indicates that the tails are not very heavy. These differences between the estimates of the tail indexes based on the parametric GPD model and the non parametric estimates of the tail indexes based on the Hill plot and the QQ-estimator indicate the difficulty in accurately assessing the heaviness of the tails.

Finally, the parameters of the model of the body of the distribution are listed in Table 7.3. We see that the tight dependence is also evident in the model of the body given in Table 7.3, as our estimates for the parameters of the bivariate normal model indicate a high correlation coefficient of 0.85349. Overall, the estimated expected log return of the log returns of the Deutsche Mark, implied by the parameter estimates of our model is 1.1308$\cdot 10^{-4}$. For the Swiss Franc, the corresponding value is 2.0617$\cdot 10^{-4}$, considerably larger. The VaR$_{5\%}$ of the log returns of the Deutsche Mark is 0.0103. In other words, our model predicts that 5% of all daily log returns of the Deutsche Mark are losses that

Table 7.3: *Parameters of the body of the Deutsche Mark and the Swiss Franc*

	DM	SF
Mean	8.3610^{-5}	2.0210^{-4}
Std. Dev.	0.0056	0.0065
Correlation		0.85349
Weight of the body		0.9550

are more severe than -0.0103. The $\text{ES}_{5\%}$ of the Deutsche Mark is 0.0156. Recall that, given that the distribution is continuous, the $\text{ES}_{5\%}$ is the expected value of the worst 5% of all observations. The numbers for the Swiss Franc are similar. The $\text{VaR}_{5\%}$ of the Swiss Franc is 0.0117 and the $\text{ES}_{5\%}$ is 0.0184. We see that, while the Swiss Franc has a larger expected log return, it is also riskier. The $\text{ES}_{5\%}$ of the Swiss Franc exceeds that of the Deutsche Mark by about 18%.

We proceeded to find solutions to the optimization problem given by (7.16). We fixed several levels of the expected log return and determined the portfolios whose expected log return matches these levels, while at the same time minimize the $\text{ES}_{5\%}$ among all such portfolios. We mentioned earlier that theoretically, we would only need to calculate the optimal portfolio for one level of expected log return. The risk and the positions in each risk factor are linear functions of the expected level of log return. The reason why we calculated several different optimal portfolios is that we used a numerical approximation to the double integral of the model density in order to calculate the distribution function of the portfolio. These approximations might result in small mistakes. By calculating several optimal portfolios for different levels of expected log return, we can assess the severity of the mistake and get a better estimate of the optimal portfolios. Table 7.4 gives an overview over two of the optimized portfolios. Remember that the proportion between the number of Deutsche Mark and the number of Swiss Francs is the same for all levels of the expected log return. The first column gives the expected level of log return of the portfolio. The expected log returns listed here are the expected log returns of the Deutsche Mark and the Swiss Franc, respectively. The second and third column give the positions in the Deutsche Mark and the Swiss Franc in the portfolio, respectively. The last column lists the estimate of the ES$_{5\%}$ of the portfolio, based on our model. We decided to quantify the riskiness of the optimal portfolio by the parameter β_0 in the following equation:

Table 7.4: *Optimized portfolios of the Deutsche Mark and the Swiss Franc for different levels of expected log returns.*

	Expected Return	Units of DM	Units of SF	Portfolio ES$_{5\%}$
1.	1.1308$\cdot 10^{-4}$	-0.4018	0.7688	0.009601
2.	2.0617$\cdot 10^{-4}$	-0.6878	1.3772	0.017480

$$\text{Expected Shortfall at 5\%} \quad = \quad \beta_0 \cdot (\text{Expected Return}) \tag{7.30}$$

By comparing the different values of the coefficient β_0 for different models, we can compare the risk measures in the optimal portfolios based on the three different models. A higher coefficient indicates higher estimates of the risk of the optimal portfolio for same levels of expected log return. In a similar fashion we also quantify the positions in the optimal portfolio in each of the financial instruments. Since the positions also depend linearly on the expected log return of the portfolio, we can write

$$\begin{aligned} \text{Number of Shares of Risk Factor 1} \quad &= \quad \beta_1 \cdot (\text{Expected Return}) \\ \text{Number of Shares of Risk Factor 2} \quad &= \quad \beta_2 \cdot (\text{Expected Return}) \end{aligned} \tag{7.31}$$

From the optimal portfolios listed in Table 7.4 we estimated these parameters for each level of expected log return. For the coefficient β_0 we obtained values between

84.37 and 85.29, based on the 11 levels of expected log returns that we considered. We see that the variance in these estimates induced by mistakes of the numerical approximations is small. We observed a similarly small variance in the estimates of β_1 and β_2. We used a least squares estimator to obtain the following single estimates for the three coefficients from the results in Table 7.4:

$$\begin{aligned} \text{Expected Shortfall at 5\%} &= 84.59 \cdot (\text{Expected Return}) \\ \text{Number of Deutsch Marks} &= -3291.5 \cdot (\text{Expected Return}) \\ \text{Number of Swiss Francs} &= 6655.5 \cdot (\text{Expected Return}) \end{aligned} \tag{7.32}$$

With the help of these coefficients, we can calculate the optimal portfolio with any desired expected log return and its Expected Shortfall at 5%.

We see that in the optimal portfolios we short the Deutsche Mark and long the Swiss Franc. For every Deutsche Mark that we sell, we have to buy, approximately, two Swiss Francs in order to minimize the risk of the portfolio.

We noted earlier that there is a close dependence between the log returns of the two currencies. A portfolio with a short position in one currency and a long position in the other currency attempts to reduce the variability on the portfolio. Assume for example that the Swiss Franc experiences a large negative log return. Almost certainly, the Deutsche Mark will also experience a large negative log return. The impact of such a large negative log return of the Swiss Franc on the portfolio log return will therefore be softened by the positive log return of the shorted Deutsche Mark position.

An important and interesting question is how much influence the model of the tails, based on the spectral measure and the GPD model for the marginal tails, has in deciding the allocation of the funds in the optimal portfolios. How much influence does the simple model of the body have? We compare the optimal portfolios based on our model with the ones based on the bivariate normal model. The parameters of the bivariate models are estimated by the sample mean, sample standard deviation and sample correlation of the dataset. We obtained the estimates of the parameters listed in Table 7.5. We see that the estimates of the expected log returns are very similar to the estimates based on our model. The estimated ES$_{5\%}$ of the Deutsche Mark, based on the normal model is 0.0146. The corresponding value for the Swiss Franc is 0.0167. As we observed for the estimates based on our model, the Swiss Franc appears to be the riskier asset, but it also seems to be the one with the larger expected log return. These estimates of the ES$_{5\%}$ are about 10% smaller than the ones that we obtained based on our model.

Table 7.5: *Parameters of the bivariate normal model of the Deutsche Mark and the Swiss Franc*

$\widehat{\mu}_{DM}$	=1.17 $\cdot 10^{-4}$	$\widehat{\sigma}_{DM}$	=0.00714
$\widehat{\mu}_{SF}$	=2.14$\cdot 10^{-4}$	$\widehat{\sigma}_{SF}$	=0.00821
Correlation: $\widehat{\rho}$=0.867			

Based on these numbers the portfolios minimizing the ESα, the VaR$_\alpha$ and the variance of the portfolio were calculated for the same 11 expected levels of log return that we used in the calculations using our mixture model of the spectral measure. Based on these 11 optimal portfolios, we calculated the least squares estimators of the coefficients β_0, β_1 and β_2. We obtained the following results:

$$\begin{aligned} \text{Expected Shortfall at 5\%} &= 70.05 \cdot (\text{Expected Return}) \\ \text{Number of Deutsch Marks} &= -4233.1 \cdot (\text{Expected Return}) \\ \text{Number of Swiss Francs} &= 6968.9 \cdot (\text{Expected Return}) \end{aligned} \tag{7.33}$$

Comparing with the results based on our model, we see that the portfolios are fairly similar to the ones that we obtained using our model. However they are not the same portfolios. The ratio between the units of the Deutsche Mark and the Swiss Franc in the portfolios is -0.60743. That means that in an optimal portfolio for every Swiss Franc that we buy, we sell 0.6 Deutsche Marks. Remember that the corresponding ratio was about -.504 for the optimal portfolios based on our model. The differences between the optimal portfolios according to our model and the optimal portfolios according to the normal model is due to the different model of the joint distribution in the tails. Our refined model, based on the spectral measure models joint large log returns differently than the bivariate normal model and therefore implies differences in the optimal portfolios.

We also see that the estimates of the expected shortfalls are significantly smaller than the ones we obtained based on our model. This is evident by comparing the respective coefficients β_0 in (7.32) and (7.33). Based on our model, we estimated $\beta_0 = 84.59$, while based on the normal model, we obtain $\beta_0 = 70.05$. As we explain below in more detail, we found that the empirical estimates of the ES$_\alpha$ were much closer to the ones predicted by our model than the ones based on the normal model. This is not surprising, since we had seen clear evidence that the left tails of both the Deutsche Mark and the Swiss Franc are heavy tailed. Therefore the bivariate model underestimates the size of large losses, since it assumes that the tails are much lighter than they truly are.

We present the results of the optimal portfolios whose expected log returns are equal to the ones used in 7.6.

Table 7.6: *Optimized portfolios of the Deutsche Mark and the Swiss Franc for different levels of expected log returns using a bivariate normal model.*

	Expected Return	Units of DM	Units of SF	Portfolio ES$_{5\%}$
1.	1.1308$\cdot 10^{-4}$	-0.4786	0.7880	0.007921
2.	2.0617$\cdot 10^{-4}$	-0.8727	1.4368	0.014443

Compare these results with the portfolios in Table 7.4. We see that the short position in the Deutsche Mark as well as the long position in the Swiss Franc are somewhat larger than in the portfolios based on our model. Also note that the estimates of the $\text{ES}_{5\%}$ of the portfolios is about 18% smaller than the ones that we obtained based on our model.

Turning to the meta t-distribution model, we first obtain the maximum likelihood estimates of the parameters of the marginal Pearson Type VII distribution. We obtained the parameter estimates listed in Table 7.7. Recall that the degrees of freedom of a Pearson Type VII distribution are equal to the tail index of the corresponding distribution. We see that the corresponding estimates are well in line with what we expect from a reasonable model. However, it is striking that the MLE of the location parameter μ for the Deutsche Mark is negative. Recall that the location parameter of a Pearson Type VII distribution is equal to its expectation. Since the sample mean of the Deutsche Mark is $1.17 \cdot 10^{-4}$, this is disturbing and certainly not very realistic. We conducted a simulation study to investigate the quality and variability of the maximum likelihood estimates of the parameters of the Pearson Type VII distribution. We created 1000 datasets, each with the same sample size as the dataset of the log returns of the Deutsch Mark and the Swiss Franc. Each dataset consisted of i.i.d realizations with a Pearson Type VII distribution with the parameters equal to the estimates of the Deutsche Mark, given in Table 7.7. We found that the distribution of the estimates of the location parameter μ can well

Table 7.7: *Parameters of the meta t distribution of the Deutsche Mark and the Swiss Franc*

$\widehat{\nu}_{DM} = 3.37$	$\widehat{\mu}_{DM} = -2.49 \cdot 10^{-5}$	$\widehat{\sigma}_{DM} = 0.00483$
$\widehat{\nu}_{SF} = 3.45$	$\widehat{\mu}_{SF} = 7.29 \cdot 10^{-5}$	$\widehat{\sigma}_{SF} = 0.00561$
Degrees of freedom of copula:		$\nu_C = 4.2594$
Correlation coefficient of P:		$\rho = 0.889$

be approximated by a normal distribution. The mean and standard deviation of the 1000 estimates of the location parameter are -2.62$\cdot 10^{-5}$ and 1.0$\cdot 10^{-4}$, respectively. The mean is very close to the true value of -2.49$\cdot 10^{-5}$. While this shows that the maximum likelihood estimator is not a bad estimation in general for the parameters of a Pearson Type VII distribution, it is not practical in our case. The estimator of the location parameter needs to be very precise, especially because the true expected mean seems to be so close to 0. We see that the standard deviation is much larger than the absolute value of the true parameter. This means that the estimates of μ are not reliable for our purpose. Indeed, we found in our simulation study that 40% of the estimates of μ have a positive sign, despite the negative sign of the true value. In the light of these results it seems that the negative estimate of the location parameter of the Deutsche Mark is the result of a an estimator that is not precise enough, given the near zero value of the parameter.

The estimated $\mathrm{ES}_{5\%}$ of the Deutsche Mark, based on the parameters in Table 7.7 is 0.017184. The Swiss Franc has an estimated $\mathrm{ES}_{5\%}$ of 0.01953. Both estimates a are a little larger than the estimates based on our model.

Despite the dubious nature of the parameter estimate of the location parameter of the marginal distribution we proceeded to use these estimates to calculate the estimates of the parameters of the copula (7.25). The parameter estimates of the copula are given in Table 7.7. The parameters reflect the close dependence in the dataset. We see that the estimate of the degree of freedom of the copula is significantly different from the estimates of the degree of freedom of the marginals. This indicates that a simple multivariate t distribution is indeed not an adequate description of the data and that a more complicated model, like the one that we used, is indeed necessary.

We calculated optimal portfolios with respect to the $\mathrm{ES}_{5\%}$ for the same 11 different levels of expected log return that we used for our model and the normal model. Based on these results we obtained the following estimates for the coefficients β_0, β_1 and β_2:

$$\begin{aligned}
\text{Expected Shortfall at 5\%} &= 89.18 \cdot (\text{Expected Return}) \\
\text{Number of Deutsch Marks} &= -9998.4 \cdot (\text{Expected Return}) \\
\text{Number of Swiss Francs} &= 10862 \cdot (\text{Expected Return})
\end{aligned} \tag{7.34}$$

As a result of the negative expected value of the Deutsche Mark, the portfolios that minimize the expected shortfall are fairly different from the one that we obtained using our model or the bivariate normal model. The optimal portfolio holds a short position of about 1.08 DM for every Swiss Franc held long. The fact that we hold a large short position in the Deutsche Mark is due to the negative expected log return of the Deutsche Mark and the close dependence between the log return of the two currencies. By holding a short position in the Deutsche Mark we are holding, according to the meta t distribution model, a position with a positive expected log return. In addition it reduces the risk of large negative log returns, since large negative log returns of the Swiss Franc are offset by large positive log return of the short position in the Deutsche Mark. The corresponding estimates of the ES$_{5\%}$ are approximately the same the ones obtained with our model, since the estimate of the coefficient β_0 is fairly close to the one that we obtained based on our model.

The portfolios that minimize the ES$_{5\%}$ for the same levels of expected log return as in Table 7.6 are given in Table 7.8.

Table 7.8: *Optimized portfolios of the Deutsche Mark and the Swiss Franc using a meta t distribution.*

	Expected Return	Units of DM	Units of SF	Portfolio ES$_{5\%}$
1.	1.1308$\cdot 10^{-4}$	-1.2290	1.1304	0.01011
2.	2.0617$\cdot 10^{-4}$	-2.2358	2.0626	0.01838

In order to compare the three models empirically, we compared the optimal portfolios with an expected log return of 4.1235$\cdot 10^{-4}$, twice the expected log return of the Swiss Franc, from each of the three models. It is customary for the purpose of comparing the performance of different models to split the dataset in a so called building sample and a validation sample. The parameters of the model are estimated based on the data in the building sample only, while the performance of the competing models is evaluated using the data in the evaluation sample. We compare the performance of the three models this way in Section 7.4. For the dataset considered in Sections 7.3.1 and 7.3.2, we found that the sample size was not sufficient to allow us to split the dataset and obtain two datasets of sufficient sample size. Remember that we need a dataset of a large sample size to obtain a sufficient number of extreme observations that can then be used to estimate the parameters of the mixture model. We therefore evaluate the competing models using the same dataset was used to estimate the parameters of the models. We found the results to be consistent with the results in Section 7.4.

Since the optimal portfolio according to our model, based on the spectral measure, and the normal model are very similar, their performance is also very similar. The optimal portfolio based on our model had an average log return of 4.29$\cdot 10^{-4}$, while the optimal portfolio based on the normal model had an average log return of 4.12$\cdot 10^{-4}$. The empirical estimate of the ES$_{5\%}$ for the portfolio based on our model is 0.0339. Our model had given us an estimated of 0.034793. The portfolio based on the normal model has an empirical ES$_{5\%}$ of 0.0326, while the corresponding estimate based on the normal model was 0.028887. We see that the normal model seems to underestimate the true risk, while the estimate from our model is very close to the empirical estimate.

In contrast to these numbers, the corresponding optimal portfolio based on the meta t distribution had an average log return of only 3.59$\cdot 10^{-4}$ and an empirical ES$_{5\%}$ of

0.0398, compared with an estimated value of 0.036757 based on the meta t model. Its average log return is significantly lower than predicted by the model and the other two portfolios. In addition it also has a much higher risk, as measured by the empirical $\mathrm{ES}_{5\%}$.

7.3.2 The Log Returns of IBM and Intel

While the log returns of the Deutsche Mark and the Swiss Franc exhibit a very tight overall dependence, the dataset of the log returns of IBM and Intel does not show such a tight dependence. This is evident from the scatter plot of the log returns, given in Figure 7.2. The same statement can be made about the dependence in the tails. The parameters of the mixture model of the spectral measure of the log returns of the Deutsche Mark and the Swiss Franc indicate a tighter dependence than the corresponding parameters for the model of the log returns of IBM and Intel.

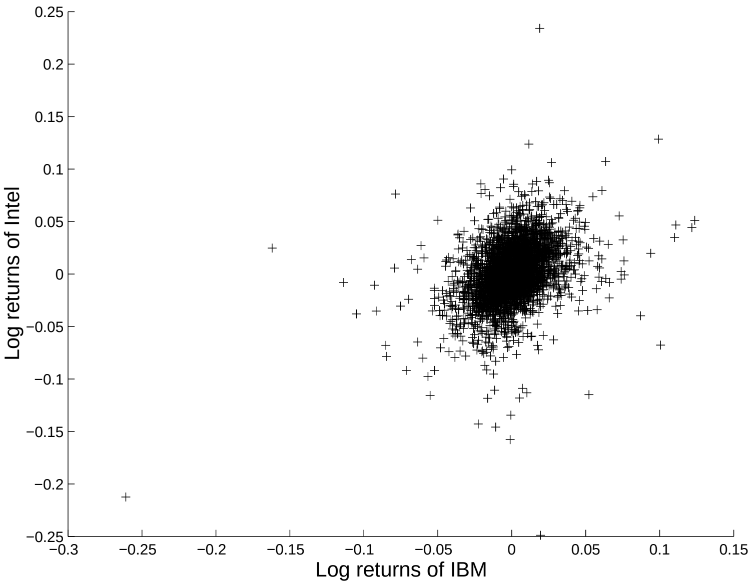

Figure 7.2: *Scatter plot of the log returns of the stocks of IBM and Intel.*

We already discussed in a Section 5.2 how we determined the model for the spectral measure of the joint distribution of the log returns of the stock prices of IBM and Intel. Table 7.9 lists the parameters of that model.

Table 7.9: *Parameters of the Spectral Measure of IBM and Intel*

Mean Direction	κ	weight
0.02	690.99	0.1155
0.08	39.75	0.1075
1.11	8.25	0.1037
1.54	281.68	0.1792
3.19	217.60	0.1548
3.88	3.74	0.1862
4.68	215.95	0.1528

The parameters of the marginal distribution of the tails, are given in Table 7.10. We see from the values in Table 7.10 that the GPD models for the marginal tails indicate

Table 7.10: *Parameters of the Marginal Model of the tails for the log returns of IBM and Intel*

IBM	ν_r	ξ_r	β_r	ν_l	ξ_l	β_l
	0.0360	0.2175	0.0129	0.0344	0.4261	0.0097
	μ_T	σ_T	p_1	p_2		
	0.0015	0.0233	0.2458	0.2613		
Intel:	ν_r	ξ_r	β_r	ν_l	ξ_l	β_l
	0.0513	0.1211	0.0155	0.0523	0.3008	0.0163
	μ_T	σ_T	p_1	p_2		
	-0.0097	0.0483	0.2215	0.1483		

that for both stocks the left and the right tails are regular varying. The left tails appear to be heavier than the right tails. The estimates of tail indexes of the left tail, based on the shape parameters of the GPD models, are 2.3 for IBM and 3.3 for Intel.

Finally, the parameters of the model of the body of the distribution are listed in Table 7.11. We see that the linear correlation between the observations in the body of the joint distribution of IBM and Intel is much smaller than the corresponding value for the Deutsche Mark and the Swiss Franc in the last section.

Table 7.11: *Parameters of the body of the Deutsche Mark and the Swiss Franc*

	IBM	Intel
Mean	1.95·10^{-4}	1.39·10^{-3}
Std. Dev.	0.013807	0.02118
Correlation		0.34775
Weight of the body		0.9572

The estimate of the $\mathrm{ES}_{5\%}$ of the log returns of IBM, based on our model is 0.0401 and the corresponding estimate for the log returns of Intel is 0.0609. Also, based on our model, the expected log return of the log returns of IBM is 2.03·10^{-4} and the corresponding value for Intel is 1.105·10^{-3}. We see that while Intel is riskier it also has a larger expected log return.

As we did in the last section with the Deutsche Mark and the Swiss Franc, we calculated the optimal portfolios for different levels of the expected log return. We then use a least squares estimator to estimate the coefficients between the expected log return and the risk measure and the positions of the two stocks in the portfolio. We obtain the following estimates:

$$\begin{aligned} \text{Expected Shortfall at 5\%} &= 54.318 \cdot (\text{Expected Return}) \\ \text{Number of shares fo IBM} &= -244.75 \cdot (\text{Expected Return}) \\ \text{Number of shares of Intel} &= 949.33 \cdot (\text{Expected Return}) \end{aligned} \tag{7.35}$$

An overview over the optimal portfolios, whose expected log returns equal the expected log returns of the two stocks, is given in Table 7.12. We see that the optimal portfolio is achieved by short selling a small amount of IBM stock short and buying the stock of Intel. For every stock of IBM that we sell, we have to buy, approximately, four stocks of Intel, in order to minimize the risk of the portfolio. Since the expected log return of the stock of Intel is about 5 times as large as the expected log return of the stock of IBM, it seems that the optimal portfolio is achieved by buying about the amount of shares of Intel necessary to achieve the desired expected log return and reduce the risk buy selling a small fraction of IBM's stock short. Even though the tail dependence between the two stocks is not as tight as the dependence between the two currencies in the last section, large negative log returns of Intel tend to happen at the same time as large negative log returns of IBM. Therefore, a short position in IBM reduces the severity of the negative log returns of the portfolio caused by the large negative log returns of Intel.

Table 7.12: *Optimized portfolios of IBM and Intel based on our model.*

	Expected Return	Units of IBM	Units of Intel	Portfolio ES$_{5\%}$
1.	2.0346$\cdot 10^{-4}$	-0.0509	0.1933	0.0111
2.	1.1058$\cdot 10^{-3}$	-0.2656	1.0489	0.0601

We compare these results with the optimal portfolios based on the bivariate normal model. We obtained the estimates of the parameters of the bivariate normal distribution that we present in Table 7.13. Comparing these estimates with the parameter estimates

Table 7.13: *Parameters of the bivariate normal model of IBM and Intel.*

$\widehat{\mu}_{IBM}$	$=2.19 \cdot 10^{-4}$	$\widehat{\sigma}_{IBM}$	$=0.01888$
$\widehat{\mu}_{Intel}$	$=1.074 \cdot 10^{-3}$	$\widehat{\sigma}_{Intel}$	$=0.02729$
Correlation: $\widehat{\rho}=0.373$			

that we obtained for the distribution of the log returns of the Deutsche Mark and the Swiss Franc, we see that the linear correlation coefficient is much smaller. This indicates that the dependence between the log returns of the two stocks is not as strong as the dependence between the two currencies. We also see that the expected log return of Intel is almost 5 times as large as the expected log return of IBM. At the same time the standard deviation of Intel is only about 50% larger than the standard deviation of IBM. The estimates of the $\text{ES}_{5\%}$ based on the normal model are 0.0387 for the log return of IBM and 0.0552 for the log returns of Intel. Both numbers are very similar to the estimates that we obtained based on our model. Based on the parameters in Table 7.13 we calculated the portfolio that minimizes the ES_α for given levels of expected log return. The resulting estimates of the coefficients β_0, β_1 and β_2 are:

$$\begin{aligned} \text{Expected Shortfall at 5\%} &= 51.235 \cdot (\text{Expected Return}) \\ \text{Number of shares fo IBM} &= -120.26 \cdot (\text{Expected Return}) \\ \text{Number of shares of Intel} &= 955.8 \cdot (\text{Expected Return}) \end{aligned} \tag{7.36}$$

The optimal portfolios whose expected log return is equal to the expected log returns of the two stocks are given in Table 7.14. The results are fairly similar to the results based on our model. We short about one share of IBM for every eight shares of Intel that we buy. The reasons appears to be to be the same as for the short positions of IBM in the optimal portfolios based on our model. Similar to the case of the Deutsche Mark and the Swiss Franc, the estimates of the ES_α of the optimal portfolios are smaller than

Table 7.14: *Optimized portfolios of IBM and Intel based on a bivariate normal model.*

	Expected Return	Units of IBM	Units of Intel	Portfolio ES$_{5\%}$
1.	2.0346$\cdot 10^{-4}$	-0.0245	0.1945	0.0104
2.	1.1058$\cdot 10^{-3}$	-0.1330	1.0569	0.0567

the estimates that we obtained with our model. This is again due to the lighter tails of the bivariate normal model. However, the coefficients β_0 are very close to each other, so that the corresponding estimates of the risk measure based on the two different models are very close.

For the meta t distribution we obtained the parameter estimates given in Table 7.15.

Table 7.15: *Parameters of the meta t distribution of IBM and Intel*

$\widehat{\nu}_{IBM} = 3.92$	$\widehat{\mu}_{IBM} = 3.36 \cdot 10^{-5}$	$\widehat{\sigma}_{IBM} = 0.01302$
$\widehat{\nu}_{Intel} = 5.26$	$\widehat{\mu}_{Intel} = 1.2731 \cdot 10^{-3}$	$\widehat{\sigma}_{Intel} = 0.021304$
Degrees of freedom of copula:		$\nu_C = 7.6499$
Correlation coefficient of P:		$\rho = 0.408$

We see that the expected log return of Intel is much larger than the one of IBM. We also see that the degrees of freedom of the distribution of Intel is considerably larger than the one of IBM. This means that the Pearson Type VII distribution indicates that the tails of Intel are much lighter than the tails of IBM. This confirms our findings based on the estimates of the shape parameters of the GPD models. However, the estimates of the tail index are very different. We mentioned before that the tail indexes that the GPD fits to the left tail of IBM and Intel implied are 2.3 and 3.3 respectively. These estimates are significantly smaller than the tail index estimates listed in Table 7.15.

Since the left tail of Intel seems to have the lighter tail according to the Pearson

Type VII model, it is a little surprising to see that the $\mathrm{ES}_{5\%}$ of IBM is actually smaller. It's value is 0.0421, compared with the $\mathrm{ES}_{5\%}$ of Intel, which is 0.0586. Both estimates are about of the same size as the estimates that we obtained based on our model and the normal model. As for the parameter estimates of the copula, both the degrees of freedom and the correlation coefficient indicate, that the dependence is not as close as the dependence of the two currencies considered in the last section.

Based on these parameters we calculated the portfolios that minimize the $\mathrm{ES}_{5\%}$ for different levels of expected log return. Based on these portfolios, we obtained the following estimates for the coefficients β_0, β_1 and β_2 as before: β_0, β_1 and β_2 are:

$$\begin{aligned} \text{Expected Shortfall at 5\%} &= 42.649 \cdot (\text{Expected Return}) \\ \text{Number of shares fo IBM} &= -438 \cdot (\text{Expected Return}) \\ \text{Number of shares of Intel} &= 796.98 \cdot (\text{Expected Return}) \end{aligned} \tag{7.37}$$

We again give the two optimal portfolios, whose expected log return equals the expected log return of the two stocks in Table 7.16. As for the portfolios based on our model and the bivariate normal distribution, we short the stock of IBM and long Intel's stock. However, the ratio of the two positions is different from the previous two cases. We only long about 2 shares of Intel for every share of IBM that we short. A possible explanation comes again from the fact that the expected log return of IBM is much smaller than the one of Intel, while its risk measure is only moderately smaller. It is therefore the best

Table 7.16: *Optimized portfolios of IBM and Intel for different levels of expected log return based on the meta t model.*

	Expected Return	Units of IBM	Units of Intel	Portfolio $\mathrm{ES}_{5\%}$
1.	2.0346$\cdot 10^{-4}$	-0.0891	0.1622	0.0087
2.	1.1058$\cdot 10^{-3}$	-0.4847	0.8813	0.0472

strategy to short the stock of IBM in an attempt to reduce the impact of the negative large log returns of the stock of Intel on the log return of the portfolio.

We again compared the resulting portfolios empirically the same we compared the corresponding portfolios of the Deutsch Mark and the Swiss Franc. The optimal portfolio with an expected log return of 3.3175$\cdot 10^{-3}$, based on our model, has an empirical $\mathrm{ES}_{5\%}$ of 0.1848. The corresponding estimated $\mathrm{ES}_{5\%}$, based on our model, is 0.1802. The average log return of the portfolio is 3.2034$\cdot 10^{-3}$. The optimal portfolio based on the normal model has a slightly larger empirical $\mathrm{ES}_{5\%}$ of 0.1910 and an average log return of 3.317$\cdot 10^{-3}$. The estimated $\mathrm{ES}_{5\%}$ based on the normal model for that portfolio is however only 0.16997. Finally the optimal portfolio based on the meta t model has a smaller empirical $\mathrm{ES}_{5\%}$ of 0.15027. Its estimated $\mathrm{ES}_{5\%}$ based on the meta t model is 0.14149. The average log return is however also much smaller, namely 2.5203$\cdot 10^{-3}$. The portfolios based on our model and the bivariate model are approximately equivalent. The optimal portfolio based on the normal model, however, has an empirical risk measure that is 12% larger than predicted by the normal model. The model based on the meta t distribution suffers from the fact that it does not achieve the desired expected log return. This is again a consequence of the unprecise estimator for the location parameter. We observed these shortcomings of the normal model and the meta t model already in the previous section for the Deutsche Mark and the Swiss Franc.

7.4 Comparison of the Models Using BMW and Siemens Stock Returns

In order to better compare the three different models, we used the BMW-Siemens dataset to conduct an empirical study by splitting the dataset. We estimated the parameters of the three models using the first 70% of the observations. We refer to this dataset with a sample size of 4302 as the “model building sample”. We calculated the optimal portfolios based on these models. We then evaluated the performance of these portfolios using the remaining 30% of the observations in the dataset. The dataset containing these observations is referred to as the “validation sample”. We calculated the portfolios that minimize the $\mathrm{ES}_{5\%}$. In addition, we also calculated the portfolios that minimize the $\mathrm{ES}_{1\%}$ and compared the performance of these portfolios as well.

7.4.1 Parameter Estimation and Calculation of the Optimal Portfolios

We first determined the appropriate model for the spectral measure of the joint distribution. Based on a Stărică plot, we determined the number of upper order statistics used in the estimation of the spectral measure. We found that $k = 60$ was the best choice. The ranks method selected 207 data points. We chose a mixture model with 7 components, based on the results from the Likelihood Ratio test with a significance level of 1%. The parameters estimates of the model are given in Table 7.17. The BIC suggested a model with 6 components, while the LR test with a significance level of 5% and the AIC suggested 8 components. We see that there are 5 components, numbered 1,3,4,5 and 7 in Table 7.17, modelling the clusters of the points found near the directions of the four axis of the cartesian coordinate system. Components 5 and 7 model the cluster located at the negative y axis. We have two components, numbered 2 and 6 modelling the dependence in the first and third quadrant, respectively.

The parameter estimates of the marginal models of the tail component of the model are listed in Table 7.18. The estimates of tail indexes of the left tail, based on the GPD models, are 5.4 for BMW and 3.1 for Siemens. The right tail of BMW also appears to

Table 7.17: *Parameters of the Spectral Measure of BMW and Siemens*

	Mean Direction	κ	weight
1.	0.0539	217.60	0.1755
2.	0.9680	5.49	0.1892
3.	1.5408	600.21	0.1425
4.	3.2584	82.66	0.1638
5.	4.6739	697.24	0.0839
6.	4.0272	10.80	0.1778
7.	4.5089	427.19	0.0673

Table 7.18: *Parameters of the Marginal Model of the tails of BMW and Siemens*

BMW:	ν_r	ξ_r	β_r	ν_l	ξ_l	β_l
	0.0394	0.2048	0.0113	0.0386	0.1852	0.0112
	μ_T	σ_T	p_1	p_2		
	-0.0045	0.0345	0.2335	0.2041		
Siemens:	ν_r	ξ_r	β_r	ν_l	ξ_l	β_l
	0.0285	-0.0156	0.0087	0.0284	0.3159	0.0086
	μ_T	σ_T	p_1	p_2		
	-0.0043	0.0264	0.2376	0.2029		

be regular varying. The corresponding tail index estimate based on the GPD model is about 4.9. On the other hand, the estimate of the shape parameter for the right tail of Siemens is negative. This indicates, that the tail has a finite right endpoint. The finite right endpoint of the GPD with parameters $\nu_r = 0.0285$, $\xi_r = -0.0156$ and $\beta_r = 0.0087$ is 0.5848. This is well outside of the range of the data, as the largest log return for the stock of Siemens in the model building dataset is 0.0730. The situation is thus similar to the case of the right tail of the Swiss Franc, discussed in Section 7.3.1.

Finally, the parameters of the model of the body of the distribution are listed in Table 7.19. The numbers are similar to the ones we observed in the case of the log returns of IBM and Intel.

Table 7.19: *Parameters of the body of the Deutsche Mark and the Swiss Franc*

	BMW	Siemens
Mean	4.18·10^{-4}	3.41·10^{-4}
Std. Dev.	0.0117	0.0090
Correlation		0.5527
Weight of the body		0.9519

Based on our model, the stock of BMW has an expected log return of 3.64·10^{-4}. The expected log return of the stock of Siemens is 2.31·10^{-4}. The ES$_{5\%}$ of the log returns of BMW is 0.0337. The corresponding estimate for the ES$_{1\%}$ is 0.0577. For Siemens, the corresponding estimates are 0.0260 for the ES$_{5\%}$ and 0.0450 for the ES$_{1\%}$. The stock of BMW is riskier, but also has a larger expected log return than the stock Siemens.

Based on this model, we determined the optimal portfolios for several different levels of expected log returns. Based on these results, we found the following relationships between the expected log return of the portfolio and the positions in the optimal portfolio and the corresponding risk measure.

$$\begin{aligned} \text{Expected Shortfall at 5\% Level} &= 90.1216 \cdot \text{(Expected Return)} \\ \text{Number of Shares of BMW} &= 2088.5 \cdot \text{(Expected Return)} \\ \text{Number of shares of Siemens} &= 1028.8 \cdot \text{(Expected Return)} \end{aligned} \tag{7.38}$$

The optimal portfolios therefore contain 2.0298 shares of BMW for every share of Siemens. Table 7.20 gives an overview over the optimal portfolios whose expected log returns are equal to the ones of the two stocks, based on our model. The expected log return of the first portfolio is equal to the expected log return of Siemens and the second has the same expected log return as BMW.

Table 7.20: *Optimized portfolios of BMW and Siemens for different levels of expected log return based on our model.*

	Expected Return	Shares of BMW	Shares of Siemens	Portfolio ES$_{5\%}$
1.	2.31·10^{-4}	0.4835	0.2381	0.02086
2.	3.64·10^{-4}	0.8350	0.3752	0.03287

	Expected Return	Shares of BMW	Shares of Siemens	Portfolio ES$_{1\%}$
1.	2.31·10^{-4}	0.4820	0.2404	0.03657
2.	3.64·10^{-4}	0.7595	0.3788	0.05763

For the portfolios that minimize the ES$_{1\%}$, we obtained the following estimates of the coefficients between the expected level of log return and the risk and the number of shares of each stock in the optimal portfolios.

$$\begin{aligned} \text{Expected Shortfall at 1\% Level} &= 158.01 \cdot \text{(Expected Return)} \\ \text{Number of Shares of BMW} &= 2082.3 \cdot \text{(Expected Return)} \\ \text{Number of shares of Siemens} &= 1038.6 \cdot \text{(Expected Return)} \end{aligned} \tag{7.39}$$

This means that for every share of Siemens we have to include 2.0049 shares of BMW in a optimal portfolio. We see that the optimal portfolios with respect to the $\mathrm{ES}_{5\%}$ and the $\mathrm{ES}_{1\%}$ are very similar. Table 7.20 again provides an overview over the optimal portfolios whose expected log returns are equal to the expected log returns of the two stocks. We see that for the portfolios whose expected log returns match the expected log return of Siemens, the risk is only about 80% of the risk of the stock of Siemens, both for the $\mathrm{ES}_{5\%}$ and the $\mathrm{ES}_{1\%}$. On the other hand, the portfolios whose expected log return equals the expected log return of BMW, the risk has only been very slightly reduced.

For the bivariate normal model, we obtained the parameter estimates given in Table 7.21. Based on these numbers we estimate that the $\mathrm{ES}_{5\%}$ of the log returns of BMW is 0.0306 and that the $\mathrm{ES}_{1\%}$ is 0.0396. The corresponding numbers of the log returns of Siemens are 0.023226 for the $\mathrm{ES}_{5\%}$ and 0.03008 for the $\mathrm{ES}_{1\%}$. While the estimates of the $\mathrm{ES}_{5\%}$ are fairly close to the ones that we obtained based on our model, the estimates of the $\mathrm{ES}_{1\%}$ are much smaller. This is due to the fact that our model has regular varying left tails, while the tails of the normal model are much lighter.

Table 7.21: *Parameters of the bivariate normal model of BMW and Siemens.*

$\widehat{\mu}_{BMW}$	= 3.54 $\cdot 10^{-4}$	$\widehat{\sigma}_{BMW}$	=0.01501
$\widehat{\mu}_{Siemens}$	= 2.38 $\cdot 10^{-4}$	$\widehat{\sigma}_{Siemens}$	=0.01138
Correlation: $\widehat{\rho}$=0.60077			

As for our model, BMW is the riskier position but also has a greater expected log return. The correlation between the log returns of the two stocks is larger than what we observed for IBM and Intel, but still smaller than the one between the Deutsche Mark and the Swiss Frank.

We obtained the following estimates of the relationships between the expected log return of the portfolio and the positions in the optimal portfolio and the corresponding risk measure.

$$\begin{aligned}\text{Expected Shortfall at 5\% Level} &= 81.25 \cdot \text{(Expected Return)}\\ \text{Number of Shares of BMW} &= 1832.2 \cdot \text{(Expected Return)}\\ \text{Number of shares of Siemens} &= 1472.5 \cdot \text{(Expected Return)}\end{aligned} \tag{7.40}$$

We see that the proportion between the number of shares of BMW and the number of shares of Siemens is fairly different from the one we saw in (7.38) for the optimal portfolios based on our model with respect to the $\text{ES}_{5\%}$. The optimal portfolio contains 1.24 shares of BMW for every share of Siemens. The estimated $\text{ES}_{5\%}$ is also consistently lower than the estimates based on our model, because our model has heavier tails than the normal model. As we mentioned before, the proportion of the number of shares is the same in the portfolio minimizing the $\text{ES}_{1\%}$ as it is in the portfolio minimizing the $\text{ES}_{5\%}$. The relationship between the $\text{ES}_{1\%}$ and the expected log return is given by the following equation.

$$\text{Expected Shortfall at 1\% Level} = 105.28 \cdot \text{(Expected Return)} \tag{7.41}$$

As for the $\text{ES}_{5\%}$, the estimates of the $\text{ES}_{1\%}$ based on our model are larger than the ones based on the normal model. For the $\text{ES}_{5\%}$, the estimates of the risk based on our model are approximately 10% larger than the ones based on the normal model. For the $\text{ES}_{1\%}$ the estimates based on our model are even 50% larger than the ones based on the normal model. We illustrate this again in Table 7.22 by listing the optimal portfolios based on the normal model with the same expected log return as the ones in Table7.20.

For the meta t distribution model, we obtained the parameters estimates presented in Table 7.23. Both marginal distributions have a similar tail index close to 3. We see that the location parameter of the model of Siemens is larger than the one of BMW. Contrary

Table 7.22: *Optimized portfolios based on the normal model.*

	Expected Return	Shares of BMW	Shares of Siemens	Portfolio $\mathrm{ES}_{5\%}$
1.	$2.31{\cdot}10^{-4}$	0.4241	0.34086	0.018808
2.	$3.64{\cdot}10^{-4}$	0.6683	0.53713	0.029638

	Expected Return	Shares of BMW	Shares of Siemens	Portfolio $\mathrm{ES}_{1\%}$
1.	$2.31{\cdot}10^{-4}$	0.4241	0.34086	0.024371
2.	$3.64{\cdot}10^{-4}$	0.6683	0.53713	0.038404

Table 7.23: *Parameters of the meta t distribution of BMW and Siemens*

$\widehat{\nu}_{BMW} = 2.84$	$\widehat{\mu}_{BMW} = 9.61 \cdot 10^{-5}$	$\widehat{\sigma}_{BMW} = 0.00935$
$\widehat{\nu}_{Siemens} = 3.033$	$\widehat{\mu}_{Siemens} = 3.176 \cdot 10^{-4}$	$\widehat{\sigma}_{Siemens} = 0.00727$
Degrees of freedom of copula:		$\nu_C = 4.9437$
Correlation coefficient of P:		$\rho = 0.63188$

to the other models, the meta t distribution model hence claims that Siemens has a larger expected log return than BMW. Based on the model for the marginal distributions the ES$_{5\%}$ of BMW is 0.0375 and the ES$_{1\%}$ is 0.0694. Siemens has an estimated ES$_{5\%}$ of 0.0276 and the ES$_{5\%}$ is estimated as 0.0499. All these numbers are comparable with the numbers we obtained based on our model. The major difference compared to the other two models is again found in the estimates of the expected log return.

Based on the parameter estimates of Table 7.23, we calculated the optimal portfolios for several different levels of expected log return. From these we obtained the following estimates of the relationships between the expected log return of the portfolio and the positions in the optimal portfolio and the corresponding risk measure, based on the meta t model.

$$\begin{aligned}
\text{Expected Shortfall at 5\% Level} &= 77.22 \cdot \text{(Expected Return)} \\
\text{Number of Shares of BMW} &= -1232 \cdot \text{(Expected Return)} \\
\text{Number of shares of Siemens} &= 3521.6 \cdot \text{(Expected Return)}
\end{aligned} \tag{7.42}$$

For the portfolios that are optimal with respect to the ES$_{1\%}$, the corresponding equations are

$$\begin{aligned}
\text{Expected Shortfall at 5\% Level} &= 139.19 \cdot \text{(Expected Return)} \\
\text{Number of Shares of BMW} &= -1268.2 \cdot \text{(Expected Return)} \\
\text{Number of shares of Siemens} &= 3532.5 \cdot \text{(Expected Return)}
\end{aligned} \tag{7.43}$$

These numbers are very different than the ones based on our model and the normal model. The stock of BMW has a smaller log return and a larger risk compared to the stock of Siemens. It is therefore not surprising to see that both for the ES$_{5\%}$ and the ES$_{1\%}$ the optimal portfolios are achieved by holding a short position in the stock of BMW and a long position in the stock of Siemens. The proportions between the number of shares held in an optimal portfolio are very similar in both cases. For every share that we sell short in a portfolio that is optimal with respect to the $\mathrm{ES}_{5\%}$, we buy about 2.85 shares of the stock of Siemens. For the portfolios that are optimal with respect to the $\mathrm{ES}_{1\%}$, the corresponding ratio is 2.25. The coefficient between the expected log return and the risk of the optimal portfolio for the $\mathrm{ES}_{1\%}$ is larger than for the normal model. Surprisingly, the same is not true for the coefficient for the $\mathrm{ES}_{5\%}$, which is smaller than its counterpart based on the normal model. Both coefficients are smaller than the corresponding coefficients based on our model. As we mentioned before, this means that the ES_{α} estimates based on the meta t distribution model are smaller than the ones based on our model, but the $\mathrm{ES}_{1\%}$ estimates are larger than the ones based on the normal distribution model. This point is illustrated in Table 7.24, which lists the optimal portfolios and the corresponding estimates of the risk measures, based on the meta t distribution model, for the same expected log returns as in Table 7.20 and 7.22.

Table 7.24: *Optimized portfolios based on the meta t model.*

	Expected Return	Shares of BMW	Shares of Siemens	Portfolio $\mathrm{ES}_{5\%}$
1.	$2.31{\cdot}10^{-4}$	-0.2851	0.8152	0.017875
2.	$3.64{\cdot}10^{-4}$	-0.4494	1.2846	0.028168

	Expected Return	Shares of BMW	Shares of Siemens	Portfolio $\mathrm{ES}_{1\%}$
1.	$2.31{\cdot}10^{-4}$	-0.29357	0.81772	0.032221
2.	$3.64{\cdot}10^{-4}$	-0.46261	1.2886	0.050773

7.4.2 Moment of Truth: Comparison of the Performance of the Models

We now compare the performance of the different portfolios based on their performance using the validation sample. It consists of the last 1844 observations of the entire dataset. Remember that these observation were not included in the model building sample. Figure 7.3 shows a scatter plot of the model building and the validation sample. We see that the dependence in the tails seems to be more pronounced in the validation sample than in the model building sample. An indication are the joint large positive and negative log returns visible in the right hand plot of Figure 7.3.

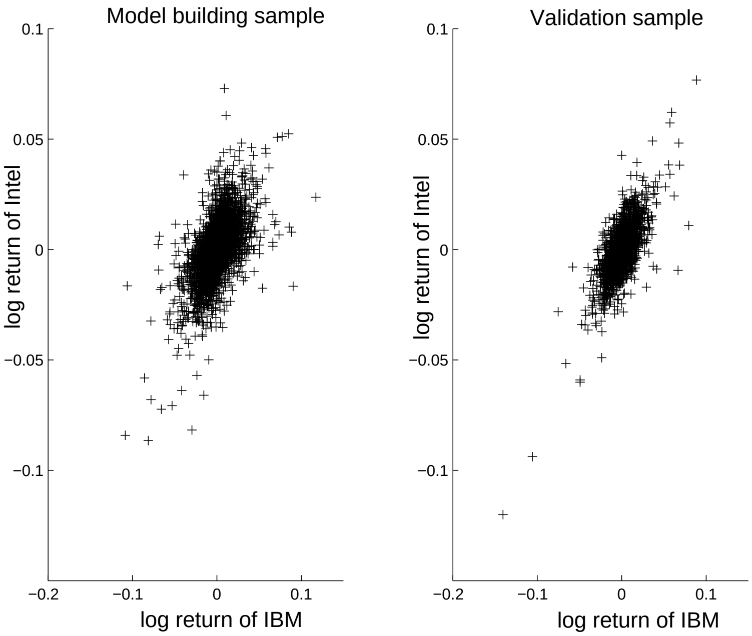

Figure 7.3: *Scatter plot of the "model building" and the "validation" sample of BMW and Siemens*

For each of the optimal portfolios that we found in the last chapter, we calculated the empirical mean and the corresponding empirical estimate of the risk measure based on the observations in the validation sample. Since the sample size of the validation sample is 1844, the empirical estimates of the $\text{ES}_{5\%}$ are based on the 92 largest negative log returns of the corresponding portfolios. This is a sample size that gives us confidence about the validity of the corresponding estimates. On the other hand, the $\text{ES}_{1\%}$ is only based on 18 observations. Since we did not regard this sample size as sufficient to obtain reliable estimates of the ES, we additionally estimated both risk measures based on a parametric model. We based these estimate on a GPD fit of the left tail of the log returns of the portfolio. We found that GPD fits based on the 100 largest negative observations provided fits to the tail distributions.

Since the portfolios and the resulting risk measures depend linearly on the expected log return, the specific level expected log return used in the analysis is irrelevant. We decided to use an expected log return of $3.64 \cdot 10^{-3}$. That is 10 times the estimated expected log return of the log return of BMW, based on our model. Table 7.25 gives an overview over the performance of the different optimal portfolios in the validation sample. We see that none of the three portfolios reaches the expected log return, $3.64 \cdot 10^{-3}$, that was predicted by the respective models. The average log returns of the portfolios based on our model and the normal model are fairly close, while the portfolio based on the meta t distribution has a much smaller average log return. We see that the empirical estimates of the $\text{ES}_{5\%}$ and the estimates of the $\text{ES}_{5\%}$ based on the GPD models are very

Table 7.25: *Performance of the optimal portfolios with respect to the $ES_{5\%}$*

	Average Return	Emp. $\text{ES}_{5\%}$	GPD $\text{ES}_{5\%}$	Predicted $\text{ES}_{5\%}$
Our Model	$2.9317 \cdot 10^{-3}$	0.31981	0.32063	0.3287
Bivariate Normal	$2.8937 \cdot 10^{-3}$	0.32852	0.32918	0.2960
Meta t	$0.6024 \cdot 10^{-3}$	0.2506	0.25164	0.2817

similar. This indicates that the estimates are indeed reliable and accurate. The differences between the estimates for the different portfolios are of a larger magnitude than the differences between the empirical estimates and the GPD based estimates. The optimal portfolio based on our model has a larger average log return and at the same time a smaller $\text{ES}_{5\%}$ than the optimal portfolio based on the normal model. Even though the differences are not dramatic, it shows that the portfolio based on our model outperforms the one based on the normal model. As for the portfolio based on the meta t model, its $\text{ES}_{5\%}$ is about 23% smaller than the $\text{ES}_{5\%}$ of the other two portfolios. But at the same time, its expected log return is only about 20% of the corresponding log returns of the other two portfolios. It is also striking that only our model was able to accurately predict the $\text{ES}_{5\%}$, based on the model building sample. The corresponding numbers, that we have already mentioned above, are again listed in the last column of Table 7.25. While the normal model has an $\text{ES}_{5\%}$ that is about 11% larger than predicted, the portfolio based on the meta t model overestimates the $\text{ES}_{5\%}$. The estimates of the $\text{ES}_{5\%}$ based on the validation sample is only about 89% of the predicted $\text{ES}_{5\%}$ based on the model building sample.

Table 7.26 gives an overview over the performance in the validation sample of the different portfolios that were optimized with respect to the $\text{ES}_{1\%}$. The picture is very similar for the portfolios that were optimized with respect to the $\text{ES}_{1\%}$. The differences

Table 7.26: *Performance of the optimal portfolios with respect to the $ES_{1\%}$*

	Average Return	Emp. $\text{ES}_{1\%}$	GPD $\text{ES}_{1\%}$	Predicted $\text{ES}_{1\%}$
Our Model	$2.9302 \cdot 10^{-3}$	0.55821	0.57654	0.57638
Bivariate Normal	$2.8937 \cdot 10^{-3}$	0.57687	0.59411	0.38404
Meta t	$0.5678 \cdot 10^{-3}$	0.41589	0.4281642	0.50773

between the models became much more accentuated. The only model that accurately predicted the $\text{ES}_{1\%}$ based on the model building sample is our model. The difference between the predicted $\text{ES}_{1\%}$ and the estimates of the $\text{ES}_{1\%}$ based on the validation are very small.

The normal model now severely underestimates the risk. This due to the fact that the normal model underestimates the heaviness of the tails. It assumes, that the portfolio distribution has a normal distribution. In reality the left tail is regular varying. Based on our GPD fits, we found that all portfolios have a regular varying left tail with tail indexes between 2.5 and 3. As a consequence the estimates of $\text{ES}_{1\%}$ based on the validation sample is about 55% larger than predicted by the normal model.

The meta t model severely overestimates the risk of the corresponding optimal portfolio. As we saw for the result with respect to the $\text{ES}_{5\%}$, the portfolio based on the meta t model has an average log return that is not even close to the expected log return that was predicted my the meta t model. We already mentioned that the estimation of the expected log return is very unreliable in the case of the meta t model. This is the reason for the poor performance of the portfolio based on the meta t model.

In conclusion, we see that our model based on the spectral measure performs much better than the other two models. While the optimal portfolios based on the normal model are fairly similar to the ones based on our model, they seem to have a slightly higher risk. The main deficit of the normal model is that it severely underestimates the risk of the portfolio, because the tails of the model are much lighter than the actual tails of the data. The meta t model is not a valid choice in the present form, since the estimated expected log returns of the marginal components are unreliable. This leads to portfolios that do not achieve the expected log return they are designed to have. In addition, despite having a much smaller average log return, they have a risk that is comparable in size to the risk of the portfolios based on our model and the normal model.

BIBLIOGRAPHY

M. ABRAMOWITZ and I. STEGUN (1972): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Supt. of Docs., U.S. G.P.O., Washington, DC.

C. ACERBI (2002): Spectral Measures of Risk: A Coherent Representation of Subjective Risk Aversion. *Journal of Banking and Finance* 26:1505–1518.

C. ACERBI, C. NORDIO and C. SIRTORI (2001): Expected Shortfall as a Tool for Financial Risk Management. Working Paper, http://www.gloriamundi.org/.

C. ACERBI and P. SIMONETTI (2002): Portfolio Optimization with Spectral Measures of Risk. *Journal of Banking and Finance* 26:1505–1518.

C. ACERBI and D. TASCHE (2002): On the coherence of Expected Shortfall. *Journal of Banking and Finance* 26:1487–1503.

H. AKAIKE (1973): Information Theory and an Extension of the Likelihood Ratio Principle. *Proceedings of the Second International Symposium of Information Theory; ed. by B. N. Petrov and F. Csaki* :257–281.

H. AKAIKE (1974): A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19:716–723.

P. ARTZNER, F. DELBAEN, J. EBER and D. HEATH (1999): Coherent Measures of Risk. *Journal of Mathematical Finance* 3:203–228.

D. BEST and N. FISHER (1981): The Bias of the Maximum Likelihood Estimators of the von Mises-Fisher Concentration Parameters. *Communications in Statistics - Simulation and Computation* B10(5):493502.

C. BIERNACKI, G. CELEUX and G. GOVAERT (1996): Assessing the Mixture Model for Clustering with the Integrated Classification Likelihood. Technical Report NO. 3521. Rhôn-Alpes: INRIA.

C. BIERNACKI and G. GOVAERT (2000): Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:719–725.

P. BILLINGSLEY (1995): *Probability and Measure*, *Third Edition*. John Wiley & Sons, New York.

S. BLYTH (1996): Out of line. *RISK* 9:82–84.

W. BREYMANN, A. DIAS and P. EMBRECHTS (2003): Dependence Structures for Multivariate High-Frequency Data in Finance. *Quantitative Finance* 3:1–14.

I. CADEZ and P. SMYTH (2000): On Model Selection and Concavity for Finite Mixture Models (Extended Abstract). In proceedings of the International Symposium on Information Theory (ISIT 2000).

J. CHEN and J. KALBFLEISCH (1996): Penalized Minimum Distance Estimates in Finite Mixture Models. *Canadian Journal of Statistics* 24:167–175.

L. DE HAAN and S. RESNICK (1993): Estimating the Limit Distribution of Multivariate Extremes. *Communications in Statistics* 9:275–309.

S. DEMARTA and A. MCNEIL (2004): The t Copula and Related Copulas. Preprint ETH Zurich, http://www.math.ethz.ch/ mcneil.

A. DEMPSTER, N. LAIRD and D. RUBIN (1977): Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38.

J. EINMAHL, L. DE HAAN and V. PITERBARG (2001): Nonparametric Estimation of the spectral measure of an extreme value distribution. *Annals of Statistics* 29:1401–1423.

P. EMBRECHTS (2000): Extreme Value Theory: Potential and Limitations as an Integrated Risk Management Tool. *Derivatives Use, Trading & Regulation* 6:449–456.

P. EMBRECHTS, C. KLÜPPELBERG and T. MIKOSCH (1997): *Modelling Extremal Events for Insurance and Finance*. Springer Verlag, Berlin.

P. EMBRECHTS, F. LINDSKOG and A. MCNEIL (2003): Modelling Dependence with Copulas and Applications to Risk Management. *In: Handbook of Heavy Tailed Distributions in Finance, ed. S. Rachev* :Chapter 8: 329–384.

P. EMBRECHTS, P. MCNEIL and D. STRAUMANN (1999): Correlation: Pitfalls and alternatives. *RISK Magazine* May:69–71.

P. EMBRECHTS, P. MCNEIL and D. STRAUMANN (2002): Correlation and Dependence in Risk Management: Pitfalls and alternatives. *In: Risk Management: Value at Risk and Beyond, ed. M.A.H. Dempster Cambridge University Press, Cambridge* :176–223.

N. FISHER (1982): Robust Estimation of the Concentration Parameter of Fisher's Distribution on the Sphere. *Applied Statistics* 31:152–154.

M. FRASER, Y. HSU and J. WALKER (1981): Identifiability of Finite Mixtures of von Mises Distributions. *Annals of Statistics* 9:1130–1131.

P. GLASSERMAN, P. HEIDELBERGER and P. SHAHABUDDIN (2002): Value-at-Risk with Heavy-Tailed Risk Factors. *Mathematical Finance* 12:239–269.

P. GREEN (1995): Reversible Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* 82:711–732.

P. GREEN and S. RICHARDSON (1997): On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of Royal Statistical Society B* 59:731–792.

H. HAUKSSON, M. DACOROGNA, T. DOMENIG, U. MULLER and G. SAMORODNITSKY (2001): Multivariate extremes, aggregation and risk estimation. *Quantitative Finance* 1:79–95.

M. ISHIGURO, Y. SAKAMOTO and G. KITAGAWA (1997): Bootstrapping Log-Likelihood and EIC, an Extension of AIC. *Annals of the Institute of Spastical Mathematics* 49:411–434.

H. JOE (1997): *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

H. JOE, R. SMITH and I. WEISSMAN (1992): Bivariate Threshold Methods for Extremes. *Journal of the Royal Statistical Society, Series B* 54:171–183.

P. JUPP and K. MARDIA (2000): *Directional Statistics*. Wiley Series in Probability and Statistics, New York.

J. KENT (1978): Limiting Behavior of the von Mises-Fisher Distribution. *Math. Proc. Cambridge Phil. Soc.* 84:531–536.

J. KENT (1983): Identifiability of Finite Mixtures for Directional Data. *Annals of Statistics* 11:984–988.

C. KLUEPPELBERG and A. MAY (1998): The dependence function for bivariate extreme value distributions - a systematic approach. Technical Report 6, available at http://www-lit.ma.tum.de/veroeff/quel/989.60009.pdf.

S. KULLBACK and R. A. LEIBLER (1951): On Information and Sufficiency. *Annals of Mathematical Statistics* 22:79–86.

M. LEADBETTER, G. LINDGERN and H. ROOTZÉN (1983): *Extremes and related properties of random sequences and processes*. Springer, Berlin.

B. LEROUX (1992): Consistent Estimation of a Mixing Distribution. *Annals of Statistics* 20:1350–1360.

B. LINDSAY and P. BASAK (1993): Multivariate Normal Mixture: A fast consistent Method of Moments. *Journal of the American Statistical Association* 88:468–476.

Y. LO, N. MENDELL and D. RUBIN (2001): Testing the number of components in a normal mixture. *Biometrica* 88:767–778.

F. LONGIN and B. SOLNIK (1998): Correlation Structure of International Equity Markets During Extremely Volatile Periods. Tech Report 646 of HEC Paris, available at http://ideas.repec.org/p/ebg/heccah/0646.html.

K. MARDIA (1972): *Statistics of Directional Distributional Data*. Academic Press, London and New York.

K. MARDIA (1975): Statistics of Directional Data (with discussion). *Journal of the Royal Statistical Society Series B* 37:349–393.

G. McLachlan and D. Peel (2000): *Finite Mixture Models*. Wiley, New York.

R. Redner (1981): Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics* 9:225–228.

R. Redner and H. Walker (1984): Mixture Densities, Maximum Likelihood And The EM Algorithm. *SIAM Review* 26:195–238.

S. Resnick (1986): *Point processes, regular variation, and weak convergence.*. Springer, New York.

S. Resnick (2002): On the Foundations of Multivariate Heavy Tail Analysis. Technical Report of the School OR&IE at Cornell University.

S. Resnick and C. Stărică (1998): Tail index estimation for dependent data. *Annals of Applied Probability* 8:1156–1183.

R. Rockafellar and S. Uryasev (2000): Optimization of conditional value-at-risk. *Journal of Risk* 2:21–41.

G. Schwarz (1978): Estimating the dimension of a model. *Annals of Statistics* 6:461–464.

J. Shao (1998): *Mathematical Statistics*. Springer Texts in Statistics. Springer, New York.

J. Shaw (1997): Beyond VaR and Stress Testing. *In VAR: Understanding an Applying Value At Risk, Risk Publication, London.* :211–224.

R. Smith (1985): Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72:67–90.

C. STĂRICĂ (1999): Multivariate extremes for models with constant conditional correlations. *Journal of Empirical Finance* 6:515–553.

D. TASCHE (2002): Expected Shortfall and Beyond. *Journal of Banking and Finance* 26:1519–1533.

J. TAWN (1988): Bivariate Extreme Value Theory: Models and Estimation. *Biometrika* 75:397–415.

D. TITTERINGTON, A. SMITH and U. MAKOV (1985): *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

J. V. USPENSKY (1937): *Introduction to mathematical probability*. McGraw-Hill Book Company, Inc., New York.

Q. VUONG (1989): Likelihood Ratio Tests For Model Selection And Non-Nested Hypothesis. *Econometrica* 57:307–333.

A. WALD (1949): Note on the consistency of the maximum-likelihood estimate. *Ann. Math. Statist.* 20:595–600.

H. WHITE (1982): Maximum Likelihood estimation of Misspecified Models. *Econometrica* 50:1–25.