Advanced Methods of Multivariate Financial Data Analysis

Habilitationsschrift

zur

Erlangung der Venia Legendi

der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2008

vorgelegt

von

Dr. rer. pol. Gabriel Frahm

aus

Naharia/Israel

Referent:	Univ.-Prof. Dr. K. Mosler
Korreferent:	Univ.-Prof. Dr. F. Schmid
Tag der Habilitation:	25.05.2009

To Franziska

Streuungsmaße

Ein Mensch der von Statistik hört,
denkt dabei nur an den Mittelwert.
Er glaubt nicht dran und ist dagegen,
ein Beispiel soll es gleich belegen:
Ein Jäger auf der Entenjagd
hat einen ersten Schuß gewagt.
Der Schuß zu hastig aus dem Rohr,
lag eine gute Handbreit vor.
Der zweite Schuß mit lautem Krach
lag eine gute Handbreit nach.
Der Jäger spricht ganz unbeschwert
voll Glauben an den Mittelwert:
„Statistisch ist die Ente tot."
Doch wär er klug und nähme Schrot –
dies sei gesagt, Ihn zu bekehren –
würde seine Chancen mehren:
Der Schuß geht ab, die Ente stürzt,
weil Streuung ihr das Leben kürzt.

Eugen Roth (1895–1976)

Preface

After my Ph.D. thesis in 2004 it was clear to me that I would like to continue my scientific work. Originally I came from capital market theory but I early decided to focus on statistics with a strong emphasis on methods of multivariate analysis. Due to the almost unlimited availability of financial market data it is quite reasonable to apply such tools as copulas, extreme value theory, generalized elliptical distributions, robust covariance matrix estimation, shrinkage estimation, Bayesian analysis, etc., to that kind of observations. Therefore the present work resumes my scientific contributions on that field over the last four years. The alert reader will observe that the topic of portfolio optimization plays a prominent role in that hodge-podge. Indeed, this is no accident since the implementation of modern portfolio theory is still under vivid discussion. Moreover, it is a perfect playground for statisticians with an aptitude for methods of multivariate analysis and indeed leads to surprising results both from the viewpoint of mathematics and economics.

First of all I would like to thank Dr. Guy Lonsdale who gave me the kind opportunity to work for the NEC Laboratories Europe (NEC Europe Ltd.) as an employee and later on as a consultant. I think that the mathematical problems which had to be solved during that period had a substantial influence on my academic work. Many discussions with practitioners opened my eyes and maybe helped me to overcome that kind of autistic thinking which is often prevalent within the scientific community per se.

At the end of 2004 I convinced Professor Karl Mosler to take me as a postdoctoral fellow on his chair of econometrics at the University of Cologne. Inexplicably that happened although I am apparently not a mathematician (at least in a formal manner). I am deeply grateful for his support. He gave me not only the joy of mathematics (which turns sometimes into suffering, as with all that beautiful things in life) but also an inspiring example of how a perfect academic advisor *should* be. Okay, everybody knows that he is brilliant but that's

not the point. Indeed, he is a wise man and maybe he understood that human ability is better unfolded by emancipating rather than restricting.

I am also thankful to my colleagues, co-authors, and all the other contributors to my work, i.e. Alexander Bade, Anna Brandt, Dr. Jadran Dobrić, Dr. Rainer Dyckerhoff, Professor Christian Genest, Professor Uwe Jaekel, Dr. Markus Junker, Professor Alexander Kempf, Carsten Körner, Christina Loley, Dr. Christoph Memmel, Dr. Dr. Gert Mittring, Dr. Julia Nasev, Walter Orth, Yulia Polyakova, Professor Donald Rubin, Christoph Scheicher, Professor Friedrich Schmid, Dr. Rafael Schmidt, and Professor David Tyler. Without their valuable contributions this work would have never been accomplished. Also I would like to thank the participants and organizers of the summer school on *Risk Theory and Related Topics* 2008 in Bedlewo. It was really a pleasure to meet so many nice people on that beautiful place. I will not forget to mention Tobias Wickern, a doctoral student of the Cologne Graduate School of Risk Management. Besides his mental capabilities he impressed me very much by his great attitude.

This work is devoted to my wonderful wife Franziska. After all that years she eventually accepted that it is not easy for me to be always at home at five o'clock in the evening. So it was her part to take care for our lovely children. Meanwhile they almost grew up and I am very happy that they still remember my name. Hence, you did a great job, Franziska! Now, after so many years lost in Platonia, it's good for me to finish my habilitation. Come on, let's see what the future will be...

Gabriel Frahm

Cologne, 18$^{\text{th}}$ December, 2008

Contents

Introduction

Since Bachelier's seminal work of 1900, probabilistic methods have been widely applied to financial data. At the beginning the primary goal was to find a mathematical description of the dynamics of asset prices on financial markets. Even though this is still a matter of particular interest – especially due to the recent turmoils after the subprime mortgage crisis – it is only one side of the coin. Statistical methods indeed can help to quantify risks. However, the other side of the coin is that statistical methods themselves are susceptible to risk as well. Besides the fact that they are often misused or misunderstood, there are many other factors which can lead to wrong conclusions. For example, a statistical model which is chosen for describing the data generating process might be wrong and even if the model is correct, a remaining problem is parameter uncertainty.

There is more to it than that. In many practical applications of statistical methods one can observe the problem of *data dredging* (also known as 'data pruning', 'data fishing', 'data snooping', 'data mining', etc.), i.e. searching for 'statistically significant' relationships in large quantities of data and/or dimensions. Another problem is to distinguish between the purpose of *statistical inference* and the goal of making an *optimal decision* based on empirical data. One might expect that an estimation procedure which is 'efficient' in statistical terms is also the best choice for the decision maker. Unfortunately, in many cases it can be shown that this assertion is wrong.

I would like to give a more or less informal example. Suppose that φ is some utility function depending on two quantities, i.e. a decision x and an unknown parameter θ. Now let $\hat{\theta}$ be some estimator for θ such that $\varphi\big(x\,;\hat{\theta}\big)$ is an efficient or at least *unbiased* estimator for the utility $\varphi\big(x\,;\theta\big)$. Typically, the decision maker tries to maximize his utility by choosing $\hat{x}^* = \arg\max_\xi \varphi\big(\xi\,;\hat{\theta}\big)$. Under quite general conditions concerning φ and $\hat{\theta}$ it holds that $\varphi\big(\hat{x}^*;\hat{\theta}\big) > \varphi\big(x\,;\hat{\theta}\big)$ (a.s.) for every *fixed* decision x. That means

$$\mathrm{E}\big\{\varphi\big(\hat{x}^*;\hat{\theta}\big)\big\} > \mathrm{E}\big\{\varphi\big(x^*;\hat{\theta}\big)\big\} = \varphi\big(x^*;\theta\big) > \varphi\big(\hat{x}^*;\theta\big)\,,$$

where $x^* = \arg\max_\xi \varphi\big(\xi\,;\theta\big)$ denotes the *optimal* decision. Hence, although $\varphi\big(\cdot\,;\hat{\theta}\big)$ is an unbiased estimator for $\varphi\big(\cdot\,;\theta\big)$, surprisingly the same function cannot produce an unbiased estimator for the utility of the decision maker's *actual* choice $\hat{x}^*$. This is dangerous because it can mislead the decision maker into taking some highly suboptimal alternative. That means the fundamental concepts of unbiasedness and efficiency, which are widely accepted in statistical inference, are not appropriate in decision theory.

The first chapter ('Estimating the tail-dependence coefficient: properties and pitfalls') is a joint work with M. Junker and R. Schmidt (Frahm et al., 2005). The so-called *tail-dependence coefficient* can be used as a measure for describing the dependence between extremal data in finance. This is an important topic since extremal dependencies between financial asset returns have dramatically increased in recent years. Therefore the tail-dependence coefficient has become a popular measure in risk management. We investigate different methods for estimating the tail-dependence coefficient which are frequently used in the literature. Our work is based on copula theory and multivariate extreme value theory. Actuaries and statisticians who are not familiar with extreme value theory often have difficulties in choosing appropriate methods for measuring or estimating the tail-dependence. One reason for that is the limited amount of (extremal) data which makes the estimation quite sensitive to the choice of the method. Another reason is the lack of literature comparing the various estimators developed in (mostly theoretical) articles related to extreme value theory. Hence, we try to partially fill this gap by surveying and comparing various methods of tail-dependence estimation.

In the second chapter ('Dependence of stock returns in bull and bear markets'), which is a joint work with J. Dobrić and F. Schmid (Dobrić et al., 2008), we present an alternative method for measuring the dependence of extreme values. Pearson's rho, i.e. the standard estimator for the linear correlation coefficient, is the most commonly used measure of dependence. However, its many shortcomings have been often documented in the literature. Pearson's rho is strongly affected by the marginal distributions of the random variables which are taken into consideration and it is also very sensitive to outliers. Further, it only quantifies the amount of *linear* dependence. In spite of that fact, Pearson's rho is used as a dependence measure in most empirical investigations of *extreme* asset returns. This can lead to wrong conclusions. As an appropriate alternative we introduce a conditional version of Spearman's rho. This is an estimator for the rank-correlation coefficient of extreme values. Our approach is based on fundamental results of copula theory. We apply

our estimator to asset returns which have been observed on the German stock market. In particular, we concentrate on the question whether dependence is significantly different in bull and bear markets.

The following chapter ('A general approach to Bayesian portfolio optimization') is a joint work with A. Bade and U. Jaekel (Bade et al., 2008). Traditional portfolio optimization strategies are susceptible to parameter uncertainty and many portfolio optimization approaches rely on rather simple assumptions about the distribution of asset returns. However, it is well-known that short-term financial data can be heavy-tailed or at least leptokurtic, tail-dependent, skewed or asymmetric in some other way. Moreover, financial time series typically exhibit volatility clusters or even long-memory; high-frequency data generally are non-stationary, have jumps, etc. This might be the reason why many authors prefer to work with long-term asset returns. However, decreasing the sampling frequency leads to a loss of statistical efficiency. Our principal goal is to present a general approach to portfolio optimization which takes account of both estimation risks and stylized facts of financial data. This is done within a Bayesian framework using contemporary methods of Markov chain Monte Carlo. By contrast, the existing literature on Bayesian portfolio optimization in general does not take stylized facts into account since many Bayesian approaches are based on a purely analytical fundament. To avoid limitations of such kind, we suggest a Metropolis-Hastings-like algorithm for simulating the posterior distribution of the unknown parameters. This is derived on the basis of empirical information obtained from time series data and prior information possibly given by an expert. By choosing a numerical rather than an analytical approach, principally we can use almost any probabilistic model for the data and parameters. At the end of Chapter 3 a realistic portfolio optimization problem is presented, which has been performed on a standard PC in reasonable time.

In Chapter 4 ('Linear statistical inference for global and local minimum variance portfolios') I provide analytic results concerning the small-sample properties of minimum variance portfolios (Frahm, 2008). At the beginning of modern portfolio theory it was usually supposed that the parameters of interest, i.e. the means and (co-)variances of asset returns, can be estimated accurately such that estimation errors remain negligible. Although this conjecture might be true for variances and covariances if the sample size is large enough compared to the number of assets, it is not an appropriate simplification for expected asset returns in most practical situations. Therefore the so-called *global minimum variance*

portfolio has been recently advocated by many authors. Its main advantage is that no expected asset returns have to be estimated and so the impact of estimation errors can be substantially reduced. However, in many practical situations investors do not aim at finding the *global* minimum variance portfolio but a minimum variance portfolio under some additional constraints besides the budget constraint. For example, portfolio managers of mutual funds often have to observe certain limits regarding their choice of portfolio weights. Such a portfolio will be referred to as a *local minimum variance portfolio*. Focusing on the small-sample rather than large-sample properties is an important issue, for I will show that large-sample approximations fail if the sample size is large but the number of observations relative to the number of assets is small. The statistical instruments used in that chapter are taken from linear regression theory under stochastic regressors. After recalling some existing hypothesis tests for the global minimum variance portfolio (Kempf and Memmel, 2006), I derive the corresponding tests for local minimum variance portfolios. Furthermore, the joint distribution of the weights of global and local minimum variance portfolios is calculated and I also present an empirical study where the given instruments are applied to stock market data.

Chapter 5 ('Dominant estimators for the global minimum variance portfolio') is a joint work with C. Memmel (Frahm and Memmel, 2008). Kempf and Memmel (2006) showed that the traditional estimator for the global minimum variance portfolio is the best unbiased estimator if the asset returns possess a multivariate normal distribution. However, as already pointed out before, that does not necessarily imply that the traditional estimator is the best choice from the investor's point of view. Indeed, we have found two estimators for the global minimum variance portfolio which *dominate* the traditional estimator with respect to the out-of-sample variance of the portfolio return. Due to the arguments given in Chapter 4, the same conclusion can be drawn for estimating local minimum variance portfolios. The methods used in Chapter 5 heavily rely on Stein-type estimation theory. In contrast to the existing shrinkage approaches which can be found in the portfolio optimization literature, our results are valid in *small* samples. We show that by using our shrinkage estimators it is possible to reduce the out-of-sample variance of the portfolio return substantially. We present not only the small-sample properties of the shrinkage estimators and some related quantities, but also their large-sample properties. Moreover, backed by the results of a recent study presented by DeMiguel et al. (2007), we derive a small-sample test for the 'naive diversification hypothesis', i.e. for deciding the question of

whether or not it is better to completely ignore time series information in favor of naive diversification.

In the following chapter ('A hypothesis test for the best investment strategy') I discuss the question whether a chosen investment strategy is *significantly* the best compared to some other candidates (Frahm, 2007). Here the notion of significance is emphasized, since for applying a statistical test it is typically assumed that the hypothesis on hand is chosen before examining the data. Otherwise the hypothesis test would suffer from a selection bias. Speaking more generally, given some empirical or simulated data it is often questionable which 'alternative' or 'decision' is the best one in terms of some objective or utility function. Hence, a favorite alternative has to be compared with some given competitors. That means various hypothesis tests have to be conducted simultaneously, which is a typical problem of multiple testing. In Chapter 6, I derive a large-sample test for the best alternative in a rather general setting. The presented test accounts for conditional heteroscedasticity and non-normality of asset returns – in contrast to the well-known Jobson-Korkie-Memmel test – and it will be demonstrated by an application to financial data.

Chapter 7 ('Asymptotic distributions of robust shape matrices and scales') leaves the scope of portfolio optimization and belongs to the general theory of robust covariance matrix estimation. In this work (**?**) I discuss the problem of shape matrix estimation. It has been frequently observed that many multivariate statistical methods like principal components analysis, canonical correlation analysis, linear discriminant analysis, and multivariate regression require the covariance matrix only up to some scaling constant. If the topic of interest is not the scale but only the *shape* of the distribution of some random vector X, it is not meaningful to focus on the asymptotic covariance matrix of an estimator for the covariance matrix of X, i.e. Σ or some other matrix being proportional to Σ. This problem is much discussed in the recent literature on robust covariance matrix estimation. I derive explicit expressions for the joint asymptotic distributions of robust shape matrix estimators and the associated estimators for the scale. This is done by using a fundamental result given by Tyler (1982) and advanced methods of multivariate analysis. This chapter also contains a generalization of a surprising result recently obtained by Paindaveine (2008) in the context of local asymptotic normality theory. More precisely, it is shown that the estimators for the shape matrix and scale are asymptotically independent for one and only one specific choice of the scale function, provided their asymptotic distribution is normal.

The last chapter ('Distribution-free shape matrix estimation for incomplete data') is a joint work with U. Jaekel (Frahm and Jaekel, 2007a). One disadvantage of most robust covariance matrix estimators is that they do not account for missing data. However, this is extremely important, since in my opinion almost any data set which can be found in practice is incomplete. Therefore we derive a shape matrix estimator which works both with complete and incomplete data. This is done by applying contemporary methods of missing data analysis. Our estimator is *distribution-free* within the class of generalized elliptical distributions (Frahm, 2004). That means it is invariant under any change of the generating variate and thus it is not bothered by heavy tails and other kinds of financial data anomalies. We show that in the complete-data case the presented estimator corresponds to Tyler's celebrated M-estimator (Tyler, 1983, 1987a). By contrast, our extension of Tyler's M-estimator turns out to be an *ML-estimator* if the data are incomplete. Thus it is possible to obtain its asymptotic properties by standard results of likelihood theory, using some arguments given in Chapter 7. We also present a fast algorithm for calculating the estimate which works well even for high-dimensional data. Further, we provide a simulation study covering the complete-data as well as the incomplete-data case using clean and contaminated data under the different missingness mechanisms MCAR, MAR, and NMAR.

Chapter 1.

Estimating the Tail-Dependence Coefficient: Properties and Pitfalls

1.1. Motivation

During the last decade, dependencies between financial asset returns have increased due to globalization effects and relaxed market regulation. However, common dependence measures such as Pearson's correlation coefficient are not always suited for a proper understanding of dependencies in financial markets; see, e.g., Embrechts et al. (2002). In particular, dependencies between extreme events such as extreme negative stock returns or large portfolio losses cause the need for alternative dependence measures to support beneficial asset-allocation strategies.

Several empirical surveys such as Ané and Kharoubi (2003) and Junker and May (2005) exhibited that the concept of tail-dependence is a useful tool to describe the dependence between extremal data in finance. Moreover, they showed that especially during volatile and bear markets, tail-dependence plays a significant role. In this context, tail-dependence is described via the so-called *tail-dependence coefficient* (TDC) introduced by Sibuya (1960). This concept is reviewed in Section 1.2.

However, actuaries and statisticians who are not familiar with *extreme value theory* (EVT) often have difficulties in choosing appropriate methods for measuring or estimating tail-dependence. One reason for that is the limited amount of (extremal) data which makes the estimation quite sensitive to the choice of method. Another reason is the lack of literature which compares the various estimators developed in (mostly theoretical) articles related to EVT. This paper tries to partially fill this gap by surveying and comparing various

methods of tail-dependence estimation. In other words, we will present the most common estimators for the TDC and compare them via a simulation study.

TDC estimators are either based on the entire set of observations or on extremal data. Regarding the latter, EVT is the natural choice for inferences on extreme values. In the one-dimensional setting, the extreme value distributions can be expressed in parametric form, as shown by Fisher and Tippett (1928). Thus it suffices to apply parametric estimation methods only. By contrast, multidimensional extreme value distributions cannot be characterized by a fully parametric model in general. This leads to more complicated estimation techniques.

Parametric estimation methods are efficient if the distribution model under consideration is true, but they suffer from biased estimates in case the underlying model is different. Nonparametric estimation procedures avoid this type of model error but come along with a larger estimation variance. Accordingly, we distinguish in Section 1.3 between the following types of TDC estimations, namely, TDC estimations which are based on:

a) a specific distribution or a family of distributions;

b) a specific copula or a family of copulas; or

c) a nonparametric model.

We discuss properties of the estimators along with possible applications and give references for further reading. Section 1.4 presents a detailed simulation study which analyzes and compares selected estimators regarding their finite sample behavior. Statistical methods testing for tail-dependence or tail-independence are not included in this work. An account on that topic can be found for instance in Draisma et al. (2004).

1.2. Preliminaries

The following approach, discussed by Sibuya (1960) and Joe (1997, p. 33) among others, represents the most common definition of tail-dependence. Let (X, Y) be a random pair with joint cumulative distribution function F and marginals G (for X) and H (for Y). The quantity

$$\lambda_U = \lim_{t \to 1^-} \mathbb{P}\{G(X) > t \mid H(Y) > t\}$$

is called the *upper tail-dependence coefficient* (upper TDC), provided the limit exists. We say that (X, Y) is *upper tail dependent* if $\lambda_U > 0$ and *upper tail independent* if $\lambda_U = 0$. Similarly, we define the lower tail-dependence coefficient by

$$\lambda_L = \lim_{t \to 0^+} \mathbb{P}\{G(X) \leq t \mid H(Y) \leq t\}.$$

Thus, the TDC roughly corresponds to the probability that one margin exceeds a high/low threshold under the condition that the other margin exceeds a high/low threshold.

The TDC can also be defined via the notion of copula, introduced by Sklar (1959). A copula C is a cumulative distribution function whose margins are uniformly distributed on $[0, 1]$. As shown by Sklar (1959), the joint distribution function F of any random pair (X, Y) with marginals G and H can be represented as

$$F(x, y) = C\{G(x), H(y)\}. \tag{1.1}$$

in terms of a copula C which is unique when G and H are continuous, as will be assumed in the sequel. Refer to Nelsen (2006) or Joe (1997) for more information on copulas.

If C is the copula of (X, Y), then

$$\lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{t} \quad \text{and} \quad \lambda_U = \lim_{t \to 1^-} \frac{1 - 2t + C(t,t)}{1-t}.$$

Another representation of the upper TDC is given by $\lambda_U = \lim_{s \to 0^+} \tilde{C}(s,s)/s$, where $\tilde{C}(1-t, 1-t) = 1 - 2t + C(t,t)$ denotes the *survival copula* of C. Thus, the upper TDC of C equals the lower TDC of its survival copula and, vice versa, the lower TDC of C is given by the upper TDC of $\tilde{C}$. Since the TDC is determined by the copula of X and Y, many copula features transfer directly to the TDC. For instance, the TDC is invariant under strictly increasing transformations of the margins.

Consider a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of observations of (X, Y). Let

$$X_n^* = \max(X_1, \ldots, X_n) \quad \text{and} \quad Y_n^* = \max(Y_1, \ldots, Y_n)$$

be the corresponding componentwise maxima. In order to have a meaningful discussion about tail-dependence in the EVT framework, we assume that F belongs to the domain of attraction of an extreme value (EV) distribution. This means that as $n \to \infty$, the joint distribution of the standardized componentwise maxima X_n^* and Y_n^* has the following limiting EV distribution (with non-degenerated margins):

$$F^n(a_n x + b_n, c_n y + d_n) \to F_{EV}(x, y)$$

for some standardizing sequences $(a_n), (c_n) > 0$ and $(b_n), (d_n) \in \mathbb{R}$. Suppose that F_{EV} has unit Fréchet margins G_{EV} and H_{EV}, i.e.,

$$G_{EV}(x) = \exp(-1/x),\ x > 0 \quad \text{and} \quad H_{EV}(y) = \exp(-1/y),\ y > 0.$$

This assumption, which is standard in the EVT framework, is similar to the assumption that the margins can be transformed into uniform distributions in the theory of copulas. Then the EV distribution possesses the following representation (Pickands, 1981):

$$F_{EV}(x,y) = \exp\left\{-\left(\frac{1}{x}+\frac{1}{y}\right) A\left(\frac{y}{x+y}\right)\right\}, \qquad x,y > 0. \tag{1.2}$$

Here $A : [0,1] \to [1/2,1]$ is a convex function such that $\max(t, 1-t) \leq A(t) \leq 1$ for every $0 \leq t \leq 1$. The function A is known as Pickands' dependence function. In the sequel, the term *dependence function* always refers to the above representation and should not be confused with the copula of a bivariate random vector.

The copula C_ℓ^*, $\ell \in \mathbb{N}$, of the componentwise maxima X_l^* and Y_l^* is related to the copula C as follows:

$$C_\ell^*(u,v) = C^\ell\left(u^{1/\ell}, v^{1/\ell}\right), \qquad 0 \leq u,v \leq 1.$$

If the diagonal section $C(t,t)$ is differentiable for $t \in (1-\varepsilon, 1)$ for some $\varepsilon > 0$, then it can be shown that

$$2 - \lambda_U = \lim_{t\to 1^+} \frac{1-C(t,t)}{1-t} = \lim_{t\to 1^+} \frac{1-C_\ell^*(t,t)}{1-t} = \lim_{t\to 1^+} \frac{dC(t,t)}{dt} = \lim_{t\to 1^+} \frac{dC_\ell^*(t,t)}{dt} \tag{1.3}$$

for all $\ell \in \mathbb{N}$. In particular, for $\ell \to \infty$ we obtain

$$C_{EV}(t,t) = F_{EV}\left\{-\frac{1}{\log(t)}, -\frac{1}{\log(t)}\right\} = t^{2A(1/2)}, \qquad 0 < t < 1,$$

where C_{EV} denotes the copula of F_{EV}. This implies the following important relationship:

$$\lambda_U = 2 - 2A\left(\frac{1}{2}\right).$$

Another representation of the EV distribution is frequently encountered in the EVT literature. If F_{EV} has unit Fréchet margins, there exists a finite spectral measure S on $\mathcal{B} = \{(x,y) :\ x,y > 0, \|(x,y)\|_2 = 1\}$, where $\|\cdot\|_2$ denotes the Euclidean norm, such that

$$F_{EV}(x,y) = \exp\left\{-\int_{\mathcal{B}} \max\left(\frac{u}{x}, \frac{v}{y}\right)\ dS(u,v)\right\}, \qquad x,y > 0,$$

with $\int_{\mathcal{B}} u\ dS(u,v) = 1$ and $\int_{\mathcal{B}} v\ dS(u,v) = 1$. This yields

$$\lambda_U = 2 - \int_{\mathcal{B}} \max(u,v)\ dS(u,v)$$

and $A(1/2) = \int_{\mathcal{B}} \max(u, v)\, dS(u, v)/2$. The estimation of the spectral measure is discussed by Joe et al. (1992), de Haan and Resnick (1993), Einmahl et al. (1993, 1997), and Capéraà and Fougères (2000), among others.

Thus *any* estimator of the upper TDC $\hat{\lambda}_U$ (the index n is dropped for notational convenience) is equivalent to some estimator $\widehat{A}_n(1/2)$ via the relationship $\hat{\lambda}_U = 2 - 2\widehat{A}_n(1/2)$. By considering the dependence function related to the survival copula, this holds also for the lower TDC. An abundant literature exists concerning the estimation of the dependence function A. See for instance de Oliveira (1984), Tawn (1988), Smith et al. (1990), Hutchinson and Lai (1990) or Coles and Tawn (1991) for fitting parametric (structural) models to A. By contrast, Pickands (1981), Deheuvels (1991), Joe et al. (1992), Abdous et al. (1999), Capéraà and Fougères (2000) or Falk and Reiss (2003) consider nonparametric estimation procedures.

Due to the invariance of the TDC of (X_ℓ^*, Y_ℓ^*) with respect to ℓ, the following estimator arises quite naturally:

$$\hat{\lambda}_U = 2 - 2\widehat{A}_m\left(\frac{1}{2}\right) = 2 - \left.\frac{\widehat{dC_m}}{dt}(t,t)\right|_{t\approx 1}, \quad 1 \le m \le n.$$

Here $\widehat{dC_m/dt}$ denotes the estimated derivative of the diagonal section of the copula C_ℓ^* from m block maxima, where each block contains $\ell = n/m$ elements of the original data set (we choose m such that $n/m \in \mathbb{N}$). The special case $m = n$ (i.e., $\ell = 1$) corresponds to n block maxima which form the original data set. Every TDC estimator has to deal with a bias-variance trade-off arising from the following two sources. The first one is the choice of the threshold t. That is, the larger t the smaller the bias (and the larger the variance) and vice versa. The second source is the number of block maxima. Thus, the larger m the smaller the variance but the larger the bias. An optimal choice of m and t, e.g., with respect to the mean squared error (MSE) of the estimator, is usually difficult to derive. A similar problem exists for univariate tail-index estimations of regular varying distributions.

In Figure 1.4, we illustrate the latter bias-variance problem via the following estimator which is motivated by (1.3) and forms the nonparametric counterpart of the parametric estimator $\hat{\chi}$ introduced in Coles et al. (1999):

$$\hat{\lambda}_U^{\mathrm{LOG}} = 2 - \frac{\log \widehat{C}_m\left(\frac{m-k}{m}, \frac{m-k}{m}\right)}{\log\left(\frac{m-k}{m}\right)}, \qquad 0 < k < m, \tag{1.4}$$

where

$$\widehat{C}_m(u, v) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left(R_{1j}/m \le u, R_{2j}/m \le v\right)$$

is called the *empirical copula*. Here $\mathbb{1}$ denotes the indicator function, while R_{1j} and R_{2j}, respectively, are the ranks of the block maxima $X^*_{\ell j}$ and $Y^*_{\ell j}$, $j = 1, \ldots, m$, $\ell = n/m$. The threshold is denoted by k. As expected, Figure 1.4 reveals that the estimation via block maxima has a lower bias but a larger variance. The bias-variance tradeoff for various thresholds can be clearly seen, too.

In order to ease the presentation we do not explicitly differentiate between block maxima and the original data set in the forthcoming sections.

1.3. TDC Estimation

The following estimation approaches are classified by the degree of prior information which is available about the distribution of the data. We will either assume a specific distribution or a class of distributions, a specific copula or a class of copulas, or we perform a completely nonparametric estimation. For notational convenience, λ will be written without the subscript L or U whenever we know that $\lambda_L = \lambda_U$. Moreover, the subscript is dropped whenever we neither specifically refer to the upper nor to the lower TDC.

1.3.1. Estimation Using a Specific Distribution

Suppose that the distribution $F(\cdot\,;\theta)$ is known. Further assume that λ can be represented via a *known* function of θ, i.e., $\lambda = \lambda(\theta)$. Also assume that F allows for tail-dependence. Then the parameter θ can be estimated via maximum-likelihood (ML), which suggests the estimator $\hat{\lambda} = \lambda(\hat{\theta})$. Under the usual regularity conditions of ML-theory, as in Casella and Berger (2002, p. 516), the functional estimator $\hat{\lambda} = \lambda(\hat{\theta})$ represents an ML-estimator which possesses the well-known consistency and asymptotic normality properties.

Example 1. Suppose that (X, Y) is bivariate t-distributed, i.e.,

$$(X, Y) \overset{\mathrm{d}}{=} \mu + \frac{Z}{\sqrt{\chi^2_\alpha/\alpha}}\,, \qquad \alpha > 0,$$

where $Z \sim \mathcal{N}(0, \Sigma)$, $\mu \in \mathbb{R}^2$, $\Sigma \in \mathbb{R}^{2\times 2}$ positive definite, and Z is stochastically independent of χ^2_α. Then Embrechts et al. (2002) show that

$$\lambda = 2\,\bar{t}_{\alpha+1}\left(\sqrt{\alpha+1}\,\sqrt{\frac{1-\rho}{1+\rho}}\right), \tag{1.5}$$

where $\bar{t}_{\alpha+1}$ is the survival function of a Student's univariate t-distribution with $\alpha+1$ degrees of freedom. The parameter $\rho = \sin(\pi\tau/2)$, expressed in terms of Kendall's tau, denotes the

correlation parameter of (X, Y). It corresponds to Pearson's correlation coefficient, when it exists. □

Obviously this estimation approach requires prior information about the joint distribution function of the data. Consequently, the TDC estimator generates good estimates (in the sense of MSE) if the proposed distribution is the right one, but it will be biased if the distribution is wrong. In other words, this type of estimation is not expected to reveal surprising results and will be, therefore, excluded from the subsequent discussion.

1.3.2. Estimation within a Class of Distributions

Instead of a specific distribution, we now suppose that F belongs to a class of distributions. Because of its popularity in theory and practice, as illustrated, e.g., by Bingham et al. (2003) and Embrechts et al. (2003), we consider the class of elliptically contoured distribution, viz.

$$(X, Y) \stackrel{\mathrm{d}}{=} \mu + \mathcal{R}\Lambda U^{(2)}, \tag{1.6}$$

where $U^{(2)}$ is a random pair uniformly distributed on the unit circle, $\mathcal{R}$ is a nonnegative random variable that is stochastically independent of $U^{(2)}$, $\mu \in \mathbb{R}^2$ is a location parameter, and $\Lambda \in \mathbb{R}^{2\times 2}$ is nonsingular. Well-known members of the latter distribution family are the multivariate normal, multivariate t and symmetric generalized hyperbolic distributions. Note that $\rho = 0$ does not correspond to independence; see, e.g., Abdous et al. (2005) for additional discussion concerning the dependence properties of this class of copulas.

In case the tail distribution of the Euclidean norm $\|(X, Y)\|_2$ is regularly varying with tail index $\alpha > 0$ [see Bingham et al. (1987) for the definition of regular variation], Schmidt (2002) and Frahm et al. (2003) show that tail-dependence is present and that relationship (1.5) still holds. In particular, we have

$$A\left(\frac{1}{2}\right) = t_{\alpha+1}\left(\sqrt{\alpha+1}\,\sqrt{\frac{1-\rho}{1+\rho}}\right).$$

Various methods for the estimation of the tail index α are discussed, e.g., in Matthys and Beirlant (2002) or Embrechts et al. (1997).

1.3.3. Estimation Using a Specific Copula

Suppose that the copula $C(\cdot\,;\theta)$ is known. Note that this is a much weaker assumption than assuming a specific distribution. The estimation of the parameter θ can be performed

in two steps. First, we transform the observations of X and Y (or the corresponding block maxima) via estimates of the marginal distribution functions G and H and fit the copula from the transformed data in a second step; the transformation is justified by (1.1). Unless stated otherwise, the marginal distribution functions will be estimated by the empirical distribution functions $\widehat{G}_n$ and $\widehat{H}_n$.

The estimation of G and H via the empirical distribution functions avoids an incorrect specification of the margins. Genest et al. (1995), as well as Shih and Louis (1995), discuss consistency and asymptotic normality of the copula parameter θ if it is estimated in this fashion. Roughly speaking, if the map between θ and λ is smooth enough, then the estimator $\hat{\lambda} = \lambda(\hat{\theta})$ is consistent and asymptotically normal provided $\hat{\theta}$ is consistent and asymptotically normal.

If $G(\cdot\,;\theta_G)$ and $H(\cdot\,;\theta_H)$ are assumed to be specific distributions, then θ_G and θ_H can be estimated, e.g., via ML methods. In particular, the *IFM method* (method of inference functions for margins) consists of estimating θ_G and θ_H via ML and, in a second step, estimating the parameter θ of the copula $C(\cdot\,;\theta)$ via ML also. However, for this approach it is necessary that the parameters θ_G and θ_H do not analytically depend on the copula parameter θ. Results about the asymptotic distribution and the asymptotic covariance matrix of this type of estimation are derived in Joe (1997, Ch. 10); see also the references therein. A simulation study (which is not included in this paper but can be obtained from the authors upon request) shows that there is not much difference between the two step and the one step estimation in terms of the MSE. Also the MSE related to the pseudo-ML and the ML-estimation via empirical margins are roughly the same in this simulation study.

Example 2. Suppose that the data stem from a bivariate t-copula

$$C\left(u, v\,;\alpha, \rho\right) = t_\alpha\left\{t_\alpha^{-1}(u), t_\alpha^{-1}(v)\,;\rho\right\},$$

where $t_\alpha(\cdot\,;\rho)$ represents the bivariate t-distribution function with α degrees of freedom and correlation parameter ρ. □

Note that elliptical copulas (i.e., copulas of elliptical random vectors) are restricted to transpositional symmetry, i.e., $C = \tilde{C}$ and thus $\lambda_L = \lambda_U$. Hence, if the TDC is estimated from the entire sample via a single copula, the elliptical copulas are not appropriate if $\lambda_L \neq \lambda_U$. For example it is well known that investors react differently to negative and positive news. In particular for asset return modeling, the symmetry assumption has to be considered with care; see, e.g., Junker (2004) for an empirical study of commodity returns

and U.S. dollar yield-returns using likelihood ratio tests. In such a case, λ_L and λ_U are better estimated by utilizing two different elliptical copulas and taking only the lower left or the upper right observations of the copula into consideration (see the example below).

Example 3. Suppose that C is a specific Archimedean copula such as the Gumbel copula

$$C_{GU}(u,v) = \exp\left[-\left\{(-\log u)^{\frac{1}{\theta}} + (-\log v)^{\frac{1}{\theta}}\right\}^{\theta}\right],$$

where $0 < \theta \leq 1$. It is easy to show that $\lambda_U(\theta) = 2 - 2^{\theta}$ and therefore $A(1/2) = 2^{\theta}/2$. Thus, λ_U may be estimated via $\lambda_U(\hat{\theta})$ where $\hat{\theta}$ is obtained from a fitted Gumbel copula. □

In general, Archimedean copulas are described by a continuous, strictly decreasing and convex generator function $\phi : [0,\, 1] \to [0,\, \infty]$ with $\phi(1) = 0$. The copula C is then given by

$$C(u,v) = \phi^{[-1]}\left\{\phi(u) + \phi(v)\right\}.$$

Here $\phi^{[-1]} : [0,\, \infty] \to [0,\, 1]$ denotes the pseudo-inverse of ϕ. The generator ϕ is called strict if $\phi(0) = \infty$ and in this case $\phi^{[-1]} = \phi^{-1}$; see Genest and MacKay (1986) or Nelsen (2006, Ch. 4).

Suppose (U, V) is distributed with Archimedean copula C with generator $\phi(\cdot\,;\theta)$ involving an unknown parameter θ. Recall that the TDC is defined along the copula's diagonal. In this context, we mention the following useful relationship

$$\mathbb{P}\{\max(U,V) \leq t\} = C(t,t) = \phi^{-1}\left\{2\,\phi(t\,;\theta)\,;\theta\right\}.$$

Example 4. Consider the following conditional distribution function:

$$\mathbb{P}(U \leq u, V \leq v \mid U, V \leq t) = \frac{C(u,v)}{C(t,t)}, \qquad 0 < t < 1,\ 0 \leq u, v \leq t.$$

Observe that we may only consider data which fall below the threshold t in order to estimate the lower TDC. The conditional distribution function of the upper right quadrant of C is similarly defined. The point is that it is useful to allow completely different conditional distributions for lower left and upper right observations of the copula. Note that the typical bias-variance trade-off appears again for the choice of the threshold t (as discussed in Section 1.1). □

1.3.4. Estimation within a Class of Copulas

Let us consider the important class of Archimedean copulas. Juri and Wüthrich (2002) have derived the following limiting result for the bivariate excess distribution of Archimedean copulas C. Define

$$F_t(x) = \frac{C\{\min(x,t),t\}}{C(t,t)}, \qquad 0 \le x \le 1,$$

where $0 < t < 1$ is a *low* threshold. Note that F_t can be also defined via the second argument of C since $C(u,v) = C(v,u)$. Now consider the "copula of small values" defined by

$$C_t(u,v) = \frac{C\left\{F_t^{-1}(u), F_t^{-1}(v)\right\}}{C(t,t)}, \tag{1.7}$$

where F_t^{-1} is the generalized inverse of F_t. It can be shown that if C has a differentiable and regularly varying generator ϕ with tail index $\alpha > 0$ then

$$\lim_{t\to 0^+} C_t(u,v) = C_{\mathrm{Cl}}(u,v\,;\alpha),$$

for every $0 \le u, v \le 1$, where

$$C_{\mathrm{Cl}}(u,v\,;\alpha) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}$$

is the Clayton copula with parameter α. One may verify that $\lambda_L = \lambda_L(\alpha) = 2^{-1/\alpha}$. Thus, the lower TDC can be estimated by fitting the Clayton copula to small values of the approximate copula realizations and set $\hat{\lambda}_L = 2^{-1/\hat{\alpha}}$.

Remarks.

i) Archimedean copulas that belong to a domain of attraction are necessarily in the domain of attraction of the Gumbel copula, which is an EV copula; see, e.g., Genest and Rivest (1989) and Capéraà and Fougères (2000). Hence, the Gumbel copula seems to be a natural choice regarding the TDC estimation if we work in an Archimedean framework.

ii) The marginal distributions of financial asset returns are commonly easier to model than the corresponding dependence structure; this is often due to the limited availability of data. Consider for instance the pricing of so-called *basket credit derivatives*. Here the marginal survival functions of the underlying credits are usually estimated via parametric hazard-rate models by utilizing observable default spreads.

The choice of an appropriate dependence structure, however, is still a debate and several approaches are currently discussed; see, e.g., Li (1999) or Laurent and Gregory (2005).

1.3.5. Nonparametric Estimation

In the present section, no parametric assumptions are made for the copula and the marginal distribution functions. TDC estimates are obtained from the empirical copula $\widehat{C}_n$. Note that the empirical copula implies the following relationship

$$\widehat{F}_n(x,y) = \widehat{C}_n\{\widehat{G}_n(x), \widehat{H}_n(y)\},$$

where $\widehat{F}_n$, $\widehat{G}_n$, and $\widehat{H}_n$ denote the empirical distributions.

In (1.4), we presented the nonparametric upper TDC estimator $\hat{\lambda}_U^{\mathrm{LOG}}$ which is based on the empirical copula. This estimator was motivated by equation (1.3). Note that if the bivariate data are stochastically independent (or comonotonic), $\hat{\lambda}_U^{\mathrm{LOG}}$ is well behaved for all thresholds k in terms of the bias, as in that case $C(t,t) = t^2$ (or $C(t,t) = t$) and thus $\hat{\lambda}_U^{\mathrm{LOG}} \approx 2 - 2 = 0$ (or $\hat{\lambda}_U^{\mathrm{LOG}} \approx 2 - 1 = 1$) holds independently of k.

Another estimator appears as a special case in Joe et al. (1992):

$$\hat{\lambda}_U^{\mathrm{SEC}} = 2 - \frac{1 - \widehat{C}_n\left(\frac{n-k}{n}, \frac{n-k}{n}\right)}{1 - \frac{n-k}{n}}, \qquad 0 < k \leq n. \tag{1.8}$$

This estimator can also be motivated by equation (1.3), which explains the superscript SEC illustrating the relationship to the secant of the copula's diagonal. Asymptotic normality and strong consistency of $\hat{\lambda}_U^{\mathrm{SEC}}$ are, e.g., addressed in Schmidt and Stadtmüller (2006).

A third nonparametric estimator is proposed below which is motivated in Capéraà et al. (1997). Let $\{(U_1, V_1), \ldots, (U_n, V_n)\}$ be a random sample obtained from the copula C. Assume that the empirical copula function approximates an EV copula C_{EV} (take block maxima if necessary) and define

$$\hat{\lambda}^{\mathrm{CFG}} = 2 - 2\exp\left[\frac{1}{n}\sum_{i=1}^{n} \log\left\{\sqrt{\log\frac{1}{U_i}\log\frac{1}{V_i}} \Big/ \log\frac{1}{\max(U_i, V_i)^2}\right\}\right].$$

1.3.6. Pitfalls

From finitely many observations $(x_1, y_1), \ldots, (x_n, y_n)$ of (X, Y), it is difficult to conclude whether (X, Y) is tail dependent or not. As for tail-index estimation, one can always specify

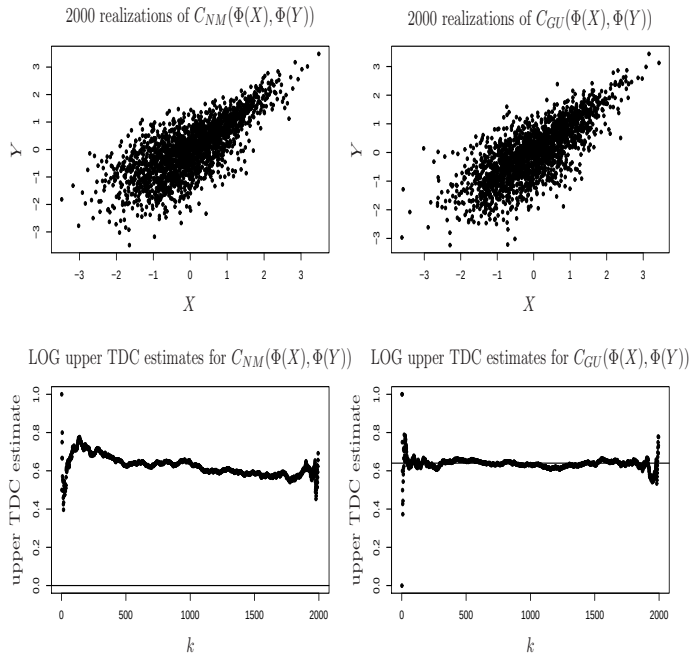

Figure 1.1.: Scatterplots of 2000 simulated data with standard normal margins and copula C_{NM} (upper left) and Gumbel copula C_{GU} (upper right), respectively. The lower plots show the corresponding TDC estimates $\hat{\lambda}_U^{\text{LOG}}$ for different choices of k. The horizontal lines indicate the true TDCs.

thin-tailed distributions which produce sample observations suggesting heavy tails even for large sample sizes. For example the upper left plot of Figure 1.1 shows the scatter plot of 2000 realizations from a distribution with standard normal univariate margins and copula C_{NM} corresponding to a mixture distribution of different bivariate Gaussian distributions, namely:

$$NM \;=\; \tfrac{7}{10}\cdot\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}0.49 & 0.245\\ 0.245 & 0.49\end{bmatrix}\right)+\tfrac{3}{10}\cdot\mathcal{N}\left(\begin{bmatrix}1\\1\end{bmatrix},\begin{bmatrix}0.49 & 0.441\\ 0.441 & 0.49\end{bmatrix}\right).$$

At first glance, the scatter plot reveals upper tail-dependence although any finite mixture of normal distributions is tail independent. The upper right plot of Figure 1.1 shows the scatter plot of 2000 realizations from a distribution with standard normal univariate margins and a Gumbel copula with $\theta = 2.25$. As expected, the sample reveals a large upper TDC of $\lambda_U = 2 - 2^{\frac{1}{\theta}} \approx 0.64$. Nevertheless, the upper left plot with $\lambda_U = 0$ looks more or less like the upper right plot. The lower two plots of Figure 1.1 give the corresponding TDC estimates of $\hat{\lambda}_U^{\text{LOG}}$ for different choices of k. It can be seen that for any choice of k the TDC estimate for copula C_{NM} has nearly the same value as the TDC estimate for the

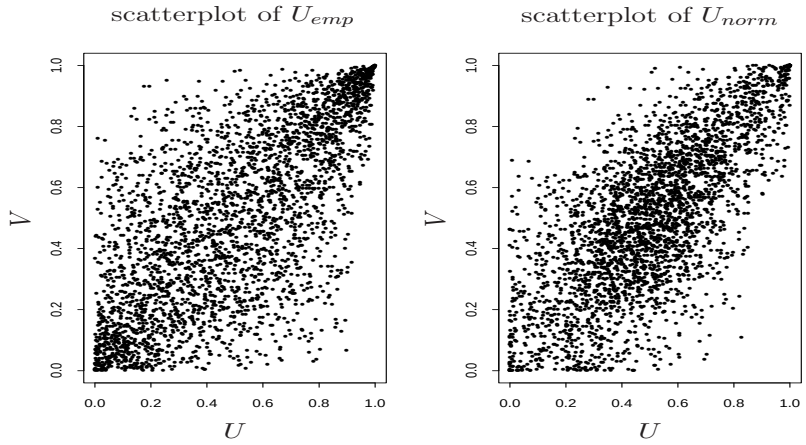

Figure 1.2.: Comparison of empirical copula densities obtained via empirical marginal distributions (left panel) and via fitted normal marginal distributions (right panel). The marginal transformations in the right panel are misspecified.

Gumbel copula. Conversely one may create samples which seem to be tail independent but they are realizations of a tail dependent distribution. Thus, the message is that one must be careful by inferring tail-dependence from a finite random sample. The best way to protect against misidentifications is the application of several estimators, test or plots to the same data set.

We address another pitfall regarding the estimation of the marginal distribution functions. The use of parametric margins instead of empirical margins bears a model risk and may lead to wrong interpretations of the dependence structure. For instance, consider 3000 realizations of a random pair with distribution function

$$H(x,y) = C_{GU}\left\{t_\nu(x), t_\nu(y); \theta\right\},$$

where t_ν denotes the univariate standard t-distribution with ν degrees of freedom and C_{GU} is the Gumbel copula with parameter θ. Set $\theta = 2$ and $\nu = 3$. In Figure 1.2, we compare the empirical copula densities which are either obtained via empirical marginal distributions or via fitted normal marginal distributions. Precisely, in the second case we plot the pairs $(G(x_i), H(y_i))$, where G and H are normal distribution functions with parameters estimated from the data. The left panel of Figure 1.2 clearly illustrates the dependence structure of a Gumbel copula. By contrast, the data in the right panel have nearly lost all the appearance for upper tail-dependence. Thus we have shown that not testing or ignoring the quality of the marginal fit can cause dramatic misinterpretation of the underlying dependence structure.

1.4. Simulation Study

In order to compare the finite sample properties of the discussed TDC estimators, we run an extensive simulation study. Each simulated data set consists of 1000 independent copies of n realizations from a random sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ having one particular distribution out of four. Three different sample sizes $n = 250, 1000, 5000$ are considered for each data set. The four different distributions are denoted by H, T, G and AG. For example the data set S_H^{250} contains realizations of 1000 samples with sample size 250 which are generated from a bivariate symmetric generalized hyperbolic distribution H. This is an elliptical distribution (see (1.6)), where $\mathcal{R}U^{(2)}$ has density

$$f(s) = \frac{K_0\left(\sqrt{1+s's}\right)}{2\pi \cdot K_0(1)}$$

and K_0 is the Bessel function of the third kind with index 0. The correlation parameter is set to $\rho = 0.5$.

Further, T refers to the bivariate t-distribution with $\nu = 1.5$ degrees of freedom and $\rho = 0.5$. Distribution G is determined by the distribution function

$$F_G(x, y) = C_G\left\{\Phi(x), \Phi(y); \vartheta, \delta\right\},$$

where Φ denotes the univariate standard normal distribution and C_G is an Archimedean copula with generator function

$$\phi_G(t; \vartheta, \delta) = \left\{\phi_{\text{Frank}}(t; \vartheta)\right\}^{\delta} = \left(-\log \frac{e^{-\vartheta t} - 1}{e^{-\vartheta} - 1}\right)^{\delta}, \qquad \vartheta \neq 0, \delta \geq 1,$$

considered by Junker and May (2005). Here, ϕ_{Frank} is the generator of the Frank copula, and values $\vartheta = -0.76$ and $\delta = 1.56$ are chosen, for reasons given below.

Finally, distribution AG is an asymmetric Gumbel copula, as defined by Tawn (1988), combined with standard normal margins, viz.

$$F_{AG}(x, y) = C_{AG}\left\{\Phi(x), \Phi(y); \theta, \phi, \delta\right\},$$

where

$$C_{AG}(u, v; \theta, \phi, \delta) = u^{1-\theta} v^{1-\phi} \exp\left(-\left[\left\{-\theta \ln(u)\right\}^{\delta} + \left\{-\phi \ln(v)\right\}^{\delta}\right]^{\frac{1}{\delta}}\right).$$

We set $\theta = 0.5, \phi = 0.9, \delta = 2.78$. For additional ways of generating asymmetric models and multi-parameter Archimedean copulas, see Genest et al. (1998).

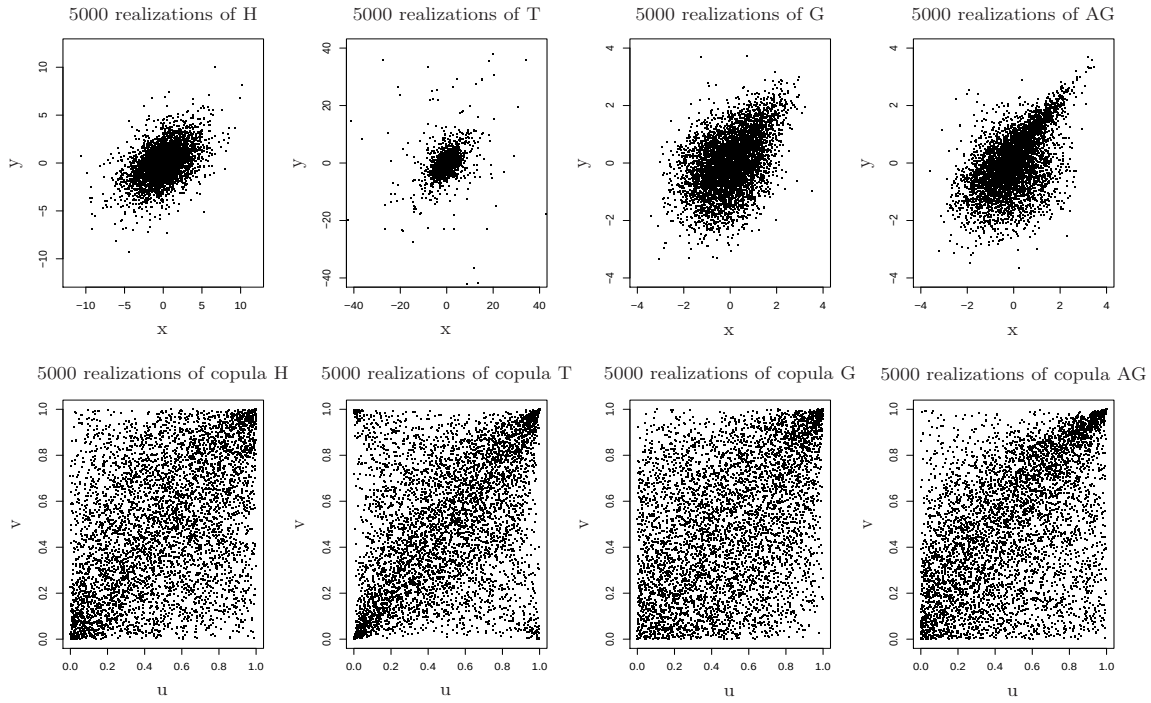

Figure 1.3.: Scatter plots of simulated distributions and corresponding empirical copula realizations.

Note that distribution H has no tail-dependence; e.g., see Schmidt (2002). Thus the set of generalized hyperbolic data is used to control the performance of the TDC estimation methods under the absence of tail-dependence. By contrast, distribution T possesses tail-dependence; see also (1.5). Further, copula C_G is lower tail independent but upper tail dependent, i.e., $\lambda_L^G = 0$ and $\lambda_U^G = 2 - 2^{1/\delta}$. The parametrization of the distributions T, G, and AG is chosen such that $\lambda_L^T = \lambda_U^T = \lambda_U^G = \lambda_U^{AG} = 0.4406$ and Kendall's tau $\tau^H = \tau^T = \tau^G = \tau^{AG} = 1/3$ in order to provide comparability of the estimation results. Figure 1.3 illustrates the different tail behavior of distributions H, T, G, and AG by presenting the scatter plots of the respective simulated data-sample with sample size $n = 5000$, together with the corresponding empirical copula realizations. Regarding the copula mapping, we use empirical marginal distribution functions.

The different estimation methods are compared via the sample means $\widehat{\mu}\left(\hat{\lambda}_n\right)$ and the sample standard deviations $\widehat{\sigma}\left(\hat{\lambda}_n\right)$ of the estimates $\hat{\lambda}_{n,i}$, $i = 1, \ldots, 1000$, depending on the sample size n. Furthermore, to analyze the bias-variance trade-off for different sample sizes and estimation methods we compare the corresponding root mean squared errors:

$$\mathrm{RMSE}\left(\hat{\lambda}_n\right) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\lambda}_{n,i} - \lambda\right)^2}. \tag{1.9}$$

Moreover, we introduce another statistical quantity called MESE (mean error to standard error):

$$\mathrm{MESE}\left(\hat{\lambda}_n\right) = \frac{\left|\widehat{\mu}\left(\hat{\lambda}_n\right) - \lambda\right|}{\widehat{\sigma}\left(\hat{\lambda}_n\right)}. \tag{1.10}$$

MESE quantifies the sample bias normalized by the sample standard error. Thus, it measures the degree of possible misinterpretation caused by considering the standard error as a criterion for the quality of the estimator. For instance, assume a situation where the standard error of the estimate is small but the bias is large. In that case the true parameter is far away from the estimate, though the approximated confidence bands suggest the opposite. This situation is represented by a large MESE. In particular, if the bias of the estimator decreases with a slower rate as the standard error (for $n \to \infty$) then MESE tends to infinity. One aim of this quantity is to investigate the danger of this sort of misinterpretation.

In the following, the TDC is estimated via the various methods discussed in Section 1.3. It is reasonable to discard those models which are obviously not compatible with the observed data. Further, we do not consider TDC estimations using a specific distribution since the results are not surprising (due to the strong distributional assumptions).

1.4.1. Estimation within a Class of Distributions

The following estimation approach is based on the expositions in Section 1.3.2. We have to estimate the tail index α and the correlation parameter ρ. For any elliptical distribution, the correlation parameter is determined by Kendall's τ via the relationship of Lindskog et al. (2003), viz. $\rho = \sin(\pi\tau/2)$. Hence, using $\hat{\tau} = (c-d)/(c+d)$, where c is the number of concordant pairs and d is the number of discordant pairs of the sample, the correlation parameter may be estimated by $\hat{\rho} = \sin(\pi\hat{\tau}/2)$. Alternatively, one can apply Tyler's M-estimator for the covariance matrix, which is completely robust within the class of elliptical distributions; see Tyler (1987a) or Frahm (2004, Ch. 4). Given the covariance matrix, the random variable $\mathcal{R}$ can be extracted by the Mahalanobis norm of (X, Y). Our pre-simulations showed that there is no essential difference regarding the finite sample properties between these two estimation procedures. Hence we use the approach via Kendall's τ for the sake of simplicity.

The tail index α could be estimated via traditional methods of EVT, i.e., by taking only extreme values or excesses of the Euclidean norm $\|(X, Y)\|_2$ into consideration. Different

methods for estimating the tail index are discussed, e.g., in Matthys and Beirlant (2002) or Embrechts et al. (1997). For our purposes, we used a Hill-type estimator with optimal sample fraction proposed by Drees and Kaufmann (1998).

For the data sets S_G and S_{AG}, we do not assume an elliptical distribution due to the obvious asymmetry of the data; see the scatter plots in Figure 1.3. Consequently we will not apply the latter estimation procedure to these data sets. The estimation results for S_H and S_T are summarized in Table 1.1.

1.4.2. Estimation Using a Specific Copula

As mentioned in Section 1.3.3, the marginal distribution functions are now estimated by their empirical counterparts, whereas the copula is chosen according to our decision. For the elliptical data sets S_H and S_T, we opted for a t-copula, which seems to be a realistic choice by glancing at the scatter plots in Figure 1.3. However, we know that the t-copula is not suitable for S_H. The TDC is estimated via relation (1.5). Regarding the data set S_G, we fit a Gumbel copula since the empirical copula, which is illustrated in Figure 1.3, shows transpositional asymmetry, i.e., the underlying copula does not seem to coincide with its survival copula. Moreover, the symmetry of the Gumbel copula with respect to the copula's diagonal appears to be satisfied by S_G, too. Here the upper TDC is estimated via $\hat{\lambda}_U^G = 2 - 2^{1/\widehat{\theta}}$. However, the original copula of S_G is not the Gumbel copula and thus the assumed model is wrong. We disregard the data set S_{AG} since it is not obvious which specific copula might be appropriate in this framework. Note that the empirical copula is even asymmetric with respect to the copula's diagonal (see Figure 1.3). The estimation results are summarized in Table 1.1.

1.4.3. Estimation within a Class of Copulas

We follow the approach given in Section 1.3.4, which is based on a result by Juri and Wüthrich (2002). The upper TDC is estimated, but in the following we refer to the lower left corner of the underlying survival copula. We choose a small threshold t for the latter copula in order to obtain the conditional copula (1.7). In order to increase the robustness of the copula estimates with respect to the threshold choice, we take the mean of estimates which correspond to 10 equidistant thresholds between $n^{-1/2}$ and $n^{-1/4}$. Note that if the margins of the underlying distribution are completely dependent, then $n^{1/2}$ data points

are expected in the copula's lower left quadrant which is determined by the threshold $t = n^{-1/2}$. For the smaller threshold $t = n^{-1/4}$, the same amount of data ($n^{1/2}$) is expected for an independence copula.

We assume that the data S_H, S_T and S_G have an Archimedean dependence structure. Since S_{AG} is not permutational symmetric, we reject the Archimedean hypothesis for this data set. We point out that the Frank copula (Frank, 1979) is the only transpositional symmetric Archimedean copula and thus suitable for S_H and S_T but it does not comprise tail-dependence. The statistical results for the data sets S_H, S_T, and S_G are provided in Table 1.1.

1.4.4. Nonparametric Estimation

No specific distributional assumptions for the upper TDC estimation are made in the present section. Recall that for $\hat{\lambda}_U^{\mathrm{LOG}}$ and $\hat{\lambda}_U^{\mathrm{SEC}}$, we have to choose the threshold k as indicated in (1.4) and (1.8). By contrast, $\hat{\lambda}_U^{\mathrm{CFG}}$ needs no additional decision regarding the threshold. This, however, goes along with the assumption that the underlying copula can be approximated by an EV copula.

The diagonal section of the copula is supposed to be smooth in the neighborhood of 1, and the second derivative of the diagonal section is expected to be small (i.e., the first derivative is approximately constant). Then $\hat{\lambda}_U^{\mathrm{SEC}}(k)$ is homogeneous for small (thresholds) k. However, k should be sufficiently large in order to decrease variance. We consider the graph $k \mapsto \hat{\lambda}_U^{\mathrm{SEC}}(k)$ in order to identify the plateau which is induced by the homogeneity. Note that $\hat{\lambda}_U^{\mathrm{LOG}}$ possesses this homogeneity property even for larger thresholds k if the diagonal section of the copula follows a power law.

The plateau is chosen according to the following heuristic plateau-finding algorithm. First, the map $k \mapsto \hat{\lambda}_k$ is smoothed by a simple box kernel with bandwidth $b \in \mathbb{N}$. That is, the means of $2b + 1$ successive points of $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ lead to the new smoothed map $\overline{\lambda}_1, \ldots, \overline{\lambda}_{n-2b}$. Here we have taken $b = \lfloor 0.005n \rfloor$ such that each moving average consists of 1% of the data, approximately. In a second step, a plateau of length $m = \lfloor \sqrt{n - 2b} \rfloor$ is defined as a vector $p_k = \left(\overline{\lambda}_k, \ldots, \overline{\lambda}_{k+m-1}\right)$, $k = 1, \ldots, n - 2b - m + 1$. The algorithm stops at the first plateau p_k which elements fulfill the condition $\sum_{i=k+1}^{k+m-1} \left|\overline{\lambda}_i - \overline{\lambda}_k\right| \leq 2\sigma$, where σ represents the standard deviation of $\overline{\lambda}_1, \ldots, \overline{\lambda}_{n-2b}$. Then the TDC estimate is set to $\hat{\lambda}_U(k) = 1/m \sum_{i=1}^{m} \overline{\lambda}_{k+i-1}$. If there is no plateau fulfilling the stopping condition, the TDC estimate is set to zero.

As outlined above, we may choose a greater bandwidth b for the $\hat{\lambda}^{\mathrm{LOG}}$ in order to reduce the variance of the estimation. However, for a better comparison we do not change b. The statistical results related to these nonparametric estimators for the data sets S_H, S_T, S_G and S_{AG} are provided in Table 1.2 and Table 1.3.

1.4.5. Discussion of the Simulation Results

We discuss the simulation results with regard to the following statistical measures: sample variance, sample bias, RMSE, and MESE.

Sample variance. The TDC estimations within the class of elliptically contoured distributions and for specific copulas show the lowest sample variances among all considered methods of TDC estimation. Of course, the small variances go along with restrictive model assumptions. Nevertheless the estimation within the class of elliptically contoured distributions has a surprisingly low variance, even though the tail index α is estimated from few extremal data. Further, a comparably small variance is obtained for the estimator $\hat{\lambda}^{\mathrm{CFG}}$ which is based on the weaker assumption of an EV copula. By contrast, the sample variances of $\hat{\lambda}^{\mathrm{SEC}}$ and $\hat{\lambda}^{\mathrm{LOG}}$ are much larger. In particular the TDC estimation within the class of Archimedean copulas, as described in Section 1.3.4, shows an exceptionally large variance. However, note that the latter three estimation methods utilize only sub-samples of extremal (excess) data. Besides, there is another explanation for the large variance of the last estimation method: Here, the TDC is estimated (in a second step) from a copula which is fitted from extremal (excess) data.

We conclude that an effective variance reduction of the TDC estimation is possible for those estimation methods which use the entire data sample.

Sample bias. It is not surprising that the estimation methods with distributional assumption have a quite low sample bias if the underlying distribution is true. See, for example, the bias related to S_T for TDC estimations within the class of elliptical distributions or for specific copulas; see also S_G for the estimation within the class of Archimedean copulas. By contrast, the estimation with regard to the sample bias performs badly if we assume an inappropriate distribution, as can be seen for the data set S_G under the estimation using a specific copula; see Section 1.3.3 and the data set S_T under the estimation within the class of Archimedean copulas. It is, however, surprising that the TDC estimation from S_H (recall that H is an elliptical distribution) shows a larger sample bias for the estimation

Method	Data set	$\hat{\mu}(\hat{\lambda}_U)$	BIAS$(\hat{\lambda}_U)$	$\hat{\sigma}(\hat{\lambda}_U)$	RMSE$(\hat{\lambda}_U)$	MESE$(\hat{\lambda}_U)$
Estimation	S_H^{250}	0.1618	0.1618^+	0.0817	0.1812^+	1.9802
for a	S_H^{1000}	0.1698	0.1698	0.0413	0.1747^+	4.1116
specific copula	S_H^{5000}	0.1739	0.1739	0.0187	0.1749	9.3141
(t- and	S_T^{250}	0.4281	-0.0125	0.0403^+	0.0422^+	0.3078
Gumbel	S_T^{1000}	0.4374	-0.0032	0.0204^+	0.0206^+	0.1403
copula)	S_T^{5000}	0.4400	-0.0006^+	0.0092^+	0.0092^+	0.0652^+
	S_G^{250}	0.3905	-0.0501^-	0.0437^+	0.0664	1.1466^-
	S_G^{1000}	0.3922	-0.0484^-	0.0212^+	0.0529	2.2819^-
	S_G^{5000}	0.3919	-0.0487^-	0.0097^+	0.0497	5.0252^-
Estimation	S_H^{250}	0.2031	0.2031	0.0588	0.2114	3.4541
within a class	S_H^{1000}	0.1815	0.1815	0.0377	0.1854	4.8143
of distributions	S_H^{5000}	0.1575	0.1575	0.0220	0.1590	7.1591
(elliptical	S_T^{250}	0.4379	-0.0027^+	0.0465	0.0466	0.0490^+
distributions)	S_T^{1000}	0.4432	0.0026^+	0.0242	0.0243	0.1041^+
	S_T^{5000}	0.4437	0.0031	0.0109	0.0113	0.2849
Estimation	S_H^{250}	0.2278	0.2278	0.1910^-	0.2972	1.1921^+
within a class	S_H^{1000}	0.1671	0.1671^+	0.1357^-	0.2152	1.2309^+
of copulas	S_H^{5000}	0.1237	0.1237^+	0.0977^-	0.1576^+	1.2657^+
(Archimedean	S_T^{250}	0.5317	0.0911	0.1864^-	0.2074^-	0.4880
copulas)	S_T^{1000}	0.5575	0.1169^-	0.1175^-	0.1657^-	0.9943^-
	S_T^{5000}	0.5701	0.1295^-	0.0647^-	0.1448^-	2.0022^-
	S_G^{250}	0.4352	-0.0054^+	0.1948^-	0.1948^-	0.0277^+
	S_G^{1000}	0.4495	0.0089^+	0.1312^-	0.1314^-	0.0548^+
	S_G^{5000}	0.4554	0.0148	0.0792^-	0.0805^-	0.1818^+

Table 1.1.: Various statistical results for the TDC estimation under a specific copula assumption or within a class of distributions or copulas. Best values of the different methods (including the nonparametric methods in Table 1.2 and Table 1.3) are ticked with a plus, worst values are ticked with a minus.

Method	Data set	$\hat{\mu}(\hat{\lambda}_U)$	BIAS$(\hat{\lambda}_U)$	$\hat{\sigma}(\hat{\lambda}_U)$	RMSE$(\hat{\lambda}_U)$	MESE$(\hat{\lambda}_U)$
Nonparametric	S_H^{250}	0.3553	0.3553	0.0444^+	0.3580	8.0008^-
estimator	S_H^{1000}	0.3558	0.3558^-	0.0229^+	0.3566^-	15.5400^-
$\hat{\lambda}_U^{\mathrm{CFG}}$	S_H^{5000}	0.3568	0.3568^-	0.0104^+	0.3570^-	34.3123^-
	S_T^{250}	0.4462	0.0056	0.0471	0.0474	0.1133
	S_T^{1000}	0.4509	0.0103	0.0234	0.0256	0.4437
	S_T^{5000}	0.4511	0.0105	0.0107	0.0150	0.9825
	S_G^{250}	0.3922	-0.0484	0.0450	0.0661^+	1.0759
	S_G^{1000}	0.3939	-0.0467	0.0216	0.0514^+	2.1593
	S_G^{5000}	0.3936	-0.0470	0.0099	0.0480	4.7443
	S_{AG}^{250}	0.4377	-0.0029	0.0480^+	0.0481^+	0.0648^+
	S_{AG}^{1000}	0.4400	-0.0006	0.0243^+	0.0243^+	0.0247^+
	S_{AG}^{5000}	0.4406	0.0000^+	0.0107^+	0.0107^+	0.0000^+

Table 1.2.: Statistical results for the nonparametric TDC estimator $\hat{\lambda}_U^{\mathrm{CFG}}$. Best values of the different methods (including the parametric methods in Table 1.1 and nonparametric methods in Table 1.3) are ticked with a plus, worst are ticked with a minus.

within the class of elliptical distributions than for the estimation with a specific copula or within the class of Archimedean copulas. We point out that the largest sample bias is observed for the nonparametric estimation methods. Further, all estimation methods, in particular $\hat{\lambda}^{\mathrm{CFG}}$, yield a large MESE value (which indicates a large-sample bias relative to the sample variance) for the data set S_H which exhibits tail-independence. In most cases, the MESE is greater than 2, which means that the true TDC value is not included in the 2σ confidence interval. Moreover, this illustrates that the sole consideration of the sample variance may lead to the fallacy of an exceptionally large TDC, even in the case of tail-independence. Thus it is absolutely necessary to test for tail-dependence in the first instance; see Ledford and Tawn (1996), Draisma et al. (2004).

RMSE. The TDC estimation using a specific copula represents the smallest RMSE if the underling copula is true, as applies, e.g., to the data set S_T. The second best RMSE for the latter data set comes from the estimation within the class of elliptically contoured distributions. This estimation goes along with a larger variance, due to the estimation of the tail index α. It is remarkable that the nonparametric estimator $\hat{\lambda}^{\mathrm{CFG}}$ (which assumes an EV copula) possesses an RMSE in the same range as the two aforementioned estimation

Method	Data set	$\hat{\mu}(\hat{\lambda}_U)$	BIAS$(\hat{\lambda}_U)$	$\hat{\sigma}(\hat{\lambda}_U)$	RMSE$(\hat{\lambda}_U)$	MESE$(\hat{\lambda}_U)$
Nonparametric	S_H^{250}	0.3636	0.3636^-	0.1016	0.3775^-	3.5787^-
estimator	S_H^{1000}	0.3056	0.3056	0.0717	0.3139	4.2622
$\hat{\lambda}_U^{\text{SEC}}$	S_H^{5000}	0.2390	0.2390	0.0932	0.2565^-	2.5644
	S_T^{250}	0.4681	0.0275	0.0800	0.0845	0.3436
	S_T^{1000}	0.4587	0.0181	0.0513	0.0545	0.3534
	S_T^{5000}	0.4463	0.0057	0.0431	0.0435	0.1322
	S_G^{250}	0.4841	0.0435	0.0796	0.0907	0.5467
	S_G^{1000}	0.4650	0.0244	0.0482	0.0541	0.5062
	S_G^{5000}	0.4453	0.0047^+	0.0603	0.0605	0.0775^+
	S_{AG}^{250}	0.5042	0.0636^-	0.0810^-	0.1029^-	0.7835^-
	S_{AG}^{1000}	0.4763	0.0357^-	0.0523^-	0.0633^-	0.6818^-
	S_{AG}^{5000}	0.4567	0.0161^-	0.0340^-	0.0376^-	0.4722^-
Nonparametric	S_H^{250}	0.3144	0.3144	0.0828	0.3251	3.7968
estimator	S_H^{1000}	0.2893	0.2893	0.0539	0.2943	5.3677
$\hat{\lambda}_U^{\text{LOG}}$	S_H^{5000}	0.2567	0.2567	0.0377	0.2595	6.8103
	S_T^{250}	0.3951	-0.0455^-	0.0727	0.0857	0.6242^-
	S_T^{1000}	0.4132	-0.0274	0.0491	0.0562	0.5569
	S_T^{5000}	0.4240	-0.0166	0.0352	0.0389	0.4704
	S_G^{250}	0.4016	-0.0390^+	0.0719	0.0818	0.5426
	S_G^{1000}	0.4098	-0.0308	0.0448	0.0544	0.6850
	S_G^{5000}	0.4229	-0.0177	0.0233	0.0293	0.7624
	S_{AG}^{250}	0.4424	0.0018^+	0.0696	0.0696	0.0259
	S_{AG}^{1000}	0.4403	-0.0003^+	0.0386	0.0386	0.0078
	S_{AG}^{5000}	0.4412	0.0006	0.0200	0.0200	0.0030

Table 1.3.: Statistical results for the nonparametric TDC estimators $\hat{\lambda}_U^{\text{SEC}}$ and $\hat{\lambda}_U^{\text{LOG}}$. Best values of the different methods (including the parametric methods in Table 1.1 and nonparametric methods in Table 1.2) are ticked with a plus, worst are ticked with a minus.

methods for the data sets S_T, S_G, and S_{AG}. By contrast, the estimators $\hat{\lambda}^{\mathrm{SEC}}$ and $\hat{\lambda}^{\mathrm{LOG}}$ have a much larger RMSE. The TDC estimation based on the class of Archimedean copulas as described in Section 1.3.4 yields by far the largest RMSE even for the (Archimedean) data set S_G. Further, the estimation using a specific copula has a similar RMSE under the wrong model assumption (see S_G) due to its low variance. An RMSE in the same range is found for the nonparametric estimators $\hat{\lambda}^{\mathrm{SEC}}$ and $\hat{\lambda}^{\mathrm{LOG}}$. This also suggest a consideration of the following statistical measure.

MESE. The MESE detects all misspecified models such as S_G under the estimation using a specific (Gumbel) copula, or S_T under the estimation within the class of Archimedean copulas. However, if the underlying model is true, then the MESE is quite low (e.g., estimation within a class of copulas or distributions for the data set S_T). In this case and for all nonparametric estimations, the MESE is usually smaller than 1, which indicates that the true TDC lies within the 1σ confidence band. Only the data set S_H represents an exception. Especially the estimator $\hat{\lambda}^{\mathrm{CFG}}$ shows an exceptionally bad performance, which is due to its small variance. Thus, the sample variance has to be considered with caution for the latter estimator.

There exists an interesting aspect regarding the estimator $\hat{\lambda}^{\mathrm{SEC}}$. Due to its geometric interpretation as the slope of the secant along the copula's diagonal (at the point $(1, 1)$), the latter estimator reacts sensitively if the extremal data are not accumulated along the diagonal. Such is the case, e.g., for the data set S_{AG} and might also explain the bad performance of $\hat{\lambda}^{\mathrm{SEC}}$ regarding the latter data set.

1.5. Conclusion

On the basis of the results of our simulation study, we have ranked the various TDC estimators according to their finite-sample performance. Table 1.4 illustrates the corresponding rankings in terms of numbers between 1 (very good performance) and 6 (very poor performance). Thereby we distinguish between a true and a wrong assumption of the underlying distribution. Moreover, we rank the estimators according to their performance under the assumption of tail-independence. The second column of Table 1.4 indicates the heaviness of the model assumptions required for each TDC estimator.

Clearly the (semi-)parametric TDC estimators perform well if the underlying distribution/copula is the right one (except the TDC estimator within a class of copulas as de-

TDC estimation methods	degree of assumptions	perform. under true assumpt.	perform. under wrong assumpt.	perform. under TDC = 0
specific copula	strong	1	6	1-2
distr. class	strong	2–3	–$^{\star)}$	3
copula class	medium	4–5	5	1–2
$\hat{\lambda}^{\mathrm{CFG}}$	weak	2–3	2–3	6
$\hat{\lambda}^{\mathrm{SEC}}$	weak	4	4	5
$\hat{\lambda}^{\mathrm{LOG}}$	weak	3	3	5

Table 1.4.: Overview of the performance of the TDC estimation methods. Grades rank from 1 to 6 with 1 excellent and 6 poor. $^{\star)}$ This TDC estimation method (via elliptical distributions) is disregarded due to the obvious asymmetry of the data arising from the wrong distributional assumption.

scribed in Section 1.3.4). However, their performance is very poor if the assumed model is wrong. Thus, we definitely recommend to test any distributional assumptions. For instance, in the case of empirical marginal distributions and specific copula, we suggest to test the goodness-of-fit of the copula via (non-)parametric procedures such as those developed in Fermanian (2005), Dobrić and Schmid (2005b) or Genest et al. (2006). Further, if one makes use of an elliptically contoured distribution, then we suggest to test for ellipticity; see, e.g., Manzotti et al. (2002). However, we do not recommend a TDC estimation as presented in Section 1.3.4, since we do not know a suitable test for Archimedean copulas and the estimator does not perform well if the assumption of an Archimedean copula is wrong. Goodness-of-fit tests within the family of Archimedean copulas are developed, e.g., in Wang and Wells (2000) or Genest et al. (2006).

Among the nonparametric estimators, the TDC estimator $\hat{\lambda}^{\mathrm{CFG}}$ does well, although we advise caution regarding the sometimes low variance relative to bias. Further, $\hat{\lambda}^{\mathrm{CFG}}$ shows a weak performance in the case of tail-independence. This estimator is followed by $\hat{\lambda}^{\mathrm{LOG}}$ and $\hat{\lambda}^{\mathrm{SEC}}$ whereas the last estimator is not robust for non-transpositional symmetric data. Further, the variance of $\hat{\lambda}^{\mathrm{LOG}}$ could be possibly reduced by enlarging the estimation kernel (see Section 1.4.4).

We conclude that, among the nonparametric TDC estimators, $\hat{\lambda}^{\mathrm{CFG}}$ shows the best performance whereas for (semi-)parametric estimations we recommend a specific copula (such as the t-copula). For the latter, we suggest to work with empirical marginal distributions.

Further, we point out that the decision for a specific distribution or class of distributions should be influenced by the visual appearance of the data, e.g., via the related scatter plots. Unfortunately, if the number of data is small (such as 250 points), it is difficult to draw sensible conclusions from the scatter plot. Moreover, the nonparametric estimators are too sensitive in case of small sample sizes. Thus, under these circumstances, a parametric TDC estimation might be favorable in order to increase the stability of the estimation although the model error could be large.

The previous simulation is based on a limited number of distributions, although we tried to incorporate a large spectrum of possible distributions. Nevertheless, according to the pitfalls in Section 1.3.6 and the statistical results for the tail independent data set H, we see that tests for tail-dependence are absolutely mandatory for every TDC estimation. Unfortunately, the current literature on this kind of test is only limited; see Coles et al. (1999), Draisma et al. (2004).

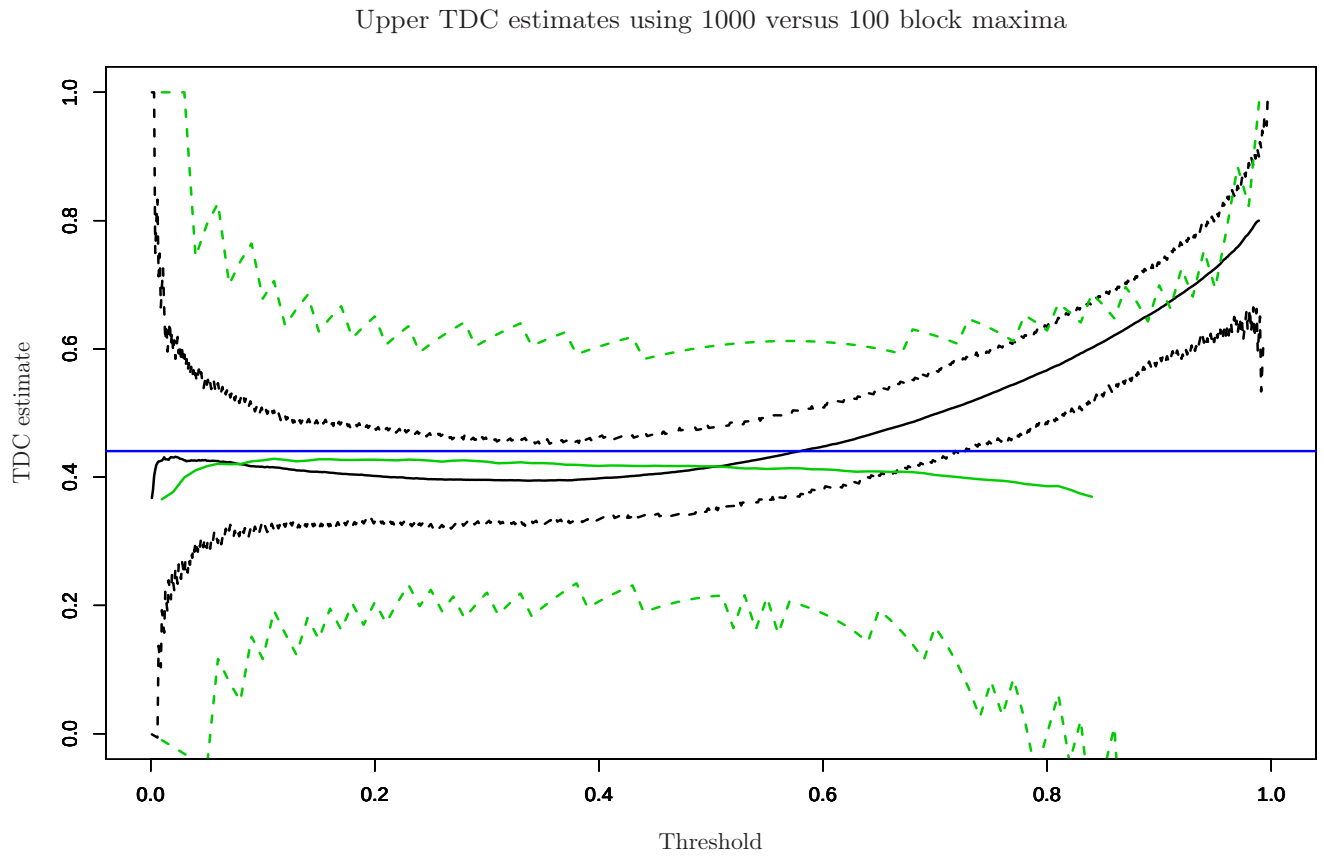

Figure 1.4.: Sample means of upper TDC estimates for various thresholds k using estimator (1.4) for 1000 samples of a bivariate t-distribution with 1.5 degrees of freedom and correlation parameter $\rho = 0.5$. The black solid line represents the case of 1000 block maxima ($\triangleq$ original data set) and the gray solid line corresponds to 100 block maxima. The related empirical 95% confidence intervals are indicated by the dashed lines. The true value of the TDC is marked by the solid straight line.

Chapter 2.

Dependence of Stock Returns in Bull and Bear Markets

2.1. Introduction

The standard estimator for the linear correlation coefficient according to Karl Pearson still seems to be the most commonly used measure of dependence of two random variables X and Y though its many shortcomings have been often documented (see, e.g., Embrechts et al., 2002). Pearson's rho is strongly affected by the marginal distributions of X and Y and its estimates are sensitive to outliers (Embrechts et al., 2002). Further, it only quantifies linear dependence though *monotone dependence* is often much more relevant. The random variables X and Y are said to possess a strong monotone dependence if there exist two real-valued and strictly increasing functions f and g such that $|\text{Corr}(f(X), g(Y))|$ is large.

It is easy to construct dependence structures where the linear correlation coefficient of X and Y is close to 0 but even so one can find two monotone transformations f and g such that $\text{Corr}(f(X), g(Y)) = 1$. Consider for instance the random variables $X = e^Z$ and $Y = e^{\sigma Z}$ with $\sigma > 0$ and $Z \sim \mathcal{N}(0, 1)$ (McNeil et al., 2005, p. 205). Since $\text{Corr}(\log X, \log Y) = 1$ they possess a *perfect* monotone dependence structure, i.e. X and Y are *comonotonic* (Nelsen, 2006, p. 32). Nevertheless, $\text{Corr}(X, Y)$ is a function of σ and can take any value between 0 ($\sigma \to \infty$) and 1 ($\sigma = 1$).

Copula theory and the dependence measures derived thereof are a convincing alternative to the linear correlation coefficient. Due to Sklar's theorem (Sklar, 1959) it is known that a joint distribution function can be split up into its *copula* (i.e. its dependence structure)

and its marginal distributions. A meaningful dependence measure should be invariant under monotone transformations of the corresponding random variables. Examples of such measures are Spearman's rho, Kendall's tau, Gini's gamma, and Blomquist's beta. In this paper we confine ourselves to the rank-correlation coefficient or its corresponding estimator, i.e. Spearman's rho. For surveys on copulas and dependence measures see, e.g., Cherubini et al. (2004), Joe (1997), and Nelsen (2006).

We investigate the contemporaneous dependence of two stock returns X and Y. In particular, we concentrate on the question whether dependence is significantly different in bull and bear markets, i.e. in case of a joint upswing or downswing. This question and related problems have been already investigated in finance literature (see, e.g., Ang and Chen, 2002, Erb et al., 1994, Fortin and Kuzmics, 2002, Junker and May, 2005, Patton, 2004, Silvapulle and Granger, 2001, Vaz de Melo Mendes, 2005). But we think that the statistical methods, in particular the use of Pearson's rho is unsatisfactory. Hence, there is space for further contributions.

Bear and bull markets are characterized as follows. A bear market is present if the two stock returns X and Y contemporaneously fall short of the $100p\%$ quantiles of their corresponding cumulative distribution functions. Analogously, a bull market is given whenever $-X$ and $-Y$ fall short of the corresponding $100q\%$ quantiles. Here p and q have to be pre-determined. The lower p-quantile of the cumulative distribution function (c.d.f.) of a stock return is commonly known as the *value-at-risk* where p is the so-called *shortfall probability*. The value-at-risk is frequently used in risk management. So it seems to be a natural choice for characterizing bull and bear markets.

Our approach is purely nonparametric. Contrary to Patton (2004) and Vaz de Melo Mendes (2005) we do not fit specific copulas to the data. Specifying the copula by some parametric model can lead to erroneous conclusions if the chosen model is wrong. From our point of view it is not necessary to rely on the parametric approach if the sample size is large enough. We are interested in financial data analysis and in that context it is easy to access many thousands of observations. By following the nonparametric approach we avoid any kind of model misspecification.

In this work we develop conditional versions of the rank-correlation coefficient to assess the dependence structure of stock returns in bull and bear markets. In contrast, some authors analyze the dependence structure of outliers in financial data by using the so-called *tail-dependence coefficient* (Fortin and Kuzmics, 2002, Junker and May, 2005). After

applying parametric methods these authors come to the conclusion that 'the empirical joint distribution of return pairs on stock indices displays high tail-dependence in the lower tail and low tail-dependence in the upper tail' (Fortin and Kuzmics, 2002). Dobrić and Schmid (2005a) as well as Frahm et al. (2005) found that estimating the tail-dependence coefficient by *nonparametric methods* can lead to very large estimation errors even if there are many observations. Hence the tail-dependence coefficient is not an appropriate alternative.

Though we focus on computational statistics and the empirical analysis of stock returns we have to introduce some statistical theory in order to have a formal basis for our testing procedure. This is done in Section 2.2, where some copula theory is presented. It allows a precise formulation of the null hypotheses to be tested. The testing procedure is described in Section 2.3. A Monte Carlo (MC) simulation is presented in Section 2.4 which shows that the procedure works well for sample sizes which are typically available in practice. In particular it is demonstrated that the test keeps the prescribed error probability of the first kind and has sufficient power to detect violations of the null hypothesis. In Section 2.5 we investigate the daily returns of stocks from the German stock index *DAX 30* between 1973-01-02 and 2008-11-14 and Section 2.6 concludes.

2.2. Some Copula Theory

In this section we introduce some notions from copula theory (Joe, 1997, Nelsen, 2006) which are required for understanding the testing procedure to be described below. Let X and Y be two random variables with joint c.d.f. $F(x,y) = \mathbb{P}(X \leq x, Y \leq y)$ and marginal cumulative distribution functions $G(x) = \mathbb{P}(X \leq x)$ and $H(y) = \mathbb{P}(Y \leq y)$ for all $x, y \in \mathbb{R}$. The quantile functions with respect to G and H are given by $G^{-1}(p) = \inf\{x\colon G(x) \geq p\}$ and $H^{-1}(p) = \inf\{y\colon H(y) \geq p\}$ for $0 \leq p \leq 1$.

Throughout this paper we assume that G and H are absolutely continuous functions. Therefore, according to Sklar's theorem (Sklar, 1959), there exists a unique copula $C : [0,1]^2 \to [0,1]$ such that

$$F(x,y) = C(G(x), H(y))\,, \qquad \forall\, x, y \in \mathbb{R}\,.$$

The function C is the joint c.d.f. of $U = G(X)$ and $V = H(Y)$. The rank-correlation coefficient of X and Y is given by

$$\rho := \mathrm{Corr}\big(U,V\big) = 12 \int\limits_{[0,1]^2} uv\, dC(u,v) - 3 = 12 \int\limits_{[0,1]^2} C(u,v)\, d(u,v) - 3\,. \qquad (2.1)$$

See Nelsen (2006, p. 167) for the latter representation of ρ.

For every fixed p with $0 < p < 1$ we define

$$A_{\mathrm{L}} := \Big\{(x,y)\colon x \le G^{-1}(p), y \le H^{-1}(p)\Big\}\,.$$

In the following we assume that $\mathbb{P}\{(X,Y) \in A_{\mathrm{L}}\} = C(p,p) > 0$. Consider the *conditional* joint c.d.f.

$$\begin{aligned} F_{\mathrm{L}}(x,y) &:= \mathbb{P}(X \le x, Y \le y \,|\, (X,Y) \in A_{\mathrm{L}}) = \frac{F(x \wedge G^{-1}(p), y \wedge H^{-1}(p))}{F(G^{-1}(p), H^{-1}(p))} \\ &= \frac{C(G(x \wedge G^{-1}(p)), H(y \wedge H^{-1}(p)))}{C(p,p)}\,, \qquad \forall\, x,y \in \mathbb{R}\,. \end{aligned}$$

The corresponding conditional marginal distribution functions are given by

$$\begin{aligned} G_{\mathrm{L}}(x) &:= \mathbb{P}(X \le x \,|\, (X,Y) \in A_{\mathrm{L}}) = F_{\mathrm{L}}(x, H^{-1}(p)) \\ &= \frac{C(G(x \wedge G^{-1}(p)), p)}{C(p,p)}\,, \qquad \forall\, x \in \mathbb{R}\,, \end{aligned}$$

and $H_{\mathrm{L}}(y)$ respectively. Since G_{L} and H_{L} are absolutely continuous, according to Sklar's theorem there exists also a unique copula $C_{\mathrm{L}}\colon [0,1]^2 \to [0,1]$ such that

$$F_{\mathrm{L}}(x,y) = C_{\mathrm{L}}(G_{\mathrm{L}}(x), H_{\mathrm{L}}(y))\,, \qquad \forall\, x,y \in \mathbb{R}\,.$$

Indeed, Juri and Wüthrich (2002) call

$$C_{\mathrm{L}}(u,v) = F_{\mathrm{L}}(G_{\mathrm{L}}^{-1}(u), H_{\mathrm{L}}^{-1}(v))\,, \qquad \forall\, u,v \in\, [0,1]\,,$$

the *extreme tail-dependence copula relative to C at the level p*. Instead we will call C_{L} *lower tail-copula* and the phrase 'relative to C at the level p' will be usually dropped for convenience.

Using the lower tail-copula we now can define the lower conditional rank-correlation coefficient, viz.

$$\rho_{\mathrm{L}} = 12 \int\limits_{[0,1]^2} uv\, dC_{\mathrm{L}}(u,v) - 3\,.$$

In the empirical part of this work this measures the monotone dependence of stock returns X and Y conditional on the bear market $(X,Y) \in A_{\mathrm{L}}$.

An analogue definition can be found for the *upper tail-copula* C_{U}. This is the lower tail-copula relative to the *survival copula* according to C (Nelsen, 2006, Section 2.6), i.e.

$$\overline{C}(u,v) := u + v - 1 + C(1-u, 1-v)\,, \qquad \forall\, u,v \in [0,1]\,,$$

at the level q $(0 < q < 1)$. The survival copula simply corresponds to the copula of $(-X, -Y)$ and thus C_{U} is the copula of $(-X, -Y)$ under the condition that $(-X, -Y) \in A_{\mathrm{U}}$. Here the area A_{U} is calculated similarly to A_{L} just by using the quantile functions of $-X$ and $-Y$ at q rather than the quantile functions of X and Y at p. Hence, the upper conditional rank-correlation coefficient ρ_{U} measures the monotone dependence of stock returns in a bull market. In the following we will have to guarantee that $A_{\mathrm{L}} \cap A_{\mathrm{U}} = \emptyset$ which is equivalent to $p + q \leq 1$.

In most cases it is not possible to derive the conditional copulas C_{L} or C_{U} in closed form. Therefore ρ_{L} and ρ_{U} cannot be calculated explicitly. However, MC simulation is a convenient tool for obtaining numerical approximations to ρ_{L} and ρ_{U} with sufficient precision. We apply this method to calculate the conditional rank-correlation coefficients for the Gauss-, t_3-, Clayton-, and Gumbel-copula (see Table 2.1 and Table 2.2). The Gauss- and t_3-copula are given by

$$C_{\mathrm{Gauss}}(u, v\,; \theta) = \Phi_\theta(\Phi^{-1}(u), \Phi^{-1}(v))\,, \qquad \forall\, u, v \in [0, 1]\,,$$

where

$$\Phi_\theta(x, y) := \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\theta^2}} \cdot \exp\left(-\frac{s^2 - 2\theta st + t^2}{2\,(1-\theta^2)}\right) ds\, dt$$

as well as

$$C_{t_3}(u, v\,; \theta) = t_{3,\theta}(t_3^{-1}(u), t_3^{-1}(v))\,, \qquad \forall\, u, v \in [0, 1]\,,$$

with

$$t_{3,\theta}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\theta^2}} \cdot \left(1 + \frac{s^2 - 2\theta st + t^2}{3\,(1-\theta^2)}\right)^{-\frac{5}{2}} ds\, dt\,,$$

where t_3 denotes Student's univariate t-distribution function with 3 degrees of freedom and $-1 < \theta < 1$. Note that the linear correlation coefficient is symbolized by the parameter θ rather than ρ. This is because to avoid possible confusions with the (unconditional) rank-correlation coefficient of C_{Gauss} or C_{t_3}. The unconditional rank-correlation coefficient of the Gauss-copula corresponds to $\rho = 6/\pi \cdot \arcsin(\theta/2)$ (Hult and Lindskog, 2002). To our knowledge there exists no such closed-form expression for the t_3-copula.

The Clayton-copula is given by

$$C_{\mathrm{Clayton}}(u, v\,; \theta) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}\,, \qquad \forall\, u, v \in [0, 1]\,,$$

with $\theta \geq 0$. In the limiting case $\theta = 0$ the Clayton-copula corresponds to the *independence* or *product copula* $\Pi(u, v) := uv$ (Nelsen, 2006, p. 11).

The Gumbel-copula can be written as

$$C_{\text{Gumbel}}(u, v\,; \theta) = \exp\Big[-\big\{(-\log u)^\theta + (-\log v)^\theta\big\}^{1/\theta}\Big]\,, \qquad \forall\, u, v \in [0,1]\,,$$

with $\theta \geq 1$. Note that for $\theta = 1$ once again the independence copula evolves. The values for θ in Table 2.2 are chosen such that the *unconditional* rank-correlation coefficient corresponds to $\rho = 0.3, 0.5, 0.7$. The relationship between θ and ρ can be obtained by numerical integration or MC simulation (see Joe, 1997, p. 147).

For our approximations of the conditional rank-correlation coefficients given in Table 2.1 and Table 2.2 we used $N_{\text{MC}} = 1000$ MC replications. In each one a sample from C with sample size $n = 10^6$ has been generated. Both for the simulation study and for the empirical study following later on we set $p = q$. Only the Clayton-copula allows for a closed-form representation of C_{L}. If C is a Clayton-copula the lower tail-copula C_{L} is equal to C for any $0 < p < 1$ (Juri and Wüthrich, 2002). That means ρ_{L} corresponds to the unconditional rank-correlation coefficient ρ.

The null hypothesis we are going to test can be formalized as

$$\begin{aligned} H_0 :&\quad \rho_{\text{L}} = \rho_{\text{U}} \\ \text{vs.}\quad H_1 :&\quad \rho_{\text{L}} \neq \rho_{\text{U}}\,, \end{aligned}$$

where some p and q with $p+q \leq 1$ are fixed. In the present framework H_1 implies that the monotone dependence of stock returns in bear markets is not the same as in bull markets.

Instead of a two-sided hypothesis test, a one-sided test like

$$\begin{aligned} H_0 :&\quad \rho_{\text{L}} \leq \rho_{\text{U}} \\ \text{vs.}\quad H_1 :&\quad \rho_{\text{L}} > \rho_{\text{U}} \end{aligned}$$

is of general interest, since H_1 implies that the monotone dependence is higher in bear markets than in bull markets.

The null hypothesis $H_0 \colon \rho_{\text{L}} = \rho_{\text{U}}$ stated above might be also of interest in another context. Both in theory and application of copulas it is sometimes questionable whether the random vector (X, Y) is *radially symmetric* or not (Nelsen, 2006, Section 2.7). Radial symmetry is a useful property which guarantees that $\rho_{\text{L}} = \rho_{\text{U}}$ for all $0 < p < 1$ since C and the corresponding survival copula coincide. In order to test the null hypothesis H_0' : 'The random vector (X, Y) is radially symmetric', one can apply the two-sided test and reject H_0' if H_0 is rejected.

Gauss-copula						
	$\theta = 0.25$		$\theta = 0.50$		$\theta = 0.75$	
$p = q$	lower	upper	lower	upper	lower	upper
0.05	.0404 (.0004)	.0407 (.0004)	.1109 (.0003)	.1114 (.0003)	.2622 (.0002)	.2624 (.0002)
0.20	.0601 (.0001)	.0601 (.0001)	.1595 (.0001)	.1593 (.0001)	.3485 (.0001)	.3483 (.0001)
0.35	.0775 (.0001)	.0774 (.0001)	.1972 (.0001)	.1973 (.0001)	.4090 (.0001)	.4091 (.0001)
0.50	.0962 (.0001)	.0962 (.0001)	.2354 (.0001)	.2356 (.0001)	.4655 (.0000)	.4656 (.0000)
t_3-copula						
	$\theta = 0.25$		$\theta = 0.50$		$\theta = 0.75$	
$p = q$	lower	upper	lower	upper	lower	upper
0.05	.3373 (.0003)	.3369 (.0003)	.4043 (.0002)	.4044 (.0002)	.5264 (.0002)	.5265 (.0002)
0.20	.3186 (.0001)	.3183 (.0001)	.3968 (.0001)	.3967 (.0001)	.5361 (.0001)	.5361 (.0001)
0.35	.2984 (.0001)	.2984 (.0001)	.3913 (.0001)	.3913 (.0001)	.5484 (.0001)	.5485 (.0001)
0.50	.2756 (.0001)	.2756 (.0001)	.3882 (.0001)	.3882 (.0001)	.5651 (.0000)	.5652 (.0000)

Table 2.1.: MC approximations to ρ_{L} and ρ_{U} for the Gauss- and t_3-copula possessing different values for θ. We use $N_{\mathrm{MC}} = 1000$ MC replications, each one generating a sample from the corresponding copula with sample size $n = 10^6$. The standard errors of the approximations are given in parentheses.

2.3. The Testing Procedure

In this section we describe the testing procedure. The first part requires independent and identically distributed (i.i.d.) data. It is well-known that short-term asset returns typically exhibit strong patterns of serial dependence. However, the i.i.d. assumption may serve as an appropriate starting point and there might exist several applications beyond financial data analysis where this assumption is adequate. Therefore it is worth to illustrate the testing procedure in the i.i.d. case. Afterwards we will drop the assumption of serially independent asset returns and explain how the test can be modified to account for the purpose of time series analysis.

Clayton-copula						
	$\theta = 0.51$		$\theta = 1.08$		$\theta = 2.13$	
$p = q$	lower	upper	lower	upper	lower	upper
0.05	.3004 (.0002)	.0025 (.0005)	.5001 (.0002)	.0018 (.0004)	.7002 (.0001)	.0035 (.0004)
0.20	.3003 (.0001)	.0040 (.0001)	.4999 (.0001)	.0113 (.0001)	.7000 (.0001)	.0318 (.0001)
0.35	.3001 (.0001)	.0130 (.0001)	.4999 (.0001)	.0356 (.0001)	.7000 (.0000)	.0906 (.0001)
0.50	.3001 (.0001)	.0298 (.0001)	.5000 (.0000)	.0764 (.0001)	.7000 (.0000)	.1783 (.0001)
Gumbel-copula						
	$\theta = 1.26$		$\theta = 1.54$		$\theta = 2.07$	
$p = q$	lower	upper	lower	upper	lower	upper
0.05	.0319 (.0004)	.3499 (.0002)	.0697 (.0003)	.4504 (.0002)	.1431 (.0003)	.5849 (.0001)
0.20	.0515 (.0001)	.3158 (.0001)	.1106 (.0001)	.4392 (.0001)	.2206 (.0001)	.5871 (.0001)
0.35	.0697 (.0001)	.2906 (.0001)	.1476 (.0001)	.4314 (.0001)	.2843 (.0001)	.5916 (.0000)
0.50	.0912 (.0001)	.2744 (.0001)	.1885 (.0001)	.4276 (.0001)	.3507 (.0000)	.5990 (.0000)

Table 2.2.: MC approximations to ρ_{L} and ρ_{U} for the Clayton- and Gumbel-copula possessing different values for θ. We use $N_{\mathrm{MC}} = 1000$ MC replications, each one generating a sample from the corresponding copula with sample size $n = 10^6$. The standard errors of the approximations are given in parentheses.

2.3.1. Independent and Identically Distributed Data

Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. As we do not assume that the marginal cumulative distribution functions G and H are known, we have to estimate them by

$$\widehat{G}_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}} \qquad \text{and} \qquad \widehat{H}_n(y) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq y\}}.$$

The corresponding estimate of the quantile function G^{-1} is given by

$$\widehat{G}_n^{-1}(p) = \inf\bigl\{x \colon \widehat{G}_n(x) \geq p\bigr\}$$

and $\widehat{H}_n^{-1}(p)$ respectively. For some fixed p and q with $p + q \leq 1$ we can define

$$\widehat{A}_{\mathrm{L}} := \Bigl\{(x, y) \colon x \leq \widehat{G}_n^{-1}(p), y \leq \widehat{H}_n^{-1}(p)\Bigr\}$$

and $\widehat{A}_{\mathrm{U}}$ in the same manner. We also define the sample sizes $n_{\mathrm{L}} := |\widehat{A}_{\mathrm{L}}|$ with respect to the lower left and $n_{\mathrm{U}} := |\widehat{A}_{\mathrm{U}}|$ with respect to the upper right area of the empirical copula (here $|\cdot|$ denotes the cardinality of a set). The observations in $\widehat{A}_{\mathrm{L}}$ and $\widehat{A}_{\mathrm{U}}$ can be used for estimating ρ_{L} and ρ_{U}. More precisely,

$$\hat{\rho}_{\mathrm{L},n} = \frac{12}{n_{\mathrm{L}}} \cdot \sum_{i \in \mathbf{I}_{\widehat{A}_{\mathrm{L}}}} \frac{\mathrm{r}_{\mathrm{L},n}(X_i)}{n_{\mathrm{L}}} \cdot \frac{\mathrm{r}_{\mathrm{L},n}(Y_i)}{n_{\mathrm{L}}} - 3\,,$$

where $\mathbf{I}_{\widehat{A}_{\mathrm{L}}}$ denotes the set of indices i such that $(X_i, Y_i) \in \widehat{A}_{\mathrm{L}}$. Although $\hat{\rho}_{\mathrm{L},n}$ is calculated on the basis of n_{L} data points, for notational convenience the number of observations is indicated by n rather than n_{L} (this is adequate since n_{L} depends on n).

Further, $\mathrm{r}_{\mathrm{L},n}(\cdot)$ is the rank of a marginal observation relative to all observations in $\widehat{A}_{\mathrm{L}}$. Note that $\mathrm{r}_{\mathrm{L},n}(X_i)/n_{\mathrm{L}} = \widehat{G}_{\mathrm{L},n}(X_i)$ and $\mathrm{r}_{\mathrm{L},n}(Y_i)/n_{\mathrm{L}} = \widehat{H}_{\mathrm{L},n}(Y_i)$. Here $\widehat{G}_{\mathrm{L},n}$ is the empirical counterpart of G_{L}, i.e.

$$\widehat{G}_{\mathrm{L},n}(x) = \frac{\widehat{C}_n(\widehat{G}_n(x \wedge \widehat{G}_n^{-1}(p)), p)}{\widehat{C}_n(p,p)}\,, \qquad \forall\, x \in \mathbb{R}\,,$$

where

$$\widehat{C}_n(u,v) := \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{1}_{\{\mathrm{r}_n(X_i)/n \le u\}} \mathbf{1}_{\{\mathrm{r}_n(Y_i)/n \le v\}}\,, \qquad \forall\, u, v \in [0,1]\,,$$

represents the empirical copula. Moreover, $\widehat{H}_{\mathrm{L},n}$ is defined respectively. The definition of the estimator $\hat{\rho}_{\mathrm{U},n}$ follows immediately, just by using the survival copula according to $\widehat{C}_n$ (which is determined by the observations in the upper right area $\widehat{A}_{\mathrm{U}}$).

Schmid and Schmidt (2006) have already shown that Spearman's rho is consistent and asymptotically normally distributed. The same holds for the conditional versions of Spearman's rho described above, i.e.

$$\sqrt{n_{\mathrm{L}}}\,(\hat{\rho}_{\mathrm{L},n} - \rho_{\mathrm{L}}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{\mathrm{L}}^2) \quad \text{and} \quad \sqrt{n_{\mathrm{U}}}\,(\hat{\rho}_{\mathrm{U},n} - \rho_{\mathrm{U}}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{\mathrm{U}}^2)$$

as $n_{\mathrm{L}}, n_{\mathrm{U}} \to \infty$, provided the lower left and upper right tail-copulas exist.

Proposition 2.1 *Let the distribution of (X,Y) be absolutely continuous. Suppose that the partial derivatives of the corresponding copula C exist and are continuous, too. Further, define $\Delta\hat{\rho}_n := \hat{\rho}_{\mathrm{L},n} - \hat{\rho}_{\mathrm{U},n}$ and $\Delta\rho := \rho_{\mathrm{L}} - \rho_{\mathrm{U}}$ with shortfall probabilities $p, q > 0$ such that $p + q \le 1$. If $C(p,p), \overline{C}(q,q) > 0$ then*

$$\sqrt{n}\,(\Delta\hat{\rho}_n - \Delta\rho) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\sigma_{\mathrm{L}}^2}{C(p,p)} + \frac{\sigma_{\mathrm{U}}^2}{\overline{C}(q,q)}\right), \qquad n \longrightarrow \infty\,.$$

Proof. Note that $n_{\mathrm{L}}/n \overset{\text{a.s.}}{\to} C(p,p)$ as $n \to \infty$. That means

$$\sqrt{n}\left(\hat{\rho}_{\mathrm{L},n} - \rho_{\mathrm{L}}\right) = \sqrt{\frac{n}{n_{\mathrm{L}}}}\,\sqrt{n_{\mathrm{L}}}\left(\hat{\rho}_{\mathrm{L},n} - \rho_{\mathrm{L}}\right) \overset{\text{d}}{\longrightarrow} \mathcal{N}\!\left(0, \frac{\sigma_{\mathrm{L}}^2}{C(p,p)}\right), \qquad n \longrightarrow \infty,$$

and the corresponding result holds also for $\hat{\rho}_{\mathrm{U},n}$. Since $p+q \leq 1$, the considered tails of the copula do not overlap. That means $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$ are stochastically independent. This leads immediately to the asymptotic variance given in the proposition. ■

In practical situations p and q have to be sufficiently large such that n_{L} and n_{U} do not become too small. A typical rule of thumb might be given by $n_{\mathrm{L}}, n_{\mathrm{U}} \geq 40$. Suppose for the moment that C corresponds to the product copula. In that case it is expected to meet $p^2 n$ observations in the lower left part of the empirical copula. That means the shortfall probabilities should be such that $p, q \geq \sqrt{40/n}$. E.g. for the sample size $n = 1000$ (that means an observation period of approximately 4 years) p and q should be not smaller than 0.2. Admittedly, financial data cannot be appropriately described by the product copula since in most cases there is some sort of positive dependence between stock returns. So there are even *more* observations in the corresponding corners of the empirical copula. Thus our rule of thumb guarantees that there are always enough data for large-sample inferences.

The asymptotic variances σ_{L}^2 and σ_{U}^2 depend on the tail-copulas C_{L} and C_{U}. In general they cannot be calculated explicitly (Schmid and Schmidt, 2006). The same holds for the asymptotic variance of $\Delta\hat{\rho}_n$, i.e. $\tau^2 := \sigma_{\mathrm{L}}^2/C(p,p) + \sigma_{\mathrm{U}}^2/\,\overline{C}(q,q)$. However, the latter can be approximated by a simple bootstrap. For conducting the hypothesis test one has to choose an appropriate significance level $\alpha > 0$ as well as the shortfall probabilities $p > 0$ and $q > 0$ such that $p + q \leq 1$. Now the test procedure reads as follows:

1. Compute $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$ from the observations in $\widehat{A}_{\mathrm{L}}$ and $\widehat{A}_{\mathrm{U}}$.

2. Compute N_{B} bootstrap replications from the entire sample. For each replication calculate $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$ as well as the corresponding difference $\Delta\hat{\rho}_n$.

3. Estimate the asymptotic variance τ^2 of $\Delta\hat{\rho}_n$ from the bootstrap and calculate the test statistic $T = \Delta\hat{\rho}_n/\sqrt{\hat{\tau}^2/n}$, where $\hat{\tau}^2$ is the bootstrap estimate of τ^2.

4a. Reject $H_0\colon \rho_{\mathrm{L}} = \rho_{\mathrm{U}}$ if

$$\left|T\right| \geq \Phi^{-1}\!\left(1 - \frac{\alpha}{2}\right),$$

where Φ denotes the standard normal c.d.f.

The one-sided hypothesis tests differ only in the fourth step from the two-sided test:

4b. Reject $H_0 \colon \rho_{\mathrm{L}} \leq \rho_{\mathrm{U}}$ or $H_0 \colon \rho_{\mathrm{L}} \geq \rho_{\mathrm{U}}$ if $T \geq \Phi^{-1}(1-\alpha)$ or $T \leq \Phi^{-1}(\alpha)$.

2.3.2. Serially Dependent Data

Now let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a stationary process $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$. It is supposed that the process exhibits a *weak dependence structure*, i.e. the two one-sided processes $\{(X_t, Y_t)\}_{t \leq 0}$ and $\{(X_t, Y_t)\}_{t \geq l}$ are asymptotically independent as $l \to \infty$ (Politis, 2003). This condition is sometimes referred to as the strong mixing or α-mixing condition and it can be shown that many time series models frequently used in theory and practice meet that requirement.

One of the referees pointed out that it is important to account for serial dependence, since $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$ are no longer independent in that case. For instance, in periods of great turbulence the lower and upper conditional Spearman's rho might be strongly correlated. However, it can be assumed that $\Delta\hat{\rho}_n$ remains asymptotically normally distributed under the weak dependence assumption of time series analysis. That means

$$\sqrt{n}\left(\Delta\hat{\rho}_n - \Delta\rho\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \tau^2_{\mathrm{LR}}\right), \qquad n \longrightarrow \infty, \tag{2.2}$$

where τ^2_{LR} represents the *long-run variance* of $\Delta\hat{\rho}_n$. This assumption seems natural, since the weak convergence property of Spearman's rho is based on the weak convergence of an empirical copula process (Schmid and Schmidt, 2006). By using a weak form of the central limit theorem from time series analysis one can argue that the weak convergence property is still satisfied under the strong mixing condition.

There exist many possible techniques for estimating the long-run variance τ^2_{LR} of the statistic $\Delta\hat{\rho}_n$, such as subsampling or block-bootstrapping (Politis, 2003). It has been shown by Politis et al. (2001) that the subsampling procedure leads to consistent estimates of the long-run variance under very mild regularity conditions. However, subsampling is probably not the best choice in our setting. The reason is that for getting an unbiased estimate for τ^2_{LR}, the number of observations within each subsample must be considerably small relative to the overall sample size. Note that in our context only a small part of each subsample can be used for calculating $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$ but for a proper approximation of the long-run variance it has to be guaranteed also that each subsample contains a sufficiently large number of usable observations.

Thus we will concentrate on a block-bootstrap procedure suggested by Künsch (1989). Consider a block length b with $0 < b < n$ and the corresponding $n - b + 1$ blocks, i.e.

$$B_i = \left\{(X_{i+1}, Y_{i+1}), \ldots, (X_{i+b}, Y_{i+b})\right\}, \qquad i = 0, 1, \ldots, n - b\,.$$

One bootstrap replication from the entire sample is given by drawing $k = \lfloor n/b \rfloor$ blocks with replacement and concatenating these blocks to form a new pseudo-series of asset returns. Note that the series consists of $l = kb \approx n$ pseudo-observations. For each of the N_{B} bootstrap replications one can calculate $\Delta\hat{\rho}_l$ and the long-run variance can be estimated from the given N_{B} realizations. Finally, the test statistic is given by

$$T = \frac{\Delta\hat{\rho}_n}{\sqrt{\hat{\tau}^2_{\mathrm{LR}}/n}}\,,$$

where $\hat{\tau}^2_{\mathrm{LR}}$ is the estimated long-run variance. Under the weak convergence property (2.2) and if $b \to \infty$ and $n/b \to \infty$, the estimator $\hat{\tau}^2_{\mathrm{LR}}$ is consistent for τ^2_{LR}. Hence, T can be used in the same way as the test statistic discussed in Section 2.3.1.

2.4. Finite-Sample Properties

In this section we investigate the finite-sample properties of the testing procedure described in Section 2.3.1. The results are obtained by MC simulations for various special cases. These are essentially defined by the copula under study. First we are interested in the rejection probability of the procedure if $H_0 : \rho_{\mathrm{L}} = \rho_{\mathrm{U}}$ is true and α is the prescribed error probability of the first kind. We consider the Gauss- and t_3-copula which belong to the class of elliptical copulas. Elliptical copulas are radially symmetric which means that the aforementioned null hypotheses is true. The selected values for the copula parameter are $\theta = 0.25, 0.5, 0.75$, the values for p are given by $p = 0.2, 0.35, 0.5$, and we validate the error probabilities $\alpha = 0.01, 0.05, 0.1$. The simulated sample size is $n = 2500$ (i.e. approximately 10 trading years), the number of bootstrap replications amounts to $N_{\mathrm{B}} = 1000$, and the number of MC replications is $N_{\mathrm{MC}} = 1000$. The results of the simulations are summarized in Panel 1 of Table 2.3. We can see that the approximated rejection probabilities satisfactorily agree with the prescribed error probabilities.

We are also interested in the *power* of the testing procedure, i.e. the probability of rejection if H_0 is wrong. For that purpose we consider the Clayton- and the Gumbel-copula. It is well-known that these copulas are not radially symmetric and thus in general $\rho_{\mathrm{L}} \neq \rho_{\mathrm{U}}$.

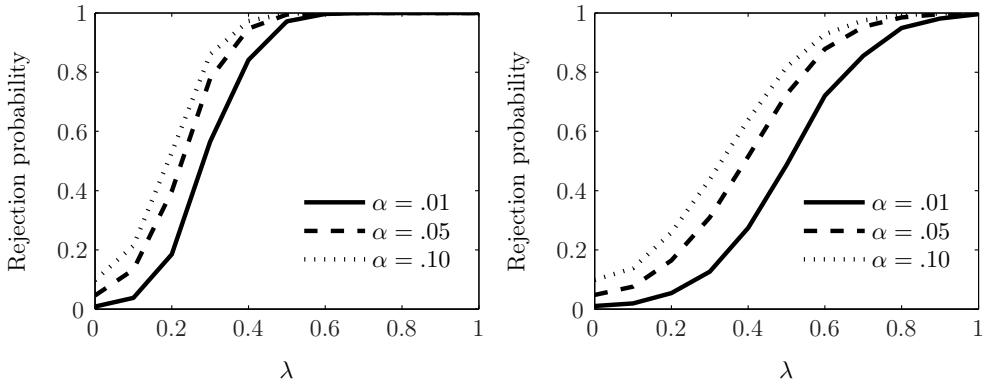

Figure 2.1.: Power functions of the two-sided hypothesis test for the mixed copulas C_{Mix1} (left hand side) and C_{Mix2} (right hand side) as a function of λ. The results are obtained by MC simulation for the sample size $n = 2500$, $N_{\text{B}} = 1000$ bootstrap replications, and $N_{\text{MC}} = 10000$ MC replications using the shortfall probability $p = q = 0.5$.

Remember that the parameter θ of both copula families (cf. p. 31) has been selected in such a way that the unconditional rank-correlation coefficients are equal to $\rho = 0.3, 0.5, 0.7$. The results of the MC simulations are given in Panel 2 of Table 2.3. It can be seen that for every fixed p and α the power is an increasing function of θ. This is because the asymmetry of the Archimedean copulas C_{Clayton} and C_{Gumbel} increases with θ (cf. Nelsen, 2006, Ch. 4). Similar results are obtained for the two one-sided tests which can be taken from Table 2.4 and Table 2.5. The rejection probabilities become very large whenever H_1 is true. In contrast, if H_0 is true our simulations produce no false rejection. For instance, consider the right-sided test $H_0 : \rho_{\text{L}} \leq \rho_{\text{U}}$ vs. $H_1 : \rho_{\text{L}} > \rho_{\text{U}}$. In that case the null hypothesis is fulfilled for the Gumbel-copula. Panel 2 of Table 2.4 shows that there is no rejection for any given unconditional rank-correlation coefficient ρ, shortfall probability p, and error probability α. In contrast, for the Clayton-copula the alternative hypothesis is true and consequently the rejection probabilities are very high (e.g. roughly 90% for $\rho = 0.3$, $p = 0.2$, and $\alpha = 0.1$). Moreover, for $\rho = 0.5$ and $\rho = 0.7$, H_0 is rejected for the Clayton-copula in almost every simulated case.

Now we want to investigate the relationship between asymmetry and power. For that purpose we consider the mixed copula

$$C_{\text{Mix1}}(u, v\,; \lambda, \theta_0, \theta_1) := \lambda C_{\text{Clayton}}(u, v\,; \theta_1) + (1 - \lambda) C_{\text{Gauss}}(u, v\,; \theta_0)\,,$$

where $0 \leq \lambda \leq 1$. Further, the copula parameters θ_0, θ_1 are such that the unconditional rank-correlation coefficients of $C_{\text{Clayton}}(u, v\,; \theta_1)$ and $C_{\text{Gauss}}(u, v\,; \theta_0)$ correspond to $\rho = 0.5$.

$H_0 : \rho_L = \rho_U$ vs. $H_1 : \rho_L \neq \rho_U$							
Panel 1		$\theta = 0.25$		$\theta = 0.50$		$\theta = 0.75$	
$p = q$	α	Gauss	t_3	Gauss	t_3	Gauss	t_3
0.20	0.10	.091 (.0091)	.083 (.0087)	.081 (.0086)	.085 (.0088)	.091 (.0091)	.093 (.0092)
	0.05	.043 (.0064)	.047 (.0067)	.039 (.0061)	.041 (.0063)	.048 (.0068)	.048 (.0068)
	0.01	.008 (.0028)	.011 (.0033)	.006 (.0024)	.007 (.0026)	.011 (.0033)	.011 (.0033)
0.35	0.10	.106 (.0097)	.081 (.0086)	.092 (.0091)	.108 (.0098)	.095 (.0093)	.087 (.0089)
	0.05	.057 (.0073)	.038 (.0060)	.049 (.0068)	.053 (.0071)	.048 (.0068)	.051 (.0070)
	0.01	.015 (.0038)	.009 (.0030)	.011 (.0033)	.007 (.0026)	.006 (.0024)	.013 (.0036)
0.50	0.10	.104 (.0097)	.088 (.0090)	.088 (.0090)	.113 (.0100)	.089 (.0090)	.098 (.0094)
	0.05	.060 (.0075)	.043 (.0064)	.035 (.0058)	.056 (.0073)	.048 (.0068)	.049 (.0068)
	0.01	.019 (.0043)	.011 (.0033)	.008 (.0028)	.008 (.0028)	.006 (.0024)	.011 (.0033)
Panel 2		$\rho = 0.30$		$\rho = 0.50$		$\rho = 0.70$	
$p = q$	α	Clayton	Gumbel	Clayton	Gumbel	Clayton	Gumbel
0.20	0.10	.815 (.0123)	.752 (.0137)	1.000 (.0000)	.965 (.0058)	1.000 (.0000)	.999 (.0010)
	0.05	.715 (.0143)	.635 (.0152)	.999 (.0010)	.938 (.0076)	1.000 (.0000)	.996 (.0020)
	0.01	.456 (.0158)	.371 (.0153)	.993 (.0026)	.800 (.0126)	1.000 (.0000)	.999 (.0010)
0.35	0.10	.988 (.0034)	.926 (.0083)	1.000 (.0000)	.999 (.0010)	1.000 (.0000)	1.000 (.0000)
	0.05	.981 (.0043)	.876 (.0104)	1.000 (.0000)	.995 (.0022)	1.000 (.0000)	1.000 (.0000)
	0.01	.928 (.0082)	.704 (.0144)	1.000 (.0000)	.980 (.0044)	1.000 (.0000)	1.000 (.0000)
0.50	0.10	.999 (.0010)	.974 (.0050)	1.000 (.0000)	1.000 (.0000)	1.000 (.0000)	1.000 (.0000)
	0.05	.999 (.0010)	.945 (.0072)	1.000 (.0000)	1.000 (.0000)	1.000 (.0000)	1.000 (.0000)
	0.01	.996 (.0020)	.837 (.0117)	1.000 (.0000)	.995 (.0022)	1.000 (.0000)	1.000 (.0000)

Table 2.3.: MC approximations of the rejection probabilities for the Gauss- and t_3-copula (Panel 1) and for the Clayton- and Gumbel-copula (Panel 2) given $H_0 : \rho_L = \rho_U$. The simulated sample size is $n = 2500$, the number of bootstrap replications corresponds to $N_B = 1000$, and the number of MC replications is $N_{MC} = 1000$. The standard errors for the approximated rejection probabilities are given in parentheses.

$H_0 \colon \rho_{\mathrm{L}} \leq \rho_{\mathrm{U}}$ vs. $H_1 \colon \rho_{\mathrm{L}} > \rho_{\mathrm{U}}$							
Panel 1		$\theta = 0.25$		$\theta = 0.50$		$\theta = 0.75$	
$p = q$	α	Gauss	t_3	Gauss	t_3	Gauss	t_3
0.20	0.10	.091 (.0091)	.096 (.0093)	.099 (.0094)	.095 (.0093)	.096 (.0093)	.099 (.0094)
	0.05	.047 (.0067)	.041 (.0063)	.040 (.0062)	.041 (.0063)	.047 (.0067)	.049 (.0068)
	0.01	.009 (.0030)	.007 (.0026)	.006 (.0024)	.007 (.0026)	.013 (.0036)	.009 (.0030)
0.35	0.10	.103 (.0096)	.093 (.0092)	.086 (.0089)	.099 (.0094)	.097 (.0094)	.094 (.0092)
	0.05	.053 (.0071)	.049 (.0068)	.038 (.0060)	.044 (.0065)	.047 (.0067)	.044 (.0065)
	0.01	.012 (.0034)	.010 (.0031)	.008 (.0028)	.006 (.0024)	.008 (.0028)	.011 (.0033)
0.50	0.10	.109 (.0099)	.092 (.0091)	.103 (.0096)	.110 (.0099)	.086 (.0860)	.094 (.0092)
	0.05	.050 (.0069)	.046 (.0066)	.046 (.0066)	.053 (.0071)	.049 (.0068)	.054 (.0071)
	0.01	.015 (.0038)	.011 (.0033)	.011 (.0033)	.011 (.0033)	.010 (.0031)	.008 (.0028)
Panel 2		$\rho = 0.30$		$\rho = 0.50$		$\rho = 0.70$	
$p = q$	α	Clayton	Gumbel	Clayton	Gumbel	Clayton	Gumbel
0.20	0.10	.899 (.0095)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
	0.05	.815 (.0123)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
	0.01	.574 (.0156)	.000 (.0000)	.997 (.0017)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
0.35	0.10	.997 (.0017)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
	0.05	.988 (.0034)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
	0.01	.961 (.0061)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
0.50	0.10	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
	0.05	.999 (.0010)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)
	0.01	.999 (.0010)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)	.000 (.0000)

Table 2.4.: MC approximations of the rejection probabilities for the Gauss- and t_3-copula (Panel 1) and for the Clayton- and Gumbel-copula (Panel 2) given $H_0 \colon \rho_{\mathrm{L}} \leq \rho_{\mathrm{U}}$. The simulated sample size is $n = 2500$, the number of bootstrap replications corresponds to $N_{\mathrm{B}} = 1000$, and the number of MC replications is $N_{\mathrm{MC}} = 1000$. The standard errors for the approximated rejection probabilities are given in parentheses.

$H_0 : \rho_L \geq \rho_U$ vs. $H_1 : \rho_L < \rho_U$							
Panel 1		$\theta = 0.25$		$\theta = 0.50$		$\theta = 0.75$	
$p = q$	α	Gauss	t_3	Gauss	t_3	Gauss	t_3
0.20	0.10	.093 (.0092)	.097 (.0094)	.089 (.0090)	.090 (.0090)	.096 (.0093)	.096 (.0093)
	0.05	.044 (.0065)	.042 (.0063)	.041 (.0063)	.044 (.0065)	.044 (.0065)	.044 (.0065)
	0.01	.009 (.0030)	.011 (.0033)	.007 (.0026)	.012 (.0034)	.006 (.0024)	.013 (.0036)
0.35	0.10	.108 (.0098)	.083 (.0087)	.100 (.0095)	.111 (.0099)	.096 (.0093)	.087 (.0089)
	0.05	.053 (.0071)	.032 (.0056)	.054 (.0071)	.064 (.0077)	.048 (.0068)	.043 (.0064)
	0.01	.013 (.0036)	.007 (.0026)	.013 (.0036)	.012 (.0034)	.005 (.0022)	.013 (.0036)
0.50	0.10	.095 (.0093)	.097 (.0094)	.103 (.0096)	.094 (.0092)	.087 (.0089)	.088 (.0090)
	0.05	.054 (.0071)	.042 (.0063)	.042 (.0063)	.060 (.0075)	.040 (.0062)	.044 (.0065)
	0.01	.016 (.0040)	.007 (.0026)	.004 (.0020)	.009 (.0030)	.003 (.0017)	.015 (.0038)
Panel 2		$\rho = 0.30$		$\rho = 0.50$		$\rho = 0.70$	
$p = q$	α	Clayton	Gumbel	Clayton	Gumbel	Clayton	Gumbel
0.20	0.10	.000 (.0000)	.856 (.0111)	.000 (.0000)	.986 (.0037)	.000 (.0000)	1.000 (.0000)
	0.05	.000 (.0000)	.752 (.0137)	.000 (.0000)	.965 (.0058)	.000 (.0000)	.999 (.0010)
	0.01	.000 (.0000)	.481 (.0158)	.000 (.0000)	.870 (.0106)	.000 (.0000)	.994 (.0024)
0.35	0.10	.000 (.0000)	.964 (.0059)	.000 (.0000)	.999 (.0010)	.000 (.0000)	1.000 (.0000)
	0.05	.000 (.0000)	.926 (.0083)	.000 (.0000)	.999 (.0010)	.000 (.0000)	1.000 (.0000)
	0.01	.000 (.0000)	.790 (.0129)	.000 (.0000)	.987 (.0036)	.000 (.0000)	1.000 (.0000)
0.50	0.10	.000 (.0000)	.991 (.0030)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)
	0.05	.000 (.0000)	.974 (.0050)	.000 (.0000)	1.000 (.0000)	.000 (.0000)	1.000 (.0000)
	0.01	.000 (.0000)	.892 (.0098)	.000 (.0000)	.996 (.0020)	.000 (.0000)	1.000 (.0000)

Table 2.5.: MC approximations of the rejection probabilities for the Gauss- and t_3-copula (Panel 1) and for the Clayton- and Gumbel-copula (Panel 2) given $H_0 : \rho_L \geq \rho_U$. The simulated sample size is $n = 2500$, the number of bootstrap replications corresponds to $N_B = 1000$, and the number of MC replications is $N_{MC} = 1000$. The standard errors for the approximated rejection probabilities are given in parentheses.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
$\overline{\hat{\rho}_{\mathrm{L},n}}$	39%	34%	34%	34%	35%
$\overline{\hat{\rho}_{\mathrm{U},n}}$	34%	27%	25%	24%	24%
$\overline{\Delta\hat{\rho}_n}$	5%	7%	9%	10%	11%
$\overline{\lvert\Delta\hat{\rho}_n\rvert}$	12%	10%	10%	11%	11%

Table 2.6.: Average conditional Spearman's rhos, differences, and absolute differences of all 435 asset combinations for different shortfall probabilities $p = q$.

Hence, the mixed copula possesses the same unconditional rank-correlation coefficient for every λ (see the formula for ρ on p. 29).

Note that $\rho_{\mathrm{L}} = \rho_{\mathrm{U}}$ is true for the Gauss-copula but for the Clayton-copula it holds that $\rho_{\mathrm{L}} > \rho_{\mathrm{U}}$ and so the mixing parameter λ determines the degree of asymmetry given by $C_{\mathrm{Mix1}}(u, v\,; \lambda, \theta_0, \theta_1)$. If one considers the two-sided hypothesis test with $H_0 : \rho_{\mathrm{L}} \leq \rho_{\mathrm{U}}$, $\lambda = 0$ means that the null hypothesis is true whereas the alternative hypothesis holds for every $\lambda > 0$. The larger λ the more often H_0 should be rejected.

A similar result is obtained for the mixed copula

$$C_{\mathrm{Mix2}}(u, v\,; \lambda, \theta_0, \theta_2) := \lambda C_{\mathrm{Gumbel}}(u, v\,; \theta_2) + (1 - \lambda) C_{\mathrm{Gauss}}(u, v\,; \theta_0)\,,$$

where θ_2 is such that the rank-correlation coefficient associated with $C_{\mathrm{Gumbel}}(u, v\,; \theta_2)$ once again amounts to $\rho = 0.5$. The corresponding power functions are given in Figure 2.1. Both figures demonstrate that the hypothesis test always keeps the prescribed error probability of the first kind and the rejection probability indeed is an increasing function of the mixing parameter λ. Similar results can be obtained for other constellations of ρ and p.

2.5. Empirical Results for German Stock Returns

Now we consider daily observations from 1973-01-02 to 2008-11-14 of the 30 stocks which are listed in the German stock index DAX 30. The stock prices have been adjusted for dividends, splits, etc. Our analyze is based on the daily log-returns of the assets (zero returns have been deleted). The maximum number of observations is given by $n = 9359$ trading days. Table 2.6 contains the sample means of the upper and lower conditional Spearman's rhos for all 435 asset combinations given the shortfall probabilities $p = q =$

$0.1, 0.2, 0.3, 0.4, 0.5$. Here $\overline{\hat{\rho}_{\mathrm{L},n}}$ denotes the mean lower and $\overline{\hat{\rho}_{\mathrm{U},n}}$ the mean upper conditional Spearman's rho, whereas $\overline{\Delta\hat{\rho}_n}$ is the mean difference and $\overline{|\Delta\hat{\rho}_n|}$ the mean absolute difference between $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$. It can be seen that the lower conditional Spearman's rhos are up to 11 points larger in average than the upper conditional Spearman's rhos. However, without a meaningful theoretical argument it is not possible to judge whether this gap between bull and bear markets is rather 'large' or 'small' and we would like to avoid such kind of statements. Instead we will discuss how much of the empirical evidence leads to significant results in our hypothesis tests.

It is worth to point out that the outcomes of the test can depend substantially on the shortfall probability p. The upper part of Figure 2.2 shows the lower and upper conditional Spearman's rho as a function of p for BASF vs. Henkel. The difference between the rhos (see the lower part of Figure 2.2) seems to be negligible if $p \leq 0.25$ but it can be very large for $p > 0.25$. The lower right part of Figure 2.2 indicates that it is easy to find a suitable p such that $H_0 \colon \rho_{\mathrm{L}} \leq \rho_{\mathrm{U}}$ can be rejected on a significance level of $\alpha = 0.05$, although in fact there are not many statistical arguments in favor of $H_1 : \rho_{\mathrm{L}} > \rho_{\mathrm{U}}$. As a counterexample consider Figure 2.3 giving the conditional Spearman's rhos of BASF vs. Thyssen. There is only a small range for p where H_0 *cannot* be rejected. That means there is a large amount of evidence for supporting H_1 but this could be easily concealed by exploiting the data. Finally, Figure 2.4 contains the conditional Spearman's rhos for BASF vs. Infineon. Only in that case data mining is impossible since there is no p for which the null hypothesis could be rejected. We conclude that the presented hypothesis tests work only if p is chosen *before* examining $\hat{\rho}_{\mathrm{L},n}$ and $\hat{\rho}_{\mathrm{U},n}$ with different shortfall probabilities. Otherwise the tests would seriously suffer from a selection bias.

2.5.1. Two-Sided Hypothesis Test

It is clear that the estimates $\hat{\rho}_{\mathrm{L}}$ and $\hat{\rho}_{\mathrm{U}}$ are different from each other for every combination of assets and we want to see whether the differences are significant. That means we test $H_0 : \rho_{\mathrm{L}} = \rho_{\mathrm{U}}$ against $H_1 : \rho_{\mathrm{L}} \neq \rho_{\mathrm{U}}$ by using the block-bootstrap procedure described in Section 2.3.2. The block length corresponds to $b = 40$ (i.e. approximately 2 trading months) and the number of bootstrap replications is $N_{\mathrm{B}} = 1000$. After computing the first estimate $\hat{\tau}^2_{\mathrm{LR},b}$, a second run with block length $b/2 = 20$ is made. This leads to the second estimate $\hat{\tau}^2_{\mathrm{LR},b/2}$ for the long-run variance and the linear combination

$$\hat{\tau}^2_{\mathrm{LR}} = 2\hat{\tau}^2_{\mathrm{LR},b} - \hat{\tau}^2_{\mathrm{LR},b/2}$$

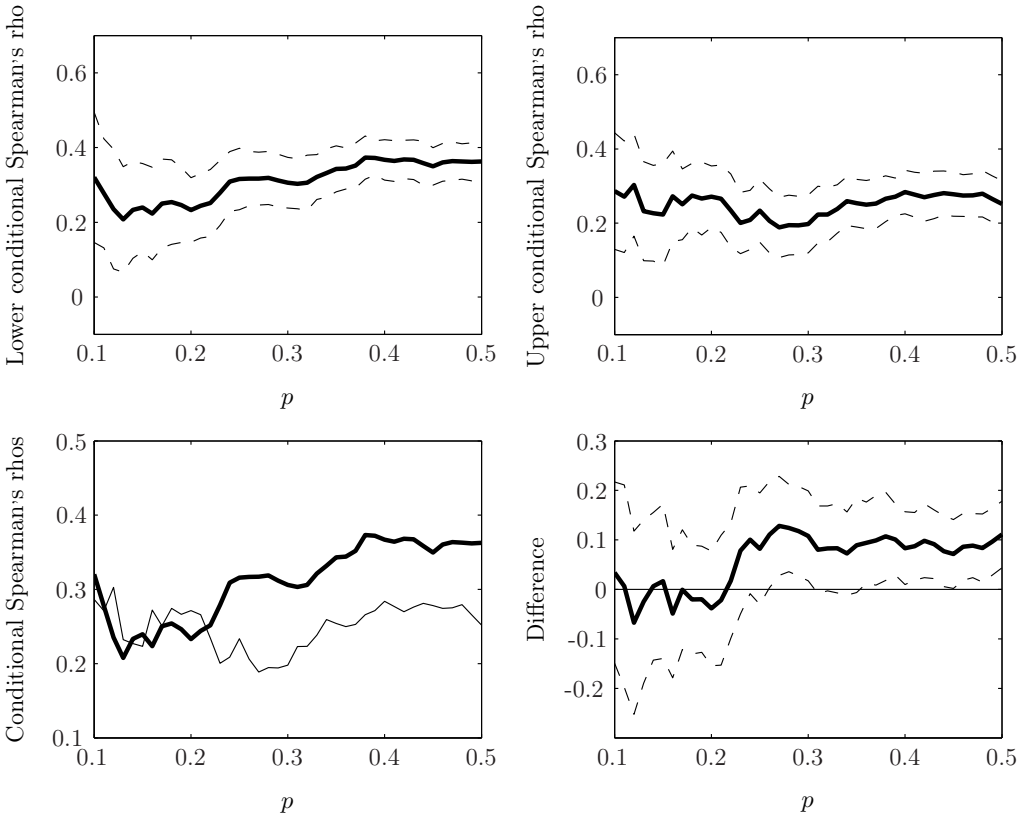

Figure 2.2.: The upper part shows the lower (left hand side) and upper (right hand side) conditional Spearman's rho as a function of $p = q$ for BASF vs. Henkel ($n = 5343$). The dashed lines represent the corresponding 95%-confidence bands. In the lower left part the lower and upper conditional Spearman's rhos are shown together where the thick line represents the lower conditional Spearman's rho. The lower right part contains the difference between the rhos and the corresponding 95%-confidence band.

is chosen as an estimate for τ^2_{LR}. Such a linear combination typically leads to more accurate estimates of the long-run variance than taking a single estimate (Politis, 2003).

Each rejection of a hypothesis test can be interpreted as an outcome of a Bernoulli experiment with parameter value $0 < \pi_i < 1$. The considered test statistics indeed depend on each other but nevertheless an unbiased estimate of the *rejection rate* $\bar{\pi} := 1/435 \sum_{i=1}^{435} \pi_i$ is given by the proportion of rejections. Since the considered tests are unbiased (see Figure 2.1), H_1 is said to be 'true in general' if $\bar{\pi} \gg \alpha$. Note that $\min_i \pi_i \leq \bar{\pi}$, i.e. the 'worst' of the 435 asset combinations cannot produce a power which is larger than $\bar{\pi}$ and so the rejection rate may serve as an upper bound for the most optimistic view in favor of H_1. Conversely, given a *one-sided* hypothesis test, $\bar{\pi} \ll \alpha$ implies that H_0 is 'true in general', since the considered hypothesis tests are not conservative but strictly decreasing in H_1 (see

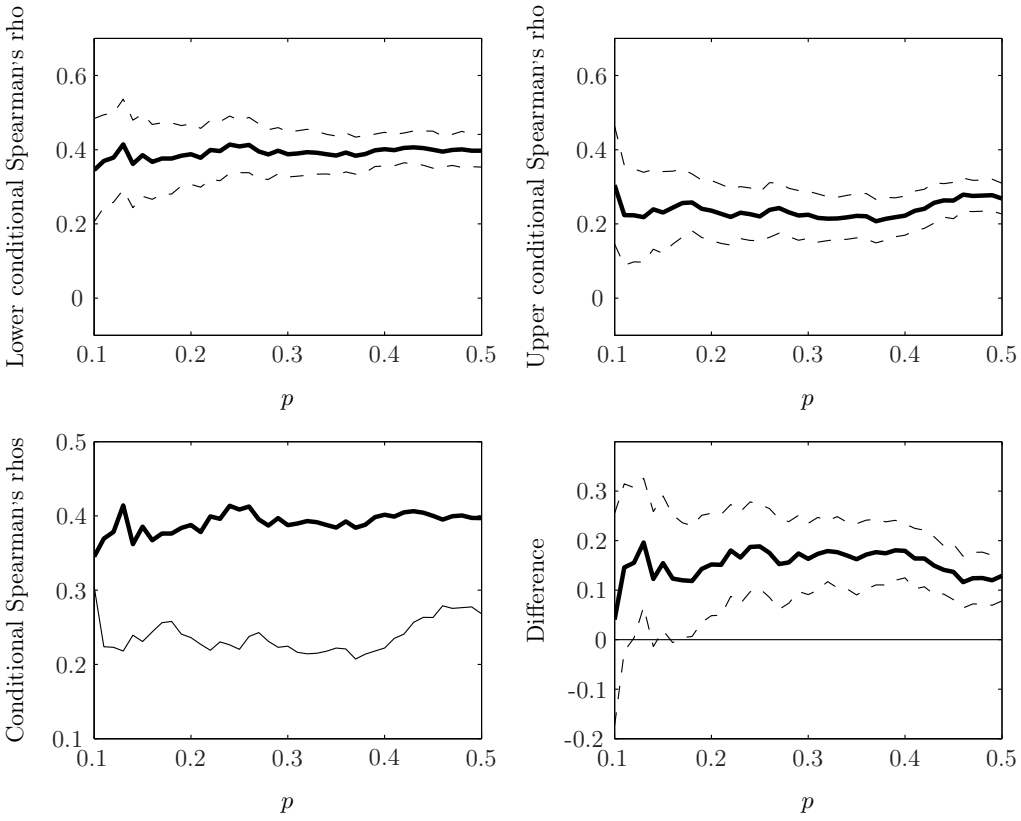

Figure 2.3.: The upper part shows the lower (left hand side) and upper (right hand side) conditional Spearman's rho as a function of $p = q$ for BASF vs. Thyssen ($n = 7884$). The dashed lines represent the corresponding 95%-confidence bands. In the lower left part the lower and upper conditional Spearman's rhos are shown together where the thick line represents the lower conditional Spearman's rho. The lower right part contains the difference between the rhos and the corresponding 95%-confidence band.

Table 2.4 and Table 2.5).

The first panel of Table 2.7 contains the proportions of rejections for all 435 asset combinations. For the shortfall probability $p = 0.1$ only 10% asset combinations exhibit significantly different Spearman's rhos on a significance level of $\alpha = 0.1$. However, it can be seen that for all $p \geq 0.2$ the proportions of rejections exceed the corresponding significance levels. Especially, if p increases the rejection rates apparently become very large and so we conclude that the lower and upper conditional rank-correlation coefficients in general are different from each other.

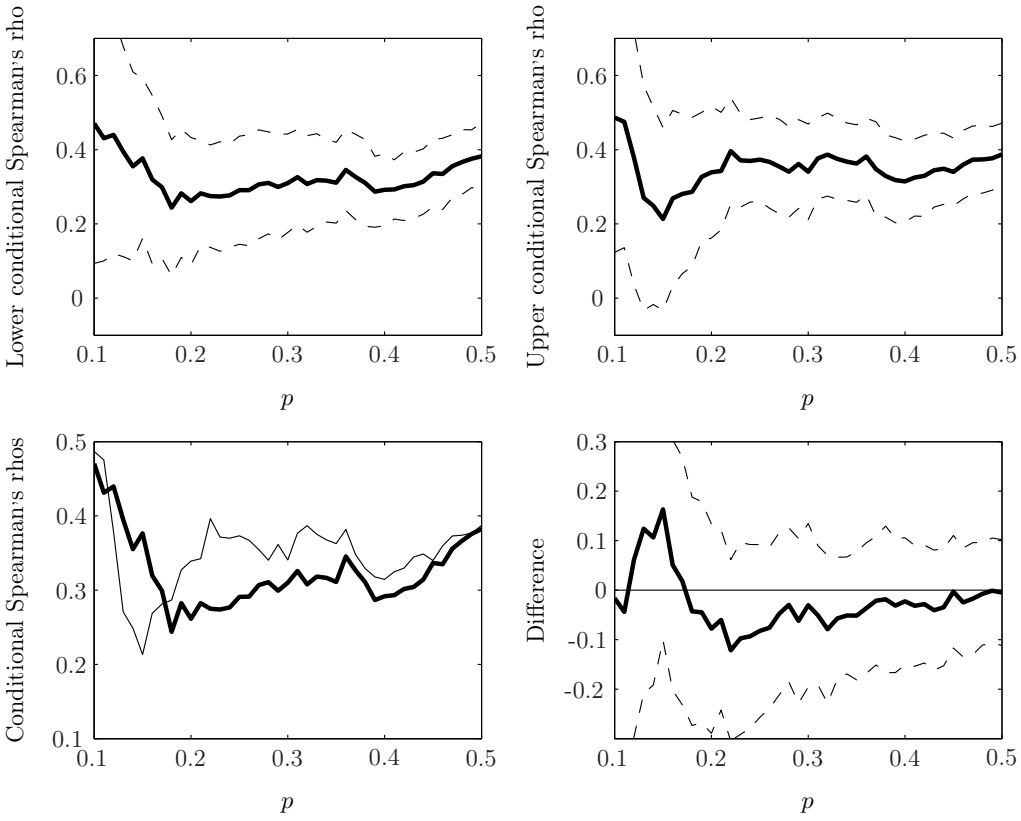

Figure 2.4.: The upper part shows the lower (left hand side) and upper (right hand side) conditional Spearman's rho as a function of $p = q$ for BASF vs. Infineon ($n = 2110$). The dashed lines represent the corresponding 95%-confidence bands. In the lower left part the lower and upper conditional Spearman's rhos are shown together where the thick line represents the lower conditional Spearman's rho. The lower right part contains the difference between the rhos and the corresponding 95%-confidence band.

2.5.2. One-Sided Hypothesis Tests

Panel 2 and 3 of Table 2.7 contain the proportions of asset combinations where the lower conditional Spearman's rho exceeds the upper conditional Spearman's rho and vice versa. For example, 67% of the asset combinations are such that $\hat{\rho}_{\mathrm{L},n} > \hat{\rho}_{\mathrm{U},n}$ given the shortfall probability $p = 0.1$ but only 17% of these combinations are significant on the significance level $\alpha = 0.1$. It is clear that not every combination with $\hat{\rho}_{\mathrm{L},n} > \hat{\rho}_{\mathrm{U},n}$ or $\hat{\rho}_{\mathrm{L},n} < \hat{\rho}_{\mathrm{U},n}$ can be significant. This holds especially if the number of observations in the lower left and upper right area of the empirical copula is small. So even if the proportion of significant combinations might appear to be somewhat small, it neither implies that most of the null hypotheses are true nor that the differences of the lower and upper conditional rank-correlation coefficients are 'small' (see the last row of Table 2.6).

Panel 1	$H_0\colon \rho_L = \rho_U$ vs. $H_1\colon \rho_L \neq \rho_U$				
α	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
0.10	.10	.27	.46	.63	.78
0.05	.05	.20	.37	.53	.69
0.01	.01	.09	.24	.36	.49
Panel 2	$H_0\colon \rho_L \leq \rho_U$ vs. $H_1\colon \rho_L > \rho_U$				
α	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
0.10	.17	.37	.57	.73	.84
0.05	.09	.25	.45	.62	.78
0.01	.02	.12	.29	.42	.57
$\hat{\rho}_{L,n} > \hat{\rho}_{U,n}$	.67	.79	.92	.95	.97
Panel 3	$H_0\colon \rho_L \geq \rho_U$ vs. $H_1\colon \rho_L < \rho_U$				
α	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
0.10	.03	.03	.01	.01	.00
0.05	.01	.02	.00	.01	.00
0.01	.00	.01	.00	.00	.00
$\hat{\rho}_{L,n} < \hat{\rho}_{U,n}$	.33	.21	.08	.05	.03

Table 2.7.: Proportions of rejections of the different hypothesis tests, shortfall probabilities, and significance levels for the 435 asset combinations. The proportions of asset combinations where $\hat{\rho}_L$ is larger or smaller than $\hat{\rho}_U$ are presented in the last rows of Panel 2 and Panel 3.

The second panel of Table 2.7 clearly reveals that the rejection rates of the hypothesis tests for various levels of p and α considerably exceed the corresponding significance levels. This effect becomes more obvious the more p increases. Hence, we have found a strong evidence for the hypothesis $H_1\colon \rho_L > \rho_U$. In contrast, for the opposite test the proportions of rejections given in Table 2.7 (Panel 3) are substantially smaller than the significance levels. Once again this clearly supports the aforementioned hypothesis.

Many empirical studies suggest that the *linear correlation coefficient* of stock returns is larger in bear markets than in bull markets (see, e.g., Ang and Chen, 2002, Erb et al., 1994). Our results of the one-sided hypothesis tests confirm these findings in the literature, where Pearson's rho is used as a dependence measure. That means in bear markets daily stock

returns depend more on each other than in bull markets. This holds even if 'dependence' is measured by Spearman's rho, which is neither susceptible to outliers nor affected by the marginal distributions of the considered random variables.

2.6. Conclusion

Several authors have investigated the dependencies of stock returns in bull and bear markets. Pearson's rho has been typically used as a canonical dependence measure. Unfortunately, it essentially depends on the marginal cumulative distribution functions of the random variables which are taken into consideration and quantifies only the degree of linear dependence. However, one is often interested in the degree of monotone rather than linear dependence. This holds especially if the marginal distributions are highly non-standard which is definitely the case when concentrating on the tails of stock return distributions. So it is crucial to find a reasonable dependence measure for the degree of monotone dependence under the condition that stock returns contemporaneously go up or down. We believe that copula theory can serve as an appropriate tool-box and suggest Spearman's rho as a dependence measure. This is in contrast to the previous literature, where e.g. conditional versions of Pearson's rho have been used for the same purpose. Moreover, our approach is purely nonparametric. Since we do not fit specific copulas to the data or suggest specific time series models, we are able to avoid any kind of model misspecification. The finite-sample performance of the proposed hypothesis tests have been demonstrated by Monte Carlo simulation. Further, an empirical study using daily returns of stocks contained in the DAX 30 has been conducted. We think that there is sufficient evidence to support the hypothesis of different degrees of monotone dependence in bull and bear markets.

Chapter 3.

A General Approach to Bayesian Portfolio Optimization

3.1. Motivation

Traditional portfolio optimization strategies are susceptible to parameter uncertainty (Jorion, 1986, Kalymon, 1971, Klein and Bawa, 1976, Markowitz, 1952, Michaud, 1989). Estimation risk is mainly driven by the uncertainty regarding the expected asset returns rather than their variances and covariances (Chopra and Ziemba, 1993). However, it can be shown that estimating the covariance matrix is also problematic if the sample size is small compared to the number of assets (Frahm, 2008, Kempf and Memmel, 2006). Many portfolio optimization approaches rely on rather simple assumptions about the distribution of asset returns. However, it is well-known that short-term financial data can be heavy-tailed or at least leptokurtic, tail-dependent, skewed or possessing other kinds of asymmetries. Financial time series typically exhibit volatility clusters or even long-memory which holds especially if log-price changes (so-called *log-returns*) of stocks, stock indices, and foreign exchange rates are considered. Moreover, high-frequency data generally are non-stationary, have jumps, and are strongly dependent.

One might argue that the stylized facts do not matter for long investment horizons since Gordin's central limit theorem (Hayashi, 2000, p. 404) takes effect even for ergodic stationary processes. For example, many applications in finance rely on the normal distribution assumption and so low-frequency data are used to estimate the expected values of long-term, such as monthly or quarterly, asset returns. Indeed, Merton (1980) showed that the estimation of expected returns generally cannot be improved by increasing the sampling

frequency. However, decreasing the sampling frequency leads to a loss of statistical efficiency since relevant information about the variances and covariances of asset returns get lost. Today's availability of high-frequency data offers new opportunities for statistical analysis, since these data include much more information than samples of low-frequency data. Nevertheless, by using high-frequency data and *ignoring* the stylized facts of empirical finance we would also obtain inaccurate estimates of the optimal portfolio weights. That means when working with high-frequency-data we need an appropriate model which accounts for the specific characteristics of the data generating process. The principal goal of this paper is to present a general approach which takes account of both estimation risks and stylized facts. Such kind of approach nowadays is feasible due to the permanent rise of computational power, especially the facilities of high-performance computing.

In order to incorporate estimation risk we rely on the Bayesian framework. This will be described in detail in Section 3.2. The Bayesian framework has several advantages. First of all we are able to make *finite-sample* inferences. This is important even for a large number of observations since the effective sample size strongly depends on the number of observations relative to the number of assets (Frahm and Jaekel, 2007b). Further, Bayesian analysis allows us to consider not only historical data but also to incorporate prior information such as expert knowledge. This can lead to more reasonable and well-diversified portfolios rather than relying on pure statistical portfolio optimization methods (Black and Litterman, 1992, Herold and Maurer, 2006, Scherer and Martin, 2007, Ch. 7). The dynamics of high-frequency data might become very complicated so that traditional estimation procedures such as maximum-likelihood estimation quickly hit the wall. In contrast, by using contemporary methods of numerical integration such as Markov chain Monte Carlo or importance sampling, calculating the Bayesian posterior distribution of some parameter is possible even for very complicated time series models (Geweke, 1989, 1995).

For the purpose of portfolio optimization we are interested in the *predictive distribution* of asset returns. The predictive distribution combines both estimation risk and market risk. Many Bayesian approaches to portfolio optimization are based on a purely analytical fundament (Garlappi et al., 2007, Jorion, 1986, Klein and Bawa, 1976, Polson and Tew, 2000, Meucci, 2005, Ch. 7). However, this is not suitable if we want to take stylized facts into account and then generally it is not possible to find the predictive distribution analytically. To avoid limitations of such kind, we suggest a Metropolis-Hastings-like

algorithm for simulating the posterior distribution of the unknown parameters. This is derived on the basis of empirical information obtained from time series data and prior information possibly given by an expert. The Markov chain Monte Carlo method belongs to the broad class of tempering algorithms which have been frequently used in natural sciences and proven to be able to simulate high-order distributions. It is therefore natural to apply them to high-order financial problems like portfolio optimization. By choosing a numerical framework, principally we can use almost any probabilistic model for the data and parameters. In Section 3.4 we will present a realistic portfolio optimization problem which has been performed on a standard PC in reasonable time.

3.2. The General Approach

3.2.1. Portfolio Optimization Problem

In the following we consider the *discrete predictive returns* of several assets after some long investment horizon. We specifically concentrate on discrete or, say, simple returns instead of log-returns for two reasons:

(1) Traditional portfolio theory is based on and can work only with discrete returns rather than, e.g., log-returns.

(2) Moreover, discrete returns usually differ substantially from log-returns if the investment horizon is long.

The latter is often neglected in literature. Moreover, we concentrate on long investment horizons since in practice investors usually do not want to liquidate or re-balance a portfolio each day or week. In contrast, we can think of, e.g., quarterly or yearly investment periods. The meaning of 'predictive' asset returns is to be understood in the Bayesian sense and will be explained later on in more detail. Roughly speaking, the distribution of predictive asset returns do not only account for market risk but also for the parameter uncertainty which is always present if the parameters of some model for the asset returns are unknown.

Let $R = (R_1, \ldots, R_d)$ be a d-dimensional vector of discrete predictive asset returns, $\mu = \mathrm{E}(R)$ the $d \times 1$ vector of predictive expected returns and $\Sigma = \mathrm{Var}(R) < \infty$ the corresponding $d \times d$ matrix of predictive variances and covariances. We are searching for

$$w = \arg\max_{v} \varphi(v'\mu, v'\Sigma\, v)\,, \qquad \text{s.t. } v \in \mathcal{C} \subset \mathbb{R}^d\,, \tag{3.1}$$

where v represents a portfolio, i.e. a vector of asset weights and φ is an appropriate objective function (i.e. φ is strongly increasing in the first and decreasing in the second argument) such as the well-known mean-variance certainty equivalent

$$\varphi(v'\mu, v'\Sigma\, v) = v'\mu - \frac{\alpha}{2} \cdot v'\Sigma\, v \tag{3.2}$$

with $\alpha \geq 0$. Note that $v'\mu$ represents the expectation and $v'\Sigma\, v$ is the variance of the predictive portfolio return of a buy-and-hold portfolio after the given investment period. The principal goal of this work is to show how the predictive moments μ and Σ (which incorporate both market and estimation risk) can be calculated if short-term asset log-returns are not normally distributed, possibly serially dependent, or exhibit other kinds of stylized facts (see below).

3.2.2. Gordin's Central Limit Theorem

Now let $(X_t \,|\, \theta)$ $(t \in \mathbb{Z})$ be a strongly stationary process representing the short-term *log-returns* of some asset with $\mathrm{E}(X_t \,|\, \theta) = \eta(\theta)$. Note that here we consider a stochastic process under some unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$. We assume also that $(X_t \,|\, \theta)$ is ergodic. Ergodicity means that any existing and finite moment of $X_t \,|\, \theta$ can be consistently estimated by using the corresponding sample moment of the time series $X_1, \ldots, X_n$ $(n \to \infty)$. This is guaranteed if $(X_t, \ldots, X_{t+k} \,|\, \theta)$ is asymptotically independent of $(X_{t-n}, \ldots, X_{t-n+l} \,|\, \theta)$ as $n \to \infty$ for all $k, l \in \mathbb{N}$ (Hayashi, 2000, p. 101). Further, we suppose that the second moments of $X_t \,|\, \theta$ exist and are finite.

However, for the central limit theorem (CLT) we need some additional assumption. More precisely, the CLT holds for the sample mean of $(X_t \,|\, \theta)$ if the centered process $(X_t - \eta(\theta) \,|\, \theta)$ satisfies Gordin's condition. Let $\mathcal{H}_t := (X_t, X_{t-1}, \ldots \,|\, \theta)$ be the history of $(X_t \,|\, \theta)$ at time $t \in \mathbb{Z}$. Roughly speaking, Gordin's condition implies that the impact of $\mathcal{H}_{t-n}$ on the conditional expectation of $X_t \,|\, \theta$ vanishes as $n \to \infty$ and also that the conditional expectations of $X_t \,|\, \theta$ do not vary too much in time (Hayashi, 2000, p. 403). In that case it is guaranteed that the CLT holds with an asymptotic or, say, *long-run variance*

$$\sigma^2_{\mathrm{L}}(\theta) := \sum_{k=-\infty}^{\infty} \gamma_\theta(k)\,,$$

where γ_θ is the autocovariance function of $(X_t \,|\, \theta)$ (Hayashi, 2000, p. 401) given the unknown parameter θ. This result can be easily extended to any d-dimensional stochastic

process (Hayashi, 2000, p. 405). Hence, in the following let $(X_t \,|\, \theta)$ be an ergodic stationary d-dimensional process satisfying Gordin's condition.

>From Gordin's CLT it follows that long-term asset log-returns typically tend to be normally distributed even if the short-term log-returns are serially dependent and heavy tailed. A broad class of time series models satisfy Gordin's condition. Hence, long-term asset log-return vectors are approximately normally distributed, i.e.

$$\log(\mathbf{1} + R) \,|\, \theta = \sum_{t=1}^{T} X_t \,|\, \theta =: X \,|\, \theta \sim \mathcal{N}_d\big\{T\eta(\theta), T\Upsilon_{\mathrm{L}}(\theta)\big\}\,, \tag{3.3}$$

where $\mathbf{1}$ represents a column vector of ones and $\log(\cdot)$ is understood as taking the logarithm of each component separately. Here $\Upsilon_{\mathrm{L}}(\theta)$ denotes the long-run covariance matrix of the stochastic process (Hayashi, 2000, p. 404) and $T \in \mathbb{N}$ represents the number of aggregated short-term log-returns or, say, the investment horizon. For example, if $X_1, \ldots, X_T$ represent daily log-returns, the sum given by Eq. 3.3 denotes a quarterly log-return if $T = 63$ and a yearly log-return in case $T = 252$.

Of course, the Gaussian distribution hypothesis holds only approximately. However, in the following the additional suffix 'approximately' or any corresponding symbol are suppressed for convenience. It is worth to mention that we generally suppose that both $\eta(\theta)$ and $\Upsilon_{\mathrm{L}}(\theta)$ can be computed either numerically or analytically under the specific time series model which is used for the short-term asset log-returns provided the model parameter θ is known. Specifically, if $(X_t - \eta(\theta) \,|\, \theta)$ is a martingale difference sequence (Hayashi, 2000, p. 104), that means if

$$\mathrm{E}(X_t \,|\, \mathcal{H}_{t-1}, \theta) = \eta(\theta)\,, \qquad \forall\, t \in \mathbb{Z}\,,$$

the components of $(X_t \,|\, \theta)$ are serially uncorrelated. In that case the long-run covariance matrix $\Upsilon_{\mathrm{L}}(\theta)$ turns out to be the *stationary* variance $\Upsilon(\theta)$ of $(X_t \,|\, \theta)$. The martingale difference property is satisfied for a broad class of time series models, such as the family of multivariate GARCH processes (Bauwens et al., 2006).

As elucidated in the introduction, estimating the moments $T\eta(\theta)$ and $T\Upsilon_{\mathrm{L}}(\theta)$ from long-term asset returns is inefficient. For example, we could estimate the quantity $T\Upsilon_{\mathrm{L}}(\theta)$ simply by applying the sample covariance matrix to the corresponding long-term asset log-returns. However in that case we would ignore a large part of the data and the resulting standard error would increase roughly by a factor of $\sqrt{T}$ relative to the approach based on high-frequency data. Hence, decreasing the sampling frequency leads to a loss of statistical efficiency.

3.2.3. Bayesian Framework

In the Bayesian framework the model parameter θ is not assumed to be fixed but it is considered as a random quantity possessing some prior distribution $p(\theta)$. The posterior distribution $p(\theta \,|\, x)$ corresponds to the distribution of θ given some observed data x. More specifically, in the following we shall interpret x as historical short-term asset log-return data. The likelihood function $\mathcal{L}(\theta\,; x) = p(x \,|\, \theta)$ represents some pre-defined probabilistic model for x. Now the posterior distribution of θ can be obtained by the Bayes formula

$$p(\theta \,|\, x) = \mathcal{L}(\theta\,; x)\, p(\theta)/p(x)\,,$$

so that the posterior involves both empirical and subjective information.

However, in Bayesian analysis the posterior distribution is not always the desired object. Instead, one can be interested in the predictive distribution of the data. Let y be some unobserved data where x and y are conditionally independent given θ. Then

$$p(y \,|\, x) = \int p(y \,|\, \theta)\, p(\theta \,|\, x)\, d\theta$$

represents the predictive distribution of y. In the following discussion this can be interpreted as the distribution of a long-term asset log-return if we take the parameter uncertainty additionally into account. Each parameter is weighted by its posterior probability, i.e. the probability of θ given the historical observations and some expert knowledge. Notice that analytical solutions for the portfolio optimization problem which are based on the predictive distribution are only available for relatively simple expressions for the prior $p(\theta)$ and the likelihood $\mathcal{L}(\theta\,; x)$.

The prior $p(\theta)$ can be either 'diffuse' or 'informative'. If the prior is *diffuse* the model parameter is assumed to possess some ad-hoc distribution such as the uniform distribution or the standard normal distribution. The prior is called *informative* if some subjective information is necessary to determine $p(\theta)$. The chosen terminology is somewhat misleading since we do not mean that diffuse priors in general are non-informative in the probabilistic sense since the posterior distribution might drastically depend on the chosen diffuse prior. Hence, we believe that Bayesian analysis is inherently subjective and since most practitioners have some basic opinions about the evolution of asset prices they might want to include that information in the optimization process (Black and Litterman, 1992). The present work heavily relies on the idea of using subjective information whenever it is possible.

One popular example of Bayesian portfolio optimization is the approach of Black and Litterman (1992). They show how to distill implicit information about the distribution of asset returns from the market by using standard results of portfolio theory. This is combined with the investor's own belief which typically leads to optimal portfolios being more robust against estimation errors than solutions obtained by pure statistical methods. However, in order to be analytically tractable, the Black-Litterman approach assumes that asset returns are normally distributed. Other Bayesian portfolio optimization techniques are given by the work of Frost and Savarino (1986) and Jorion (1986). They all share the same disadvantage, namely that an analytic expression of the predictive distribution or optimal portfolio is only available by imposing unrealistic assumptions on the underlying data or otherwise being inefficient, since they have to be applied by using low-frequency data.

Scherer and Martin (2007, Ch. 7) suggest to apply so-called *conjugate* priors in Bayesian portfolio optimization. These are informative priors which, after multiplying with the likelihood, lead to a posterior distribution that is of the same type as the chosen likelihood function. Again, this limitation can be motivated by the requirement to obtain analytically tractable expressions for the posterior distribution. However, unrealistic assumptions about the distribution of empirical data are necessary in general and the set of possible prior distributions is substantially restricted. In particular, conjugate priors often are not available if the assumption of normally distributed asset returns is relaxed. Scherer and Martin (2007, Ch. 7) refer to a Markov chain Monte Carlo method (which will be discussed later on in Section 3.3) to simulate the posterior distribution of the mean and variance of a single asset return. In this work we will show how this idea can be extended to incorporate arbitrary prior information given the asset returns are not normally distributed.

For choosing some likelihood function for θ we have to consider an appropriate model for the data, that means to take account for the stylized facts of empirical finance. These can be subsumed by the following anomalies (see McNeil et al., 2005, p. 117):

(1) Short-term asset returns are heavy-tailed and particularly not Gaussian.

(2) Asset returns are not independent and identically distributed although they show little serial correlation.

(3) In contrast, squared asset returns show strong serial correlation.

(4) Asset volatility varies over time and appears in clusters.

There are several alternatives to deal with these phenomena. For instance, GARCH processes (Bollerslev, 1986, Engle, 1982) can be used to model volatility clusters. Another possibility is to work with stochastic volatility models (Barndorff-Nielsen et al., 2002, Jacquier et al., 1994, 2004).

3.2.4. Predictive Moments

In the last section we mentioned that the parameter θ is considered as a random quantity and from Section 3.2.2 we know that

$$X \,|\, \theta \sim \mathcal{N}_d\big\{T\eta(\theta), T\Upsilon_{\mathrm{L}}(\theta)\big\}\,,$$

where $X \,|\, \theta$ denotes a long-term log-return vector given the unknown parameter θ. Hence, the vector of long-term discrete returns is given by

$$R \,|\, \theta = \exp(X \,|\, \theta) - \mathbf{1}\,,$$

where $\exp(\cdot)$ shall be interpreted as a component-wise function. Thus each component of $R \,|\, \theta$ is log-normally distributed and it can be easily shown that

$$\mathrm{E}(R \,|\, \theta) = \exp\big[T\big\{\eta(\theta) + \mathrm{diag}(\Upsilon_{\mathrm{L}}(\theta))/2\big\}\big] - \mathbf{1}$$

and

$$\mathrm{Var}(R \,|\, \theta) = \exp\Big[T\big\{\eta(\theta)\mathbf{1}' + \mathbf{1}\eta(\theta)' + D(\theta)\big\}\Big] \odot \Big[\exp\big\{T\Upsilon_{\mathrm{L}}(\theta)\big\} - \mathbf{1}\mathbf{1}'\Big]\,,$$

where $\odot$ denotes the Hadamard (i.e. component-wise) product, and

$$D(\theta) = \frac{\mathrm{diag}\{\Upsilon_{\mathrm{L}}(\theta)\}\mathbf{1}' + \mathbf{1}\mathrm{diag}\{\Upsilon_{\mathrm{L}}(\theta)\}'}{2}\,.$$

Finally, we obtain the predictive moments of the long-term log-return vector by the law of total expectations and the variance decomposition theorem, viz.

$$\mu = \mathrm{E}(R) = \mathrm{E}\big\{\mathrm{E}(R \,|\, \theta)\big\}$$

and

$$\Sigma = \mathrm{E}\big\{\mathrm{Var}(R \,|\, \theta)\big\} + \mathrm{Var}\big\{\mathrm{E}(R \,|\, \theta)\big\}\,.$$

Interestingly, the conditional means of the discrete returns are also determined by the long-run variances. Moreover, predictive expectations and variances of discrete returns are *nonlinear* functions of the investment horizon T. Hence, the investment horizon can have a substantial impact on the optimal portfolio. In Section 3.3 we will see how the predictive moments can be approximated by Monte Carlo simulation.

3.3. Numerical Implementation

Now we will discuss several Markov chain Monte Carlo algorithms for simulating the posterior distribution $p(\theta \mid x)$ even if this has a rather complicated analytical structure. There is a big number of different simulation techniques like for instance *importance sampling* (Gamerman and Lopes, 2006, Ch. 3.4). However, we got the best simulation results in reasonable time using a Markov chain Monte Carlo algorithm, which will be presented in the following sections. In our case we want to use Markov chains only to sample from a complex posterior distribution. Hence, we have to guarantee that the stationary distribution of the considered Markov chain corresponds to $p(\theta \mid x)$.

3.3.1. Gibbs Sampling

A simple approach is known as *Gibbs sampling*. That means for simulating θ we could principally start with some initial parameter vector $\theta = (\theta_1, \ldots, \theta_p)$ and draw a new realization θ_1' of the first component from the conditional distribution of θ_1 given $\theta_2, \ldots, \theta_p$. Then we can take the new parameter vector $(\theta_1', \theta_2, \ldots, \theta_p)$ into consideration and simulate the second component of θ by drawing from the distribution of θ_2 under the new condition $\theta_1', \theta_3, \ldots, \theta_p$, etc., until we obtain the parameter vector $\theta' = (\theta_1', \ldots, \theta_p')$. If the same procedure is repeated with θ' and so on we obtain a Markov chain whose stationary distribution corresponds to the posterior distribution of θ. Scherer and Martin (2007, Ch. 7) give an example of how to use Gibbs sampling for simulating the posterior distribution of the mean and variance of a normally distributed single asset return by using a conjugate prior. However, in our case this is not a useful approach since drawing from the conditional posterior distributions of θ is not substantially easier than drawing directly from $p(\theta \mid x)$.

3.3.2. Metropolis-Hastings Algorithm

Another MCMC scheme which is frequently used in Bayesian statistics is the Metropolis-Hastings algorithm (Hastings, 1970, Metropolis et al., 1953). An application to the Bayesian analysis of stochastic volatility models is presented by Jacquier et al. (2004). The Metropolis-Hastings algorithm is very similar to the Gibbs sampler, but unlike that, it does not require to sample from the conditional stationary distribution. In contrast, the sampling part is completely reduced to sampling from an arbitrary *proposal distribution* which is easy to draw from. The stationary distribution is then only needed to calculate

the *acceptance probability* of each new state in the chain, which comes from the proposal distribution. This is why we choose a Metropolis-Hastings-like algorithm to simulate the distribution of $\theta|x$. First, we will present the Metropolis-Hastings algorithm and after that an extension called *parallel tempering* will be discussed.

Assume there exists some *target distribution* $\pi(\theta)$ which shall be simulated. The current state of the chain will be denoted by ϕ. In case of the Metropolis-Hastings algorithm, the simulation is done by introducing an 'easy to draw from' *proposal distribution* $q(\phi, \phi')$ which denotes the distribution of a proposal to move from state ϕ to state ϕ'. However, the actual probability to move from ϕ to ϕ' is determined by the acceptance probability

$$\alpha(\phi, \phi') = \min\left\{1, \frac{\pi(\phi')\, q(\phi', \phi)}{\pi(\phi)\, q(\phi, \phi')}\right\}. \tag{3.4}$$

Note that if we have a symmetric proposal distribution, the acceptance probability is simply given by $\alpha = \min\{1, \pi(\phi')/\pi(\phi)\}$.

The probability density of a new state ϕ' given an old state ϕ, that is the so-called *transition kernel* $K(\phi, \phi')$ (Gamerman and Lopes, 2006, p. 194) of the Markov chain, is given by

$$K(\phi, \phi') = q(\phi, \phi')\, \alpha(\phi, \phi') + \delta(\phi' - \phi)\left(1 - \int q(\phi, \xi)\, \alpha(\phi, \xi)\, d\xi\right),$$

where δ is the Dirac distribution. It can be shown that for the acceptance probability given by Eq. 3.4, the *detailed balance condition*

$$\pi(\phi)K(\phi, \phi') = \pi(\phi')K(\phi', \phi)$$

is satisfied for all ϕ and ϕ'. Thus we obtain a reversible Markov chain (Gamerman and Lopes, 2006, Ch. 4.6). That means by the presented Metropolis-Hastings algorithm in fact we are able to simulate realizations from the target distribution π.

3.3.3. Parallel Tempering

Though the Metropolis-Hastings algorithm is very powerful, one big problem can easily occur: The Markov chain can get stuck in local optima for a very long time. Assume for instance a univariate bi-modal distribution. If the chain is currently in a region around one of the two modes, there is almost no incentive to move to the region around the other mode, since the acceptance probability $\alpha(\phi, \phi')$ approaches zero if $\pi(\phi')$ is much smaller than $\pi(\phi)$. To avoid this problem, the idea of *heated* equilibrium distributions has been

introduced. Instead of simulating only one stationary distribution $\pi(\theta)$ at a time, m parallel chains are used, each having an equilibrium distribution

$$\pi_i(\theta) \propto \pi_1(\theta)^{(1/T_i)}, \qquad \forall\, i = 1, \ldots, m\,,$$

where T_i is the *temperature* of the distribution $\pi_i(\theta)$. The temperature of the desired stationary distribution $\pi_1(\theta)$ is $T_1 = 1$. At each iteration of the algorithm, an exchange between the states ϕ_i and ϕ_j of chain i and j is proposed. The acceptance probability of this swap is

$$\alpha_{ij}(\phi_i, \phi_j) = \min\left\{1, \frac{\pi_i(\phi_j)\,\pi_j(\phi_i)}{\pi_i(\phi_i)\,\pi_j(\phi_j)}\right\}.$$

One disadvantage of this method is that only the outcome of chain 1 contains samples from the desired distribution and all the other samples are dropped. However, especially for very complex distributions the advantage of not getting stuck in local modes overcomes the disadvantage of high computational effort. For further details and applications of tempering algorithms see for instance Gamerman and Lopes (2006, Ch. 6 and Ch. 7).

In our case the stationary distribution which has to be simulated is the posterior distribution of the model parameters which can become very complex. In our empirical study we will use $m = 2$ different chains. For the proposal distribution we choose a composite distribution $q(\theta, \theta')$ by taking account of the specific domains of the different components of θ. Of course we could also choose a proposal distribution which probably leads to realizations outside of Θ but, however, if some parameter is proposed to exceed the parameter set, the prior probability and thus also the acceptance probability becomes zero. Hence, it cannot happen that we get some realizations of θ such that $\theta \notin \Theta$.

Our implementation of the parallel tempering algorithm is as follows:

1. Create the initial parameter vectors θ_1 and θ_2 .

2. Repeat the following steps very often:

   a) Generate θ_1' and θ_2' by randomly drawing from the proposal distributions.

   b) Calculate $p(\theta_1' \,|\, x) \propto \mathcal{L}(\theta_1' \,; x)\, p(\theta_1')$ and $p(\theta_2' \,|\, x) \propto \mathcal{L}(\theta_2' \,; x)\, p(\theta_2')$.

   c) Calculate

   $$\alpha_1 = \min\left\{1, \frac{p(\theta_1' \,|\, x)\, q(\theta_1', \theta_1)}{p(\theta_1 \,|\, x)\, q(\theta_1, \theta_1')}\right\}$$

   and

   $$\alpha_2 = \min\left\{1, \frac{p(\theta_2' \,|\, x)^{(1/T_2)} q(\theta_2', \theta_2)}{p(\theta_2 \,|\, x)^{(1/T_2)} q(\theta_2, \theta_2')}\right\}.$$

d) Set $\theta_1 = \theta_1'$ with probability α_1, and $\theta_2 = \theta_2'$ with probability α_2, otherwise keep the old θ_1 or θ_2, respectively.

e) Swap the states θ_1 and θ_2 of the chains with probability

$$\alpha_{12}(\theta_1, \theta_2) = \min\left\{1, \frac{p(\theta_2 \mid x)\, p(\theta_1 \mid x)^{(1/T_2)}}{p(\theta_1 \mid x)\, p(\theta_2 \mid x)^{(1/T_2)}}\right\}.$$

As mentioned above we only consider the realizations of the first chain which are obtained after some burning-in phase.

3.4. Empirical Study

In this section we will present an empirical study based on the framework developed in the previous sections. First, we create a model for high-frequency asset log-returns by taking account of stylized facts. It is a multivariate extension of the GARCH model developed by Bollerslev (1986). A comprehensive overview on different multivariate GARCH (MGARCH) models is given in Bauwens et al. (2006). MGARCH processes are martingale difference sequences and so Gordin's condition (see Section 3.2.2) is automatically satisfied. Further, the predictive moments (see Section 3.2.4) can be easily calculated by the MCMC algorithm discussed in Section 3.3. After the data generating process is developed, we present the chosen prior information for the unknown model parameter θ. Then we will apply our method to time series data to find optimal portfolios.

3.4.1. Modeling the Distribution of Asset Log-Returns

In this section we will describe a way for modeling the distribution of daily asset log-returns. We will concentrate on risky assets. The risk-free asset or, say, money market account does not possess any market risk per definition. That means we do not need any stochastic model and so there exists no parameter uncertainty.

In order be provide a flexible framework for the asset returns, we rely on the broad class of elliptically symmetric distributions. A d-dimensional random vector X is said to be *elliptically symmetric distributed* (Cambanis et al., 1981) if and only if

$$X \overset{\mathrm{d}}{=} \eta + \Gamma \mathcal{R} U$$

with $\eta \in \mathbb{R}^d$ being a location vector, $\Gamma \in \mathbb{R}^{d \times k}$ is a transformation matrix, U a k-dimensional random vector uniformly distributed on the unit hypersphere, and $\mathcal{R}$ is a

non-negative random variable stochastically independent of U. The positive semi-definite matrix $\Omega := \Gamma\Gamma'$ is referred to as the *dispersion matrix* of X and $\mathcal{R}$ is called its *generating variate*. By choosing $\mathcal{R}$ properly, we are able to account for stylized facts like heavy tails. Further, it can be shown that

$$V := \mathrm{Var}(X) = \mathrm{E}(\mathcal{R}^2)/k \cdot \Omega$$

is the covariance matrix of X provided $\mathrm{E}(\mathcal{R}^2) < \infty$.

A d-dimensional MGARCH process (X_t) is characterized by

$$X_t \,|\, \mathcal{H}_{t-1} \stackrel{\mathrm{d}}{=} \eta + V_t^{\frac{1}{2}} \epsilon_t \,,$$

where η is a $d \times 1$ vector of time-independent expected log-returns, V_t is a function only of $\mathcal{H}_{t-1}$ and denotes the $d \times d$ positive definite conditional covariance matrix of the log-return vector X_t, and ϵ_t is an independent and identically distributed $d \times 1$ vector of perturbations with $\mathrm{E}(\epsilon_t) = 0$ and covariance matrix $\mathrm{Var}(\epsilon_t) = I_d$. If ϵ_t is assumed to be spherically distributed, i.e. elliptically symmetric with location 0 and dispersion proportional to I_d, then the MGARCH model perfectly fits into the class of elliptically symmetric distributions.

There are various specifications of the time-dependent covariance matrix V_t. For a thorough discussion of MGARCH processes see Bauwens et al. (2006). Since MGARCH specifications often require a huge number of parameters and are hardly applicable to practical problems, for complexity reduction we suggest to use a principal components model for the asset log-returns. The underlying idea of principal components is that most of the dynamics of the observed data can be explained by a small number of uncorrelated factors. The spectral decomposition theorem assures that the covariance matrix V of an elliptically symmetric distributed random vector X can be decomposed into $V = \mathcal{O}\Lambda\mathcal{O}'$, where

- Λ is the diagonal matrix of the eigenvalues $\lambda_1, \ldots, \lambda_d$ of V and
- $\mathcal{O}$ is an orthogonal $d \times d$ matrix containing the associated eigenvectors.

By applying this decomposition for the vector of asset log-returns we can specify the MGARCH model as

$$X_t \,|\, \mathcal{H}_{t-1} \stackrel{\mathrm{d}}{=} \eta + \mathcal{O}\Lambda_t^{\frac{1}{2}} \epsilon_t$$

and define

$$Y_t := \Lambda_t^{\frac{1}{2}} \epsilon_t = \mathcal{O}'(X_t - \eta) \,.$$

This reduces the number of required model parameters tremendously, since the elements of Y_t are uncorrelated per definition. However, we have to presume that the eigenvectors do not change over time. Speaking economically, the factors which drive the dynamics of the asset log-returns do not change but the *impact* of each factor can vary over time. For modeling the components of Λ_t we can simply assume that Y_t consists of d unrelated univariate GARCH(1,1) processes. The resulting process is sometimes called *orthogonal* GARCH (Bauwens et al., 2006).

Principally, we can choose any elliptically symmetric distribution for modeling the perturbation ϵ_t as long as the corresponding density function can be computed either numerically or analytically. However, here we assume that ϵ_t is multivariate t-distributed, i.e.

$$\epsilon_t \sim t_d\left(0\,, \frac{\nu-2}{\nu}\cdot I_d\,, \nu\right)$$

with $\nu > 2$ degrees of freedom and the dispersion matrix is such that $\mathrm{Var}(\epsilon_t) = I_d$. Hence, the random vector $X_t \,|\, \mathcal{H}_{t-1}$ possesses the density

$$p(x_t \,|\, \mathcal{H}_{t-1}) = \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{\nu}{2})}\cdot\sqrt{\frac{\det\Lambda_t^{-1}}{(\nu\pi)^d}}\cdot\left(1+\frac{(x_t-\eta)'\mathcal{O}\Lambda_t^{-1}\mathcal{O}'(x_t-\eta)}{\nu-2}\right)^{-\frac{d+\nu}{2}},$$

where Λ_t is a diagonal $d\times d$ matrix with main diagonal elements

$$\lambda_{it} = \gamma_i + \alpha_i Y_{i,t-1}^2 + \beta_i\lambda_{i,t-1}\,, \qquad i = 1,\ldots,d\,, \tag{3.5}$$

representing the conditional variances of the d principal components. Note that the orthogonal matrix $\mathcal{O}$ $(d\times d)$ contains $\binom{d}{2}$ free parameters and there are $3d$ GARCH parameters. Altogether, the resulting data generating process contains only $d\,(d+7)/2+1$ parameters.

3.4.2. Modeling the Prior Information

There are several ways to implement prior information. In case of a diffuse prior there is no explicit information that is incorporated into the prior distribution. This is often done to get an analytical expression for the posterior distribution and so to obtain an analytical result for the optimal portfolio. However, it can be shown that the diffuse prior approach can lead to paradox results (Berger, 2006) and the concrete choice of the diffuse prior can have a substantial impact on the optimal decision. Therefore, as already mentioned, it is suggested to use informative priors whenever it is possible.

Our hierarchical approach is very general. First of all note that our model parameters are given by $\eta, \alpha, \beta, \lambda, \mathcal{O}, \nu$. Here η $(d\times 1)$ is the vector of expected asset log-returns, α $(d\times 1)$

and β ($d \times 1$) contain the GARCH(1,1) parameters according to (3.5) and the $d \times 1$ vector λ contains the *unconditional* variances $\lambda_1, \ldots, \lambda_d$, i.e.

$$\lambda_i = \frac{\gamma_i}{1 - \alpha_i - \beta_i}, \qquad i = 1, \ldots, d.$$

Note that the parameters $\gamma_i = \lambda_i (1 - \alpha_i - \beta_i)$ $(i = 1, \ldots, d)$ follow implicitly from α, β, and λ. That means we use the following re-parameterization of Eq. 3.5:

$$\lambda_{it} = \lambda_i (1 - \alpha_i - \beta_i) + \alpha_i Y_{i,t-1}^2 + \beta_i \lambda_{i,t-1}, \qquad i = 1, \ldots, d.$$

We will substitute $\mathcal{O}$ by an estimate based on the sample covariance matrix of the time series data. That means $\mathcal{O}$ is fixed for the sake of simplicity. Finally, the number of degrees of freedom ν is set to 3 to account for the typical heavy tails of daily log-returns. We did not observe any improvements by introducing some prior distribution for ν. Hence, we obtain the parameter vector $\theta = (\eta, \alpha, \beta, \lambda)$ and suppose that they are a priori stochastically independent, i.e.

$$p(\theta) = p(\eta)\, p(\alpha)\, p(\beta)\, p(\lambda).$$

Since $\alpha, \beta \in (0, 1)$ we decided to use flat priors for α and β where the components of α and β are assumed to be mutually independent. So the prior for θ can be simply expressed as $p(\theta) = p(\eta)\, p(\lambda)$.

Also the components of λ are assumed to be mutually independent but each one follows a gamma distribution, i.e. $\lambda_i \sim \Gamma(\kappa_2, \lambda_0/\kappa_2)$ $(i = 1, \ldots, d)$ and $\lambda_0, \kappa_2 > 0$. Hence, we expect a priori that each principal component has the same proportion of total variation. Note that $\mathrm{E}(\lambda_i) = \lambda_0$ is constant but $\mathrm{Var}(\lambda_i) = \lambda_0^2/\kappa_2$. That means κ_2 can be interpreted as the investor's confidence that the unconditional variances of the principal components indeed correspond to λ_0. In our empirical study we choose $\lambda_0 = 0.2^2/T$ and $\kappa_2 = 2$.

For the expected values of the daily log-returns we use the prior proposed by Jorion (1986), i.e.

$$\eta \,|\, V \sim \mathcal{N}_d(\eta_0, V/\kappa_1),$$

where η_0 is a vector of prior expected returns. We decided to choose $\eta_0 = 0$ since sample means of daily log-returns are typically close to zero (McNeil et al., 2005, p. 117). The scale parameter κ_1 represents the confidence of the investor in their a priori assumption concerning η and can be seen as a virtual sample size. For instance, if there are $n = 1260$ observations (i.e. 5 trading years) then $\kappa_1 = 1260$ would mean that the investor trusts in their own belief about η as much as the empirical evidence given by the time series.

	USA	UK	JPN	ITA	GER	FRA	CAN
$\hat{\mu}$	5.98%	12.50%	12.78%	17.63%	14.27%	14.53%	20.97%
$\hat{\sigma}$	16.07%	13.70%	21.55%	14.68%	23.44%	17.90%	22.10%

Table 3.1.: Descriptive statistics of yearly discrete returns.

Note that $V = \mathcal{O}\Lambda\mathcal{O}'$ where $\mathcal{O}$ is fixed and Λ is random. Hence, we can write Jorion's prior equivalently as

$$\eta \,|\, \Lambda \sim \mathcal{N}_d(0\,, \mathcal{O}\Lambda\mathcal{O}'/\kappa_1)$$

such that $p(\eta) = p(\eta \,|\, \Lambda)\, p(\Lambda)$ can be easily calculated, since

$$p(\Lambda) = p(\lambda) \propto \prod_{i=1}^{d} \lambda_i^{\kappa_2 - 1} \, \exp\!\left(-\frac{\kappa_2 \lambda_i}{\lambda_0}\right)$$

and

$$p(\eta \,|\, \Lambda) \propto \exp\!\left(-\frac{\kappa_1}{2} \cdot \eta' \mathcal{O}\Lambda^{-1}\mathcal{O}'\eta\right).$$

3.4.3. Data Description

In our empirical study we use daily log-returns of seven MSCI stock indices of the countries USA, UK, Japan, Italy, Germany, France, and Canada. The indices are adjusted by dividends, splits, etc. and are calculated on the basis of USD stock prices. We have $n = 1260$ daily observations ranging from 2001-12-03 to 2006-09-29 and the whole sample is divided chronologically into 5 subsets where each subset contains 252 observations. In Table 3.1 we can see the sample means and standard deviations of the yearly discrete returns of each country. In our study we assume that the investment horizon corresponds to 1 year, i.e. $T = 252$ and the quantities given in Table 3.1 are based on the available 5 observations of yearly discrete returns. Of course, since the sample size is very small, these values are strongly affected by estimation errors.

The process $(X_t \,|\, \theta)$ of daily log-returns is assumed to be an ergodic stationary martingale difference sequence as described in Section 3.2.2. Hence, both the sample mean $\hat{\eta}$ and the sample covariance matrix $\widehat{\Upsilon}$ of the daily log-returns are strongly consistent estimators for $\eta(\theta)$ and $\Upsilon_{\mathrm{L}}(\theta)$, respectively. Now we can also estimate the first and second moments of yearly discrete returns by using the formulas given in Section 3.2.4 based on daily log-returns, viz.

$$\widehat{\mathrm{E}}(R \,|\, \theta) = \exp\!\Big[252\,\big\{\hat{\eta} + \mathrm{diag}\big(\widehat{\Upsilon}\big)/2\big\}\Big] - \mathbf{1}$$

	USA	UK	JPN	ITA	GER	FRA	CAN
$\hat{\mu}$	6.29%	13.41%	13.46%	18.54%	15.23%	15.73%	20.71%
$\hat{\sigma}$	17.42%	19.60%	24.13%	20.42%	28.14%	24.52%	19.29%

Table 3.2.: Descriptive statistics of yearly discrete returns based on daily log-returns.

$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$	$\hat{\lambda}_7$
6.45 e-4	1.64 e-4	0.94 e-4	0.48 e-4	0.32 e-4	0.21 e-4	0.14 e-4
63.35 %	16.09 %	9.21 %	4.75 %	3.17 %	2.09 %	1.33 %

Table 3.3.: Eigenvalues of the sample covariance matrix of daily log-returns.

and

$$\widehat{\mathrm{Var}}(R \,|\, \theta) = \exp\Big[252 \left\{\hat{\eta}\mathbf{1}' + \mathbf{1}\hat{\eta}' + \widehat{D}\right\}\Big] \odot \Big[\exp\big\{252\,\widehat{\Upsilon}\big\} - \mathbf{1}\mathbf{1}'\Big]\,,$$

where

$$\widehat{D} = \frac{\mathrm{diag}\big\{\widehat{\Upsilon}\big\}\mathbf{1}' + \mathbf{1}\mathrm{diag}\big\{\widehat{\Upsilon}\big\}'}{2}\,.$$

The corresponding values are given in Table 3.2. Note that there are only slight differences between the results in Table 3.1 and Table 3.2 regarding the means but for the standard deviations the results can differ substantially.

Table 3.3 contains the eigenvalues of $\widehat{\Upsilon}$ as well as their proportions of the total variation. As described earlier, each eigenvalue can be interpreted as the unconditional variance of a principal component. In our case, the first component (i.e. the systematic risk of the market) almost explains two third of the total variation and the impact of the other components are relatively small. Similar results for financial data have been frequently observed in literature (see, e.g., Plerou et al., 1999). Note that our prior expectation for λ_i corresponds to $\lambda_0 = 0.2^2/252$ =1.59 e-4, which reflects a relatively conservative assumption relative to the empirical results. For the confidence in λ_0 we choose the parameter $\kappa_2 = 2$ which leads to an a priori standard deviation of λ_i roughly corresponding to 1.12 e-4 ($i = 1, \ldots, d$).

3.4.4. Results

In this section we present the results of our simulation. Our main objective is to demonstrate the practical applicability of our approach. We want to show how prior information

can be used to account for estimation risk – even if the underlying model is complex – and to obtain well-diversified portfolios. The parameter κ_1, which reflects the investor's confidence in their prior assumption about the expected log-returns, is varied in order to see how expert knowledge determines the optimal portfolio. Asset return variances and covariances can be estimated quite good by using short-term asset returns. In contrast, it is well-known that portfolio selection is very sensitive to expected asset returns which cannot be estimated accurately. Hence, investors preferably have a strong confidence about expected asset returns in order to reduce estimation risk. This is the reason why we kept $\kappa_2 = 2$ fixed, which indicates that there is only little confidence in the prior information about the eigenvalues.

We performed standard Markowitz portfolio selection (Markowitz, 1952). Our objective function is the traditional mean-variance certainty equivalent given by Eq. 3.2 where we choose a risk aversion of $\alpha = 1$. In many practical situations constraints are included in the optimization problem. For instance, investors might be willing to forbid short-selling. Other constraints might be given by legal issues and so on. We do not want to provide optimal portfolios for each imaginable investor, but instead we present a flexible framework which can be adapted to most kinds of situations.

Each additional constraint limits the space of alternatives. Therefore, in the first part of the study (P1) we have only one constraint, namely the budget constraint $\mathcal{C}_{\mathrm{B}} : w'\mathbf{1} = 1$. The short-selling constraint $\mathcal{C}_{\mathrm{S}} : w \geq 0$ is additionally considered in the second part of the study (P2). In our study we are searching for the optimal portfolio given by (3.1) using the objective function

$$\varphi(v) = v'\mu - \frac{1}{2} \cdot v'\Sigma\, v\,, \qquad \text{s.t. } v \in \mathcal{C}\,,$$

where $\mathcal{C} = \mathcal{C}_1 = \mathcal{C}_{\mathrm{B}}$ in P1 and $\mathcal{C} = \mathcal{C}_2 = \mathcal{C}_{\mathrm{B}} \cap \mathcal{C}_{\mathrm{S}}$ in P2.

Table 3.4 contains our results of the portfolio optimization. These can be compared with the portfolio weights obtained by traditional Markowitz optimization, i.e. searching for the *Markowitz portfolio* (MP), viz.

$$\mathrm{MP} = \arg\max_v v'\widehat{\mathrm{E}}(R \,|\, \theta) - \frac{1}{2} \cdot v'\widehat{\mathrm{Var}}(R \,|\, \theta)v\,, \qquad \text{s.t. } v \in \mathcal{C}\,,$$

and the so-called *global minimum variance portfolio* (MVP), i.e.

$$\mathrm{MVP} = \arg\min_v v'\widehat{\mathrm{Var}}(R \,|\, \theta)v\,, \qquad \text{s.t. } v \in \mathcal{C}\,.$$

The MVP has been advocated by many authors as an alternative to the traditional mean-variance optimal portfolio since there are no expected asset returns which have to be estimated and thus the impact of estimation errors can be substantially reduced (Frahm, 2008).

As we can see in Table 3.4 the Markowitz portfolios tend to overrate assets with large expected returns. In P1 the MP suggests a short-selling of 484.01% of USA and investing 487.05% in CAN - a strategy which would certainly not be pursued in practice. When short-selling is forbidden, all the available capital is invested in CAN. Compared to that the two minimum variance portfolios are far more diversified. However, it can be clearly seen that these portfolios are not optimal in the sense of expected return maximization, since the asset with the smallest estimated return, USA, possesses the highest weight in both minimum variance portfolios.

The optimal portfolios in case $\kappa_1 = 1$, which almost corresponds to a diffuse prior information about the expected asset returns, are similar to the Markowitz portfolios. However, using an appropriate model for high-frequency data apparently leads to slight changes of the expected returns, variances, and covariances which alters the optimal portfolios. Nevertheless, the optimal portfolio for $\kappa_1 = 1$ in P2 is the same as in the empirical case, where all the capital is invested in CAN.

The more confident the investor is about the expected asset returns, the more the optimal portfolios tend to be diversified. In case $\kappa_1 = 1260$ the investor relies on their prior assumption about the expected returns as much as on the empirical information. The optimal portfolio in P1 does not possess weights which are such excessive as for traditional Markowitz optimization or in the case $\kappa_1 = 1$. For instance the amount of capital invested in CAN reduces to 404.76%. In P2 not all the capital is put into CAN anymore. Instead, 14.70% is invested in JPN now. The reason for that is that the expected predictive asset returns are shrunk towards the prior assumption $\eta_0 = 0$. So increasing the confidence in prior information clearly reduces estimation risk. This effect even strengthens when κ_1 is further increased.

In fact, $\kappa_1 = 6300$ is a configuration which can be seen as typical for practical investment problems. Here the investor trusts their own assumption about the expected returns 5 times more than the empirical information. Recall that we use a time series of daily log-returns lasting 5 years, which means that the estimation of yearly expected returns is based on 5 observations. So from a practical point of view, when it comes to estimating

empirical	USA	UK	JPN	ITA	GER	FRA	CAN
$\hat{\mu}$	6.29%	13.41%	13.46%	18.54%	15.23%	15.73%	20.71%
$\hat{\sigma}$	17.42%	19.60%	24.13%	20.42%	28.14%	24.52%	19.29%
MP_1	−484.01%	−195.93%	−13.18%	373.62%	−2.45%	−65.10%	487.05%
MP_2	0%	0%	0%	0%	0%	0%	100%
MVP_1	50.37%	37.72%	20.13%	43.27%	−28.86%	−28.46%	5.84%
MVP_2	42.49%	19.17%	23.88%	4.64%	0%	0%	9.83%
$\kappa_1 = 1$	USA	UK	JPN	ITA	GER	FRA	CAN
μ	5.25%	12.44%	19.57%	16.29%	13.32%	14.52%	25.29%
σ	17.96%	20.66%	27.29%	21.39%	29.02%	25.80%	21.53%
w_1	−553.87%	−208.89%	54.80%	200.23%	−25.00%	2.05%	630.68%
w_2	0%	0%	0%	0%	0%	0%	100%
1260	USA	UK	JPN	ITA	GER	FRA	CAN
μ	4.76%	9.06%	14.02%	11.77%	11.94%	11.52%	15.26%
σ	17.57%	19.39%	24.97%	19.97%	27.98%	24.39%	18.83%
w_1	−373.32%	−190.63%	62.17%	98.56%	64.67%	33.79%	404.76%
w_2	0%	0%	14.70%	0%	0%	0%	85.30%
2520	USA	UK	JPN	ITA	GER	FRA	CAN
μ	4.51%	7.73%	10.81%	9.67%	10.74%	10.01%	11.67%
σ	17.31%	18.93%	23.91%	19.33%	27.41%	23.81%	17.93%
w_1	−278.63%	−159.50%	52.44%	55.36%	79.64%	44.51%	306.18%
w_2	0%	0%	18.69%	0%	0%	0%	81.31%
6300	USA	UK	JPN	ITA	GER	FRA	CAN
μ	4.06%	5.86%	7.00%	7.05%	9.05%	7.91%	7.55%
σ	17.14%	18.36%	22.48%	18.66%	26.53%	22.96%	17.11%
w_1	−147.49%	−108.20%	40.34%	8.29%	88.43%	49.15%	169.47%
w_2	0%	0%	19.46%	0%	36.04%	0%	44.50%
12600	USA	UK	JPN	ITA	GER	FRA	CAN
μ	2.96%	4.15%	5.11%	4.84%	6.72%	5.70%	4.88%
σ	16.80%	18.00%	22.10%	18.15%	25.83%	23.37%	16.55%
w_1	−76.83%	−60.97%	43.01%	−6.15%	71.30%	35.85%	93.78%
w_2	0%	0%	31.10%	0%	43.28%	0%	25.62%

Table 3.4.: Empirical and predictive moments of yearly discrete returns as well as the corresponding portfolio weights for the constraints $\mathcal{C}_1$ and $\mathcal{C}_2$.

expected returns it makes sense to trust far more in expert knowledge than in time series information. The optimal portfolio in P2 is more diversified than the Markowitz portfolio

on the one hand. On the other hand, in contrast to the MVP, it also takes account for the expected predictive returns and the investor's will to reap the profit.

The optimal portfolios for $\kappa_1 = 12600$ are even more diversified. However, here almost all of the empirical information about the expected returns is lost, since the confidence in the corresponding prior assumption is 10 times higher than the empirical evidence.

3.5. Conclusion

We develop an approach to incorporate the stylized facts of high-frequency financial data and arbitrary prior information into the portfolio optimization process. Our approach is characterized by rather weak assumptions about the underlying stochastic process. Using Gordin's central limit theorem, we are able to approximate the distribution of asset log-returns of long investment horizons by the normal distribution. In order to avoid estimation risk, we rely on the Bayesian framework which allows us to include subjective prior information such as expert knowledge. By using a Markov chain Monte Carlo algorithm we simulate the posterior distribution of the unknown model parameters and after that we calculate the first two moments of the discrete predictive asset returns after the given investment period. In a last step, we perform a standard portfolio optimization using these predictive moments, which incorporate both empirical information contained in the data and subjective prior information of the investor.

We give a practical example to demonstrate the applicability of our approach to real-world problems. For that purpose, we use 7 time series of daily log-returns. For the data generating process, we propose an orthogonal MGARCH model. The investor's subjective prior information about expected asset returns and eigenvalues of the covariance matrix is modeled using a hierarchical approach. The suggested portfolios show that prior assumptions have a substantial impact on the optimal decision. Our portfolios become well-diversified compared to the outcomes of traditional portfolio optimization strategies and reflect the investor's assessment about the market. The computational performance of our algorithm encourages applying our approach to higher-dimensional problems in practice, where both empirical information contained in time series and expert knowledge are available.

Chapter 4.

Linear Statistical Inference for Global and Local Minimum Variance Portfolios

4.1. Motivation

During the past decades traditional portfolio optimization has often been criticized since it does not account for estimation risk (Jorion, 1986, Kalymon, 1971, Klein and Bawa, 1976, Michaud, 1989). At the beginning of modern portfolio theory (Markowitz, 1952) it was usually supposed that the parameters of interest, i.e. the means and (co-)variances of asset returns can be estimated accurately such that estimation errors remain negligible. Although this conjecture might be true for variances and covariances if the sample size is large enough compared to the number of assets, it is not an appropriate simplification for expected asset returns in most practical situations (Chopra and Ziemba, 1993, Kempf and Memmel, 2002, Merton, 1980). Nowadays many portfolio optimization procedures which take the parameter uncertainty into account can be found in the literature (Black and Litterman, 1992, Frost and Savarino, 1986, Herold and Maurer, 2006, Kan and Zhou, 2007, Scherer, 2004).

Consider a d-dimensional random vector $R = (R_1, \ldots, R_d)$ of asset excess returns at the end of some investment horizon. The excess return of an asset corresponds to the asset return minus the risk-free interest rate and in the following I will usually drop the prefix 'excess' for convenience. It is assumed that the vector of asset returns is multivariate normally distributed, i.e. $R \sim \mathcal{N}_d(\mu, \Sigma)$, where μ $(d \times 1)$ is an unknown vector of expected asset returns and Σ $(d \times d)$ is an unknown positive-definite matrix containing their variances and covariances.

The *tangential portfolio* (TP) is defined as the portfolio of risky assets which maximizes the Sharpe ratio (see Figure 4.1), i.e.

$$\underset{(d\times 1)}{w_{\mathrm{TP}}} := \arg\max_{v} \mu' v / \sqrt{v' \Sigma\, v}$$

such that the budget constraint $1'v = 1$ is satisfied. Here $v = (v_1, \ldots, v_d)$ symbolizes a vector of portfolio weights and 1 is a vector of ones or the one scalar, respectively. In the following '$(x_1, \ldots, x_d)$' indicates a d-tuple which is understood to be a d-dimensional column vector, whereas '$[\, x_1 \ \cdots \ x_d \,]$' (without the commas) is a d-dimensional row vector, i.e. $(x_1, \ldots, x_d) \equiv [\, x_1 \ \cdots \ x_d \,]'$.

An (mean-variance) *efficient portfolio* (EP) can be characterized in terms of the typical mean-variance utility function (or, more precisely, certainty equivalent), i.e.

$$\underset{(d\times 1)}{w_{\mathrm{EP}}} := \arg\max_{v} \left(\mu' v - \alpha/2 \cdot v' \Sigma\, v \right)$$

for some risk-aversion parameter $\alpha > 0$. If the EP satisfies the budget constraint, it can be found on the efficient frontier, i.e. the upper part of the hyperbola given in Figure 4.1. Otherwise it is located on the capital market line.

A rather simple alternative to the TP or some other EP is given by the so-called *global minimum variance portfolio* (GMVP). This is defined as

$$\underset{(d\times 1)}{w} := \arg\min_{v} v' \Sigma\, v$$

under the budget constraint $1'v = 1$. The GMVP can be viewed as an EP after setting $\alpha = \infty$. Any portfolio which minimizes the variance of the portfolio return $R'v$ under some *additional* constraints for the portfolio weights will be called *local minimum variance portfolio* (LMVP).

It is well-known that $w_{\mathrm{TP}} = \Sigma^{-1}\mu / (1'\Sigma^{-1}\mu)$ and $w = \Sigma^{-1}1 / (1'\Sigma^{-1}1)$ (a closed-form expression for the LMVP under a set of linear equality constraints for the portfolio weights can be found in Section 4.3.1). The TP strongly depends on the vector μ of expected asset returns and the same holds for the EP if the investor has a relatively low risk aversion (that means if α is small). In contrast, the GMVP as well as any LMVP is not determined by the unknown parameter μ. However, a LMVP in general will be *inefficient* which is shown by Figure 4.1.

The GMVP has been advocated by many authors (Jagannathan and Ma, 2003, Kempf and Memmel, 2006, Ledoit and Wolf, 2003). On the one hand choosing the GMVP is closely

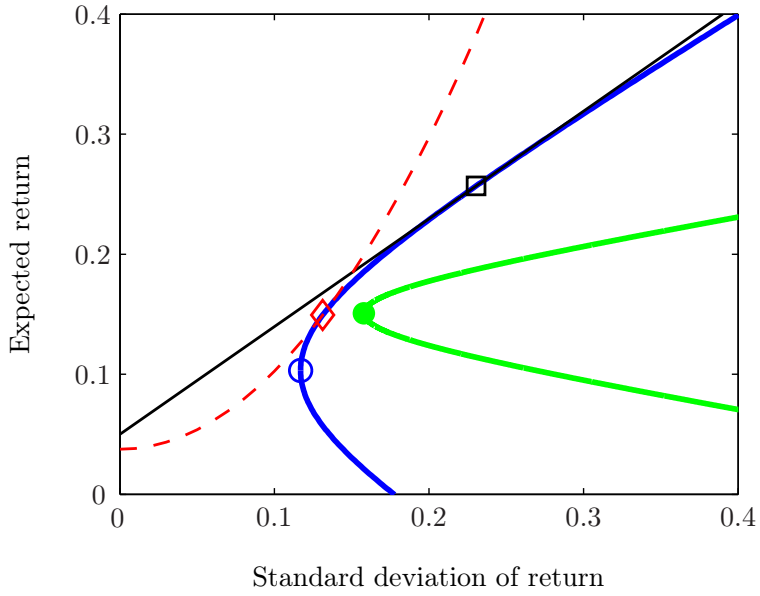

Figure 4.1.: Capital market line (straight), utility isoline (dashed), TP (□), EP (◊), GMVP (○), and LMVP (●).

related to the basic idea of Markowitz (1952), i.e. searching for an efficient portfolio by diversification. On the other hand there are no expected asset returns which have to be estimated and so the impact of estimation errors can be substantially reduced. However, one might ask why it should be appropriate to search for a minimum variance portfolio if the investor is interested in maximizing a mean-variance utility function or the Sharpe ratio according to Tobin's two-fund separation theorem (Tobin, 1958). Thus I would like to explain now the main idea of the present work.

The *suggested* TP can differ substantially from the true one in the presence of estimation risk. Put another way, its *realized* (but not the suggested) Sharpe ratio can be very small since the expected asset returns are unknown and then it might be better to search for some minimum variance portfolio. In particular, the constraints for a LMVP can be chosen in such a way that large volatility assets are preferred (recall that the variances and covariances of asset returns can be much better estimated than their expectations). If some branch contains a larger *risk premium* than another (e.g. the IT sector bears more risk than the finance sector), an investor could be simply willing to reap the profit by choosing the corresponding LMVP. Now this is probably closer to the TP or another EP than the GMVP, although the LMVP is inefficient (see Figure 4.1).

Since there are no expected asset returns which have to be estimated for the LMVP, its

realized Sharpe ratio is hopefully larger than the realized Sharpe ratio of the suggested TP. In fact some authors argue that even if portfolio restrictions are binding (which is indicated e.g. by the small hyperbola in Figure 4.1) they can increase the *out-of-sample* performance (Frost and Savarino, 1988, Jagannathan and Ma, 2003). This is because restricting portfolio weights forces diversification and the investor's decision becomes less vulnerable to estimation risk. Hence, the advertising motto for minimum variance portfolios could be 'A bird in the hand is worth two in the bush'.

Another argument for restricting portfolio weights is that people might have *prior knowledge* apart from empirical data. For instance, investors often believe that some industry sector, region or stock market will 'outperform' another and so they might wish to take the opportunity. Moreover, in many practical situations an investor *must not* choose a mean-variance efficient portfolio. For example, portfolio managers of mutual funds often have to observe certain limits regarding their choice of portfolio weights. This is a typical situation in *top down portfolio management*. That means the set of available assets is divided into some subsets of assets, each subset is divided into some further subsets, etc. These subsets are generally referred to as *asset classes*, according to some industry sector, rating or regional classification. Now, top down portfolio management means that the amount of capital is allocated to the top level partition at first. Given the portfolio weights for that partition, somebody has to choose some optimal portfolio weights for the subsequent asset classes, etc., so that each of the succeeding decisions are limited by the preceding allocations.

As already pointed out by Black and Litterman (1992) as well as Herold and Maurer (2006), combining historical data with 'expert knowledge' (which is usually done in practice) or drawing up some guidelines which must be observed by the decision maker can lead to more reasonable and well-diversified portfolios rather than relying on pure statistical portfolio optimization methods. In this work I will assume that the portfolio weights are generally restricted by a set of linear equality constraints. Thus one might be interested in testing linear hypotheses for the corresponding LMVP rather than the GMVP. I will present standard hypothesis tests for global and local minimum variance portfolios as well as the small-sample distributions of the estimated portfolio weights.

The present work is focused on small-sample rather than large-sample properties but the latter can be easily deduced from the former ones. This is an important issue for I will show that large-sample approximations fail if the sample size is large but the number of

observations relative to the number of assets is small. As already mentioned I will concentrate on linear equality constraints though it is clear that in many practical situations inequality constraints play an important role. However, the statistical properties of portfolio weights satisfying inequality constraints cannot be studied by standard econometric methods (Geweke, 1986, Gouriéroux et al., 1982, Wolak, 1987). Investigating the role of linear inequality constraints is left for future research.

In the next section I recall some standard hypothesis tests for the GMVP. The following section deals with hypothesis tests for local minimum variance portfolios. It is shown that, after a suitable transformation of the data, the corresponding tests follow immediately by applying the results of Section 4.2. In Section 4.4 the joint distribution of the weights of global and local minimum variance portfolios is derived. The first two moments of an unbiased estimator for the expected portfolio return are also presented. Section 4.5 contains an empirical study where the following results are applied to stock market data and Section 4.6 concludes the present work.

4.2. The Global Minimum Variance Portfolio

4.2.1. Theoretical Foundation

Note that $w = \Sigma^{-1}1/(1'\Sigma^{-1}1)$ is a nonlinear function of Σ. However, Kempf and Memmel (2006) noticed that minimizing the variance of the portfolio return can be viewed as a linear regression problem. The return of the GMVP can be written as

$$(1 - w_2 - \ldots - w_d)R_1 + w_2R_2 + \ldots + w_dR_d = \eta + \varepsilon\,, \tag{4.1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. By defining $\beta_1 := \eta$, $\beta_j := w_j$, $\Delta R_j := R_1 - R_j$ for $j = 2, \ldots, d$, and $u := \varepsilon$, Eq. 4.1 becomes equivalent to

$$R_1 = \beta_1 + \beta_2\Delta R_2 + \ldots + \beta_d\Delta R_d + u\,. \tag{4.2}$$

Note that this is a linear regression equation with *stochastic* regressors but the joint normality assumption guarantees that the usual results of econometric theory still hold in this context.

The following proposition is a standard result of linear regression theory. It is crucial for understanding the basic idea of the subsequent derivations and thus it is recalled here for convenience.

Proposition 4.1 *Let $Z = (Z_1, \ldots, Z_d)$ be a d-dimensional random vector with positive-definite covariance matrix. Consider the vector*

$$\underset{(d\times 1)}{\beta} = (\beta_1, \ldots, \beta_d) := \arg\min_b \mathrm{E}\big\{(Z_1 - b_1 - b_2 Z_2 - \ldots - b_d Z_d)^2\big\},$$

where $b = (b_1, \ldots, b_d)$ and define

$$u := Z_1 - \beta_1 - \beta_2 Z_2 - \ldots - \beta_d Z_d\,.$$

The vector β exists and is uniquely defined. More precisely, the subvector $\beta^{\mathrm{s}} := (\beta_2, \ldots, \beta_d)$ is given by

$$\beta^{\mathrm{s}} = \mathrm{Var}(Z^{\mathrm{s}})^{-1}\mathrm{Cov}(Z_1, Z^{\mathrm{s}})\,,$$

where $Z^{\mathrm{s}} := (Z_2, \ldots, Z_d)$, $\mathrm{Var}(Z^{\mathrm{s}})$ $((d-1)\times(d-1))$ is the covariance matrix of Z^{s}, and $\mathrm{Cov}(Z_1, Z^{\mathrm{s}})$ is the $(d-1)\times 1$ vector of covariances between Z_1 and Z_j $(j = 2, \ldots, d)$. Moreover, the parameter β_1 is given by

$$\beta_1 = \mathrm{E}(Z_1) - \mathrm{E}(Z^{\mathrm{s}})'\beta^{\mathrm{s}}$$

and it holds that $\mathrm{E}(u) = 0$ as well as $\mathrm{Cov}(X_j, u) = 0$ for $j = 2, \ldots, d$.

The parameters $\beta_1, \ldots, \beta_d$ in Eq. 4.2 are chosen in such a way that $\mathrm{E}(u) = 0$ holds and $\mathrm{Var}(u) = E(u^2)$ is minimal, i.e. $\mathrm{Cov}(\Delta R_j, u) = 0$ $(j = 2, \ldots, d)$. So it has been shown that Eq. 4.2 indeed is a proper linear regression equation satisfying the standard assumptions of linear regression theory, especially the *strict exogeneity assumption* (Hayashi, 2000, p. 7). For that reason it is possible to develop several exact hypothesis tests for the GMVP by standard methods of econometrics (cf. Kempf and Memmel, 2006).

The next corollary states that the converse of Proposition 4.1 is true.

Corollary 4.2 *Let $Z = (Z_1, \ldots, Z_d)$ be a d-dimensional random vector with positive-definite covariance matrix. Search for some numbers $b_1, \ldots, b_d$ such that $\mathrm{E}(u^*) = 0$ and $\mathrm{Cov}(Z_j, u^*) = 0$ for $j = 2, \ldots, d$, where*

$$u^* := Z_1 - b_1 - b_2 Z_2 - \ldots - b_d Z_d\,.$$

The vector $b = (b_1, \ldots, b_d)$ exists and is uniquely defined by $b = \beta$ where β is given by Proposition 4.1.

The proof of that corollary follows immediately from the proof of Proposition 4.1 (see the appendix) and noting that the linear equation

$$0 = \mathrm{Cov}(Z^{\mathrm{s}}, u^*) = \mathrm{Cov}(Z_1, Z^{\mathrm{s}}) - \mathrm{Var}(Z^{\mathrm{s}})b^{\mathrm{s}}$$

has a unique solution (due to the positive definiteness of $\mathrm{Var}(Z^{\mathrm{s}})$). Corollary 4.2 implies that the strict exogeneity assumption is satisfied only if the error u has minimum variance. Later on it is shown that for that reason the standard test statistics for the GMVP in general must not be applied for testing a LMVP.

4.2.2. Statistical Inference

Of course, in practice the weights of the GMVP are unknown, i.e. they have to be estimated from historical data. Let

$$\underset{(n\times d)}{\mathbf{R}} := \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1d} \\ R_{21} & R_{22} & \cdots & R_{2d} \\ \vdots & \vdots & & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{nd} \end{bmatrix}$$

be a sample of $n > d$ independent copies of R. Now define

$$\underset{(n\times d)}{\mathbf{X}} := \begin{bmatrix} 1 & X_{12} & \cdots & X_{1d} \\ 1 & X_{22} & \cdots & X_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n2} & \cdots & X_{nd} \end{bmatrix},$$

where $X_{ij} := R_{i1} - R_{ij}$ $(i = 1, \ldots, n,\ j = 2, \ldots, d)$ and $\mathbf{Y} := (Y_1, \ldots, Y_n)$ $(n \times 1)$ with $Y_i := R_{i1}$ $(i = 1, \ldots, n)$. Similarly, I will also write $X := (1, X_2, \ldots, X_d)$ $(d \times 1)$, $X^{\mathrm{s}} := (X_2, \ldots, X_d)$ $((d-1) \times 1)$, and $Y \equiv R_1$ (1×1).

According to the standard notation of linear regression theory the linear model represented by Eq. 4.2 is given by

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}\,,$$

where $\beta = (\beta_1, \ldots, \beta_d)$ $(d \times 1)$ contains the weights $\beta_2, \ldots, \beta_d$ of the GMVP – except for the first one – as well as the expected return β_1 of the GMVP. Here $\mathbf{u} := (u_1, \ldots, u_n)$ is an $n \times 1$ vector of unobservable errors. Hence, the *ordinary least squares* (OLS) estimator for β can be calculated by

$$\hat{\beta}_{\mathrm{OLS}} = \big(\hat{\eta}, \hat{w}_2, \ldots, \hat{w}_d\big) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

In fact the weights of the GMVP can be estimated by

$$\hat{w}^{\mathrm{s}} := \left(\hat{w}_2, \ldots, \hat{w}_d\right) = \widehat{\Omega}^{-1}\hat{\omega}\,,$$

where $\widehat{\Omega}$ is the sample covariance matrix of X^{s} and $\hat{\omega}$ is the $(d-1)\times 1$ vector of the sample covariances between Y and X_j $(j = 2, \ldots, d)$. The random vector

$$\hat{w} := \left(1 - 1'\hat{w}^{\mathrm{s}}, \hat{w}^{\mathrm{s}}\right)$$

is the best unbiased estimator for the GMVP in the context of normally distributed asset returns (Kempf and Memmel, 2006). Note that if the normal distribution assumption for the asset returns is dropped, in general it cannot be guaranteed that the standard assumptions of linear regression theory are satisfied and thus $\hat{w}$ might become inefficient.

Kempf and Memmel (2006) showed that $\hat{w} = \widehat{\Sigma}^{-1}1/(1'\widehat{\Sigma}^{-1}1)$, i.e. $\hat{w}$ corresponds to the traditional GMVP estimator, where the $d \times d$ matrix

$$\widehat{\Sigma} := \mathbf{R}'\mathbf{R}/n - \overline{\mathbf{r}}\,\overline{\mathbf{r}}'$$

represents the sample covariance matrix and $\overline{\mathbf{r}} := \mathbf{R}'1/n$ $(d \times 1)$ is the sample mean vector of R. Further, also the OLS estimator for the expected GMVP return corresponds to its traditional estimator, i.e. $\hat{\eta} = \overline{\mathbf{r}}'\hat{w}$.

The relation between the OLS estimator $\hat{\beta}_{\mathrm{OLS}}$ and the residual vector $\hat{\mathbf{u}}$ $(n \times 1)$ can be represented by

$$R_1 = \hat{\eta} + \hat{w}_2\Delta R_2 + \ldots + \hat{w}_d\Delta R_d + \hat{u}$$

or – according to the usual notation of linear regression theory – as

$$\mathbf{Y} = \mathbf{X}\hat{\beta}_{\mathrm{OLS}} + \hat{\mathbf{u}}\,.$$

Let $\hat{\sigma}^2_{\mathrm{OLS}} := \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-d)$ be the unbiased OLS estimator for σ^2. It holds that

$$\hat{\sigma}^2 := \hat{w}'\widehat{\Sigma}\,\hat{w} = 1/\left(1'\widehat{\Sigma}^{-1}1\right) = \frac{n-d}{n}\cdot\hat{\sigma}^2_{\mathrm{OLS}}\,,$$

where $\hat{\sigma}^2$ is the traditional estimator for the variance of the GMVP return.

Now consider the fundamental least squares problem

$$(\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b) \rightarrow \min_b! \tag{4.3}$$

under the additional constraint $Hb = h$, where H $(q \times d)$ is a matrix with $\operatorname{rk} H = q \leq d$ and h $(q \times 1)$ some arbitrary vector. The solution of this minimization problem is given by the *restricted least squares* (RLS) estimator

$$\hat{\beta}_{\mathrm{RLS}} := \arg\min_b (\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b), \qquad \text{s.t. } Hb = h. \tag{4.4}$$

In the following I will write $\hat{\beta}_{\mathrm{RLS}} = (\hat{\eta}^*, \hat{w}_2^*, \ldots, \hat{w}_d^*)$ and correspondingly

$$R_1 = \hat{\eta}^* + \hat{w}_2^* \Delta R_2 + \ldots + \hat{w}_d^* \Delta R_d + \hat{u}^* \tag{4.5}$$

or more compactly

$$\mathbf{Y} = \mathbf{X}\hat{\beta}_{\mathrm{RLS}} + \hat{\mathbf{u}}^*$$

to indicate that $\hat{\mathbf{u}}^*$ $(n \times 1)$ is the residual vector with respect to the RLS estimator and not to the OLS estimator. The RLS estimator can be calculated explicitly by applying the Lagrange method (Greene, 2003, p. 100). However, in Section 4.3.2 I will present an alternative method which is more useful in the context of portfolio optimization.

Here only inhomogeneous regressions are taken into consideration and so both $\hat{\mathbf{u}}$ and $\hat{\mathbf{u}}^*$ have zero means. That is to say (4.3) indeed leads to the local minimum *variance* portfolio satisfying the given restriction $Hb = h$. However, in contrast to the unrestricted case, each column of $\mathbf{X}$ is correlated with $\hat{\mathbf{u}}^*$ in general. More precisely, $\mathbf{X}'\hat{\mathbf{u}}^* \neq 0$ if the linear restrictions are binding. This is an empirical consequence of Corollary 4.2. In the following I will write

$$\hat{w}^* := \left(1 - 1'\hat{w}^{*\mathrm{s}}, \hat{w}^{*\mathrm{s}}\right), \tag{4.6}$$

where $\hat{w}^{*\mathrm{s}} := (\hat{w}_2^*, \ldots, \hat{w}_d^*)$.

An exact or, say, small-sample hypothesis test against $H_0 : H\beta = h$ is given by the next theorem. For an alternative representation of that F-test and some applications to financial data see Kempf and Memmel (2006).

Theorem 4.3 *Let $\hat{w}$ be the traditional estimator for the GMVP $w = (w_1, \ldots, w_d)$ and $\hat{w}^*$ the RLS estimator given by Eq. 4.6. Further, let η be the expected return of the GMVP. If $H\beta = h$ with $\beta = (\eta, w_2, \ldots, w_d)$ it holds that*

$$\frac{n-d}{q} \cdot \frac{(\hat{w} - \hat{w}^*)'\widehat{\Sigma}(\hat{w} - \hat{w}^*)}{\hat{\sigma}^2} \sim F_{q,n-d},$$

where $\hat{\sigma}^2$ denotes the traditional estimator for the variance of the GMVP return.

A similar F-test for the TP (or any other efficient portfolio which is proportional to the TP) has been obtained by Britten-Jones (1999). The result given in Theorem 4.3 does not follow from this F-test since Britten-Jones requires the existence of a risk-free asset and the considered portfolios always lie on the capital market line but not on the efficient frontier.

Another important hypothesis is given by $H_0 : \sigma^2 \geq \sigma_0^2$ (for some $\sigma_0^2 > 0$) which can be tested by the next theorem (cf. Kempf and Memmel, 2006).

Theorem 4.4 *Consider the traditional estimator $\hat{\sigma}^2$ for the variance σ^2 of the GMVP return. It holds that*

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-d}\,.$$

This is a standard result from linear regression theory (Greene, 2003, p. 50) after noting that $\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/n$ and so the proof can be skipped. The parameter uncertainty concerning the variance σ^2 of the GMVP return can be quantified by $\sigma^2 \,|\, \hat{\sigma}^2 \sim \hat{\sigma}^2 n/\chi^2_{n-d}$ either from a *fiducial* (Rao, 1965, Section 5b.5) or *Bayesian* perspective (by using Jeffreys' prior distribution for σ^2), where the estimate $\hat{\sigma}^2$ is considered as fixed. Since $\mathrm{E}(n/\chi^2_{n-d}) = n/(n-d-2)$, it follows that

$$\mathrm{E}\big(\sigma^2 \,|\, \hat{\sigma}^2\big) \approx \frac{\hat{\sigma}^2}{1 - 1/Q}\,,$$

with $Q := n/d > 1$, i.e. the estimation risk essentially depends on the sample size relative to the number of assets. Hence, the capital market is said to be *high-dimensional* if Q – which can be interpreted as the *effective* size of a multivariate sample – is small. In that case small-sample inference must be applied even if the number of observations is large.

Usually an investor not only wants to know whether the variance of the GMVP is bounded by some number σ_0^2 but also to test against $H_0 : \eta \leq \eta_0$, where η represents the true expected return of the GMVP. This can be done by applying the next theorem.

Theorem 4.5 *Consider the traditional estimators $\hat{\eta}$ for the expected GMVP return η and $\hat{\sigma}^2$ for the variance of the GMVP return. It holds that*

$$\frac{\hat{\eta} - \eta}{\sqrt{\big\{\hat{\sigma}^2(1 + \overline{\mathbf{r}}\,'\widehat{\Sigma}^{-1}\overline{\mathbf{r}}) - \hat{\eta}^2\big\}/(n-d)}} \sim t\,(n-d)\,,$$

where $t(n-d)$ denotes Student's t-distribution with $n-d$ degrees of freedom.

The latter theorem completes the repertoire of standard hypothesis tests for the GMVP. In the next section it is shown that the same repertoire can be used also for local minimum variance portfolios after a suitable transformation of the data.

4.3. Local Minimum Variance Portfolios

4.3.1. Theoretical Foundation

Consider the LMVP

$$\underset{(d\times 1)}{w^*} = \left(w_1^*, \ldots, w_d^*\right) := \arg\min_v \operatorname{Var}(R'v)\,, \qquad \text{s.t. } Fv = f\,, \tag{4.7}$$

where the budget constraint $1'v = 1$ is also satisfied. Here f is a $q \times 1$ vector and F is a $q \times d$ matrix ($q < d$) such that the stacked $(q+1) \times d$ matrix $(1', F)$ has rank $q+1$. Both f and F are assumed to be non-random. Using the definitions from above this can be formulated as a least squares problem, i.e.

$$\underset{(d\times 1)}{\beta^*} := \arg\min_b \mathrm{E}\left\{(Y - X'b)^2\right\} \tag{4.8}$$

under a set of linear restrictions affecting only the parameters $b_2, \ldots, b_d$ (i.e. the portfolio weights without the first one). However, due to Corollary 4.2 this would not lead to a proper linear regression equation, say

$$R_1 = \beta_1^* + \beta_2^* \Delta R_2 + \ldots + \beta_d^* \Delta R_d + u^*\,, \tag{4.9}$$

since u^* generally depends on the regressors $\Delta R_2, \ldots, \Delta R_d$. So the standard test statistics which have been provided in Section 4.2.2 cannot be applied. However, in the following it will be shown how to reformulate (4.8) such that the standard hypothesis tests become applicable.

Consider a matrix $\mathcal{T}$ $(d \times (d-q))$ such that

$$\begin{bmatrix} 1' \\ F \end{bmatrix} \mathcal{T} = \begin{bmatrix} 1' \\ f1' \end{bmatrix}\,.$$

Then the condition $F\mathcal{T}v = f$ is satisfied for any vector $v \in \mathbb{R}^{d-q}$ with $1'v = 1$. Moreover, it is guaranteed that $1'\mathcal{T}v = 1$, i.e. the budget constraint holds also for $\mathcal{T}v \in \mathbb{R}^d$. Now the LMVP can be simply found by searching for the GMVP with respect to the *transformed* asset return vector

$$R^* = (R_1^*, \ldots, R_{d-q}^*) := \mathcal{T}'R\,.$$

Hence, the least squares problem given by (4.8) can be reformulated as

$$\underset{((d-q)\times 1)}{\alpha} := \arg\min_{a} \mathrm{E}\big\{(Y^* - X^{*\prime}a)^2\big\}\,.$$

Here $Y^* := R_1^*$ and $X^* := (1, X_2^*, \ldots, X_{d-q}^*)$ with $X_j^* := R_1^* - R_j^*$ for $j = 2, \ldots, d-q$. The corresponding modified linear model

$$R_1^* = \alpha_1 + \alpha_2 \Delta R_2^* + \ldots + \alpha_{d-q} \Delta R_{d-q}^* + u^* \tag{4.10}$$

is quite similar to the linear regression equation 4.9. However, the vector α can be chosen *without any restriction* from $\mathbb{R}^{d-q}$ so that $\mathrm{Var}(u^*)$ becomes minimal and it is always guaranteed that the condition $Fw^* = f$ is satisfied after the re-parameterization

$$w^* := \mathcal{T}(1 - 1'\alpha^{\mathrm{s}}, \alpha^{\mathrm{s}})\,,$$

where $\alpha^{\mathrm{s}} := (\alpha_2, \ldots, \alpha_{d-q})$. Eq. 4.10 in fact represents a proper linear regression equation, i.e. $\mathrm{E}(u^*) = 0$ and $\mathrm{Cov}(X_j^*, u^*) = 0$ for $j = 2, \ldots, d-q$.

The LMVP is given by

$$w^* = \frac{\mathcal{T}(\mathcal{T}'\Sigma\,\mathcal{T})^{-1}1}{1'(\mathcal{T}'\Sigma\,\mathcal{T})^{-1}1}$$

and the quantity $\mathcal{T}$ can be derived as follows. Assume that the $(q+1) \times d$ matrix

$$\overline{F} := \begin{bmatrix} 1' \\ F \end{bmatrix} = \begin{bmatrix} \overline{F}_1 & \overline{F}_2 \end{bmatrix}$$

is structured in such a way that $\overline{F}_1$ is a nonsingular $(q+1) \times (q+1)$ matrix and $\overline{F}_2$ is a $(q+1) \times (d-q-1)$ matrix. A structure like this can be always found by a permutation of the columns of F since this has full row rank. Similarly, consider the partition

$$\mathcal{T} = \begin{bmatrix} \mathcal{T}_1 \\ \mathcal{T}_2 \end{bmatrix},$$

where $\mathcal{T}_1$ is a $(q+1) \times (d-q)$ and $\mathcal{T}_2$ is a $(d-q-1) \times (d-q)$ matrix.

Recall that $\mathcal{T}$ has to be such that $\overline{F}\,\mathcal{T} = (1', f1')$. In the following let

$$\mathcal{T}_2 = \begin{bmatrix} 0 & I_{d-q-1} \end{bmatrix} \tag{4.11}$$

so that

$$\overline{F}\,\mathcal{T} = \overline{F}_1\mathcal{T}_1 + \begin{bmatrix} 0 & \overline{F}_2 \end{bmatrix} = \begin{bmatrix} 1' \\ f1' \end{bmatrix}.$$

That means

$$\mathcal{T}_1 = \overline{F}_1^{-1}\left(\begin{bmatrix} 1' \\ f1' \end{bmatrix} - \begin{bmatrix} 0 & \overline{F}_2 \end{bmatrix}\right). \tag{4.12}$$

Note that for the special case $\overline{F} = 1'$, i.e. if there is no additional restriction at all, it holds that $\mathcal{T} = I_d$.

4.3.2. Statistical Inference

In Section 4.2.2 the minimization problem given by Eq. 4.4 has been considered, which involves the expected return estimate $\hat{\beta}_{\mathrm{RLS},1} = \hat{\eta}^* = \overline{\mathbf{r}}\,'\hat{w}^*$. Note that the $q \times d$ matrix H refers to the expected GMVP return β_1 and the GMVP weights *without* the first one. However, in practical situations linear constraints possibly involve the first portfolio weight by considering the vector $w^+ := (\eta, w_1, \ldots, w_d)$. That means the null hypothesis is given by $H_0 \colon Gw^+ = g$ where G is a $q \times (d+1)$ matrix with $\operatorname{rk} G = q$ and g is an arbitrary $q \times 1$ vector. In fact, in that case the LMVP w^* defined by Eq. 4.7 has to be found under the budget constraint $1'v = 1$ and

$$G \begin{bmatrix} \overline{\mathbf{r}}\,' \\ I_d \end{bmatrix} v = g\,.$$

That means (4.4) can be solved in the same manner as (4.7) if the sample mean vector $\overline{\mathbf{r}}$ is included in the linear constraint $Fv = f$. Thus any *Markowitz portfolio*

$$\underset{(d\times 1)}{w_{\mathrm{M}}} = \arg\min_v v'\Sigma\, v\,, \qquad \text{s.t. } \mu' v = \eta_0$$

can be represented as a GMVP after a suitable transformation of the data. However, since in that case the linear constraint is stochastic, the presented methods of statistical inference cannot be applied.

Due to the preceding theoretical arguments the parameter vector α can be readily estimated by the OLS estimator

$$\hat{\alpha}_{\mathrm{OLS}} = \left(\hat{\alpha}_{\mathrm{OLS},1}, \ldots, \hat{\alpha}_{\mathrm{OLS},d-q}\right) := (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*, \tag{4.13}$$

where

$$\underset{(n\times(d-q))}{\mathbf{X}^*} := \begin{bmatrix} 1 & X^*_{12} & \cdots & X^*_{1,d-q} \\ 1 & X^*_{22} & \cdots & X^*_{2,d-q} \\ \vdots & \vdots & & \vdots \\ 1 & X^*_{n2} & \cdots & X^*_{n,d-q} \end{bmatrix} \tag{4.14}$$

and $\mathbf{Y}^* := (Y_1^*, \ldots, Y_n^*)$ $(n \times 1)$.

The relationship between the residual vector $\hat{\mathbf{u}}^*$ $(n \times 1)$ and the OLS estimator $\hat{\alpha}_{\mathrm{OLS}}$ can be represented by

$$\mathbf{Y}^* = \mathbf{X}^* \hat{\alpha}_{\mathrm{OLS}} + \hat{\mathbf{u}}^* \,.$$

After defining $\hat{\alpha}^{\mathrm{s}}_{\mathrm{OLS}} := (\hat{\alpha}_{\mathrm{OLS},2}, \ldots, \hat{\alpha}_{\mathrm{OLS},d-q})$, the OLS estimator for w^* corresponds to

$$\hat{w}^* := \mathcal{T}\big(1 - 1'\hat{\alpha}^{\mathrm{s}}_{\mathrm{OLS}}, \hat{\alpha}^{\mathrm{s}}_{\mathrm{OLS}}\big) \tag{4.15}$$

and $\hat{\alpha}_{\mathrm{OLS},1} = \hat{\eta}^*$ is the estimator for the expected return of the LMVP. Hence, $\hat{w}^*$ turns out to be the best unbiased estimator for the corresponding LMVP.

Any null hypothesis concerning the *local* minimum variance portfolio can be implemented in the same way as described at the beginning of this section. Let $w^{*+} := (\eta^*, w_1^*, \ldots, w_d^*)$ be the parameter vector of the LMVP and consider the null hypothesis $H_0^* : Cw^{*+} = c$, where C is some $p \times (d+1)$ matrix with $\operatorname{rk} C = p \leq d - q$ and c is an arbitrary $p \times 1$ vector. This is similar to the null hypothesis $H_0 : Gw^+ = g$. However, for H_0^* there are only $d - q$ degrees of freedom left since the LMVP has been already characterized by q linear restrictions. Of course it has also to be guaranteed that H_0^* does not imply the linear restrictions of the LMVP or the budget constraint. More precisely, consider the linear system of equations

$$\begin{bmatrix} 0 & 1' \\ {\scriptstyle (1\times 1)} & {\scriptstyle (1\times d)} \\ 0 & F \\ {\scriptstyle (q\times 1)} & {\scriptstyle (q\times d)} \\ C_1 & C_2 \\ {\scriptstyle (p\times 1)} & {\scriptstyle (p\times d)} \end{bmatrix} \begin{bmatrix} \eta^* \\ w_1^* \\ \vdots \\ w_d^* \end{bmatrix} = \begin{bmatrix} 1 \\ {\scriptstyle (1\times 1)} \\ f \\ {\scriptstyle (q\times 1)} \\ c \\ {\scriptstyle (p\times 1)} \end{bmatrix}$$

with $p + q \leq d$. Now it has to be guaranteed that the $(p + q + 1) \times (d + 1)$ matrix on the left hand side possesses full row rank.

The restricted minimum variance portfolio according to H_0^* is denoted by $\hat{w}^{**}$ and can be calculated as described for the null hypothesis H_0 without using the Lagrange method. Moreover, the standard hypothesis tests derived in Section 4.2.2 can be applied to local minimum variance portfolios just by transforming the asset returns $R_1, \ldots, R_d$ into the portfolio returns $R_1^*, \ldots, R_{d-q}^*$. Then it holds that

$$\frac{n - d + q}{p} \cdot \frac{(\hat{w}^* - \hat{w}^{**})' \widehat{\Sigma} \, (\hat{w}^* - \hat{w}^{**})}{\hat{\sigma}^{*2}} \sim F_{p,n-d+q} \,, \tag{4.16}$$

provided H_0^* is not binding, as well as

$$\frac{n\hat{\sigma}^{*2}}{\sigma^{*2}} \sim \chi^2_{n-d+q}$$

and

$$\frac{\hat{\eta}^* - \eta^*}{\sqrt{\left\{\hat{\sigma}^{*2}(1 + \overline{\mathbf{r}}'\mathcal{T}(\mathcal{T}'\widehat{\Sigma}\,\mathcal{T})^{-1}\mathcal{T}'\,\overline{\mathbf{r}}) - \hat{\eta}^{*2}\right\}/(n-d+q)}} \sim t\,(n-d+q)\,.$$

That means

1. the F-distribution given in Theorem 4.3,
2. the χ^2-distribution from Theorem 4.4, and
3. the t-distribution presented in Theorem 4.5

simply capture q additional degrees of freedom, where q is the number of linear equalities characterizing the LMVP. Hence, imposing linear restrictions is a simple dimension reduction technique which reduces the parameter uncertainty of portfolio optimization. A similar effect can be also observed for linear inequality constraints like setting upper bounds for the portfolio weights or using short-selling constraints. This is confirmed by several simulation and out-of-sample studies (Eichhorn et al., 1998, Frost and Savarino, 1988, Grauer and Shen, 2000, Jagannathan and Ma, 2003).

It is worth to point out that the GMVP as well as any LMVP can exhibit large positive or negative weights which are not caused by estimation errors. Asset returns in general are dominated by a large principal component representing the *market* or *systematic* risk. There often exist some assets – typically belonging to the finance sector – which strongly depend on the market risk and have a relatively small amount of idiosyncratic risk. In that case extreme negative portfolio weights occur as a matter of principle (Green and Hollifield, 1992). Thus, placing short-selling constraints on the portfolio weights can increase the out-of-sample variance of the portfolio return. Of course, this holds also if linear equality constraints are considered. Nevertheless, Jagannathan and Ma (2003) argue that the negative effect of restricting portfolio weights is usually outweighed by the positive effect of reducing estimation risk. This question will be treated analytically in a different paper.

4.4. Distribution of the Estimated Portfolio Weights

In the following section I will concentrate on the small-sample distribution of the estimated weights of global and local minimum variance portfolios. This is only loosely connected to hypothesis testing but the small-sample distribution of the estimated portfolio weights might be of interest in its own right.

4.4.1. Preliminary Definitions

For the sake of simplicity from now on I will ignore the standard notation of linear regression theory. Recall that $\hat{w}$ denotes the estimator for the GMVP whereas $\hat{w}^*$ is the estimator for some LMVP. Correspondingly, w symbolizes the true GMVP and w^* is the true LMVP. The expected return of the GMVP is denoted by η whereas the expected return of the LMVP is given by η^*. Moreover, σ^2 is the variance of the GMVP return whereas σ^{*2} symbolizes the variance of the LMVP return. The corresponding traditional estimators for these quantities are given by $\hat{\eta}$, $\hat{\eta}^*$, $\hat{\sigma}^2$, and $\hat{\sigma}^{*2}$.

In the following $t_k(a, B, \nu)$ (where $t(\cdot) \equiv t_1(\cdot)$) stands for the k-variate t-distribution with $\nu > 0$ degrees of freedom, location vector a $(k \times 1)$, and positive-semidefinite dispersion matrix B $(k \times k)$, i.e.

$$a + \frac{\zeta}{\sqrt{\chi^2_\nu / \nu}} \sim t_k(a, B, \nu)\,,$$

where $\zeta \sim \mathcal{N}_k(0, B)$ is stochastically independent of χ^2_ν. Here $\zeta \sim B^{1/2}\xi$ with $\xi \sim \mathcal{N}_k(0, I_k)$ and $B^{1/2}$ is some matrix such that $B^{1/2}B^{1/2\prime} = B$.

By defining the $(d-1) \times d$ matrix $\Delta := \begin{bmatrix} 1 & -I_{d-1} \end{bmatrix}$ it follows that $\Delta R = X^{\mathrm{s}}$ and thus $\Omega := \Delta\Sigma\Delta'$ denotes the covariance matrix of X^{s}. Analogously, in the context of local minimum variance portfolios the notation $R^* = \mathcal{T}'R$ and $\Delta R^* = X^{*\mathrm{s}}$ will be used. Further, $\Omega^* := \Delta\Sigma^*\Delta'$ is the covariance matrix of $X^{*\mathrm{s}}$, where $\Sigma^* := \mathcal{T}'\Sigma\mathcal{T}$ denotes the covariance matrix of R^*.

4.4.2. Global Minimum Variance Portfolio

The next theorem provides the small-sample distribution of the traditional estimator for the GMVP. Another variant of this theorem can be found in Okhrin and Schmid (2006) and so the proof is skipped.

Theorem 4.6 *Let $w = (w_1, \ldots, w_d)$ be the GMVP of d assets and $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_d)$ the corresponding traditional estimator given a sample of asset returns with size $n \geq d$. It holds that*

$$(\hat{w}_2, \ldots, \hat{w}_d) \sim t_{d-1}\Big((w_2, \ldots, w_d), \frac{\sigma^2}{n-d+1} \cdot \Omega^{-1}, n-d+1\Big),$$

where Ω is the covariance matrix of ΔR and $\sigma^2 = w'\Sigma w$ is the variance of the GMVP return.

An unbiased estimator for the covariance matrix of $\hat{w}^{\mathrm{s}} = (\hat{w}_2, \ldots, \hat{w}_d)$ is provided by the next corollary.

Corollary 4.7 *Consider a sample of asset returns with size $n \geq d+2$ and let $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_d)$ be the traditional estimator for the GMVP. Then the matrix*

$$\widehat{\mathrm{Var}}\big\{(\hat{w}_2, \ldots, \hat{w}_d)\big\} := \frac{\hat{\sigma}^2}{n-d} \cdot \widehat{\Omega}^{-1}$$

is an unbiased estimator for the covariance matrix of $\hat{w}^{\mathrm{s}} = (\hat{w}_2, \ldots, \hat{w}_d)$, where $\widehat{\Omega}$ is the sample covariance matrix of ΔR and $\hat{\sigma}^2$ is the traditional estimator for the variance of the GMVP return.

Note that $\hat{w}_1 = 1 - 1'\hat{w}^{\mathrm{s}}$ and from Theorem 4.6 it follows that the GMVP estimator $\hat{w}$ is t-distributed with mean w, dispersion matrix $\sigma^2\Delta'\Omega^{-1}\Delta/(n-d+1)$, and $n-d+1$ degrees of freedom. From Proposition 1 of Okhrin and Schmid (2006) it follows that $\sigma^2\Delta'\Omega^{-1}\Delta = \sigma^2\Sigma^{-1} - ww'$ and thus

$$\hat{w} \sim t_d\Big(w, \big(\sigma^2\Sigma^{-1} - ww'\big)/(n-d+1), n-d+1\Big).$$

Moreover, Corollary 4.7 implies that

$$\widehat{\mathrm{Var}}(\hat{w}) := \big(\hat{\sigma}^2\widehat{\Sigma}^{-1} - \hat{w}\hat{w}'\big)/(n-d) \tag{4.17}$$

is an unbiased estimator for the covariance matrix of $\hat{w}$.

A stochastic representation for $\hat{\eta}$, i.e. the traditional estimator for the expected return of the GMVP could be found after some calculation. However, this is cumbersome and not useful for econometric purposes. In contrast, the first two moments of the distribution of $\hat{\eta}$ can be easily derived. First of all recall that $\overline{\mathbf{r}}$ and $\widehat{\Sigma}$ are stochastically independent. Thus

$$\mathrm{E}(\hat{\eta}) = \mathrm{E}\big\{\mathrm{E}\big(\overline{\mathbf{r}}'\hat{w} \,|\, \widehat{\Sigma}\big)\big\} = \mathrm{E}(\mu'\hat{w}) = \mu'w = \eta\,.$$

Further, it holds that

$$\begin{aligned}\mathrm{Var}(\hat{\eta}) &= \mathrm{E}\big\{\mathrm{Var}\big(\overline{\mathbf{r}}'\hat{w} \,|\, \widehat{\Sigma}\big)\big\} + \mathrm{Var}\big\{\mathrm{E}\big(\overline{\mathbf{r}}'\hat{w} \,|\, \widehat{\Sigma}\big)\big\} \\ &= \mathrm{E}(\hat{w}'\Sigma\,\hat{w}/n) + \mu'\mathrm{Var}(\hat{w})\mu\,,\end{aligned}$$

and after some calculation it follows from Theorem 4.6 that

$$\mathrm{E}(\hat{w}'\Sigma\,\hat{w}) = \frac{n-2}{n-d-1}\cdot\sigma^2\,.$$

That means if $n \geq d+2\,$,

$$\mathrm{Var}(\hat{\eta}) = \mu'\mathrm{Var}(\hat{w})\mu + \frac{n-2}{n-d-1}\cdot\frac{\sigma^2}{n}\,,$$

where

$$\mathrm{Var}(\hat{w}) = (\sigma^2\Sigma^{-1} - ww')/(n-d-1)\,.$$

Note that σ^2/n is the variance of $\overline{\mathbf{r}}'w$, i.e. the variance of the expected GMVP return if w would be known but the expected asset returns $\mu_1,\ldots,\mu_d$ unknown. That means the estimation risk concerning the expected GMVP return can be decomposed into two parts, viz.

1. one part carrying the estimation risk of the portfolio weights and
2. another part for the estimation risk concerning the expected returns.

More precisely, the variance of $\hat{\eta}$ is an affine-linear transformation of $\sigma^2/n\,$, where $(n-2)/(n-d-1) \geq 1$ and $\mu'\mathrm{Var}(\hat{w})\mu \geq 0\,$.

4.4.3. Local Minimum Variance Portfolios

From the previous discussion it is clear that any LMVP can be found in the same manner as the GMVP after transforming the asset return vector R into the portfolio return vector R^*. Recall that the LMVP estimator $\hat{w}^*$ can be written as $\hat{w}^* = \mathcal{T}(1 - 1'\hat{\alpha}^{\mathrm{s}}_{\mathrm{OLS}}, \hat{\alpha}^{\mathrm{s}}_{\mathrm{OLS}})$ (see Section 4.3.2), where

$$\hat{\alpha}^{\mathrm{s}}_{\mathrm{OLS}} \sim t_{d-q-1}\Big(\alpha^{\mathrm{s}}, \frac{\sigma^{*2}}{n-d+q+1}\cdot\Omega^{*-1}, n-d+q+1\Big)\,.$$

Thus it holds that

$$\hat{w}^* \sim t_d\Big(w^*, \big(\sigma^{*2}\mathcal{T}\Sigma^{*-1}\mathcal{T}' - w^*w^{*\prime}\big)/(n-d+q+1), n-d+q+1\Big)\,.$$

Similarly, the remaining assertions follow from the theorems and corollaries already derived for the GMVP, simply by substituting d by $d-q$, η (or $\hat{\eta}$) by η^* (or $\hat{\eta}^*$), and σ^2 (or $\hat{\sigma}^2$) by σ^{*2} (or $\hat{\sigma}^{*2}$). For example, according to Eq. 4.17 it follows that

$$\widehat{\mathrm{Var}}(\hat{w}^*) := \big(\hat{\sigma}^{*2}\mathcal{T}\widehat{\Sigma}^{*-1}\mathcal{T}' - \hat{w}^*\hat{w}^{*\prime}\big)/(n-d+q)$$

is an unbiased estimator for the covariance matrix of $\hat{w}^*$. Moreover, $\mathrm{E}(\hat{\eta}^*) = \eta^*$ and

$$\mathrm{Var}(\hat{\eta}^*) = \mu'\mathrm{Var}(\hat{w}^*)\mu + \frac{n-2}{n-d+q-1}\cdot\frac{\sigma^{*2}}{n}\,,$$

where

$$\mathrm{Var}(\hat{w}^*) = \big(\sigma^{*2}\mathcal{T}\Sigma^{*-1}\mathcal{T}' - w^*w^{*\prime}\big)/(n-d+q-1)\,.$$

4.5. Empirical Study

The following empirical study is based on daily asset prices between 1980-01-01 and 2003-11-26 of the 500 stocks listed by the S&P 500 stock index on 26$^{\text{th}}$ November 2003. The data have been kindly provided by Thomson Financial Datastream and the considered asset prices are adjusted for dividends, splits, etc. However, only for 285 stocks the asset prices are available over the whole sample period. The residual 215 time series exhibit missing values caused by IPO's or M&A's during the sample period and are not considered in this study. Moreover, 274 firms could be found to belong to one of 10 industry sectors according to S&P's *Global Industry Classification Standard* (GICS). The other 11 stocks have been also removed from the study.

The risk-free interest rate is calculated by the secondary market 3-month US treasury bill rate (p.a.). The investment period is supposed to be 21 days (i.e. one trading month) and so the corresponding yields have been divided by 12. For example, the treasury bill rate on 1$^{\text{st}}$ January, 1980, corresponds to 12.04% and so the risk-free interest rate between 1980-01-01 and 1980-01-22 is set to 1%. The interest rates are used to calculate the excess returns of each asset.

The sample contains $n = 296$ monthly excess returns for each of the $d = 274$ firms. The estimated expected return of the GMVP corresponds to $\hat{\eta} = 0.18\%$, whereas $\hat{\sigma} = 0.8\%$ is its estimated standard deviation. The latter is obtained by the biased traditional estimator $\hat{\sigma}^2$. After adjusting for the bias the estimated standard deviation corresponds to

$$\sqrt{\frac{n}{n-d}}\cdot\hat{\sigma} = \sqrt{\frac{1}{1-1/Q}}\cdot\hat{\sigma} = 2.94\%$$

Table 4.1.: Industry sectors and numbers of assets.

Industry sector	Assets
Consumer Discretionary	54
Energy	14
Consumer Staples	31
Financial	38
Health Care	22
Industrial	40
Information Technology	20
Materials	24
Telecommunications	5
Utilities	26
Σ	274

with effective sample size $Q = n/d = 1.08$. Hence, the considered capital market is high-dimensional and the small-sample bias is tremendously large although there are 296 observations.

For the purpose of dimension reduction a pre-allocation is done by aggregating the stocks within each industry sector. More precisely, the asset returns of the firms belonging to the industry sector 'Consumer Discretionary' (see Table 4.1) are equally weighted by $1/54$, the asset returns belonging to 'Energy' by $1/14$ and so on. Hence, after the pre-allocation there remain 10 portfolios which can be interpreted as sector indices. The estimate for the expected return of the corresponding GMVP (see Table 4.2) amounts to $\hat{\eta} = 0.33\%$, whereas the estimated standard deviation is $\hat{\sigma} = 3.62\%$. Now there are only $d = 10$ assets (which are the sector indices), $Q = 29.6$ and so the curse of dimension is lifted. Hence, the estimate for σ based on the unbiased estimate for σ^2 corresponds to 3.68%, which is quite similar to $\hat{\sigma}$.

By applying Theorem 4.3 one can test for example against the null hypothesis $H_0 : w = 1/d = 0.1 \cdot 1$, i.e. that the GMVP corresponds to the *trivial portfolio*. Thus $q = d - 1 = 9$, $n - d = 286$, $\hat{w}^* = 0.1 \cdot 1$, and the F-statistic corresponds to

$$\frac{n-d}{q} \cdot \frac{(\hat{w} - \hat{w}^*)' \widehat{\Sigma} (\hat{w} - \hat{w}^*)}{\hat{\sigma}^2} = 15.7536 > 1.9127 = F^{-1}_{F,9,286}(1-\alpha)$$

Table 4.2.: Estimated weights of the GMVP and corresponding standard errors in parentheses.

Sector	Weight	Sector	Weight
Consumer Discretionary	−14.67% (10.24%)	Industrial	27.07% (13.10%)
Energy	14.33% (4.29%)	Information Technology	−4.70% (4.09%)
Consumer Staples	35.50% (9.79%)	Materials	−5.81% (9.64%)
Financial	−17.14% (7.89%)	Telecommunications	18.18% (5.34%)
Health Care	6.02% (7.19%)	Utilities	41.22% (5.88%)

with $\alpha = 0.05$. Hence, H_0 can be rejected which means that for the purpose of risk minimization it is not sufficient to choose the trivial portfolio.

The next null hypothesis is given by $H_0 \colon \sigma^2 \geq \sigma_0^2 = (0.2)^2/12 = 0.33\%$. Due to Theorem 4.4 the test statistic is given by

$$\frac{n\hat{\sigma}^2}{\sigma_0^2} = 116.2874 < 247.8302 = F^{-1}_{\chi^2,286}(\alpha)\,.$$

That means the GMVP has a sufficiently low risk of return (i.e. $12\sigma^2 < (0.2)^2$). Another null hypothesis is given by $H_0 : \eta \leq \eta_0 = 0.02/12 = 0.17\%$. For the t-test based on Theorem 4.5 one has to calculate the t-statistic

$$\frac{\hat{\eta} - \eta_0}{\sqrt{\left\{\hat{\sigma}^2(1 + \overline{\mathbf{r}}\,'\widehat{\Sigma}^{-1}\overline{\mathbf{r}}) - \hat{\eta}^2\right\}/(n-d)}} = 0.7352 \not> 1.6502 = F^{-1}_{t,286}(1-\alpha)\,,$$

and so the null hypothesis *cannot* be rejected. Although $\hat{\eta}$ is twice the size of η_0, the estimate for the expected return of the GMVP is not *significantly* larger than 0.02/12 or, equivalently, $12\eta > 0.02$. This is a typical problem of performance measurement (Frahm, 2007).

Now suppose that an investor wants to put 80% into the sectors 'Energy' and 'Information Technology' and he is searching for the corresponding LMVP. The matrix F given by (4.7) corresponds to the row vector $[\,0\;1\;0\;0\;0\;0\;1\;0\;0\;0\,]$, $f = 0.8$, the matrix

$$\overline{F}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Table 4.3.: Estimated weights of the LMVP and corresponding standard errors in parentheses.

Sector	Weight	Sector	Weight
Consumer Discretionary	-19.48% (12.66%)	Industrial	-18.37% (15.57%)
Energy	51.30% (3.81%)	Information Technology	28.70% (3.81%)
Consumer Staples	77.92% (11.36%)	Materials	-21.99% (11.82%)
Financial	-20.03% (9.76%)	Telecommunications	5.84% (6.50%)
Health Care	-16.77% (8.61%)	Utilities	32.88% (7.23%)

possesses full rank and

$$\overline{F}_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The transformation matrix $\mathcal{T}$ can be simply calculated by (4.11) and (4.12) and the estimated weights of the LMVP are given in Table 4.3. Further, the estimate for the expected LMVP return corresponds to $\hat{\eta}^* = 0.45\%$ and $\hat{\sigma}^* = 4.49\%$ for the standard deviation. Both the risk and expected return are apparently higher for the LMVP than for the GMVP. This effect has been already motivated in Section 4.1 and indicated by Figure 4.1.

Similar to the F-test conducted above, the null hypothesis is now that the industry sectors are equally weighted (except for 'Energy' and 'Information Technology'). Here $p = 7$, $n - d + 1 = 287$, and it can be found that $\hat{w}_2^{**} = 64.82\%$, $\hat{w}_7^{**} = 15.18\%$, $\hat{w}_1^{**}, \hat{w}_3^{**}, \ldots, \hat{w}_6^{**}, \hat{w}_8^{**}, \ldots, \hat{w}_{10}^{**} = 2.50\%$. The F-statistic given by (4.16) amounts to

$$\frac{n-d+q}{p} \cdot \frac{(\hat{w}^* - \hat{w}^{**})' \widehat{\Sigma} \, (\hat{w}^* - \hat{w}^{**})}{\hat{\sigma}^{*2}} = 18.4532 > 2.0416 = F_{F,7,287}^{-1}(1-\alpha)\,.$$

That means the LMVP is not a trivial one.

Further, the χ^2-test against the null hypothesis $H_0 \colon \sigma^{*2} \geq \sigma_0^{*2} = 0.33\%$ leads to

$$\frac{n\hat{\sigma}^{*2}}{\sigma_0^{*2}} = 178.8119 < 248.7615 = F_{\chi^2,287}^{-1}(\alpha)$$

and so also the LMVP risk of return is sufficiently low. However, for the t-test against $H_0 \colon \eta^* \leq \eta_0^* = 0.17\%$ it holds that

$$\frac{\hat{\eta}^* - \eta_0^*}{\sqrt{\left\{\hat{\sigma}^{*2}(1 + \overline{\mathbf{r}}\,' \mathcal{T} (\mathcal{T}' \widehat{\Sigma}\, \mathcal{T})^{-1} \mathcal{T}' \, \overline{\mathbf{r}}) - \hat{\eta}^{*2}\right\}/(n-d+q)}} = 1.0689\,,$$

whereas $F^{-1}_{t,287}(1-\alpha) = 1.6502$. Once again it is not possible to proof that the expected excess return of the LMVP is significantly large whilst the t-value obtained for the LMVP (1.0689) exceeds the t-value of the GMVP (0.7352).

4.6. Conclusion

Traditional portfolio optimization does not take estimation risk into account. Many empirical and numerical studies show that estimation risk is a substantial drawback of pure statistical portfolio optimization techniques. This is an important problem in practice, particularly when the sample size compared to the number of assets is small. In the present work it has been shown that estimation risk can be simply reduced by imposing linear constraints on the portfolio weights. Small-sample hypothesis tests for global and local minimum variance portfolios have been derived by linear regression theory. Further, the joint distribution of the weights as well as the first two moments of the estimator for the expected return of the global or some local minimum variance portfolio have been calculated. The presented results hold in small samples, which is an important fact since large-sample approximations fail if the sample size is large but the number of observations relative to the number of assets is small. Hence, the estimation risk of global and local minimum variance portfolios can be readily controlled by applying the given instruments even in the context of high-dimensional data.

Appendix

Proof of Proposition 4.1

Since

$$\begin{aligned} \mathrm{E}\big\{(Z_1 - b_1 - Z^{\mathrm{s}\prime} b^{\mathrm{s}})^2\big\} &= \mathrm{Var}(Z_1 - b_1 - Z^{\mathrm{s}\prime} b^{\mathrm{s}}) + \big\{\mathrm{E}(Z_1 - b_1 - Z^{\mathrm{s}\prime} b^{\mathrm{s}})\big\}^2 \\ &= \mathrm{Var}(Z_1 - Z^{\mathrm{s}\prime} b^{\mathrm{s}}) + \big\{\mathrm{E}(Z_1) - b_1 - \mathrm{E}(Z^{\mathrm{s}})' b^{\mathrm{s}}\big\}^2, \end{aligned}$$

where $b^{\mathrm{s}} := (b_2, \ldots, b_d)$, it is clear that $\beta_1 = \mathrm{E}(Z_1) - \mathrm{E}(Z^{\mathrm{s}})'\beta^{\mathrm{s}}$ and thus $\mathrm{E}(u) = 0$. That means the minimization problem can be solved equivalently by minimizing

$$\mathrm{E}\big\{(Z_1^* - b_2 Z_2^* - \ldots - b_d Z_d^*)^2\big\}, \tag{4.18}$$

where $Z_j^* := Z_j - \mathrm{E}(Z_j)$ for $j = 1, \ldots, d$. Now define $Z^{*\mathrm{s}} := (Z_2^*, \ldots, Z_d^*)$ so that (4.18) corresponds to

$$\mathrm{E}\big\{(Z_1^* - Z^{*\mathrm{s}\prime} b^{\mathrm{s}})^2\big\} = \mathrm{Var}(Z_1) - 2\mathrm{Cov}(Z_1, Z^{\mathrm{s}})' b^{\mathrm{s}} + b^{\mathrm{s}\prime}\mathrm{Var}(Z^{\mathrm{s}}) b^{\mathrm{s}}\,.$$

Due to the positive definiteness of $\mathrm{Var}(Z)$ also $\mathrm{Var}(Z^{\mathrm{s}})$ is positive-definite. Hence, this is a simple quadratic minimization problem and its unique solution is given by

$$\beta^{\mathrm{s}} = \mathrm{Var}(Z^{\mathrm{s}})^{-1}\mathrm{Cov}(Z_1, Z^{\mathrm{s}})\,.$$

Now calculate the $(d-1) \times 1$ vector of covariances between u and Z_j $(j = 2, \ldots, d)$, i.e.

$$\begin{aligned}\mathrm{Cov}(Z^{\mathrm{s}}, u) &= \mathrm{Cov}(Z^{\mathrm{s}}, Z_1 - \beta_1 - Z^{\mathrm{s}\prime}\beta^{\mathrm{s}}) \\ &= \mathrm{Cov}(Z_1, Z^{\mathrm{s}}) - \mathrm{Var}(Z^{\mathrm{s}})\beta^{\mathrm{s}} = 0\,.\end{aligned}$$

Q.E.D.

Proof of Theorem 4.3

From linear regression theory (Greene, 2003, p. 102) it is known that

$$\frac{n-d}{q} \cdot \frac{\hat{\mathbf{u}}^{*\prime}\hat{\mathbf{u}}^* - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} \sim F_{q,n-d}\,.$$

Since Eq. 4.5 constitutes an inhomogeneous regression it holds that $\hat{\eta}^* = \bar{\mathbf{r}}\,'\hat{w}^*$ and hence $\hat{\mathbf{u}}^* = \big(\mathbf{R} - 1\bar{\mathbf{r}}\,'\big)\hat{w}^*$. That means

$$\hat{\mathbf{u}}^{*\prime}\hat{\mathbf{u}}^*/n = \hat{w}^{*\prime}\big(\mathbf{R} - 1\bar{\mathbf{r}}\,'\big)'\big(\mathbf{R} - 1\bar{\mathbf{r}}\,'\big)\hat{w}^*/n = \hat{\sigma}^{*2},$$

where $\hat{\sigma}^{*2} := \hat{w}^{*\prime}\widehat{\Sigma}\,\hat{w}^*$. Since $\hat{\sigma}^2 = \hat{w}'\widehat{\Sigma}\,\hat{w}$ and $\hat{w} = \widehat{\Sigma}^{-1}1/(1'\widehat{\Sigma}^{-1}1)$, it follows that

$$\hat{\sigma}^{*2} = \hat{\sigma}^2 + (\hat{w} - \hat{w}^*)'\widehat{\Sigma}\,(\hat{w} - \hat{w}^*)\,.$$

Note also that $\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/n$ and thus

$$\frac{\hat{\mathbf{u}}^{*\prime}\hat{\mathbf{u}}^* - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} = \frac{(\hat{w} - \hat{w}^*)'\widehat{\Sigma}\,(\hat{w} - \hat{w}^*)}{\hat{\sigma}^2}\,,$$

which leads to the desired F-statistic. Q.E.D.

Proof of Theorem 4.5

From linear regression theory (Greene, 2003, p. 51) it follows that

$$\frac{\hat{\eta} - \eta}{\sqrt{n\hat{\sigma}^2 \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{11}/(n-d)}} \sim t\,(n-d)\,,$$

where $\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{11}$ denotes the upper left component of $(\mathbf{X}'\mathbf{X})^{-1}$, i.e.

$$\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{11} = \left\{n - n\overline{\mathbf{x}}'(\mathbf{X}^{\mathrm{s}\prime}\mathbf{X}^{\mathrm{s}})^{-1}\overline{\mathbf{x}}n\right\}^{-1} = \frac{\left\{1 - n\overline{\mathbf{x}}'(\mathbf{X}^{\mathrm{s}\prime}\mathbf{X}^{\mathrm{s}})^{-1}\overline{\mathbf{x}}\right\}^{-1}}{n}\,,$$

where $\mathbf{X}^{\mathrm{s}}$ $(n \times (d-1))$ symbolizes the regressor matrix $\mathbf{X}$ without the column of ones. Note that $\mathbf{X}^{\mathrm{s}\prime}\mathbf{X}^{\mathrm{s}} = n\big(\widehat{\Omega} + \overline{\mathbf{x}}\,\overline{\mathbf{x}}'\big)$ and due to the binomial inverse theorem (Press, 2005, p. 23) it holds that

$$n\,(\mathbf{X}^{\mathrm{s}\prime}\mathbf{X}^{\mathrm{s}})^{-1} = \big(\widehat{\Omega} + \overline{\mathbf{x}}\,\overline{\mathbf{x}}'\big)^{-1} = \widehat{\Omega}^{-1} - \frac{\widehat{\Omega}^{-1}\overline{\mathbf{x}}\,\overline{\mathbf{x}}'\widehat{\Omega}^{-1}}{1 + \overline{\mathbf{x}}'\widehat{\Omega}^{-1}\overline{\mathbf{x}}}\,.$$

That is

$$1 - n\overline{\mathbf{x}}'(\mathbf{X}^{\mathrm{s}\prime}\mathbf{X}^{\mathrm{s}})^{-1}\overline{\mathbf{x}} = 1 - \overline{\mathbf{x}}'\widehat{\Omega}^{-1}\overline{\mathbf{x}} + \frac{(\overline{\mathbf{x}}'\widehat{\Omega}^{-1}\overline{\mathbf{x}})^2}{1 + \overline{\mathbf{x}}'\widehat{\Omega}^{-1}\overline{\mathbf{x}}} = \frac{1}{1 + \overline{\mathbf{x}}'\widehat{\Omega}^{-1}\overline{\mathbf{x}}}$$

and thus

$$\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{11} = \frac{1 + \overline{\mathbf{r}}\,'\Delta'\widehat{\Omega}^{-1}\Delta\overline{\mathbf{r}}}{n}\,.$$

Since $\hat{\sigma}^2\Delta'\widehat{\Omega}^{-1}\Delta = \hat{\sigma}^2\widehat{\Sigma}^{-1} - \hat{w}\hat{w}'$ and $\hat{\eta} = \overline{\mathbf{r}}\,'\hat{w}$, it follows that

$$n\hat{\sigma}^2\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{11} = \hat{\sigma}^2 + \overline{\mathbf{r}}\,'(\hat{\sigma}^2\widehat{\Sigma}^{-1} - \hat{w}\hat{w}')\,\overline{\mathbf{r}} = \hat{\sigma}^2(1 + \overline{\mathbf{r}}\,'\widehat{\Sigma}^{-1}\overline{\mathbf{r}}) - \hat{\eta}^2\,.$$

Q.E.D.

Proof of Corollary 4.7

Theorem 4.6 implies that the covariance matrix of $(\hat{w}_2, \ldots, \hat{w}_d)$ is given by

$$\mathrm{Var}\{(\hat{w}_2, \ldots, \hat{w}_d)\} = \frac{\sigma^2}{n-d-1} \cdot \Omega^{-1}.$$

From Wishart theory it follows that $\widehat{\Omega}^{-1} \sim W_{d-1}^{-1}\big((\Omega/n)^{-1}, n+d-1\big)$ (Press, 2005, p. 117). Hence, it holds that

$$\mathrm{E}\big(\widehat{\Omega}^{-1}\big) = \frac{(\Omega/n)^{-1}}{(n+d-1) - 2\,(d-1) - 2} = \frac{n}{n-d-1} \cdot \Omega^{-1}$$

(Press, 2005, p. 119). Moreover, from linear regression theory (Greene, 2003, p. 56) it is known that $\hat{\sigma}^2_{\mathrm{OLS}} = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-d)$ is a conditionally unbiased estimator for σ^2. That means

$$\begin{aligned}\mathrm{E}\left(\frac{\hat{\sigma}^2_{\mathrm{OLS}}}{n} \cdot \widehat{\Omega}^{-1}\right) &= \mathrm{E}\left\{\mathrm{E}\left(\frac{\hat{\sigma}^2_{\mathrm{OLS}}}{n} \cdot \widehat{\Omega}^{-1} \,|\, \widehat{\Omega}^{-1}\right)\right\} = \mathrm{E}\left(\frac{\sigma^2}{n} \cdot \widehat{\Omega}^{-1}\right) \\ &= \frac{\sigma^2}{n-d-1} \cdot \Omega^{-1} = \mathrm{Var}\{(\hat{w}_2, \ldots, \hat{w}_d)\}\end{aligned}$$

and note that $\hat{\sigma}^2_{\mathrm{OLS}} = n/(n-d) \cdot \hat{\sigma}^2$. Q.E.D.

Chapter 5.

Dominant Estimators for the Global Minimum Variance Portfolio

5.1. Introduction

When implementing portfolio optimization according to Markowitz (1952), one needs to estimate the expected asset returns as well as the corresponding variances and covariances. If the parameter estimates are based only on time series information, the suggested portfolio tends to be far removed from the optimum. For this reason, there is a broad literature which addresses the question of how to reduce estimation risk in portfolio optimization. In a recent study, DeMiguel et al. (2007) compare portfolio strategies which differ in the treatment of estimation risk. It turns out that none of the strategies suggested in the literature is significantly better than naive diversification, i.e. taking the equally weighted portfolio. Further, the study conducted by DeMiguel et al. (2007) confirms that the considered strategies perform better than the traditional implementation of Markowitz optimization, which means replacing the unknown parameters by their sample counterparts.

The global minimum variance portfolio (GMVP) has been frequently advocated in the literature (Frahm, 2008, Jagannathan and Ma, 2003, Kempf and Memmel, 2006, Ledoit and Wolf, 2003) because it is completely independent of the expected asset returns, which have been found to be the principal source of estimation risk (Chopra and Ziemba, 1993, Merton, 1980). We present two estimators for the GMVP which *dominate* the traditional estimator with respect to the out-of-sample variance of the portfolio return. Due to the arguments set forth by Frahm (2008), the same conclusion can be drawn for estimating local minimum variance portfolios, i.e. minimum variance portfolios where the portfolio weights are subject to other linear equality constraints besides the budget constraint.

Okhrin and Schmid (2006), Kempf and Memmel (2006) and Frahm (2008) all explore the properties of the traditional GMVP estimator by assuming jointly normally distributed asset returns. They derive the small-sample distribution of the estimated portfolio weights and give a closed-form expression for the out-of-sample variance of the portfolio return. In contrast, Bayesian and shrinkage approaches have a long tradition in the implementation of modern portfolio optimization. Jobson and Korkie (1979) and Jorion (1986) introduce shrinkage estimators for the expected returns. Frost and Savarino (1986) generalize these estimators by also including the variances and covariances. Furthermore, DeMiguel et al. (2007), Garlappi et al. (2007), Golosnoy and Okhrin (2007) as well as Kan and Zhou (2007) present some shrinkage estimators for the weights of mean-variance optimal portfolios, whereas Ledoit and Wolf (2003) introduce a shrinkage estimator for the covariance matrix of stock returns and apply their results to the estimation of the GMVP.

Our work is related to these shrinkage approaches. However, it differs in two important aspects. First, we derive *feasible* estimators, and our dominance results turn out to be valid even in small samples. The shrinkage approaches presented by the aforementioned authors can only be justified for a large number of observations. As pointed out by Frahm (2008), large-sample results can be misleading in the context of portfolio optimization since, even if the sample size is large, the number of observations can be small compared to the number of assets. Second, in contrast to Ledoit and Wolf (2003) we do not seek to obtain a better covariance matrix estimator but instead to reduce the out-of-sample variance of the portfolio return, which seems to be the major goal when searching for a minimum variance portfolio.

Another method of alleviating the impact of estimation risk is to impose certain restrictions on the estimated covariance matrix or portfolio weights. Examples for restrictions on the covariance matrix are the single index model of Sharpe (1963) and the constant correlation model suggested by Elton and Gruber (1973). Jagannathan and Ma (2003) show that imposing short-sales constraints on the GMVP is equivalent to assuming a special structure of the covariance matrix. Frahm (2008) analyzes linear equality constraints on the portfolio weights and proves that linear restrictions reduce estimation risk. All these approaches have in common the fact that the restrictions may be binding and so the true GMVP does not need to be attained if the length of the time series approaches infinity. Nevertheless, in an empirical study presented by Chan et al. (1999) it has been shown that the reduction of estimation risk typically outweighs the loss caused by applying 'wrong' restrictions.

Shrinkage estimators reduce the estimation risk as well. However, in addition they have the appealing property of converging towards the optimal portfolio weights as the sample size grows to infinity.

Our contribution to the literature is threefold. First, we derive two shrinkage estimators for the GMVP that dominate the traditional estimator with respect to the out-of-sample variance of the portfolio return. Second, we present not only the small-sample properties of the shrinkage estimators and some related quantities, but also their large-sample properties for fixed d and $n \to \infty$ as well as $n, d \to \infty$ and $n/d \to q \leq \infty$. The latter kind of asymptotic behavior becomes relevant when analyzing the estimators in large asset universes. Third, backed by the results of DeMiguel et al. (2007), we derive a small-sample test for the *naive diversification hypothesis*, i.e. for deciding the question of whether or not it is better to completely ignore time series information in favor of naive diversification.

5.2. Preliminaries

5.2.1. Notation and Assumptions

Suppose that the investment universe consists of d assets and the investor is searching for a buy-and-hold portfolio which will be liquidated after one period. We will consider the asset excess returns $R_t = (R_{1t}, \ldots, R_{dt})$ for $t = 1, \ldots, n$,[1] i.e. the asset returns minus the corresponding risk-free interest rates. Nevertheless we will drop the prefix 'excess' for convenience and make the following assumptions:

A1. The asset returns are jointly normally distributed, i.e. $R_t \sim \mathcal{N}_d(\mu, \Sigma)$ for $t = 1, \ldots, n$ with $\mu \in \mathbb{R}^d$ and positive-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$.

A2. The mean vector μ and the covariance matrix Σ are unknown.

A3. The asset returns are serially independent.

A4. The sample size exceeds the number of assets, more precisely $n \geq d + 2$.

A5. There exist at least four assets, i.e. $d \geq 4$.

[1] In the following '$(x_1, \ldots, x_d)$' indicates a d-tuple, i.e. a d-dimensional column vector.

The GMVP w is defined as the solution of the minimization problem

$$\min_{v\in\mathbb{R}^d} v'\Sigma\, v\,, \qquad \text{s.t. } v\,'\mathbf{1} = 1\,. \tag{5.1}$$

Here $\mathbf{1}$ denotes a vector of ones. Since Σ is positive-definite, the GMVP is unique and the solution of this minimization problem corresponds to $w = \Sigma^{-1}\mathbf{1}/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$. The traditional estimator $\hat{w}_{\mathrm{T}}$ for the GMVP consists in replacing the unknown covariance matrix Σ with the sample covariance matrix $\widehat{\Sigma}$, i.e.

$$\widehat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n}\big(R_t - \bar{R}\big)\big(R_t - \bar{R}\big)'\,, \tag{5.2}$$

where $\bar{R} = 1/n\sum_{t=1}^{n} R_t$ represents the sample mean vector of $R_1,\ldots,R_n$. The variance of the GMVP return corresponds to $\sigma^2 = w'\Sigma\, w = 1/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$ and its traditional estimator is given by $\hat{\sigma}^2_{\mathrm{T}} = \hat{w}'_{\mathrm{T}}\widehat{\Sigma}\,\hat{w}_{\mathrm{T}} = 1/(\mathbf{1}'\widehat{\Sigma}^{-1}\mathbf{1})$.

Since the portfolio weights always add up to 1, it is possible to omit one element of the portfolio weights vector without losing information. We choose to omit the first element and define $w^{\mathrm{ex}} := (w_2,\ldots,w_d)$. For convenience we introduce the $(d-1)\times d$ matrix $\Delta := [\,\mathbf{1}\ -I_{d-1}\,]$. By using the operator Δ, we can easily switch between the two notations. For instance, note that $(v_1 - v_2) = -\Delta'(v_1^{\mathrm{ex}} - v_2^{\mathrm{ex}})$ for all vectors $v_1, v_2 \in \mathbb{R}^d$ whose elements add up to 1. Moreover, the following relationship will be useful in the subsequent discussion:

$$(v_1 - v_2)'A\,(v_1 - v_2) = (v_1^{\mathrm{ex}} - v_2^{\mathrm{ex}})'B\,(v_1^{\mathrm{ex}} - v_2^{\mathrm{ex}}) \tag{5.3}$$

with $B := \Delta A\Delta'$ for any $d\times d$ matrix A. A key note of the present work is that

$$v'\Sigma\, v = \sigma^2 + (v-w)'\Sigma\,(v-w) = \sigma^2 + (v^{\mathrm{ex}} - w^{\mathrm{ex}})'\Omega\,(v^{\mathrm{ex}} - w^{\mathrm{ex}}) \tag{5.4}$$

for every vector $v\in\mathbb{R}^d$ with $v\,'\mathbf{1} = 1$, where Ω is defined as $\Omega := \Delta\Sigma\Delta'$. The first equality in (5.4) can be obtained by noting that $\Sigma w = \mathbf{1}/(\mathbf{1}'\Sigma^{-1}\mathbf{1})$ and thus $v'\Sigma\, w = 1/(\mathbf{1}'\Sigma^{-1}\mathbf{1}) = \sigma^2$. The second equality follows from the arguments given above.

In the following $\chi^2_k(\lambda)$ denotes a noncentral χ^2-distributed random variable with $k\in\mathbb{N}$ degrees of freedom and noncentrality parameter $\lambda \geq 0$. This means $\chi^2_k(\lambda) \sim X'X$ with $X \sim \mathcal{N}_k(\theta, I_k)$ and $\theta\in\mathbb{R}^k$, where the noncentrality parameter is defined as $\lambda := \theta'\theta/2$. By contrast, χ^2_k stands for a central χ^2-distributed random variable (i.e. $\lambda = 0$) and we also define $\chi^r_k(\lambda) := \big\{\chi^2_k(\lambda)\big\}^{r/2}$ for any $r\in\mathbb{Z}$. Moreover, let $\chi^2_{k_1}(\lambda)$ and $\chi^2_{k_2}$ with $k_1, k_2\in\mathbb{N}$ be stochastically independent. Then $F_{k_1,k_2}(\lambda) \sim (k_2/k_1)\big(\chi^2_{k_1}(\lambda)/\chi^2_{k_2}\big)$ has a noncentral F-distribution with k_1 and k_2 degrees of freedom as well as noncentrality parameter $\lambda\geq 0$.

Now suppose that $X_1, \ldots, X_m$ are m independent copies of $X \sim \mathcal{N}_q(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ denotes a vector of zeros and Σ is a positive-definite $q \times q$ matrix. Then the $q \times q$ random matrix $W_q(\Sigma, m) \sim \sum_{i=1}^{m} X_i X_i'$ possesses a q-dimensional Wishart distribution with covariance matrix Σ and m degrees of freedom. Furthermore, $x^+ := \max\{x, 0\}$ denotes the positive part and $x^- := -\min\{x, 0\}$ the negative part of $x \in \mathbb{R}$. Let A be some positive-definite $q \times q$ matrix. Then $A^{\frac{1}{2}}$ represents the unique symmetrical $q \times q$ matrix such that $A = A^{\frac{1}{2}} A^{\frac{1}{2}}$. Finally, $x \propto y$ means 'x is proportional to y' and $\|\cdot\|$ denotes the Euclidean norm.

5.2.2. Important Theorems

Let us now provide some important theorems which will come in handy in the following sections. First, we present some elementary small-sample properties of the traditional estimator for the GMVP and its related quantities. A proof can be found in Kempf and Memmel (2006).

Lemma 5.1 (Kempf and Memmel (2006)) *Under assumptions A1 to A3 and $n > d$, the sample covariance matrix $\widehat{\Omega}$ of ΔR, the traditional estimator $\hat{w}_{\mathrm{T}}^{\mathrm{ex}}$ for the GMVP (except for the first portfolio weight), and the traditional estimator $\hat{\sigma}_{\mathrm{T}}^2$ for the minimum variance σ^2 satisfy the following properties:*

P1. $n\,\widehat{\Omega} \sim W_{d-1}(\Omega, n-1)$, *where* $\widehat{\Omega} := \frac{1}{n}\sum_{t=1}^{n}\big(\Delta R - \Delta\bar{R}\big)\big(\Delta R - \Delta\bar{R}\big)'$.

P2. $\hat{w}_{\mathrm{T}}^{\mathrm{ex}} \,|\, \widehat{\Omega} \sim \mathcal{N}_{d-1}\big(w^{\mathrm{ex}}, \sigma^2 \widehat{\Omega}^{-1}/n\big)$.

P3. $n\hat{\sigma}_{\mathrm{T}}^2/\sigma^2 \sim \chi^2_{n-d}$.

P4. $\hat{\sigma}_{\mathrm{T}}^2$ *is stochastically independent of* $\widehat{\Omega}$ *and* $\hat{w}_{\mathrm{T}}^{\mathrm{ex}}$.

The following theorem will play the central role in the development of the shrinkage estimator and its dominance property.

Theorem 5.2 *Consider a $q \times q$ random matrix $W \sim W_q\big(\Omega, m\big)$, where Ω is a positive-definite $q \times q$ matrix, $q \geq 3$ and $m \geq q + 2$, a q-dimensional random vector X with $X \,|\, W \sim \mathcal{N}_q\big(\omega, W^{-1}\big)$, where $\omega \in \mathbb{R}^q$ is an unknown parameter, and a random variable $\chi^2 \sim \chi^2_k$ with $k \geq 2$, which is stochastically independent of W and X. Furthermore,*

consider a non-stochastic vector $x \in \mathbb{R}^q$. *For all* $0 < c < 2\,(q-2)/(k+2)$, *the shrinkage estimator*

$$X_{\mathrm{S}} = x + \left(1 - \frac{c\,\chi^2}{(X-x)'W(X-x)}\right)(X-x)$$

dominates the estimator X *with respect to the loss function*

$$\mathcal{L}_{\omega,\Omega}\big(\hat{\omega}\big) = \big(\hat{\omega}-\omega\big)'\Omega\,\big(\hat{\omega}-\omega\big)\,, \tag{5.5}$$

i.e. $\mathrm{E}\big\{(X_{\mathrm{S}}-\omega)'\Omega\,(X_{\mathrm{S}}-\omega)\big\} < \mathrm{E}\big\{(X-\omega)'\Omega\,(X-\omega)\big\}$. *In case* $x=\omega$ *the expected loss of the shrinkage estimator becomes minimal if and only if* $c=(q-2)/(k+2)$.

Proof: See the appendix.

Note that Theorem 5.2 coincides with the well-known result developed by Stein (1956) if W is substituted by the identity matrix I_q. Other extensions of Stein's theorem, which can be found in the literature, require that W correspond to a non-stochastic but observable matrix Ω, or at least that W be stochastically independent of X where Ω is unobservable (Judge and Bock (1978, p. 177), Srivastava and Bilodeau (1989), and Press (2005, p. 189)). By contrast, we allow X to depend on a Wishart-distributed random matrix W, but the matrix Ω given in Theorem 5.2 remains unobservable.

Theorem 5.2 also clarifies why the shrinkage constant $c=(q-2)/(k+2)$ is a natural choice. Although any constant within the interval given in Theorem 5.2 would lead to a dominant estimator, only $c=(q-2)/(k+2)$ turns out to be the best choice if the reference vector x corresponds to the unknown parameter ω. The same value for c remains optimal in the variants of Stein's theorem where W is non-stochastic or stochastically independent of X.

5.2.3. Out-of-Sample Variance

The out-of-sample variance of the return of a stochastic portfolio $\hat{v}$ is defined as

$$\mathrm{Var}\big(\hat{v}\,'R\big) = \mathrm{E}\big\{\mathrm{Var}(\hat{v}\,'R\,|\,\hat{v})\big\} + \mathrm{Var}\big\{\mathrm{E}(\hat{v}\,'R\,|\,\hat{v})\big\} = \mathrm{E}\big(\hat{v}'\Sigma\,\hat{v}\big) + \mu'\mathrm{Var}\big(\hat{v}\big)\mu\,.$$

This means the total variance of the portfolio $\hat{v}$ can be split into a within variance $\mathrm{E}\big(\hat{v}'\Sigma\,\hat{v}\big)$ and a between variance $\mu'\mathrm{Var}\big(\hat{v}\big)\mu$. Due to (5.4), it holds that

$$\mathrm{Var}\big(\hat{v}\,'R\big) = \sigma^2 + \mathrm{E}\big\{(\hat{v}-w)'\Sigma\,(\hat{v}-w)\big\} + \mu'\mathrm{Var}\big(\hat{v}\big)\mu\,. \tag{5.6}$$

Hence, the minimum variance σ^2 is a lower bound for the out-of-sample variance of any given portfolio $\hat{v}$. Interestingly, the between variance $\mu'\mathrm{Var}(\hat{v})\mu$ vanishes whenever the expected asset returns are equal to each other, i.e. $\mu = \eta\,\mathbf{1}$ for any $\eta \in \mathbb{R}$. This can be seen by noting that $\mathrm{Var}(\hat{v}) = \Delta'\mathrm{Var}(\hat{v}^{\mathrm{ex}})\Delta$ and $\Delta\mu = \mathbf{0}$ if $\mu = \eta\,\mathbf{1}$.

Kempf and Memmel (2006) showed that – concerning the traditional estimator $\hat{w}$ for the GMVP – the second part of (5.6) corresponds to

$$\mathrm{E}\big\{(\hat{w}_{\mathrm{T}} - w)'\Sigma\,(\hat{w}_{\mathrm{T}} - w)\big\} = \frac{d-1}{n-d-1}\cdot\sigma^2\,.$$

The factor $(d-1)/(n-d-1)$ is large whenever the sample size n is small compared to the number of assets d. For $n, d \to \infty$ but $n/d \to q$ with $1 < q \leq \infty$, this factor tends to $1/(q-1)$. Hence even in large samples the contribution of the estimation risk to the out-of-sample variance is not negligible if the 'effective sample size' q is small. For instance, given an investment universe with $d = 50$ assets and a history of $n = 100$ monthly observations, the additional variance caused by the estimation risk is $1/(100/50 - 1) = 100\%$.

From the small-sample distribution of $\hat{w}$ presented by Frahm (2008), it follows that the third part of (5.6) corresponds to

$$\mu'\mathrm{Var}\big(\hat{w}_{\mathrm{T}}\big)\mu = \frac{r_{\max}^2 - r_{\mathrm{GMVP}}^2}{n-d-1}\cdot\sigma^2\,,$$

where $r_{\max}$ denotes the Sharpe ratio of the tangential portfolio $\Sigma^{-1}\mu\,/(\mathbf{1}'\Sigma^{-1}\mu)$ and r_{GMVP} the Sharpe ratio of the GMVP.[2] This means it holds that

$$\mathrm{Var}\big(\hat{w}_{\mathrm{T}}'R\big) = \left(1 + \frac{d-1}{n-d-1} + \frac{r_{\max}^2 - r_{\mathrm{GMVP}}^2}{n-d-1}\right)\cdot\sigma^2\,.$$

In most practical situations the difference of $r_{\max}^2$ and r_{GMVP}^2 turns out to be much smaller than the numerator $d-1$ (and even vanishes if $\mu = \eta\,\mathbf{1}$).

Generally, in real-world asset markets the expected returns presumably do not differ so greatly in the cross-section; the between variance is therefore very small compared to the within variance. Hence we believe that the between variance $\mu'\mathrm{Var}\big(\hat{v}\big)\mu$ for any portfolio $\hat{v}$ is negligible in most practical situations and will concentrate in the following on reducing the within variance $\mathrm{E}\big(\hat{v}'\Sigma\,\hat{v}\big)$. Note that each realization of $\hat{v}'\Sigma\,\hat{v}$ represents the *actual variance* of the return belonging to the portfolio $\hat{v}$, which has been chosen on the basis of historical observations, for instance. Then due to (5.4), each realization of $(\hat{v}-w)'\Sigma\,(\hat{v}-w)$ can be interpreted as that part of the actual variance which is caused by estimation risk. In the subsequent analysis this quantity will be referred to as the *loss* of $\hat{v}$.

[2] The Sharpe ratio of a portfolio is the expected excess return divided by the standard deviation.

5.3. The Dominant Estimators

5.3.1. Small-Sample Properties

We now present the shrinkage estimator for the GMVP that dominates the traditional estimator. Kempf and Memmel (2006) show that the traditional estimator is the best unbiased estimator in the case of jointly normally distributed asset returns.[3] However, as already discussed earlier, this estimator can lead to a huge out-of-sample variance of the portfolio return compared to σ^2, i.e. the smallest of all possible portfolio return variances.

In this section we will use the following notation. Let $\hat{w}_{\mathrm{A}}$ be an arbitrary portfolio. Then $\sigma^2_{\mathrm{A}} = \hat{w}'_{\mathrm{A}} \Sigma \, \hat{w}_{\mathrm{A}}$ is the actual variance of the portfolio return, whereas $\hat{\sigma}^2_{\mathrm{A}} = \hat{w}'_{\mathrm{A}} \widehat{\Sigma} \, \hat{w}_{\mathrm{A}}$ denotes the corresponding estimator. This notation will be used both for stochastic and non-stochastic portfolios, i.e. if w_{A} is a non-stochastic portfolio, it holds that $\sigma^2_{\mathrm{A}} = w'_{\mathrm{A}} \Sigma \, w_{\mathrm{A}}$ and $\hat{\sigma}^2_{\mathrm{A}} = w'_{\mathrm{A}} \widehat{\Sigma} \, w_{\mathrm{A}}$.

Theorem 5.3 *Suppose that the assumptions A1 to A5 are satisfied. Let $\hat{w}_{\mathrm{T}}$ be the traditional estimator for the GMVP w, whereas $w_{\mathrm{R}} \in \mathbb{R}^d$ with $w'_{\mathrm{R}} \mathbf{1} = 1$ denotes an arbitrary reference portfolio. Consider the shrinkage estimator*

$$\hat{w}_{\mathrm{S}} = \kappa_{\mathrm{S}} w_{\mathrm{R}} + \big(1 - \kappa_{\mathrm{S}}\big) \hat{w}_{\mathrm{T}} \tag{5.7}$$

with

$$\kappa_{\mathrm{S}} = \frac{d-3}{n-d+2} \cdot \frac{1}{\hat{\tau}_{\mathrm{R}}} \, ,$$

where $\hat{\tau}_{\mathrm{R}} = \big(\hat{\sigma}^2_{\mathrm{R}} - \hat{\sigma}^2_{\mathrm{T}}\big) / \hat{\sigma}^2_{\mathrm{T}}$ is the estimated relative loss of the reference portfolio w_{R} . The shrinkage estimator $\hat{w}_{\mathrm{S}}$ dominates $\hat{w}_{\mathrm{T}}$ with respect to the loss function $\mathcal{L}_{w,\Sigma}(\hat{v}) = (\hat{v} - w)' \Sigma \, (\hat{v} - w)$, i.e.

$$\mathrm{E}\big\{(\hat{w}_{\mathrm{S}} - w)' \Sigma \, (\hat{w}_{\mathrm{S}} - w)\big\} < \mathrm{E}\big\{(\hat{w}_{\mathrm{T}} - w)' \Sigma \, (\hat{w}_{\mathrm{T}} - w)\big\} \, .$$

Proof: See the appendix.

The estimator suggested in Theorem 5.3 exhibits the typical structure of James-Stein-type shrinkage estimators. It is a weighted average of a given reference portfolio and the traditional estimator for the GMVP. The better the reference portfolio fits the actual

[3] An estimator is called *best* if its covariance matrix attains the Rao-Cramér lower bound.

GMVP, the smaller the out-of-sample variance of the shrinkage estimator will be. When it comes to portfolio diversification without any subjective or empirical information as well as restrictions on the portfolio weights, the *naive portfolio* $w_{\mathrm{N}} := \mathbf{1}/d$ can be viewed as a natural choice for the reference portfolio. Due to the arguments given by DeMiguel et al. (2007), there are even doubts as to whether time series information can add useful information at all, and so $w_{\mathrm{R}} = w_{\mathrm{N}}$ might serve as a rule. We will come back to this point in Section 5.4.

Theorem 5.4 *Under the assumptions of Theorem 5.3, the distribution of the relative loss*

$$\tau_{\mathrm{S}} = \frac{\sigma_{\mathrm{S}}^2 - \sigma^2}{\sigma^2}$$

of the shrinkage estimator for the GMVP given by (5.7) depends only on the number of observations n, the number of assets d, and the relative loss $\tau_{\mathrm{R}} = (\sigma_{\mathrm{R}}^2 - \sigma^2)/\sigma^2$ of the reference portfolio. More precisely, τ_{S} can be represented stochastically by

$$\tau_{\mathrm{S}} = \left\| \kappa_{\mathrm{S}}\theta - \big(1 - \kappa_{\mathrm{S}}\big) V^{-\frac{1}{2}}\xi \right\|^2, \tag{5.8}$$

with any $\theta \in \mathbb{R}^{d-1}$ such that $\theta'\theta = \tau_{\mathrm{R}}$, $\xi \sim \mathcal{N}_{d-1}(\mathbf{0}, I_{d-1})$, $V \sim W_{d-1}(I_{d-1}, n-1)$, and

$$\kappa_{\mathrm{S}} = \frac{d-3}{n-d+2} \cdot \frac{\chi_{n-d}^2}{\big(\theta + V^{-\frac{1}{2}}\xi\big)' V \big(\theta + V^{-\frac{1}{2}}\xi\big)}\,.$$

Here ξ, V, and χ_{n-d}^2 are supposed to be mutually independent.

Proof: See the appendix.

Due to Theorem 5.3, the shrinkage estimator is dominant in the sense that $\mathrm{E}\big(\tau_{\mathrm{S}}\big) < \mathrm{E}\big(\tau_{\mathrm{T}}\big)$, where $\tau_{\mathrm{T}} = (\sigma_{\mathrm{T}}^2 - \sigma^2)/\sigma^2$ represents the relative loss of the traditional estimator for the GMVP. It can be shown that the expected relative loss of the shrinkage estimator is a strictly increasing function of τ_{R} and its infimum is attained if and only if $\tau_{\mathrm{R}} = 0$. Note that $\tau_{\mathrm{R}} = 0$ or, equivalently, $\theta = \mathbf{0}$ holds if and only if $w_{\mathrm{R}} = w$, since Σ is positive-definite. In that case it turns out that

$$\mathrm{E}\big(\tau_{\mathrm{S}}\big) = \left(1 - \frac{d-3}{d-1} \cdot \frac{n-d}{n-d+2}\right) \frac{d-1}{n-d-1}\,.$$

By contrast, $\mathrm{E}\big(\tau_{\mathrm{S}}\big) \to \mathrm{E}(\tau_{\mathrm{T}})$ for $\tau_{\mathrm{R}} \to \infty$.

Following the arguments given by Judge and Bock (1978, p. 182), we can try to reduce the out-of-sample variance of the suggested estimator by restricting κ_{S} to values smaller

than or equal to 1, i.e. by taking $\kappa_{\mathrm{M}} := \min\{\kappa_{\mathrm{S}}, 1\}$ instead of κ_{S}. Then the corresponding shrinkage estimator is given by

$$\hat{w}_{\mathrm{M}} := \kappa_{\mathrm{M}} w_{\mathrm{R}} + \big(1 - \kappa_{\mathrm{M}}\big)\hat{w}_{\mathrm{T}}\,. \tag{5.9}$$

The shrinkage constant κ_{M} can only attain values between 0 and 1, which prevents $\hat{w}_{\mathrm{M}}$ from having the opposite sign of $\hat{w}_{\mathrm{T}}$ whenever $\hat{\tau}_{\mathrm{R}}$ is small, i.e. whenever the traditional estimate of the GMVP is close to the reference portfolio. The next theorem states that the modified shrinkage estimator does, in fact, lead to a better out-of-sample performance.

Theorem 5.5 *Under the assumptions of Theorem 5.3 and given the notation of Theorem 5.4, the distribution of the relative loss*

$$\tau_{\mathrm{M}} = \frac{\sigma_{\mathrm{M}}^2 - \sigma^2}{\sigma^2}$$

of the modified shrinkage estimator for the GMVP given by (5.9) depends only on the number of observations n, the number of assets d, and the relative loss τ_{R} of the reference portfolio. More precisely, τ_{M} can be represented stochastically by

$$\tau_{\mathrm{M}} = \Big\|\kappa_{\mathrm{M}}\theta - \big(1 - \kappa_{\mathrm{M}}\big)V^{-\frac{1}{2}}\xi\Big\|^2\,, \tag{5.10}$$

with $\kappa_{\mathrm{M}} = \min\{\kappa_{\mathrm{S}}, 1\}$, and it holds that

$$\mathrm{E}\big(\tau_{\mathrm{M}}\big) < \mathrm{E}\big(\tau_{\mathrm{S}}\big) < \mathrm{E}\big(\tau_{\mathrm{T}}\big)\,.$$

Proof: See the appendix.

The stochastic representations (5.8) and (5.10) can be used, for instance, for evaluating the out-of-sample performances of the presented shrinkage estimators by Monte Carlo simulation. Theorem 5.5 asserts that the modified shrinkage estimator dominates not only the traditional estimator but also the simple shrinkage estimator given by (5.7). Moreover, it can be shown that the expected relative loss of $\hat{w}_{\mathrm{M}}$ corresponds to

$$\mathrm{E}\big(\tau_{\mathrm{M}}\big) = \mathrm{E}\left[\left\{\left(1 - \frac{d-3}{n-d+2}\cdot\frac{\chi_{n-d}^2}{\chi_{d+1}^2}\right)^+\right\}^2\right]\frac{d-1}{n-d-1}$$

in the event that $\tau_{\mathrm{R}} = 0$.

Our results about the superiority of the presented shrinkage estimators require the asset universe to consist of at least four assets. By contrast, if there are only two or three assets,

one should draw on the traditional estimator. It is worth pointing out that the methodology presented here can be easily applied to the estimation of local minimum variance portfolios. As has been shown by Frahm (2008), any d-dimensional asset universe can be transformed into a $(d-q)$-dimensional asset universe such that q linear equality constraints (besides the budget constraint) are implicitly satisfied for each portfolio of the $d-q$ available assets. In that case assumptions A4 and A5 have to be changed to $n \geq d-q+2$ and $d \geq q+4$. Furthermore, the chosen reference portfolio must satisfy the given linear restrictions.

5.3.2. Large-Sample Properties

In the previous section, we investigated the small-sample properties of the relative losses of the shrinkage estimators $\hat{w}_{\mathrm{S}}$ and $\hat{w}_{\mathrm{M}}$. Due to Theorem 5.4 and Theorem 5.5, it can be seen that the expected relative losses of the shrinkage estimators as well as the traditional estimator tend to zero if the number of assets d is fixed but $n \to \infty$. However, that does not mean that the presented shrinkage estimators are always asymptotically equivalent to the traditional estimator. This is confirmed by the next theorem.

Theorem 5.6 *Under assumptions A1 to A3 it holds that*

$$\sqrt{n}\begin{bmatrix}\hat{w}_{\mathrm{T}}-w\\ \hat{w}_{\mathrm{S}}-w\\ \hat{w}_{\mathrm{M}}-w\end{bmatrix} \xrightarrow{\mathrm{d}} \begin{bmatrix}1\\ \mathbb{1}_{\{\tau_{\mathrm{R}}=0\}}\left(1-\frac{d-3}{\xi'\xi}\right)+\mathbb{1}_{\{\tau_{\mathrm{R}}>0\}}\\ \mathbb{1}_{\{\tau_{\mathrm{R}}=0\}}\left(1-\frac{d-3}{\xi'\xi}\right)^{+}+\mathbb{1}_{\{\tau_{\mathrm{R}}>0\}}\end{bmatrix}\Lambda\xi\,, \qquad n\longrightarrow\infty\,,$$

where Λ *is a* $d\times(d-1)$ *matrix such that* $\Lambda\Lambda' = \sigma^2\Sigma^{-1} - ww'$ *and* $\xi \sim \mathcal{N}_{d-1}\big(\mathbf{0}, I_{d-1}\big)$.

Proof: See the appendix.

For instance, from the last theorem it follows that

$$\sqrt{n}\left(\hat{w}_{\mathrm{T}}-w\right) \xrightarrow{\mathrm{d}} \mathcal{N}_d\big(\mathbf{0}, \sigma^2\Sigma^{-1}-ww'\big)\,, \qquad n\longrightarrow\infty\,,$$

and the shrinkage estimators are asymptotically equivalent to the traditional estimator, i.e.

$$\sqrt{n}\left(\hat{w}_{\mathrm{T}}-\hat{w}_{\mathrm{S}}\right)\xrightarrow{\mathrm{p}}\mathbf{0} \qquad \text{and} \qquad \sqrt{n}\left(\hat{w}_{\mathrm{T}}-\hat{w}_{\mathrm{M}}\right)\xrightarrow{\mathrm{p}}\mathbf{0}\,, \qquad n\longrightarrow\infty\,, \tag{5.11}$$

only if $w_{\mathrm{R}} \neq w$.[4] The last theorem also implies that if $w_{\mathrm{R}} = w$ and the sample size is large (compared to the number of assets), the modified shrinkage estimate corresponds to the true GMVP roughly with probability $F_{\chi^2_{d-1}}\big(d-3\big)$. Admittedly, this might be regarded as purely theoretical, since it has to be assumed that $w_{\mathrm{R}} \neq w$ in most practical situations, with $\hat{w}_{\mathrm{M}}$ then being asymptotically equivalent to $\hat{w}_{\mathrm{T}}$ in the sense given above.

So far we have focused on the expected relative losses of the estimators for the GMVP but, as already mentioned, these quantities vanish if the sample size tends to infinity. However, due to the next theorem it is possible to make statements about the relative loss itself if d is fixed but n tends to infinity.

Theorem 5.7 *Under assumptions A1 to A3 it holds that*

$$n \begin{bmatrix} \tau_{\mathrm{T}} \\ \tau_{\mathrm{S}} \\ \tau_{\mathrm{M}} \end{bmatrix} \xrightarrow{\mathrm{d}} \begin{bmatrix} 1 \\ \mathbb{1}_{\{\tau_{\mathrm{R}}=0\}}\Big(1-\frac{d-3}{\chi^2_{d-1}}\Big)^2 + \mathbb{1}_{\{\tau_{\mathrm{R}}>0\}} \\ \mathbb{1}_{\{\tau_{\mathrm{R}}=0\}}\Big\{\Big(1-\frac{d-3}{\chi^2_{d-1}}\Big)^+\Big\}^2 + \mathbb{1}_{\{\tau_{\mathrm{R}}>0\}} \end{bmatrix} \chi^2_{d-1}\,, \qquad n \longrightarrow \infty\,.$$

Proof: See the appendix.

This theorem asserts that the relative losses are super-consistent. It is worth pointing out that, even if the expected relative losses of the shrinkage estimators presented here are always smaller than the expected loss of the traditional estimator (which follows from Theorem 5.4 and Theorem 5.5), a given realization of τ_{S} may turn out to be greater than τ_{T}. Surprisingly, Theorem 5.7 implies that, only if $w_{\mathrm{R}} = w$, the probability of this event does not vanish (even asymptotically) but tends to $F_{\chi^2_{d-1}}\big\{(d-3)/2\big\} > 0$. For example, if there exist $d = 5$ assets, this adverse effect occurs with a probability of approximately 9%. However, the same theorem confirms that $\tau_{\mathrm{M}} > \tau_{\mathrm{T}}$ is asymptotically impossible. This is another advantage of the modified shrinkage estimator over the simple one.

As already discussed earlier, it might be criticized that in many practical applications of portfolio theory the number of assets is large compared to the number of observations. In the following we will investigate the asymptotic distribution of the relative loss assuming that $n, d \to \infty$ but $n/d \to q$ with $1 < q \leq \infty$. Here the relative loss of the reference portfolio is assumed to be constant; recall that the number q can be interpreted as the

[4] The proof of Theorem 5.6 reveals that (5.11) can be even strengthened to almost sure convergence.

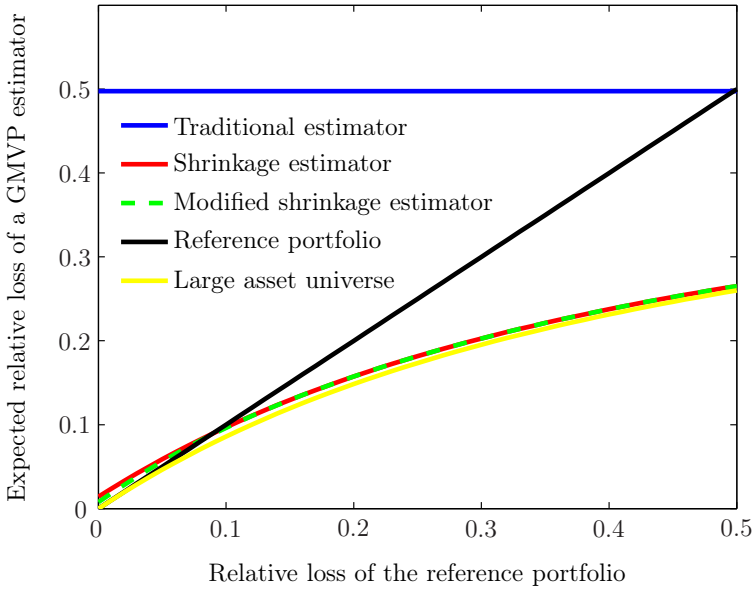

Figure 5.1.: Expected relative losses of the traditional (blue), simple (red) and modified (dashed green) shrinkage estimator for $n = 300$ and $d = 100$ as well as the relative loss of the reference portfolio (black) and the asymptotic loss function $L(\tau_{\mathrm{R}}\,, 3)$ (yellow).

effective sample size. The following theorem asserts that if the asset universe is large, the relative losses of all GMVP estimators are no longer super-consistent.

Theorem 5.8 *Under assumptions A1 to A3 it holds that*

$$\tau_{\mathrm{T}} \xrightarrow{\text{a.s.}} \frac{1}{q-1}$$

as $n, d \to \infty$ *but* $n/d \to q$ *with* $1 < q \leq \infty$ *. Moreover, concerning the shrinkage estimators for the GMVP it holds that*

$$\kappa_{\mathrm{S}}\,, \kappa_{\mathrm{M}} \xrightarrow{\text{a.s.}} \frac{1}{1+q\tau_{\mathrm{R}}}$$

as well as

$$\tau_{\mathrm{S}}\,, \tau_{\mathrm{M}} \xrightarrow{\text{a.s.}} L\big(\tau_{\mathrm{R}}, q\big) := \frac{\tau_{\mathrm{R}}}{(1+q\tau_{\mathrm{R}})^2} + \left(1 - \frac{1}{1+q\tau_{\mathrm{R}}}\right)^2 \frac{1}{q-1}$$

as $n, d \to \infty$ *but* $n/d \to q$ *with* $1 < q \leq \infty$ *.*

Proof: See the appendix.

It can be shown that the *asymptotic loss function* L is increasing in τ_{R}, and it holds that $L\big(\tau_{\mathrm{R}}, q\big) < 1/(q-1)$ whenever $q < \infty$, i.e. the shrinkage estimators dominate the traditional estimator with respect to the asymptotic loss if not only the number of observations

	$n \to \infty\,,\ d < \infty$		$n \to \infty\,,\ d \to \infty\,,\ n/d \to q$		
	$q = \infty$		$q < \infty$		$q = \infty$
	$\tau_{\mathrm{R}} = 0$	$\tau_{\mathrm{R}} > 0$	$\tau_{\mathrm{R}} = 0$	$\tau_{\mathrm{R}} > 0$	$\tau_{\mathrm{R}} \geq 0$
τ_{T}	0	0	$\frac{1}{q-1} > 0$	$\frac{1}{q-1} > 0$	0
τ_{S}	0	0	0	$0 < L(\tau_{\mathrm{R}}, q) < \frac{1}{q-1}$	0
τ_{M}	0	0	0	$0 < L(\tau_{\mathrm{R}}, q) < \frac{1}{q-1}$	0
$n\tau_{\mathrm{T}}$	χ^2_{d-1}	χ^2_{d-1}	∞	∞	∞
$n\tau_{\mathrm{S}}$	$\left(1 - \frac{d-3}{\chi^2_{d-1}}\right)^2 \chi^2_{d-1}$	χ^2_{d-1}	0	∞	∞
$n\tau_{\mathrm{M}}$	$\left\{\left(1 - \frac{d-3}{\chi^2_{d-1}}\right)^+\right\}^2 \chi^2_{d-1}$	χ^2_{d-1}	0	∞	∞

Table 5.1.: Large-sample properties of the relative losses of $\hat{w}_{\mathrm{T}}$, $\hat{w}_{\mathrm{S}}$, and $\hat{w}_{\mathrm{M}}$.

but also the number of assets tend to infinity and the effective sample size remains finite. Moreover, it turns out that $L\big(\tau_{\mathrm{R}}, q\big) > \tau_{\mathrm{R}}$ if and only if

$$\tau_{\mathrm{R}} < \frac{1}{q} \cdot \frac{2-q}{q-1}\,. \tag{5.12}$$

Therefore, the shrinkage estimators dominate the reference portfolio *uniformly* if $q \geq 2$ (see Figure 5.1). Conversely, in terms of the asymptotic loss they become uniformly worse than w_{R} as q tends to 1 from above, since the right-hand side of (5.12) then tends to infinity. The large-sample properties of the relative losses of the GMVP estimators $\hat{w}_{\mathrm{T}}$, $\hat{w}_{\mathrm{S}}$, and $\hat{w}_{\mathrm{M}}$ are summarized in Table 5.1.

5.3.3. The Link to Covariance Matrix Estimation

Jagannathan and Ma (2003) analyze short-sales constraints as a means of lessening the impact of estimation errors on the sample covariance matrix. They show that using short-sales constraints is equivalent to transforming the sample covariance matrix and taking this quantity for calculating the GMVP on the basis of the unconstrained traditional estimator for the GMVP. The following theorem states that a similar argument holds for the shrinkage estimators presented earlier.

Theorem 5.9 *For any reference portfolio w_{R} there exists a positive-definite $d \times d$ matrix Σ_{R}^{-1} such that $w_{\mathrm{R}} \propto \Sigma_{\mathrm{R}}^{-1}\mathbf{1}$ as well as $\mathbf{1}'\Sigma_{\mathrm{R}}^{-1}\mathbf{1} = \mathbf{1}'\widehat{\Sigma}^{-1}\mathbf{1}$, where $\widehat{\Sigma}$ is the sample covariance*

matrix given by Eq. 5.2 and it is assumed that $n > d$. *The shrinkage estimators for the GMVP can be calculated by using*

$$\widehat{\Sigma}_{\mathrm{S}}^{-1} := \kappa_{\mathrm{S}}\Sigma_{\mathrm{R}}^{-1} + \big(1-\kappa_{\mathrm{S}}\big)\widehat{\Sigma}^{-1} \qquad \textit{and} \qquad \widehat{\Sigma}_{\mathrm{M}}^{-1} := \kappa_{\mathrm{M}}\Sigma_{\mathrm{R}}^{-1} + \big(1-\kappa_{\mathrm{M}}\big)\widehat{\Sigma}^{-1}$$

for the traditional GMVP estimator, i.e.

$$\hat{w}_{\mathrm{S}} = \frac{\widehat{\Sigma}_{\mathrm{S}}^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\Sigma}_{\mathrm{S}}^{-1}\mathbf{1}} \qquad \textit{and} \qquad \hat{w}_{\mathrm{M}} = \frac{\widehat{\Sigma}_{\mathrm{M}}^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\Sigma}_{\mathrm{M}}^{-1}\mathbf{1}}\,.$$

Proof: See the appendix.

The random matrices $\widehat{\Sigma}_{\mathrm{S}}$ and $\widehat{\Sigma}_{\mathrm{M}}$ can be interpreted as shrinkage estimators for the unknown covariance matrix Σ. However, $\widehat{\Sigma}_{\mathrm{M}}$ is positive-definite, a trait that does not hold for $\widehat{\Sigma}_{\mathrm{S}}$ in general. Any other matrix which is proportional to $\widehat{\Sigma}_{\mathrm{S}}$ or $\widehat{\Sigma}_{\mathrm{M}}$ would lead to the same shrinkage estimators for the GMVP, but the expressions given in Theorem 5.9 satisfy a convenient scaling condition, i.e. $\mathbf{1}'\widehat{\Sigma}_{\mathrm{S}}^{-1}\mathbf{1} = \mathbf{1}'\widehat{\Sigma}_{\mathrm{M}}^{-1}\mathbf{1} = \mathbf{1}'\Sigma_{\mathrm{R}}^{-1}\mathbf{1} = \mathbf{1}'\widehat{\Sigma}^{-1}\mathbf{1} = 1/\hat{\sigma}_{\mathrm{T}}^2$.

Similar shrinkage estimators for the covariance matrix have been already suggested by Ledoit and Wolf (2001, 2003). However, the estimators given in Theorem 5.9 differ from the estimators introduced by Ledoit and Wolf in two aspects:

1. Their shrinkage constants depend on unobservable quantities which have to be estimated from empirical data. Even if the suggested covariance matrix estimators dominate the sample covariance matrix asymptotically, it is not clear why the dominance result should be valid in small samples. By contrast, our shrinkage approach focuses on the small-sample properties of the resulting portfolio weights.

2. Ledoit and Wolf shrink the covariance matrix itself, whereas our approach is based on shrinking its inverse. By shrinking the covariance matrix, it is possible to allow for $n \leq d$, i.e. the aforementioned authors are able to apply their approach to asset universes where the number of assets exceed the number of observations.

So far our methodology consists of shrinking the traditional GMVP estimator towards some non-stochastic reference portfolio w_{R}. However, all the presented results remain valid if w_{R} is a stochastic portfolio satisfying the budget constraint and being stochastically independent of the historical observations which are used for calculating $\hat{w}_{\mathrm{T}}$.[5] Nevertheless, in the following we will concentrate on the special case $w_{\mathrm{R}} = w_{\mathrm{N}} = \mathbf{1}/d$.

[5]For example, w_{R} could be interpreted as a portfolio which has been suggested by a layman.

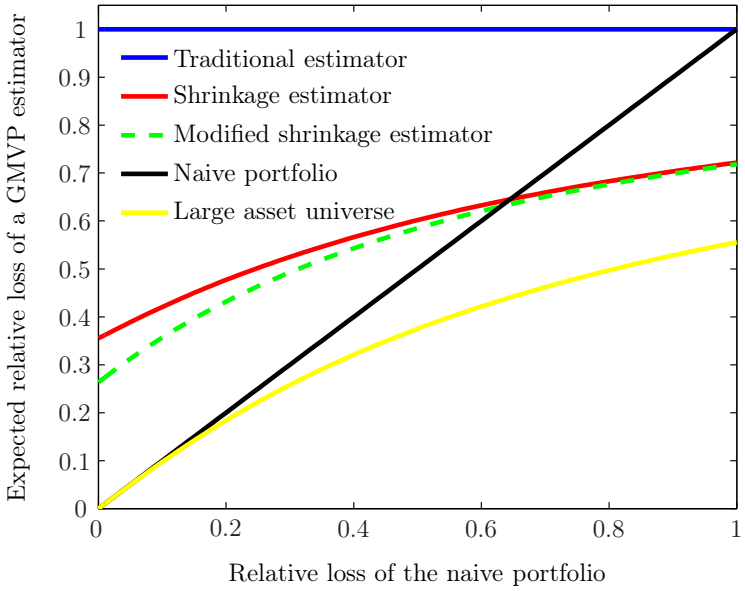

Figure 5.2.: Expected relative losses of the traditional (blue), simple (red) and modified shrinkage (dashed green) estimator for $n = 20$ and $d = 10$ as well as the relative loss of the naive portfolio (black) and the asymptotic loss function $L(\tau_{\mathrm{R}}, q)$ with $q = 2$ (yellow).

5.4. Naive Diversification vs. Portfolio Optimization

5.4.1. A Small-Sample Simulation Study

DeMiguel et al. (2007) raise the question of whether optimizing a portfolio using time series information is worthwhile to begin with. They do not even refer to the fact that asset returns typically exhibit structural breaks, serial correlations in the higher moments, and heavy tails. According to these authors, the estimation error outweighs the potential gain of portfolio optimization, even if the asset returns are normally distributed and serially independent. In this section we address a similar question: Does it pay to strive for the GMVP by using time series information or is it better to renounce parameter estimation altogether and put the money straight away into the naive portfolio?

In order to revisit this question, we may focus on the expected relative loss which is caused by a given GMVP estimator. Due to Theorem 5.5 and the arguments given in Section 5.3.2, we will concentrate on the modified shrinkage estimator $\hat{w}_{\mathrm{M}}$ and choose the naive portfolio w_{N} as a reference portfolio. Although closed-form expressions for τ_{M} in large samples and asset universes have been already presented in Section 5.3.2, the relative loss can only be simulated, e.g. by using Equations 5.8 and 5.10, if the sample is small. Figure

5.2 contains the expected relative losses of the four different portfolio strategies, i.e. naive diversification, traditional estimation, as well as simple and modified shrinkage estimation for $n = 20$ observations and $d = 10$ assets. The x-axis denotes the relative loss τ_{N} of the naive portfolio, whereas the y-axis accounts for the expected relative losses of the different portfolio strategies depending on τ_{N}. Note that (according to Theorem 5.4) the expected relative loss of the traditional estimator does not depend on τ_{N} but only on the number n of observations and the number d of assets.

It can be seen that the expected relative loss of the traditional estimator corresponds to 100%. Due to Theorem 5.4 and Theorem 5.5 it is clear that the expected relative losses of the shrinkage estimators are always below the expected relative loss of the traditional estimator. This is also confirmed by Figure 5.2. Particularly if τ_{N} is small, i.e. the true GMVP does not differ too greatly from the naive portfolio (which serves as an anchor point for $\hat{w}_{\mathrm{S}}$ and $\hat{w}_{\mathrm{M}}$), the shrinkage estimators are more favorable than the traditional estimator.

Figure 5.2 also indicates the *critical relative loss* τ_{N}^{*} of the naive portfolio with respect to the modified shrinkage estimator $\hat{w}_{\mathrm{M}}$. This is that point on the x-axis where the modified shrinkage estimator leads to the same expected relative loss as naive diversification. As indicated by Figure 5.2, this critical value is about 63%. For example if there are 5 years of quarterly asset returns and 10 stocks on the market, naive diversification would be better as long as $\tau_{\mathrm{N}} < 63\%$. Suppose that the standard deviation of the GMVP return corresponds to $\sigma = 10\%$, whereas its counterpart related to the naive portfolio amounts to 11% (per quarter). In that case, the relative loss of naive diversification is $\tau_{\mathrm{N}} = (0.11/0.10)^2 - 1 = 21\%$, whereas the expected relative loss caused by the modified shrinkage estimator roughly amounts to $\mathrm{E}\big(\tau_{\mathrm{M}}\big) = 43\%$. Therefore, it would not pay to use the modified shrinkage estimator in that case. In contrast, if the naive portfolio leads to a standard deviation of 13%, it holds that $\tau_{\mathrm{N}} = (0.13/0.10)^2 - 1 = 69\% > \tau_{\mathrm{N}}^{*}$ and so the modified shrinkage estimator is slightly better than the naive portfolio. Note that traditional estimation is always worse than naive diversification in all such cases.

Table 5.2 lists some critical relative losses of naive diversification for different combinations of n and d. For example, if 10 years of monthly asset return observations are available (i.e. $n = 120$) and the stock market consists of $d = 50$ assets, one should use the modified shrinkage estimator if and only if the variance of the naive portfolio return is at least 21% greater than the variance of the GMVP return. Depending on the length of the time series

$n \setminus d$	5	10	25	50	100
12	52% (550%)	847% (99261%)	—	—	—
24	16% (111%)	40% (334%)	—	—	—
36	9% (59%)	19% (132%)	152% (1809%)	—	—
60	5% (30%)	9% (58%)	28% (209%)	420% (7806%)	—
120	2% (13%)	4% (24%)	8% (57%)	21% (161%)	377% (5202%)

Table 5.2.: Critical relative losses of the naive portfolio with respect to the modified shrinkage estimator for different combinations of n and d. The parentheses under the critical relative losses contain the critical thresholds of $\hat{\tau}_{\mathrm{N}}$ for testing the naive diversification hypothesis at a significance level of $\alpha = 5\%$.

and the number of assets, the modified shrinkage estimator is able to reduce the relative loss of naive diversification. However, the table also indicates that, if the number of assets is large compared to the number of observations, naive diversification is apparently the best strategy, which reconfirms the naive diversification hypothesis of DeMiguel et al. (2007).

5.4.2. Testing the Naive Diversification Hypothesis

For applying the decision rule discussed above, one needs two numbers, i.e.

1. the critical relative loss of the naive portfolio with respect to the modified shrinkage estimator and
2. the relative loss of the naive portfolio.

The critical relative loss can be calculated by Monte Carlo simulation (as it was done to obtain Table 5.2), whereas the actual relative loss of the naive portfolio is not observable and needs to be estimated from the history. The next theorem provides the distribution of its empirical counterpart $\hat{\tau}_{\mathrm{N}}$ or, more generally, $\hat{\tau}_{\mathrm{R}}$ (see also Theorem 5.3).

Theorem 5.10 *Under assumptions A1 to A3 and $n > d$, the estimator $\hat{\tau}_{\mathrm{R}} = \big(\hat{\sigma}^2_{\mathrm{R}} - \hat{\sigma}^2_{\mathrm{T}}\big)/\hat{\sigma}^2_{\mathrm{T}}$ for the relative loss of the reference portfolio is conditionally noncentrally F-distributed, more precisely*

$$\hat{\tau}_{\mathrm{R}} \sim \frac{d-1}{n-d} \cdot F_{d-1,n-d}\big(\tau_{\mathrm{R}} \chi^2_{n-1}/2\big)\,.$$

Proof: See the appendix.

With Theorem 5.10, it is possible to test whether one should invest in the naive portfolio or to apply a GMVP estimator, i.e.

$$H_0\colon \tau_{\mathrm{N}} \leq \tau_{\mathrm{N}}^* \text{ vs.}$$
$$H_1\colon \tau_{\mathrm{N}} > \tau_{\mathrm{N}}^* .$$

The test statistic is given by $\hat{\tau}_{\mathrm{N}} = \big(\hat{\sigma}_{\mathrm{N}}^2 - \hat{\sigma}_{\mathrm{T}}^2\big)/\hat{\sigma}_{\mathrm{T}}^2$ and according to Theorem 5.10, H_0 can be rejected whenever the realization of $\hat{\tau}_{\mathrm{N}}$ exceeds the upper α-quantile ($0 < \alpha < \frac{1}{2}$) of the cumulative distribution function of

$$\frac{d-1}{n-d} \cdot F_{d-1,n-d}\big(\tau_{\mathrm{N}}^* \chi_{n-1}^2/2\big) ,$$

which can be also calculated by Monte Carlo simulation.[6]

Critical thresholds for this hypothesis test at a significance level of $\alpha = 5\%$ are presented in Table 5.2. For instance, suppose that the asset universe consists of 50 assets and the investor can observe 10 years of monthly asset returns. Then the naive diversification hypothesis can be only rejected if $\hat{\tau}_{\mathrm{N}} > 161\%$. Note that this is by far greater than the theoretical value of the critical relative loss $\tau_{\mathrm{N}}^* = 21\%$, since the distribution of $\hat{\tau}_{\mathrm{N}}$ is considerably skewed to the right.

We consciously formulate the hypothesis test in such a way that the naive portfolio has to be rejected but not the portfolio based on some GMVP estimator. Therefore, for typical significance levels like $\alpha = 1\%, 5\%, 10\%$, our decision rule favors naive diversification. More precisely, if H_0 can be rejected, the considered GMVP estimator significantly leads to a better out-of-sample performance but if H_0 is not rejected, from a statistical point of view it cannot be assumed that naive diversification is better. However, in that case the naive portfolio can be justified either empirically, e.g. because of the well-known stylized facts of financial data, or due to the arguments given by DeMiguel et al. (2007). In other words: if it is not possible to guarantee that a statistical method will lead to a better result but it is likely that the outcome will become worse, the naive portfolio can be justified by the principle of insufficient reason (against naive diversification).

[6] This hypothesis test can be adapted to any GMVP estimator if its expected relative loss $\mathrm{E}(\tau) < \infty$ depends only on n, d, and τ_{N} and provided $\tau_{\mathrm{N}} \mapsto \mathrm{E}(\tau)$ has only one intersection point with $\tau_{\mathrm{N}} \mapsto \tau_{\mathrm{N}}$.

5.5. Conclusion

We present two shrinkage estimators for the GMVP that dominate the traditional estimator under the assumption of serially independent and identically normally distributed asset returns. Their small-sample and their large-sample properties alike have been investigated. The presented shrinkage estimators considerably reduce the out-of-sample variance of the portfolio return compared to the traditional estimator, especially if the asset universe is large. In addition, we provide a hypothesis test to decide whether one should invest in a portfolio based on estimators for the GMVP or in the naive portfolio. This decision depends only on three quantities: the number of observations, the number of assets, and the relative loss (compared to the GMVP) caused by naive diversification. Further research could include, for instance, an empirical investigation of the presented shrinkage estimators.

Appendix

Lemma 5.11 *For any $\lambda \geq 0$ it holds that*

$$\mathrm{E}\Big\{\chi_q^{-2}(\lambda)\Big\} = q\,\mathrm{E}\Big\{\chi_{q+2}^{-4}(\lambda)\Big\} + 2\lambda\,\mathrm{E}\Big\{\chi_{q+4}^{-4}(\lambda)\Big\}, \tag{5.13}$$

and if $q \geq 3$,

$$(q-2)\,\mathrm{E}\Big\{\chi_q^{-2}(\lambda)\Big\} = (q-2\lambda)\,\mathrm{E}\Big\{\chi_{q+2}^{-2}(\lambda)\Big\} + 2\lambda\,\mathrm{E}\Big\{\chi_{q+4}^{-2}(\lambda)\Big\}. \tag{5.14}$$

Proof: Eq. 5.13 follows immediately from Theorem 2 in Judge and Bock (1978, p. 322) by setting $\phi(x) = x^{-2}$, $A = I_q$, and $\theta \in \mathbb{R}^q$ such that $\lambda = \theta'\theta/2$. Similarly, with $\phi(x) = x^{-1}$,

$$1 = q\,\mathrm{E}\Big\{\chi_{q+2}^{-2}(\lambda)\Big\} + 2\lambda\,\mathrm{E}\Big\{\chi_{q+4}^{-2}(\lambda)\Big\} = (q-2)\,\mathrm{E}\Big\{\chi_q^{-2}(\lambda)\Big\} + 2\lambda\,\mathrm{E}\Big\{\chi_{q+2}^{-2}(\lambda)\Big\}$$

for any $q \geq 3$, which leads to (5.14). Q.E.D.

Lemma 5.12 *Consider a $q \times q$ random matrix $V \sim W_q\big(I_q, m\big)$ with $q \geq 3$ and $m \geq q+2$. Further, define $\lambda := \theta'\theta/2$ and $\hat{\lambda} := \theta' V \theta/2$ for some $\theta \in \mathbb{R}^q$. Then it holds that*

$$\mathrm{E}\left[\left(\mathrm{tr}\,V^{-1} - \frac{\lambda}{\hat{\lambda}}\cdot q\right)\mathrm{E}\Big\{\chi_{q+2}^{-2}(\hat{\lambda})\,|\,V\Big\}\right] = \frac{q-1}{m-q-1}\cdot\mathrm{E}\left[(q-2)\cdot\frac{\lambda}{\hat{\lambda}}\cdot\mathrm{E}\Big\{\chi_q^{-2}(\hat{\lambda})\,|\,V\Big\}\right]$$

and

$$\begin{aligned}\mathrm{E}\left[\left(\mathrm{tr}\,V^{-1} - \frac{\lambda}{\hat{\lambda}}\cdot q\right)\mathrm{E}\Big\{\chi_{q+2}^{-4}(\hat{\lambda})\,|\,V\Big\}\right] \;=\; & \frac{q-1}{m-q-1}\cdot\mathrm{E}\left[\frac{\lambda}{\hat{\lambda}}\cdot\mathrm{E}\Big\{\chi_q^{-2}(\hat{\lambda})\,|\,V\Big\}\right] - \\ & \frac{q-1}{m-q-1}\cdot\mathrm{E}\left[2\lambda\,\mathrm{E}\Big\{\chi_{q+2}^{-4}(\hat{\lambda})\,|\,V\Big\}\right].\end{aligned}$$

Proof: Consider the function $h\big(2\hat{\lambda}\big) = \mathrm{E}\big\{\chi_{q+2}^{-2}(\hat{\lambda}) \,|\, V\big\}$ and note that, after rotating θ, it holds that $2\hat{\lambda} = \theta'\theta\chi^2$ for some random variable $\chi^2 \sim \chi_m^2$. Then, due to Theorem 6 in Judge and Bock (1978, p. 324),

$$\mathrm{E}\Big\{\Big(\operatorname{tr} V^{-1}\Big)h\big(2\hat{\lambda}\big)\Big\} = \frac{q\,(m-2)}{m-q-1}\cdot\mathrm{E}\bigg\{\frac{h\big(2\hat{\lambda}\big)}{\chi^2}\bigg\} + \frac{2\,(q-1)}{m-q-1}\cdot\mathrm{E}\Big\{\theta'\theta\,h'\big(2\hat{\lambda}\big)\Big\}\,,$$

where h' denotes the first derivative of h with respect to $2\hat{\lambda}$. Since $\lambda/\hat{\lambda} = 1/\chi^2$,

$$\mathrm{E}\bigg\{\bigg(\operatorname{tr} V^{-1} - \frac{\lambda}{\hat{\lambda}}\cdot q\bigg)h\big(2\hat{\lambda}\big)\bigg\} = \frac{q-1}{m-q-1}\cdot\bigg[q\,\mathrm{E}\bigg\{\frac{h\big(2\hat{\lambda}\big)}{\chi^2}\bigg\} + 2\,\theta'\theta\,\mathrm{E}\Big\{h'\big(2\hat{\lambda}\big)\Big\}\bigg]\,, \quad (5.15)$$

where

$$h'\big(2\hat{\lambda}\big) = \frac{1}{2}\cdot\frac{\mathrm{dE}\big\{\chi_{q+2}^{-2}(\hat{\lambda})\,|\,V\big\}}{\mathrm{d}\hat{\lambda}} = \frac{1}{2}\cdot\Big[\mathrm{E}\big\{\chi_{q+4}^{-2}(\hat{\lambda})\,|\,V\big\} - \mathrm{E}\big\{\chi_{q+2}^{-2}(\hat{\lambda})\,|\,V\big\}\Big]\,,$$

which follows from the derivative rule on page 327 in Judge and Bock (1978). After substituting $h'(2\hat{\lambda})$ in (5.15) and some re-arrangement, we obtain

$$\mathrm{E}\bigg[\bigg(\operatorname{tr} V^{-1} - \frac{\lambda}{\hat{\lambda}}\cdot q\bigg)\,\mathrm{E}\Big\{\chi_{q+2}^{-2}(\hat{\lambda})\,|\,V\Big\}\bigg] = \\ \frac{q-1}{m-q-1}\cdot\mathrm{E}\bigg[\frac{\lambda}{\hat{\lambda}}\bigg[(q-2\hat{\lambda})\,\mathrm{E}\Big\{\chi_{q+2}^{-2}(\hat{\lambda})\,|\,V\Big\} + 2\hat{\lambda}\mathrm{E}\Big\{\chi_{q+4}^{-2}(\hat{\lambda})\,|\,V\Big\}\bigg]\bigg]\,.$$

Now the first statement of the lemma appears immediately after applying (5.14). Similarly, by allowing for the function $h\big(2\hat{\lambda}\big) = \mathrm{E}\big\{\chi_{q+2}^{-4}(\hat{\lambda})\,|\,V\big\}$ and using (5.13), the second statement of the lemma becomes valid. Q.E.D.

Proof of Theorem 5.2

The loss function $\mathcal{L}_{\omega,\Omega}$ can be re-formulated as

$$\mathcal{L}_{\omega,\Omega}\big(\hat{\omega}\big) = \big(\hat{\omega}-\omega\big)'\Omega\,\big(\hat{\omega}-\omega\big) = \big(\hat{\theta}-\theta\big)'\big(\hat{\theta}-\theta\big) = \mathcal{L}_{\theta}\big(\hat{\theta}\big)\,,$$

where $\hat{\theta} := \Omega^{\frac{1}{2}}(\hat{\omega}-x)$ and $\theta := \Omega^{\frac{1}{2}}(\omega-x)$. Accordingly, the random vector X is transformed into $Y := \Omega^{\frac{1}{2}}(X-x)\,|\,V \sim \mathcal{N}_q\big(\theta, V^{-1}\big)$ with $V := \Omega^{-\frac{1}{2}}W\Omega^{-\frac{1}{2}} \sim W_q\big(I_q, m\big)$ and similarly

$$Y_{\mathrm{S}} := \Omega^{\frac{1}{2}}(X_{\mathrm{S}}-x) = \Big(1 - \frac{c\,\chi^2}{Y'VY}\Big)Y\,.$$

After some elementary transformations, it turns out that

$$\mathcal{L}_{\theta}(Y_S) = \mathcal{L}_{\theta}(Y) - \bigg\{2c\chi^2\cdot\frac{Y'(Y-\theta)}{Y'VY} - c^2\chi^4\cdot\frac{Y'Y}{(Y'VY)^2}\bigg\}\,.$$

This means the random variable Y_S dominates Y if and only if

$$\mathrm{E}\big\{\mathcal{L}_\theta(Y) - \mathcal{L}_\theta(Y_S)\big\} = 2ck\mathcal{E}_1 - c^2 k\,(k+2)\,\mathcal{E}_2 > 0\,, \tag{5.16}$$

where

$$\mathcal{E}_1 := \mathrm{E}\bigg\{\frac{Y'(Y-\theta)}{Y'VY}\bigg\} \qquad \text{and} \qquad \mathcal{E}_2 := \mathrm{E}\bigg\{\frac{Y'Y}{(Y'VY)^2}\bigg\}.$$

Hence, the dominance result is satisfied for all c with $0 < c < 2/(k+2)\cdot\mathcal{E}_1/\mathcal{E}_2$ and, to prove the theorem, it has to be shown that $\mathcal{E}_1/\mathcal{E}_2 \geq (q-2)$. Now we define $Z := V^{\frac{1}{2}}Y$ and $\zeta := V^{\frac{1}{2}}\theta$ so that $Z\,|\,V \sim \mathcal{N}_q(\zeta, I_q)$. Then it holds that

$$\frac{Y'(Y-\theta)}{Y'VY}\,|\,V \sim \frac{Z'V^{-1}(Z-\zeta)}{Z'Z}\,|\,V \qquad \text{and} \qquad \frac{Y'Y}{(Y'VY)^2}\,|\,V \sim \frac{Z'V^{-1}Z}{(Z'Z)^2}\,|\,V\,.$$

By setting $\phi(x) = x^{-1}$ in Theorem 1 and Theorem 2 of Judge and Bock (1978, pp. 321–322) and allowing for $\lambda = \theta'\theta/2$ and $\hat{\lambda} = \theta'V\theta/2$ it follows that

$$\mathrm{E}\bigg\{\frac{Y'(Y-\theta)}{Y'VY}\,|\,V\bigg\} = \Big(\operatorname{tr} V^{-1}\Big)\mathrm{E}\Big\{\chi^{-2}_{q+2}(\hat{\lambda})\,|\,V\Big\} + 2\lambda\,\mathrm{E}\Big\{\chi^{-2}_{q+4}(\hat{\lambda})\,|\,V\Big\} - 2\lambda\,\mathrm{E}\Big\{\chi^{-2}_{q+2}(\hat{\lambda})\,|\,V\Big\}.$$

Similarly, by setting $\phi(x) = x^{-2}$ in Theorem 2 given by Judge and Bock (1978, p. 322), we find that

$$\mathrm{E}\bigg\{\frac{Y'Y}{(Y'VY)^2}\,|\,V\bigg\} = \Big(\operatorname{tr} V^{-1}\Big)\mathrm{E}\Big\{\chi^{-4}_{q+2}(\hat{\lambda})\,|\,V\Big\} + 2\lambda\,\mathrm{E}\Big\{\chi^{-4}_{q+4}(\hat{\lambda})\,|\,V\Big\}.$$

After some re-arrangement and an application of (5.14) we obtain

$$\mathrm{E}\bigg(\frac{Y'(Y-\theta)}{Y'VY}\,|\,V\bigg) = (q-2)\cdot\frac{\lambda}{\hat{\lambda}}\cdot\mathrm{E}\Big\{\chi^{-2}_{q}(\hat{\lambda})\,|\,V\Big\} + \Big(\operatorname{tr} V^{-1} - \frac{\lambda}{\hat{\lambda}}\cdot q\Big)\,\mathrm{E}\Big\{\chi^{-2}_{q+2}(\hat{\lambda})\,|\,V\Big\}.$$

Moreover, with an application of (5.13) it also turns out that

$$\mathrm{E}\bigg(\frac{Y'Y}{(Y'VY)^2}\,|\,V\bigg) = \frac{\lambda}{\hat{\lambda}}\cdot\mathrm{E}\Big\{\chi^{-2}_{q}(\hat{\lambda})\,|\,V\Big\} + \Big(\operatorname{tr} V^{-1} - \frac{\lambda}{\hat{\lambda}}\cdot q\Big)\,\mathrm{E}\Big\{\chi^{-4}_{q+2}(\hat{\lambda})\,|\,V\Big\}.$$

Now, from Lemma 5.12 it follows that $\mathcal{E}_1 = (q-2)\,\mathcal{E}_2 + \varepsilon$ with

$$\varepsilon := \frac{(q-1)(q-2)}{m-q-1}\cdot 2\lambda\,\mathrm{E}\bigg[\mathrm{E}\Big\{\chi^{-4}_{q+2}(\hat{\lambda})\,|\,V\Big\}\bigg] \geq 0\,.$$

Since $\mathcal{E}_1 \geq (q-2)\,\mathcal{E}_2$ with $\mathcal{E}_2 > 0$ it follows that $\mathcal{E}_1/\mathcal{E}_2 \geq (q-2)$. For $x = \omega$ it holds that $\lambda = 0$ and thus $\mathcal{E}_1 = (q-2)\,\mathcal{E}_2$. This means the optimal constant c of the quadratic function given by (5.16) does not depend on $\mathcal{E}_1$ or $\mathcal{E}_2$. Further, it is unique and corresponds to $c = (q-2)/(k+2)$. Q.E.D.

Proof of Theorem 5.3

Lemma 5.1 and Theorem 5.2 can be brought together by the following substitutions: $m = n-1$, $q = d-1$, $W = n\,\widehat{\Omega}/\sigma^2$, $X = \hat{w}_{\mathrm{T}}^{\mathrm{ex}}$, $\chi^2 = n\,\hat{\sigma}_{\mathrm{T}}^2/\sigma^2$, $k = n-d$, and $x = w_{\mathrm{R}}^{\mathrm{ex}}$. Then the constant

$$c = \frac{q-2}{k+2} = \frac{d-3}{n-d+2}$$

leads to a dominant shrinkage estimator $\hat{w}_{\mathrm{S}}^{\mathrm{ex}}$ for w^{ex}, viz.

$$\hat{w}_{\mathrm{S}}^{\mathrm{ex}} = w_{\mathrm{R}}^{\mathrm{ex}} + \left(1 - \frac{d-3}{n-d+2} \cdot \frac{\hat{\sigma}_{\mathrm{T}}^2}{(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})'\widehat{\Omega}\,(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})}\right)\big(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}}\big)\,.$$

Note that

$$(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})'\widehat{\Omega}\,(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}}) = (\hat{w}_{\mathrm{T}} - w_{\mathrm{R}})'\widehat{\Sigma}\,(\hat{w}_{\mathrm{T}} - w_{\mathrm{R}})$$

and thus

$$\frac{\hat{\sigma}_{\mathrm{T}}^2}{(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})'\widehat{\Omega}\,(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})} = \frac{\hat{\sigma}_{\mathrm{T}}^2}{(\hat{w}_{\mathrm{T}} - w_{\mathrm{R}})'\widehat{\Sigma}\,(\hat{w}_{\mathrm{T}} - w_{\mathrm{R}})} = \frac{\hat{\sigma}_{\mathrm{T}}^2}{\hat{\sigma}_{\mathrm{R}}^2 - \hat{\sigma}_{\mathrm{T}}^2} = \frac{1}{\hat{\tau}_{\mathrm{R}}}\,.$$

Due to $\hat{w}_{\mathrm{S}} = \mathbf{e}_1 - \Delta'\hat{w}_{\mathrm{S}}^{\mathrm{ex}}$ it follows that

$$\hat{w}_{\mathrm{S}} = w_{\mathrm{R}} + \left(1 - \frac{d-3}{n-d+2} \cdot \frac{1}{\hat{\tau}_{\mathrm{R}}}\right)\big(\hat{w}_{\mathrm{T}} - w_{\mathrm{R}}\big) = \kappa_{\mathrm{S}}\, w_{\mathrm{R}} + \big(1 - \kappa_{\mathrm{S}}\big)\,\hat{w}_{\mathrm{T}}\,.$$

Q.E.D.

Proof of Theorem 5.4

After some calculations we find that

$$\tau_{\mathrm{S}} = \tau_{\mathrm{R}} - 2\big(1-\kappa_{\mathrm{S}}\big)a + \big(1-\kappa_{\mathrm{S}}\big)^2 b\,,$$

where

$$\kappa_{\mathrm{S}} = \frac{d-3}{n-d+2} \cdot \frac{n\hat{\sigma}_{\mathrm{T}}^2/\sigma^2}{(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})'(n\widehat{\Omega}/\sigma^2)(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})}\,,$$

$$a = \frac{(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})'\Omega\,(w^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})}{\sigma^2} \qquad \text{and} \qquad b = \frac{(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})'\Omega\,(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})}{\sigma^2}\,.$$

With $\theta = \Omega^{\frac{1}{2}}/\sigma\,(w^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})$, $\xi \sim \mathcal{N}_{d-1}(\mathbf{0}, I_{d-1})$, and $V \sim W_{d-1}(I_{d-1}, n-1)$, the shrinkage constant κ_{S} can be represented by

$$\kappa_{\mathrm{S}} = \frac{d-3}{n-d+2} \cdot \frac{\chi_{n-d}^2}{\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)}$$

as well as $a = \theta'\big(\theta + V^{-\frac{1}{2}}\xi\big)$ and $b = \big(\theta + V^{-\frac{1}{2}}\xi\big)'\big(\theta + V^{-\frac{1}{2}}\xi\big)$, where ξ, V, and χ^2_{n-d} are mutually independent. Hence, τ_{S} is equal to the expression given on the right hand side of (5.8). Moreover, it holds that

$$\tau_{\mathrm{S}} = \big\|\mathcal{O}\big\{\kappa_{\mathrm{S}}\theta - \big(1-\kappa_{\mathrm{S}}\big)V^{-\frac{1}{2}}\xi\big\}\big\|^2 = \big\|\kappa_{\mathrm{S}}\eta - \big(1-\kappa_{\mathrm{S}}\big)\mathcal{O}V^{-\frac{1}{2}}\xi\big\|^2$$

with $\eta := \mathcal{O}\theta$ for any orthogonal $(d-1)\times(d-1)$ matrix $\mathcal{O}$; note also that κ_{S} is a function of $V^{-\frac{1}{2}}\xi$ only through the quadratic form

$$\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big) = \big(\eta + \mathcal{O}V^{-\frac{1}{2}}\xi\big)'\big(\mathcal{O}V\mathcal{O}'\big)\big(\eta + \mathcal{O}V^{-\frac{1}{2}}\xi\big)\,.$$

The random matrix V has a radial distribution, i.e. $\mathcal{O}V\mathcal{O}' \sim V$ as well as $\mathcal{O}V^{-1}\mathcal{O}' \sim V^{-1}$. Similarly, ξ has a spherical distribution, i.e. $\mathcal{O}\xi \sim \xi$. It follows that $\mathcal{O}V^{-\frac{1}{2}}\mathcal{O}' \sim V^{-\frac{1}{2}}$ and thus $\mathcal{O}V^{-\frac{1}{2}}\xi \sim V^{-\frac{1}{2}}\xi$. This means for any rotation η of θ it holds that

$$\tau_{\mathrm{S}} \sim \big\|\kappa_{\mathrm{S}}\eta - \big(1-\kappa_{\mathrm{S}}\big)V^{-\frac{1}{2}}\xi\big\|^2\,.$$

Ergo, the distribution of τ_{S} depends only on n, d, and $\tau_{\mathrm{R}} = \theta'\theta$. Q.E.D.

Proof of Theorem 5.5

From the proof of Theorem 5.4 it follows that the distribution of τ_{M}, too, is only a function of n, d, and τ_{R}. To prove that $\mathrm{E}(\tau_{\mathrm{M}}) < \mathrm{E}(\tau_{\mathrm{S}})$, the relative loss of the simple shrinkage estimator can be written as

$$\tau_{\mathrm{S}} = \tau_{\mathrm{R}} - 2\,\theta'V^{-\frac{1}{2}}\big(1-\kappa_{\mathrm{S}}\big)\big(V^{\frac{1}{2}}\theta + \xi\big) + \big(1-\kappa_{\mathrm{S}}\big)^2\,\|V^{\frac{1}{2}}\theta + \xi\|_V^2\,.$$

Since $\big(1-\kappa_{\mathrm{S}}\big) = \big(1-\kappa_{\mathrm{S}}\big)^+ - \big(1-\kappa_{\mathrm{S}}\big)^-$, the relative loss of the modified shrinkage estimator becomes

$$\tau_{\mathrm{M}} = \tau_{\mathrm{S}} - 2\,\theta'V^{-\frac{1}{2}}\big(1-\kappa_{\mathrm{S}}\big)^-\big(V^{\frac{1}{2}}\theta + \xi\big) - \big\{\big(1-\kappa_{\mathrm{S}}\big)^-\big\}^2\,\|V^{\frac{1}{2}}\theta + \xi\|_V^2\,.$$

Here it holds that

$$\mathrm{E}\Big[\big\{\big(1-\kappa_{\mathrm{S}}\big)^-\big\}^2\,\|V^{\frac{1}{2}}\theta + \xi\|_V^2\Big] > 0$$

and from Theorem 1 given by Judge and Bock (1978, pp. 321) it follows that

$$\mathrm{E}\Big\{\theta'V^{-\frac{1}{2}}\big(1-\kappa_{\mathrm{S}}\big)^-\big(V^{\frac{1}{2}}\theta + \xi\big)\Big\} = \tau_{\mathrm{R}}\mathrm{E}\left[\left\{1 - \frac{d-3}{n-d+2}\cdot\frac{\chi^2_{n-d}}{\chi^2_{d+1}(\tau_{\mathrm{R}}\chi^2_{n-1}/2)}\right\}^-\right] \geq 0\,.$$

That means $\mathrm{E}\big(\tau_{\mathrm{M}}\big) < \mathrm{E}\big(\tau_{\mathrm{S}}\big)$. The second inequality $\mathrm{E}\big(\tau_{\mathrm{S}}\big) < \mathrm{E}\big(\tau_{\mathrm{T}}\big)$ is a direct consequence of Theorem 5.3. Q.E.D.

Proof of Theorem 5.6

The traditional estimator for the GMVP without the first portfolio weight can be represented by $\hat{w}_{\mathrm{T}}^{\mathrm{ex}} = w^{\mathrm{ex}} + \sigma\,\Omega^{-\frac{1}{2}}V^{-\frac{1}{2}}\xi$, where $V \sim W_{d-1}(I_{d-1}, n-1)$ is stochastically independent of $\xi \sim \mathcal{N}_{d-1}(\mathbf{0}, I_{d-1})$. Since $\sqrt{n}\,V^{-\frac{1}{2}} = \big(V/n\big)^{-\frac{1}{2}} \xrightarrow{\text{a.s.}} I_{d-1}$ as $n \to \infty$, it holds that

$$\sqrt{n}\,\big(\hat{w}_{\mathrm{T}}^{\mathrm{ex}} - w^{\mathrm{ex}}\big) \xrightarrow{\text{a.s.}} \sigma\,\Omega^{-\frac{1}{2}}\xi\,, \qquad n \longrightarrow \infty\,.$$

The presented expression for the asymptotic normality of $\hat{w}_{\mathrm{T}} = \mathbf{e}_1 - \Delta'\hat{w}_{\mathrm{T}}^{\mathrm{ex}}$ follows from the relationship $\sigma^2\Delta'\Omega^{-1}\Delta = \sigma^2\Sigma^{-1} - ww'$ (Frahm, 2008). Further, the shrinkage estimator can be represented by

$$\hat{w}_{\mathrm{S}}^{\mathrm{ex}} = w_{\mathrm{R}}^{\mathrm{ex}} + \left\{1 - \frac{d-3}{n-d+2} \cdot \frac{\chi^2_{n-d}}{\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)}\right\}\Big\{\big(w^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}}\big) + \sigma\,\Omega^{-\frac{1}{2}}V^{-\frac{1}{2}}\xi\Big\}\,,$$

where $\theta = \Omega^{\frac{1}{2}}/\sigma\,(w^{\mathrm{ex}} - w_{\mathrm{R}}^{\mathrm{ex}})$ and $\theta'\theta = \tau_{\mathrm{R}}$. Following the proof of Theorem 5.4 it can be assumed that $\theta = \big(\sqrt{\tau_{\mathrm{R}}}, \mathbf{0}\big)$ without loss of generality. Since

$$\frac{\theta' V \theta}{n} = \tau_{\mathrm{R}} \cdot \frac{\chi^2_{n-1}}{n} \xrightarrow{\text{a.s.}} \tau_{\mathrm{R}}\,, \quad \frac{2\theta' V^{\frac{1}{2}}\xi}{n} = 2\theta'(V/n)^{\frac{1}{2}}\xi/\sqrt{n} \xrightarrow{\text{a.s.}} 0\,, \quad \frac{\xi'\xi}{n} \xrightarrow{\text{a.s.}} 0\,, \quad n \longrightarrow \infty\,,$$

it follows that $\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)/n \xrightarrow{\text{a.s.}} \tau_{\mathrm{R}}$ as well as $\chi^2_{n-d}/n \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$. Hence, in the event that $\tau_{\mathrm{R}} > 0$ it holds that

$$\sqrt{n} \cdot \frac{d-3}{n-d+2} \cdot \frac{\chi^2_{n-d}/n}{\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)/n} \cdot \big(w_{\mathrm{R}}^{\mathrm{ex}} - w^{\mathrm{ex}}\big) \xrightarrow{\text{a.s.}} \mathbf{0}\,, \qquad n \longrightarrow \infty\,.$$

Further, as already mentioned above, $\sqrt{n}\,\sigma\,\Omega^{-\frac{1}{2}}V^{-\frac{1}{2}}\xi \xrightarrow{\text{d}} \sigma\,\Omega^{-\frac{1}{2}}\xi$ and so

$$\left\{1 - \frac{d-3}{n-d+2} \cdot \frac{\chi^2_{n-d}/n}{\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)/n}\right\}\sqrt{n}\,\sigma\,\Omega^{-\frac{1}{2}}V^{-\frac{1}{2}}\xi \xrightarrow{\text{a.s.}} \sigma\,\Omega^{-\frac{1}{2}}\xi$$

as $n \to \infty$. By contrast, if $\tau_{\mathrm{R}} = 0$ and thus $\theta = \mathbf{0}$ as well as $w^{\mathrm{ex}} = w_{\mathrm{R}}^{\mathrm{ex}}$,

$$\frac{d-3}{n-d+2} \cdot \frac{\chi^2_{n-d}}{\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)} = \frac{d-3}{n-d+2} \cdot \frac{\chi^2_{n-d}}{\xi'\xi}$$

and since $\chi^2_{n-d}/(n-d+2) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$,

$$\sqrt{n}\,\big(\hat{w}_{\mathrm{S}}^{\mathrm{ex}} - w^{\mathrm{ex}}\big) \xrightarrow{\text{a.s.}} \left(1 - \frac{d-3}{\xi'\xi}\right)\sigma\,\Omega^{-\frac{1}{2}}\xi\,, \qquad n \longrightarrow \infty\,.$$

Similar arguments hold for the modified shrinkage estimator, since

$$\min\left\{\sqrt{n} \cdot \frac{d-3}{n-d+2} \cdot \frac{\chi^2_{n-d}/n}{\big(\theta + V^{-\frac{1}{2}}\xi\big)'V\big(\theta + V^{-\frac{1}{2}}\xi\big)/n}\,, \sqrt{n}\right\} \xrightarrow{\text{a.s.}} 0\,, \qquad n \longrightarrow \infty\,,$$

if $\tau_{\mathrm{R}} > 0$ and otherwise

$$\min\left\{\frac{d-3}{n-d+2}\cdot\frac{\chi^2_{n-d}}{\xi'\xi}\,,1\right\}\xrightarrow{\text{a.s.}}\min\left\{\frac{d-3}{\xi'\xi}\,,1\right\},\qquad n\longrightarrow\infty\,.$$

Q.E.D.

Proof of Theorem 5.7

Due to Eq. 5.3 it will suffice to concentrate on the GMVP estimators without the first portfolio weight for calculating the relative losses, e.g.

$$n\tau_{\mathrm{T}}=\frac{\sqrt{n}\,(\hat{w}^{\mathrm{ex}}_{\mathrm{T}}-w^{\mathrm{ex}})'\Omega\sqrt{n}\,(\hat{w}^{\mathrm{ex}}_{\mathrm{T}}-w^{\mathrm{ex}})}{\sigma^2}\,.$$

Now the theorem follows immediately by applying the Continuous Mapping Theorem to the results which are given in the proof of Theorem 5.6 and noting that

$$\left[\mathbb{1}_{\{\tau_{\mathrm{R}}=0\}}X+\mathbb{1}_{\{\tau_{\mathrm{R}}>0\}}\right]^2=\mathbb{1}_{\{\tau_{\mathrm{R}}=0\}}X^2+\mathbb{1}_{\{\tau_{\mathrm{R}}>0\}}$$

for any random variable X. Q.E.D.

Proof of Theorem 5.8

Due to the proof of Theorem 5.6 it holds that

$$\tau_{\mathrm{T}}=\frac{(\hat{w}^{\mathrm{ex}}_{\mathrm{T}}-w^{\mathrm{ex}})'\Omega\,(\hat{w}^{\mathrm{ex}}_{\mathrm{T}}-w^{\mathrm{ex}})}{\sigma^2}=\xi'V^{-1}\xi=\frac{\chi^2_{d-1}}{\chi^2_{n-d+1}}$$

with $\chi^2_{d-1}:=\xi'\xi$ and $\chi^2_{n-d+1}:=\chi^2_{d-1}/\xi'V^{-1}\xi$. Note that $(n-d)\to\infty$ as $n,d\to\infty$ and $n/d\to q$. That means

$$\tau_{\mathrm{T}}=\frac{d}{n-d}\cdot\frac{\chi^2_{d-1}/d}{\chi^2_{n-d+1}/(n-d)}\xrightarrow{\text{a.s.}}\frac{1}{q-1}\,,\qquad n,d\longrightarrow\infty\,,\,n/d\longrightarrow q\,.$$

For proving the almost sure convergence of the shrinkage constants κ_{S} and κ_{M}, consider $\theta=\left(\sqrt{\tau_{\mathrm{R}}}\,,\mathbf{0}\right)$ and suppose that $V^{\frac{1}{2}}$ is the Cholesky root of V, i.e.

$$\theta'V^{\frac{1}{2}}\xi=\sqrt{\tau_{\mathrm{R}}}\,\chi_{n-1}\xi_1\,.$$

Furthermore, note that $(d-3)/(n-d+2)\to 1/(q-1)$, $\chi^2_{n-d}/(n-d)\xrightarrow{\text{a.s.}}1$,

$$\frac{\theta'V\theta}{n-d}=\tau_{\mathrm{R}}\cdot\frac{\chi^2_{n-1}}{n}\cdot\frac{n}{n-d}\xrightarrow{\text{a.s.}}\frac{q\tau_{\mathrm{R}}}{q-1}\,,\quad\frac{2\theta'V^{\frac{1}{2}}\xi}{n-d}=2\sqrt{\tau_{\mathrm{R}}}\cdot\frac{\chi_{n-1}\xi_1}{n-d}\xrightarrow{\text{a.s.}}0$$

as well as

$$\frac{\xi'\xi}{n-d} = \frac{\xi'\xi}{d} \cdot \frac{d}{n-d} \xrightarrow{\text{a.s.}} \frac{1}{q-1}\,, \qquad n,d \longrightarrow \infty\,,\, n/d \longrightarrow q\,.$$

Now, by applying the Continuous Mapping Theorem, we obtain $\kappa_{\mathrm{S}}\,, \kappa_{\mathrm{M}} \xrightarrow{\text{a.s.}} 1/(1+q\tau_{\mathrm{R}})$ as $n,d \to \infty$ and $n/d \to q\,$. Similarly, note that

$$2\theta' V^{-\frac{1}{2}}\xi = 2\sqrt{\tau_{\mathrm{R}}} \cdot \frac{\xi_1}{\chi_{n-d+1}} = 2\sqrt{\tau_{\mathrm{R}}} \cdot \frac{n-d}{\chi_{n-d+1}} \cdot \frac{\xi_1}{n} \cdot \frac{n}{n-d} \xrightarrow{\text{a.s.}} 0$$

and $\xi' V^{-1}\xi \xrightarrow{\text{a.s.}} 1/(q-1)$ as $n,d \to \infty$ and $n/d \to q\,$. By relying on (5.8) and (5.10) it turns out that

$$\tau_{\mathrm{S}}\,, \tau_{\mathrm{M}} \xrightarrow{\text{a.s.}} \frac{\tau_{\mathrm{R}}}{1+q\tau_{\mathrm{R}}} - \left(1 - \frac{1}{1+q\tau_{\mathrm{R}}}\right)\tau_{\mathrm{R}} + \left(1 - \frac{1}{1+q\tau_{\mathrm{R}}}\right)^2 \left(\tau_{\mathrm{R}} + \frac{1}{q-1}\right).$$

After a little calculation it can be found that the limit corresponds to the asymptotic loss function $L\big(\tau_{\mathrm{R}}, q\big)$ which is given in the theorem. Q.E.D.

Proof of Theorem 5.9

Since $w_{\mathrm{R}}'\mathbf{1} = 1 > 0$, the angle between w_{R} and $\mathbf{1}$ is acute. Therefore, there exists an orthogonal $d \times d$ matrix $\mathcal{O}$ such that both $\mathcal{O}w_{\mathrm{R}}$ and $\mathcal{O}\mathbf{1}$ belong to the set $\{x \in \mathbb{R}^d : x > \mathbf{0}\}$. That means there also exists a positive-definite diagonal $d \times d$ matrix Λ such that $\mathcal{O}\mathbf{1} = \Lambda\mathcal{O}w_{\mathrm{R}}$, i.e. $w_{\mathrm{R}} = A\mathbf{1}$ with $A := \mathcal{O}'\Lambda^{-1}\mathcal{O}$ being positive-definite. The matrix Σ_{R}^{-1} can be obtained by re-scaling A such that the condition $\mathbf{1}'\Sigma_{\mathrm{R}}^{-1}\mathbf{1} = \mathbf{1}'\widehat{\Sigma}^{-1}\mathbf{1} > 0$ is satisfied. Now the rest of the theorem can be verified by substituting $\widehat{\Sigma}^{-1}$ by the given expressions for $\widehat{\Sigma}_{\mathrm{S}}^{-1}$ and $\widehat{\Sigma}_{\mathrm{M}}^{-1}$ within the traditional GMVP estimator. Q.E.D.

Proof of Theorem 5.10

Due to the proof of Theorem 5.4 it can be seen that

$$\hat{\tau}_{\mathrm{R}} = \frac{\big(V^{\frac{1}{2}}\theta + \xi\big)'\big(V^{\frac{1}{2}}\theta + \xi\big)}{\chi_{n-d}^2};$$

note that $\theta' V \theta = \tau_{\mathrm{R}}\chi_{n-1}^2\,.$ Q.E.D.

Chapter 6.

A Hypothesis Test for the Best Investment Strategy

Motivation

In many practical situations we cannot calculate some number analytically. Then it is often possible to use Monte Carlo simulation for approximating the desired quantity. Standard large-sample theory can be applied for controlling such kind of approximations. Now suppose that we are searching for the maximum of some unknown and analytically untractable quantities. Thus we could choose the largest outcome given by Monte Carlo simulation. However, since we take the best result from a set of given outcomes there is some sort of selection bias and it is not evident if our choice is *significantly* better or at least not worse than any other. The same problem frequently occurs in statistical inference or decisions under uncertainty when searching for the 'best alternative' such as portfolio optimization. In the following I will derive a large-sample test for the best alternative in a rather general setting. The presented test is demonstrated by an application to financial data. It is shown that the Jobson-Korkie-Memmel test for the Sharpe ratios of two asset portfolios can be generalized to ergodic stationary stochastic processes satisfying Gordin's condition. The resulting test for the best alternative accounts for conditional heteroscedasticity and non-normality of asset returns in contrast to the Jobson-Korkie-Memmel test.

6.1. Testing for the Best Alternative

6.1.1. Basic Assumptions and Notation

Let $\mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$ be an unknown vector of quantities and we are searching for the *best alternative*

$$i^* := \arg\max_j \{\mu_j \colon\ j = 1, \ldots, d\}\,.$$

It is worth to mention that i^* does not need to be unique. That means there can be several equivalent and optimal alternatives. In contrast, let $i \in \{1, \ldots, d\}$ be our specific *choice*, i.e. we believe that there is no other alternative better than μ_i. We will set $i = 1$ for notational convenience and without loss of generality. Hence, we want to support the alternative hypothesis

$$H_1 \colon\ \mu_1 \geq \mu_2, \ldots, \mu_d$$

vs. the null hypothesis $H_0 \colon \neg H_1$. If we can reject H_0, our choice turns out to be significantly optimal among all given alternatives.

Let (X_n) be a sequence of d-dimensional random vectors such that

$$a_n\big(X_n - \mu\big) \xrightarrow{\mathrm{d}} \xi\,, \qquad n \longrightarrow \infty\,,$$

where (a_n) is some sequence of real numbers growing to infinity and ξ is a d-dimensional random vector. It is supposed that the cumulative distribution function (c.d.f.) of ξ does not depend on μ. By Cramér's theorem (Davidson, 1994, p. 355) it follows that $X_n \to_{\mathrm{p}} \mu$ as $n \to \infty$. Hence, we can think of X_n as a convenient approximation of μ if n is large. Due to the Central limit theorem (CLT) we will typically encounter $a_n = \sqrt{n}$ and ξ has a multivariate normal distribution with zero mean and covariance matrix Σ.

6.1.2. Test Procedure

A crucial point of the following test is that i must be fixed *without* examining X_n or say, more precisely, the choice must not depend on the data which are used for testing the aforementioned hypothesis. Otherwise the presented method would suffer from a selection bias. Indeed, this is not a serious drawback of the procedure. For instance, consider a Monte Carlo simulation. In that case we can simply run the process (X_n) a first time so as to choose the largest component of X_n, that is

$$i = \arg\max_j \{X_{jn} \colon\ j = 1, \ldots, d\}\,.$$

After that we start a new run of (X_n) and apply the following test with respect to the choice made by the first run. In case of historical data we can simply divide the overall sample into two sub-samples, i.e. a calibration and a validation sample. Then the choice can be made by using the calibration sample, whereas the test has to be applied to the validation sample.

Define the $(d-1) \times d$ matrix

$$\Delta := \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

and note that due to the Continuous mapping theorem (Davidson, 1994, p. 355) we obtain

$$a_n\big(\Delta X_n - \Delta\mu\big) \xrightarrow{\mathrm{d}} \Delta\xi\,, \qquad n \longrightarrow \infty\,.$$

Now the alternative hypothesis can be compactly written as $H_1 \colon \Delta\mu \geq 0$. In case $d = 2$ we will obtain a simple Gauss-type test for the null hypothesis $H_{02} \colon \mu_1 < \mu_2$. In the general multivariate case the global hypothesis H_1 can be supported whenever $H_{1j} : \mu_1 \geq \mu_j$ survives after each comparison with $j = 2, \ldots, d$. This is an important implication of the following theorem.

Theorem 6.1 *Let $\zeta = (\zeta_1, \ldots, \zeta_k)$ be a random vector and consider $Z = \eta + \zeta$ where $\eta \in \mathbb{R}^k$ but not $\eta \geq 0$. Let λ_j be the β-quantile of ζ_j for $j = 1, \ldots, k$ and $0 < \beta < 1$. Then $\mathbb{P}\big(Z > \lambda\big) \leq 1 - \beta$ with $\lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$.*

Proof. At least one component of η must be negative, say $\eta_j < 0$. Now the assertion follows immediately by noting that $\mathbb{P}(Z > \lambda) \leq \mathbb{P}(Z_j > \lambda_j) \leq 1 - \beta$. ■

In our case η represents $\Delta\mu$, $k = d - 1$, $\beta = 1 - \alpha$ with $0 < \alpha < 1$, $\zeta = \Delta\xi/a_n$, and $Z = \Delta X_n$. Hence, we can reject H_0 if $\Delta X_n > \lambda$ or, following the usual notation of large-sample theory, $T := a_n \Delta X_n > \tau$, where $\tau = (\tau_1, \ldots, \tau_{d-1}) := a_n\lambda$. The $(d-1) \times 1$ vector τ contains the $(1-\alpha)$-quantiles of $\Delta\xi$. Theorem 6.1 guarantees that our choice is significantly optimal among all given alternatives whenever it is significantly better or not worse than every other candidate on the *same* level α. That means if each pairwise test $H_0 \colon \mu_1 < \mu_j$ vs. $H_1 \colon \mu_1 \geq \mu_j$ possesses a significance level of α then the overall test $H_1 \colon \mu_1 \geq \mu_2, \ldots, \mu_d$ vs. $H_0 \colon \neg H_1$ works on the same level.

In many practical situations we do not know the exact c.d.f. of $\Delta\xi$. However, we can often calculate or simulate the c.d.f. of ξ_n, where (ξ_n) is some sequence of d-dimensional random vectors such that $\xi_n \to_{\mathrm{d}} \xi$ as $n \to \infty$. This can be used for a large-sample approximation of the critical thresholds $\tau_1, \ldots, \tau_{d-1}$. For instance, suppose that $X_1, \ldots, X_n$ is a sample of independent copies of a random vector X with mean vector μ and positive definite covariance matrix Σ. We assume that μ and Σ are unknown. From the CLT we know that

$$\sqrt{n} \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i - \mu \right) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \Sigma\big)\,, \qquad n \longrightarrow \infty\,.$$

For brevity we may denote the sample mean vector by $\overline{X}_n = \big(\overline{X}_{1n}, \ldots, \overline{X}_{dn}\big)$. Now we try to reject $H_{0j}\colon \mu_1 < \mu_j$ by applying the one-sided Gauss test

$$T_{j-1} := \sqrt{n} \cdot \big(\overline{X}_{1n} - \overline{X}_{jn} \big) > \tau_{j-1}\,,$$

for $j = 2, \ldots, d$, where

$$\tau_{j-1} := \sqrt{\sigma_1^2 + \sigma_j^2 - 2\sigma_{1j}} \cdot \Phi^{-1}(1-\alpha)\,.$$

Here σ_j^2 represents the variance of the jth component of X $(j = 1, \ldots, d)$, σ_{1j} is the covariance between its first and jth component $(j = 2, \ldots, d)$, and Φ^{-1} denotes the quantile function of the standard normal distribution. Note that the parameters of Σ are unknown but we can substitute Σ by the sample covariance matrix

$$\widehat{\Sigma}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} \big(X_i - \overline{X}_n\big)\big(X_i - \overline{X}_n\big)'$$

because – due to the i.i.d. assumption – the sample covariance matrix is strongly consistent for Σ. Hence, by the Cramér-Wold device (Davidson, 1994, p. 405) it follows that

$$\widehat{\Sigma}_n^{\frac{1}{2}} Y \xrightarrow{\mathrm{d}} \Sigma^{\frac{1}{2}} Y \sim \mathcal{N}(0, \Sigma)\,, \qquad n \longrightarrow \infty\,,$$

where $Y \sim \mathcal{N}(0, I_d)$ and $\Sigma^{\frac{1}{2}}$ denotes a $d \times d$ matrix such that $\Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}\prime} = \Sigma$. That means the critical thresholds $\tau_1, \ldots, \tau_{d-1}$ can be readily approximated by using the sample variances and covariances and we obtain the usual one-sided Gauss test for a joint sample, viz.

$$\frac{\overline{X}_{1n} - \overline{X}_{jn}}{\sqrt{\big(\hat{\sigma}_1^2 + \hat{\sigma}_j^2 - 2\hat{\sigma}_{1j}\big)/n}} > \Phi^{-1}(1-\alpha)\,.$$

If this inequality is satisfied for every $j = 2, \ldots, d$, the first alternative is significantly optimal among all given alternatives.

6.2. Application to Financial Data

6.2.1. General Conditions

Let $P_t \overset{\text{a.s.}}{>} 0$ be the price of an asset at time $t \in \mathbb{Z}$ so that $R_t := P_t/P_{t-1} - 1$ represents the corresponding asset return from $t-1$ to t. It is assumed that (R_t) is strongly stationary and ergodic with $\mathrm{E}(R_t) = \eta$ and $\mathrm{Var}(R_t) = \sigma^2 < \infty$. Ergodicity means that any existing and finite moment of R_t can be consistently estimated by the corresponding sample moment of (R_t). This is guaranteed if $(R_t, \ldots, R_{t+k})$ is asymptotically independent of $(R_{t-n}, \ldots, R_{t-n+l})$ as $n \to \infty$ for all $k, l \in \mathbb{N}$, whilst the components of the considered random vectors generally depend on each other (Hayashi, 2000, p. 101). For the CLT we need some additional restrictions. More precisely, the CLT holds for the sample mean of (R_t) if the centered process $(R_t - \eta)$ satisfies *Gordin's condition*. Let $\mathcal{H}_t := (R_t, R_{t-1}, \ldots)$ be the history of (R_t) at time $t \in \mathbb{Z}$. Roughly speaking, Gordin's condition implies that the impact of $\mathcal{H}_{t-n}$ on the conditional expectation of R_t vanishes as $n \to \infty$ and also that the conditional expectations of R_t do not vary too much in time (Hayashi, 2000, p. 403). In that case it is guaranteed that the CLT holds with an asymptotic or, say, *long-run variance* $\sigma_{\mathrm{L}}^2 := \sum_{k=-\infty}^{\infty} \gamma(k)$ (Hayashi, 2000, p. 401), where γ is the autocovariance function of (R_t). This can be easily extended to any d-dimensional stochastic process (Hayashi, 2000, p. 405) and applied to a broad class of standard time series models. There exist several alternative criteria for the CLT in the context of time series analysis which can be found, e.g., in Brockwell and Davis (1991, p. 213) and Hamilton (1994, p. 195). However, to my knowledge Gordin's condition represents the most unrestrictive set of assumptions concerning the serial dependence structure of a stochastic process (Eagleson, 1975).

It is worth to note that the number of dimensions d is supposed to be fixed or at least $n, d \to \infty$ such that $n/d \to \infty$. If n/d tends to a finite number, the CLT may become invalid and other interesting issues arise from *Random matrix theory* (Bai, 1999). However, if the number of observations relative to the number of assets is large enough, the sample mean is approximately normally distributed under the aforementioned conditions. We additionally assume that the asset return R_t possesses a finite fourth moment and that Gordin's condition is satisfied not only for $(R_t - \eta)$ but also for $\{(R_t - \eta)^2 - \sigma^2\}$. Consider the random variable $X := R/\sigma$ and suppose that the risk-free interest rate is constant and zero without loss of generality. The *Sharpe ratio* $\mu := \eta/\sigma$ (see, e.g., Campbell et al., 1997, p. 188) is frequently used as a performance measure in the finance literature. Now

I will derive a hypothesis test for judging whether a certain portfolio possesses the largest Sharpe ratio among a set of given portfolios. As mentioned before, this will be done under quite general assumptions about the serial dependence structure of the asset returns. This problem has been also addressed by Ledoit and Wolf (2008) as well as Schmid and Schmidt (2007) in a bivariate setting. However, note that the following test is motivated by the multivariate point of view and the primary goal is to avoid a selection bias.

6.2.2. Asymptotic Distributions

Concerning the sample mean $\hat{\eta}$ we obtain

$$\sqrt{n} \cdot \big(\hat{\eta} - \eta\big) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \sigma_{\mathrm{L}}^2\big)\,, \qquad n \longrightarrow \infty\,.$$

The sample variance $\hat{\sigma}^2$ represents a consistent estimator for the stationary variance σ^2 but for estimating the long-run variance σ_{L}^2 we need to estimate the autocovariance function γ of (R_t). Actually, there exist many ways for estimating long-run variances and covariances (Andrews, 1991, Ledoit and Wolf, 2008). This is not the primary concern of the present work and for the sake of simplicity we can choose a simple box-kernel type estimator, viz.

$$\hat{\sigma}_{\mathrm{L}}^2 := \hat{\sigma}^2 + 2\sum_{k=1}^{l} \hat{\gamma}(k)\,,$$

where $\hat{\gamma}$ is the sample autocovariance function of (R_t) (Hayashi, 2000, p. 142) and $l < n$. However, many empirical studies confirm that $\gamma(k) \approx \hat{\gamma}(k) \approx 0$ for $k \neq 0$ and so we can expect that $\hat{\sigma}_{\mathrm{L}}^2 \approx \hat{\sigma}^2$. The standard error of $\hat{\eta}$ is given by $\epsilon(\hat{\eta}) = \sigma_{\mathrm{L}}/\sqrt{n}$ and this can be estimated by $\hat{\epsilon}(\hat{\eta}) = \hat{\sigma}_{\mathrm{L}}/\sqrt{n}$.

Since $\{(R_t - \eta)^2 - \sigma^2\}$ satisfies Gordin's condition, the sample variance $\hat{\sigma}^2$ is also asymptotically normally distributed, viz.

$$\sqrt{n} \cdot \big(\hat{\sigma}^2 - \sigma^2\big) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \upsilon_{\mathrm{L}}\big)\,, \qquad n \longrightarrow \infty\,.$$

The long-run variance υ_{L} of the squared centered asset returns can be estimated by

$$\hat{\upsilon}_{\mathrm{L}} := \hat{\kappa}(0) + 2\sum_{k=1}^{l} \hat{\kappa}(k)\,,$$

where $\hat{\kappa}$ denotes the sample autocovariance function of $\{(R_t - \eta)^2\}$. Typically, asset returns are conditionally heteroscedastic and thus υ_{L} can become relatively large. This is also confirmed by the following empirical study. We consider monthly excess returns of the

	Canada	France	Germany	Italy	Japan	UK	USA
$\hat{\sigma}_{\mathrm{L}}^2/\hat{\sigma}^2$	1.1334	1.3834	1.2356	1.9596	2.1995	0.9883	1.0505
$\hat{\upsilon}_{\mathrm{L}}/\hat{\kappa}(0)$	2.1004	1.8611	2.3553	1.8195	2.0844	2.5268	2.0429

Table 6.1.: Estimated long-run variances divided by sample variances.

MSCI stock indices for the G7 countries Canada, France, Germany, Italy, Japan, UK and USA from January 1970 to September 2006. The sample size corresponds to $n = 456$ and the risk-free interest rate is calculated by the secondary market 3-month US treasury bill rate. Further, the considered indices are adjusted by dividends, splits, etc. and are calculated on the basis of USD stock prices.

For estimating the long-run variances we have to choose an appropriate lag length $l \in \mathbb{N}$. Figure 6.1 shows the empirical autocorrelations for the squared centered excess returns of the MSCI indices and the *equally weighted portfolio* (EWP) up to $l = 12$. The Ljung-Box test leads to a rejection of the null hypothesis $H_0 \colon \rho(1) = \ldots = \rho(12) = 0$ in every case except for the EWP, France, and Italy. That means there is a strong evidence of conditional heteroscedasticity for monthly asset returns and we may choose $l = 12$ as an appropriate lag length. Now, Table 6.1 contains the estimated long-run variances divided by the corresponding sample variances. In most cases the long-run variances of the asset returns roughly correspond to the stationary variances, whereas the long-run variances of the squared asset returns are quite twice as large as the stationary ones. Hence, it is not appropriate to ignore the effect of heteroscedasticity when analyzing the volatility of monthly asset returns.

By applying the well-known 'delta method' we obtain

$$\sqrt{n} \cdot \left(\hat{\sigma} - \sigma\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\upsilon_{\mathrm{L}}}{4\sigma^2}\right), \qquad n \longrightarrow \infty \,.$$

The standard error of $\hat{\sigma}$ is given by $\epsilon(\hat{\sigma}) := \sqrt{\upsilon_{\mathrm{L}}}\,/(2\sqrt{n}\,\sigma)$ and its estimator can be denoted by $\sqrt{\hat{\upsilon}_{\mathrm{L}}}\,/(2\sqrt{n}\,\hat{\sigma})$. The Sharpe ratio can be estimated by $\hat{\mu} := \hat{\eta}/\hat{\sigma}$ which is also asymptotically normally distributed since

$$\sqrt{n} \cdot \left(\begin{bmatrix} \hat{\eta} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \eta \\ \sigma^2 \end{bmatrix} \right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \begin{bmatrix} \sigma_{\mathrm{L}}^2 & \varrho_{\mathrm{L}} \\ \varrho_{\mathrm{L}} & \upsilon_{\mathrm{L}} \end{bmatrix}\right), \qquad n \longrightarrow \infty \,,$$

where ϱ_{L} represents the long-run covariance between R_t and $(R_t - \eta)^2$. After applying

	EWP	Canada	France	Germany	Italy	Japan	UK	USA
$\hat{\eta}$	.0051 (.0026)	.0048 (.0027)	.0062 (.0035)	.0053 (.0031)	.0036 (.0047)	.0058 (.0044)	.0061 (.0030)	.0042 (.0021)
$\hat{\sigma}$	.0437 (.0025)	.0545 (.0038)	.0640 (.0040)	.0603 (.0042)	.0718 (.0039)	.0633 (.0035)	.0638 (.0091)	.0436 (.0029)
$\hat{\mu}$	.1177 (.0620)	.0886 (.0509)	.0967 (.0550)	.0880 (.0523)	.0507 (.0646)	.0909 (.0696)	.0955 (.0479)	.0959 (.0503)

Table 6.2.: Means, standard deviations, and Sharpe ratios for the monthly excess returns of the G7 MSCI indices and the EWP.

once again the delta method we obtain

$$\sqrt{n}\cdot\left(\hat{\mu}-\mu\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,\frac{\sigma_{\mathrm{L}}^2}{\sigma^2}-\frac{\mu\varrho_{\mathrm{L}}}{\sigma^3}+\frac{\mu^2\upsilon_{\mathrm{L}}}{4\sigma^4}\right),\qquad n\longrightarrow\infty\,,$$

and the standard error of $\hat{\mu}$ can be estimated in the same manner as $\epsilon(\hat{\eta})$ or $\epsilon(\hat{\sigma})$. Schmid and Schmidt (2007) obtain the same asymptotic variance under the assumption of a so-called 'α-mixing process'. As already mentioned this assumption is somewhat more restrictive than Gordin's condition. Schmid and Schmidt (2007) also provide closed-form expressions for the asymptotic variance of the Sharpe ratio in case of a stochastic volatility and a GARCH model.

Table 6.2 contains the estimated means, standard deviations, and Sharpe ratios for the monthly excess returns of the G7 MSCI indices and the EWP. The corresponding standard error estimates $\hat{\epsilon}(\hat{\eta})$, $\hat{\epsilon}(\hat{\sigma})$, and $\hat{\epsilon}(\hat{\mu})$ are given in the parentheses. Obviously, the standard errors for the Sharpe ratios are big despite of the large number of observations. This is a common problem in performance measurement. Now we want to derive an appropriate hypothesis test for the best alternative, i.e. the best performing asset. Without any previous look at the data we may expect that the EWP possesses the largest Sharpe ratio due to the effect of *international diversification* (see, e.g., Jorion, 1985). That means the variance of the EWP return should be relatively small. Indeed, this can be verified in Table 6.2. Hence, the EWP may serve as the benchmark portfolio and we want to know if its estimated Sharpe ratio $\hat{\mu}_1 = 0.1177$ is *significantly* larger (or at least not smaller) than any other Sharpe ratio.

Also the 2-dimensional random vector $(\hat{\mu}_1, \hat{\mu}_j)$ $(j = 2, \ldots, d)$ is asymptotically normally distributed, i.e.

$$\sqrt{n}\cdot\left(\begin{bmatrix}\hat{\mu}_1\\ \hat{\mu}_j\end{bmatrix}-\begin{bmatrix}\mu_1\\ \mu_j\end{bmatrix}\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,\begin{bmatrix}\vartheta_1^2 & \vartheta_{1j}\\ \vartheta_{j1} & \vartheta_j^2\end{bmatrix}\right),\qquad n\longrightarrow\infty\,.$$

After some calculation we obtain

$$\vartheta_{1j} = \frac{\omega_{\mathrm{L}1j}}{\sigma_1\sigma_j} - \frac{\mu_j\sigma_1\omega_{\mathrm{L}2j} + \mu_1\sigma_j\omega_{\mathrm{L}3j}}{2\sigma_1^2\sigma_j^2} + \frac{\mu_1\mu_j\omega_{\mathrm{L}4j}}{4\sigma_1^2\sigma_j^2}$$

for $j = 2, \ldots, d$. Here $\omega_{\mathrm{L}1j}$ represents the long-run covariance between R_{1t} and R_{jt}, $\omega_{\mathrm{L}2j}$ is the long-run covariance between R_{1t} and $(R_{jt} - \eta_j)^2$, $\omega_{\mathrm{L}3j}$ is the long-run covariance between $(R_{1t} - \eta_1)^2$ and R_{jt}, whereas $\omega_{\mathrm{L}4j}$ is the long-run covariance between $(R_{1t} - \eta_1)^2$ and $(R_{jt} - \eta_j)^2$. Now it follows that

$$\sqrt{n} \cdot \left\{(\hat{\mu}_1 - \hat{\mu}_j) - (\mu_1 - \mu_j)\right\} \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \vartheta_1^2 + \vartheta_j^2 - 2\vartheta_{1j}\big)\,, \qquad n \longrightarrow \infty\,.$$

Table 6.3 contains the values of the test statistic, i.e. $T_{j-1} = \sqrt{n} \cdot (\hat{\mu}_1 - \hat{\mu}_j)$ for $j = 2, \ldots, 8$, the standard errors calculated on the basis of the long-run variances and covariances, and the corresponding 'p-values'. There exists no country with a Sharpe ratio being significantly smaller than the Sharpe ratio of the EWP.

The Jobson-Korkie-Memmel test (Jobson and Korkie, 1981, Memmel, 2003) is frequently used in the finance literature for comparing the Sharpe ratios of two asset portfolios. For applying this test we have to assume that the asset returns are serially independent and multivariate normally distributed. In that case there is no need to distinguish between long-run, stationary, and conditional variances and covariances of asset returns since these quantities simply coincide. That means $\sigma_{\mathrm{L}1}^2 = \sigma_1^2$, $\sigma_{\mathrm{L}j}^2 = \sigma_j^2$, and $\omega_{\mathrm{L}1j} = \sigma_{1j}$ $(j = 2, \ldots, d)$. Further, by applying some standard results of multivariate analysis (see, e.g., Muirhead, 1982, p. 43) we obtain $\varrho_{\mathrm{L}1} = \varrho_{\mathrm{L}j} = 0$, $\upsilon_{\mathrm{L}1} = 2\sigma_1^4$, $\upsilon_{\mathrm{L}j} = 2\sigma_j^4$, $\omega_{\mathrm{L}2j} = \omega_{\mathrm{L}3j} = 0$, and $\omega_{\mathrm{L}4j} = 2\sigma_{1j}^2$ $(j = 2, \ldots, d)$ so that

$$\sqrt{n} \cdot \big((\hat{\mu}_1 - \hat{\mu}_j) - (\mu_1 - \mu_j)\big) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 2\,(1 - \rho_{1j}) + \frac{\mu_1^2 + \mu_j^2 - 2\mu_1\mu_j\rho_{1j}^2}{2}\right)$$

as $n \to \infty$, where $\rho_{1j} := \sigma_{1j}/(\sigma_1\sigma_j)$ for $j = 2, \ldots, d$. The latter expression for the asymptotic variance can be found also in Memmel (2003).

Table 6.4 once again contains the values of the test statistic T_{j-1} and the corresponding standard errors, but now calculated on the basis of sample variances and covariances according to the Jobson-Korkie-Memmel test. The star indicates that the corresponding Sharpe ratio difference is significantly nonnegative on a 5% level. We conclude that the MSCI index 'Italy' appears to be significantly worse than the EWP of all MSCI indices. However, this result is based on the wrong assumption of normality and serial independence of monthly asset returns. All in all it seems to be very difficult to validate portfolio

	Canada	France	Germany	Italy	Japan	UK	USA
T	.6214 (.9305)	.4478 (.5473)	.6355 (.7453)	1.4320 (.9452)	0.5729 (1.0180)	.4739 (.7208)	.4661 (.8499)
p	.2521	.2066	.1969	.0649	.2868	.2554	.2917

Table 6.3.: Performance test based on long-run variances and covariances.

	Canada	France	Germany	Italy	Japan	UK	USA
T	.6214 (.7413)	.4478 (.6058)	.6355 (.6972)	1.4320* (.7915)	0.5729 (.8599)	.4739 (.7066)	.4661 (.7614)
p	.2009	.2299	.1810	.0352	.2526	.2512	.2702

Table 6.4.: Jobson-Korkie-Memmel performance test.

strategies only by historical data. Instead, the strategies should be extensively validated by the application of Monte Carlo methods (see, e.g., Memmel, 2004, Section 5.2) rather than historical simulation. We can use the presented hypothesis test to judge whether a suggested portfolio strategy dominates some other strategies significantly, as already mentioned in Section 6.1.2.

6.3. Conclusion

In many practical situations we are searching for the best alternative among several candidates. If our decision is based on historical or simulated data there is some sort of selection bias and it is not evident if our choice is significantly optimal over all given alternatives. This problem frequently occurs in statistical inference or decisions under uncertainty such as portfolio optimization. Of course, such kind of decisions have to be reliable and thus we need a strong statistical fundament to justify our choice. In the present work a large-sample test for the best alternative has been derived in a rather general setting and it has been demonstrated by an application to financial data. It was shown that the traditional Jobson-Korkie-Memmel test can be generalized to ergodic stationary stochastic processes satisfying Gordin's condition. The presented hypothesis test accounts for conditional heteroscedasticity and non-normality of asset returns. We find that ignoring these kinds of stylized facts of empirical finance can lead to false rejections of the null hypothesis and misleading decisions.

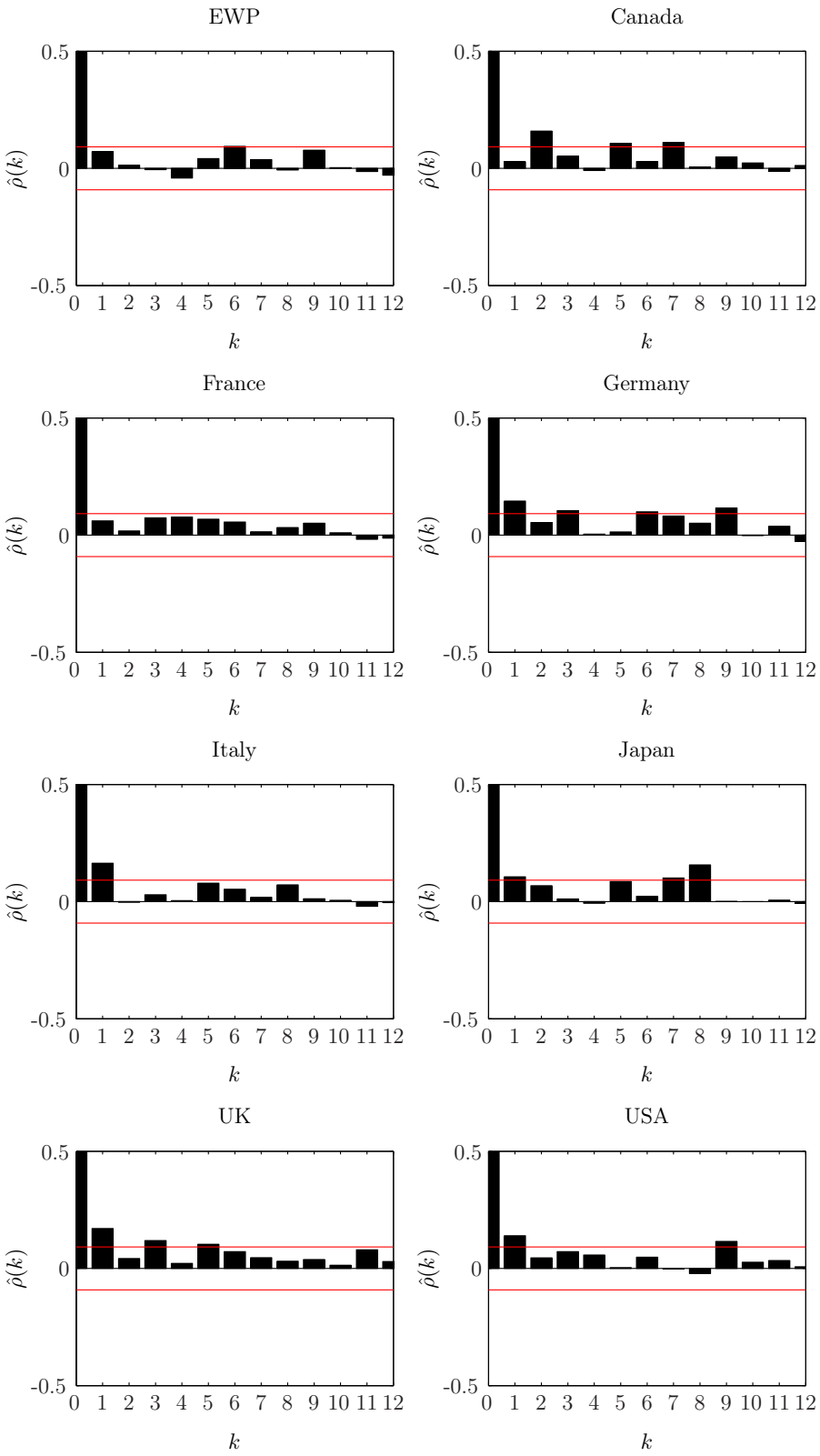

Figure 6.1.: Correlograms for the squared centered excess returns of the G7 MSCI indices and the EWP. The critical thresholds for the null hypothesis $H_0\colon \rho(k) = 0$ $(k \neq 0)$ on the 5% level are indicated by the horizontal lines.

Chapter 7.

Asymptotic Distributions of Robust Shape Matrices and Scales

7.1. Motivation

Since the seminal paper by Maronna (1976), covariance matrix estimation has become a popular branch of robust statistics. Several techniques have been developed for calculating the asymptotic distributions of robust covariance matrix estimators such as the radial distribution approach of Tyler (1982) and the approach based on influence functions (Hampel et al., 1986). Moreover, in recent years deep insights have been gained from the viewpoint of local asymptotic normality (LAN) theory (Hallin et al., 2006, Hallin and Paindaveine, 2006a,b).

Let X be a d-dimensional random vector possessing an elliptically symmetric distribution, i.e. it can be represented by $X = \mu + \Lambda\mathcal{R}U$, where U is a k-dimensional random vector, uniformly distributed on the unit hypersphere, $\mathcal{R}$ is a nonnegative random variable that is stochastically independent of U, $\mu \in \mathbb{R}^d$, and $\Lambda \in \mathbb{R}^{d\times k}$ (Cambanis et al., 1981, Fang et al., 1990, p. 42). It is assumed that $\mathcal{R}$ and U are unobservable quantities. The positive-semidefinite matrix $\Sigma := \Lambda\Lambda'$ is called the *dispersion matrix* and $\mathcal{R}$ is the *generating variate* of X. If $\mathrm{E}(\mathcal{R}^2) < \infty$, the covariance matrix of X is given by $\mathrm{Var}(X) = \mathrm{E}(\mathcal{R}^2)/k \cdot \Sigma$, whereas if $\mathrm{E}(\mathcal{R}^2) = \infty$, the linear dependence structure of X can be further described using the dispersion matrix Σ although $\mathrm{Var}(X)$ is not defined.

In general I will assume that Σ is positive-definite, i.e. $\mathrm{r}(\Lambda) = d \leq k$. In the robust statistics literature (Tyler, 1982, Bilodeau and Brenner, 1999, Ch. 13) and in the context of LAN theory (Hallin and Paindaveine, 2006a, Paindaveine, 2008) it is often supposed that

the distribution of $\mathcal{R}$ is absolutely continuous. Then the density of X can be written as $p(x) = \sqrt{\det \Sigma^{-1}}\, g\{(x-\mu)'\Sigma^{-1}(x-\mu)\}$, where the so-called *density generator* $g\colon \mathbb{R}^+ \to \mathbb{R}_0^+$ depends on x only through the quadratic form $(x-\mu)'\Sigma^{-1}(x-\mu)$. It can be shown (Frahm, 2004, p. 9) that the density function of $\mathcal{R}$ is given by $f(r) \propto r^{d-1} g(r^2)$.

Tatsuoka and Tyler (2000) wrote that 'The assumption of an elliptically symmetric distribution is often made simply because of its mathematical tractability'. Nevertheless, the class of elliptically symmetric distributions is a natural extension of the multivariate normal distribution. Moreover, the elliptical distribution assumption is fundamental in multivariate analysis and the results presented in this work generally require that the data are elliptically symmetrically distributed. However, there is one exception where the data are only assumed to have a *generalized elliptical distribution* (Frahm, 2004, Ch. 3). This will be treated in more detail below.

Note that $X = \mu + \Lambda\mathcal{R}U = \mu + V\mathcal{S}U$ with $\mathcal{S} := \mathcal{R}/\tau$, $V := \tau\Lambda$, and $\tau > 0$. This means that if X possesses the dispersion matrix Σ, there always exists an equivalent representation of X with dispersion matrix $\tau^2\Sigma$, so this can be only identified if the distribution of $\mathcal{R}$ is somehow restricted. However, many multivariate statistical methods like principal components analysis, canonical correlation analysis, linear discriminant analysis, and multivariate regression require the covariance or dispersion matrix only up to some scaling constant. This has been frequently observed in the literature (Croux and Haesbroeck, 1999, Hallin and Paindaveine, 2006a, Oja, 2003, Paindaveine, 2008, Taskinen et al., 2006). If the topic of interest is not the scale but only the *shape* of the distribution of X, it is not meaningful to focus on the asymptotic covariance matrix (ACM) of an estimator for Σ, $\mathrm{Var}(X)$ or another matrix $\Gamma \propto \Sigma$ (i.e. $\Gamma = \tau^2\Sigma$, where τ is a *constant* and thus not determined by Σ).

Therefore I will concentrate on robust estimators for the *shape matrix* of X (Oja, 2003, Paindaveine, 2008). The associated estimators for the scale are investigated concomitantly. I will derive explicit expressions for their joint asymptotic distributions. The paper is organized as follows. Section 7.2 introduces the notation and provides some helpful prerequisites concerning homogeneous functions. The question of how to choose an appropriate scale is investigated in Section 7.3. This section also contains the main results concerning the joint asymptotic distributions of estimators for the shape matrix and scale. In Section 7.4 it is shown how to calculate the asymptotic distributions of such estimators on the basis of some well-known robust covariance matrix estimators, namely M-, R-, and S-estimators.

7.2. Prerequisites

7.2.1. Notation

The following notation will be used in the sequel. The $d^2 \times d^2$ identity matrix is symbolized by I_{d^2}. Let $\mathbf{e}_{ij}$ be the $d \times d$ matrix with 1 in the ijth position and zeros elsewhere. The $d^2 \times d^2$ matrix J_{d^2} is defined as $J_{d^2} := \sum_{i=1}^{d} \mathbf{e}_{ii} \otimes \mathbf{e}_{ii}$, where '$\otimes$' denotes the Kronecker product (Schott, 1997, p. 253). The $n \times m$ matrix A' denotes the transpose of an $m \times n$ matrix A. In contrast, if f is an $\mathbb{R}$-valued function on an open subset of $\mathbb{R}$, then $f'(x)$ stands for the derivative of f at $x \in \mathbb{R}$. Further, the *commutation matrix* K_{d^2} is the $d^2 \times d^2$ matrix given by $K_{d^2} := \sum_{i,\,j=1}^{d} \mathbf{e}_{ij} \otimes \mathbf{e}_{ji}$ (Schott, 1997, p. 277).

For any symmetric $d \times d$ matrix A, the d^2-dimensional vector $\operatorname{vec} A$ is obtained by stacking the columns of A on top of each other, whereas $\operatorname{vech} A$ denotes the $d(d+1)/2$-dimensional vector obtained by stacking only the elements of the lower triangular part of A. Further, the *duplication matrix* is the $d^2 \times d(d+1)/2$ matrix D_d such that $D_d \operatorname{vech} A = \operatorname{vec} A$ (Schott, 1997, p. 283). Then it holds that $D_d^{+} \operatorname{vec} A = \operatorname{vech} A$, where the $d(d+1)/2 \times d^2$ matrix D_d^{+} is the Moore-Penrose inverse of D_d (Schott, 1997, p. 284). Let I_0 be defined as the $\{d(d+1)/2 - 1\} \times d(d+1)/2$ matrix $I_0 := [\,0 \;\; I_{d(d+1)/2-1}\,]$ and $N_d := I_0 D_d^{+}$, so that $\operatorname{vech}_0 A := N_d \operatorname{vec} A$ is the vech of A deprived of its first component A_{11} (Hallin and Paindaveine, 2006a).

I will frequently calculate the differential of an $\mathbb{R}^m$-valued function f, i.e. $\mathrm{d}f = \mathcal{J}_f \partial x$, where $\mathcal{J}_f := \partial f(x)/\partial x' \in \mathbb{R}^{m \times n}$ denotes the Jacobi matrix of f at $x \in \mathbb{R}^n$. Suppose that x represents the vec of a symmetric matrix. Then each off-diagonal element in the lower triangular part of that matrix represents an implicit function of the corresponding off-diagonal element in the upper triangular part and vice versa. However, I will not take the symmetry into consideration when calculating the partial derivatives of f. Otherwise, to adjust for the redundancies caused by the symmetry it would be necessary to apply the operator $(I_{d^2} + J_{d^2})/2$ on the partial differentials ∂x when calculating the total differential $\mathrm{d}f$. Hence, to avoid additional notation and tedious calculations of implicit derivatives, the Jacobi matrix $\mathcal{J}_f$ is understood to be the matrix of partial derivatives of f which are obtained by ignoring the symmetry condition. In the present context this poses no problem since $\mathcal{J}_f$ is always used only in combination with ∂x.

7.2.2. Homogeneous Functions

Consider a differentiable $\mathbb{R}^m$-valued function h of $x \in \mathbb{R}^n$. The function h is said to be *homogeneous* of degree $\nu \in \mathbb{R}$ if $h(\alpha x) = \alpha^\nu h(x)$ for all $x \in \mathbb{R}^n$ and $\alpha > 0$. Due to the Euler relation it holds that $\mathcal{J}_h x = \nu h(x)$. A function f is said to be *scale-invariant* if it is homogeneous of degree 0, i.e. $f(\alpha x) = f(x)$ for all $\alpha > 0$. This means that $\mathcal{J}_f x = 0$ and if h is homogeneous of degree 1, it holds that $\mathcal{J}_h x = h(x)$. In the following a homogeneous function is always understood to be homogeneous of degree 1. Note that the partial derivatives of any homogeneous function are scale-invariant.

Let $\mathcal{P}^d$ be the set of all symmetric positive-definite $d \times d$ matrices and $\varphi\colon \mathcal{P}^d \to \mathbb{R}^k$ a scale-invariant function, i.e. $\varphi(\alpha\Gamma) = \varphi(\Gamma)$ for all $\alpha > 0$ and $\Gamma \in \mathcal{P}^d$. In particular, consider a scale-invariant function $\Omega(\Gamma) = \Gamma/\sigma^2(\Gamma)$, where $\sigma^2 : \mathcal{P}^d \to \mathbb{R}^+$ is an homogeneous function, i.e. $\sigma^2(\alpha\Gamma) = \alpha\sigma^2(\Gamma) > 0$. It is supposed that the so-called *scale function* σ^2 is differentiable at any point $\Gamma \in \mathcal{P}^d$ and also that $\sigma^2(I_d) = 1$. Then $\sigma^2(\Gamma)$ is called the *scale* of Γ. The matrix $\Omega(\Gamma)$ will be called the *shape matrix* (with respect to the scale function σ^2) belonging to Γ. I will write $\sigma^2 \equiv \sigma^2(\Gamma)$ and $\Omega \equiv \Omega(\Gamma)$ whenever these quantities cannot be confused with the corresponding functions.

Note that $\sigma^2(\Omega) = 1$ and $\varphi \circ \Omega = \varphi$, since $\varphi\{\Omega(\Gamma)\} = \varphi\{\Gamma/\sigma^2(\Gamma)\} = \varphi(\Gamma)$. For instance, the correlation matrix produced by Γ is scale-invariant and thus it can be derived from any shape matrix Ω. Hence, whenever Ω_n is an estimator for Ω, an estimator for $\varphi(\Gamma)$ is simply given by $\varphi(\Omega_n)$. This is a formal justification of directing one's attention to shape matrices (Frahm and Jaekel, 2007a, Hallin and Paindaveine, 2006a, Oja, 2003, Paindaveine, 2008, Taskinen et al., 2006). General robustness and efficiency properties of scale-invariant functions have been investigated by Tyler (1983).

7.3. Asymptotic Distributions

7.3.1. The Choice of the Scale Function

In most cases asymptotic normality of robust estimators μ_n and Γ_n for the mean vector and covariance matrix can be guaranteed by the usual regularity conditions given in the robust statistics literature. Typically μ_n and Γ_n are also asymptotically independent. In the present work it is shown that the asymptotic independence of an estimator Ω_n for the shape matrix and an associated estimator σ_n^2 for the scale can only be guaranteed for one

and only one scale function σ^2. A similar result in the context of LAN theory has been obtained by Paindaveine (2008) (see below).

Let Γ_n be some estimator for $\Gamma \propto \Sigma$ where n represents the sample size. The corresponding shape matrix estimator is given by $\Omega_n := \Gamma_n/\sigma^2(\Gamma_n)$. At first glance the choice of the scale function σ^2 might be considered as arbitrary and the following variants can often be observed in the literature (Paindaveine, 2008):

(**S1**) Frahm (2004, p. 64), Hallin et al. (2006), Hallin and Paindaveine (2006b), Hettmansperger and Randles (2002) as well as Randles (2000) simply choose $\sigma^2(\Gamma) = \Gamma_{11}$ so that $\Omega_{11} = 1$.

(**S2**) Dümbgen (1998), Frahm and Jaekel (2007b) as well as Tyler (1987a) take the scale function $\sigma^2(\Gamma) = (\operatorname{tr}\Gamma)/d$ so that $\operatorname{tr}\Omega = d$.

(**S3**) Dümbgen and Tyler (2005), Hallin and Paindaveine (2008, 2009), Paindaveine (2008), Salibian-Barrera et al. (2006), Taskinen et al. (2006) as well as Tatsuoka and Tyler (2000) postulate $\sigma^2(\Gamma) = (\det\Gamma)^{1/d}$ so that $\det\Omega = 1$.

Paindaveine (2008) considers the latter normalization as *canonical* since this is the only one where the Fisher information matrix with respect to the mean vector, shape matrix and scale is block diagonal if the distribution of X or, more precisely, the corresponding experiment is LAN (van der Vaart, 1998, Ch. 7).

The scale functions defined by **S2** and **S3** correspond to the arithmetic and geometric means of the eigenvalues of Γ, respectively. Hence, another possible scale function is given by the harmonic mean of the eigenvalues of Γ, i.e.

(**S4**) $\sigma^2(\Gamma) = d/(\operatorname{tr}\Gamma^{-1})$ so that $\operatorname{tr}\Omega^{-1} = d$.

It is worth pointing out that shape matrices are not affine equivariant, since

$$\Omega(V\Gamma V') = \frac{V\Gamma V'}{\sigma^2(V\Gamma V')} = \frac{\sigma^2(\Gamma)}{\sigma^2(V\Gamma V')} \cdot V\Omega(\Gamma)V'$$

for any nonsingular $d \times d$ matrix V and generally $\sigma^2(\Gamma)$ does not correspond to $\sigma^2(V\Gamma V')$. This is not surprising because even after an affine-linear transformation of the data, the shape matrix has to satisfy the scaling condition $\sigma^2(\Omega) = 1$ and so the equality $\Omega(V\Gamma V') = V\Omega(\Gamma)V'$ cannot be guaranteed in general. However, a natural requirement is that the

equivariance property holds at least for all transformations V with $\sigma^2(VV') = 1$. This means that if not the scale but only the shape of the distribution of X is affected by V, the shape matrix should remain equivariant.

More generally, it can be required (Tyler, 2002) that

$$\Omega(V\Gamma V') = \frac{V\Omega(\Gamma)V'}{\sigma^2(VV')}\,,$$

i.e. $\sigma^2(V\Gamma V') = \sigma^2(VV')\,\sigma^2(\Gamma)$. Interestingly, from the scale functions considered in **S1**–**S4** only the canonical one (**S3**) satisfies this kind of affine equivariance property. This is another argument in favor of the determinant-based normalization proposed by Paindaveine (2008).

The previous arguments as well as a thorough discussion in Hallin and Paindaveine (2006a) show that the choice of the scale function must be driven by statistical considerations and should be handled carefully.

Lemma 7.1 *Let $\Omega(\Gamma) = \Gamma/\sigma^2(\Gamma)$ be a $d \times d$ shape matrix and σ^2 a scale function. Then*

$$\mathcal{J}_\Omega := \frac{\partial \operatorname{vec}\Omega(\Gamma)}{\partial(\operatorname{vec}\Gamma)'} = \frac{1}{\sigma^2}\left\{I_{d^2} - \operatorname{vec}\Omega\,\mathcal{J}_{\sigma^2}\right\},$$

where

$$\mathcal{J}_{\sigma^2} := \frac{\partial\sigma^2(\Gamma)}{\partial(\operatorname{vec}\Gamma)'} = \frac{\partial\sigma^2(\Omega)}{\partial(\operatorname{vec}\Omega)'}\,.$$

Proof. By the product rule it follows that

$$\mathcal{J}_\Omega = \frac{1}{\sigma^2}\cdot\frac{\partial\operatorname{vec}\Gamma}{\partial(\operatorname{vec}\Gamma)'} - \frac{\operatorname{vec}\Gamma}{\sigma^4}\cdot\mathcal{J}_{\sigma^2} = \frac{1}{\sigma^2}\left\{I_{d^2} - \operatorname{vec}\Omega\,\mathcal{J}_{\sigma^2}\right\}.$$

Since the partial derivatives of a homogeneous function are scale-invariant, it holds that $\mathcal{J}_{\sigma^2} = \partial\sigma^2(\Omega)/\partial(\operatorname{vec}\Omega)'$. ∎

In the following I will write $\Psi := I_{d^2} - \operatorname{vec}\Omega\,\mathcal{J}_{\sigma^2}$ for notational convenience.

7.3.2. Main Results

Let $\mathcal{Q}$ be a symmetric random $d \times d$ matrix. A symmetric random $d \times d$ matrix $\mathcal{M}$ is said to possess a *radial distribution* if $\mathcal{O}\mathcal{M}\mathcal{O}' \sim \mathcal{M}$ for any orthogonal $d \times d$ matrix $\mathcal{O}$ (Tyler, 1982). In the following let $\mathcal{N}$ be a symmetric random $d \times d$ matrix with finite second moments. It is supposed that $\mathcal{N}$ is of the radial type with respect to a symmetric positive-definite $d \times d$ matrix Γ. This means that $T\mathcal{N}T'$ has a radial distribution whenever the $d \times d$

matrix T is such that $T'T = \Gamma^{-1}$. Further, let (Γ_n) be a sequence of symmetric positive-definite random $d \times d$ matrices and (σ_n^2) an associated sequence with $\sigma_n^2 := \sigma^2(\Gamma_n)$, where σ^2 is a scale function. Moreover, consider the sequence (Ω_n) of symmetric positive-definite random $d \times d$ matrices with $\Omega_n := \Gamma_n/\sigma_n^2$.

Theorem 7.2 *Let σ^2 be a scale function and $\Omega \equiv \Omega(\Gamma) = \Gamma/\sigma^2(\Gamma)$ the shape matrix belonging to Γ. Further, let (a_n) be a sequence of real numbers increasing to infinity such that $a_n(\operatorname{vec}\Gamma_n - \operatorname{vec}\Gamma) \to_{\mathrm{d}} \operatorname{vec}\mathcal{Q}$ as $n \to \infty$ with $\mathrm{E}(\operatorname{vec}\mathcal{Q}) = 0$ and*

$$\operatorname{Var}(\operatorname{vec}\mathcal{Q}) = \gamma_1(I_{d^2} + K_{d^2})(\Gamma \otimes \Gamma) + \gamma_2(\operatorname{vec}\Gamma)(\operatorname{vec}\Gamma)', \tag{7.1}$$

where $\gamma_1 \geq 0$ and $\gamma_2 \geq -2\gamma_1/d$. Then it follows that

$$a_n\left(\begin{bmatrix}\sigma_n^2\\ \operatorname{vec}\Omega_n\end{bmatrix} - \begin{bmatrix}\sigma^2\\ \operatorname{vec}\Omega\end{bmatrix}\right) \xrightarrow{\mathrm{d}} \xi\,, \qquad n \longrightarrow \infty\,,$$

where $\sigma^2 \equiv \sigma^2(\Gamma)$, ξ is a (d^2+1)-dimensional random vector with $\mathrm{E}(\xi) = 0$, and

$$\operatorname{Var}(\xi) = \begin{bmatrix}\mathcal{V}(\sigma_n^2) & \mathcal{V}(\sigma_n^2, \Omega_n)\\ \mathcal{V}(\sigma_n^2, \Omega_n)' & \mathcal{V}(\Omega_n)\end{bmatrix}.$$

More specifically,

$$\mathcal{V}(\sigma_n^2) = \sigma^4\left\{2\gamma_1\mathcal{J}_{\sigma^2}(\Omega \otimes \Omega)\mathcal{J}_{\sigma^2}' + \gamma_2\right\}$$

with $\mathcal{J}_{\sigma^2} = \partial\sigma^2(\Omega)/\partial(\operatorname{vec}\Omega)'$ and $\sigma^4 = \{\sigma^2(\Gamma)\}^2$,

$$\mathcal{V}(\sigma_n^2, \Omega_n)' = 2\gamma_1\sigma^2\,\Psi(\Omega \otimes \Omega)\mathcal{J}_{\sigma^2}'\,,$$

with $\Psi = I_{d^2} - \operatorname{vec}\Omega\,\mathcal{J}_{\sigma^2}$, and

$$\mathcal{V}(\Omega_n) = \gamma_1\Psi(I_{d^2} + K_{d^2})(\Omega \otimes \Omega)\Psi'\,.$$

Proof. The vector $\{\sigma^2(\Gamma), \operatorname{vec}\Omega(\Gamma)\}$ is differentiable at $\operatorname{vec}\Gamma$ and thus

$$a_n\left(\begin{bmatrix}\sigma_n^2\\ \operatorname{vec}\Omega_n\end{bmatrix} - \begin{bmatrix}\sigma^2\\ \operatorname{vec}\Omega\end{bmatrix}\right) \xrightarrow{\mathrm{d}} \xi := \mathcal{J}_{\sigma^2,\,\Omega}\operatorname{vec}\mathcal{Q}\,, \qquad n \longrightarrow \infty\,,$$

where $\mathcal{J}_{\sigma^2,\,\Omega}$ is defined as $\partial\{\sigma^2(\Gamma), \operatorname{vec}\Omega(\Gamma)\}/\partial(\operatorname{vec}\Gamma)'$. From $\mathrm{E}(\operatorname{vec}\mathcal{Q}) = 0$ it follows that $\mathrm{E}(\xi) = 0$ and the variance of the first element of ξ is given by $\mathcal{V}(\sigma_n^2) = \mathcal{J}_{\sigma^2}\operatorname{Var}(\operatorname{vec}\mathcal{Q})\mathcal{J}_{\sigma^2}'$. Since σ^2 is a homogeneous function it holds that $\mathcal{J}_{\sigma^2}\operatorname{vec}\Gamma = \sigma^2$. Note also that $\mathcal{J}_{\sigma^2}(I_{d^2} + K_{d^2}) = 2\mathcal{J}_{\sigma^2}$ and thus

$$\mathcal{V}(\sigma_n^2) = 2\gamma_1\mathcal{J}_{\sigma^2}(\Gamma \otimes \Gamma)\mathcal{J}_{\sigma^2}' + \gamma_2\sigma^4 = \sigma^4\left\{2\gamma_1\mathcal{J}_{\sigma^2}(\Omega \otimes \Omega)\mathcal{J}_{\sigma^2}' + \gamma_2\right\}.$$

Similarly, the covariances between the first element of ξ and its residual elements are given by $\mathcal{V}(\sigma_n^2, \Omega_n) = \mathcal{J}_{\sigma^2}\mathrm{Var}(\mathrm{vec}\,\mathcal{Q})\Psi'/\sigma^2$. Since Ω is a scale-invariant function of Γ, due to Euler's relation it holds that $(\mathrm{vec}\,\Gamma)'\Psi' = 0$ and thus

$$\mathcal{V}(\sigma_n^2, \Omega_n) = \gamma_1 \mathcal{J}_{\sigma^2}(I_{d^2} + K_{d^2})(\Gamma \otimes \Gamma)\Psi'/\sigma^2 = 2\gamma_1\sigma^2\mathcal{J}_{\sigma^2}(\Omega \otimes \Omega)\Psi' \,. \tag{7.2}$$

The expression for the variances and covariances of the residual elements of ξ , i.e. $\mathcal{V}(\Omega_n)$ follows by a straightforward application of the arguments given above. ■

The next proposition ensures that the preceding theorem is applicable to any case where Γ_n represents an affine equivariant covariance matrix estimator and the data stem from an elliptically symmetric distribution.

Proposition 7.3 *Let σ^2 be a scale function and $\Omega \equiv \Omega(\Gamma) = \Gamma/\sigma^2(\Gamma)$ the shape matrix belonging to Γ. Further, let (a_n) be a sequence of real numbers increasing to infinity such that $a_n(\mathrm{vec}\,\Gamma_n - \mathrm{vec}\,\Gamma) \to_{\mathrm{d}} \mathrm{vec}\,\mathcal{N}$ as $n \to \infty$. Here $\mathrm{E}(\mathrm{vec}\,\mathcal{N}) = 0$ and $\mathcal{N}$ is of the radial type with respect to the matrix Γ. Then the conditions of Theorem 7.2 are satisfied.*

Proof. It is only necessary to show that the second-moment condition (7.1) is satisfied. Since $\mathcal{N}$ is of the radial type, this follows immediately from Corollary 1 of Tyler (1982). ■

In the following Γ_n can be interpreted as a covariance matrix estimator. Due to the central limit theorem, in most practical situations it can be found that $a_n = \sqrt{n}$ and the random vector $\mathrm{vec}\,\mathcal{N}$ is multivariate normally distributed. A well-known exception is the *minimum volume ellipsoid* (MVE-)estimator (Rousseeuw, 1985). This is only $\sqrt[3]{n}$-consistent and its asymptotic distribution is non-normal (Davies, 1992). Nonetheless, whenever Γ_n is affine equivariant and the data stem from an elliptically symmetric distribution, the limiting random matrix $\mathcal{N}$ is of the radial type (Tyler, 1982). Hence, Proposition 7.3 is applicable to a wide range of covariance matrix estimators.

An important consequence of Theorem 7.2 is that the asymptotic distribution of Ω_n is only driven by the number γ_1 . This means that γ_2 has no impact on the asymptotic distribution of Ω_n . Hence, the asymptotic relative efficiency of some shape matrix estimator Ω_{1n} compared to another shape matrix estimator Ω_{2n} (i.e. the two estimators are based on the *same* scale function σ^2 but different covariance matrix estimators) can be simply calculated as the ratio γ_{12}/γ_{11}, where γ_{11} is the γ_1 of Ω_{1n} and γ_{12} is the γ_1 of Ω_{2n} (Tyler, 1983).

Corollary 7.4 *Suppose that the conditions of Theorem 7.2 are satisfied and σ^2 corresponds to the scale function given by **S3**. Then it holds that*

$$\mathcal{V}(\sigma_n^2) = \sigma^4\left(\frac{2\gamma_1}{d} + \gamma_2\right), \qquad \mathcal{V}(\sigma_n^2, \Omega_n)' = 0\,,$$

and

$$\mathcal{V}(\Omega_n) = \gamma_1\,(I_{d^2} + K_{d^2})(\Omega \otimes \Omega) - \frac{2\gamma_1}{d} \cdot (\operatorname{vec}\Omega)(\operatorname{vec}\Omega)'\,. \tag{7.3}$$

In particular, if $\operatorname{vec}\mathcal{Q}$ is multivariate normally distributed, the quantities σ_n^2 and Ω_n are asymptotically independent.

Proof. Note that

$$\mathcal{J}_{\sigma^2} = \frac{\sigma^2}{d\det\Gamma} \cdot \frac{\partial \det\Gamma}{\partial(\operatorname{vec}\Gamma)'} = \frac{\sigma^2}{d} \cdot (\operatorname{vec}\Gamma^{-1})' = (\operatorname{vec}\Omega^{-1})'/d\,.$$

Due to Theorem 7.2 the asymptotic variance $\mathcal{V}(\sigma_n^2)$ is given by

$$\mathcal{V}(\sigma_n^2) = \sigma^4\left\{2\gamma_1\mathcal{J}_{\sigma^2}(\Omega\otimes\Omega)\mathcal{J}_{\sigma^2}' + \gamma_2\right\}$$

and note that $(\Omega\otimes\Omega)\mathcal{J}_{\sigma^2}' = \operatorname{vec}\Omega/d$. Moreover, $\mathcal{J}_{\sigma^2}\operatorname{vec}\Omega = 1$, which means that $\mathcal{V}(\sigma_n^2) = \sigma^4(2\gamma_1/d + \gamma_2)$. Further,

$$\mathcal{V}(\sigma_n^2, \Omega_n)' = 2\gamma_1\sigma^2\Psi(\Omega\otimes\Omega)\mathcal{J}_{\sigma^2}' = 2\gamma_1\sigma^2\,\Psi\operatorname{vec}\Omega/d\,.$$

Due to Euler's relation it holds that $\Psi\operatorname{vec}\Omega = 0$ and thus $\mathcal{V}(\sigma_n^2, \Omega_n)' = 0$. This means that σ_n^2 and Ω_n are asymptotically uncorrelated or even independent if $\operatorname{vec}\mathcal{Q}$ is multivariate normally distributed. Finally, the expression for $\mathcal{V}(\Omega_n)$ follows by a straightforward calculation after noting that $\mathcal{J}_{\sigma^2}(\Omega\otimes\Omega)\mathcal{J}_{\sigma^2}' = 1/d$. ■

Theorem 7.5 *Suppose that the conditions of Theorem 7.2 are satisfied with $\gamma_1 > 0$. Then the scale function given by **S3** is the only one where σ_n^2 and Ω_n are asymptotically uncorrelated.*

Proof. Paindaveine (2008) shows that the determinant-based scale function given by **S3** is the only one where the Fisher information is a block diagonal matrix if the family of elliptically symmetric distributions considered is LAN. Suppose that the data are multivariate normally distributed. Then Theorem 7.2 applies to the sample covariance matrix and it is clear that the family of multivariate normal distributions is LAN. The Fisher information is the inverse of the ACM of σ_n^2 and Ω_n (which can be obtained after re-shaping

Ω_n to avoid singularity (Hallin and Paindaveine, 2006a,b)). Hence, there is no other scale function such that (7.2) vanishes. Since the latter is only an algebraic statement, the same must hold for any other distribution under the conditions of Theorem 7.5. ■

Theorem 7.5 extends the main result of Paindaveine (2008) which has been obtained in the context of LAN theory. Similarly, it can be shown that the canonical scale function is the only one which admits the simple representation of the ACM of a shape matrix estimator given by Eq. 7.3. In fact, this ACM exhibits the same desirable form as the ACM of any affine equivariant covariance matrix estimator according to Theorem 7.5 and Eq. 7.1. The operators Ψ and $\mathcal{J}_{\sigma^2}$ corresponding to the remaining scale functions defined by **S1**, **S2**, and **S4** are now given for convenience without an explicit derivation.

ad S1. $\mathcal{J}_{\sigma^2} = \mathbf{e}_1'$, where $\mathbf{e}_1$ is the $d^2 \times 1$ vector with 1 in the first position and zeros elsewhere, so that $\Psi = I_{d^2} - \operatorname{vec}\Omega\,\mathbf{e}_1'$.

ad S2. $\mathcal{J}_{\sigma^2} = (\operatorname{vec} I_d)'/d$ and thus $\Psi = I_{d^2} - (\operatorname{vec}\Omega)(\operatorname{vec} I_d)'/d$ (see also Theorem 5 in Sirkiä et al., 2007).

ad S4. It can be shown that $\mathcal{J}_{\sigma^2} = d/(\operatorname{tr}\Gamma^{-1})^2 \cdot (\operatorname{vec}\Gamma^{-2})' = (\operatorname{vec}\Omega^{-2})'/d$, where $\Gamma^{-2} := \Gamma^{-1}\Gamma^{-1}$ and $\Omega^{-2} := \Omega^{-1}\Omega^{-1}$, i.e. $\Psi = I_{d^2} - (\operatorname{vec}\Omega)(\operatorname{vec}\Omega^{-2})'/d$.

If a shape matrix estimator Ω_{1n} defined via a scale function σ_1^2 is *renormalized* by applying some other scale function σ_2^2 to Ω_{1n}, its ACM simply corresponds to

$$\mathcal{V}(\Omega_{2n}) = \gamma_1 \Psi_2 (I_{d^2} + K_{d^2})(\Omega_2 \otimes \Omega_2)\Psi_2', \tag{7.4}$$

where $\Psi_2 = I_{d^2} - \operatorname{vec}\Omega_2\,\mathcal{J}_{\sigma_2^2}$ and Ω_2 is the shape matrix belonging to Γ with respect to the scale function σ_2^2. This means that the first normalization has no impact on the asymptotic distribution of Ω_{2n}.

7.4. Robust Covariance Matrix Estimation

In the following I will present some well-known robust covariance matrix estimators (i.e. M-, R-, and S-estimators) which satisfy the aforementioned conditions and calculate the joint asymptotic distributions of the corresponding estimators for the shape matrix and scale. It is neither possible nor reasonable to study here all existing robust covariance

matrix estimators (for some contemporary overviews see, e.g., Zuo, 2006, Maronna et al., 2006, Ch. 6), but the essential concept might become clear from the subsequent discussion. Let Γ_n be an affine equivariant estimator which is consistent for Γ. Due to the general result of Tyler (1982), in most practical situations Γ_n is asymptotically normally distributed with ACM $\mathcal{V}(\Gamma_n) = \gamma_1(I_{d^2}+K_{d^2})(\Gamma\otimes\Gamma)+\gamma_2(\operatorname{vec}\Gamma)(\operatorname{vec}\Gamma)'$, where $\gamma_1 \geq 0$ and $\gamma_2 \geq -2\gamma_1/d$ usually depend on the generating variate $\mathcal{R}$. In the following I will only present the numbers γ_1 and γ_2. The $\sqrt{n}$-convergence to the normal law is implicitly assumed. Hence, Theorem 7.5 implies that the canonical scale function is the only one where the estimators for the shape matrix and scale are asymptotically independent. As a counterexample consider the MVE-estimator. This is not $\sqrt{n}$-consistent and asymptotically normally distributed (Davies, 1992). However, since the MVE-estimator is affine equivariant and the rate of convergence does not matter, the corresponding MVE-estimators for the shape matrix and scale remain asymptotically uncorrelated (under the elliptical distribution assumption).

Throughout this section it is supposed that the unknown location vector $\mu \in \mathbb{R}^d$ can be substituted by some $\sqrt{n}$-consistent estimate (here, too, it has already been demonstrated by Rousseeuw (1985) that the MVE-estimator for the location is only $\sqrt[3]{n}$-consistent and its asymptotic distribution is non-normal). In most cases – under mild regularity conditions concerning the distribution of X (see, e.g., Hallin and Paindaveine, 2006b, Tyler, 1987a, Bilodeau and Brenner, 1999, Ch. 13) – it can be shown that the resulting covariance matrix estimator is asymptotically normally distributed, possessing an ACM of the form which is required in Theorem 7.2. Hence, in the following $X_1, \ldots, X_n$ will represent *centered* i.i.d. random vectors for simplicity and without loss of generality.

7.4.1. M-Estimation

An *M-estimator* for Γ (Maronna, 1976) is defined as a solution of

$$\Gamma_n = \frac{1}{n}\sum_{t=1}^{n} w\big(X_t'\Gamma_n^{-1}X_t\big)X_tX_t',$$

where $w : \mathbb{R}^+ \to \mathbb{R}_0^+$ satisfies a set of general conditions (Maronna, 1976, Bilodeau and Brenner, 1999, Section 13.4.1). The estimator Γ_n is strongly consistent for the matrix $\Gamma = \mathrm{E}\{w(X'\Gamma^{-1}X)XX'\}$ which is related to the dispersion matrix of X by $\Gamma = \tau^2\Sigma$, where $\tau > 0$ is such that $\mathrm{E}\{\psi(\mathcal{R}^2/\tau^2)\} = d$ with $\psi(t) := tw(t)$. The numbers γ_1 and γ_2 can be calculated using $\gamma_1 = (d+2)^2\psi_1/(d+2\psi_2)^2$ and

$$\gamma_2 = \frac{(\psi_1-1) - 2\,(\psi_2-1)\,\psi_1\{d+(d+4)\,\psi_2\}/(d+2\psi_2)^2}{\psi_2^2},$$

where $\psi_1 := \mathrm{E}\{\psi^2(\mathcal{R}^2/\tau^2)\}/\{d\,(d+2)\}$ and $\psi_2 := \mathrm{E}\{\psi'(\mathcal{R}^2/\tau^2)\,\mathcal{R}^2\}/(d\tau^2)$ (Tyler, 1982, Bilodeau and Brenner, 1999, p. 223).

If X possesses a continuous elliptical distribution and Σ_n is the corresponding *ML-estimator* for the dispersion matrix Σ, it holds that $\gamma_1 = \{d\,(d+2)/4\}/\mathrm{E}\{h^2(\mathcal{R}^2)\}$ and $\gamma_2 = -2\gamma_1\,(1-\gamma_1)/\{2+d\,(1-\gamma_1)\}$, where $h(t) := t\,\partial \log g(t)/\partial t$. If $X \sim \mathcal{N}_d(0,\Sigma)$ and Σ_n represents the sample covariance matrix, it holds that $\gamma_1 = 1$ and $\gamma_2 = 0$. Otherwise the sample covariance matrix is an M-estimator where $\psi(t) = t$. This means that $\mathrm{E}(\mathcal{R}^2/\tau^2) = \mathrm{E}\{\psi(\mathcal{R}^2/\tau^2)\} = d$, $\psi_1 = d/(d+2)\cdot \mathrm{E}(\mathcal{R}^4)/\mathrm{E}^2(\mathcal{R}^2)$, and $\psi_2 = 1$, so $\gamma_1 = \psi_1$ and $\gamma_2 = \gamma_1 - 1$ if $\mathcal{R}$ has a finite fourth moment.

Now special attention is devoted to Tyler's M-estimator (Tyler, 1983, 1987a)

$$T_n = \frac{d}{n}\sum_{t=1}^{n}\frac{X_t X_t'}{X_t' T_n^{-1} X_t} = \frac{d}{n}\sum_{t=1}^{n}\frac{S_t S_t'}{S_t' T_n^{-1} S_t}\,, \tag{7.5}$$

where $S_t := X_t/\|X_t\|$, $\|\cdot\|$ denotes the Euclidean norm, and it is only supposed that $\mathbb{P}(\mathcal{R} > 0) = 1$. Note that T_n is not affected by the realizations of the generating variate $\mathcal{R}$, since $S = X/\|X\| = \mathcal{R}\Lambda U/\|\mathcal{R}\Lambda U\| = \Lambda U/\|\Lambda U\|$ (a.s.).

This means that Tyler's M-estimator is *distribution-free* in the context of elliptically symmetric distributions. This has been already observed by Tyler (1987b). Frahm and Jaekel (2007a,b) pointed out that the distribution-free property even holds within the class of generalized elliptical distributions. A random vector is said to have a generalized elliptical distribution if its generating variate $\mathcal{R}$ can be negative and might depend on U (Frahm, 2004, p. 46). This feature allows for the modeling of various kinds of asymmetries (**?**Frahm, 2004, Section 3.4). For instance it can be shown that any *skew-elliptical distribution* (Liu and Dey, 2004) belongs to the class of generalized elliptical distributions (Frahm, 2004, p. 47).

Tyler's M-estimator (7.5) is unique up to a scaling constant. Hence, in fact T_n is a genuine *shape matrix* estimator since it can only be calculated with some suitable scale function σ^2 such that $\sigma^2(T_n) = 1$. Originally, Tyler (1987a,b) applied the trace-based scale function given by **S2**, whereas in Tatsuoka and Tyler (2000) the authors prefer to use the canonical normalization **S3**. For the purpose of calculating the asymptotic distribution, Tyler (1987a,b) focuses on $\overline{T}_n := d/(\mathrm{tr}\,\Sigma^{-1}T_n)\cdot T_n$, which means that he defines the scale of T_n via Σ by $\sigma^2(T_n) = \mathrm{tr}\,\Sigma^{-1}T_n/d$. This leads to $\sigma^2(\overline{T}_n) = \sigma^2(\Sigma) = 1$ for any positive-definite $d \times d$ matrix Σ.

Note that in contrast to some normalization according to **S1**–**S4**, the shape matrix estimator $\overline{T}_n$ is indeed affine equivariant and consequently its ACM (Tyler, 1987b) exhibits the simple structure suggested by Eq. 7.1, namely

$$\mathcal{V}(\overline{T}_n) = \frac{d+2}{d} \cdot (I_{d^2} + K_{d^2})(\Sigma \otimes \Sigma) - \frac{2\,(d+2)}{d^2} \cdot (\operatorname{vec}\Sigma)(\operatorname{vec}\Sigma)'\,. \tag{7.6}$$

Since Σ represents a shape matrix with respect to Tyler's scale function, this ACM in fact corresponds to the ACM given by Eq. 7.3 with $\gamma_1 = (d+2)/d$. Furthermore, the Jacobian of Tyler's scale function is given by $\mathcal{J}_{\sigma^2} = (\operatorname{vec}\Sigma^{-1})'/d$ and this actually corresponds to the Jacobian of the *canonical* scale function (see the proof of Corollary 7.4). This means that by using Tyler's scale function in association with some other affine equivariant covariance matrix estimator, the corresponding estimators for the shape matrix and scale become asymptotically uncorrelated. This seems to contradict Theorem 7.5. However, note that Tyler's σ^2 in general does not meet the natural requirement $\sigma^2(I_d) = 1$ and unfortunately $\overline{T}_n$ cannot be applied in practical situations, since σ^2 is determined by the unknown parameter Σ.

An alternative way of obtaining the desired ACM of Tyler's M-estimator is as follows. Note that T_n is simply an M-estimator with $\psi(t) = d$. This means that $\psi_1 = d/(d+2)$ and $\psi_2 = 0$, so $\gamma_1 = (d+2)/d$ and γ_2 is not defined (since σ^2 cannot be estimated by T_n). Hence, due to Theorem 7.2, the ACM of T_n generally corresponds to $\mathcal{V}(T_n) = (d+2)/d \cdot \Psi(I_{d^2} + K_{d^2})(\Omega \otimes \Omega)\Psi'$. Moreover, due to Corollary 7.4 the ACM of Tyler's M-estimator, based on the *canonical* scale function, corresponds to (7.6) where Σ has to be substituted by Ω.

7.4.2. R-Estimation

The R-estimator for the shape matrix has been introduced by Hallin et al. (2006). Consider Tyler's M-estimator T_n which is normalized according to **S1**, i.e. the upper left element corresponds to 1. The R-estimator is based on a *discretized version* of T_n. Suppose that x is a component of T_n. Then it can be discretized by $x^{\#} := \operatorname{sgn} x/\sqrt{n}\,\lceil\sqrt{n}\,|x|\rceil$ (Hallin et al., 2006), where $\lceil y \rceil$ denotes the smallest integer not smaller than $y \in \mathbb{R}$. The corresponding discretized version of Tyler's M-estimator is denoted by $T_n^{\#}$. Hallin and Paindaveine (2006b) also define $U_t := (T_n^{\#})^{-1/2}X_t/\|(T_n^{\#})^{-1/2}X_t\|$. Here $A^{-1/2}$ denotes a positive-definite $d \times d$ matrix such that $A^{-1/2}A^{-1/2\prime} = A^{-1}$, where A^{-1} is the

inverse of a symmetric positive-definite $d \times d$ matrix A. Further, R_t represents the rank of $\|(T_n^{\#})^{-1/2} X_t\|$ with respect to the sample $X_1, \ldots, X_n$.

Let $f_{\mathcal{S}} \colon \mathbb{R}^+ \to \mathbb{R}_0^+$ be the density function of some imaginary generating variate $\mathcal{S}$, whereas $f_{\mathcal{R}}$ refers to the true generating variate $\mathcal{R}$. Consider the cumulative distribution functions $F_{\mathcal{S}}(x) = \int_0^x f_{\mathcal{S}}(r)\, dr$ and $F_{\mathcal{R}}$ respectively. Here both $\mathcal{R}$ and $\mathcal{S}$ are absolutely continuous and satisfy some weak regularity conditions which guarantee local asymptotic normality (Hallin and Paindaveine, 2006b). As already mentioned before, the density function of $\mathcal{S}$ is given by $f_{\mathcal{S}}(r) \propto r^{d-1} g_{\mathcal{S}}(r^2)$, where $g_{\mathcal{S}}$ is the density generator of $\mathcal{S}$. However, in the following consider the function $f_{\mathcal{S}}^*(r) := r^{-(d-1)} f_{\mathcal{S}}(r) = g_{\mathcal{S}}(r^2)$ and for $0 < p < 1$ define $K_{\mathcal{S}}(p) := \psi_{\mathcal{S}}\{F_{\mathcal{S}}^{-1}(p)\}\, F_{\mathcal{S}}^{-1}(p)$, where $F_{\mathcal{S}}^{-1}$ is the quantile function of $\mathcal{S}$ and $\psi_{\mathcal{S}}(x) := -f_{\mathcal{S}}^{*\prime}(x)/f_{\mathcal{S}}^*(x)$. Now, the so-called *cross-information coefficient* (Hallin et al., 2006) is given by

$$\mathcal{I}_{\mathcal{R},\mathcal{S}} := \int_0^1 K_{\mathcal{R}}(p)\, K_{\mathcal{S}}(p)\, dp\,. \tag{7.7}$$

Also define

$$\Delta_n := M_d \left(T_n^{\#} \otimes T_n^{\#}\right)^{-1/2} \sum_{t=1}^{n} \left\{ K_{\mathcal{S}}\Big(\frac{R_t}{n+1}\Big) \operatorname{vec}\big(U_t U_t'\big) - \frac{\overline{K}_{\mathcal{S}}}{d} \cdot \operatorname{vec} I_d \right\}$$

with $\overline{K}_{\mathcal{S}} := 1/n \sum_{t=1}^{n} K_{\mathcal{S}}(t/(n+1))$. The $\{d\,(d+1)/2 - 1\} \times d^2$ matrix M_d symbolizes the Moore-Penrose inverse of N_d' (where N_d is such that $N_d \operatorname{vec} A = \operatorname{vech}_0 A$). Further, let $\Psi_n := I_{d^2} - \operatorname{vec} T_n^{\#}\, \mathbf{e}_1'$ and $Q_n := N_d \Psi_n (I_{d^2} + K_{d^2})(T_n^{\#} \otimes T_n^{\#}) \Psi_n' N_d'$. Now the R-estimator Ω_n is defined in terms of the vech_0 operator, namely

$$\operatorname{vech}_0 \Omega_n = \operatorname{vech}_0 T_n^{\#} + \frac{d\,(d+2)}{2n} \cdot \widehat{\mathcal{I}}_{\mathcal{R},\mathcal{S},n}^{-1}\, Q_n \Delta_n\,,$$

where $\widehat{\mathcal{I}}_{\mathcal{R},\mathcal{S},n}$ represents some consistent estimator for the cross-information coefficient (7.7) (Hallin et al., 2006). The upper left element of Ω_n is set to 1.

Thereafter, following the arguments of Hallin and Paindaveine (2006a) and Paindaveine (2008), one can apply a renormalization by using the canonical scale function and the ACM of the resulting R-estimator readily follows by applying Eq. 7.4 with $\gamma_1 = d\,(d+2)\,\mathcal{I}_{\mathcal{S},\mathcal{S}}/\mathcal{I}_{\mathcal{S},\mathcal{R}}^2$. In particular, if $\mathcal{S} \sim \mathcal{R}$, it holds that $\gamma_1 = d\,(d+2)/\mathcal{I}_{\mathcal{R},\mathcal{R}}$ with $\mathcal{I}_{\mathcal{R},\mathcal{R}} = \int_0^1 K_{\mathcal{R}}^2(p)\, dp = \mathrm{E}(\psi_{\mathcal{R}}^2(\mathcal{R})\, \mathcal{R}^2)$. From $\psi_{\mathcal{R}}(r)\, r = -2r^2 g'(r^2)/g(r^2)$ it follows that $\psi_{\mathcal{R}}^2(r)\, r^2 = 4h^2(r^2)$, where h has already been defined in Section 7.4.1. Recall that the function h is used for calculating the ACM of an ML-estimator. This means that if $\mathcal{S} \sim \mathcal{R}$, the R-estimator has the same limiting distribution as the corresponding ML-estimator and thus it becomes asymptotically efficient.

7.4.3. S-Estimation

The S-estimator for the dispersion matrix (Davies, 1987) can be defined as

$$\Gamma_n = \arg\min_{\Upsilon \in \mathcal{P}^d} \det \Upsilon$$

subject to

$$\frac{1}{n}\sum_{t=1}^{n} \rho\Big(\sqrt{X_t' \Upsilon^{-1} X_t}\Big) = \alpha\rho(\infty)\,,$$

where $0 < \alpha < 1$ and $\rho\colon \mathbb{R}^+ \to \mathbb{R}_0^+$ has to be bounded, increasing, and sufficiently smooth (Croux and Haesbroeck, 1999, Tyler, 2002, Bilodeau and Brenner, 1999, Section 13.4.2). The chosen constraint guarantees that Γ_n is consistent for $\Gamma = \tau^2\Sigma$, where $\tau > 0$ is such that $\mathrm{E}\{\rho(\mathcal{R}/\tau)\} = \alpha\rho(\infty)$.

Let ψ be the first and ψ' the second derivative of ρ. It is assumed that

$$\mathrm{E}\{\psi'(\mathcal{R}/\tau)\} > 0 \qquad \text{and} \qquad \mathrm{E}\{\psi'(\mathcal{R}/\tau)\,\mathcal{R}^2/\tau + (d+1)\,\psi(\mathcal{R}/\tau)\,\mathcal{R}\} > 0\,.$$

Then the numbers γ_1 and γ_2 are given by

$$\gamma_1 = \frac{d\,(d+2)\,\mathrm{E}\{\psi^2(\mathcal{R}/\tau)\,\mathcal{R}^2\}}{\mathrm{E}^2\{\psi'(\mathcal{R}/\tau)\,\mathcal{R}^2/\tau + (d+1)\,\psi(\mathcal{R}/\tau)\,\mathcal{R}\}}$$

and

$$\gamma_2 = \frac{4\tau^2 \mathrm{Var}\{\rho(\mathcal{R}/\tau)\}}{\mathrm{E}^2\{\psi(\mathcal{R}/\tau)\,\mathcal{R}\}} - \frac{2\gamma_1}{d}$$

(Davies, 1987, Lopuhaä, 1989, Bilodeau and Brenner, 1999, p. 225).

7.5. Conclusion

In most practical situations the matter of interest is only the covariance matrix up to some unknown scaling constant. In that case covariance matrix estimation can be reduced to shape matrix estimation and so it is adequate to focus on the asymptotic distribution of a given shape matrix estimator. In the present work robust estimators for the shape matrix and its associated scale have been investigated. I derived explicit expressions for their joint asymptotic distributions and generalized a result which has been recently obtained in the context of local asymptotic normality theory. The given instruments are applicable to a wide range of problems in multivariate analysis such as principal components analysis, canonical correlation analysis, linear discriminant analysis, and multivariate regression.

Chapter 8.

Distribution-Free Shape Matrix Estimation for Incomplete Data

8.1. Introduction

During the last decades robust covariance matrix estimation has become a popular branch of robust statistics. Many different estimation approaches have been established until today. For a broad overview on robust statistics see Hampel et al. (1986), Huber (2003), and Maronna et al. (2006). In the literature there exist many robust techniques to insulate from the 'bad influence' of outliers. For the subsequent discussion it is important to distinguish between the robust, the nonparametric, and the distribution-free estimation approach which are often mixed-up.

1. The robust approach produces an estimator which is less sensitive to *contaminated data* such as outliers or clusters.

2. The nonparametric approach is based on the *empirical distribution* of a random quantity without specifying some parametrical model.

3. The distribution-free approach leads to an estimator whose *finite-sample distribution* is invariant against some part of the data-generating process.

Robust approaches (a) can be inherently parametric. For example, M-estimators are always constructed on the basis of some parametric (pseudo-)model (Hampel et al., 1986, p. 230). However, M-theory specifically allows for discrepancies between the pseudo-model and the true model. If the pseudo-model appropriately accounts for outliers or clusters

then typically the resulting estimator will become robust. Further, the nonparametric approach (b) can be justified by the Glivenko-Cantelli theorem or, more generally, by the fundamental properties of empirical processes (van der Vaart, 1998, Ch. 19). However, if the data are not normally distributed this approach can lead to highly non-robust estimators. For instance, the sample covariance matrix can be derived as a nonparametric estimator and it is well-known that this is not robust against outliers.

For understanding the distribution-free approach (c) consider a linear regression model

$$Y_t = \beta_1 + \beta_2 X_{t2} + \ldots + \beta_m X_{tm} + u_t\,, \qquad t = 1, \ldots, n\,,$$

where

$$X = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1m} \\ 1 & X_{22} & \cdots & \cdots & X_{2m} \\ \vdots & \vdots & & & \vdots \\ 1 & X_{n2} & X_{n3} & \cdots & X_{nm} \end{bmatrix}$$

represents an $n \times m$ matrix ($n > m$) of stochastic regressors and $u = (u_1, \ldots, u_n)$ is a spherically distributed random vector being stochastically independent of X. The OLS-estimator

$$\hat{\sigma}^2 = \frac{\hat{u}^\mathsf{T}\hat{u}}{n-m}\,,$$

for the residual variance σ^2 is independent of X since $\hat{u}$ is the complement of the orthogonal projection $X(X^\mathsf{T}X)^{-1}X^\mathsf{T}u$, i.e. $\hat{u} = (I_n - X(X^\mathsf{T}X)^{-1}X^\mathsf{T})u$. That means $\hat{u}^\mathsf{T}\hat{u}$ and so $\hat{\sigma}^2$ is only determined by u and so it represents a distribution-free estimator with respect to X. Thus distribution freeness primarily has nothing to do with robustness and especially such estimators might result from parametric models.

Of course, nonparametric methods can be called 'distribution-free' since they do not require any or only some weak assumptions about the sampling distribution. Here the notion 'distribution-free' shall indicate that the requirements concerning the underlying distribution are minimal and therefore Rao (1965, p. 422) suggested to use the terms 'nonparametric' and 'distribution-free' synonymously. However, following the arguments of Kendall and Sundrum (1953), we think that it is more convenient if the attribute 'distribution-free' refers to the finite-sample distribution of the resulting estimator which of course has direct consequences for the associated confidence intervals and hypothesis tests. If a distribution-free estimator is invariant against the 'contaminating part' of the data-generating process, i.e. the underlying mechanism which is responsible for outliers or clusters, it is a *completely*

robust estimator. The main purpose of the present work is to derive a completely robust estimator for the shape matrix of a generalized elliptically distributed random vector if the data are incomplete.

The notion of robust 'covariance matrix' estimation is somewhat ambiguous since many robust statistical procedures actually do not estimate the covariance matrix but some other matrix being proportional to the covariance matrix. Indeed, a large number of multivariate statistical methods like principal component analysis, canonical correlation analysis, linear discriminant analysis, and multivariate regression require the covariance matrix only up to a scaling constant (Oja, 2003, Paindaveine, 2008, **?**). Let $\mathcal{P}^d$ be the set of all symmetric positive-definite $d \times d$ matrices and $\varphi\colon \mathcal{P}^d \to \mathbb{R}^k$ a scale-invariant function, i.e. $\varphi(\alpha\Sigma) = \varphi(\Sigma)$ for all $\alpha > 0$ and $\Sigma \in \mathcal{P}^d$. Now consider a scale-invariant function $\Omega(\Sigma) = \Sigma/\sigma^2(\Sigma)$ where σ^2 is a positive homogeneous function, i.e. $\sigma^2(\alpha\Sigma) = \alpha\sigma^2(\Sigma) > 0$. The matrix $\Omega(\Sigma)$ will be called a *shape matrix* associated with Σ. Note $\varphi \circ \Omega = \varphi$ since

$$\varphi\big\{\Omega(\Sigma)\big\} = \varphi\left\{\frac{\Sigma}{\sigma^2(\Sigma)}\right\} = \varphi(\Sigma)\,.$$

For instance the correlation matrix associated with some positive-definite matrix Σ is scale-invariant and thus it can be derived from any shape matrix Ω. Since the most problems of multivariate statistics are scale-invariant, in the following discussion we will concentrate on shape matrix rather than covariance matrix estimation. The reader should keep in mind that whenever $\widehat{\Omega}$ is a distribution-free estimator for the shape matrix Ω, a distribution-free estimator for $\varphi(\Sigma)$ (e.g. the parameters of a linear regression) is given by $\varphi(\widehat{\Omega})$.

We will present a shape matrix estimator which is distribution-free within the class of *generalized elliptical distributions* (Frahm, 2004). In the complete-data case the presented estimator corresponds to Tyler's M-estimator (Tyler, 1983, 1987a). However, in the context of missing data there is still not much work on robust covariance matrix estimation. One of the few available approaches is to maximize the observed-data likelihood function of heavy-tailed or contaminated data (Little, 1988). This is a parametrical approach and the aim of the present work is to formulate an alternative distribution-free estimation approach for missing data. It turns out that the resulting estimator is a completely robust ML-estimator which can be viewed as a non-trivial generalization of Tyler's M-estimator. Thus it is possible to obtain its asymptotic properties by standard results of likelihood theory. We present a fast algorithm for calculating the estimate which works well even for high-dimensional data. Further, we provide a simulation study covering the complete-data

as well as the incomplete-data case using clean and contaminated data under the different missingness mechanisms MCAR, MAR, and NMAR.

8.2. Elliptical Distributions

8.2.1. Elliptically Symmetric Distributions

Consider a d-dimensional elliptically symmetric distributed random vector X. That is X can be represented by

$$X = \mu + \Lambda \mathcal{R} U \,, \tag{8.1}$$

where U is a k-dimensional random vector, uniformly distributed on the unit hypersphere, $\mathcal{R}$ is a nonnegative random variable being stochastically independent of U, $\mu \in \mathbb{R}^d$, and $\Lambda \in \mathbb{R}^{d \times k}$ (Cambanis et al., 1981, Fang et al., 1990, p. 42). The random variable $\mathcal{R}$ is called the *generating variate* of X and in the special case $\mu = 0$ and $\Lambda = \sigma I_d$ with $\sigma > 0$, the random vector X is spherically distributed on $\mathbb{R}^d$.

Note that the number k principally can be larger than d. For instance, consider some latent factor model for a d-dimensional random vector X, i.e.

$$X = \mu + LF + \varepsilon \,,$$

where $L \in \mathbb{R}^{d \times f}$ with $f < d$ is a matrix of *factor loadings*, F is an f-dimensional random vector of *common factors*, and ε is a d-dimensional random vector of *idiosyncratic risks* such that (F, ε) is spherically distributed on $\mathbb{R}^{f+d}$.

In case $k > d$ there always exists a reduced form representation of X. The distribution of X is determined by Λ only through the *dispersion matrix* $\Sigma := \Lambda\Lambda^\mathsf{T}$ (Cambanis et al., 1981). Let the $d \times r$ matrix Γ with $r \leq d$ be a root of Σ, i.e. $\Gamma\Gamma^\mathsf{T} = \Sigma$. Then

$$X = \mu + \Gamma \mathcal{S} U \,,$$

where U is now an r-dimensional random vector on the unit hypersphere. Further, $\mathcal{S} := \mathcal{R}\sqrt{\beta}$ with $\beta \sim \mathrm{Beta}(r/2, (k-r)/2)$ is stochastically independent of $\mathcal{R}$ and U (Cambanis et al., 1981). In contrast, if $k < d$ the dispersion matrix cannot be positive-definite which might also happen in case $k \geq d$. However, in the following we will generally assume that Σ is positive-definite, which means that $\mathrm{r}\,\Lambda = d$ and refer to a *full-rank representation* of

X (Cambanis et al., 1981) whenever this is elliptically symmetric distributed. Hence, it is assumed without loss of generality that Λ is a nonsingular square matrix.

The random vector X possesses a density function if and only if $\mathcal{R}$ is absolutely continuous and then the density of X is given by

$$p(x) = \sqrt{\det \Sigma^{-1}}\, g\big\{(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\big\}\,. \tag{8.2}$$

The so-called *density generator* g is a nonnegative function on the set of all positive real numbers which depends on x only through the quadratic form $(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)$. Conversely, if some random vector X has an elliptically contoured density of the form (8.2) it is elliptically symmetric distributed. Consider for instance the density generator $g(z) = (2\pi)^{-d/2}\exp(-z/2)$ of the multivariate normal distribution. Moreover, the density generator g can be calculated from the density of $\mathcal{R}$ by the relation

$$g(z) = \frac{\Gamma(d/2)}{2\pi^{d/2}} \cdot z^{-\frac{d-1}{2}} f\big(\sqrt{z}\,\big)\,, \qquad z > 0\,,$$

where f represents the density function of $\mathcal{R}$ (Frahm, 2004, p. 9).

Depending on the chosen density function of $\mathcal{R}$, elliptically symmetric distributions allow for the modeling of exponentially decaying density functions (e.g. the multivariate normal distribution), *heavy tails* either with finite variance (e.g. the multivariate t-distribution with $\nu > 2$ degrees of freedom) or infinite variance (e.g. multivariate symmetric α-stable or, say, sub-Gaussian distributions), and *semi-heavy tails* (e.g. multivariate symmetric generalized hyperbolic distributions). Note that if the generating variate $\mathcal{R}$ is heavy-tailed possessing some *tail index* α (Mikosch, 2003), the corresponding random vector is also regularly varying with the same tail index (Hult and Lindskog, 2002, Frahm, 2006).

8.2.2. Skew-Elliptical Distributions

Elliptically symmetric distributions suffer from the lack of skewness or, more generally, different kinds of asymmetries which can be often observed from empirical data. Thus a popular generalization of elliptically symmetric distributions is provided by the class of *skew-elliptical distributions* (Liu and Dey, 2004). Consider a $(d+1)$-dimensional elliptically distributed random vector $X^* = (Z, Y)$ with location vector $\mu^* = (0, \mu)$ and dispersion matrix

$$\Sigma^* := \begin{bmatrix} 1 & \beta^{\mathsf{T}} \\ \beta & \Sigma \end{bmatrix},$$

where Y is a d-dimensional random vector associated with the location vector μ and dispersion matrix Σ. Now the random vector $X := (Y \mid Z > 0)$ is said to be *skew-elliptically distributed* (Branco and Dey, 2001). Here β is a skewness parameter whereas anymore μ and Σ are the location vector and the dispersion matrix of X. Thus any skew-elliptical distribution can be obtained by *hidden truncation* (Arnold and Beaver, 2004), since only that part of the distribution of X^* is recognized where the hidden variable Z is positive. If X^* is multivariate normally distributed, X is said to be *skew-normally distributed* (Azzalini and Dalla Valle, 1996). For a broad exposition on skew-elliptical distributions see Genton (2004).

8.2.3. Generalized Elliptical Distributions

Finally, a d-dimensional random vector X is said to be *generalized elliptically distributed* if it can be represented by Eq. 8.1 where U is a k-dimensional random vector, uniformly distributed on the unit hypersphere, $\mathcal{R}$ is a random variable, $\mu \in \mathbb{R}^d$, and $\Lambda \in \mathbb{R}^{d \times k}$ (Frahm, 2004, p. 46). In contrast to the former representation of elliptically symmetric distributions, now the generating variate $\mathcal{R}$ may depend on the unit random vector U and it can be negative, too. By using that property it is easy to show that the class of skew-elliptical distributions belongs to the class of generalized elliptical distributions. More precisely, any skew-elliptical random vector with parameters μ and Σ is generalized elliptically distributed possessing the same location vector and dispersion matrix (Frahm, 2004, p. 47).

It is worth to point out that for generalized elliptical distributions in general there exists no full-rank representation unless the distribution is elliptically symmetric. This is due to the fact that $\mathcal{R}$ might depend on U. Further results on generalized elliptical distributions can be found in Frahm (2004, Ch. 3) and a discussion in the context of high-dimensional data is given by Frahm and Jaekel (2007b). Figure 8.1 documents that generalized elliptical distributions provide a fairly nice fit to financial data.

Now we will present a generalized elliptical distribution which will play a major role in the following discussion.

Theorem 8.1 *Let Λ be a $d \times k$ matrix with $\mathrm{r}\,\Lambda = d$ and U a k-dimensional random vector, uniformly distributed on the unit hypersphere. The density of the unit random*

vector $V = \Lambda U/\|\Lambda U\|$ *corresponds to*

$$\psi(v) = \frac{\Gamma(d/2)}{2\pi^{d/2}} \cdot \sqrt{\det \Sigma^{-1}}\, \sqrt{v^\mathsf{T}\Sigma^{-1}v}^{\,-d},$$

where $\|\cdot\|$ *denotes the Euclidean norm,* $\Sigma = \Lambda\Lambda^\mathsf{T}$*, and* v *is such that* $\|v\| = 1$*.*

Proof. Frahm (2004, pp. 59–60). ■

Note that the random vector V is generalized elliptically distributed with generating variate $\|\Lambda U\|^{-1}$. Its distribution is sometimes referred to as the *angular central Gaussian distribution on the sphere* (Tyler, 1987b, Kent and Tyler, 1988, Mardia and Jupp, 2000, p. 182) or the *offset normal distribution* (Mardia and Jupp, 2000, p. 178) but we will call it simply *spectral density function*. This is justified by the following corollary.

Corollary 8.2 *The extremal positions of* ψ *in Theorem 8.1 are given by the space of normalized eigenvectors of* Σ*, i.e. for any point* v *on the unit hypersphere such that* $\Sigma v = \lambda v$*, the value* $\psi(v)$ *is a local extremum of* ψ *and vice versa.*

Proof. Frahm and Jaekel (2007b). ■

As a direct consequence of Corollary 8.2, the local extrema of ψ can be calculated by

$$\psi(v) = \frac{\Gamma(d/2)}{2\pi^{d/2}} \cdot \sqrt{\det(\lambda\Sigma^{-1})}\,,$$

where λ is an eigenvalue of Σ and v is the corresponding normalized eigenvector.

8.3. Distribution-Free Shape Matrix Estimation

Consider a d-dimensional random vector X around a center $\mu \in \mathbb{R}^d$. Now define a matrix Ω such that

$$\Omega = d \cdot \mathrm{E}\left\{ \frac{(X-\mu)(X-\mu)^\mathsf{T}}{(X-\mu)^\mathsf{T}\Omega^{-1}(X-\mu)} \right\} \tag{8.3}$$

with $\det \Omega = 1$. If Ω exists and is unique we will call that matrix the shape matrix of X. The following theorem asserts that Ω indeed is well-defined if X is generalized elliptically distributed.

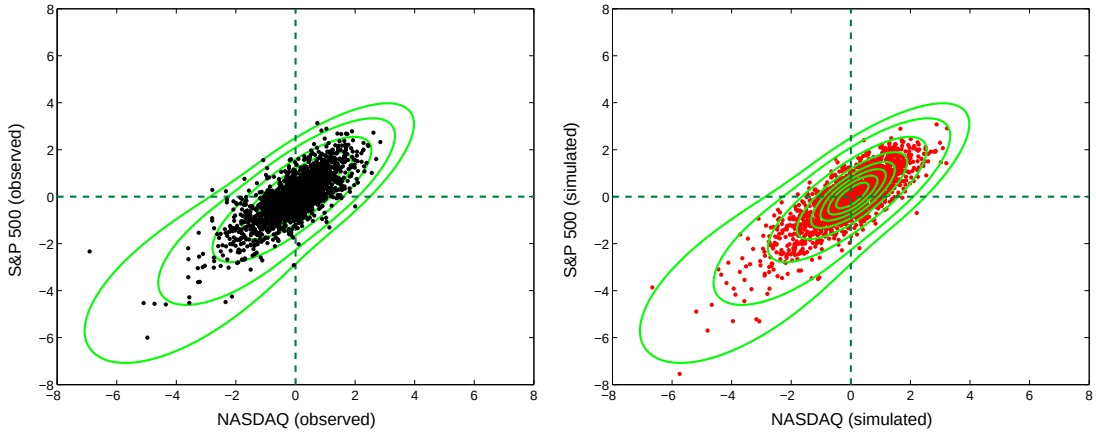

Figure 8.1.: Observed GARCH(1,1)-residuals of NASDAQ and S&P 500 daily log-returns from 1993-01-01 to 2000-06-30 (left hand) and simulated generalized elliptically distributed residuals ($n = 1892$) (right hand). The density contours of the chosen model (Frahm and Jaekel, 2007b) are given by the green curves.

Theorem 8.3 *Let X be a d-dimensional generalized elliptically distributed random vector with location vector $\mu \in \mathbb{R}^d$, positive-definite dispersion matrix $\Sigma \in \mathbb{R}^{d \times d}$, and generating variate $\mathcal{R}$ with $\mathbb{P}(\mathcal{R} = 0) = 0$. Then the shape matrix Ω exists and corresponds to*

$$\Omega = \frac{\Sigma}{(\det \Sigma)^{1/d}} .$$

Proof. By the definition of generalized elliptical distributions it follows that

$$\frac{(X-\mu)(X-\mu)^{\mathsf{T}}}{(X-\mu)^{\mathsf{T}}\Omega^{-1}(X-\mu)} \overset{\text{a.s.}}{=} \frac{\Lambda U U^{\mathsf{T}} \Lambda^{\mathsf{T}}}{U^{\mathsf{T}}\Lambda^{\mathsf{T}}\Omega^{-1}\Lambda U} . \tag{8.4}$$

Since ΛU is elliptically symmetric distributed it can be assumed that $\Lambda \in \mathbb{R}^{d \times d}$ without loss of generality (cf. Section 8.2.1). By setting Ω to $\Lambda\Lambda^{\mathsf{T}} = \Sigma$, the right hand side of Eq. 8.4 becomes $\Lambda U U^{\mathsf{T}} \Lambda^{\mathsf{T}}$. Note that $\mathrm{E}(UU^{\mathsf{T}}) = I_d/d$ (Fang et al., 1990, p. 34), so that

$$d \cdot \mathrm{E}\left\{ \frac{(X-\mu)(X-\mu)^{\mathsf{T}}}{(X-\mu)^{\mathsf{T}}\Omega^{-1}(X-\mu)} \right\} = \Lambda\Lambda^{\mathsf{T}} = \Omega .$$

Re-scaling Ω by the constant $(\det \Sigma)^{-1/d}$ leads to the desired result. ∎

Note that the shape matrix Ω is a scale-invariant function of Σ. Further, the shape matrix exists and is unique for any random vector X which is continuously distributed on $\mathbb{R}^d$ (Tyler, 1987a). Since

$$\frac{(X-\mu)(X-\mu)^{\mathsf{T}}}{(X-\mu)^{\mathsf{T}}\Omega^{-1}(X-\mu)} = \frac{SS^{\mathsf{T}}}{S^{\mathsf{T}}\Omega^{-1}S} , \qquad S := \frac{X-\mu}{\|X-\mu\|} ,$$

it becomes clear that for the existence and uniqueness of Ω, the projection S has only to be continuously distributed on the *unit hypersphere* $\{x \in \mathbb{R}^d : \|x\| = 1\}$. Here it must be presumed that the distribution of X has no atom at $\mu \in \mathbb{R}^d$, i.e. $\mathbb{P}(X = \mu) = 0$. Within the class of generalized elliptical distributions S is continuously distributed by definition.

In the following we will concentrate on shape matrix estimation. It is supposed that the location vector μ is known or that it can be properly estimated. That issue will be discussed later on in more detail. For the time being we will restrict on *centered* random vectors $X_1, \ldots, X_n$ and their corresponding realizations $x_1, \ldots, x_n$ without loss of generality.

8.3.1. The Complete-Data Case

Tyler's M-Estimator

If X is generalized elliptically distributed, outliers are produced by extreme realizations of the generating variate $\mathcal{R}$. Note that by definition such values can be clustered in arbitrary directions in $\mathbb{R}^d$ since $\mathcal{R}$ may depend on U. The discussion in Section 8.2 shows that the class of generalized elliptical distributions is huge. While this gives us the flexibility to adapt to specific characteristics of the data, it can be a problem in many practical situations where neither the distribution family of $\mathcal{R}$ nor the dependence structure between $\mathcal{R}$ and U are known. In the following we will focus on estimating the shape matrix Ω *without* specifying the joint distribution function of $\mathcal{R}$ and U. This allows us to separate the linear dependence structure from non-linear dependencies such as tail-dependence possibly caused by $\mathcal{R}$. As already mentioned the location vector μ is supposed to be known and μ is set to zero without loss of generality. Any other parameters like the linear operator Λ, the parameters concerning the distribution of the generating variate $\mathcal{R}$ and even the functional form of the joint distribution of $\mathcal{R}$ and U are supposed to be unknown.

In the following let Σ be positive-definite and $\mathbb{P}(\mathcal{R} = 0) = 0$. Due to the definition of X it holds that

$$S = \frac{X}{\|X\|} = \frac{\mathcal{R}\Lambda U}{\|\mathcal{R}\Lambda U\|} \overset{\text{a.s.}}{=} \pm\frac{\Lambda U}{\|\Lambda U\|} = \pm V\,, \tag{8.5}$$

where $\pm := \operatorname{sgn}(\mathcal{R})$ and $V = \Lambda U/\|\Lambda U\|$. The key observation is that the random vector V does not depend on the absolute value of $\mathcal{R}$. In particular, it is completely robust against *extreme* outcomes of the generating variate. However, the sign of $\mathcal{R}$ still remains in Eq. 8.5 and indeed this might moreover depend on U. The unit random vector S represents the

direction of X on the unit hypersphere. It contains all necessary information for estimating the shape matrix. Note that in the univariate case $S = \mathrm{sgn}(X)$. However, since the shape matrix quantifies only the linear dependence structure of X but not its scale, multivariate data are necessary for estimating Ω. That means it is always required that $d > 1$ in the following discussion.

Tyler's M-estimator (Tyler, 1983, 1987a) is defined via the fixed point equation

$$T = \frac{d}{n} \sum_{t=1}^{n} \frac{X_t X_t^\mathsf{T}}{X_t^\mathsf{T} T^{-1} X_t} \tag{8.6}$$

and can be regarded as a 'sample version' of the shape matrix defined by (8.3). Since T is defined only up to a scaling constant, this fixed-point equation has to be solved by the additional constraint $\det T = 1$ and in that case T becomes an appropriate estimator for the shape matrix Ω. Of course, any other constraint like, e.g., $\Sigma_{11} = 1$ or $\mathrm{tr}\,\Sigma = d$ would also work but the determinant-based normalization has several statistical advantages which are discussed by **?** and Paindaveine (2008). Note that for estimating Ω it is not necessary to know the sign of $\mathcal{R}$ since $\pm^2 = +$. From Eq. 8.5 it can be seen that Tyler's M-estimator is invariant under any change of the generating variate $\mathcal{R}$, i.e. it is *distribution-free* within the class of generalized elliptically distributed random vectors. Moreover, it is strongly consistent and asymptotically normally distributed provided that X possesses a continuous distribution on $\mathbb{R}^d$ (Tyler, 1987a).

Important results concerning the existence of Tyler's M-estimator for *any* kind of distributions were established by Tyler (1987a) as well as Kent and Tyler (1988, 1991). For instance, if the data are contaminated at some point in $\mathbb{R}^d$, the rate of contamination must not exceed $1/d$ (Kent and Tyler, 1988). Further, Kent and Tyler (1988) proved that for any given sample $x_1, \ldots, x_n \neq 0$ the fixed-point solution T exists and the sequence (T_i) defined by the fixed-point iteration scheme

$$T_{i+1} = \frac{d}{n} \sum_{t=1}^{n} \frac{x_t x_t^\mathsf{T}}{x_t^\mathsf{T} T_i^{-1} x_t}, \qquad i = 0, 1, \ldots, \tag{8.7}$$

converges to $\sigma^2 T$ provided the data stem from a continuous distribution on $\mathbb{R}^d$ and $n > d$. The initial value T_0 can be any positive-definite $d \times d$ matrix and $\sigma^2 > 0$ is a scaling constant depending on the initial value T_0. For estimating the shape matrix Ω the normalization $(\det T_N)^{-1/d} T_N$ can be applied just after performing a sufficiently large number N of iterations (Kent and Tyler, 1991). If the data are contaminated at some point in $\mathbb{R}^d$ the

convergence of this algorithm is guaranteed if the rate of contamination is smaller than $1/d$ (Kent and Tyler, 1988).

Tyler's M-estimator is a robust estimator and its robustness properties (i.e. breakdown point, maximum bias and variance) were already investigated by Adrover (1998), Dümbgen and Tyler (2005), Maronna and Yohai (1990), as well as Tyler (1983, 1987a). In particular, it has been shown that the Dirac contamination breakdown point of T corresponds to $1/d$ (Maronna and Yohai, 1990) whereas for *any* kind of contamination it is between $1/(d+1)$ and $1/d$ (Adrover, 1998) if the data are elliptically symmetric distributed. Due to the arguments given above the same holds for generalized elliptical distributions, too.

The Spectral Density Approach

Tyler originally derived his estimator as an M-estimator by using a Huber-type weight function (Tyler, 1983, 1987a) but T is also an *ML-estimator* if the spectral density given by Eq. 8.1 is taken into consideration. This important fact has been already noticed by Tyler (1987b) and, in a somewhat different context related to generalized elliptical distributions, by Frahm (2004).

Theorem 8.4 *Let $X_1, \ldots, X_n$ be a sample of independent copies of a d-dimensional generalized elliptically distributed and centered random vector X with positive-definite dispersion matrix $\Sigma \in \mathbb{R}^{d\times d}$ and generating variate $\mathcal{R}$ such that $\mathbb{P}(\mathcal{R}=0)=0$. Consider the unit random vector $S = X/\|X\|$ and the corresponding sample $S_1, \ldots, S_n$ with $n > d$. Then Tyler's M-estimator T exists almost surely and it is an ML-estimator with respect to the likelihood function*

$$\mathcal{L}\big(\Sigma\,; V_1, \ldots, V_n\big) = \prod_{t=1}^{n} \psi\big(V_t\,;\Sigma\big)\,,$$

where $V_1, \ldots, V_n$ are defined according to Eq. 8.5 and ψ is the spectral density function given by Theorem 8.1. Furthermore, T satisfies the log-likelihood equation

$$\frac{\partial \log \mathcal{L}(T\,; V_1, \ldots, V_n)}{\partial \Sigma} = \sum_{t=1}^{n} \frac{\partial \log \psi(V_t\,;T)}{\partial \Sigma} = 0\,.$$

Proof. The arguments for the existence and thus positive definiteness of Tyler's M-estimator can be found in Kent and Tyler (1988). Consider the log-likelihood function

$$\log \mathcal{L}(\Sigma\,; V_1, \ldots, V_n) = \sum_{t=1}^{n} \log \psi(V_t\,;\Sigma) = c + \frac{n}{2}\cdot \log\det \Sigma^{-1} - \frac{d}{2}\sum_{t=1}^{n} \log\big(S_t^{\mathsf{T}}\Sigma^{-1}S_t\big),$$

where c is a constant and note that $\psi(V_t) = \psi(S_t)$, since ψ is an even function. The partial derivative of $\log \mathcal{L}(\Sigma\,; V_1, \ldots, V_n)$ with respect to the inverse Σ^{-1} is given by

$$\frac{\partial \log \mathcal{L}}{\partial \Sigma^{-1}} = \frac{n}{2} \cdot \big(2\Sigma - \operatorname{diag} \Sigma\big) - \frac{d}{2} \sum_{t=1}^{n} \left(\frac{2 S_t S_t^{\mathsf{T}} - \operatorname{diag}(S_t S_t^{\mathsf{T}})}{S_t^{\mathsf{T}} \Sigma^{-1} S_t} \right) = M - \operatorname{diag} M/2\,,$$

where

$$M := n\Sigma - d \cdot \sum_{t=1}^{n} \frac{S_t S_t^{\mathsf{T}}}{S_t^{\mathsf{T}} \Sigma^{-1} S_t}\,.$$

Since the set of all positive-definite matrices with unit determinant is open, T is a stationary point of the log-likelihood function so that

$$nT - d \cdot \sum_{t=1}^{n} \frac{S_t S_t^{\mathsf{T}}}{S_t^{\mathsf{T}} T^{-1} S_t} = 0\,,$$

which is equivalent to Eq. 8.6. ∎

Recall that ψ is an even function, i.e. $\psi(-s) = \psi(s)$ for every s with $\|s\| = 1$. That means Tyler's M-estimator indeed maximizes the likelihood function for a sample $V_1, \ldots, V_n$ of independent copies of the unit random vector V given by Eq. 8.5 even though the corresponding realizations of V are given only up to the corresponding signs.

Now, many statistical properties of Tyler's M-estimator can be derived on the basis of likelihood theory. For instance it can be shown that T is asymptotically efficient among all distribution-free estimators for the shape matrix of a generalized elliptical distribution (Frahm, 2004, Ch. 5). Due to Theorem 8.1 and Theorem 8.4, T is said to be a *spectral estimator*. It is asymptotically normally distributed and its asymptotic covariance matrix is given by **?** under the specific constraint $\det T = 1$.

The problem of estimating the location vector μ has been already investigated by Tyler (1987a) under quite general conditions. Suppose that X has a continuous distribution on $\mathbb{R}^d$ and μ is estimated by a consistent estimator $\hat{\mu}_n$ which is used for centralizing the data. If $\hat{\mu}_n$ converges to μ at an appropriate rate and X is not too much concentrated around μ then T is still consistent and asymptotically normally distributed (Tyler, 1987a). Otherwise even small perturbations of $\hat{\mu}$ would lead to wrong projections of X to the unit hypersphere, i.e. $(X - \hat{\mu})/\|X - \hat{\mu}\|$. If the regularity conditions hold, Tyler's M-estimator possesses the *same* asymptotic covariance matrix as if μ was known. However, we admit that it is not easy to find a consistent estimator for μ if the distribution of X is asymmetric. In particular some of Tyler's conditions for the asymptotic normality are violated when X has an asymmetric distribution. Hence, if μ is unknown the spectral density approach can

be clearly defended if the data are elliptically symmetric distributed whereas there is no formal justification in case of generalized elliptical distributions.

8.3.2. The Incomplete-Data Case

Now Tyler's M-estimator will be generalized to the case of incomplete data by using the well-developed likelihood theory for missing data. This is not a trivial generalization since Tyler originally argued on the basis of M-estimation theory (Tyler, 1983). The key observation is that Tyler's M-estimator in fact is an ML-estimator (Frahm, 2004, Tyler, 1987b) and then methods of missing-data analysis have to be applied carefully. The difficult part is to derive the score function under incomplete data and to formulate an appropriate algorithm for finding its root. First of all we will recapitulate the fundamental background of missing-data analysis which is necessary for understanding the subsequent derivations. A comprehensive introduction to that topic is given by Little and Rubin (2002) as well as Schafer (1997).

Ignorable Missingness Patterns

Let x be some realized data and m an ensemble of zeros and ones indicating which part of x is observed and which is missing. According to the missingness pattern m let x_o be the observed and x_m the unobserved data. The observed part O of the complete data X is a *random index* whereas o denotes some realization of O according to the missingness pattern m which is a realization of M. Sometimes it is helpful to interpret m as a function $m\colon x \mapsto x_o$. Further, M and X are random quantities possessing the joint distribution $p(m, x\,;\theta)$. Here $\theta \in \Theta \subset \mathbb{R}^k$ is some unknown parameter. The marginal distribution of m and x_o corresponds to

$$p(m, x_o\,;\theta) = \int p(m, x_o, x_m\,;\theta)\, dx_m\,.$$

Suppose that the parameter θ has to be estimated. All available information are given by m and x_o though $p(m, x\,;\theta)$ is the underlying sampling distribution of the experiment. However, under the standard assumptions of likelihood theory, the likelihood function

$\mathcal{L}(\theta\,;m,x_o) = p(m,x_o\,;\theta)$ turns out to be Fisher consistent for θ, since

$$\begin{aligned}
\mathrm{E}\left\{\frac{\partial \log \mathcal{L}(\theta\,;m,x_o)}{\partial\theta}\right\} &= \int\int\int \frac{\partial \log\, p(m,x_o\,;\theta)}{\partial\theta}\cdot p(m,x_o,x_m\,;\theta)\,dx_m\,dx_o\,dm \\
&= \int\int \frac{\partial \log\, p(m,x_o\,;\theta)}{\partial\theta}\cdot p(m,x_o\,;\theta)\,dx_o\,dm \\
&= \int\int \frac{\partial p(m,x_o\,;\theta)}{\partial\theta}\,dx_o\,dm \\
&= \frac{\partial}{\partial\theta}\int\int p(m,x_o\,;\theta)\,dx_o\,dm = \frac{\partial 1}{\partial\theta} = 0\,.
\end{aligned}$$

Note that

$$\mathcal{L}(\theta\,;m,x_o) = p(m\,;\theta)\,p(x_o\,|\,m\,;\theta) = p(x_o\,;\theta)\,p(m\,|\,x_o\,;\theta)\,,$$

where $p(x_o\,;\theta)$ denotes the marginal distribution of the observed data X_o and o is the realized index of observations.

Now suppose that the missingness pattern is not determined by the model parameter under the observed data. That means $p(m\,|\,x_o\,;\theta)$ is invariant under a change of θ. In that case the missingness pattern is not relevant and it can be ignored for maximum likelihood estimation, since

$$\mathcal{L}(\theta\,;m,x_o) \propto p(x_o\,;\theta) = \mathcal{L}_o(\theta\,;x_o)\,.$$

Hence, for estimating θ it is sufficient to concentrate on the marginal distribution of X_o. This is the *ignorability assumption* of missing-data analysis and $\mathcal{L}_o(\theta\,;x_o)$ is called the *observed-data likelihood function* (Schafer, 1997, Section 2.3.1).

For justifying the ignorability assumption, the conditional distribution

$$p(m\,|\,x_o\,;\theta) = \int p(m\,|\,x_o,x_m\,;\theta)\,p(x_m\,|\,x_o\,;\theta)\,dx_m$$

has to be examined. In many circumstances it can be assumed that the distribution of M depends on the complete data X but *not* on the specific parameter θ. For example, non-responses in questionnaires might be determined by the individual outcomes x_o and x_m but it is unlikely that the missingness pattern depends on the model parameter θ per se. The so-called *distinctness assumption* of missing-data analysis conveys that $p(m\,|\,x_o,x_m\,;\theta)$ is not determined by θ. If the distinctness assumption can be accepted it follows that

$$p(m\,|\,x_o\,;\theta) = \int p(m\,|\,x_o,x_m)\,p(x_m\,|\,x_o\,;\theta)\,dx_m\,.$$

Now there are two non-excluding cases where the ignorability assumption is satisfied, viz.

1. either $p(x_m\,|\,x_o\,;\theta)$ is not determined by θ

2. or $p(m \,|\, x_o, x_m)$ is not determined by x_m .

ad a. The distribution of the complete data X is determined by θ. However, if $p(x_o, x_m \,; \theta) = p(x_o \,; \theta)\, p(x_m)$, then $p(x_m \,|\, x_o \,; \theta)$ is not driven by θ and the ignorability assumption is satisfied. That means if the unobserved data are independent of the observed data and do not contain any information about the unknown parameter, the missing data can be ignored.

ad b. If $p(m \,|\, x_o, x_m)$ is not determined by the unobserved data x_m, it holds that

$$p(m \,|\, x_o \,; \theta) = \int p(m \,|\, x_o)\, p(x_m \,|\, x_o \,; \theta)\, dx_m = p(m \,|\, x_o)\,.$$

This condition is clearly satisfied if M is stochastically independent of X_M. In that case x_m is said to be *missing at random* (MAR) (Little and Rubin, 2002, p. 12). However, it is worth to point out that the ignorability assumption holds true if the invariance property is satisfied only for the *realized* missingness pattern m (Schafer and Graham, 2002). That means for estimating the parameter θ by the marginal distribution of the observed data it is not relevant whether some of the observed data x_o would be MAR or not MAR (NMAR) in case that they were missing. Nevertheless, for *likelihood inference* in the context of missing data, the MAR assumption must be satisfied (Kenward and Molenberghs, 1998). If M is not only independent of the unobserved data X_M but also of the observed data X_O, the missing data are *missing completely at random* (MCAR) (Little and Rubin, 2002, p. 12).

In the next section the spectral density approach will be adapted to incomplete data, but before that we have to discuss an important drawback of missing-data analysis. Let $y = g(x_o)$ be some measurable function of the observed data and $q(y \,; \theta)$ the corresponding density. A naive application of missing-data analysis would suggest to estimate θ by using the observed-data likelihood function related to $q(y \,; \theta)$, i.e. $\mathcal{L}_y(\theta \,; y) = q(y \,; \theta)$, instead of $\mathcal{L}_o(\theta \,; x_o)$ if x_m is MAR. For example, this approach is suitable if $\mathcal{L}_y$ leads to a robust or even distribution-free estimator for θ (see Section 8.3.2). In that case it has to be guaranteed that

$$\mathcal{L}_y(\theta \,; m, y) = q(y \,; \theta)\, p(m \,|\, y \,; \theta) \propto q(y \,; \theta) = \mathcal{L}_y(\theta \,; y)$$

with $p(m \,|\, y \,; \theta) = p(m \,|\, g(x_o) \,; \theta)$.

Note that in most practical situations it is unlikely that the missingness is 'triggered' by the transformed data y rather than the original data x, in particular since y is usually calculated

after observing the data. Suppose that the function g is not injective. The distribution of X_o under the condition $g(X_o) = y$ in general is a function of θ and thus if the distribution of m given x_o and x_m is only determined by the observed data, $p(m \,|\, g(x_o)\,;\theta)$ might be essentially determined by the parameter θ. Hence, the distinctness assumption usually cannot be accepted when working with transformed data if the transformation is not one-to-one. For instance consider a projection of a d-dimensional random vector onto some subspace or manifold in $\mathbb{R}^d$. In that case the missingness pattern is no more ignorable. Only if the missing data are MCAR the distinctness assumption remains plausible since it is simply assumed that $p(m \,|\, g(x_o)\,;\theta) = p(m\,;\theta) = p(m)$.

The Spectral Density Approach

Lemma 8.5 *Let X be a d-dimensional generalized elliptically distributed and centered random vector with dispersion matrix $\Sigma \in \mathbb{R}^{d\times d}$. Consider the partitions*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \textit{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where X_1 is an r-dimensional sub-vector of X. It holds that X_1 is a generalized elliptically distributed and centered random vector with dispersion matrix $\Sigma_{11} \in \mathbb{R}^{r\times r}$.

Proof. Write $X_1 = \mathcal{I}X$ with $\mathcal{I} := [\, I_r \; 0\,]$ $(r \times d)$. That means $X_1 = \mathcal{I}\Lambda\mathcal{R}U$ and note that $\mathcal{I}\Lambda\Lambda^{\mathsf{T}}\mathcal{I}^{\mathsf{T}} = \mathcal{I}\Sigma\mathcal{I}^{\mathsf{T}} = \Sigma_{11}$. ∎

More generally, let $X = (X_o, X_m)$ be a generalized elliptically distributed random vector which is divided into an observed and an unobserved part according to some fixed missingness pattern m. Correspondingly, the vector $x = (x_o, x_m)$ denotes a realization of the complete data, where x_o is an r-dimensional sub-vector of x. Further, let $s = x_o/\|x_o\|$ be the observed data projected to the unit hypersphere in $\mathbb{R}^r$ and $S = X_o/\|X_o\|$ the corresponding random vector. From the preceding lemma it is known that the distribution of $V =_{\text{a.s.}} \pm S$ is given by

$$\psi(v) = \frac{\Gamma(r/2)}{2\pi^{r/2}} \cdot \sqrt{\det \Sigma_o^{-1}}\, \sqrt{s^{\mathsf{T}}\Sigma_o^{-1}s}^{\,-r},$$

where Σ_o denotes that part of Σ which is related to the observation x_o. Once again the generating variate of X_o does not play any role for estimating Σ since it is canceled out by the projection onto the unit hypersphere (see Eq. 8.5).

Now consider a sample of possibly dependent and not identically generalized elliptically distributed random vectors $X_1 \ldots, X_n$, where only the realizations $x_{o1}, \ldots, x_{on}$ of the sub-vectors $X_{o1}, \ldots, X_{on}$ can be observed. More precisely, it is assumed that X_t $(t = 1, \ldots, n)$ can be represented as described in Section 8.2.3, where $\mu = 0$ without loss of generality, $\Sigma = \Lambda\Lambda^{\mathsf{T}}$ is positive-definite, and the distribution of X_t has no atom at zero. Hence, the observed data can be written as

$$X_{ot} = \mathcal{I}_t X_t = \mathcal{I}_t \Lambda \mathcal{R}_t U_t = \Lambda_t \mathcal{R}_t U_t \,, \qquad t = 1, \ldots, n \,,$$

where $\mathcal{I}_t$ is a matrix which converts X_t into X_{ot} and $\Lambda_t := \mathcal{I}_t \Lambda$. Now it is only assumed that the angular parts $U_1, \ldots, U_n$ are mutually independent, whereas the joint distribution of the radial parts $\mathcal{R}_1, \ldots, \mathcal{R}_n$ is irrelevant. That means the generating variates might depend on each other and do not need to be identically distributed. This feature especially allows for several kinds of multivariate autoregressive conditional heteroscedasticity imposed by the variation of $\mathcal{R}$ in time (Bade et al., 2008). Moreover, in the following it is supposed that the missing data $x_{m1}, \ldots, x_{mn}$ are MCAR.

Let Σ_t be the sub-matrix of Σ associated with the observation x_{ot} and $s_t := x_{ot}/\|x_{ot}\|$ $(t = 1, \ldots, n)$ the corresponding projection on the unit hypersphere. Moreover, let d_t be the number of components of that observation. Then the observed-data likelihood function is given by

$$\mathcal{L}_s(\Sigma \,; v_1, \ldots, v_n) = \prod_{t=1}^{n} \psi\big(v_t \,; \Sigma_t\big) \propto \prod_{t=1}^{n} \sqrt{\det \Sigma_t^{-1}} \, \sqrt{s_t^{\mathsf{T}} \Sigma_t^{-1} s_t}^{\,-d_t} \,. \tag{8.8}$$

Note that $v_t = \pm s_t$ is not a one-to-one function of x_{ot} and due to the arguments given at the end of Section 8.3.2 it has to be supposed that the missing data $x_{m1}, \ldots, x_{mn}$ are MCAR. Then a proper ML-estimate of Σ can be obtained by maximizing $\mathcal{L}_s(\Sigma \,; v_1, \ldots, v_n)$, provided the number of observations is large enough. The observed-data log-likelihood function

$$\log \mathcal{L}_s(\Sigma \,; v_1, \ldots, v_n) = c + \frac{1}{2} \sum_{t=1}^{n} \log \det \Sigma_t^{-1} - \frac{1}{2} \sum_{t=1}^{n} d_t \log\big(s_t^{\mathsf{T}} \Sigma_t^{-1} s_t\big) \tag{8.9}$$

can be used alternatively, where c is some constant.

Since $\mathcal{L}_s$ is a scale-invariant function, i.e. $\mathcal{L}_s(\alpha\Sigma) = \mathcal{L}_s(\Sigma)$ for every $\alpha > 0$, the scale of Σ has to be fixed by the additional constraint $\det \Sigma = 1$. This leads to the spectral estimator for the shape matrix Ω , i.e. $\widehat{\Omega} = (\det \widehat{\Sigma}\,)^{-1/d} \widehat{\Sigma}$, where $\widehat{\Sigma}$ denotes the ML-estimator for Σ .

Figure 8.2 shows a spectral estimate for a sample of multivariate t-distributed data with 2 degrees of freedom possessing a monotone missingness pattern (Little and Rubin, 2002,

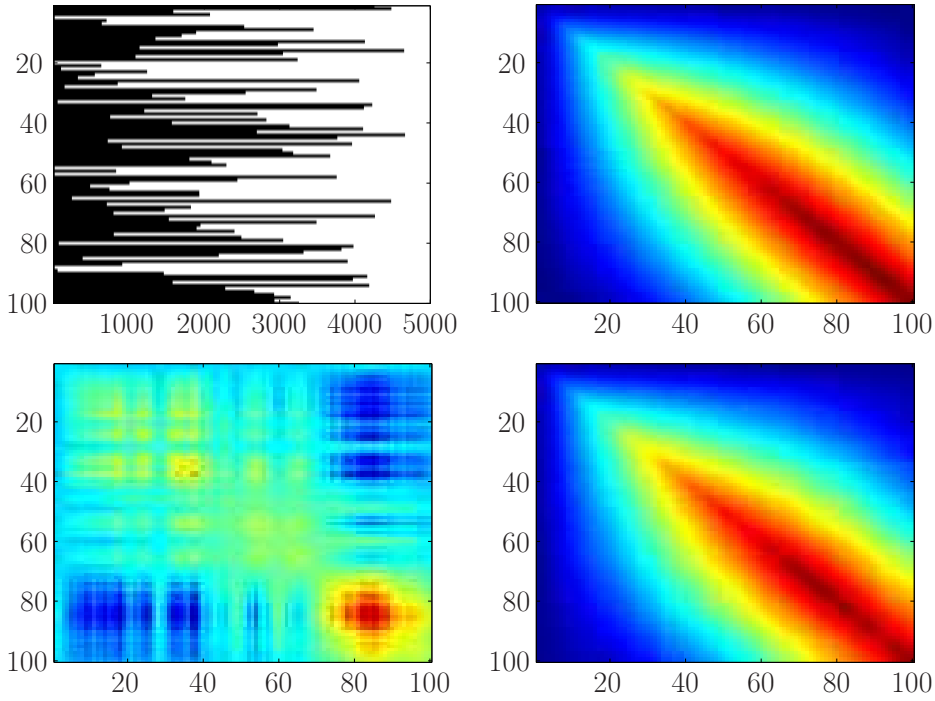

Figure 8.2.: Missingness pattern (upper left) of a simulated sample of 5000 independent copies of a 100-dimensional t-distributed random vector with 2 degrees of freedom, location vector $\mu = 0$, and dispersion matrix Σ proportional to the shape matrix given on the upper right (violet cells indicate small numbers and red cells large numbers). There are 215184 missing values (43% of the sample). The corresponding spectral estimate is given on the lower right, whereas the ML-estimate based on the normal distribution assumption can be found on the lower left. The computational time for the spectral estimate on a standard PC (3 GHz CPU) amounts to 1 minute and 58 seconds.

p. 5). This can be compared with the corresponding ML-estimate based on the normal distribution assumption. Obviously, the Gaussian estimator is not robust against extreme realizations of the multivariate t-distribution. In Figure 8.3 the same experiment is done with multivariate normally distributed data. The spectral estimate looks pretty much the same as the Gaussian one in agreement with the simulation study discussed in Section 8.5, showing that the loss of efficiency is small even for normally distributed data.

The unknown location vector μ can be replaced by some appropriate estimate $\hat{\mu}_n$ for centralizing the data. We suggest to use the *factored likelihood method* described by Little and Rubin (2002, Ch. 7) as well as Schafer (1997, Section 6.5). This leads to the maximum of an observed-data likelihood function where the data are assumed to follow a multivariate normal distribution. It is guaranteed that $\hat{\mu}_n$ is asymptotically unbiased and normally

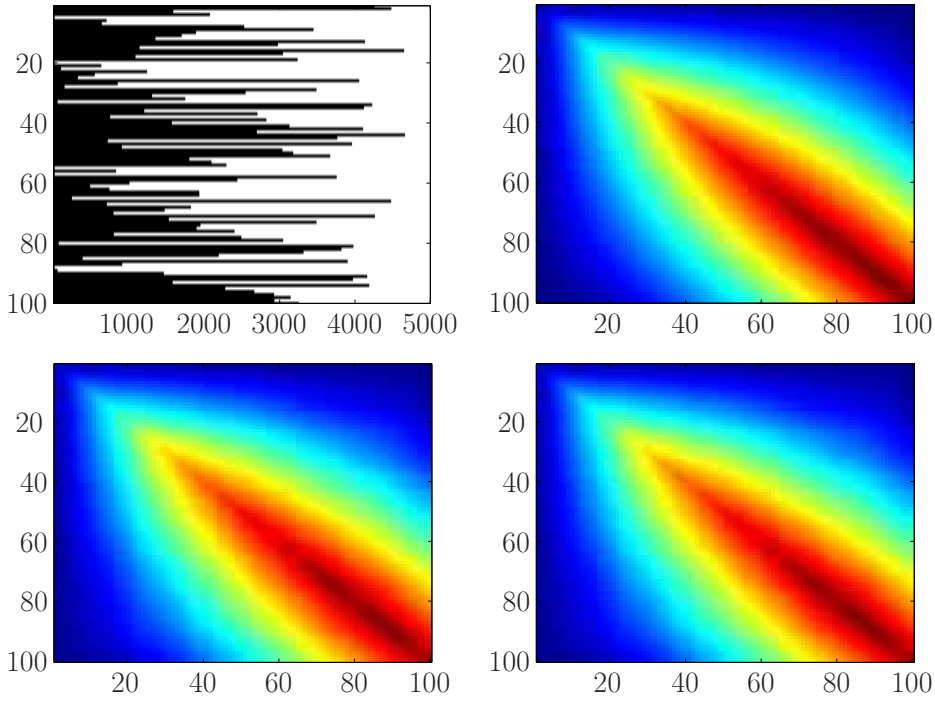

Figure 8.3.: Missingness pattern (upper left) of a simulated sample of 5000 independent copies of a 100-dimensional normally distributed random vector with location vector $\mu = 0$ and dispersion matrix Σ proportional to the shape matrix given on the upper right (see also Figure 8.2). The corresponding spectral estimate is given on the lower right, whereas the ML-estimate based on the normal distribution assumption can be found on the lower left. The computational time for the spectral estimate on a standard PC (3 GHz CPU) amounts to 2 minutes and 19 seconds.

distributed for $n \to \infty$ under the typical regularity conditions of ML-theory. Thus it is *not* recommended to use an 'older method' (Schafer and Graham, 2002) of missing-data analysis since these procedures generally produce asymptotically biased estimates if the missing data are not MCAR. Of course, for applying our ML-approach it has to be presumed that the centered data are *asymptotically* independent of each other. This is true if $\hat{\mu}_n$ converges to μ at an appropriate rate. As mentioned earlier, for the complete-data case this has been already investigated by Tyler (1987a). Presumably the same arguments hold also in the incomplete-data case but at present we cannot provide a formal proof for this conjecture. However, in our numerical simulations we did not encounter any problems concerning the spectral estimator caused by the estimation of μ provided the missing data are MCAR and the data stem from an elliptically symmetric distribution.

We would like to clarify the main purpose of our method. Of course, modern estimation

procedures of missing-data analysis (Little and Rubin, 2002, Schafer and Graham, 2002) could be efficiently applied for estimating the shape matrix if the *true data-generating process* was known. Traditional ML-theory works only if the proposed model is correct. In contrast, by M-theory the asymptotic distribution of covariance or shape matrix estimators can be calculated if the suggested model does not correspond to the true one (**?**Maronna, 1976, Tyler, 1982). However, there are some remaining difficulties regarding traditional robust covariance matrix estimation. For evaluating the asymptotic distribution of an M-estimator, generally some nuisance parameters have to be estimated (Tyler, 1982). Other robust estimation procedures are based on geometrical approaches (Visuri, 2001, Ch. 3) and cannot be generalized to the missing-data problem. To the best of our knowledge, M-estimators for incomplete data have been only discussed by Little (1988).

Little (1988) suggests to maximize the observed-data likelihood functions of heavy-tailed or contaminated data. Usually this leads to robust estimates of the shape matrix and obviously the method is similar to the spectral density approach. However, Little's estimators are based on a multivariate t-distribution or a contaminated normal-distribution assumption and so his approach is *parametrical* rather than distribution-free. With a parametrical approach one has only limited information about the asymptotic distribution of the estimators if the model is misspecified. This can be avoided for the most part by using the spectral estimator due to its invariance property discussed above. The only conditions which have to be guaranteed are that

(1) the sample consists of data which are generalized elliptically distributed (even serial dependence is allowed in the weak sense described above),

(2) the missing part of the sample is MCAR, and

(3) μ either is known or can be properly estimated by missing-data analysis.

Asymptotic Distribution of the Spectral Estimator

In the complete-data case it can be shown (**?**) that

$$\sqrt{n}\left(\widehat{\Omega}-\Omega\right) \xrightarrow{\mathrm{d}} \mathcal{N}_{d\times d}\big\{0, V(\Omega)\big\}\,, \qquad n\longrightarrow\infty\,,$$

where

$$V(\Omega)=\frac{d+2}{d}\cdot\left\{(I_{d^2}+K_{d^2})(\Omega\otimes\Omega)-2/d\cdot(\mathrm{vec}\;\Omega)(\mathrm{vec}\;\Omega)^{\mathsf{T}}\right\},$$

and K_{d^2} denotes the $d^2 \times d^2$ *commutation matrix* (Schott, 1997, p. 277). Further, the d^2-dimensional vector vec Ω is obtained by stacking the columns of Ω on top of each other.

In the incomplete-data case $\log \mathcal{L}_s$ is a proper log-likelihood function under the conditions given in Section 8.3.2 and so the spectral estimator turns out to be asymptotically unbiased, normally distributed, and consistent. For calculating its asymptotic covariance matrix, the Fisher information has to be calculated either by the score function or the Hessian of $\log \mathcal{L}_s$. The following proposition can be used for calculating the score function.

Proposition 8.6 *Let v be a d-dimensional vector with unit length and $\Sigma \in \mathbb{R}^{d\times d}$ positive-definite. The partial derivative of $\log \psi(v\,;\Sigma)$ with respect to Σ is given by*

$$\frac{\partial \log \psi(v\,;\Sigma)}{\partial \Sigma} = \left(d \cdot \frac{\Sigma^{-1} v v^{\mathsf{T}} \Sigma^{-1}}{v^{\mathsf{T}} \Sigma^{-1} v} - \Sigma^{-1} \right) - \frac{1}{2} \cdot \operatorname{diag}\left(d \cdot \frac{\Sigma^{-1} v v^{\mathsf{T}} \Sigma^{-1}}{v^{\mathsf{T}} \Sigma^{-1} v} - \Sigma^{-1} \right).$$

Proof. Frahm (2004, p. 70). ∎

The Fisher information of an observed data point $S_t = X_{ot}/\|X_{ot}\|$ $(t = 1, \ldots, n)$ is given by the $\binom{d+1}{2} \times \binom{d+1}{2}$ matrix

$$\mathcal{F}_t(\Sigma) = \mathrm{E}\left\{ \operatorname{vech}\left(\frac{\partial \log \psi(V_t\,;\Sigma_t)}{\partial \Sigma} \right) \operatorname{vech}\left(\frac{\partial \log \psi(V_t\,;\Sigma_t)}{\partial \Sigma} \right)^{\mathsf{T}} \right\}, \tag{8.10}$$

where the vech-operator converts the lower triangular part of a symmetric matrix to a column vector. Note that S_t refers only to the observed part of the d-dimensional random vector X_t realized for $t \in \{1, \ldots, n\}$ and thus $\log \psi(V_t\,;\Sigma_t)$ is invariant against that part of Σ which is not related to the available observation. That means there exists a $d_t \times d_t$ matrix

$$\frac{\partial \log \psi(V_t\,;\Sigma_t)}{\partial \Sigma_t} = \left(d_t \cdot \frac{\Sigma_t^{-1} S_t S_t^{\mathsf{T}} \Sigma_t^{-1}}{S_t^{\mathsf{T}} \Sigma_t^{-1} S_t} - \Sigma_t^{-1} \right) - \frac{1}{2} \cdot \operatorname{diag}\left(d_t \cdot \frac{\Sigma_t^{-1} S_t S_t^{\mathsf{T}} \Sigma_t^{-1}}{S_t^{\mathsf{T}} \Sigma_t^{-1} S_t} - \Sigma_t^{-1} \right),$$

but here the $d \times d$ matrix $\partial \log \psi(V_t\,;\Sigma_t)/\partial \Sigma$ has to be considered. The latter contains zeros according to each element of Σ which does not belong to the sub-matrix Σ_t.

Now

$$\sqrt{n}\left(\operatorname{vech} \widehat{\Sigma} - \operatorname{vech} \Sigma\right) \xrightarrow{\mathrm{d}} \mathcal{N}_{\binom{d+1}{2}}\left\{0, \mathcal{F}(\Sigma)^{-1}\right\}, \qquad n \longrightarrow \infty,$$

where

$$\mathcal{F}(\Sigma) := \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathcal{F}_t(\Sigma)$$

denotes the asymptotic average Fisher information. However, in the following we will use the vec-operator which converts the *whole* matrix to a column vector. This can be

obtained after a preceding vech-operation by $\operatorname{vec}\Sigma = D_d \operatorname{vech}\Sigma$, where D_d represents the $d^2 \times d(d+1)/2$ *duplication matrix* (Schott, 1997, p. 283). Hence, the asymptotic distribution can be written as

$$\sqrt{n}\,(\widehat{\Sigma} - \Sigma) \xrightarrow{\mathrm{d}} \mathcal{N}_{d\times d}\big\{0, D_d \mathcal{F}(\Sigma)^{-1} D_d^{\mathsf{T}}\big\}\,, \qquad n \longrightarrow \infty\,.$$

However, up to here we considered some artificial derivations since we missed to account for the normalization $\Sigma \to \Omega$. Interestingly, such a normalization has a substantial impact on the asymptotic distribution of $\widehat{\Sigma}$ (**?**Paindaveine, 2008). Following the arguments of **?** it can be concluded that

$$\sqrt{n}\,(\widehat{\Omega} - \Omega) \xrightarrow{\mathrm{d}} \mathcal{N}_{d\times d}\big\{0, \Psi D_d \mathcal{F}(\Omega)^{-1} D_d^{\mathsf{T}} \Psi^{\mathsf{T}}\big\}\,, \qquad n \longrightarrow \infty\,,$$

with $\Psi := I_{d^2} - (\operatorname{vec}\Omega)(\operatorname{vec}\Omega^{-1})^{\mathsf{T}}/d$.

Note that in contrast to other M-estimators the asymptotic distribution of the spectral estimator is only determined by the dimensions $d_1, \ldots, d_n$ of the observed data $x_{o1}, \ldots, x_{on}$ and the true shape matrix Ω. More precisely, it is possible to assess the asymptotic distribution of the spectral estimator without any parametrical assumption concerning the generating distribution, i.e. the distribution of $\mathcal{R}$ and so there are no nuisance parameters which have to be estimated for statistical inference. This is also confirmed by the results presented in Section 8.5. That means the asymptotic covariance matrix of the spectral estimator solely follows from the spectral estimate itself which is an important advantage compared to other M-estimators.

From Section 8.3.2 it becomes clear that under the MCAR assumption both the score functions and the Hessians belonging to $\mathcal{L}_s(\Sigma\,; m, v_1, \ldots, v_n)$ and $\mathcal{L}_s(\Sigma\,; v_1, \ldots, v_n)$ correspond to each other. For the application of large-sample theory in the context of missing data we follow the arguments given by Kenward and Molenberghs (1998) as well as Schafer and Graham (2002). That is we suggest to estimate the Fisher information in a nonparametric way by the *observed data* rather than calculating the expectations given by Eq. 8.10 analytically or numerically. Hence, an appropriate estimate for $\mathcal{F}_t(\Omega)$ is

$$\widehat{\mathcal{F}}_t(\widehat{\Omega}) = \operatorname{vech}\left(\frac{\partial \log\psi(v_t\,; \widehat{\Omega}_t)}{\partial\Omega}\right)\operatorname{vech}\left(\frac{\partial \log\psi(v_t\,; \widehat{\Omega}_t)}{\partial\Omega}\right)^{\mathsf{T}},$$

where v_t can be replaced by the observed data point $s_t = x_{ot}/\|x_{ot}\|$ and $\widehat{\Omega}_t$ is the part of $\widehat{\Omega}$ which is related to that observation. The asymptotic average Fisher information can be consistently estimated by

$$\widehat{\mathcal{F}}(\widehat{\Omega}) := \frac{1}{n}\sum_{t=1}^{n} \widehat{\mathcal{F}}_t(\widehat{\Omega})$$

and so a large-sample approximation is given by

$$\sqrt{n}\,\big(\widehat{\Omega} - \Omega\big) \,\dot{\sim}\, \mathcal{N}_{d\times d}\big\{0, \widehat{\Psi} D_d\, \widehat{\mathcal{F}}\big(\widehat{\Omega}\big)^{-1} D_d^{\mathsf{T}} \widehat{\Psi}^{\mathsf{T}}\big\}\,,$$

where $\widehat{\Psi} := I_{d^2} - \big(\operatorname{vec}\,\widehat{\Omega}\big)\big(\operatorname{vec}\,\widehat{\Omega}^{-1}\big)^{\mathsf{T}}/d\,.$

8.4. Numerical Implementation

In the following it is assumed that there exists a positive-definite matrix $\widehat{\Omega}$ which maximizes the observed-data likelihood function given by (8.8). A necessary condition for the existence under a monotone missingness pattern can be found later on in Theorem 8.8. Note that the dispersion matrix Σ is symmetric. This is the reason why the main diagonal has to be subtracted in Proposition 8.6. Since the set of all symmetric positive-definite matrices with unit determinant is open, $\widehat{\Omega}$ is a stationary point of the observed-data log-likelihood function (8.9) and the main diagonal part from Proposition 8.6 can be omitted. It follows that the log-likelihood equation can be written as

$$\sum_{t=1}^{n} \left[d_t \cdot \frac{\widehat{\Omega}_t^{-1} s_t s_t^{\mathsf{T}} \widehat{\Omega}_t^{-1}}{s_t^{\mathsf{T}} \widehat{\Omega}_t^{-1} s_t} \right] - \sum_{t=1}^{n} \left[\widehat{\Omega}_t^{-1} \right] = 0\,. \tag{8.11}$$

As already mentioned in Section 8.3.2, since $\partial \log \mathcal{L}_s / \partial \Sigma$ is a $d \times d$ matrix, each $d_t \times d_t$ matrix

$$d_t \cdot \frac{\widehat{\Omega}_t^{-1} s_t s_t^{\mathsf{T}} \widehat{\Omega}_t^{-1}}{s_t^{\mathsf{T}} \widehat{\Omega}_t^{-1} s_t} \qquad \text{and} \qquad \widehat{\Omega}_t^{-1}$$

in Eq. 8.11 has to be inflated by zeros according to the positions of $\widehat{\Omega}$ which do not belong to $\widehat{\Omega}_t$ such that the aforementioned $d_t \times d_t$ matrices become the $d \times d$ matrices

$$\left[d_t \cdot \frac{\widehat{\Omega}_t^{-1} s_t s_t^{\mathsf{T}} \widehat{\Omega}_t^{-1}}{s_t^{\mathsf{T}} \widehat{\Omega}_t^{-1} s_t} \right] \qquad \text{and} \qquad \left[\widehat{\Omega}_t^{-1} \right]\,.$$

Now define a function $F\colon\ \mathcal{P}^d \to \mathbb{R}^{d\times d}$, where $\mathcal{P}^d$ denotes the set of all symmetric and positive-definite $d \times d$ matrices, by

$$F(\Sigma) := \Sigma f(\Sigma) \Sigma\,,$$

where

$$f(\Sigma) := \sum_{t=1}^{n} \left[d_t \cdot \frac{\Sigma_t^{-1} s_t s_t^{\mathsf{T}} \Sigma_t^{-1}}{s_t^{\mathsf{T}} \Sigma_t^{-1} s_t} \right] - \sum_{t=1}^{n} \left[\Sigma_t^{-1} \right] \tag{8.12}$$

for all $\Sigma \in \mathcal{P}^d$. The spectral estimator $\widehat{\Omega}$ corresponds to a fixed-point solution of

$$G(\Sigma) := \Sigma + \alpha F(\Sigma) = \Sigma\,, \tag{8.13}$$

where $0 < \alpha \leq 1/n$. The next proposition guarantees that not only Σ but also $G(\Sigma)$ is positive-definite if the random vector X is continuously distributed.

Proposition 8.7 *Let $x_1, \ldots, x_n$ be a realized sample of independent copies of a d-dimensional centered random vector X possessing a continuous distribution on $\mathbb{R}^d$. Further, let $x_{o1}, \ldots, x_{on}$ be the corresponding sample of observations following an arbitrary missingness pattern. Denote the number of complete observations by $m \leq n$ and consider the map*

$$G(\Sigma) = \Sigma + \alpha F(\Sigma),$$

where Σ is a symmetric and positive-definite $d \times d$ matrix and $0 < \alpha \leq 1/n$. If $n \geq m \geq d$, the $d \times d$ matrix $G(\Sigma)$ is symmetric and positive-definite, too.

Proof. Since the data are continuously distributed and $m \geq d$, the $d \times d$ matrix

$$\alpha\Sigma\left(\sum_{t=1}^{n}\left[d_t \cdot \frac{\Sigma_t^{-1} s_t s_t^{\mathsf{T}} \Sigma_t^{-1}}{s_t^{\mathsf{T}} \Sigma_t^{-1} s_t}\right]\right)\Sigma$$

is positive-definite almost surely. Hence, it suffices to prove that

$$\Sigma - \frac{1}{n} \cdot \Sigma\left(\sum_{t=1}^{n}\left[\Sigma_t^{-1}\right]\right)\Sigma = \frac{1}{n}\sum_{t=1}^{n}\left(\Sigma - \Sigma\left[\Sigma_t^{-1}\right]\Sigma\right) \tag{8.14}$$

is positive-semidefinite. Note that without loss of generality

$$\Sigma - \Sigma\left[\Sigma_{11}^{-1}\right]\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} - \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{bmatrix}.$$

Since Σ is positive-definite the remaining Schur complement is positive-definite, too, i.e. the matrix given by Eq. 8.14 is positive-semidefinite. ∎

Note that in the complete-data case it follows that

$$G(\Sigma) = \frac{d}{n}\sum_{t=1}^{n}\frac{s_t s_t^{\mathsf{T}}}{s_t^{\mathsf{T}}\Sigma^{-1}s_t}$$

by setting $\alpha = 1/n$ and so the fixed-point problem given by (8.13) is equivalent to (8.6), i.e. finding Tyler's M-estimator. According to Tyler's fixed-point algorithm already discussed in Section 8.3.1, some initial value $\Sigma_{(0)}$ has to be chosen. In the next step calculate

$$\Sigma_{(1)} = \Sigma_{(0)} + \alpha F(\Sigma_{(0)})$$

and substitute $\Sigma_{(0)}$ by $\Sigma_{(1)}$, etc. Hence, our fixed-point algorithm is given by the sequence

$$\Sigma_{(i+1)} = \Sigma_{(i)} + \alpha F(\Sigma_{(i)}), \qquad i = 0, 1, \ldots . \tag{8.15}$$

After performing a sufficiently large number N of iterations, the spectral estimate can be approximated by

$$\widehat{\Omega} \approx \frac{\Sigma_{(N)}}{(\det \Sigma_{(N)})^{1/d}} .$$

The initial value can be chosen as $\Sigma_{(0)} = I_d$ and for minimizing the number of iterations it is recommended to take the upper bound $\alpha = 1/n$. In the following we will concentrate on the numerical evaluation of $f(\Sigma_{(i)})$ and for notational convenience we write $f_i \equiv f(\Sigma_{(i)})$. Beforehand it is worth to point out that any observation $x_t = 0 \in \mathbb{R}^d$ or $x_t \in \mathbb{R}$ should be discarded. The former argument is clear from the preceding discussion and the latter argument follows immediately by noting that

$$\left[d_t \cdot \frac{\Sigma_t^{-1} s_t s_t^{\mathsf{T}} \Sigma_t^{-1}}{s_t^{\mathsf{T}} \Sigma_t^{-1} s_t} \right] = \left[\Sigma_t^{-1} \right]$$

in case $s_t = \pm 1$. That means univariate data do not contain any valuable information for solving the log-likelihood equation and in the following n shall quantify the number of *useful* observations.

If there are many observations sharing the same missing components, these observations should be put together. This holds especially if the missingness pattern is monotone. Then the matrix Σ_t^{-1} in Eq. 8.12 has to be calculated only once for all observations possessing the same missingness. However, from a numerical perspective it is rather inefficient to compute the inverse Σ_t^{-1} if d_t is large, particularly if there are many observations with only a few missing values. If there are, e.g., 1000 realizations of a 100-dimensional random vector with 10% of its components missing at random, the inverse of a 90×90 sub-matrix of Σ has to be computed for each realization. Suppose that each inverse could not be re-used, since if the missingness pattern is irregular it is unlikely that two realizations share the same missingness. However, once the full inverse of Σ has been computed, it can be used for evaluating the inverses of the sub-matrices Σ_t $(t = 1, \ldots, n)$ more efficiently. Consider the partition

$$\Sigma^{-1} = \begin{bmatrix} A & B^{\mathsf{T}} \\ B & C \end{bmatrix} ,$$

where the $d_t \times d_t$ matrix A occupies the same range in Σ^{-1} as Σ_t in Σ. Then the inverse of Σ_t can be calculated by the Schur complement

$$\Sigma_t^{-1} = A - B^{\mathsf{T}} C^{-1} B .$$

That means instead of calculating the inverse of the $d_t \times d_t$ matrix Σ_t, only the inverse of the $(d - d_t) \times (d - d_t)$ matrix C has to be calculated. In the case discussed above,

this corresponds to the solutions of 10×10 rather than 90×90 linear systems for 1000 observations. Of course, this is only recommended for each observation where $d - d_t$, i.e. the number of missing values is small, since otherwise it could be more efficient to calculate the inverse of Σ_t by another method (see below).

If the missingness pattern is monotone, we suggest to use the *sweep operator* (Beaton, 1964, Goodnight, 1979) for calculating the inverses of the sub-matrices. A sweep operation on a symmetric and positive-definite $d \times d$ matrix Σ is a simple manipulation of Σ (Little and Rubin, 2002, p. 221) which produces another symmetric and positive-definite matrix. There also exists an inverse function which can be used for reversing a previous sweep operation. By applying the sweep and the reverse sweep operator iteratively, the inverse of a sub-matrix Σ_t can be efficiently calculated from the inverse of another similar sub-matrix of Σ which is already given by a preceding step.

The fixed-point iteration scheme given by (8.15) always produces symmetric matrices, analytically, but our own experience shows that the numerical computation of inverses (or, in a more efficient implementation, solutions of linear systems) leads to roundoff errors that make the iterations slightly asymmetric. The asymmetric component is tiny in the beginning, but can blow up especially in higher dimensions after 50 or 100 iterations. This can be easily avoided by symmetrizing $f_i \to (f_i + f_i^{\mathsf{T}})/2$ in every iteration.

Of course, the fixed-point algorithm given by (8.15) can only work if the missingness is not too strong. In the following we give a necessary condition for the existence of the spectral estimate provided the missingness pattern is monotone.

Theorem 8.8 *Let $x_1, \ldots, x_n$ be a realized sample of independent copies of a d-dimensional centered random vector X possessing an arbitrary distribution on $\mathbb{R}^d$. Further, let $x_{o1}, \ldots, x_{on}$ be the corresponding sample of observations following a monotone missingness pattern. Denote the number of complete observations by $m \leq n$. Then the spectral estimate exists only if $n \geq m \geq d$.*

Proof. Suppose that the spectral estimate $\widehat{\Omega}$ exists. Then Eq. 8.11 can be written as

$$\sum_{t=1}^{n} \left[\widehat{\Omega}_t^{-1} y_t y_t^{\mathsf{T}} \widehat{\Omega}_t^{-1} \right] - \sum_{t=1}^{n} \left[\widehat{\Omega}_t^{-1} \right] = 0 \,,$$

where $y_t := (d_t / x_{ot}^{\mathsf{T}} \widehat{\Omega}_t^{-1} x_{ot})^{1/2} \, x_{ot}$. Note that this corresponds to the log-likelihood equation for centered and weighted observations $y_1, \ldots, y_n$ under the normal distribution assumption. That means the spectral estimator can only exist for $x_{o1}, \ldots, x_{on}$ if the Gaussian

estimate exists for $y_1, \ldots, y_n$. The latter can be obtained by factorizing the observed-data likelihood function (Schafer, 1997, Ch. 6.5.1) and following the method described by Schafer (1997, Ch. 6.5.2) where μ has to be set to zero. Due to elementary properties of the sweep operator it becomes clear that the first sweep operation leads to a singular Schur complement in case $m < d$. Hence, after finishing all necessary sweep operations the final reverse sweep operation cannot produce a nonsingular covariance matrix estimate. That is $\widehat{\Omega}$ cannot exist which contradicts the assertion at the beginning of the proof. ■

In a separate work the authors will discuss necessary and sufficient conditions for the existence of the spectral estimator given more general missingness patterns. The mathematical details are rather complicated and would go beyond the scope of the present work. However, for convenience we would like to mention that for the existence of a spectral estimate it is sufficient to have a continuous distribution on $\mathbb{R}^d$ possessing an arbitrary missingness pattern with $m > d$ complete observations.

8.5. Simulation Study

In the following simulation study the spectral estimator is compared with the shape matrix estimator based on the normal distribution assumption. More precisely, if $X_1, \ldots, X_n$ is a sample of a d-dimensional centered random vector X then

$$\widehat{\Sigma} := \frac{1}{n} \sum_{t=1}^{n} X_t X_t^{\mathsf{T}}$$

represents the sample covariance matrix and $\widehat{\Omega}_{\mathrm{G}} := (\det \widehat{\Sigma})^{-1/d} \widehat{\Sigma}$ denotes the corresponding Gaussian shape matrix estimator for the complete-data case. In the incomplete-data case we will use the same symbol for the observed-data ML-estimator based on the normal distribution assumption presented by Little and Rubin (2002, Ch. 7.4) as well as Schafer (1997, Ch. 6.5). As already mentioned, we prefer to calculate the observed-data ML-estimator by the factored likelihood method since this leads to a fast and reliable algorithm. From now on the spectral estimator will be denoted by $\widehat{\Omega}_{\mathrm{sp}}$, whereas $\widehat{\Omega}$ represents an arbitrary shape matrix estimator. We distinguish between situations where the location vector μ is known and where it is unknown. In the complete-data case μ is estimated by the sample mean vector and in the incomplete-data case the corresponding estimate is provided by the factored likelihood method. These estimates are used both for calculating the Gaussian as well as the spectral estimate.

The *bias* of a shape matrix estimator is defined as the $d \times d$ matrix

$$\mathrm{B}\big(\widehat{\Omega}\big) := \mathrm{E}\big(\widehat{\Omega} - \Omega\big)\,.$$

The estimator $\widehat{\Omega}$ is called *unbiased* if each element of $\mathrm{B}\big(\widehat{\Omega}\big)$ corresponds to zero. However, the following tables contain only the number

$$\mathrm{b}(\widehat{\Omega}) := \frac{1}{d^2} \sum_{i,j=1}^{d} \big|\mathrm{B}\big(\widehat{\Omega}\big)_{ij}\big|$$

for the shape matrix estimators which have been taken into consideration.

Further, the *mean squared error* (MSE) of the corresponding estimator is given by the scalar

$$\mathrm{MSE}\big(\widehat{\Omega}\big) := \mathrm{E}\left[\frac{\mathrm{tr}\big\{(\widehat{\Omega} - \Omega)(\widehat{\Omega} - \Omega)^{\mathsf{T}}\big\}}{d^2}\right].$$

This can be interpreted as the average mean squared error of all components of $\widehat{\Omega}$.

The shape matrix estimators $\widehat{\Omega}_{\mathrm{G}}$ and T or $\widehat{\Omega}_{\mathrm{sp}}$ are compared by the *relative efficiencies*

$$\mathrm{re}_{\,T/\mathrm{G}} := \frac{\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})}{\mathrm{MSE}(T)} \qquad \text{and} \qquad \mathrm{re}_{\,\mathrm{sp}/\mathrm{G}} := \frac{\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})}{\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})}\,.$$

The simulation study is used to test against the null hypothesis $H_0\colon \mathrm{B}(\widehat{\Omega}) = 0$ (that means the considered shape matrix estimator is unbiased) by Hotelling's T^2 as well as

$$H_0\colon \mathrm{re}_{\,T/\mathrm{G}} \leq 1 \qquad \text{or} \qquad H_0\colon \mathrm{re}_{\,\mathrm{sp}/\mathrm{G}} \leq 1$$

(that is T or $\widehat{\Omega}_{\mathrm{sp}}$ are not more efficient than $\widehat{\Omega}_{\mathrm{G}}$) by applying the delta method.

8.5.1. Complete-Data Case

For investigating the large-sample properties of the shape matrix estimators we simulate 1000 samples containing 10000 independent copies of a 3-dimensional t-distributed random vector with location vector $\mu = 0$, dispersion matrix $\Sigma = I_3$, and $\nu = 2, 3, 5, 10, \infty$ degrees of freedom. Note that for $\nu = \infty$ the data follow a joint normal distribution. Moreover, we consider 3 additional scenarios where the normal distributions are contaminated at $(10, 10, 10) \in \mathbb{R}^3$. We add an amount of $\lfloor n\alpha/(1-\alpha) \rfloor$ contaminating data to the sample. The number α will be referred to as the *rate of contamination* and in the simulation study we consider $\alpha = 0.01, 0.05, 0.1$. That means there are 8 scenarios $t_2, t_3, t_5, t_{10}, t_\infty, c_{1\%}, c_{5\%}$ and $c_{10\%}$ for each situation where μ is either known or unknown.

Table 8.1.: Simulation study for the complete-data case where t_ν indicates a 3-dimensional t-distribution with ν degrees of freedom and t_∞ stands for a 3-dimensional normal distribution. Further, c_α indicates a contaminated 3-dimensional normal distribution where α is the rate of contamination. The standard errors are given in parentheses and bold numbers are significant on a 95%-level.

μ unknown	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_\mathrm{G})$	**.0642**	**.0037**	**.0006**	.0003	.0002	**.1409**	**.2061**	**.2287**
$\mathrm{b}(T)$	.0004	.0003	.0003	.0001	.0002	**.0033**	**.0082**	**.0142**
$\mathrm{MSE}(\widehat{\Omega}_\mathrm{G})$	.5867 (.1690)	.0091 (.0029)	.0003 (.0000)	.0001 (.0000)	.0001 (.0000)	.0351 (.0016)	.1085 (.0100)	.1646 (.0206)
$\mathrm{MSE}(T)$	.0002 (.0000)	.0002 (.0000)	.0002 (.0000)	.0002 (.0000)	.0002 (.0000)	.0001 (.0000)	.0008 (.0002)	.0055 (.0019)
$\mathrm{re}_{T/\mathrm{G}}$	**3075.6** (844.33)	**48.315** (15.422)	**1.6756** (.0700)	.8439 (.0174)	.6301 (.0115)	**488.04** (13.597)	**136.67** (19.655)	**29.8451** (7.6420)
μ known	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_\mathrm{G})$	**.0607**	**.0030**	.0006	.0003	.0004	**.1413**	**.2079**	**.2317**
$\mathrm{b}(T)$	.0003	**.0006**	.0003	.0004	.0003	**.0027**	**.0045**	**.0054**
$\mathrm{MSE}(\widehat{\Omega}_\mathrm{G})$	.7080 (.3054)	.0075 (.0015)	.0003 (.0000)	.0001 (.0000)	.0001 (.0000)	.0354 (.0016)	.1123 (.0106)	.1746 (.0229)
$\mathrm{MSE}(T)$	.0002 (.0000)	.0002 (.0000)	.0002 (.0000)	.0002 (.0000)	.0002 (.0000)	.0001 (.0000)	.0001 (.0000)	.0003 (.0001)
$\mathrm{re}_{T/\mathrm{G}}$	**3784.9** (1635.6)	**40.889** (8.1236)	**1.7053** (.0937)	.7975 (.0166)	.5952 (.0113)	**580.11** (16.709)	**928.82** (46.853)	**669.80** (65.748)

The results are given in Table 8.1. If the data are uncontaminated only the Gaussian estimator turns out to be significantly biased for small values of ν, whereas Tyler's M-estimator seems to be unbiased (except for the case where μ is known and the t-distribution has 3 degrees of freedom). In contrast, for contaminated data a significant bias is present in every scenario. However, due to its robustness property, Tyler's M-estimator is much more efficient than the Gaussian estimator if the data are contaminated but its relative efficiency generally decreases if the rate of contamination increases.

Tyler's M-estimator is more efficient than the Gaussian one as long as $\nu < 10$ or the data are contaminated. Note that estimating the unknown location vector has no essential impact on the large-sample properties of T and $\widehat{\Omega}_\mathrm{G}$ if the data are uncontaminated. In contrast, for contaminated data the relative efficiency of Tyler's M-estimator can increase substantially if μ is known. Further, the MSE of T does not depend on the parameter ν of the t-distribution since T is distribution-free if the data are uncontaminated. In that case

Table 8.2.: Simulation study for the incomplete-data case where the missing data are MCAR. The symbols and numbers can be interpreted as in Table 8.1.

μ unknown	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_{\mathrm{G}})$	**.0704**	**.0088**	.0006	.0003	.0006	**.6228**	**1.9553**	**3.3099**
$\mathrm{b}(\widehat{\Omega}_{\mathrm{sp}})$	.0006	.0008	.0004	.0005	.0009	**.0435**	**.4356**	**1.8112**
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})$	.5538 (.1760)	.0167 (.0046)	.0008 (.0000)	.0004 (.0000)	.0003 (.0000)	.4387 (.0003)	3.9139 (.0030)	11.067 (.0082)
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})$	.0006 (.0000)	.0005 (.0000)	.0005 (.0000)	.0005 (.0000)	.0005 (.0000)	.0035 (.0000)	.2189 (.0004)	3.3623 (.0046)
$\mathrm{re}_{\mathrm{sp/G}}$	**1006.8** (309.27)	**30.743** (8.4188)	**1.5640** (.0644)	.7552 (.0191)	.5518 (.0110)	**125.95** (1.4111)	**17.883** (.0313)	**3.2916** (.0036)

μ known	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_{\mathrm{G}})$	**.0717**	**.0056**	**.0010**	.0007	.0007	**.6291**	**2.0295**	**3.5667**
$\mathrm{b}(\widehat{\Omega}_{\mathrm{sp}})$	.0006	.0008	.0007	**.0011**	.0012	**.0348**	**.1956**	**.4735**
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})$	.7958 (.2877)	.0099 (.0015)	.0008 (.0000)	.0004 (.0000)	.0003 (.0000)	.4472 (.0004)	4.2120 (.0032)	12.8365 (.0091)
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})$	.0005 (.0000)	.0005 (.0000)	.0005 (.0000)	.0005 (.0000)	.0005 (.0000)	.0025 (.0000)	.0522 (.0002)	.2602 (.0005)
$\mathrm{re}_{\mathrm{sp/G}}$	**1523.2** (548.31)	**19.278** (2.8462)	**1.5158** (.0574)	.7379 (.0180)	.5476 (.0115)	**181.20** (2.5016)	**80.685** (.2573)	**49.3280** (.0897)

its asymptotic relative efficiency can be calculated analytically (**?**) by

$$\lim_{n\to\infty} \mathrm{re}_{T/G} = \frac{d}{d+2} \cdot \frac{\nu-2}{\nu-4}, \qquad \nu > 4\,. \tag{8.16}$$

This is fairly reflected by the relative efficiencies in t_5, t_{10}, and t_∞ which are presented in Table 8.1. Eq. 8.16 states that the relative efficiency of T depends on the number of dimensions and whenever $d > \nu - 4$, Tyler's M-estimator is more efficient than the Gaussian estimator. Thus if the number of dimensions is large enough, T should be preferred even if the data apparently follow a multivariate normal distribution.

8.5.2. Incomplete-Data Case

Again we simulate 1000 samples containing 10000 independent copies of a 3-dimensional t-distributed random vector using the same parameters as in Section 8.5.1 and we also add some contaminating data as described in the previous section. For the simulation study we consider the three different missingness mechanisms MCAR, MAR, and NMAR. Let $\mathbf{x}$ $(3 \times n)$ be a realized sample. Some of the data in the first row of $\mathbf{x}$ are missing. This is denoted by $m_t = 1$ if $\mathbf{x}_{1t}$ is missing and $m_t = 0$ if it is observed $(t = 1, \ldots, n)$. The missing

data are MCAR if the missingness pattern $M = (m_1, \ldots, m_n)$ is stochastically independent of the sample. In contrast, if the distribution of M depends only on the observed part of $\mathbf{x}$, the missing data are MAR and if the missingness is determined by the unobserved part of the sample, the missing data are NMAR.

For the MCAR case we simulate a $1 \times n$ vector Y which is stochastically independent of the sample. Each component of Y is Student t-distributed with ν degrees of freedom (denoted by $Y_t \sim t_\nu$) and $Y_1, \ldots, Y_n$ are mutually independent. The element $\mathbf{x}_{1t}$ is considered as missing whenever $y_t < t_\nu^{-1}(0.75)$. Further, for the MAR case $\mathbf{x}_{1t}$ is missing if $\mathbf{x}_{2t} < t_\nu^{-1}(0.75)$ and for the NMAR case this element is unobserved whenever $\mathbf{x}_{1t} < t_\nu^{-1}(0.75)$. So approximately 75% of the uncontaminated data in the first row of $\mathbf{x}$ are missing for each missingness mechanism. Table 8.2 contains the results of the MCAR case and, accordingly, Table 8.3 and Table 8.4 report the outcomes of the MAR and NMAR case.

In the MCAR case the spectral estimator is not significantly biased (except for the case where μ is known and the t-distribution has 10 degrees of freedom) provided the data are uncontaminated. Once again the Gaussian estimator is significantly biased if ν is small. For contaminated data the mean squared errors as well as the biases given in Table 8.1 and 8.2 are very different which indicates that the missing data effect has a substantial impact on the shape matrix estimators. Note that for uncontaminated data the MSE of the spectral estimator turns out to be constant for any $\nu > 0$ (cf. Section 8.3.2). Although the relative efficiencies of the spectral estimator are smaller in Table 8.2, in the MCAR case the overall picture is not substantially different from the complete-data case which is illustrated in Table 8.1.

If the missing data are MAR (see Table 8.3) the spectral estimator becomes biased in every scenario. This is not surprising since in Section 8.3.2 it has been already mentioned that the MCAR assumption is required when working with projected data. Also the Gaussian estimator turns out to be biased whenever $\nu < \infty$ (except for $\nu = 2$ where the hypothesis test probably has not enough power). Indeed, from missing-data analysis it is known that the latter should be asymptotically unbiased if the missing data are MAR but here it is presumed that the data follow a joint normal distribution. Otherwise the observed-data ML-estimator may become biased and in fact this is indicated throughout our simulations as long as the normal distribution assumption is not satisfied. However, if the missing data are purely MAR, Table 8.3 shows that the spectral estimator generally is dominated by the Gaussian one if the data are uncontaminated but it dominates the Gaussian one

Table 8.3.: Simulation study for the incomplete-data case where the missing data are MAR. The symbols and numbers can be interpreted as in Table 8.1.

μ unknown	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_{\mathrm{G}})$	.0809	**.0585**	**.0311**	**.0142**	.0010	**.7780**	**2.2971**	**3.8509**
$\mathrm{b}(\widehat{\Omega}_{\mathrm{sp}})$	**.0605**	**.0862**	**.0945**	**.0985**	**.1025**	**.3214**	**.5996**	**2.1882**
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})$	1.2978 (.9681)	.0311 (.0062)	.0049 (.0002)	.0013 (.0000)	.0006 (.0000)	.7060 (.0007)	5.4421 (.0044)	15.069 (.0114)
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})$	.0788 (.0098)	.0379 (.0060)	.0292 (.0001)	.0308 (.0001)	.0328 (.0001)	.2191 (.0006)	.4388 (.0009)	4.9420 (.0090)
$\mathrm{re}_{\mathrm{sp/G}}$	16.471 (12.4244)	.8193 (.0884)	.1673 (.0061)	.0437 (.0009)	.0175 (.0005)	**3.2218** (.0063)	**12.401** (.0205)	**3.0492** (.0047)

μ known	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_{\mathrm{G}})$	.1301	**.0564**	**.0310**	**.0145**	.0004	**.5977**	**1.9569**	**3.4417**
$\mathrm{b}(\widehat{\Omega}_{\mathrm{sp}})$	**.0847**	**.0903**	**.0949**	**.0983**	**.1019**	**.1335**	**.2884**	**.5976**
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})$	6.7343 (5.8811)	.0189 (.0008)	.0039 (.0001)	.0010 (.0000)	.0002 (.0000)	.4086 (.0004)	3.9213 (.0028)	11.959 (.0085)
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})$	.0227 (.0001)	.0257 (.0001)	.0283 (.0001)	.0301 (.0001)	.0323 (.0001)	.0314 (.0001)	.1017 (.0003)	.4206 (.0009)
$\mathrm{re}_{\mathrm{sp/G}}$	296.72 (259.12)	.7374 (.0333)	.1391 (.0026)	.0338 (.0007)	.0076 (.0002)	**13.016** (.0383)	**38.555** (.0950)	**28.434** (.0518)

if they are contaminated. Hence, the Gaussian estimator generally is to be preferred if the data are uncontaminated and the missing part is MAR but not MCAR. In the NMAR case (see Table 8.4) the Gaussian estimator is biased even if the data are jointly normally distributed. Like in the MAR case, the Gaussian estimator should be preferred whenever the data are uncontaminated (except for very heavy tails) but the spectral estimator is more efficient if they are contaminated.

8.6. Conclusion

We presented a distribution-free approach for estimating the shape matrix of incomplete multivariate data leading to the spectral estimator. This is particularly appropriate if the data stem from the class of generalized elliptical distributions including both the class of elliptically symmetric and skew-elliptical distributions. We showed that in the complete-data case the spectral estimator corresponds to Tyler's M-estimator whereas in the incomplete-data case it can be represented as an observed-data ML-estimator. In both cases the estimator is invariant under arbitrary changes of the generating variate of the generalized

Table 8.4.: Simulation study for the incomplete-data case where the missing data are NMAR. The symbols and numbers can be interpreted as in Table 8.1.

μ unknown	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_{\mathrm{G}})$	**.0635**	**.0476**	**.1139**	**.1614**	**.2029**	**.7969**	**2.4495**	**4.1247**
$\mathrm{b}(\widehat{\Omega}_{\mathrm{sp}})$	**.0056**	**.0922**	**.1456**	**.1781**	**.2063**	**.2184**	**.6145**	**2.3627**
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})$	.6388 (.3768)	.1617 (.1395)	.0404 (.0003)	.0789 (.0002)	.1236 (.0002)	.6890 (.0006)	6.1380 (.0047)	17.268 (.0131)
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})$	.0015 (.0001)	.0269 (.0002)	.0645 (.0002)	.0956 (.0003)	.1279 (.0003)	.0901 (.0002)	.4203 (.0008)	5.7230 (.0078)
$\mathrm{re}_{\mathrm{sp/G}}$	**412.69** (221.09)	6.0198 (5.1941)	.6260 (.0035)	.8255 (.0021)	.9663 (.0016)	**7.6431** (.0169)	**14.602** (.0222)	**3.0173** (.0033)

μ known	t_2	t_3	t_5	t_{10}	t_∞	$c_{1\%}$	$c_{5\%}$	$c_{10\%}$
$\mathrm{b}(\widehat{\Omega}_{\mathrm{G}})$	.0827	**.1051**	**.1027**	**.1008**	**.0985**	**.5399**	**1.7264**	**3.0425**
$\mathrm{b}(\widehat{\Omega}_{\mathrm{sp}})$	**.1617**	**.1718**	**.1799**	**.1860**	**.1924**	**.2147**	**.2910**	**.3997**
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{G}})$	1.0266 (.4442)	.0504 (.0013)	.0409 (.0002)	.0390 (.0001)	.0371 (.0001)	.3635 (.0002)	3.0745 (.0019)	9.3694 (.0056)
$\mathrm{MSE}(\widehat{\Omega}_{\mathrm{sp}})$	.1080 (.0003)	.1231 (.0003)	.1367 (.0003)	.1475 (.0004)	.1576 (.0004)	.1429 (.0003)	.1251 (.0002)	.2149 (.0004)
$\mathrm{re}_{\mathrm{sp/G}}$	**9.5069** (4.1142)	.4094 (.0109)	.2992 (.0014)	.2641 (.0008)	.2351 (.0005)	**2.5436** (.0060)	**24.567** (.0433)	**43.609** (.0686)

elliptical distribution. That means the underlying mechanism which is responsible for outliers or clusters can be eliminated and our estimator becomes completely robust. We also derived its asymptotic distribution under the MCAR assumption. An important argument in favor of the spectral estimator is that if the data stem from a generalized elliptical distribution, no nuisance parameters need to be estimated for assessing its asymptotic distribution and its asymptotic covariance matrix solely follows from the spectral estimate itself. Moreover, we developed a fast algorithm for calculating the spectral estimate and gave some practical advice for its numerical implementation. A simulation study for the complete-data and the incomplete-data case reveals that for contaminated data the spectral estimator should be always preferred. The same holds if the data are uncontaminated but heavy-tailed and the missing part of the sample is MCAR. In contrast, if the uncontaminated data possess a moderate tail index and the number of dimensions is small, the Gaussian approach seems to serve its purpose.

Summary

In Chapter 1 I discussed the potential drawbacks of TDC estimation. It turned out that the (semi-)parametric TDC estimators perform well if the underlying distribution or copula is the right one. By contrast, their performance is very poor if the assumed model is wrong. Hence, model risk is an inherent problem of TDC estimation. Further, the nonparametric estimators exhibit a large bias in case of tail-independence.

In Chapter 2 I presented an alternative measure for the extremal dependence of financial data, namely a conditional version of Spearman's rho. This is based on a purely nonparametric approach and so it is possible to avoid any kind of model misspecification. An empirical study using daily returns of stocks contained in the DAX 30 revealed that there is sufficient evidence to support the hypothesis of different degrees of monotone dependence in bull and bear markets.

A numerical approach to incorporate the stylized facts of high-frequency financial data and arbitrary prior information into portfolio optimization has been developed in Chapter 3. It is characterized by rather weak assumptions about the underlying stochastic process. I gave a practical example to demonstrate its applicability to real-world problems. The resulting portfolios became well-diversified compared to the outcomes of traditional portfolio optimization strategies.

Exact hypothesis tests for global and local minimum variance portfolios as well as their small-sample distributions have been derived in Chapter 4. It has been shown that estimation risk can be simply reduced by imposing linear constraints on the portfolio weights. All the presented results hold in small samples, which is an important fact since large-sample approximations fail if the sample size is large but the number of observations relative to the number of assets is small.

In Chapter 5 I presented two shrinkage estimators for the GMVP that dominate the traditional estimator. Their small-sample and their large-sample properties alike have been

investigated. The estimators considerably reduce the out-of-sample variance of the portfolio return compared to the traditional estimator, especially if the asset universe is large. In addition, I provided a hypothesis test to decide whether one should invest in a portfolio based on estimators for the GMVP or in the naive portfolio.

I presented a hypothesis test for the best investment strategy in Chapter 6 and demonstrated the test by an application to financial data. For this purpose I generalized the Jobson-Korkie-Memmel test considering ergodic stationary stochastic processes satisfying Gordin's condition. It turned out that ignoring the stylized facts of empirical finance can lead to wrong decisions.

In Chapter 7 I derived the joint asymptotic distributions of robust estimators for shape matrices and their associated scales. I also generalized an important result from local asymptotic normality theory. The given instruments are applicable to a wide range of problems in multivariate analysis such as principal components analysis, canonical correlation analysis, linear discriminant analysis, and multivariate regression.

Finally, in Chapter 8 I presented a distribution-free approach for estimating the shape matrix if the data are incomplete. I showed that in the complete-data case the resulting estimator corresponds to Tyler's M-estimator, whereas in the incomplete-data case it is an ML-estimator. In both cases the estimator is invariant under arbitrary changes of the generating variate of a generalized elliptical distribution. I also derived its asymptotic distribution and developed a fast algorithm for calculating the desired estimate.

Bibliography

B. Abdous, K. Ghoudi, and A. Khoudraji (1999), 'Non-parametric estimation of the limit dependence function of multivariate extremes', *Extremes* **2**, pp. 243–265.

J.G. Adrover (1998), 'Minimax bias-robust estimation of the dispersion matrix of a multivariate distribution', *Annals of Statistics* **26**, pp. 2301–2320.

D.W.K. Andrews (1991), 'Heteroskedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica* **59**, pp. 817–858.

T. Ané and C. Kharoubi (2003), 'Dependence structure and risk measure', *Journal of Business* **76**, pp. 411–438.

A. Ang and J. Chen (2002), 'Asymmetric correlations of equity portfolios', *Journal of Financial Economics* **63**, pp. 443–494.

B.C. Arnold and R.J. Beaver (2004), 'Elliptical Models Subject to Hidden Truncation or Selective Sampling', in: M.G. Genton, ed., 'Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality', chapter 6, Chapman & Hall.

A. Azzalini and A. Dalla Valle (1996), 'The multivariate skew-normal distribution', *Biometrika* **83**, pp. 715–726.

M.L. Bachelier (1900), *Théorie de la spéculation*, Annales scientifiques de l'É.N.S., 3^e série, tome 17, p. 21–86, Gauthier-Villars.

A. Bade, G. Frahm, and U. Jaekel (2008), 'A general approach to Bayesian portfolio optimization', *Mathematical Methods of Operations Research* pp. DOI: 10.1007/s00186–008–0271–4.

Z.D. Bai (1999), 'Methodologies in spectral analysis of large dimensional random matrices, a review', *Statistica Sinica* **9**, pp. 611–677.

O.E. Barndorff-Nielsen, E. Nicolato, and N. Shepard (2002), 'Some recent developments in stochastic volatility modelling', *Quantitative Finance* **2**, pp. 11–23.

L. Bauwens, S. Laurent, and J.V.K. Rombouts (2006), 'Multivariate GARCH models: a survey', *Journal of Applied Econometrics* **21**, pp. 79–109.

A.E. Beaton (1964), 'The use of special matrix operations in statistical calculus', Research bulletin RB-64-51, Educational Testing Service, Princeton, NJ.

J. Berger (2006), 'The case for objective Bayesian analysis', *Bayesian Analysis* **1**, pp. 385–402.

M. Bilodeau and D. Brenner (1999), *Theory of Multivariate Statistics*, Springer.

N.H. Bingham, C.M. Goldie, and J.L. Teugels (1987), 'Regular variation', in: 'Encyclopedia of Mathematics and its Applications', volume 27, Cambridge University Press.

N.H. Bingham, R. Kiesel, and R. Schmidt (2003), 'Semi-parametric models in finance: econometric applications', *Quantitative Finance* **3**, pp. 426–441.

F. Black and R. Litterman (1992), 'Global portfolio optimization', *Financial Analysts Journal* **48**, pp. 28–43.

T. Bollerslev (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**, pp. 307–327.

M.D. Branco and D.K. Dey (2001), 'A general class of multivariate skew-elliptical distributions', *Journal of Multivariate Analysis* **79**, pp. 99–113.

M. Britten-Jones (1999), 'The sampling error in estimates of mean-variance efficient portfolio weights', *Journal of Finance* **54**, pp. 655–671.

P.J. Brockwell and R.A. Davis (1991), *Time Series: Theory and Methods*, Springer, second edition.

S. Cambanis, S. Huang, and G. Simons (1981), 'On the theory of elliptically contoured distributions', *Journal of Multivariate Analysis* **11**, pp. 368–385.

J.Y. Campbell, A.W. Lo, and A.C. MacKinlay (1997), *The Econometrics of Financial Markets*, Princeton University Press.

P. Capéraà and A.-L. Fougères (2000), 'Estimation of a bivariate extreme value distribution', *Extremes* **3**, pp. 311–329.

P. Capéraà, A.-L. Fougères, and C. Genest (1997), 'A nonparametric estimation procedure for bivariate extreme value copulas', *Biometrika* **84**, pp. 567–577.

G. Casella and R.L. Berger (2002), *Statistical Inference*, Duxbury Press, second edition.

L.K.C. Chan, J. Karceski, and J. Lakonishok (1999), 'On portfolio optimization: Forecasting covariances and choosing the risk model', *Review of Financial Studies* **12**, pp. 937–974.

U. Cherubini, E. Luciano, and W. Vecchiato (2004), *Copula Methods in Finance*, John Wiley.

V.K. Chopra and W.T. Ziemba (1993), 'The effect of errors in means, variances, and covariances on optimal portfolio choice', *Journal of Portfolio Management* **19**, pp. 6–11.

S.G. Coles, J.E. Heffernan, and J.A. Tawn (1999), 'Dependence measures for extreme value analyses', *Extremes* **2**, pp. 339–365.

S.G. Coles and J.A. Tawn (1991), 'Modelling multivariate extreme events', *Journal of the Royal Statistical Society, Series B* **53**, pp. 377–392.

C. Croux and G. Haesbroeck (1999), 'Influence function and efficiency of the minimum covariance determinant scatter matrix estimator', *Journal of Multivariate Analysis* **71**, pp. 161–190.

J. Davidson (1994), *Stochastic Limit Theory*, Oxford University Press.

P.L. Davies (1987), 'Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices', *Annals of Statistics* **15**, pp. 1269–1292.

P.L. Davies (1992), 'The asymptotics of Rousseeuw's minimum volume ellipsoid estimator', *Annals of Statistics* **20**, pp. 1828–1843.

P. Deheuvels (1991), 'On the limiting behaviour of the Pickands estimator for bivariate extreme value distributions', *Statistics and Probability Letters* **12**, pp. 429–439.

V. DeMiguel, L. Garlappi, and R. Uppal (2007), 'Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy?', *Review of Financial Studies*, URL: http://rfs.oxfordjournals.org/cgi/content/abstract/hhm075v1.

L. Dümbgen (1998), 'On Tyler's M-functional of scatter in high dimension', *Annals of the Institute of Statistical Mathematics* **50**, pp. 471–491.

L. Dümbgen and D.E. Tyler (2005), 'On the breakdown properties of some multivariate M-functionals', *Scandinavian Journal of Statistics* **32**, pp. 247–264.

J. Dobrić, G. Frahm, and F. Schmid (2008), 'Dependence of stock returns in bull and bear markets', Discussion paper, University of Cologne, Department of Economic and Social Statistics, Germany.

J. Dobrić and F. Schmid (2005a), 'Nonparametric estimation of the lower tail dependence λ_L in bivariate copulas', *Journal of Applied Statistics* **32**, pp. 387–407.

J. Dobrić and F. Schmid (2005b), 'Testing goodness of fit for parametric families of copulas – application to financial data', *Communications in Statistics: Simulation and Computation* **34**, pp. 1053–1068.

G. Draisma, H. Drees, A. Ferreira, et al. (2004), 'Bivariate tail estimation: dependence in asymptotic independence', *Bernoulli* **10**, pp. 251–280.

H. Drees and E. Kaufmann (1998), 'Selecting the optimal sample fraction in univariate extreme value estimation', *Stochastic Processes and Their Applications* **75**, pp. 149–172.

G.K. Eagleson (1975), 'On Gordin's Central limit theorem for stationary processes', *Journal of Applied Probability* **12**, pp. 176–179.

D. Eichhorn, F. Gupta, and E. Stubbs (1998), 'Using constraints to improve the robustness of asset allocation', *Journal of Portfolio Management* **24**, pp. 41–48.

J. Einmahl, L. de Haan, and X. Huang (1993), 'Estimating a multidimensional extreme value distribution', *Journal of Multivariate Analysis* **47**, pp. 35–47.

J. Einmahl, L. de Haan, and A.K. Sinha (1997), 'Estimating the spectral measure of an extreme value distribution', *Stochastic Processes and Their Applications* **70**, pp. 143–171.

E.J. Elton and M. Gruber (1973), 'Estimating the dependence structure of share prices - Implications for portfolio selection', *Journal of Finance* **28**, pp. 1203–1232.

P. Embrechts, C. Klüppelberg, and T. Mikosch (1997), *Modelling Extremal Events (for Insurance and Finance)*, Springer.

P. Embrechts, F. Lindskog, and A. McNeil (2003), 'Modelling dependence with copulas and applications to risk management', in: S. Rachev, ed., 'Handbook of Heavy Tailed Distributions in Finance', pp. 329–384, Elsevier.

P. Embrechts, A.J. McNeil, and D. Straumann (2002), 'Correlation and dependence in risk management: properties and pitfalls', in: M. Dempster, ed., 'Risk Management: Value at Risk and Beyond', Cambridge University Press.

R.F. Engle (1982), 'Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation', *Econometrica* **50**, pp. 987–1007.

C.B. Erb, C.R. Harvey, and T.E. Viskanta (1994), 'Forecasting international equity correlations', *Financial Analysts Journal* **50**, pp. 32–45.

M. Falk and R.D. Reiss (2003), 'Efficient estimation of the canonical dependence function', *Extremes* **6**, pp. 61–82.

KT. Fang, S. Kotz, and KW. Ng (1990), *Symmetric Multivariate and Related Distributions*, Chapman & Hall.

J.-D. Fermanian (2005), 'Goodness-of-fit tests for copulas', *Journal of Multivariate Analysis* **95**, pp. 119–152.

R.A. Fisher and L.H.C. Tippett (1928), 'Limiting forms of the frequency distribution of the largest or smallest member of a sample', *Proceedings of the Cambridge Philosophical Society* **24**, pp. 180–190.

I. Fortin and C. Kuzmics (2002), 'Tail-dependence in stock return pairs', *International Journal of Intelligent Systems in Accounting, Finance & Management* **11**, pp. 89–107.

G. Frahm (2004), *Generalized Elliptical Distributions: Theory and Applications*, Ph.D. thesis, University of Cologne, Department of Economic and Social Statistics, Germany.

G. Frahm (2006), 'On the extremal dependence coefficient of multivariate distributions', *Statistics and Probability Letters* **76**, pp. 1470–1481.

G. Frahm (2007), 'Testing for the best alternative with an application to performance measurement', Discussion paper, University of Cologne, Department of Economic and Social Statistics, Germany.

G. Frahm (2008), 'Linear statistical inference for global and local minimum variance portfolios', *Statistical Papers* pp. DOI: 10.1007/s00362–008–0170–z.

G. Frahm and U. Jaekel (2007a), 'Distribution-free shape matrix estimation for incomplete data', Discussion paper, University of Cologne, Department of Economic and Social Statistics, Germany.

G. Frahm and U. Jaekel (2007b), 'Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance', Discussion paper, University of Cologne, Department of Economic and Social Statistics, Germany.

G. Frahm, M. Junker, and R. Schmidt (2005), 'Estimating the tail-dependence coefficient: properties and pitfalls', *Insurance: Mathematics and Economics* **37**, pp. 80–100.

G. Frahm, M. Junker, and A. Szimayer (2003), 'Elliptical copulas: applicability and limitations', *Statistics and Probability Letters* **63**, pp. 275–286.

G. Frahm and C. Memmel (2008), 'Dominating estimators for the global minimum variance portfolio', Discussion paper, University of Cologne, Department of Economic and Social Statistics, Germany.

M.J. Frank (1979), 'On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$', *Aequationes Mathematicae* **19**, pp. 194–226.

P.A. Frost and J.E. Savarino (1986), 'An empirical Bayes approach to efficient portfolio selection', *Journal of Financial and Quantitative Analysis* **21**, pp. 293–305.

P.A. Frost and J.E. Savarino (1988), 'For better performance: constrain portfolio weights', *Journal of Portfolio Management* **14**, pp. 29–34.

D. Gamerman and H.F. Lopes (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall.

L. Garlappi, R. Uppal, and T. Wang (2007), 'Portfolio selection with parameter and model uncertainty: a multi-prior approach', *Review of Financial Studies* **20**, pp. 41–81.

C. Genest, K. Ghoudi, and L.-P. Rivest (1995), 'A semiparametric estimation procedure of dependence parameters in multivariate families of distributions', *Biometrika* **82**, pp. 543–552.

C. Genest, K. Ghoudi, and L.-P. Rivest (1998), 'Comment on the article of E.W. Frees and E.A. Valdez entitled 'Understanding relationships using copulas.", *North American Actuarial Journal* **2**, pp. 143–149.

C. Genest and R.J. MacKay (1986), 'Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données', *The Canadian Journal of Statistics* **14**, pp. 145–159.

C. Genest, J.-F. Quessy, and B. Rémillard (2006), 'Goodness-of-fit procedures for copula models based on the probability integral transformation', *Scandinavian Journal of Statistics* **33**, pp. 337–366.

C. Genest and L.-P. Rivest (1989), 'A characterization of Gumbel's family of extreme value distributions', *Statistics and Probability Letters* **8**, pp. 207–211.

M.G. Genton, ed. (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman & Hall.

J. Geweke (1986), 'Exact inference in the inequality constraint normal linear regression model', *Journal of Applied Econometrics* **1**, pp. 127–141.

J. Geweke (1989), 'Bayesian inference in econometric models using Monte Carlo integration', *Econometrica* **57**, pp. 1317–1339.

J. Geweke (1995), 'Bayesian comparison of econometric models', Working paper 532, Federal Reserve Bank of Minneapolis, Minnesota.

V. Golosnoy and Y. Okhrin (2007), 'Multivariate shrinkage for optimal portfolio weights', *The European Journal of Finance* **13**, pp. 441–458.

J.H. Goodnight (1979), 'A tutorial on the SWEEP operator', *American Statistician* **33**, pp. 149–158.

C. Gouriéroux, A. Holly, and A. Monfort (1982), 'Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters', *Econometrica* **50**, pp. 63–80.

R. Grauer and F. Shen (2000), 'Do constraints improve portfolio performance?', *Journal of Banking and Finance* **24**, pp. 1253–1274.

R. Green and B. Hollifield (1992), 'When will mean-variance efficient portfolios be well diversified?', *Journal of Finance* **47**, pp. 1785–1809.

W.H. Greene (2003), *Econometric Analysis*, Prentice Hall.

L. de Haan and S.I. Resnick (1993), 'Estimating the limit distribution of multivariate extremes', *Stochastic Models* **9**, pp. 275–309.

M. Hallin, H. Oja, and D. Paindaveine (2006), 'Semiparametrically efficient rank-based inference for shape. II. Optimal R-estimation of shape', *Annals of Statistics* **34**, pp. 2757–2789.

M. Hallin and D. Paindaveine (2006a), 'Parametric and semiparametric inference for shape: the role of the scale functional', *Statistics and Decisions* **24**, pp. 327–350.

M. Hallin and D. Paindaveine (2006b), 'Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity', *Annals of Statistics* **34**, pp. 2707–2756.

M. Hallin and D. Paindaveine (2008), 'Optimal rank-based tests for homogeneity of scatter', *Annals of Statistics* **36**, pp. 1261–1298.

M. Hallin and D. Paindaveine (2009), 'Optimal tests for homogeneity of covariance, scale, and shape', *Journal of Multivariate Analysis* **100**, pp. 422–444.

J.D. Hamilton (1994), *Time Series Analysis*, Princeton University Press.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, et al. (1986), *Robust Statistics*, John Wiley.

W.K. Hastings (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, pp. 97–109.

F. Hayashi (2000), *Econometrics*, Princeton University Press.

U. Herold and R. Maurer (2006), 'Portfolio choice and estimation risk – a comparison of Bayesian to heuristic approaches', *ASTIN Bulletin* **36**, pp. 135–160.

T.P. Hettmansperger and R.H. Randles (2002), 'A practical affine equivariant multivariate median', *Biometrika* **89**, pp. 851–860.

P.J. Huber (2003), *Robust Statistics*, John Wiley.

H. Hult and F. Lindskog (2002), 'Multivariate extremes, aggregation and dependence in elliptical distributions', *Advances in Applied Probability* **34**, pp. 587–608.

T.P. Hutchinson and C.D. Lai (1990), *Continuous Bivariate Distributions*, Rumsby Scientific Publishing.

E. Jacquier, N.G. Polson, and P.E. Rossi (1994), 'Bayesian analyis of stochastic volatility models', *Journal of Business and Economic Statistics* **12**, pp. 371–389.

E. Jacquier, N.G. Polson, and P.E. Rossi (2004), 'Bayesian analysis of stochastic volatility models withfat-tails and correlated errors', *Journal of Econometrics* **122**, pp. 185–212.

R. Jagannathan and T. Ma (2003), 'Risk reduction in large portfolios: Why imposing the wrong constraints helps', *Journal of Finance* **58**, pp. 1651–1683.

J.D. Jobson and B. Korkie (1979), 'Improved estimation for Markowitz portfolios using James-Stein type estimators', in: 'Proceedings of the American Statistical Association (Business and Economic Statistics)', volume 1, pp. 279–284.

J.D. Jobson and B. Korkie (1981), 'Performance hypothesis testing with the Sharpe and Treynor measures', *Journal of Finance* **36**, pp. 889–908.

H. Joe (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall.

H. Joe, R.L. Smith, and I. Weissman (1992), 'Bivariate threshold models for extremes', *Journal of the Royal Statistical Society, Series B* **54**, pp. 171–183.

P. Jorion (1985), 'International portfolio diversification with estimation risk', *Journal of Business* **58**, pp. 259–278.

P. Jorion (1986), 'Bayes-Stein estimation for portfolio analysis', *Journal of Financial and Quantitative Analysis* **21**, pp. 279–292.

G.G. Judge and M.E. Bock (1978), *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*, North-Holland Publishing Company.

M. Junker (2004), *Modelling, Estimating and Validating Multidimensional Distribution Functions: with Applications to Risk Management*, Ph.D. thesis, Technical University of Kaiserslautern, Germany.

M. Junker and A. May (2005), 'Measurement of aggregate risk with copulas', *Econometrics Journal* **8**, pp. 428–454.

A. Juri and M. Wüthrich (2002), 'Copula convergence theorems for tail events', *Insurance: Mathematics and Economics* **30**, pp. 405–420.

B.A. Kalymon (1971), 'Estimation risk in the portfolio selection model', *Journal of Financial and Quantitative Analysis* **6**, pp. 559–582.

R. Kan and G. Zhou (2007), 'Optimal portfolio choice with parameter uncertainty', *Journal of Financial and Quantitative Analysis* **42**, pp. 621–656.

A. Kempf and C. Memmel (2002), 'Schätzrisiken in der Portfoliotheorie', in: J.M. Kleeberg and H. Rehkugler, eds., 'Handbuch Portfoliomanagement', pp. 893–919, Uhlenbruch.

A. Kempf and C. Memmel (2006), 'Estimating the global minimum variance portfolio', *Schmalenbach Business Review* **58**, pp. 332–348.

M.G. Kendall and R.M. Sundrum (1953), 'Distribution-free methods and order properties', *Review of the International Statistical Institute* **21**, pp. 124–134.

J.T. Kent and D.E. Tyler (1988), 'Maximum likelihood estimation for the wrapped Cauchy distribution', *Journal of Applied Statistics* **15**, pp. 247–254.

J.T. Kent and D.E. Tyler (1991), 'Redescending M-estimates of multivariate location and scatter', *Annals of Statistics* **19**, pp. 2102–2119.

M.G. Kenward and G. Molenberghs (1998), 'Likelihood based frequentist inference when data are missing at random', *Statistical Science* **13**, pp. 236–247.

R.W. Klein and V.S. Bawa (1976), 'The effect of estimation risk on optimal portfolio choice', *Journal of Financial Economics* **3**, pp. 215–231.

H.R. Künsch (1989), 'The jackknife and the bootstrap for general stationary observations', *Annals of Statistics* **17**, pp. 1217–1241.

J.-P. Laurent and J. Gregory (2005), 'Basket default swaps, CDOs and factor copulas', *Journal of Risk* **7**.

A.W. Ledford and J.A. Tawn (1996), 'Statistics for near independence in multivariate extreme values', *Biometrika* **83**, pp. 169–187.

O. Ledoit and M. Wolf (2001), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of Multivariate Analysis* **88**, pp. 365–411.

O. Ledoit and M. Wolf (2003), 'Improved estimation of the covariance matrix of stock returns with an application to portfolio selection', *Journal of Empirical Finance* **10**, pp. 603–621.

O. Ledoit and M. Wolf (2008), 'Robust performance hypothesis testing with the Sharpe ratio', Working paper 320, University of Zurich, Institute for Empirical Research in Economics.

D.X. Li (1999), 'The valuation of basket credit derivatives', Technical report, CreditMetrics Monitor, JPMorgan.

F. Lindskog, A.J. McNeil, and U. Schmock (2003), 'Kendall's tau for elliptical distributions', in: G. Bol, G. Nakhaeizadeh, S.T. Rachev, et al., eds., 'Credit Risk - Measurement, Evaluation and Management', Physica.

R.J. Little (1988), 'Robust estimation of the mean and covariance matrix from data with missing values', *Applied Statistics* **37**, pp. 23–38.

R.J. Little and D.B. Rubin (2002), *Statistical Analysis with Missing Data*, John Wiley, second edition.

J. Liu and D.K. Dey (2004), 'Skew-Elliptical Distributions', in: M.G. Genton, ed., 'Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality', chapter 3, Chapman & Hall.

H.P. Lopuhaä (1989), 'On the relation between S-estimators and M-estimators of multivariate location and covariance', *Annals of Statistics* **17**, pp. 1662–1683.

A. Manzotti, F.J. Perez, and A.J. Quiroz (2002), 'A statistic for testing the null hypothesis of elliptical symmetry', *Journal of Multivariate Analysis* **81**, pp. 274–285.

K.V. Mardia and P.E. Jupp (2000), *Directional Statistics*, John Wiley.

H.M. Markowitz (1952), 'Portfolio selection', *Journal of Finance* **7**, pp. 77–91.

R. Maronna, D. Martin, and V. Yohai (2006), *Robust Statistics*, John Wiley.

R. Maronna and V. Yohai (1990), 'The maximum bias of robust covariances', *Communications in Statistics: Theory and Methods* **19**, pp. 3925–3933.

R.A. Maronna (1976), 'Robust M-estimators of multivariate location and scatter', *Annals of Statistics* **4**, pp. 51–67.

G. Matthys and J. Beirlant (2002), 'Adaptive threshold selection in tail index estimation', in: P. Embrechts, ed., 'Extremes and Integrated Risk Management', pp. 37–49, Risk Books.

A.J. McNeil, R. Frey, and P. Embrechts (2005), *Quantitative Risk Management*, Princeton University Press.

C. Memmel (2003), 'Performance hypothesis testing with the Sharpe ratio', *Finance Letters* **1**, pp. 21–23.

C. Memmel (2004), *Schätzrisiken in der Portfoliotheorie*, Ph.D. thesis, University of Cologne, Department of Economic and Social Statistics, Germany.

R.C. Merton (1980), 'On estimating the expected return on the market: An exploratory investigation', *Journal of Financial Economics* **8**, pp. 323–361.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, et al. (1953), 'Equation of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, pp. 1087–1092.

A. Meucci (2005), *Risk and Asset Allocation*, Springer.

R.O. Michaud (1989), 'The Markowitz optimization enigma: Is 'optimized' optimal?', *Financial Analysts Journal* **45**, pp. 31–42.

T. Mikosch (2003), 'Modeling dependence and tails of financial time series', in: B. Finkenstaedt and H. Rootzén, eds., 'Extreme Values in Finance, Telecommunications, and the Environment', Chapman & Hall.

R.J. Muirhead (1982), *Aspects of Multivariate Statistical Theory*, John Wiley.

R.B. Nelsen (2006), *An Introduction to Copulas*, Springer, second edition.

H. Oja (2003), 'Multivariate M-estimates of location and shape', in: R. Höglund, M. Jäntti, and G. Rosenqvist, eds., 'Statistics, Econometrics and Society. Essays in Honor of Leif Nordberg.', Statistics Finland.

Y. Okhrin and W. Schmid (2006), 'Distributional properties of portfolio weights', *Journal of Econometrics* **134**, pp. 235–256.

J.T. de Oliveira (1984), 'Bivariate models for extremes', in: J.T. de Oliveira, ed., 'Statistical Extremes and Applications', pp. 131–153, Reidel, Dordrecht.

D. Paindaveine (2008), 'A canonical definition of shape', *Statistics and Probability Letters* **78**, pp. 2240–2247.

A.J. Patton (2004), 'On the out-of-sample importance of skewness and asymmetric dependence for asset allocation', *Journal of Financial Econometrics* **2**, pp. 130–168.

J. Pickands (1981), 'Multivariate extreme value distributions', in: 'Proceedings of the 43rd session of the International Statistical Insitute', pp. 859–878.

V. Plerou, P. Gopikrishnan, B. Rosenow, et al. (1999), 'Universal and nonuniversal properties of cross correlations in financial time series', *Physical Review Letters* **83**, pp. 1471–1474.

D.N. Politis (2003), 'The impact of bootstrap methods on time series analysis', *Statistical Science* **18**, pp. 219–230.

D.N. Politis, J.P. Romano, and M. Wolf (2001), 'On the asymptotic theory of subsampling', *Statistica Sinica* **11**, pp. 1105–1124.

N.G. Polson and B.V. Tew (2000), 'Bayesian portfolio selection: An empirical analysis of the S&P 500 index 1970-1996', *Journal of Business and Economic Statistics* **18**, pp. 164–173.

S.J. Press (2005), *Applied Multivariate Analysis*, Dover Publications, second edition.

R.H. Randles (2000), 'A simpler, affine-invariant, multivariate, distribution-free sign test', *Journal of the American Statistical Association* **95**, pp. 1263–1268.

C.R. Rao (1965), *Linear Statistical Inference and Its Applications*, John Wiley.

P. Rousseeuw (1985), 'Multivariate estimation with high breakdown point', in: W. Grossmann, G. Pflug, I. Vincze, et al., eds., 'Mathematical Statistics and Applications', pp. 283–297, Reidel, Dordrecht.

M. Salibian-Barrera, S. van Aelst, and G. Willems (2006), 'Principal components analysis based on multivariate MM-estimators with fast and robust bootstrap', *Journal of the American Statistical Association* **101**, pp. 1198–1211.

J.L. Schafer (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall.

J.L. Schafer and J.W. Graham (2002), 'Missing data: our view of the state of the art', *Psychological Methods* **7**, pp. 147–177.

B. Scherer (2004), 'Resampled efficiency and portfolio choice', *Financial Markets and Portfolio Management* **18**, pp. 382–398.

B. Scherer and R.D. Martin (2007), *Introduction to Modern Portfolio Optimization with NuOPT, S-PLUS and S+Bayes*, Springer, second edition.

F. Schmid and R. Schmidt (2006), 'Multivariate extensions of Spearman's rho and related statistics', *Statistics and Probability Letters* **77**, pp. 407–416.

F. Schmid and R. Schmidt (2007), 'Statistical inference for Sharpe's ratio', Preprint, University of Cologne, Department of Economic and Social Statistics, Germany.

R. Schmidt (2002), 'Tail dependence for elliptically contoured distributions', *Mathematical Methods of Operations Research* **55**, pp. 301–327.

R. Schmidt and U. Stadtmüller (2006), 'Nonparametric estimation of tail dependence', *Scandinavian Journal of Statistics* **33**, pp. 307–335.

J.R. Schott (1997), *Matrix Analysis for Statistics*, John Wiley.

W.F. Sharpe (1963), 'A simplified model for portfolio analysis', *Management Science* **9**, pp. 277–293.

J.H. Shih and T.A. Louis (1995), 'Inferences on the association parameter in copula models for bivariate survival data', *Biometrics* **51**, pp. 1384–1399.

M. Sibuya (1960), 'Bivariate extreme statistics', *Annals of the Institute of Statistical Mathematics* **11**, pp. 195–210.

P. Silvapulle and C.W.J. Granger (2001), 'Large returns, conditional correlation and portfolio diversification: a value-at-risk approach', *Quantitative Finance* **1**, pp. 542–551.

S. Sirkiä, S. Taskinen, and H. Oja (2007), 'Symmetrized M-estimators of multivariate scatter', *Journal of Multivariate Analysis* **98**, pp. 1611–1629.

A. Sklar (1959), 'Fonctions de répartition à n dimensions et leurs marges', Research report, University of Paris, Institute of Statistics, France.

R.L. Smith, J.A. Tawn, and H.K. Yuen (1990), 'Statistics of multivariate extremes', *International Statistical Review* **58**, pp. 47–58.

M.S. Srivastava and M. Bilodeau (1989), 'Stein estimation under elliptical distributions', *Journal of Multivariate Analysis* **28**, pp. 247–259.

C. Stein (1956), 'Inadmissability of the usual estimator for the mean of a multivariate normal distribution', in: 'Proceedings of the 3rd Berkeley Symposium on Probability and Statistics', volume 1, pp. 197–206.

S. Taskinen, C. Croux, A. Kankainen, et al. (2006), 'Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices', *Journal of Multivariate Analysis* **97**, pp. 1219–1243.

K.S. Tatsuoka and D.E. Tyler (2000), 'On the uniqueness of S-functionals and M-functionals under nonelliptical distributions', *Annals of Statistics* **28**, pp. 1219–1243.

J.A. Tawn (1988), 'Bivariate extreme value theory: models and estimation', *Biometrika* **75**, pp. 397–415.

J. Tobin (1958), 'Liquidity preferences as behavior towards risk', *Review of Economic Studies* **25**, pp. 325–333.

D.E. Tyler (1982), 'Radial estimates and the test for sphericity', *Biometrika* **69**, pp. 429–436.

D.E. Tyler (1983), 'Robustness and efficiency properties of scatter matrices', *Biometrika* **70**, pp. 411–420.

D.E. Tyler (1987a), 'A distribution-free M-estimator of multivariate scatter', *Annals of Statistics* **15**, pp. 234–251.

D.E. Tyler (1987b), 'Statistical analysis for the angular central Gaussian distribution on the sphere', *Biometrika* **74**, pp. 579–589.

D.E. Tyler (2002), 'High breakdown point multivariate M-estimation', *Estadística* **54**, pp. 213–247.

A.W. van der Vaart (1998), *Asymptotic Statistics*, Cambridge University Press.

B. Vaz de Melo Mendes (2005), 'Asymmetric extreme interdependence in emerging equity markets', *Applied Stochastic Models in Business and Industry* **21**, pp. 483–498.

S. Visuri (2001), *Array and Multichannel Signal Processing Using Nonparametric Statistics*, Ph.D. thesis, Helsinki University of Technology, Signal Processing Laboratory, Finland.

W. Wang and M.T. Wells (2000), 'Model selection and semiparametric inference for bivariate failure-time data', *Journal of the American Statistical Association* **95**, pp. 62–76.

F.A. Wolak (1987), 'An exact test for multiple inequality and equality constraints in the linear regression model', *Journal of the American Statistical Association* **399**, pp. 782–793.

Y. Zuo (2006), 'Robust location and scatter estimators in multivariate analysis', in: J. Fan and H.L. Koul, eds., 'Frontiers of Statistics (in honor of Professor P.J. Bickel's 65th birthday)', pp. 467–490, Imperial College.